Data Science Technologies in Economics and Finance: A Gentle Walk-In

Luca Barbaglia, Sergio Consoli, Sebastiano Manzan, Diego Reforgiato Recupero, Michaela Saisana, and Luca Tiozzo Pezzoli

Abstract This chapter is an introduction to the use of data science technologies in the fields of economics and finance. The recent explosion in computation and information technology in the past decade has made available vast amounts of data in various domains, which has been referred to as *Big Data*. In economics and finance, in particular, tapping into these data brings research and business closer together, as data generated in ordinary economic activity can be used towards effective and personalized models. In this context, the recent use of data science technologies for economics and finance provides mutual benefits to both scientists and professionals, improving forecasting and nowcasting for several kinds of applications. This chapter introduces the subject through underlying technical challenges such as data handling and protection, modeling, integration, and interpretation. It also outlines some of the common issues in economic modeling with data science technologies and surveys the relevant big data management and analytics solutions, motivating the use of data science methods in economics and finance.

1 Introduction

The rapid advances in information and communications technology experienced in the last two decades have produced an explosive growth in the amount of information collected, leading to the new era of big data [31]. According to [26], approximately three billion bytes of data are produced every day from sensors, mobile devices, online transactions, and social networks, with 90% of the data in

Authors are listed in alphabetic order since their contributions have been equally distributed.

L. Barbaglia · S. Consoli (✉) · S. Manzan · M. Saisana · L. Tiozzo Pezzoli
European Commission, Joint Research Centre, Ispra (VA), Italy
e-mail: sergio.consoli@ec.europa.eu

D. Reforgiato Recupero
Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy

S. Consoli et al. (eds.), *Data Science for Economics and Finance*,
https://doi.org/10.1007/978-3-030-66891-4_1

the world having been created in the last 3 years alone. The challenges in storage, organization, and understanding of such a huge amount of information led to the development of new technologies across different fields of statistics, machine learning, and data mining, interacting also with areas of engineering and artificial intelligence (AI), among others. This enormous effort led to the birth of the new cross-disciplinary field called "Data Science," whose principles and techniques aim at the automatic extraction of potentially useful information and knowledge from the data. Although data science technologies have been successfully applied in many different domains (e.g., healthcare [15], predictive maintenance [16], and supply chain management [39], among others), their potentials have been little explored in economics and finance. In this context, devising efficient forecasting and nowcasting models is essential for designing suitable monetary and fiscal policies, and their accuracy is particularly relevant during times of economic turmoil. Monitoring the current and the future state of the economy is of fundamental importance for governments, international organizations, and central banks worldwide. Policy-makers require readily available macroeconomic information in order to design effective policies which can foster economic growth and preserve societal well-being. However, key economic indicators, on which they rely upon during their decision-making process, are produced at low frequency and released with considerable lags—for instance, around 45 days for the Gross Domestic Product (GDP) in Europe—and are often subject to revisions that could be substantial. Indeed, with such an incomplete set of information, economists can only approximately gauge the actual, the future, and even the very recent past economic conditions, making the nowcasting and forecasting of the economy extremely challenging tasks. In addition, in a global interconnected world, shocks and changes originating in one economy move quickly to other economies affecting productivity levels, job creation, and welfare in different geographic areas. In sum, policy-makers are confronted with a twofold problem: timeliness in the evaluation of the economy as well as prompt impact assessment of external shocks.

Traditional forecasting models adopt a mixed frequency approach which bridges information from high-frequency economic and financial indexes (e.g., industrial production or stock prices) as well as economic surveys with the targeted low-frequency variable, such as the GDP [28]. An alternative could be dynamic factor models which, instead, resume large information in few factors and account of missing data by the use of Kalman filtering techniques in the estimation. These approaches allow the use of impulse-responses to assess the reaction of the economy to external shocks, providing general guidelines to policy-makers for actual and forward-looking policies fully considering the information coming from abroad. However, there are two main drawbacks to these traditional methods. First, they cannot directly handle huge amount of unstructured data since they are tailored to structured sources. Second, even if these classical models are augmented with new predictors obtained from alternative big data sets, the relationship across variables is assumed to be linear, which is not the case for the majority of the real-world cases [21, 1].

Data science technologies allow economists to deal with all these issues. On the one hand, new big data sources can integrate and augment the information carried by publicly available aggregated variables produced by national and international statistical agencies. On the other hand, machine learning algorithms can extract new insights from those unstructured information and properly take into consideration nonlinear dynamics across economic and financial variables. As far as big data is concerned, the higher level of granularity embodied on new, available data sources constitutes a strong potential to uncover economic relationships that are often not evident when variables are aggregated over many products, individuals, or time periods. Some examples of novel big data sources that can potentially be useful for economic forecasting and nowcasting are: retail consumer scanner price data, credit/debit card transactions, smart energy meters, smart traffic sensors, satellite images, real-time news, and social media data. Scanner price data, card transactions, and smart meters provide information about consumers, which, in turn, offers the possibility of better understanding the actual behavior of macro aggregates such as GDP or the inflation subcomponents. Satellite images and traffic sensors can be used to monitor commercial vehicles, ships, and factory tracks, making them potential candidate data to nowcast industrial production. Real-time news and social media can be employed to proxy the mood of economic and financial agents and can be considered as a measure of perception of the actual state of the economy.

In addition to new data, alternative methods such as machine learning algorithms can help economists in modeling complex and interconnected dynamic systems. They are able to grasp hidden knowledge even when the number of features under analysis is larger than the available observations, which often occurs in economic environments. Differently from traditional time-series techniques, machine learning methods have no "a priori" assumptions about the stochastic process underlying the state of the economy. For instance, deep learning [29], a very popular data science methodology nowadays, is useful in modeling highly nonlinear data because the order of nonlinearity is derived or learned directly from the data and not assumed as is the case in many traditional econometric models. Data science models are able to uncover complex relationships, which might be useful to forecast and nowcast the economy during normal time but also to spot early signals of distress in markets before financial crises.

Even though such methodologies may provide accurate predictions, understanding the economic insights behind such promising outcomes is a hard task. These methods are black boxes in nature, developed with a single goal of maximizing predictive performance. The entire field of data science is calibrated against out-of-sample experiments that evaluate how well a model trained on one data set will predict new data. On the contrary, economists need to know how models may impact in the real world and they have often focused not only on predictions but also on model inference, i.e., on understanding the parameters of their models (e.g., testing on individual coefficients in a regression). Policy-makers have to support their decisions and provide a set of possible explanations of an action taken; hence, they are interested on the economic implication involved in model predictions. Impulse response functions are a well-known instruments to assess the impact of a shock

in one variable on an outcome of interest, but machine learning algorithms do not support this functionality. This could prevent, e.g., the evaluation of stabilization policies for protecting internal demand when an external shock hits the economy. In order to fill this gap, the data science community has recently tried to increase the transparency of machine learning models in the literature about *interpretable AI* [22]. Machine learning applications in economics and finance can now benefit from new tools such as Partial Dependence plots or Shapley values, which allow policy-makers to assess the marginal effect of model variables on the predicted outcome. In summary, data science can enhance economic forecasting models by:

- Integrating and complementing official key statistic indicators by using new real-time unstructured big data sources
- Assessing the current and future economic and financial conditions by allowing complex nonlinear relationships among predictors
- Maximizing revenues of algorithmic trading, a completely data-driven task
- Furnishing adequate support to decisions by making the output of machine learning algorithms understandable

This chapter emphasizes that data science has the potential to unlock vast productivity bottlenecks and radically improve the quality and accessibility of economic forecasting models, and discuss the challenges and the steps that need to be taken into account to guarantee a large and in-depth adoption.

2 Technical Challenges

In recent years, technological advances have largely increased the number of devices generating information about human and economic activity (e.g., sensors, monitoring, IoT devices, social networks). These new data sources provide a rich, frequent, and diversified amount of information, from which the state of the economy could be estimated with accuracy and timeliness. Obtaining and analyzing such kinds of data is a challenging task due to their size and variety. However, if properly exploited, these new data sources could bring additional predictive power than standard regressors used in traditional economic and financial analysis.

As the data size and variety augmented, the need for more powerful machines and more efficient algorithms became clearer. The analysis of such kinds of data can be highly computationally intensive and has brought an increasing demand for efficient hardware and computing environments. For instance, Graphical Processing Units (GPUs) and cloud computing systems in recent years have become more affordable and are used by a larger audience. GPUs have a highly data parallel architecture that can be programmed using frameworks such as CUDA[1] and OpenCL.[2] They

[1] NVIDIA CUDA: https://developer.nvidia.com/cuda-zone.

[2] OpenCL: https://www.khronos.org/opencl/.

consist of a number of cores, each with a number of functional units. One or more of these functional units (known as *thread processors*) process each thread of execution. All thread processors in a core of a GPU perform the same instructions, as they share the same control unit. Cloud computing represents the distribution of services such as servers, databases, and software through the Internet. Basically, a provider supplies users with on-demand access to services of storage, processing, and data transmission. Examples of cloud computing solutions are the Google Cloud Platform,[3] Microsoft Azure,[4] and Amazon Web Services (AWS).[5]

Sufficient computing power is a necessary condition to analyze new big data sources; however, it is not sufficient unless data are properly stored, transformed, and combined. Nowadays, economic and financial data sets are still stored in individual silos, and researchers and practitioners are often confronted with the difficulty of easily combining them across multiple providers, other economic institutions, and even consumer-generated data. These disparate economic data sets might differ in terms of data granularity, quality, and type, for instance, ranging from free text, images, and (streaming) sensor data to structured data sets; their integration poses major legal, business, and technical challenges. Big data and data science technologies aim at efficiently addressing such kinds of challenges.

The term "big data" has its origin in computer engineering. Although several definitions for big data exist in the literature [31, 43], we can intuitively refer to data that are so large that they cannot be loaded into memory or even stored on a single machine. In addition to their large *volume*, there are other dimensions that characterize big data, i.e., *variety* (handling with a multiplicity of types, sources and format), *veracity* (related to the quality and validity of these data), and *velocity* (availability of data in real time). Other than the four big data features described above, we should also consider relevant issues as data trustworthiness, data protection, and data privacy. In this chapter we will explore the major challenges posed by the exploitation of new and alternative data sources, and the associated responses elaborated by the data science community.

2.1 Stewardship and Protection

Accessibility is a major condition for a fruitful exploitation of new data sources for economic and financial analysis. However, in practice, it is often restricted in order to protect sensitive information. Finding a sensible balance between accessibility and protection is often referred to as *data stewardship*, a concept that ranges from properly collecting, annotating, and archiving information to taking a "long-term care" of data, considered as valuable digital assets that might be reused in

[3] Google Cloud: https://cloud.google.com/.

[4] Microsoft Azure: https://azure.microsoft.com/en-us/.

[5] Amazon Web Services (AWS): https://aws.amazon.com/.

future applications and combined with new data [42]. Organizations like the World Wide Web Consortium (W3C)[6] have worked on the development of interoperability guidelines among the realm of open data sets available in different domains to ensure that the data are FAIR (F*indable*, A*ccessible*, I*nteroperable*, and R*eusable*).

Data protection is a key aspect to be considered when dealing with economic and financial data. Trustworthiness is a main concern of individuals and organizations when faced with the usage of their financial-related data: it is crucial that such data are stored in secure and privacy-respecting databases. Currently, various privacy-preserving approaches exist for analyzing a specific data source or for connecting different databases across domains or repositories. Still several challenges and risks have to be accommodated in order to combine private databases by new anonymization and pseudo-anonymization approaches that guarantee privacy. Data analysis techniques need to be adapted to work with encrypted or distributed data. The close collaboration between domain experts and data analysts along all steps of the data science chain is of extreme importance.

Individual-level data about credit performance is a clear example of sensitive data that might be very useful in economic and financial analysis, but whose access is often restricted for data protection reasons. The proper exploitation of such data could bring large improvements in numerous aspects: financial institutions could benefit from better credit risk models that identify more accurately risky borrowers and reduce the potential losses associated with a default; consumers could have easier access to credit thanks to the efficient allocation of resources to reliable borrowers, and governments and central banks could monitor in real time the status of their economy by checking the health of their credit markets. Numerous are the data sets with anonymized individual-level information available online. For instance, mortgage data for the USA are provided by the Federal National Mortgage Association (Fannie Mae)[7] and by the Federal Home Loan Mortgage Corporation (Freddie Mac):[8] they report loan-level information for millions of individual mortgages, with numerous associated features, e.g., repayment status, borrower's main characteristics, and granting location of the loan (we refer to [2, 35] for two examples of mortgage-level analysis in the US). A similar level of detail is found in the European Datawarehouse,[9] which provides loan-level data of European assets about residential mortgages, credit cards, car leasing, and consumer finance (see [20, 40] for two examples of economic analysis on such data).

[6] World Wide Web Consortium (W3C): https://www.w3.org/.

[7] Federal National Mortgage Association (Fannie Mae): https://www.fanniemae.com.

[8] Federal Home Loan Mortgage Corporation (Freddie Mac): http://www.freddiemac.com.

[9] European Datawarehouse: https://www.eurodw.eu/.

2.2 Data Quantity and Ground Truth

Economic and financial data are growing at staggering rates that have not been seen in the past [33]. Organizations today are gathering large volume of data from both proprietary and public sources, such as social media and open data, and eventually use them for economic and financial analysis. The increasing data volume and velocity pose new technical challenges that researchers and analysts can face by leveraging on data science. A general data science scenario consists of a series of observations, often called instances, each of which is characterized by the realization of a group of variables, often referred to as attributes, which could take the form of, e.g., a string of text, an alphanumeric code, a date, a time, or a number. Data volume is exploding in various directions: there are more and more available data sets, each with an increasing number of instances; technological advances allow to collect information on a vast number of features, also in the form of images and videos.

Data scientists commonly distinguish between two types of data, unlabeled and labeled [15]. Given an attribute of interest (label), unlabeled data are not associated with an observed value of the label and they are used in unsupervised learning problems, where the goal is to extract the most information available from the data itself, like with clustering and association rules problems [15]. For the second type of data, there is instead a label associated with each data instance that can be used in a supervised learning task: one can use the information available in the data set to predict the value of the attribute of interest that have not been observed yet. If the attribute of interest is categorical, the task is called classification, while if it is numerical, the task is called regression [15]. Breakthrough technologies, such as deep learning, require large quantities of labelled data for training purposes, that is data need to come with annotations, often referred to as *ground truth* [15].

In finance, e.g., numerous works of unsupervised and supervised learning have been explored in the fraud detection literature [3, 11], whose goal is to identify whether a potential fraud has occurred in a certain financial transaction. Within this field, the well-known Credit Card Fraud Detection data set[10] is often used to compare the performance of different algorithms in identifying fraudulent behaviors (e.g., [17, 32]). It contains 284,807 transactions of European cardholders executed in 2 days of 2013, where only 492 of them have been marked as fraudulent, i.e., 0.17% of the total. This small number of positive cases need to be consistently divided into training and test sets via stratified sampling, such that both sets contain some fraudulent transactions to allow for a fair comparison of the out-of-sample forecasting performance. Due to the growing data volume, it is more and more common to work with such highly unbalanced data set, where the number of positive cases is just a small fraction of the full data set: in these cases, standard econometric analysis might bring poor results and it could be useful investigating rebalancing

[10] https://www.kaggle.com/mlg-ulb/creditcardfraud.

techniques like undersampling, oversampling or a combination of the both, which could be used to possibly improve the classification accuracy [15, 36].

2.3 Data Quality and Provenance

Data quality generally refers to whether the received data are fit for their intended use and analysis. The basis for assessing the quality of the provided data is to have an updated metadata section, where there is a proper description of each feature in the analysis. It must be stressed that a large part of the data scientist's job resides in checking whether the data records actually correspond to the metadata descriptions. Human errors and inconsistent or biased data could create discrepancies with respect to what the data receiver was originally expecting. Take, for instance, the European Datawarehouse presented in Sect. 2.1: loan-level data are reported by each financial institution, gathered in a centralized platform and published under a common data structure. Financial institutions are properly instructed on how to provide data; however, various error types may occur. For example, rates could be reported as fractions instead of percentages, and loans may be indicated as defaulted according to a definition that varies over time and/or country-specific legislation.

Going further than standard data quality checks, *data provenance* aims at collecting information on the whole data generating process, such as the software used, the experimental steps undertaken in gathering the data or any detail of the previous operations done on the raw input. Tracking such information allows the data receiver to understand the source of the data, i.e., how it was collected, under which conditions, but also how it was processed and transformed before being stored. Moreover, should the data provider adopt a change in any of the aspect considered by data provenance (e.g., a software update), the data receiver might be able to detect early a structural change in the quality of the data, thus preventing their potential misuse and analysis. This is important not only for the reproducibility of the analysis but also for understanding the reliability of the data that can affect outcomes in economic research. As the complexity of operations grows, with new methods being developed quite rapidly, it becomes key to record and understand the origin of data, which in turn can significantly influence the conclusion of the analysis. For a recent review on the future of data provenance, we refer, among others, to [10].

2.4 Data Integration and Sharing

Data science works with structured and unstructured data that are being generated by a variety of sources and in different formats, and aims at integrating them into big data repositories or Data Warehouses [43]. There exists a large number of standardized ETL (Extraction, Transformation, and Loading) operations that

help to identify and reorganize structural, syntactic, and semantic heterogeneity across different data sources [31]. Structural heterogeneity refers to different data and schema models, which require integration on the schema level. Syntactic heterogeneity appears in the form of different data access interfaces, which need to be reconciled. Semantic heterogeneity consists of differences in the interpretation of data values and can be overcome by employing semantic technologies, like graph-based knowledge bases and domain ontologies [8], which map concepts and definitions to the data source, thus facilitating collaboration, sharing, modeling, and reuse across applications [7].

A process of integration ultimately results in consolidation of duplicated sources and data sets. Data integration and linking can be further enhanced by properly exploiting information extraction algorithm, machine learning methods, and Semantic Web technologies that enable context-based information interpretation [26]. For example, authors in [12] proposed a semantic approach to generate industry-specific lexicons from news documents collected within the Dow Jones DNA dataset,[11] with the goal of dynamically capturing, on a daily basis, the correlation between words used in these documents and stock price fluctuations of industries of the Standard & Poor's 500 index. Another example is represented by the work in [37], which has used information extracted from the *Wall Street Journal* to show that high levels of pessimism in the news are relevant predictors of convergence of stock prices towards their fundamental values.

In macroeconomics, [24] has looked at the informational content of the Federal Reserve statements and the guidance that these statements provide about the future evolution of monetary policy.

Given the importance of data-sharing among researchers and practitioners, many institutions have already started working toward this goal. The European Commission (EC) has launched numerous initiatives, such as the EU Open Data[12] and the European Data[13] portals directly aimed at facilitating data sharing and interoperability.

2.5 Data Management and Infrastructures

To manage and analyze the large data volume appearing nowadays, it is necessary to employ new infrastructures able to efficiently address the four big data dimensions of volume, variety, veracity, and velocity. Indeed, massive data sets require to be stored in specialized distributed computing environments that are essential for building the data pipes that slice and aggregate this large amount of information. Large unstructured data are stored in distributed file systems (DFS), which join

[11] Dow Jones DNA: https://www.dowjones.com/dna/.

[12] EU Open Data Portal: https://data.europa.eu/euodp/en/home/.

[13] European Data Portal: https://www.europeandataportal.eu/en/homepage.

together many computational machines (nodes) over a network [36]. Data are broken into blocks and stored on different nodes, such that the DFS allows to work with partitioned data, that otherwise would become too big to be stored and analyzed on a single computer. Frameworks that heavily use DFS include Apache Hadoop[14] and Amazon S3,[15] the backbone of storage on AWS. There are a variety of platforms for wrangling and analyzing distributed data, the most prominent of which perhaps is Apache Spark.[16] When working with big data, one should use specialized algorithms that avoid having all of the data in a computer's working memory at a single time [36]. For instance, the MapReduce[17] framework consists of a series of algorithms that can prepare and group data into relatively small chunks (Map) before performing an analysis on each chunk (Reduce). Other popular DFS platforms today are MongoDB,[18] Apache Cassandra,[19] and ElasticSearch,[20] just to name a few. As an example in economics, the authors of [38] presented a NO-SQL infrastructure based on ElasticSearch to store and interact with the huge amount of news data contained in the Global Database of Events, Language and Tone (GDELT),[21] consisting of more than 8 TB of textual information from around 500 million news articles worldwide since 2015. The authors showed an application exploiting GDELT to construct news-based financial sentiment measures capturing investor's opinions for three European countries: Italy, Spain, and France [38].

Even though many of these big data platforms offer proper solutions to businesses and institutions to deal with the increasing amount of data and information available, numerous relevant applications have not been designed to be dynamically scalable, to enable distributed computation, to work with nontraditional databases, or to interoperate with infrastructures. Existing cloud infrastructures will have to massively invest in solutions designed to offer dynamic scalability, infrastructures interoperability, and massive parallel computing in order to effectively enable reliable execution of, e.g., machine learning algorithms and AI techniques. Among other actions, the importance of cloud computing was recently highlighted by the EC through its European Cloud Initiative,[22] which led to the birth of the European Open Science Cloud,[23] a trusted open environment for the scientific community for

[14] Apache Hadoop: https://hadoop.apache.org/.

[15] Amazon AWS S3: https://aws.amazon.com/s3/.

[16] Apache Spark: https://spark.apache.org/.

[17] https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html.

[18] MongoDB: https://www.mongodb.com/.

[19] Apache Cassandra: https://cassandra.apache.org/.

[20] ElasticSearch: https://www.elastic.co/.

[21] GDELT website: https://blog.gdeltproject.org/.

[22] European Cloud Initiative: https://ec.europa.eu/digital-single-market/en/%20european-cloud-initiative.

[23] European Open Science Cloud: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.

storing, sharing, and reusing scientific data and results, and of the European Data Infrastructure,[24], which targets the construction of an EU super-computing capacity.

3 Data Analytics Methods

Traditional nowcasting and forecasting economic models are not dynamically scalable to manage and maintain big data structures, including raw logs of user actions, natural text from communications, images, videos, and sensors data. This high volume of data is arriving in inherently complex high-dimensional formats, and their use for economic analysis requires new tool sets [36]. Traditional techniques, in fact, do not scale well when the data dimensions are big or growing fast. Relatively simple tasks such as data visualization, model fitting, and performance checks become hard. Classical hypothesis testing aimed to check the importance of a variable in a model (T-test), or to select one model across different alternatives (F-test), have to be used with caution in a big data environment [26, 30]. In this complicated setting, it is not possible to rely on precise guarantees upon standard low-dimensional strategies, visualization approaches, and model specification diagnostics [36, 26]. In these contexts, social scientists can benefit from using data science techniques and in recent years the efforts to make those applications accepted within the economic modeling space have increased exponentially. A focal point consists in opening up the black-box machine learning solutions and building interpretable models [22]. Indeed, data science algorithms are useless for policy-making when, although easily scalable and highly performing, they turn out to be hardly comprehensible. Good data science applied to economics and finance requires a balance across these dimensions and typically involves a mix of domain knowledge and analysis tools in order to reach the level of model performance, interpretability, and automation required by the stakeholders. Therefore, it is good practice for economists to figure out what can be modeled as a prediction task and reserving statistical and economic efforts for the tough structural questions. In the following, we provide an high-level overview of maybe the two most popular families of data science technologies used today in economics and finance.

3.1 Deep Machine Learning

Despite long-established machine learning technologies, like Support Vector Machines, Decision Trees, Random Forests, and Gradient Boosting have shown high potential to solve a number of data mining (e.g., classification, regression) problems around organizations, governments, and individuals. Nowadays the

[24]European Data Infrastructure: https://www.eudat.eu/.

technology that has obtained the largest success among both researchers and practitioners is *deep learning* [29]. Deep learning is a general-purpose machine learning technology, which typically refers to a set of machine learning algorithms based on learning data representations (capturing highly nonlinear relationships of low level unstructured input data to form high-level concepts). Deep learning approaches made a real breakthrough in the performance of several tasks in the various domains in which traditional machine learning methods were struggling, such as speech recognition, machine translation, and computer vision (object recognition). The advantage of deep learning algorithms is their capability to analyze very complex data, such as images, videos, text, and other unstructured data.

Deep hierarchical models are Artificial Neural Networks (ANNs) with deep structures and related approaches, such as Deep Restricted Boltzmann Machines, Deep Belief Networks, and Deep Convolutional Neural Networks. ANN are computational tools that may be viewed as being inspired by how the brain functions and applying this framework to construct mathematical models [30]. Neural networks estimate functions of arbitrary complexity using given data. Supervised Neural Networks are used to represent a mapping from an input vector onto an output vector. Unsupervised Neural Networks are used instead to classify the data without prior knowledge of the classes involved. In essence, Neural Networks can be viewed as generalized regression models that have the ability to model data of arbitrary complexities [30]. The most common ANN architectures are the multilayer perceptron (MLP) and the radial basis function (RBF). In practice, sequences of ANN layers in cascade form a deep learning framework. The current success of deep learning methods is enabled by advances in algorithms and high-performance computing technology, which allow analyzing the large data sets that have now become available. One example is represented by robot-advisor tools that currently make use of deep learning technologies to improve their accuracy [19]. They perform stock market forecasting by either solving a regression problem or by mapping it into a classification problem and forecast whether the market will go up or down.

There is also a vast literature on the use of deep learning in the context of time series forecasting [29, 6, 27, 5]. Although it is fairly straightforward to use classic MLP ANN on large data sets, its use on medium-sized time series is more difficult due to the high risk of overfitting. Classical MLPs can be adapted to address the sequential nature of the data by treating time as an explicit part of the input. However, such an approach has some inherent difficulties, namely, the inability to process sequences of varying lengths and to detect time-invariant patterns in the data. A more direct approach is to use recurrent connections that connect the neural networks' hidden units back to themselves with a time delay. This is the principle at the base of Recurrent Neural Networks (RNNs) [29] and, in particular, of Long Short-Term Memory Networks (LSTMs) [25], which are ANNs specifically designed to handle sequential data that arise in applications such as time series, natural language processing, and speech recognition [34].

In finance, deep learning has been already exploited, e.g., for stock market analysis and prediction (see e.g. [13] for a review). Another proven ANNs approach for financial time-series forecasting is the Dilated Convolutional Neural Network presented in [9], wherein the underlying architecture comes from DeepMind's WaveNet project [41]. The work in [5] exploits an ensemble of Convolutional Neural Networks, trained over Gramian Angular Fields images generated from time series related to the Standard & Poor's 500 Future index, where the aim is the prediction of the future trend of the US market.

Next to deep learning, *reinforcement learning* has gained popularity in recent years: it is based on a paradigm of learning by trial and error, solely from rewards or punishments. It was successfully applied in breakthrough innovations, such as the AlphaGo system[25] of Deep Mind that won the Go game against the best human player. It can also be applied in the economic domain, e.g., to dynamically optimize portfolios [23] or for financial assert trading [18]. All these advanced machine learning systems can be used to learn and relate information from multiple economic sources and identify hidden correlations not visible when considering only one source of data. For instance, combining features from images (e.g., satellites) and text (e.g., social media) can yield to improve economic forecasting.

Developing a complete deep learning or reinforcement learning pipeline, including tasks of great importance like processing of data, interpretation, framework design, and parameters tuning, is far more of an art (or a skill learnt from experience) than an exact science. However the job is facilitated by the programming languages used to develop such pipelines, e.g., R, Scala, and Python, that provide great work spaces for many data science applications, especially those involving unstructured data. These programming languages are progressing to higher levels, meaning that it is now possible with short and intuitive instructions to automatically solve some fastidious and complicated programming issues, e.g., memory allocation, data partitioning, and parameters optimization. For example, the currently popular Gluon library[26] wraps (i.e., provides higher-level functionality around) MXNet,[27] a deep learning framework that makes it easier and faster to build deep neural networks. MXNet itself wraps C++, the fast and memory-efficient code that is actually compiled for execution. Similarly, Keras,[28] another widely used library, is an extension of Python that wraps together a number of other deep learning frameworks, such as Google's TensorFlow.[29] These and future tools are creating a world of user friendly interfaces for faster and simplified (deep) machine learning [36].

[25] Deep Mind AlphaGo system: https://deepmind.com/research/case-studies/alphago-the-story-so-far.

[26] Gluon: https://gluon.mxnet.io/.

[27] Apache MXNet: https://mxnet.apache.org/.

[28] Keras: https://keras.io/.

[29] TensorFlow: https://www.tensorflow.org/.

3.2 Semantic Web Technologies

From the perspectives of data content processing and mining, textual data belongs to the so-called unstructured data. Learning from this type of complex data can yield more concise, semantically rich, descriptive patterns in the data, which better reflect their intrinsic properties. Technologies such as those from the Semantic Web, including Natural Language Processing (NLP) and Information Retrieval, have been created for facilitating easy access to a wealth of textual information. The Semantic Web, often referred to as "Web 3.0," is a system that enables machines to "understand" and respond to complex human requests based on their meaning. Such an "understanding" requires that the relevant information sources be semantically structured [7]. Linked Open Data (LOD) has gained significant momentum over the past years as a best practice of promoting the sharing and publication of structured data on the Semantic Web [8], by providing a formal description of concepts, terms, and relationships within a given knowledge domain, and by using Uniform Resource Identifiers (URIs), Resource Description Framework (RDF), and Web Ontology Language (OWL), whose standards are under the care of the W3C.

LOD offers the possibility of using data across different domains for purposes like statistics, analysis, maps, and publications. By linking this knowledge, interrelations and associations can be inferred and new conclusions drawn. RDF/OWL allows for the creation of triples about anything on the Semantic Web: the decentralized data space of all the triples is growing at an amazing rate since more and more data sources are being published as semantic data. But the size of the Semantic Web is not the only parameter of its increasing complexity. Its distributed and dynamic character, along with the coherence issues across data sources, and the interplay between the data sources by means of reasoning, contribute to turning the Semantic Web into a complex, big system [7, 8].

One of the most popular technology used to tackle different tasks within the Semantic Web is represented by NLP, often referred to with synonyms like text mining, text analytics, or knowledge discovery from text. NLP is a broad term referring to technologies and methods in computational linguistics for the automatic detection and analysis of relevant information in unstructured textual content (free text). There has been significant breakthrough in NLP with the introduction of advanced machine learning technologies (in particular deep learning) and statistical methods for major text analytics tasks like: linguistic analysis, named entity recognition, co-reference resolution, relations extraction, and opinion and sentiment analysis [15].

In economics, NLP tools have been adapted and further developed for extracting relevant concepts, sentiments, and emotions from social media and news (see, e.g., [37, 24, 14, 4], among others). These technologies applied in the economic context facilitate data integration from multiple heterogeneous sources, enable the development of information filtering systems, and support knowledge discovery tasks.

4 Conclusions

In this chapter we have introduced the topic of data science applied to economic and financial modeling. Challenges like economic data handling, quality, quantity, protection, and integration have been presented as well as the major big data management infrastructures and data analytics approaches for prediction, interpretation, mining, and knowledge discovery tasks. We summarized some common big data problems in economic modeling and relevant data science methods.

There is clear need and high potential to develop data science approaches that allow for humans and machines to cooperate more closely to get improved models in economics and finance. These technologies can handle, analyze, and exploit the set of very diverse, interlinked, and complex data that already exist in the economic universe to improve models and forecasting quality, in terms of guarantee on the trustworthiness of information, a focus on generating actionable advice, and improving the interactivity of data processing and analytics.

References

1. Aruoba, S. B., Diebold, F. X., & Scotti, C. (2009). Real-time measurement of business conditions. *Journal of Business & Economic Statistics, 27*(4), 417–427.
2. Babii, A., Chen, X., & Ghysels, E. (2019). Commercial and residential mortgage defaults: Spatial dependence with frailty. *Journal of Econometrics, 212*, 47–77.
3. Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2015). *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection.* Chichester: John Wiley & Sons.
4. Barbaglia, L., Consoli, S., & Manzan, S. (2020). Monitoring the business cycle with fine-grained, aspect-based sentiment extraction from news. In V. Bitetta et al. (Eds.), *Mining Data for Financial Applications (MIDAS 2019), Lecture Notes in Computer Science* (Vol. 11985, pp. 101–106). Cham: Springer. https://doi.org/10.1007/978-3-030-37720-5_8
5. Barra, S., Carta, S., Corriga, A., Podda, A. S., & Reforgiato Recupero, D. (2020). Deep learning and time series-to-image encoding for financial forecasting. *IEEE Journal of Automatica Sinica, 7*, 683–692.
6. Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, B., Maddix, D. C., Türkmen, C., Gasthaus, J., Bohlke-Schneider, M., Salinas, D., Stella, L., Callot, L., & Januschowski, T. (2020). Neural forecasting: Introduction and literature overview. CoRR, abs/2004.10240.
7. Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., Lerer, A., & Sheets, D. (2006). Tabulator: Exploring and analyzing linked data on the semantic web. In *Proc. 3rd International Semantic Web User Interaction Workshop (SWUI 2006).*
8. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The story so far. *International Journal on Semantic Web and Information Systems, 5*, 1–22.
9. Borovykh, A., Bohte, S., & Oosterlee, C. W. (2017). Conditional time series forecasting with convolutional neural networks. *Lecture Notes in Computer Science, 10614*, 729–730.
10. Buneman, P., & Tan, W.-C. (2019). Data provenance: What next? *ACM SIGMOD Record, 47*(3), 5–16.
11. Carta, S., Fenu, G., Reforgiato Recupero, D., & Saia, R. (2019). Fraud detection for e-commerce transactions by employing a prudential multiple consensus model. *Journal of Information Security and Applications, 46*, 13–22.

12. Carta, S., Consoli, S., Piras, L., Podda, A. S., & Reforgiato Recupero, D. (2020). Dynamic industry specific lexicon generation for stock market forecast. In G. Nicosia et al. (Eds.), *Machine Learning, Optimization, and Data Science (LOD 2020), Lecture Notes in Computer Science* (Vol. 12565, pp. 162–176). Cham: Springer. https://doi.org/10.1007/978-3-030-64583-0_16
13. Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications, 83*, 187–205.
14. Consoli, S., Tiozzo Pezzoli, L., & Tosetti, E. (2020). Using the GDELT dataset to analyse the Italian bond market. In G. Nicosia et al. (Eds.), *Machine learning, optimization, and data science (LOD 2020), Lecture Notes in Computer Science* (Vol. 12565, pp. 190–202). Cham: Springer. https://doi.org/10.1007/978-3-030-64583-0_18.
15. Consoli, S., Reforgiato Recupero, D., & Petkovic, M. (2019). *Data science for healthcare - Methodologies and applications*. Berlin: Springer Nature.
16. Daily, J., & Peterson, J. (2017). Predictive maintenance: How big data analysis can improve maintenance. In *Supply chain integration challenges in commercial aerospace* (pp. 267–278). Cham: Springer.
17. Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159–166). Piscataway: IEEE.
18. Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2017). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems, 28*(3), 653–664.
19. Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. In *IJCAI International Joint Conference on Artificial Intelligence* (Vol. 2015, pp. 2327–2333).
20. Ertan, A., Loumioti, M., & Wittenberg-Moerman, R. (2017). Enhancing loan quality through transparency: Evidence from the European central bank loan level reporting initiative. *Journal of Accounting Research, 55*(4), 877–918.
21. Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics, 55*(4), 665–676.
22. Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA 2018)* (pp. 80–89).
23. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.
24. Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics, 99*, S114–S133.
25. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.
26. Jabbour, C. J .C., Jabbour, A. B. L. D. S., Sarkis, J., & Filho, M. G. (2019). Unlocking the circular economy through new business models based on large-scale data: An integrative framework and research agenda. *Technological Forecasting and Social Change, 144*, 546–552.
27. Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., & Callot, L. (2020). Criteria for classifying forecasting methods. *International Journal of Forecasting, 36*(1), 167–177.
28. Kuzin, V., Marcellino, M., & Schumacher, C. (2011). MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting, 27*(2), 529–542.
29. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature, 521*(7553), 436–444.
30. Marwala, T. (2013). *Economic modeling using Artificial Intelligence methods*. Heidelberg: Springer.
31. Marx, V. (2013). The big challenges of big data. *Nature, 498*, 255–260.
32. Oblé, F., & Bontempi, G. (2019). Deep-learning domain adaptation techniques for credit cards fraud detection. In *Recent Advances in Big Data and Deep Learning: Proceedings of the INNS Big Data and Deep Learning Conference* (Vol. 1, pp. 78–88). Cham: Springer.

33. OECD. (2015). Data-driven innovation: Big data for growth and well-being. *OECD Publishing, Paris*. https://doi.org/10.1787/9789264229358-en
34. Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting, 36*(3), 1181–1191.
35. Sirignano, J., Sadhwani, A., & Giesecke, K. (2018). Deep learning for mortgage risk. Technical report, Working paper available at SSRN: https://doi.org/10.2139/ssrn.2799443
36. Taddy, M. (2019). *Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions*. New York: McGraw-Hill, US.
37. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139–1168.
38. Tiozzo Pezzoli, L., Consoli, S., & Tosetti, E. (2020). Big data financial sentiment analysis in the European bond markets. In V. Bitetta et al. (Eds.), *Mining Data for Financial Applications (MIDAS 2019), Lecture Notes in Computer Science* (Vol. 11985, pp. 122–126). Cham: Springer. https://doi.org/10.1007/978-3-030-37720-5_10
39. Tiwari, S., Wee, H. M., & Daryanto, Y. (2018). Big data analytics in supply chain management between 2010 and 2016: Insights to industries. *Computers & Industrial Engineering, 115*, 319–330.
40. Van Bekkum, S., Gabarro, M., & Irani, R. M. (2017). Does a larger menu increase appetite? Collateral eligibility and credit supply. *The Review of Financial Studies, 31*(3), 943–979.
41. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). WaveNet: A generative model for raw audio. *CoRR, abs/1609.03499*.
42. Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, 1.
43. Wu, X., Zhu, X., Wu, G., & Ding, W. (2014). Data mining with Big Data. *IEEE Transactions on Knowledge and Data Engineering, 26*(1), 97–107.