Data Science for Entrepreneurship Research: Studying Demand Dynamics for Entrepreneurial Skills in the Netherlands

Jens Prüfer and Patricia Prüfer[1]

31 January 2019

The recent rise of big data and artificial intelligence (AI) is changing markets, politics, organizations, and societies. It also affects the domain of research. Supported by new statistical methods that rely on computational power and computer science --- data science methods --- we are now able to analyze data sets that can be huge, multidimensional, unstructured, and are diversely sourced. In this paper, we describe the most prominent data science methods suitable for entrepreneurship research and provide links to literature and Internet resources for self-starters. We survey how data science methods have been applied in the entrepreneurship research literature. As a showcase of data science techniques, based on a dataset of 95% of all job vacancies in the Netherlands over a 6-year period with 7.7 million data points, we provide an original analysis of the demand dynamics for entrepreneurial skills in the Netherlands. We show which entrepreneurial skills are particularly important for which type of profession. Moreover, we find that demand for both entrepreneurial and digital skills has increased for managerial positions, but not for others. We also find that entrepreneurial skills were significantly more demanded than digital skills over the entire period 2012-2017 and that the absolute importance of entrepreneurial skills has even increased more than digital skills for managers, despite the impact of datafication on the labor market. We conclude that further studies of entrepreneurial skills in the general population --- outside the domain of entrepreneurs --- is a rewarding subject for future research.

Keywords: Data science, Machine learning, Entrepreneurship, Entrepreneurial skills, Big Data, Artificial Intelligence

JEL-codes: L26, C50, C55, C87, O32

1. Introduction

We are drowning in data. 90 percent of the world's data today has been created in the last two years alone.[2] Most of it is unstructured text, images, and videos, which is hard to categorize, let alone understand, for human beings.[3] There are sensor data in (self-driving) cars, smart home and office equipment, social media data, mobile data, data on Internet and browsing behavior, or digital camera images, to name just a few. This explosion of data is accompanied by tremendous progress in data science methods, which can make sense of all the available information. Those methods,

[1] J. Prüfer: Department of Economics, CentER, TILEC, Tilburg University; j.prufer@uvt.nl. P. Prüfer: CentERdata, Tilburg University; p.prufer@uvt.nl. Both: P.O. Box 90153, 5000 LE Tilburg, The Netherlands. We are grateful to Freek van Gils, George Knox, and Marcia den Uijl for comments on an earlier draft and to Pradeep Kumar and Chayanin Wipusanawan for valuable research assistance. All errors are our own.

[2] https://public.dhe.ibm.com/common/ssi/ecm/wr/en/wrl12345usen/watson-customer-engagement-watson-marketing-wr-other-papers-and-reports-wrl12345usen-20170719.pdf

[3] https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/

are fueled by artificial intelligence (AI). And this may just be the beginning (Taddy, 2018). The McKinsey Global Institute recently projected that the adoption of AI by firms may follow an S-curve pattern --- a slow start given the investment associated with learning and deploying the technology, and then acceleration driven by competition and improvements in complementary capabilities (Bughin et al., 2018). At the macro level, they expect that AI could potentially deliver additional economic output of around U$13 trillion by 2030, boosting global GDP by about 1.2 percent a year. The increased output from efficiency gains and innovations could be passed to workers in the form of wages and to entrepreneurs and firms in the form of profits.[4]

These rapid and ongoing changes in the economic, political, and social spheres also affect the domain of research. Massively improved AI offers better, i.e. cheaper, prediction (Agrawal et al., 2018). Improved prediction capabilities allow us to work with huge data sets that are representative for entire populations, simply because they contain nearly complete data on that population (see Section 4 for an example). Even more, the statistical methods relying on AI --- data science methods --- allow us to tackle novel types of questions, such as: How to study the role of geographic and social proximity for entrepreneurial interactions by using huge social media data sets, i.e. from Twitter, instead of traditional case studies? How can we classify the personalities of more than 1000 CEOs, identify the entrepreneurial ones, and study to what extent being entrepreneurial has positive effects on firm performance? To what degree are entrepreneurial skills and personality traits helpful for workers in all kinds of sectors and jobs? Crucially, these questions --- if they could be asked at all --- could not be seriously studied, let alone be answered, by traditional empirical methods that have been taught in graduate schools in economics and management in the past decades.

In their seminal article, Shane and Venkataraman (2000) defined entrepreneurship as the identification, evaluation, and exploitation of opportunities. Shane (2012) underlined that entrepreneurship is a process, not a one-time event. The questions listed above relate to opportunities initiated by very recent technological progress, which by itself is an ongoing process. Thus, the very object of entrepreneurship research changes along the development of the technological frontier. Today, due to the availability of much more data and computer power, this frontier is shaped strongly by the state of data science techniques. Today, we can analyze and interpret large amounts of complex and unstructured data and make predictions based on correlations and inductive modeling.

Researchers can benefit by understanding and --- where appropriate --- embracing statistical methods that are driven by AI algorithms. This process has already started and has had disruptive effects on the social sciences, such as economics (Einav and Levin, 2014) and management (George, Haas, & Pentland, 2014). It has created the new field of *computational social science,*

[4] Note, however, that in practice this redistribution of profits is often not occurring. Profits are accumulated and stay mostly in the top ‘superstar firms‘ (Mayer-Schönberger and Ramge, 2018).

which may reveal new patterns of individual and group behavior and allow to model economic and social interactions more precisely (Lazer et al, 2009).

We contribute to the entrepreneurship literature in two dimensions. First, in the next section, we describe the most prominent data science methods suitable for entrepreneurship research. The goal is to give the interested reader a concise overview over what is possible technically today, with enough input and references to start educating oneself. Section 2 is complemented by the Appendix, where we provide links to literature and Internet resources and where we also delineate key technical terms and list the most relevant text mining tools and download resources for self-starters. Our second contribution comes in Sections 3 and 4. Section 3 surveys how data science methods have been applied in the entrepreneurship research literature and sketch how they have been used to study important research questions that could not --- or not to the same extent --- be studied without these techniques. Along these lines, in Section 4 we provide an original analysis of a data set with 7.7 million data points and study the dynamics of demand for entrepreneurial skills in the Dutch population. In Section 5, we conclude by discussing opportunities and risks of data science techniques and relate them to traditional empirical research methods and theory.

2. Data science methods for entrepreneurship research

Background

In conventional statistical research, you start with the formulation and testing of hypotheses with the help of data, assuming that the data are generated by a given stochastic data model. In data science, by contrast, you churn large volumes of data looking for patterns by using algorithmic models and treating the data mechanism as unknown.[5] Thus, data science "not only provides new tools, it solves a different problem" (Mullainathan and Spiess, 2017, p.88) and is able to discover complex structures that were not specified in advance (Breiman, 2001). In other words, whereas conventional statistics is deductive, data science is inductive: the approaches are complementary.

Data science relies heavily on computational power and computer science to derive knowledge from the unprecedented, exponentially growing, complex and unstructured data, so-called "big data." By making software autonomous or using iterative feedback to discover associations in data, we can find generalizable patterns and anomalies. Thus, instead of teaching machines to do things, the goal of data science is to design them to "think" for themselves and then allow them access to the mass of available data so they could learn. Moreover, while the human brain can associate two or three dimensions of information with each other, algorithms allow hundreds of dimensions. This leads to a system searching for much more fine-grained associations, clusters, and classifications,

[5] This section mildly overlaps with one section in Prüfer and Prüfer (2018). That paper, however, is significantly shorter and focuses on institutional economics, not entrepreneurship research and the dynamics of entrepreneurial skills.

extracting meaningful information from the data. As a next step, an understandable structure can be developed to facilitate data-driven decision making.

Due to the nature of 'big data' and the complexity of the algorithms used, data science often requires special ways of data storage, accessibility, and processing. Analyses are often done by using multiple computers and multiple calculation units, so-called "high-performance computing," for instance, Hadoop clusters and Spark-Streaming, or parallel virtual environments.[6] Usually the basic steps for analysis include writing an algorithm, setting up an automated process (script), and linking it with open data protocols and Application Programming Interfaces (APIs). Collecting large amounts of unstructured information often generates a complex information set. With the help of visualization techniques and tools, such as chord charts and network graphs, we can observe clusters within that information and present results of data analyses.

Of course, traditional data sources such as surveys and large administrative data sets (old data) can be analyzed and interpreted with the help of data science techniques, too. The computational power of these techniques allows for a much broader and varied search on existing data, which may lead to the revelation of new patterns and insights even in traditional data sources. A notable example is the use of machine learning techniques on the huge United States Patent and Trademark Office (USPTO) database. Various papers have shown that these methods can improve inventor disambiguation from this database and, thereby, help to add a more accurate understanding of inventor careers (Li et al., 2014; Ventura et al., 2015). Machine learning algorithms cannot only match patents more correctly to inventors, they can also include more information from other useful data sources, for example co-authorships, collaboration variables, and geographic location. Based on this information, "large-scale innovation studies across time and space with visualization of inventor mobility across the United States" (Li et al., 2014, p. 941) are possible with much lower error rates than before. Similarly, disambiguation approaches based on machine learning are more consistent across contexts as they can cope better with varying features and detect the best features automatically and more precisely (Ventura et al., 2015).

Key data science methods

A multitude of different tools and techniques are available, of which we highlight the most interesting ones for entrepreneurship research. In general, *Python,* currently the fastest growing (general purpose) programming language, features a large range of very effective scripts and open source libraries for these tasks.

Machine Learning

Within the field of data science, machine learning (ML) is an advanced field of research dealing with the techniques that teach computers to learn without being programmed explicitly (Samuel, 1959). ML is not a synonym for AI, though; it is technically a branch of AI. AI, in fact, is a much

[6] See Box A1 in the Appendix for detailed explanations of these, and more, technical terms.

broader concept, in which machines mimic cognitive functions of learning and problem solving. Therefore, AI algorithms and machines are able to adapt to different situations and to carry out tasks in a way that we would consider "smart" or "intelligent", that is, with human-like cognitive functions (OECD, 2017, Taddy, 2018).

Within ML, the two most important categories are *supervised learning* and *unsupervised learning*. *Supervised ML* is the name of a set of advanced algorithms that use information from known results, so-called *labels*, to optimize predictions. Technically, in a *supervised* learning task a computer learns a relation between some observed input (usually a vector of many predictors) and some desired output (one outcome variable of interest) (Hastie et al., 2009). A supervised learning algorithm analyzes the labeled training data and produces an inferred function to map novel (test) data. Supervised learning helps to predict unseen patterns and to understand which input best predicts the outcome to assess the quality of previously tested predictions/inferences. Therefore, it also serves to reduce the "curse of dimensionality", for example by using an algorithm for dimensionality reduction such as *Principal Component Analysis,* where variables that are meaningless in explaining a desired target variable or are possibly correlated, are eliminated by the statistical procedure of orthogonal transformation.

Depending on the type of data, one can choose from *regression* and *classification* techniques within supervised ML. If one has to predict continuous values, regression techniques are the way to go, while classification techniques are used in discrete settings; they identify which set of categories (classes) a new observation belongs to. An easy to interpret and widely used classification method is a *decision tree*. Starting from the root, the training observations are split up as heterogeneously as possible into two subgroups. At each node, the algorithm examines which variable it can best split into two new nodes. In this way, the data is split up further and further, until a *stop criterion* is met (for example, less than n training observations per node). Depending on the values of the variables, each observation ultimately falls into one class (i.e. a single leaf). The results of a decision tree can be interpreted and graphically displayed relatively easily. However, decision trees are prone to instability: a relatively small change in the data can result in another tree. Thereby, a decision tree has a large "generalization error," a phenomenon that is also called "overfitting the data," meaning that it can contain nodes that have been created by specific cases in the training data set, making the model poorly generalizable to other data. This means that the model can only perfectly rationalize a specific outcome based on the given training data but is not able to predict variants that were not used for training.[7]

[7] One way to overcome instability is to use *ensemble methods*, such as a *Random Forest* (RF). This is a tree-based supervised learning technique, in which a large number of decision trees are combined to arrive at the final prediction. Thereby, the method is more stable than a single decision tree. If the target variable that we want to predict is categorical, the final outcome is determined by means of 'majority voting'. In other words: the outcome of most trees is considered to be the final outcome. The collection of trees is called random, because each tree is trained on a random selection of variables and observations. When multiple models are combined in a large model, we speak of an ensemble model. Combining many loose decision trees into an ensemble model results in higher precision and in more stable predictions. Therefore, a RF

Deep learning (DL) is a special class of supervised learning algorithms that is frequently used for feature extraction from complex, multidimensional data such as images. For instance, Google uses DL to automatically suggest the next word(s) of a search term when one has started typing a word. DL uses so-called *(artificial) neural networks*, which allow computers to more closely mimic human brains while still being faster, more accurate and less biased. Neural networks are especially suited for deriving patterns from (highly) non-linear processes. Depending on the form of the model underlying the DL algorithm, a neural network falls either within the category of *supervised learning* or within unsupervised learning, which we will explain below.

In a pioneering example, Tan and Koh (1996) trained a neural network based on information from psychological, demographic, and family characteristics to predict entrepreneurial inclination. Results from a survey administered among 200 business undergraduates served as training and testing data to model entrepreneurial inclination in an individual. Then, the neural network predicted inclination in any other person based on knowledge of the imputed social and psychological correlates. In this early case, the ML algorithm had an accuracy of 80% for predicting entrepreneurial inclination in individuals not encountered before.

Machine learning can also be performed unsupervised. Then it is used to learn and establish baseline profiles for different entities. In *unsupervised ML,* 'natural' groups or *clusters* of observations are made, whereby observations that are 'equal' or ‘close’ to each other, belong to the same group. This allows trends and patterns in data to be properly mapped out, for instance, when customers have to be grouped into different segments based on their characteristics so that services can be tailored individually (Alsayat and El-Sayed, 2016). In this type of *cluster analysis* it is necessary to optimize the number of clusters and to thoroughly investigate the stability of the clusters. The latter can be done by adding noise or using multiple algorithms to check whether a certain change in data gives rise to a new cluster.

For the clustering, a distance metric is often used as (in)equality score. This can be the Euclidean distance or another distance function. Whereas most distances can only be used for numerical and complete data, the *Gower's remote function* can deal with both categorical and missing data. The more data there is, the more computationally expensive the choice of an algorithm that minimizes distance. Frequently used algorithms are *K-means* and *K-modes* which work with centroids as distance measures and are, thereby, less computationally expensive in terms of the best distance metric.

A very early example of unsupervised learning using a neural network for clustering is Rutherford et al. (2001). This paper uses a so-called *self-organizing map (SOM)* approach to study the relation of firm size with firm success and survival. Using information from the National Survey of Small Business Finances (NSSBF), Rutherford et al. (2001) classify small firms (having less than 500

generally provides much better predictions than a decision tree. Other well-known techniques are *Gradient Boosting* and *Support Vector Machines (SVM)*. More information on these (and other) techniques, see Hastie et al. (2009) or Provost and Fawcett (2013). There you can also find a discussion on *performance metrics* to test which of the available models works best for a given dataset and research question.

employees) into multiple groups based on size and ownership as well as firm characteristics. The 4,637 small firms in their sample cluster naturally into two distinct groups: a larger group with 3,311 members of very small firms and a smaller group with 1,326 members of larger (but still small) firms. Given that these two groups differ significantly on other background characteristics, this early paper provides evidence that differences in firm size and structures matter to predict antecedents of firm survival and success.

Yet another category is *reinforcement learning* (RL), which differs from standard supervised learning because correct input/output pairs are never presented, nor sub-optimal actions explicitly corrected. Thus, in reinforcement learning, there is no answer. Instead, the reinforcement agent decides how to perform the given task. The only training data given as feedback to the algorithm is in the form of rewards and punishments. In the absence of training data, it is bound to learn from its experience. Calvano et al. (2018) use RL to experiment with AI pricing agents interacting repeatedly in a controlled environment (computer-simulated marketplaces). Their algorithmic price setting experiments shows that when replacing human decision-making even relatively simple pricing algorithms systematically learn to play sophisticated collusive strategies without communicating with each other at all.

Text analytics and web data scraping

In addition, data science methods are also suitable for obtaining information from unstructured data, often scraped from the Internet. This is very useful because about 80% of big data is available in unstructured text form, for example in blogs, websites, and social media (Cogburn and Hine, 2017). This way, all data sources that relate to natural language can be used, such as open answers, text files, notes from customer contacts, reports or e-mails. There are several useful tools and techniques for handling text, semantic, and social data to extract valuable information from these sources. Here, we describe how and what we can infer from the data and discuss useful techniques for mining and analyzing text data to discover interesting patterns, extract useful knowledge, and support decision making. Even more information can be found in section 4.

In addition, Internet log files and the metadata of search engines can provide interesting information about trends over time. Search engines register when and where a search query was performed in their search logs and process this information for the answers provided to subsequent related search queries.[8] The numbers of searches on certain topics and the presented order of search results often show interesting patterns, which Google Trends makes use of, for instance. In a recent book, Stephens-Davidowitz (2017) presents research that uses different kinds of Internet data: Google Trends, online search data, information on views and clicks, and even patterns of swipes in mobile apps. A famous example is what happened after Facebook introduced the "News Feed" in 2006. With this function, users would get automated updates of the activities of all their friends. It provoked immediate fierce protests of nearly a million users but Facebook did not remove the

[8] The economic consequences of this usage are studied in Prüfer and Schottmüller (2017).

News Feed. The company had what Stephenson-Davidowitz calls the "digital truth serum" (2017, p.154): numbers on clicks and visits increased tremendously after the introduction of the News Feed. In his book, the author provides many more examples on how to use Internet data to derive new insights in human nature and behavior, especially for sensitive issues such as sexual orientation, sexism, customers' revealed preferences, and stereotypes.

Mining, clustering, and analyzing these unstructured data sources requires the use of analytical techniques for natural language. This so-called *Natural Language Processing* (NLP) can be performed in different programming languages, for example, Python or R, and researchers can use well-established packages and toolboxes. *Sentiment analysis*, for instance, can extract subjective information from language, while *topic modelling* can discover the abstract 'topics' in a collection of documents. Other techniques, such as *Named Entity Recognition* (NER) or *Part-of-Speech* (POS) *tagging,* recognize entities such as organizations, people, locations, dates, time, or currency (NER) or word types such as verb, noun, etc. (POS) in text. Box A2 in the Appendix lists the most common concepts and tools in general and Section 4 exemplifies the steps one has to take when working with text data from online sources.

3. Applying data science to entrepreneurship research

In this section, we highlight some recent papers using data science methods for research questions on various aspects of entrepreneurial characteristics, processes, and entrepreneurship (success). ML and/or text analytics have been applied to issues such as funding (via venture capital and via crowdfunding), (product) innovation, inventors' disambiguation, and entrepreneurial traits. The fundamental contributions of these studies fall in two categories: the utilization of new sources of information and data, advancing the *data frontier*; and applying novel techniques to existing data and/or problems, thereby advancing the *knowledge frontier*. We now take them in order.

A classical question in entrepreneurship research relates to the factors predicting a startups' success (Stuart and Abetti, 1987, Hisrich et al., 2007). Today, the role of online and social media communication and information for the development, identification, and success of entrepreneurial activities and agents has received a lot of attention. To arrive at deeper, richer, and more fine-grained insights on the entrepreneurial mindset, so-called *digital footprints* from social media are increasingly used. Lee et al. (2017) measure overconfidence of CEOs by classifying their messages sent on Twitter. They distinguish 'professional CEOs' and 'founder CEOs' and find that the latter use more optimistic language on Twitter and during earnings conference calls. Founder CEOs are also more likely to issue earnings forecasts that are too high.

Aggarwal and Singh (2013) show that social media can also be used as a means to an end for entrepreneurial success. They study company blogs across multiple stages of venture capitalists' decision making and find that blogging can help managers in getting their products and services selected at the screening stage, but that, beyond that, blogging does not help directly. The authors show that blogs can help indirectly in the last stage of the venture capital process when negotiating

a contract with the venture capitalist: blogs (with good coverage) attract the attention of competing venture capitalists, which drives up venture prices, and hence improves the blogger's outside option.

Since the success of the managerial 'upper echelons' perspective, it is rather undisputed that the individual characteristics and values of decision makers have a significant impact on the performance of firms (Andrews 1980, Hambrick and Mason 1984). A key question is how certain managerial characteristics translate into better performance. A popular empirical approach to this question has been to measure actual behavior of decision makers through real-time personal observation (Mintzberg, 1973). This time-consuming procedure, however, creates the problem of small sample sizes and suffers from selection issues.

Bandiera et al. (2017) tackle the issue by developing new methodology: First, via daily phone calls with 1,114 CEOs or their assistants, they collected 42,233 data points about the decision makers' diaries. Then they employed an unsupervised learning algorithm (a latent Dirichlet allocation, LDA), which provides them with a complete probabilistic description of time-use patterns, despite the high dimensionality of their data set. The algorithm posits that the actual behavior of each CEO is a mixture of a small number of "pure" behaviors, and that the creation of each activity is attributable to one of these pure behaviors. In their case, the algorithm finds two "pure" behaviors and generates a one-dimensional behavior index that represents a CEO as a convex combination of the two pure behaviors. Following Kotter (1999), they classify the first pure behavior as "manager" and the second as "leader:" "manager" refers to more time of the CEO spent in meetings with production-level workers and one-to-one meetings with firm employees or suppliers; "leader" refers to more time spent with top-executives and in interactions with several participants and functions from inside and outside the firm together. Kotter associated "managers" with a focus on monitoring and implementation tasks, whereas 'leaders focus on the creation of organizational alignment and communication across a broad variety of characteristics. Clearly, the characterization of "leaders" is related to the characterization of entrepreneurial skills in the entrepreneurship literature (see Section 4).

As a final step, Bandiera et al. (2017) correlate their managerial behavior index with firms' balance sheet data and find that "leader" CEOs are more likely to be found in larger and more productive firms: an increase of the behavior index by one standard deviation is associated with an increase of 7% in sales, controlling for a battery of factors. This not only suggests that decision makers with entrepreneurial characteristics can also do well in more established organizations. More important for the study at hand, Bandiera et al. (2017) show an innovative way how to use data science techniques to give a more robust answer to an existing question involving personal characteristics of decision makers. There is a lot of scope to apply this to a host of questions in the entrepreneurship literature.

Obschonka et al. (2017a) use Twitter data to identify the personality traits of superstar entrepreneurs and compare them to the characteristics of superstar managers, "a hitherto understudied population in entrepreneurship research" (p.14). To do this, the authors use a sample

of 106 Twitter accounts of (superstar) entrepreneurs and managers. They analyze information from these accounts by using a novel language-based personality assessment tool that is capable of dealing with the huge number of observations from social media data.[9] Up to now, traditional, survey-based methods, such as a standard Big Five questionnaire, have been used to assess an individual's personality traits. In contrast to these subjective and self-reported measures, digital footprints, where individuals willingly and unwillingly spread (personal) information to a large and diverse audience, can be used to derive objective and accurate information, revealing individuals' true preferences. Obschonka et al. (2017a) show that this new tool delivers valid results for univariate and multivariate analyses of personality differences between (superstar) entrepreneurs and (superstar) managers and that, surprisingly and contrary to earlier findings, the latter category shows more entrepreneurial characteristics than the former one.[10]

Tata et al. (2017) use Twitter data to arrive at the 'psycholinguistics of entrepreneurship' and demonstrate that even though entrepreneurs are fundamentally different from the general population, also the organizational life cycle matters for the emotions and sentiments attached to entrepreneurship and to the work-life balance in general. The use of language as a robust means for revealing individuals' (work-life) concerns, motives, traits, and emotions is not new to the field. For instance, Tausczik and Pennebaker (2010) have shown that language is a robust means for revealing individuals' work-life concerns and emotions. "Entrepreneurial emotion" is a topic in itself and describes a package of feelings that often come with being an entrepreneur (Cardon et al., 2012), a topic that has gained increased importance through big data and AI as enablers of new self-employed businesses: "Approximately 150 million workers in North America and Western Europe have left the relatively stable confines of organizational life --- sometimes by choice, sometimes not --- to work as independent contractors" (Petriglieri et al., 2018). However, by using Twitter data for these analyses, Tata et al. (2017) are able to overcome several limitations of traditional data sources, such as surveys. Social media data can not only avoid response and recall biases; they also offer a real-time window into peoples' thoughts over long periods, for more actors than any existing alternative, at any point in time, and across diverse geographical locations. Moreover, content analysis of Twitter data allows collecting information on emotions, constructs, and concerns simultaneously.

Wang et al. (2017) use Twitter data for yet another type of entrepreneurship research. They apply social network analysis to entrepreneurial networks in the US to identify and locate entrepreneurs jointly with important regional subtleties within the network. They find that although Twitter enables interactions across geographically (and socially) distant locations, the highest intensity can

[9] The tool, *Receptiviti*, is used for top-down language analysis along a large number of psychological metrics. Receptiviti is the commercial variant of the Linguistic Inquiry and Word Count (LIWC) text analysis platform, which allows for an assessment of language and text for psychological purposes in more than 80 languages. See Box A2 in the appendix for more details.

[10] In another interesting paper, Obschonka and Fisch (2017) use a NLP approach on Twitter data to analyze whether entrepreneurial personalities are increasingly more numerous and more influential in political leadership. They test the underlying hypothesis, that an entrepreneurial personality benefits from the rise of the 'entrepreneurial society,' on US President Donald J. Trump, who was an entrepreneur before.

be detected in regional interactions characterized by similar socioeconomic and demographic profiles. This suggests that, even in our digitally connected world, geographic and social proximity are important for entrepreneurial interactions. Hence, earlier results about the important role of social relationships for entrepreneurship are still valid. See the work of Olav Sorenson (Rickne et al., 2018). For instance, Sorenson (2018) shows that both professional and private social relationships are original reasons for industry concentrations in a small number of places, even when firms do not benefit from this clustering. Wang et al. (2017) extend the research on the relevance of networks in various ways: they simultaneously examine the types of actors engaged in digital networks and the specific regions that are active on the Twitter entrepreneurship domain. Moreover, they analyze the regional characteristics that explain the intensity of activity on this social media platform. The use of big data allows for social network analyses on a much larger scale than when using data from primary survey collection efforts. Thereby, Wang et al. (2017) offer a seminal example of bringing data science into the entrepreneurial social networks literature that has mostly been dominated by case studies (Greve and Salaff, 2003). They also demonstrate how the incorporation of additional quantitative and qualitative information can mitigate issues of representativeness inherent in social media data.

Data science methods have also been applied to assess performance of crowdsourcing and crowdfunding platforms. Crowdsourcing taps into a crowd with talent to get a whole bunch of business projects or tasks solved by dividing them into microtasks. These microprojects assigned to a skilled on-demand workforce provide many business opportunities, but can also deliver interesting research questions. Crowdfunding, on the other hand, is a unique form of entrepreneurial finance that combines elements of private and public equity (Cummings et al., 2019). It taps into the power of the crowd to acquire financial support. Platforms match individuals or entities in need of funding with individuals or groups willing to contribute financially, often in the form of microfunding.[11] Apart from being interesting sources of big data, these platforms themselves can be viewed as big data and datafication phenomena as a result of the ongoing digitization and automation. Both innovative developments use mass collaboration, mostly via online tools, to accomplish certain goals, for example the funding of an idea or a project. There is serious money involved in crowdfunding[12] and, therefore, reliable predictions on the success rates of these products or projects are important. Obviously, big data and data science methods play an

[11] The idea of crowdfunding also exists in the financing of research, amongst others due to high rejection rates prevalent in (scientific) research. Various scientific crowdfunding platforms emerged, for example *Experiment.com* (https://experiment.com/) or *SciFund Challenge* (https://scifundchallenge.org/). Vachelard et al. (2016) wrote: "Crowdfunding represents an attractive new option for funding research projects, especially for students and early-career scientists or in the absence of governmental aid in some countries. The number of successful science-related crowdfunding campaigns is growing, which demonstrates the public's willingness to support and participate in scientific projects" (p. 1).

[12] According to the *Crowdfunding Industry Report*, global crowdfunding was expected to reach $ 34,4 billion in 2015 (http://crowdexpert.com/crowdfunding-industry-statistics/), while the *Crowdfund Campus* reports that "in 2016, equity raised from crowdfunding passed VC funding for the first time, and, by 2025, the World Bank Report estimates that global investment through crowdfunding will reach $93 billion." (https://crowdfundcampus.com/blog/2017/01/crowdfunding-in-2017-three-key-trends/).

important role also for the internal business processes at a crowdfunding or crowdsourcing platform. With social media promotions, statistics on earlier projects, market dynamics, and other activities, a huge amount of data is generated, which can predict the success of ideas or products based on past analytics results.[13]

Hoornaert et al. (2017) build a ML model to predict the success and failure of business and product ideas generated within the crowd based on 3C's: its *content*, the *contributor* proposing it, and the *crowd's* feedback on the idea. A nonlinear, supervised algorithm identifies the variables that are most predictive of an idea's distinctiveness and successful implementation. The authors find that considering immediately available information about the content and contributor improves the ranking performance by around 25% over random idea selection, while adding crowd-related information that accumulates over time further improves performance by nearly up to 50%. The last C, crowd feedback is, thus, the best predictor, but also the one that needs most time to develop.

Courtney et al. (2017) use data from *Kickstarter*, a large and popular crowdfunding portal. They examine the interplay of three signal types obtained from different sources within the platform on the viability of a certain idea: the direct actions a startup takes regarding a proposed idea and/or product (the content), its characteristics (mainly crowdfunding experience; the contributor), as well as third-party endorsements (sentiments expressed in backer comments; the crowd). For the last type of signal, the authors implement a novel sentiment analysis technique, with which the underlying tone of textual comments by backers can be derived. This allows for a continuous feedback measure of a large and heterogeneous group of individuals commenting on a project, a major improvement on the dichotomous variable that is usually used to measure third-party endorsements (Courtney et al., 2017).

On a higher level, Hartmann et al. (2016) connect data science and entrepreneurship. They derive a taxonomy of business models used by start-up firms that rely on data as a key resource for business, which they call data-driven business models. Their taxonomy consists of six different types of such business models among start-ups and thereby develops a basis for understanding how start-ups build business models that capture value from data as a key resource.

Whereas the above cited papers use new (big) data sources, Hoberg and Phillips (2016) apply a novel technique, text analysis, to study an existing administrative database.[14] They use the product descriptions that firms filed with the US Securities and Exchange Commission (SEC) to develop new time-varying industry classifications. These new, more flexible measures of industry

[13] On the other hand, even data collection itself can be crowdsourced. This method originates from the scientific world where the first known case of crowdsourcing, taking place around 150 years earlier than Wikipedia, is the Oxford English Dictionary. The aim of this dictionary, to list all the words known in the English language with their definition and explanation of usage, could only be reached after people all over the world contributed to it (https://dictionarylab.stanford.edu/crowdsourcing-oed). Another great example of crowdsourcing in practice is *OpenStreetMap*—an alternative to *GoogleMaps* launched in 2004. Since then, more than 1 million mappers have worked together to collect and supply data (https://www.openstreetmap.org/).

[14] Li et al. (2014) and Ventura et al. (2015), discussed in Section 3, fall in the same category as Hoberg and Phillips (2016).

membership are better suited to explain differences in key characteristics across industries, such as profitability, sales growth, and market risk. Information on the text-based network classification is also informative about identifying rival firms. Moreover, these classifications show endogenously how industries and their competitors change due to external shocks and how R&D activities and advertisement are endogenously adjusted to the behavior of relevant competitors. Hoberg and Phillips (2016) combine two central ideas: first, that the product features and bundles a firm offers can be consistently derived from SEC product descriptions, and that these descriptions can be used to assign a spatial location based on product descriptions, generating a Hotelling-like product location space for these firms. Second, this study uses text analysis to build a network of firms, in which the similarity of each firm to every other firm is calculated by firm-by-firm pairwise word similarity scores using the original product descriptions. Based on these pairwise similarity scores, firms are grouped into industries and the general industry classification can be interpreted as an unrestricted network of firms. There, a firm's competitors are analogous to a group of friends on social media, with each firm having its own distinctive set of competitors.

4. A case in point: using NLP to study dynamics of entrepreneurial skill demand in a large population

Complementing the (admittedly selective) broad literature review in the previous section, now we go into some depth. To exemplify the general statements made above, we offer an original analysis of a novel big data set by using various data science methods. Specifically, we study the consequences of the ongoing technological and economic developments on the demand for entrepreneurial skills.

Developing entrepreneurial skills is increasingly seen as important to foster entrepreneurship (Baumol et al., 2007). Several recent articles picked up the call and approached the topic from several disciplinary and methodological angles.[15] As the purpose of this section is not to review the literature on entrepreneurial skills (for that see the cited articles) but to exemplify data science methods, we restrict our notion to two observations. First, there is no generally accepted delineation, let alone definition, of "entrepreneurial skills." Second, one of the restrictions of existing studies using traditional empirical methods is the small number of available data points: the cited articles report sample sizes of 39, 523, and 1126 subjects, respectively. Consequently, it is hard to draw general, robust lessons that can be applied to different contexts than those studied. An additional characteristic of these articles, which is interrelated with sample size, is that they focus on (would-be) entrepreneurs, which does not allow to make statements about the importance of and demand for entrepreneurial skills in the general population. Here, we try to alleviate these constraints.

[15] These include RezaeiZadeh (2016), Obschonka et al. (2017b), and Rosique-Blasco et al. (2018).

The starting point is that digitization, automation, and the development of new (adaptable) technologies have an increasing impact on the labor market. The boundary between 'ICT jobs' and other professions in which ICT-related skills are required is becoming increasingly blurred. Moreover, the specific skills demanded and the tasks that have to be fulfilled in all occupations have changed considerably in recent years (Spitz-Oener, 2006).

These changes have led to increased demand for employees with sufficient digital skills in many countries, including the Netherlands (ROA, 2017). For employees who, in the longer term, cannot acquire the necessary digital skills through training and retraining, suitable measures and career development paths are required that avoid insufficient qualifications and, eventually, unemployment. In contrast, research has shown that employees can adapt sufficiently to the changes on the labor market and that the negative effects of digitization and automation might be exaggerated, as many jobs may change but also new jobs will be created (Autor, 2015; Arntz et al., 2016). Given that innovative capacity is directly related to economic growth, a lack of people with sufficient digital, technical, and ICT skills, in combination with a broader set of so-called 21st century skills, limits innovative capacity (Obschonka et al., 2017b). Alternatively, abundance of these types of employees helps to mitigate the negative effects on innovative capacity and the labor market (Elliott, 2017; McAfee and Brynjolfsson, 2017).

What are the dynamics in skill demand on the labor market? What are the consequences for different occupations and for employees with different educational backgrounds and different levels of expertise? How do they affect certain types of professions such as managers, ICT professionals, and employees in non-IT/technical jobs? Prüfer et al. (2019) answer these questions by making use of a novel approach of 'labor market analytics' in which information from online vacancies, thus from unstructured (big) Internet data, is combined with information from labor market forecasts, that is, with structured data from administrative sources.[16] Thereby, an innovative and very rich source of information, as well as a unique dataset is created with which the authors analyze the impact of digitization and automation on the labor market in general, on specific economic sectors, on 371 different occupations, and on 3 types of professions. Prüfer et al. (2019) measure the change in skills requirements over time by taking into account digital, technical and ICT skills compared to general cognitive and non-cognitive skills.

In an original extension to Prüfer et al. (2019), the current section derives insights on the consequences of the ongoing digitization and automatization on the dynamics of *entrepreneurial skills*. We distinguish three types of professions: managers, ICT jobs, and non-ICT jobs. This helps to understand how the requirements have changed over time and among types of professions and, thus, not only provides insights into ongoing skills dynamics, but also on the need for additional qualifications and retraining of specific groups.

[16] The World Economic Forum (2018) together with the Boston Consulting Group and Burning Glass Technologies used a similar approach for the US labor market. On top of their approach, Prüfer et al. (2019) analyze dynamics in the demand for skills in various professions based on vacancy data.

This approach is not without caveats, either. Vacancy data are not necessarily representative and we do not know who applies for a certain vacancy and who is employed in the end. On the other hand, (online) vacancies give a much more fine-grained and real-time picture of labor market demand. This data source can provide information over a longer period of time, for a larger sample, and across various locations. Moreover, vacancy data are less prone to response and recall bias, which are eminent in survey data --- even more so as it is fairly expensive to place a (clearly visible and widely distributed) vacancy. Finally, vacancy data are much cheaper than other sources of information such as questionnaires within a representative sample or register data that have to be linked from multiple sources.

Data and methods

Data

We use data from the vacancy database *Jobfeed*, which is administered by TextKernel, a tech company.[17] This online job portal contains more than 95% of all vacancies published on the Dutch labor market in the last ten years. Therefore, it offers a nearly complete --- and hence nearly representative --- data set of online job ads in the Netherlands. Jobfeed searches the Internet for new vacancies on a daily basis and applies ML algorithms to crawl for vacancies and filter out redundancies. The data mainly contain (unstructured) text, but Jobfeed also extracts structured data such as profession, education, location and company name.

We use data for a period of 6 years, from January 2012 until December 2017, in total about 7.7 million vacancies. Most of the vacancies are written in Dutch; about 8% are in English. As long as a candidate or job description is available, we use all vacancies in our analyses; this holds for 7.32 million vacancies relating to 371 different occupations. The candidate and job descriptions contain relevant information about required skills, experience, and education. In addition, we use information gathered from multiple sources, including the Occupational Information Network (O*NET), an online database with information about the knowledge, skills, tasks, training and experience required for a large number of occupations. Another data source is ISCO (International Standard Classification of Occupations; version ISCO-2008), a classification of 436 professions supplied by the International Labor Organization (ILO). In the ISCO-08 classification, a profession has a skill level (1 to 4) and is a combination of the nature of the work, the required training, and the required experience. Other sources for skills data we used include the EU skills framework, Stackoverflow, and Dbpedia, Wikipedia's skills database.[18]

Methods

As the collected vacancies from the Internet consist of unstructured text, we apply *Natural Language Processing* (NLP) techniques. However, an initial step is *data pre-processing* of the

[17] See https://www.textkernel.com/hr-software/jobfeed/.
[18] More information can be found on the various websites: https://www.onetonline.org/help/onet/database; http://dbpedia.org/page/Category:Skills; https://stackoverflow.com/tags?page=1&tab=popular.

vacancy texts, which helps to improve text-mining results. An important step of pre-processing is the removal of *stop words*, such as articles and prepositions. These words often appear in the candidate and job descriptions, but do not describe skills, education, knowledge or experience. Examples of this are articles and prepositions. Standard Dutch and English stop word lists exist to remove these stop words. In addition, we have identified high-frequency words that do not provide information about the profile, for instance 'experience' or 'knowledge,' and removed all this non-usable information from our dataset.[19]

Moreover, we removed structured fields in the Jobfeed database, such as e-mail addresses, telephone numbers and links to websites, by using a so-called *regular expression,* a sequence of characters that define a search pattern. Using for instance the popular library re (for regular expression operations) in Python allows us to match or search sequences of characters by checking if a given word/phrase is present in a text.[20] Hence, it is useful for dictionary-based skill extraction. It is also useful for text cleaning operations by matching the specified sequence of characters, for example website links or email addresses, which we then remove because this type of information is not relevant for our analysis and could even have negative effects (for instance, web links could be incorrectly recognized as HTML-skills).

A final step is to normalize the text because in unstructured data words can appear in various forms, such as ‘required,’ ‘require,’ and ‘requiring’. There are also derived words with similar meaning, such as ‘entrepreneurial,’ ‘entrepreneur,’ and ‘entrepreneurship.’ The purpose of *text normalization* is to reduce inflections (i.e. derivations) of a word into a common basic form to arrive at a single canonical form the text might not have had before. The form of text normalization that we apply is called *stemming*, in which ends of words are hacked by applying a heuristic process. To do this in Dutch language we apply an existing algorithm.[21]

The specific NLP tool we use for this project is the *bag-of-words* model. This model helps to retrieve information from an unstructured data source by representing a text as the bag (multiset) of its words, disregarding grammar and word order, but keeping information on the frequency of each word and using it as a feature for training a classifier. To make text suitable for analysis, we transformed it into a vector of numbers that relate to the meaning of each word and how it relates to other words. We then applied a mathematical distance measure to calculate the difference (distance) between all the words in our text fragments.

After the pre-processing steps, we can finally extract all necessary information from our text data. Therefore, we categorize the mentioned skills into two unique lists: *digital and technical skills* and *other skills*. Because the vacancies are partly in English, we use both Dutch and English skill labels. We also included as many different forms and expressions of skills as possible based on the frequency of words in the vacancy texts. In addition, to make the extraction process of skills more reliable and robust, the entire list of skills is normalized and divided into two parts --- skills that

[19] Natural Language Toolkit in Python, https://www.nltk.org/
[20] Regular expression: https://docs.python.org/3/library/re.html .
[21] This is called the *Dutch Snowball stemmer*, available in the Python NLTK package.

contain one character, one word or an abbreviation, and a second list with skills with more than one word. For both categories, the skills are searched within one vacancy. If the exact skill is found in the text, it is counted and if a skill occurs several times within one vacancy, this counts as one. A *unigram model* was used for the first category. In this model, the text (candidate and job description) is fragmented word by word. The text is first cleaned up partly, for example by removing brackets and conversing everything into lowercase. Also noise related to line breaks, special characters and white space is removed again by using regular expression. The splitting into words then only needs to happen on a single space while the words can be looked up in the list of skills.[22] For the skills from the second category, skills with more than one word (so-called bigrams or trigrams), we used regular expressions to match the skills after having done the necessary cleaning.

As mentioned above, there is no generally accepted definition of entrepreneurial skills. Moreover, there are semantic problems, for example, one job ad could mention 'solution-oriented,' whereas another one requires applicants to be 'capable of solving problems;' often multiple skills fall into the same category. Therefore, within the *other skills* list, we created 11 broader categories for the entrepreneurial skills reflecting frequently mentioned skills in the framework of 21st century skills and in the entrepreneurship literature (see Table 1).

Table 1: Categories of Entrepreneurial skills with examples

Category	Skill Examples
Critical thinking	Reasoning/ability to reason, Research, Judgment and decision making, Critical thinking, Systems analysis, Systems evaluation, Business analysis, Business modelling, Business process improvement
Creativity	Creative, Innovation, Originality
Collaboration	Active listening, Team-oriented, Participation in discussions, Collaboration, Ability to work together
Communication	Speaking/Oral communication, Writing, Reporting, Reading comprehension, Written understanding, Bilingual/Multi-lingual(Dutch, German, French, English), Presentation Skills
Computational thinking	Mathematics, Analytical, Science, Econometrics, Statistics
Flexibility	Adapting, Flexibility, Ability to adjust

[22] This approach helps to prevent that skills are incorrectly recognized on the basis of only part of a word or sentence, thus to avoid any spurious matching of skills. A special exception is the 'Microsoft Word' skill. This skill is sometimes referred to only as 'Word'. But the Dutch word 'word' and 'Word' is also common. Only the word 'Word', case sensitive, is recognized as a skill, with the exception of cases where it is followed by 'you' or 'then'. The same problem occurs with the programming language 'C', which cannot be erroneously recognized when asked for a driving license C or the Dutch nursing diploma C.

Leadership	Coordination, Negotiation, Leadership, Delegating, Coaching, Persuasiveness, Ability to lead a team/group
Self-starter	Self-motivated, Initiative, Proactive, Entrepreneurship, Inquisitive, Enthusiastic, Independence, Curious go-getter
Problem solving	Root Cause Analysis, Problem management, Problem Sensitivity, Problem solving, Solution-oriented, Perseverance
Active learning	Active learning, Learning strategies, Learning assessment and evaluation, Development management, Eager to learn, Ability to learn
Planning and organization	Time management, Risk management, Organization design and implementation, Project management, Facility Management, Strategic thinking, Systemic thinking, Change management, Program management, Sustainability strategy, Requirements definition and management, Requirement gathering, Monitoring

Results

Figure 1 shows the ranking of entrepreneurial skills in all vacancies that require at least one entrepreneurial skill. This ranking is based on the cumulative fraction of appearance of the skills of a certain category in all vacancies. Thus, it is the total number of skills appearing in the job descriptions normalized by the total number of jobs of that year in that category. The more often the skills from a certain category are demanded in vacancies, the higher the rank of this category on our heat map (and the darker the color). In other words, this shows the (change in) total demand for the skills in the different categories.

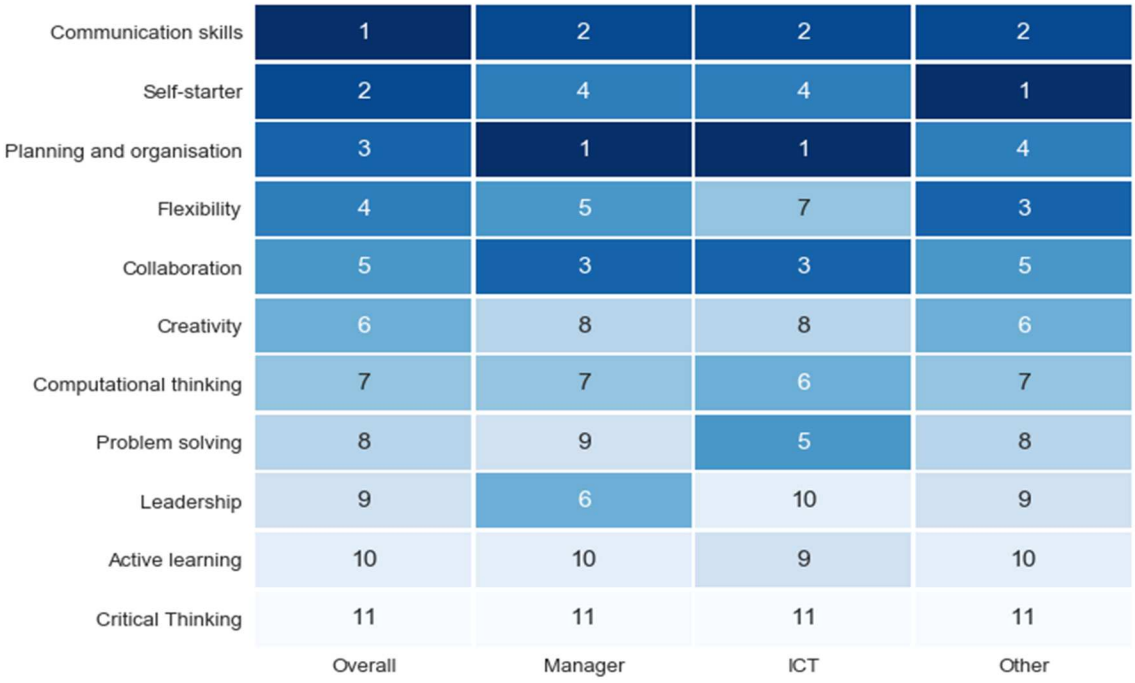
Overall, *communications skills* are in highest demand in the years 2015-2017, followed-shortly by *self-starter skills*.[23] *Planning and organization skills*, also including the project management skills "agile" and "scrum," rank highly for managers and ICT professionals.[24] Other skills categories that are more relevant in these two occupation types than in general are the well-known entrepreneurial skills *collaboration* and *leadership*. Surprisingly, *creativity* and *flexibility* are less demanded than overall, although the difference is less pronounced for flexibility. In contrast, *self-starter skills* are

[23] The majority of vacancies (more than 80%) are related to other professions (including the category 'unknown'), while managerial occupations account for 7% of all vacancies and ICT jobs for about 10%. Communication skills and self-starter skills do not differ much from each other in the vacancies for other professions, while self-starter skills rank only fourth for managers and ICT professions. Therefore, it is possible that communication skills are number 2 in all the three types of professions, but rank number 1 overall.

[24] Scrum and Agile are project management methods (http://www.mountaingoatsoftware.com/agile/scrum)

ranked first for other professions; *flexibility* comes third, while *planning and organization skills* end up on the fourth position.

Figure 1: Ranking of entrepreneurial skills overall and per job type (2015-2017)



Moving to the dynamic dimension of our study, if we look at the trend in entrepreneurial skills between 2012 and 2017 (Figures 2 and 3), we observe an increase in the demand for cooperation (related to *communication* (by factor 1.0) and *collaboration* (by factor 1.4) *skills*) and in skills for *planning and organization* (by factor 1.4), *self-starter* (by factor 1.0), *computational thinking* (by factor 1.2), *problem solving* (by factor 1.2), and *active learning* (by factor 1.7). *Flexibility* (by factor 1.1) and *leadership skills* (by factor 1.0) are also in increasing in demand, while the remaining skills remain more or less stable. Overall, the demand for *active learning skills* is rising most in this period, indicating an increasing need for employees that are intrinsically interested in achieving higher skill levels and in lifelong learning. This also highlights the repercussions from the ongoing digitization and automation, which lead to faster technological change and, therefore, impose higher demand for a highly skilled, self-managing, and continuously learning labor force.

Figure 2: Dynamics in entrepreneurial skills 2012-2017 (top categories)

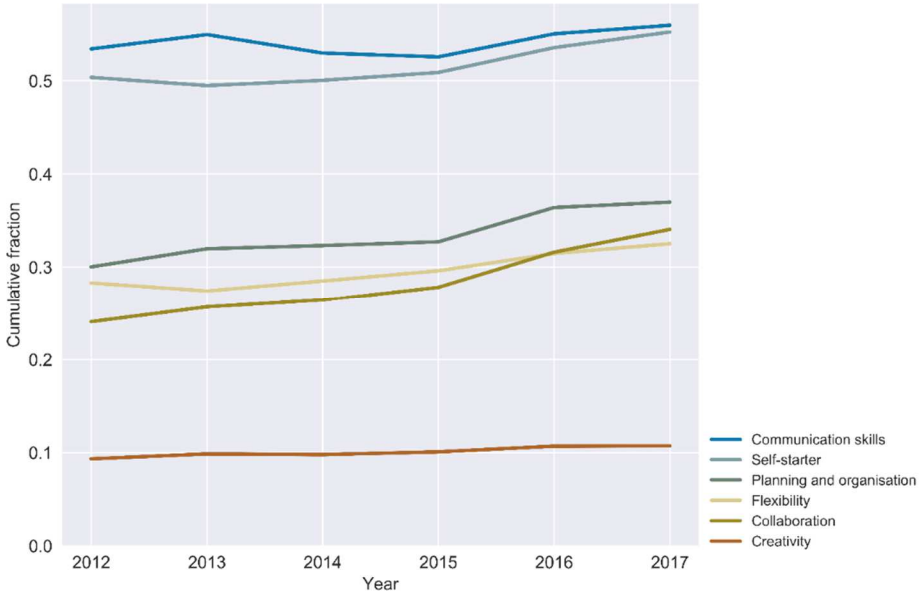
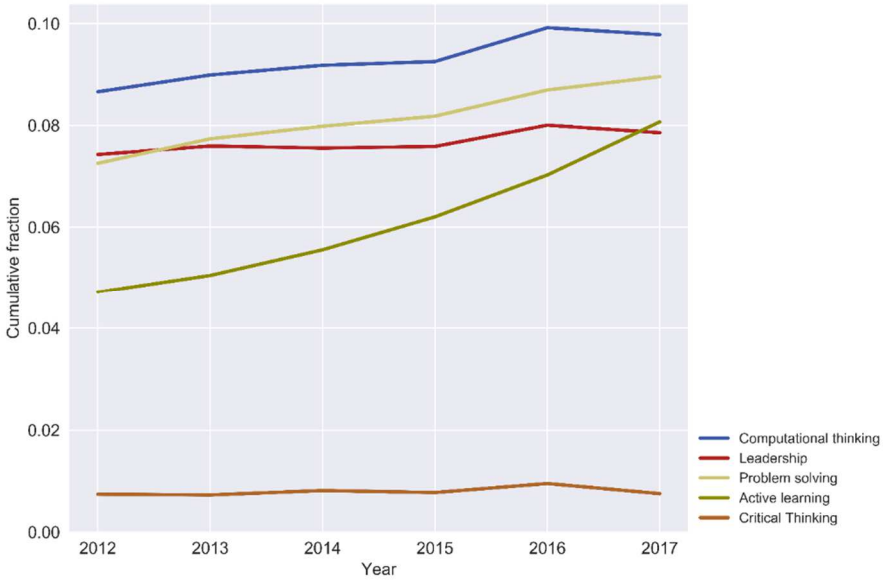
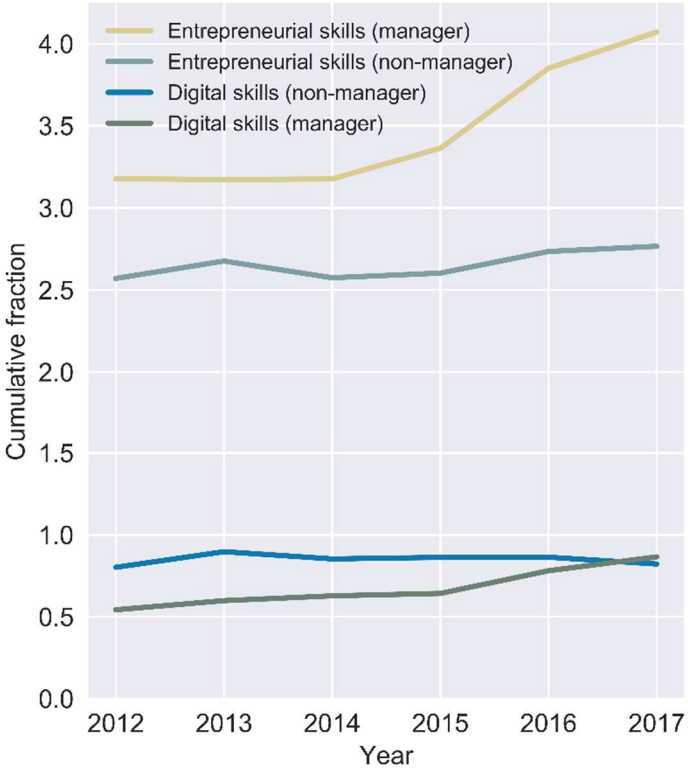


Figure 3: Dynamics in entrepreneurial skills 2012-2017 (bottom categories)



Within the class of entrepreneurial skills, we thus find that there is an increase in the demand for Communication, Collaboration, Computational thinking, Planning and organizational, Self-starter, Problem solving, and Active learning skills, highlighting the importance of the so-called 21st Century skills.

Figure 4: Dynamics in entrepreneurial and digital skills for different types of professions[25]



Comparing the dynamics in entrepreneurial skills to the dynamics in digital skills and making a distinction between managers and non-managerial professions, we find that demand for *entrepreneurial skills* has increased by a factor of 1.3 for managers between 2012 and 2017 (Figure 4) (from a cumulative fraction of 3.17 to 4.07). The demand for this type of skills has also increased slightly for non-managerial occupations (combining ICT/technical job and non-ICT/technical jobs). For *digital skills*, we find an increase of factor 1.6 for managers (from a cumulative fraction of .54 to .87) but none for the other occupation types. Prüfer et al. (2019) explain the latter result by steeply increasing demand for skills related to 'Digital transformation' and 'Big data and analytics'.

The relatively larger increase for managers' digital skills (due to their low baseline demand in 2012) notwithstanding, Figure 4 shows that the cumulative fraction of entrepreneurial skills demanded by managers is significantly larger than the cumulative fraction of managers' demanded digital skills. Moreover, the absolute demand increase for managers' entrepreneurial skills over the 5-year period studied (.9 points) is also larger than the absolute increase for their digital skills (.4 points).

Summarizing, we conclude that both entrepreneurial and digital skills are in increased demand for managerial positions in the Netherlands over the entire period 2012-2017. Given the hugely growing importance of datafication and our finding that, amongst digital skills, those on 'Digital transformation' and 'Big data and analytics' are most valued by employers, one could expect that demand for digital skills would increase most. Our empirical results, however, show the opposite: entrepreneurial skills were significantly more relevant over the six-year period studied. Moreover, the absolute importance of this skill type in managerial job vacancies has increased even more than digital skills'.

[25] If the value of the cumulative fraction is more than 1, on average more than 1 skill appears per job description in that specific category.

5. Discussion and conclusion: opportunities and risks for researchers

The ongoing datafication, coupled with gigantic technological progress in the domain of AI, is changing all aspects of our lives: work, politics, community interactions, economic transactions, and many more. Agrawal et al. (2018, p.194) summarize:

"AI can lead to disruption because incumbent firms often have weaker economic incentives than startups to adopt the technology. AI-enabled products are often inferior at first because it takes time to train a prediction machine to perform as well as a hard-coded device that follows human instructions rather than learning on its own. However, once deployed, an AI can continue to learn and improve, leaving its unintelligent competitors' products behind. It is tempting for established companies to take a wait-and-see approach, standing on the sidelines and observing the progress in AI applied to their industry. That may work for some companies, but others may find it difficult to catch up once their competitors get ahead in the training and deployment of AI tools."

Now, substitute "researchers" for "firms"/"companies" in this quotation and "research projects" for "products."

The disruption occurring at the economy-level is mirrored in the world of research, fueled by developments in data science methods. Distinguishing themselves from traditional statistics and econometrics, these methods use algorithmic models and treat the data mechanism as unknown in order to discover complex structures that were not specified in advance. Where conventional statistics is deductive, data science is inductive. These inductive methods facilitate the automated collection of information, especially on, but not restricted to, the Internet. Via text analysis, computers can learn to understand the meaning of words, relate them to each other, and analyze them at scales that otherwise would require the help of hordes of research assistants. The new techniques and technologies also allow to use many more (unstructured) real-time data sources to conduct analyses that would not have been possible otherwise, for instance by using sensor data from mobile devices (Blumenstock et al., 2015). By making reliance on subjective and self-reported surveys largely unnecessary and substituting these sources with objective data on revealed preferences, they improve the accuracy, robustness and, hence, the value of entrepreneurship research.

Given that these methods are usually freely available and relatively easy to learn, data science techniques thereby contribute to a democratization of empirical research tools, where scholars or students with fewer resources have a higher chance to compete with established researchers from resource-rich countries regarding the types of research questions they can study.

However, given the current state of data science methods, they cannot completely substitute human creativity and research design skills.[26] According to Agarwal et al. (2018), AI algorithms are better than humans at factoring in complex interactions among different indicators if enough data are

[26] This may be less important in hard sciences and may also change in entrepreneurship research once an *Artificial General Intelligence* is developed - which is expected to take 10-100 years (OECD, 2017).

available. If this condition does not hold, however, humans are often better than machines when understanding the data generation process confers a prediction advantage. In the social sciences, data science methods appear to be especially well suited for first, inductive analyses that guide further research efforts. This occurs, for instance, by pointing researchers at relevant correlations and helping them to design better (field) experiments, to make better comparisons between more precise populations of interest, and to reveal behavior that was difficult to detect previously (Monroe et al., 2015). The inductive, data-driven approach can also point theorists at the key variables of interest for a specific question that deserve being modeled. This may alleviate the need for expert interviews or the use of small, unrepresentative surveys to obtain a first understanding of the main influence factors for a given research question. In Section 4, we showed the advantages of this approach --- and the details how to apply it to a specific question from the domain of entrepreneurship research, the demand dynamics of entrepreneurial skills. Our study, based on a dataset of 95% of all job vacancies in the Netherlands over a 6-year period with 7.7 million data points, has visualized that with data science methods we can study questions that could not have been studied on smaller, non-representative data sets. It has allowed us to state that demand for both entrepreneurial and digital skills has increased for managerial positions but that entrepreneurial skills were significantly more relevant over the entire period 2012-2017 and that the absolute importance of entrepreneurial skills has even increased more than digital skills'. This finding may serve as motivation for more research on the role of entrepreneurial skills in the general population --- and not only among (would-be) entrepreneurs.

Moreover, data science techniques may also reduce the risk that theorists fall victim to confirmation bias (Mahmoodi et al., 2017). Dreaming ahead, this may lead to a norm for the best theoretical researchers having to motivate their models by the results of big data analyses. Notably, data science methods are no substitute for theoretical research or conventional statistics. They complement those established methodologies. A fruitful avenue for further research is to combine big data and ML with administrative and survey data. In all social sciences, data science techniques have been largely applied to Internet data (often by scraping and analyzing big social media data sets). Entrepreneurship research is no exception, as Section 3 has shown. However, this approach ignores both potential selection effects that are due to differences between online (social media) users and the entire population and measurement errors that are due to the unreliability of social media data as a representative measure of social phenomena. Comparing the results of a (small) representative survey with results of (big) unrepresentative data, of which the representativeness can even be assessed empirically, therefore looks like an ideal way forward for empirical research.[27]

Just as all technologies based on AI, data science methods come with risks. Agarwal et al. (2018) conclude their insightful book on the consequences of AI by focusing on three trade-offs. The first

[27] Hal Varian (2014, p.23), Google's Chief Economist, comments: "A good predictive model can be better than a randomly chosen control group, which is usually thought to be the gold standard." Stephens-Davidowitz (2017, p.255) notes: "[E]ven a spectacularly successful Big Data organization like Facebook sometimes makes use of […] a small survey."

is *productivity versus distribution*. Bughin et al. (2018) note: "A key challenge is that adoption of AI could widen [performance and outcome] gaps between countries, companies, and workers." Applied to research, data science methods can increase the number, breadth, and speed of questions we can work on, increasing our productivity. But researchers who neglect technological progress or who miss the train may feel very disadvantaged as some traditional methods may be dominated by data science techniques. Consequently, there may be a watershed moment for every researcher, where she either invests some time to familiarize herself with data science methods (for these the above-described democratization of research tools may kick in), or not (which saves time and effort in the short run but may come at significant risk for the relevance of her research in the long run).

The second trade-off is *innovation versus competition*. In business, the successes of Google and Facebook, both of which are highly data-driven firms that have embraced AI early, have shown that data-driven markets display first-mover advantages and are prone to market tipping. Importantly, watching the dismal fate of their competitors underlines how important it is not to fall behind.[28] To some degree, data science methods could introduce a similar spiral, where those researchers who embrace them early could produce higher-quality research, which may have positive feedback effects on their consecutive projects. As long as data sets from one project can be merged and, hence, be partly reused in future projects, the prediction power of those researchers' models might outcompete latecomers repeatedly, discouraging entry of new researchers in their fields.[29] The quality of top-researchers' work might be(come) stellar but the competitive supply of answers to important research questions might decrease, giving the top researchers significant opinion leadership.

The third trade-off is *performance versus privacy*. Using AI successfully depends on huge amounts of data because it is the very power of personalization of services and inference about an individual's preferences and characteristics that can be made if only sufficient data about other individuals are available.[30] But the benefits of aggregate data may come at individuals' costs, especially for privacy.[31] Doing research by analyzing big data sets with data science methods is subject to the same trade-off as running a firm in a data-driven market. Therefore, such research is subjects to the same laws. As a direct policy response to datafication and AI, the *General Data Protection Regulation* (GDPR) has become effective in the EU in May 2018, regulating the legal

[28] Prüfer and Schottmüller (2017) provide more empirical details, rationalize dominant firms' strategies, and introduce "data-driven indirect network effects" as the source of market tipping on data-driven markets.
[29] Our exercise in Section 4 about the dynamics of demand for entrepreneurial skills, despite all its flaws and omissions, may produce relevant intuition for this point: if it is possible to study the entire population of a country in one research project, the value of studying small sample sizes (with traditional methods) may diminish.
[30] For instance, Facebook offers marketers targeting of more than 29,000 categories of users. As the firms has multidimensional data on its users, it is easy to place a given individual in one category even if some data are missing (https://www.propublica.org/article/facebook-doesnt-tell-users-everything-it-really-knows-about-them).
[31] See Acquisti et al. (2016) for an overview of the privacy literature and Dengler and Prüfer (2018) for a rationalization of consumers' privacy choices even if they have no exogenous taste for privacy.

use of privacy-sensitive data, especially those relating to Internet services.[32] The GDPR is already affecting researchers doing empirical research that uses data from the EU or about EU citizens.[33]

Crucially, the one-to-one translation of the three trade-offs listed by Agarwal et al. (2018) from the business to the research domain is subject to further scrutiny. For instance, it is unclear whether empirical research using data science methods is subject to the same indirect network effects as competition on data-driven markets (which leads to market tipping and one highly dominant firm per market).

By contrast, what is certainly true is that we as researchers need to keep up the standards of verifiability, reliability, and replicability of research results. However, this is particularly difficult when ML algorithms are used because, by definition, the algorithm is learning: it adapts based on feedback.[34] Therefore, it is harder than with conventional research methods to reproduce predictions (read: results) based on ML. What is necessary, thus, is to make the decision-making processes of algorithms more transparent. This would facilitate trust in the new technologies and replicability would be easier.

One option to achieve this goal is to build algorithms with an internal self-evaluation or calibration stage such that the machine can test its own certainty and report back to the researcher. One attempt in this direction is the *Automatic Statistician*, which was developed at Cambridge University.[35] The tool is set up with funding from Google and helps researchers to analyze their datasets while also providing a report in a human-understandable form that explains what it is doing and how certain it is about its predictions. This technology is related to a recent development within ML, *Automated Machine Learning* (AutoML). This approach tackles the fundamental problems of accountability and verifiability. Here, ML methods and hyper-parameter settings are automatically selected and, thereby, reduce the necessity of handcrafted human interventions. Apart from substantial performance improvements, AutoML can provide evaluations of all tested methods and specifications. Thereby, it can help non-experts to effectively and reliably apply ML techniques.

In all social sciences, including entrepreneurship research, there is a lot of ground to cover.

[32] Regulation (EU) 2016/679 of the European Parliament and of the European Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (http://ec.europa.eu/justice/data-protection/).
[33] On the positive side, as all of us are also data subjects, not just researchers, our personalities and digital footprints are protected much better in the EU than in other jurisdictions.
[34] Recently, a Google employee mentioned in personal conversation that the algorithm of Google's search engine would be changed about 2,500 times per year. While the exact number is irrelevant, the high frequency of changes, which complicates accountability for an algorithm's results, is not.
[35] See https://www.automaticstatistician.com/index/.

References

Acquisti, A., Taylor, C. and Wagman, L. (2016), The Economics of Privacy, *Journal of Economic Literature* 54(2): 442-92.

Aggarwal, R. and Singh, H. (2013), Differential Influence of Blogs across Different Stages of Decision Making: The Case of Venture Capitalists, *MIS Quarterly*, *37*(4): 1033–1112.

Agrawal, A., Gans, J. and A. Goldfarb (2018), *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press. Cambridge, MA.

Alsayat, A. and El-Sayed, H. (2016), Social media analysis using optimized K-Means clustering, *Proceedings of 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications* (SERA).

Andrews, K.R. (1980), *The Concept of Corporate Strategy,* 2nd ed., Irwin, Homewood, Illinois.

Arntz, M, Gregory, T. and U. Zierahn (2016), The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis, *OECD Social, Employment and Migration Working Papers*, No. 189, OECD Publishing: Paris.

Autor, D. (2015), Why Are There Still So Many Jobs? The History and Future of Workplace Automation, *Journal of Economic Perspectives* 29(3), 3-30.

Bandiera, O,. Prat, A., Hanse, S. and R. Sadun (2017), CEO Behavior and Firm Performance, Harvard Business School Working Paper 17-083.

Baumol, W.J., Litan, R.E. and C.J. Schramm (2007), *Good capitalism, bad capitalism, and the economics of growth and prosperity*. New Haven, CT: Yale University Press.

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

Bishop, C. (2011), *Pattern Recognition and Machine Learning*, Springer, New York.

Breiman, L. (2001), Statistical Modeling: The Two Cultures, *Statistical Science*, 16(3): 199-231.

Brynjolfsson, E., Mitchell, T. and Rock, D. (2018), What Can Machines Learn and What Does It Mean for Occupations and the Economy? *AEA Papers and Proceedings*, 108: 43-47.

Bughin , J., Seong, J. Manyika, J., Chui, M., and R. Joshi (2018), Notes from the AI frontier: Modeling the impact of AI on the world economy, Discussion Paper McKinsey Global Institute.

Calvano, E., Calzolari, G., Denicolo, V. and S. Pastorello (2018), Artificial Intelligence, Algorithmic Pricing and Collusion, mimeo, University of Bologna.

Cardon, M.S., Foo, M.D., Shepherd, D. and J. Wiklund (2012), Exploring the heart: entrepreneurial emotion is a hot topic, *Entrepreneurship Theory and Practice*, 36(1): 1-10.

Cogburn, D. and Hine, M. (2017), Introduction to Text Mining in Big Data Analytics, *Proceedings of the 50th Hawaii International Conference on System Sciences*, HICSS 2017.

Courtney, C., Dutta, S. and Li, Y. (2017), Resolving Information Asymmetry: Signaling, Endorsement, and Crowdfunding Success, *Entrepreneurship Theory and Practice*, *41*(2): 265–290.

Cummings, M.E., Rawhouser, H., Vismara, S. and E.L. Hamilton (2019), An equity crowdfunding research agenda: evidence from stakeholder participation in the rulemaking process, *Small Business Economics*, forthcoming.

Dean, J. and Ghemawat, S. (2004), MapReduce: Simplified Data Processing on Large Clusters, *OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA*: 137-150.

Dengler, S. and Prüfer, J. (2018), Consumers' Privacy Choices in the Era of Big Data, *TILEC Discussion Paper No. 2018-014, CentER Discussion Paper No. 2018-012*.

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

Elliott, S. (2017), Computers and the Future of Skill Demand, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/9789264284395-en.

George, G., Haas, M. and Pentland, A. (2014), Big data and management, *Academy of Management Journal*, 57(2): 321-32.

Greve, A. and Salaff, J. (2003), Social networks and entrepreneurship, *Entrepreneurship Theory and Practice*, 28: 1-22.

Hambrick, D.C. and P.A. Mason (1984), Upper Echelons: The Organization as a Reflection of Its Top Managers, *Academy of Management Review*, 9: 193-206.

Hartmann, P.M., Zaki, M., Feldmann, N. and A. Neely (2016), Capturing value from big data–a taxonomy of data-driven business models used by start-up firms, International Journal of Operations & Production Management, 36(10): 1382-1406.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Hisrich, R., Langan-Fox, J., Grant, S. (2007) Entrepreneurship research and practice: a call to action for psychology. *American Psychologist*, 62(6): 575-589.

Hoberg, G. and Phillips, G. (2016), Text-Based Network Industries and Endogenous Product Differentiation, *Journal of Political Economy*, *124*(5): 1423–1465.

Hoornaert, S., Ballings, M., Malthouse, E. C. and Van den Poel, D. (2017), Identifying New Product Ideas: Waiting for the Wisdom of the Crowd or Screening Ideas in Real Time, *Journal of Product Innovation Management*, *34*(5): 580–597.

Kotter, J.P. (1999), *John Kotter on What Leaders Really Do*. Harvard Business School Press, Boston.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009), Computational Social Science, *Science,* 323(5915): 721-723.

Lee, J., Hwang, B. and H. Che (2017), Are Founder CEOs more Overconfident than Professional CEOs? Evidence from S&P 1500 Companies, *Strategic Management Journal*, 38: 751-769.

Li, G., Lai, R., D'Amour, A., Doolin, D., Sun, Y., Torvik, V., Yu, A., and Fleming, L. (2014), Disambiguation and Co-Authorship Networks of the U.S. Patent Inventor Database (1975–2010), *Research Policy*, *43*(6): 941–955.

Mahmoodi, J., Leckelt, M., Van Zalk, M., Geukes, K. and Black, M. (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.

Mayer-Schönberger, V. and T. Ramge (2018), *Reinventing Capitalism in the Age of Big Data*, John Murray, London.

McAfee, A. and Brynjolfsson, E. (2017), *Machine - Platform – Crowd: Harnessing our Digital Future*, New York: Norton.

Mintzberg, H. (1973), *The Nature of Managerial Work*. Harper and Row, New York.

Monroe, B.L., Pan, J., Roberts, M.E., Sen, M., and Sinclair, B. (2015), No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science, *PS-Political Science and Politics*, 48(1):71-74.

Mullainathan, S. and Spiess, J. (2017), Machine Learning: An Applied Econometric Approach, *Journal of Economic Perspectives*, 31(2): 87-106.

Murphy, K. (2012), *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge.

OECD (2017), *OECD Digital Economy Outlook 2017*, OECD Publishing, Paris, ch.7.

Obschonka, M. and Fisch, C. (2017), Entrepreneurial Personalities in Political Leadership, *Small Business Economics*: 1–19.

Obschonka, M., Fisch, C. and Boyd, R. (2017a), Using Digital Footprints in Entrepreneurship Research: A Twitter-Based Personality Analysis of Superstar Entrepreneurs and Managers, *Journal of Business Venturing Insights*, *8*: 13–23.

Obschonka, M., Hakkarainen, K., Lonka, K. and K. Salmela-Aro (2017b), Entrepreneurship as a twenty-first century skill: entrepreneurial alertness and intention in the transition to adulthood, *Small Business Economics*, 48: 487-501.

Petriglieri, G., Ashford, S.J. and A. Wrzesniewski (2018), Thriving in the Gig Economy, *Harvard Business Review*, March-April: 140-143.

Provost, F. and T. Fawcett (2013), *Data Science for Business*, O'Reilly, Sebastopol, CA.

Prüfer, J. and Schottmüller, C. (2017) Competing with Big Data, *CentER Discussion Paper* No. 2017-007.

Prüfer, J. and Prüfer, P. (2018), Data Science for Institutional and Organizational Economics, in: *A Research Agenda for New Institutional Economics*, Claude Ménard and Mary M. Shirley (eds.), Edward Elgar Publishers, Cheltenham, UK: 248-259.

Prüfer, P., Kumar, P. and M. den Uijl (2019), Arbeidsmarktonderzoek Digitalisering in Topsectoren, mimeo, CentERdata, Tilburg.

RezaeiZadeh, M., Hogan, M., O'Reilly, J., Cunningham, J. and E. Murphy (2017),Core entrepreneurial competencies and their interdependencies: insights from a study of Irish and Iranian entrepreneurs, university students and academics, *International Entrepreneurship and Management Journal*, 13: 35-73.

Rickne, A., Ruef, M. and K. Wennberg (2018), The socially and spatially bounded relationships of entrepreneurial activity: Olav Sorenson—recipient of the 2018 Global Award for Entrepreneurship Research, *Small Business Economics*, 51(3): 515-525.

ROA (2017), De Arbeidsmarkt naar Opleiding en Beroep tot 2022, ROA-R-2017/10, Maastricht.

Rogers, E. (1962), *Diffusion of Innovations*, Free Press, New York.

Rosique-Blasco, M., Madrid-Guijarro, A. and D. García-Pérez-de-Lema (2018), The effects of personal abilities and self-efficacy on entrepreneurial intentions, *International Entrepreneurship and Management Journal,* 14: 1025-1052.

Rutherford, M., McMullen, P. and Oswald, S. (2001), Examining the Issue of Size and the Small Business: A Self Organizing Map Approach, *Journal of Business and Economic Studies*, 7(2): 64–79.

Samuel, A. (1959), Some Studies in Machine Learning Using the Game of Checkers, *IBM Journal*, 3(3), 535-554.

Shane, S. (2012), Reflections on the 2010 AMR decade award: Delivering on the promise of entrepreneurship as a field of research. *Academy of Management Review*, 37(1): 10-20.

Shane, S. and Venkataraman, S. (2000), The promise of entrepreneurship as a field of research, *Academy of Management Review*, 25: 217-226.

Sorenson, O. (2018), Social networks and the geography of entrepreneurship, *Small Business Economics*, 51(3): 527-537.

Spitz-Oener, A. (2006), Technical Change, Job Tasks, and Rising Educational Demands: Looking Outside the Wage Structure, *Journal of Labor Economics,* 24, 235-270.

Stephenson-Davidowitz, S. (2017), *Everybody Lies – Big Data, New Data, and What the Internet can tell us about who we really are*, Harper Collins, New York.

Stuart, R. and P.A. Abetti (1987), Start-up ventures: Towards the prediction of initial success, *Journal of Business Venturing*, 2(3): 215-230.

Taddy, M. (2018), The Technological Elements of Artificial Intelligence*, NBER Working Paper* No. 24301.

Tan, S. and Koh, H. C. (1996), Modelling Entrepreneurial Inclination with an Artificial Neural Network, *Journal of Small Business & Entrepreneurship*, *13*(2): 14–24.

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

Tausczik, Y.R. and J.W. Pennebaker (2010), The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1): 24-54.

The Economist (2018), *No PhD, no problem – New schemes teach the masses to build AI*, October 25, San Francisco.

Vachelard, J., Gambarra-Soares, T., Augustini, G., Riul, P. and Maracaja-Coutinho, V. (2016), A Guide to Scientific Crowdfunding, *PLoS Biology*, 14(2): 1-7.

Varian, H. (2014), Big Data: New Tricks for Econometrics, *Journal of Economic Perspectives*, 28(2): 3-27.

Ventura, S., Nugent, R. and Fuchs, E. (2015), Seeing the Non-Stars: (Some) Sources of Bias in Past Disambiguation Approaches and a New Public Tool Leveraging Labeled Records, *Research Policy*, *44*(9): 1672–1701.

Wang, F., Mack, E. and Maciewjewski, R. (2017), Analyzing Entrepreneurial Social Networks with Big Data, *Annals of the American Association of Geographers*, *107*(1): 130–150.

World Economic Forum (2018), *Towards a Reskilling Revolution - A Future of Jobs for All*, Davos, Zwitserland.

Appendix

Getting started yourself

Using data science for your own research does not require a PhD or other academic credentials in that field (The Economist 2018). Bishop (2011), Hastie et al. (2009), Murphy (2012), and Provost and Fawcett (2013) are excellent entry points in book form. Moreover, many high-quality online resources are available, for which good knowledge of basic linear algebra and probability theory is a big help. Useful resources to learn these methods are the video lectures of Andrew Ng at Stanford University, GitHub repositories, and Coursera, Udacity, Udemy, or edX courses. To gain practical experience with various kinds of challenging data, data science enthusiasts can try many open projects available at Kaggle. It can really help to learn fast and hone the practical skills related data science further.

Kaggle, the biggest data science community in the world, is actually itself a crowdsourcing initiative for data science.[36] Companies that need help sorting data turn to Kaggle, a new platform that leverages the data science crowd via competitions. These data scientists work to solve a company's data questions in an attempt to win the company-sponsored financial reward or just for the pleasure of showing off. Founded in 2010, Kaggle boasts a community of "tens of thousands" experts from over 100 countries and 200 universities in any fields related to data science. The Kaggle ranking has become an essential metric in the world of data science. Some employers have begun listing a Kaggle rank as an essential qualification and Facebook uses Kaggle competitions as part of its recruiting strategy. An interview for a job as a data scientist at Facebook is the prize.

Hadoop Ecosystem

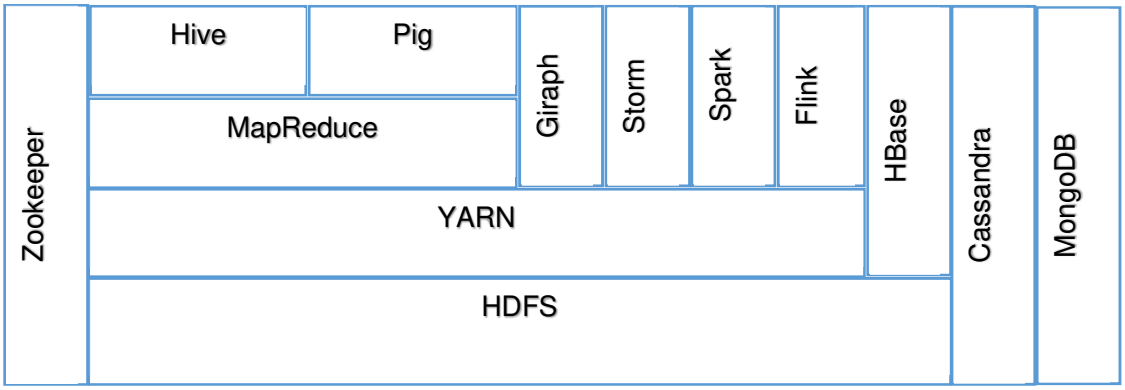
The big data open-source movement potentially commenced in 2004 when Google published a paper on their in-house processing framework popularly known as *MapReduce* (Dean and Ghemawat, 2004). Later, Yahoo released an open-source implementation based on this framework called *Hadoop*. Subsequently, many other frameworks and tools were released as open-source projects. As of now, there are more than 100 open-source projects for big data, a fast-growing number.

The following layered diagram (also called *stacked diagram*) organizes the capability or functionality of the components in the layer. In a layer diagram, a component uses the functionality of the components in the layer below it. Normally components at the same layer do not communicate.

Key Points of this framework:

1. The Hadoop distributed file system (HDFS) is the foundation for many big data frameworks as it provides scalable and reliable storage.

[36] https://www.kaggle.com/competitions.



2. Hadoop *YARN* provides flexible scheduling and resource management over the HDFS storage.
3. *MapReduce* is a programming model that simplifies parallel computing.
4. *Hive* and *Pig* are two additional programming models on top of MapReduce. Hive was created at Facebook to issue SQL-like queries using MapReduce on their data in HDFS. Pig was created at Yahoo to model data flow based programs using MapReduce.
5. *Giraph* was built for processing large-scale graphs efficiently. For example, Facebook uses Giraph to analyze the social graphs of its users.
6. *Storm, Spark,* and *Flink* (In-memory processing) are useful for real time and in memory processing of big data on top of the YARN resource scheduler and HDFS.
7. *Cassandra, MongoDB* and *HBase* are NoSQL databases. Cassandra was created at Facebook. Facebook also used HBase for its messaging platform.
8. *Zookeeper* was created at Yahoo. It is a centralized management system for synchronization, configuration and to ensure high availability of all these tools.

The Hadoop ecosystem consists of a growing number of open-source tools. Providing opportunities to pick the right tool for the right tasks for better performance and lower costs. We describe some tools in further detail and recommend optimal use in the following table.

Box A1: Technical terms and tools for big data applications and data science[37]

Application Programming Interfaces (APIs)	A set of subroutine definitions, protocols, and tools for building application software. In general, a set of clearly defined methods of communication between various software components. An API may be for a web-based system, operating system, database system, or computer hardware. The use of open APIs has resulted in an exponential growth of user-generated data (via apps and software programs) as an API reports any of its use, e.g. for web scraping, back to the API provider. Hence, it has given direct and indirect boost to big data and analytics.

[37] Not all of the terms explained are mentioned in this paper. Given that these are all frequently used terms, we include them anyway as a service to the reader.

Cassandra	Apache Cassandra is a free and open-source distributed NoSQL database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure.
Flink	An open source framework for processing data in both real time mode and batch mode. It provides several benefits such as fault-tolerant and large-scale computation. Its programming model is similar to MapReduce. In contrast to MapReduce, it offers additional high-level functions such as join, filter and aggregation.
Flume	A highly distributed, reliable, robust, fault tolerant and configurable tool, which collects streaming data (log data) from various web servers to HDFS.
Git	Free and open source software used as a version control system for tracking changes in computer files and coordinating work on those files among multiple people.
Hadoop	An open-source software library that establishes a framework for the distributed processing of big data using simple programming models. Hadoop has two main components: Hadoop *Distributed File System (HDFS)* a scalable distributed file system for storing large files over distributed machines in a reliable and efficient way; and *MapReduce* programming, a model to process huge data in-parallel on large clusters (thousands of nodes) in a reliable, fault-tolerant manner.
HBase	A column-oriented NoSQL database management system that runs on top of Hadoop Distributed File System (HDFS) similar to Google's Bigtable and well suited for well suited for sparse data sets, which are common in many big data use cases.
Hive	Built on top of Hadoop and allows SQL developers to reading, writing, and managing large datasets in distributed storage using Hive Query Language (HQL) statements similar to standard SQL.
Library	A library is collection of pre-written programs, scripts, or functions that can be loaded on disk for immediate use. All of the available functions within a (software) library can be used within the program body without explicitly defining them. With the help of his these libraries, one can implement (complex) algorithms by writing few line of codes.
Pig	An abstraction over MapReduce and tool/platform to analyze larger sets of data representing them as data flows. It is generally used with Hadoop to perform the data manipulation operations in Hadoop. Pig is amenable and well suited for parallelization and, thus, provides capability to handle very large data sets.
Spark	An open-source cluster computing framework operating on distributed data collections (in-memory distributed data analysis platform; without a storage component such as Hadoop) primarily targeted at speeding up batch analysis jobs, iterative ML jobs, interactive queries, and graph processing. The Spark Big Data platform can be combined with other analytics platforms, such as *Databricks*, which uses the well-known programming language *Scala*. An extension of the core Spark API is *Spark Streaming* that enables scalable, high-throughput, fault-tolerant stream processing of data.
Storm	An open source framework for processing large structured and unstructured data in real time. Storm is a fault tolerant framework that is suitable for real time data analysis, ML, sequential and iterative computation. Storm is geared for real time applications while the Hadoop is effective for batch applications.
Tableau	Is a commercial package that can be very useful to visualize big data and to get actionable insights in fast and efficient way.
Virtual environment software	Refers to any software, program, or system that implements, manages, and controls multiple virtual environment instances. The software is installed within an organization's existing IT infrastructure and controlled from within the organization itself. At its core, the main purpose of (Python) virtual environments is to create an isolated environment for (Python) projects.

	This means that each project can have its own dependencies, regardless of what dependencies every other project has.

Box A2: Techniques and tools for text mining

Basic Text Mining	*Natural Language Toolkit (NLTK)* is a widely used open-source toolkit for text mining and NLP. It has several handy tools, gives access to many text corpora, and to the most suitable algorithms for such tasks. [http://www.nltk.org/]
Web Scraping	*BeautifulSoup* is a tool to work with web-based data. It facilitates the scraping, parsing, and reading of web data, as well as data access using web APIs in different formats of data, for example in HTML, XML, and JSON formats. [https://www.crummy.com/software/BeautifulSoup/bs4/doc/]
Text Classification	One of the important and typical tasks in *supervised* machine learning. Assigning categories to documents, which can be web pages, library books, media articles, etc. has many applications, for instance, spam filtering, e-mail routing, or sentiment analysis. Several toolkits are available for supervised text classification. *Scikit-learn*, an open-source machine learning library in Python, is a prominent one. [http://scikit-learn.org/stable/]
Information Extraction (IE)	An important task for natural language understanding and making sense of textual data. The main goal of IE is to identify and extract fields of interest from free text. It is the first step in converting the unstructured text to more structured forms. The so-called *Stanford NLP* is a suite of very useful NLP tools for IE. [https://nlp.stanford.edu/software/]
Semantic Similarity & Topic Modelling	Algorithms to detect semantic similarity are used to group similar words into semantic concepts that have the same meaning, or appear to have the same meaning. For example, currency – money – coin are semantically similar. One of the resources useful for semantic similarity is WordNet, which is a semantic dictionary of words interlinked by semantic relationships.[38] Topic modeling is a widely used text-mining tool for discovering hidden patterns in a text body. A good topic model for example gives 'school', 'university', 'college', 'teacher', 'professor' for a topic "Education".
Sentiment Analysis	Opinion mining (sometimes known as sentiment analysis or emotion AI) refers to the use of NLP to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely used to analyze reviews, survey responses, and online and social media discussions. There are two ways to perform sentiment analysis: the lexicon-based approach and the machine-learning approach. For both approaches, different tools and algorithms exist as well as databases of positive and negative words.[39]

[38] *WordNet* was developed for English but exists for many other languages today, too. WordNet includes rich linguistic information e.g. part of speech, different meanings of the same word, synonyms, words with same meaning, hypernyms and hyponyms. WordNet is freely available in NLTK (http://www.nltk.org/howto/wordnet.html) or on the website of Princeton University (https://wordnet.princeton.edu/wordnet/download/). It is extensively used in many natural language processing tasks and, more broadly, in text mining tasks.

[39] For example, Liu and Hu's opinion lexicon contains around 6,800 positive and negative opinion or sentiment words for English: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon. *SentiWordNet* is a good lexical resource for opinion mining that assigns three sentiment scores (positivity, negativity, and objectivity) to the words in WordNet (http://sentiwordnet.isti.cnr.it/). *NLTK* and *TextBlob* are Python libraries that are frequently used for sentiment analysis based on machine learning. TextBlob is built

Linguistic Inquiry and Word Count (LIWC)	LIWC is an application of computer-based text analysis tools in psychology. Its two features are: the processing component and the dictionaries. The processing feature is the program that opens a series of text files such as essays, poems, blogs, novels, and social media data and then analyzes each file word by word. Each word in a given text file is compared with the dictionary file. This tool reflects how language correlates with emotional state, social relationships, thinking styles and individual differences. [http://liwc.wpengine.com/]

on the top of NLTK, is more convenient than NLTK for new users, and has a lot of functionality in NLP tasks. Similar libraries are also available in R and RapidMiner.