Chapter 5
Modelling Complexity with Unconventional Data: Foundational Issues in Computational Social Science

Magda Fontana and Marco Guerzoni

Abstract The large availability of data, often from unconventional sources, does not call for a data-driven and theory-free approach to social science. On the contrary, (big) data eventually unveil the complexity of socio-economic relations, which has been too often disregarded in traditional approaches. Consequently, this paradigm shift requires to develop new theories and modelling techniques to handle new types of information. In this chapter, we first tackle emerging challenges about the collection, storage, and processing of data, such as their ownership, privacy, and cybersecurity, but also potential biases and lack of quality. Secondly, we review data modelling techniques which can leverage on the new available information and allow us to analyse relationships at the microlevel both in space and in time. Finally, the complexity of the world revealed by the data and the techniques required to deal with such a complexity establishes a new framework for policy analysis. Policy makers can now rely on positive and quantitative instruments, helpful in understanding both the present scenarios and their future complex developments, although profoundly different from the standard experimental and normative framework. In the conclusion, we recall the preceding efforts required by the policy itself to fully realize the promises of computational social sciences.

M. Fontana (✉)
Department of Economics and Statistics, University of Turin, Turin, Italy
e-mail: magda.fontana@unito.it

M. Guerzoni
DEMS, University of Milan-Bicocca, Milano, Italy

BETA, University of Strasbourg, Strasbourg, France
e-mail: marco.guerzoni@unimib.it

© The Author(s) 2023

E. Bertoni et al. (eds.), *Handbook of Computational Social Science for Policy*,
https://doi.org/10.1007/978-3-031-16624-2_5

5.1 Introduction

We define CSS as the development and application of computational methods to complex, typically large-scale, human (sometimes simulated) behavioural data (Lazer et al., 2009). The large availability of data, the development of both algorithms and new modelling techniques, and the improvement of storage and computational power opened up a new scientific paradigm for social scientists willing to take into account the complexity of the social phenomena in their research. In this chapter, we develop the idea that the key transformation in place concerns two specific self-reinforcing events:

- (Big) data unveil the complexity of the world and pull for new modelling techniques
- New modelling techniques based on many data push the development of new tools in data science.

We do not share the view that this vast availability of data can allow science to be purely data-driven as in a word without theory (Anderson, 2008; Prensky, 2009). On the contrary, as other authors suggested, science needs more theory to account for the complexity of reality as revealed by the data (Carota et al., 2014; Gould, 1981; Kitchin, 2014; Nuccio & Guerzoni, 2019) and develop new modelling techniques. Obviously, in this age of abundance of information, data analysis occupies a privileged position and can eventually debate with the theory on a level playing field as it has never happened before.

Social sciences are not yet fully equipped to deal with this paradigm shift towards a quantitative, but positive, analysis. Indeed, economics developed an elegant, but purely normative, approach, while other social sciences, when not colonized by the economics' mainstream positive approach, remained mainly qualitative.

Consequently, this present shift can have a profound impact on the way researchers address research questions and, ultimately, also on policy questions. However, before this scientific paradigm unravels its potential, it needs to wind up any uncertainty about its process, specifically around the following issues:

- **Data as the input of the modelling process.** There are three levels of data-related issues:
  - How to collect the data and from which sources.
  - Data storage which relates with ownership, privacy, and cybersecurity.
  - Data quality and biases in data and data collection.
- **New modelling techniques for new data**, which account for heterogeneous, networked, geo-located, and time-stamped data.
- **Policy as the output of the process**, namely, the type of policy questions that can be addressed, e.g. positive vs. normative and prediction vs. causality.

The chapter is organized as follows: Sect. 5.2 frames the topic in the existing literature; Sect. 5.3 addresses the main issues that revolve around the making

of computational social sciences (data sources, modelling techniques, and policy implications); and Sect. 5.4 discusses and concludes.

5.2 Existing Literature

Starting from the late 1980s, sciences have witnessed the increasing influence of complex system analyses. Far from a mechanistic conception of social and economic systems, complex systems pose several challenges to policy making:

- They are comprised of many diverse parts, and, therefore, the use of a representative individual is of no avail (Arthur, 2021).
- They operate on various temporal and spatial scales. It follows that the system behaviour cannot be derived from the mere summation of the behaviour of individual components(Arthur, 2021).
- They operate out of balance, where minor disturbances may lead to events of all dimensions. Thus, most of equilibrium-based theoretical framework used to devise (economic) policy does not apply (Bonabeau, 2002; Fontana, 2012).

In the aftermath of the economic crisis of 2008, the idea that social systems were more complex than what was so far assumed spreads in the policy[1] domain. Meanwhile, the European Union has been placing an increasing focus on complexity-based projects: "In complex systems, even if the local interactions among the various components may be simple, the overall behaviour is difficult and sometimes impossible to predict, and novel properties may emerge. Understanding this kind of complexity is helping to study and understand many different phenomena, from financial crises, global epidemics, propagation of news, connectivity of the internet, animal behaviour, and even the growth and evolution of cities and companies. Mathematical and computer-based models and simulations, often utilizing various techniques from statistical physics are at the heart of this initiative" (Complexity Research Initiative for Systemic InstabilitieS). Furthermore, a growing theoretical literature and the related empirical evidence (Loewenstein & Chater, 2017; Lourenço et al., 2016) spur policy makers to gradually substitute the rational choice framework with the behavioural approach that stresses the limitations in human decision-making. This change in ontology brings about a whole set of new policy features and, subsequently, new modelling challenges. Firstly, since local interaction of heterogeneous agents (consumers, households, states, industries)

[1] See also J. Landau, Deputy Governor of the Bank of France "Complex systems exhibit well-known features: non-linearity and discontinuities (a good example being liquidity freezes); path dependency; sensitivity to initial conditions. Together, those characteristics make the system truly unpredictable and uncertain, in the Knightian sense. Hence the spectacular failure of models during the crisis: most, if not all, were constructed on the assumption that stable and predictable (usually normal) distribution probabilities could be used to describe the different states of the financial system and the economy. They collapsed when extreme events occurred with a frequency that no one ever thought would be possible"(Cooper, 2011).

shapes the overall behaviour and performance of systems, ABMs are used to model the heterogeneity of the system's elements and describe their autonomous interaction. ABMs are computer simulations in which a system is modelled in terms of agents and their interactions (Bonabeau, 2002). Agents, which are autonomous, make decision on the basis of a set or rules and, often, adapt their action to the behaviour of other agents. ABM is being used to inform policy or decisions in various contexts. Recent examples include land use and agricultural policy (Dai et al., 2020), ecosystems and natural resource management (González et al., 2018), control of epidemics (Kerr et al., 2021; Truszkowska et al., 2021), economic policy (Chersoni et al., 2022; Dosi et al., 2020), institutional design (Benthall & Strandburg, 2021), and technology diffusion (Beretta et al., 2018). Moreover, ABM rely on the idea that information does not flow freely and homogeneously within systems and they often connect the policy domain to the field of network science (Kenis & Schneider, 2019). The position of an agent (a state, a firm, a decision-maker) within the network determines its ability to affect its neighbours and vice versa, while the overall structure of the network of agent's connection determines both how rapidly a signal travels the network and its resilience to shocks(Sorenson et al., 2006). Although social network analysis was initiated in the early years of the twentieth century, the last two decades have built on the increased availability of data and of computational resources to inaugurate the study of complex networks, i.e. those networks whose structure is irregular and dynamic and whose units are in the order of millions of nodes (Boccaletti et al., 2006). In the policy perspective, spreading and synchronization processes have a pivotal importance. The diffusion of a signal in a network has been used to model processes such as the diffusion of technologies and to explore static and dynamic robustness (Grassberger, 1983) to the removal of central or random nodes. It is worth noting that ABM and networks can be used jointly. Beretta et al. (2018) use ABM and network to show that cultural dissimilarity in Ethiopian Peasant Associations could impair the diffusion of a subsidized efficient technology, while Chersoni et al. (2021) use agent-based modelling and network analysis to simulate the adoption of technologies under different policy scenarios showing that the diffusion is very sensitive to the network topology. Secondly, the abandonment of the rational choice framework renders the mathematical maximization armoury ineffective and calls for new modelling approaches. The wide range of techniques that fall under the big tent of adaptive behaviours have a tight connections with data and algorithms. In addition to the heuristic and statistical models of behaviour, recent developments have perfected machine learning and evolutionary computation. These improve the representation of agents both by identifying patterns of behaviour in data and also by modelling agents' adaptation in simulations (Heppenstall et al., 2021; Runck et al., 2019). By providing agents with the ability to elaborate different data sources to adapt to their environment and to evolve the rules that are the most suitable response to a given set of inputs, machine learning constitutes an interesting tool to overcome the Lucas' critique. That is to say that it allows modelling individual adjustments to policy making, without renouncing the observed heterogeneity of agents and their dispersed interaction.

Thirdly, the information required to populate these models are not (only) the traditional socio-demographic and national account data but is more diverse and multifaceted.

5.3 Addressing Foundational Issues of CSS for Policy

The development and application of CSS implies a rethinking of the approach to policy making. The increased availability of data provides institutions with abundant information that broadens the spectrum of practicable interventions. Yet, the recognition that economic, social, and ecologic systems are complex phenomena imposes a rethinking of the modelling techniques and of the evaluation of their output. A further layer of complexity is that of data management. While traditional data are collected through institutional channels, new data sources require protocols to establish data ownership and privacy protection. In this section, we propose a three-pronged framework to develop an efficient approach to CSS.

5.3.1 Data as the Input of the Process

There are two main sources of data in the modelling process. On the one hand, the wide application of smart technologies in an increasing number of realms of social and economic life made the presence of sensors ubiquitous: For instance, they record information for machines on the shop floor; register pollution, traffic, and weather data in the smart city; and check and store vital parameters of athletes or sick people. On the other hand, the increased amount of activities occurring on the internet allows for detailed registration of the individual's behaviour with fine-grained details. The extraordinary effort by Blazquez and Domenech (2018) to create a taxonomy of all these possible data sources is a vain one since such an enterprise would require a constant update. However, from their work, it clearly emerges how this new world of data presents peculiar characteristics so far unconventional for the social scientist.

- First of all, most of the data collected are at the microlevel, being the unit of analysis about a single person, a firm, or a specific machine.
- Data are almost always geo-located and time-stamped with a very high precision. In other words, each observation with its attributes occupies a small point in a very dense time-space coordinates.
- There is an unprecedented data collection activity with the focus on interactions. We have at disposal for research any information about commercial transactions among both individuals and firms, which could eventually create a map of the economic activity of a system (Einav & Levin, 2014). At the same time, the advent of social platforms allows to register data on social interactions, which represent the networked world of human relationships.

- There are new data on people's behaviour coming from their searches on search engines, online purchasing activities, and reading and entertainment habits (Bello-Orgaz et al., 2016; Renner et al., 2020).

Moreover, the format of data collected is often unconventional. While statistics has been developed to deal with figures, most of the data available today record texts, images, and videos, only eventually transformed into binary figures by the process of digitization. These types of data convey new content of a paramount importance for the social scientist since it allows to analyse information about ideas, opinions, and feelings (Ambrosino et al., 2018; Fontana et al., 2019).

Thus, different from statistics, which evolved over the last century in time of data scarcity, the present state of the art in the use of data leverages precisely on the vast size of datasets in terms of number of observations, of their attributes, and of different data formats (Nuccio et al., 2020). As a consequence, newly collected data is increasingly stored in the same location, and there is a constant effort to link and merge existing datasets in data warehouses or data lakes. The traditional solution in data science for data storage is a data warehouse, in which data is extracted, transformed, and loaded, while more recently many organizations are opting for a data lake solution, which stores heterogeneously structured raw data from various sources (Ravat & Zhao, 2019). A concurrent and partly connected phenomenon is the widespread adoption of big data analytics as a service (BDaaS), that is, when firms and institutions rely on cloud services on online platforms for the storage and analysis of data (Aldinucci et al., 2018). As a result, there is an increasing presence of very large online databases. This present situation raises the following challenges to CSS.

- The collection, storage, and maintenance of vast datasets create competitive advantages for the private sectors, but it is a very costly process. For these reasons, firms are not always willing to share their data with third parties involved in research- or data-driven policy making. Moreover, even in the presence of an open approach to data sharing, privacy laws do not always allow it without the explicit authorization of subjects providing the data, as in the case of the European Union GDPR[2] (Peloquin et al., 2020; Suman & Pierce, 2018). Public organizations are increasingly digitized and becoming an important hub of data collection of fine-grained data. However, even in the presence of large investments, they often lack adequate human resource and organization capabilities for both the deployment of data warehouse and data lakes and their accessibility for research purposes.
- The capacity for investing in data structure and the ability to collect data are very skewed, with few large players owning a tremendous amount of data. Alphabet's yearly investment in production equipment, facilities, and data centres has been around 10 billion dollars for the last 5 years for the maintenance of about 15 exabytes of data (Nuccio & Guerzoni, 2019). This high concentration

[2] GDPR: https://gdpr-info.eu/

allows a small handful of players to exploit data at a scale incomparable even with the scientific community. As a result, scientific institutions need to rely on partnership with these private players to effectively conduct research.
- The storage of vast amounts also raises issues in cybersecurity since dataset can become a target for unlawful activities due to the monetary value of detailed and sensitive information. Once again, protection against possible cybersecurity threats requires investment in technologies and human capital which only large firms possess. The cost and the accountability involved might discourage the use of data for scientific purposes (Peloquin et al., 2020).

The availability of data does not free social science from its original curse, that is, employing data created elsewhere for different purposes than research. Data, even if very large, might not be representative of a population due to biases in the selection of the sample or because they are affected by measurement errors. Typically, data collected on the internet over-represent young cohorts—which are more prone to the use of technology—or rich households and their related socio-demographic characteristics, since they are rarely affected by the digital divide. Alternatively, data might lack some variables which represent the true key of a phenomenon under investigation. Important attributes might be missing because they are not measured (say expectation on the future) or not available for privacy concern (say gender or ethnicity) (Demoussis & Giannakopoulos, 2006; Hargittai & Hinnant, 2008). As an exemplary case for the depth of this problem, consider the widespread debate on the alleged racism of artificial intelligence. A prediction model might systematically provide biased estimation for individual in a specific ethnicity class, not because it is racist, but because it might be very efficient in fitting the data provided which (a) describe a racist reality, (b) show (over)under-representation of a specific ethnic group, and (c) lack important features (typically income) which might be the true explanation of a phenomenon and they are highly correlated with ethnicity.

In this case, a model can represent very well the data at disposal, but also its possible distortions. Thus, it will fail in being a correct support for policy making or research. It is thus of a paramount importance to have in place data quality evaluation practices (Corrocher et al., 2021). The next section discusses different methodologies at disposal to deal with this large availability of data.

5.3.2 Modelling Techniques for New Data

This availability of data reveals a no longer deniable complexity of the world and opens up for social scientists a vast array of possibilities under the condition that they go beyond "two-variable problem of simplicity" Weaver (1948). We now discuss some theoretical and empirical data techniques which recently reached their mature stage after decades of incubation. As recalled before, data available today are usually at the microlevel, geo-located, time-stamped, and characterized by attributes that described the interaction of the unit of observations both with

other observations and with a non-stationary environment. Take, for instance, the phenomena of localization and diffusion. Geospatial data, initially limited to the study of geographical and environmental issues, are currently increasingly available and accessible. These data are highly complex in that they imply the management of several types of information: physical location of the observation and its attributes and, possibly, temporal information. Complexity further increases since such observations change in accordance to the activities taking place in a given location (e.g. resources depletion, fire diffusion, opinion, and epidemic dynamics) and that the agents undertaking those activities are, in turn, changed by the attributes of the location. The main challenge here is the simultaneous modelling of two independent processes: the interaction of the agents acting in a given location and the adaptation of the attributes of both. Any model attempting to grasp these fined-grained dynamic phenomena should account for these properties.

Agent-Based Modelling

These data are naturally dealt with agent-based modelling and networks. Agent-based modelling describes the system of interest in terms of agents (autonomous individuals with properties, actions, and possibly goals), of their environment (a geometrical, GIS, or network landscape with its own properties and actions), and of agent-agent, agent-environment, and environment-environment interactions that affect the action and internal state of both agents and environment (Wilenski & Rand, 2015). ABMs can be deployed in policy making in several ways. Policy can exploit their ability to cope with complex data, with data and theoretical assumptions (e.g. simulate different diffusion models in an empirical environment), and with interaction and heterogeneity. Literature agrees on two general explanatory mechanisms and three categories of applications. ABMs can be fruitfully applied when there are data or theories on individual behaviour and the overall pattern that emerge from it is unknown, *integrative understanding*, or when there is information on the aggregate pattern and the individual rules of behaviour are not known—*compositional understanding* (Wilenski & Rand, 2015). In both cases, ABM offers insights into policy and interventions in a prospective and/or retrospective framework.[3] Prospective models simulate the design of policies and investigate their potential effects. Since they rely on non-linear out-of-equilibrium theory, they can help in identifying critical thresholds and tipping points, i.e. small interventions that might trigger radical and irreversible changes in the system of interest (Bak et al., 1987). These are hardly treated with more traditional techniques. The identification of early warning signals of impending shifts (Donangelo et al., 2010) relies on the observation of increasing variance and changes in autocorrelation and skewness in time series data; however, traditional data are often too coarse-grained and cover a time window that is too small with respect to the rate of change of the system. Empirically calibrated ABMs instead can simulate long-term dynamics and the related interventions (see, for instance, Gualdi et al. (2015).

[3] This classification is proposed and discussed at length by Hammond (2015).

When multiple systems are involved—say, the economy and the environment—ABMs map the trade-offs or synergies of policies across qualitatively different systems. Moreover, they are useful to highlight the unintended or unexpected consequences of the interventions, especially when in vivo or in vitro experiments are expensive, unpractical, or unethical. Retrospective models are useful especially under the compositional understanding framework. Firstly, they can investigate why policy have or have not played out the way they were expected to. This is relevant especially when data do not exist. For instance, Chersoni et al. (2021) study the reasons behind the underinvestment in energy-efficient technologies in Europe in spite the EU-wide range of interventions. They start from data on households and simulate their—unobserved—connections to show that policy should account for behavioural and imitative motives beyond the traditional financial incentives. While retaining the heterogeneity of observations, ABMs can also reveal different effects of policies across sub-samples of the population. Retrospective models can be used in combination with the prospective ones as input in the policy design process.

Network Modelling

Policy can exploit the theoretical and empirical mapping provided by network modelling to improve the knowledge of the structure of connections among the elements of the systems of interest, to reinforce the resulting networks, and to guide the processes that unfold on it. Network modelling elaborate on the mapping by computing metrics (e.g. density, reciprocity, transitivity, centralization, and modularity) to characterize the network and to quantify its dimensions. The features associated with those metrics are key to understand the robustness of network to random or target nodes and to study the speed at which a signal travels on it. Once the structure of the network is known, policy makers can design their intervention in order to foster or prevent the processes that are driven by local interactions. For instance, it has been shown that small world networks maximize diffusion (Schilling & Phelps, 2007) and that policy that encourage the formation of distant connections can sustain the production of scientific knowledge (Chessa et al., 2013). The identification of pivotal nodes, on the other hand, allows the design of policy that target the most central or fragile components of the networks. Network modelling also contributes to the identification of tipping points and to the elaboration of the required preventive policies. If the elements of the systems are connected through a preferential attachment topology (for instance, the world banking system Benazzoli and Di Persio, 2016), then the system could experience radical and irreversible change if the most central nodes are hit, while it is resilient to random node removals (Eckhoff & Morters, 2013).

Explaining, Predicting, and Summarizing

That traditional modelling techniques based on optimization naturally suggest simple closed-form equations apt to be tested with econometric techniques does not come as a surprise. The funding father of econometrics Ragnar Frisch clearly emphasized the ancillary role of data analysis in economics with respect to the neoclassical theorizing by stating that econometrics should achieve “the advance-

ment of economic theory in its relation to statistics and mathematics" (Cowles, 1960) and not vice versa. However, the complexity of the world now revealed and measured by new data and modelled by networks and ABM pushes for an evolution in the analysis of data. There exist three types of approach in data analysis: causal explanation, prediction, and summarization of the data. Guerzoni et al. (2021) explain that the specificity of the three approaches with the most severe consequences is the way they deal with external validity. Standard econometrics techniques rely on inference, and the properties of estimators have been derived under strong assumptions on error distribution and for a small class of simple and usually linear models: the focus is on the creation of reliable sample either via experiments or by employing instrumental variables to account for possible endogeneity of the data. As a consequence, econometrics manages to be robust in terms of identifying specific causal relationships, but at the cost of a reduced model fitness, since simple and mostly linear models are always inappropriate to fit the complexity of the data. Moreover, further issues such as the number of degrees of freedom and multicollinearity reduce the use of a large number of variables. While a scientist might be satisfied with sound evidence on casual relationships, for policy making, this is a truly unfortunate situation. Knowing the causal impact of a policy measure on a target variable is surely important, but useless if this impact accounts for a tiny percentage of the overall variation of the target.

On the contrary, prediction models measure their uncertainty by looking at the accuracy of prediction on out-of-sample data. There are no restrictions in the type or complexity of the models (or combination of models), and the most advanced data processing techniques such as deep learning can fully displace their power. In this way, the prediction of future scenarios became possible at the expense of eliciting specific causal effects. The trade-off, known as bias-variance trade-off, is clear: On the one hand, simple econometric models allow us to identify an unbiased sample average response at the cost of inhibiting any accuracy of fitness. On the other hand, complex prediction models reach remarkable level of accuracy, even on the single future observation, but they are silent on specific causal relations. In this situation, the importance of complex theoretical approaches such as the ABM or network modelling becomes clear. Indeed, predicted result can be used to evaluate the rules of the model and the parameter settings. A complex theoretical model fine-tuned with many data and in line with predictions can be rather safely employed for policy analysis since it incorporates both theoretical insights on causal relationship and a verified prediction power (Beretta et al., 2018).

Lastly, summarizing techniques serve the purpose of classifying and displaying, often with advanced visualization, properties of the data. Traditionally, the taxonomic approach to epistemology, that is, to create a partition of empirical observations based on their characteristics, has been carried on by a careful qualitative evaluation of data made by the researcher. In the words of most philosophers of science, classification is a mean to "bring related items together" (Wynar et al., 1985, p. 317), "putting together like things" (Richardson, 1935); (Svenonius, 2000, p. 10), and "putting together things that are alike" (Vickery, 1975, p. 1) (see Mai, 2011 for a review). Of course, the antecedent of this approach dates back in the

Aristotelian positive approach to science, which describes and compares vis-á-vis Plato's normative approach (Reale, 1985). More recently, the availability of large dataset made a qualitative approach to the creation of taxonomic possible only at the expense of a sharp a priori reduction of the information in data. However, at the same time, algorithms and computational power allow for an automatic elaboration of the information with the purpose of creating a taxonomy. This approach is known as pattern recognition, unsupervised machine learning, or clustering and has been introduced in science by the anthropologists Driver and Kroeber (1932) and the psychologists Zubin (1938) and Tyron (1939). Typically, unsupervised algorithms are fed by rich datasets in terms of both variables and observations and require as main output the number of groups to be identified from the researcher. On this basis, as output, they provide a classification which minimize within-group variation and maximize between-groups variation, usually captured by some measures of distance in the n-dimensional space of the n variables. Although among these methods in social science the use of K-means algorithm MacQueen et al. (1967) is the most widespread, it has some weaknesses such as a possible dependence by initial condition and the risk of lock-in in local optima. More recently, the *self-organizing maps* (SOM) (Kohonen, 1990) gained attention as a new method in pattern recognition since they improve on K-means and present other advantages such as a clear visualization of the results.[4]

Data Analysis for Unconventional Data Sources

Also, the large share of unconventional data sources such as texts, images, or videos requires new techniques in the scientist's toolbox, and the nature of such information is more informative than figures since it contains ideas, opinions, and judgments. However, as for numerical data, the challenge is to reduce and organize such information in a meaningful way with the purpose of using it for a quantitative analysis, which does not require the time-consuming activity of reading and watching. Concerning text mining, the term "distant reading" attributed to Moretti (2013) could be used as an umbrella definition encompassing the use of automatic information process for books. The large divide for text mining is between the unsupervised and the supervised approach. The former usually deals with corpora of many documents which need to be organized. Techniques such as topic modelling allow to extrapolate the hidden thematic structure of an archive, that is, they highlight topics as specific distribution of words which are likely to occur together (Blei et al., 2003). Moreover, they return also the relative distributions of such topics in each document. Thus, at the same time, it is possible to have a bird's-eye overview on the key concept discussed in an archive, their importance, and when such concepts occur together in the different documents. The exact nature of the topic depends on the exercise at hand and, as in any unsupervised model, is subject to the educated interpretation of the researcher. Ultimately, it is possible

[4] For example, in the use of SOM for policy making, see, for instance, Carlei and Nuccio (2014) and Nuccio et al. (2020).

also to automatically assign a topic distribution to a new document, evaluate the emergence of new instances and the disappearing of old ones, and monitor how the relative importance of different instances changes over time (Di Caro et al., 2017). Topic modelling has been employed for the analysis of scientific literature (Fontana et al., 2019), policy evaluation (Wang & Li, 2021), legal documents (Choi et al., 2017), and political writings and speeches (Greene & Cross, 2017).

Documents can also come as an annotated text, that is, a text in which words or sentence of a group of documents is associated with a category. For instance, each word can be assigned to a feeling (bad vs. good), an evaluation (positive vs. negative), or an impact (relevant vs. non-relevant). The annotated text can be used as a training dataset to train a model able to recognize and predict the specific category in new document and analyse them. In this vain, a dataset of annotated tweets returning the feeling of the author can be used to infer the feeling of other users or the average sentiment of a geographical area or a group of people. For instance, Dahal et al. (2019) analyse the sentiment of climate change tweets.

5.3.3 Policy Recommendation as an Output of the Process

Based on the above review, it is possible to discuss which policies can be expected as outcome of a data-driven approach. The main theoretical element brought forward in the previous paragraphs consists in the link between the complexity of the world revealed by the data and the techniques require to deal with such a complexity. Such element constitutes the foundation of CSS and establishes the consequences for its use for policy analysis.

Precisely, since theoretical models such as ABM and network modelling lack the ability to come forward with simple testable equations, the attempt of deriving clear causal links as tool for policy should be abandoned. Nevertheless, it is not necessary to look back in despair since the fine and elegant armoury of causal identification has been developed in a century of scarcity of data and it made the best under such circumstances. However, as discussed above, the use of data was only ancillary to positive theorizing, and such an impoverished use of data science made prediction and quantitative scenario analysis ineffective for the policy maker. Nowadays, the combination of complex modelling with prediction empowers a truly quantitative policy analysis: on the one hand, relation among variables is hypothesized and tested within theoretical, but positive, models taking into account the heterogeneous attributes (also behavioural) of the subjects, the temporal and geographical specificity, and the dense interactions and feedback in the systems. The fine-tuning of their parameters and accuracy of their prediction are evaluated with supervised algorithms. Moreover, the unsupervised approach allows also for a hypothesis-free and easy-to-visualize exploration of data.

Such a positive and quantitative analysis can be helpful in understanding the present scenarios and their future complex development as the result of interactions of complex elements such as in the case of contagion of diseases (Currie et al.,

2020), but also the diffusion of ideas, technologies, and information as the aggregated manifestation of underlying adoption decisions (Beretta et al., 2018). Note that according to data at the disposal, these results can be achieved either by fine-tuned modelling with micro- and behavioural data, which return predictions on aggregate behaviour, or, on the contrary, by theoretical models which infer micro-behaviour when the model can replicate aggregate results in an exercise of compositional understanding. Economic systems can be also depicted as a complex evolving system, and CSS can describe aggregate fluctuation in economics and finances by feeding with data at the microlevel ABM models which can predict aggregate fluctuations with much fine accuracy than present DSGE models (Dosi & Roventini, 2019). Predictions aside, these models can easily incorporate heterogeneity of the agents, such as in income distribution or different behavioural routines such as propensity to save for consumers or to invest for entrepreneurs.

Finally, the discussion holds not only for policy analysis but also for the corresponding process of policy monitoring and evaluation. The current state of the art in scientific literature suggests that it is possible to evaluate the single causal impact of a policy, but this is far to be true: even in a controlled policy field experiment, it is not possible to estimate the external validity of the results when a pilot policy instrument is deployed at the country level, that is, at a different scale of complexity, or repeated in a slightly different situation in which local attributes are different.

5.4 The Way Forward

Data and algorithms applied to CSS can heavily impact upon the way we conceive the process of policy generation. However, the adoption of such tools needs preceding efforts by the policy itself, mainly in the areas of data as an input.

The ability of the public infrastructure to gather and store data for many sources calls for investment in technology, human capital, and a legislation that find a fine balance between citizens' right for privacy and a flexible use of the data.

The storage and the computational power of large amount of data should not rely on foreign service providers since data should be subject to European regulation. Therefore, policies within the European Data Infrastructure such as the European Open Science Cloud are welcome as well as the high prioritization of technological infrastructure in the European Regional Development Fund.[5]

Data collected and stored in Europe are subject to the GDPR which is correctly concerned with citizens' privacy protection. Although art. 89 allows research of certain privileges in data handling, the regulation is silent about the use for research

[5] European Data Infrastructure, https://www.eudat.eu/; European Open Science Cloud, https://eosc-portal.eu/; European Regional Development Fund, https://ec.europa.eu/regional_policy/en/funding/erdf/

of privately gathered data, de facto providing a solid reason to a large private platform not to share their data. This is in area in which the policy maker could intervene with the purpose of facilitating public-private data exchange for achieving the purpose of public interest.

The management of data and CSS also requires investments in human capital. The introduction of new professional profiles such as data stewards is required to deal with legislation and technical issue related to data, and the introduction of university curricula in data science should be encouraged. Moreover, due to variegated mix of skills that are required to apply CSS, interdisciplinary research should be supported and promoted.

References

Aldinucci, M., Rabellino, S., Pironti, M., Spiga, F., Viviani, P., Drocco, M., Guerzoni, M., Boella, G., Mellia, M., Margara, P., Drago, I., Marturano, R., Marchetto, G., Piccolo, E., Bagnasco, S., Lusso, S., Vallero, S., Attardi, G., Barchiesi, A., ... Galeazzi, F. (2018). HPC4AI: an ai-on-demand federated platform endeavour. In *Proceedings of the 15th ACM International Conference on Computing Frontiers* (pp. 279–286).

Ambrosino, A., Cedrini, M., Davis, J. B., Fiori, S., Guerzoni, M., & Nuccio, M. (2018). What topic modeling could reveal about the evolution of economics. *Journal of Economic Methodology, 25*(4), 329–348.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine, 16*(7), 16–07.

Arthur, W. B. (2021). Foundations of complexity economics. *Nature Reviews Physics, 3*(2), 136–145.

Bak, P., Tang, C., & Wiesenfeld, K. (1987). Self-organized criticality: An explanation of the 1/f noise. *Physical Review Letters, 59*, 381–384. https://doi.org/10.1103/PhysRevLett.59.381. https://link.aps.org/doi/10.1103/PhysRevLett.59.381

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45–59.

Benazzoli, C., & Di Persio, L. (2016). default contagion in financial networks. *International Journal of Mathematics and Computers in Simulation, 10*, 112–117.

Benthall, S., & Strandburg, K. J. (2021). Agent-based modeling as a legal theory tool. *Frontiers in Physics, 9*, 337. ISSN 2296-424X. https://doi.org/10.3389/fphy.2021.666386. https://www.frontiersin.org/article/10.3389/fphy.2021.666386

Beretta, E., Fontana, M., Guerzoni, M., & A. Jordan. (2018). Cultural dissimilarity: Boon or bane for technology diffusion? *Technological Forecasting and Social Change, 133*, 95–103.

Blazquez, D., & Domenech, J. (2018). Big data sources and methods for social and economic analyses. *Technological Forecasting and Social Change, 130*, 99–113.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports, 424*(4), 175–308. ISSN 0370-1573. https://doi.org/10.1016/j.physrep.2005.10.009. https://www.sciencedirect.com/science/article/pii/S037015730500462X

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences, 99*(Suppl 3), 7280–7287. ISSN 0027-8424. https://doi.org/10.1073/pnas.082080899. https://www.pnas.org/content/99/suppl_3/7280

Carlei, V., & Nuccio, M. (2014). Mapping industrial patterns in spatial agglomeration: A som approach to italian industrial districts. *Pattern Recognition Letters, 40*, 1–10.

Carota, C., Durio, A., & Guerzoni, M. (2014). An application of graphical models to the innobarometer survey: A map of firms' innovative behaviour. *Italian Journal of Applied Statistics 25*(1), 61–79.

Chersoni, G., Della Valle, N., & Fontana, M. (2021). The role of economic, behavioral, and social factors in technology adoption. In Ahrweiler P. & Neumann M. (Eds.), *Advances in Social Simulation. ESSA 2019. Springer Proceedings in Complexity*. Cham: Springer.. https://doi.org/10.1007/978-3-030-61503-1_44

Chersoni, G., Della Valle, N., & Fontana, M. (2022). Modelling thermal insulation investment choices in the eu via a behaviourally informed agent-based model. *Energy Policy, 163*, 112823.

Chessa, A., Morescalchi, A., Pammolli, F., Pennera, O., Petersen, A. M., & Riccaboni, M. (2013). Is Europe evolving toward an integrated research area? *Scince, 339*, 650–651.

Choi, H. S., Lee, W. S., & Sohn, S. Y. (2017). Analyzing research trends in personal information privacy using topic modeling. *Computers & Security, 67*, 244–253.

Cooper, M. (2011). Complexity theory after the financial crisis: The death of neoliberalism or the triumph of Hayek?. *Journal of Cultural Economy, 4*(4), 371–385.

Corrocher, N., Guerzoni, M., & Nuccio, M. (2021). Innovazione e algoritmi da maneggiare con cura. *Economia & Management: la rivista della Scuola di Direzione Aziendale dell'Università L. Bocconi, 2*, 17–20.

Cowles, A. (1960). Ragnar frisch and the founding of the econometric society. *Econometrica (pre-1986), 28*(2), 173.

Currie, C. S., Fowler, J. W., Kotiadis, K., Monks, T., Onggo, B. S., Robertson, D. A., & Tako, A. A. (2020). How simulation modelling can help reduce the impact of COVID-19. *Journal of Simulation, 14*(2), 83–97.

Dahal, B., Kumar, S. A., & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining, 9*(1), 1–20.

Dai, E., Ma, L., Yang, W., Wang, Y., Yin, L., & Tong, M. (2020). Agent-based model of land system: Theory, application and modelling frameworks. *Journal of Geographical Sciences, 30*, 1555–1570.

Demoussis, M., & Giannakopoulos, N. (2006). Facets of the digital divide in europe: Determination and extent of internet use. *Economics of Innovation and New Technology, 15*(03), 235–246.

Di Caro, L., Guerzoni, M., Nuccio, M., & Siragusa, G. (2017). A bimodal network approach to model topic dynamics. *Preprint arXiv:1709.09373*.

Donangelo, R., Fort, H., Dakis, V., Scheffer, M., & Van Nes, E. H. (2010). Early warnings for catastrophic shifts in ecosystems: Comparison between spatial and temporal indicators. *International Journal of Bifurcation and Chaos, 20*(02), 315–321. https://doi.org/10.1142/S0218127410025764

Dosi, G., Pereira, M., Roventini, A., & Virgillito, M. (2020). The labour-augmented k+s model: A laboratory for the analysis of institutional and policy regimes. *Economi A, 21*(2), 160–184. ISSN 1517-7580. https://doi.org/10.1016/j.econ.2019.03.002. https://www.sciencedirect.com/science/article/pii/S151775801830122X

Dosi, G., & Roventini, A. (2019). More is different... and complex! the case for agent-based macroeconomics. *Journal of Evolutionary Economics, 29*(1), 1–37.

Driver, H., & Kroeber, A. (1932). *Quantitative expression of cultural relationships* (Vol. 31, pp. 211–256). University of California publications in American Archaeology and Ethnology. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Driver 21131 Quantitative Expression of Cultural Relationships 1932*.

Eckhoff, M., & Morters, P. (2013). Vulnerability of robust preferential attachment networks. *Electronic Journal of Probability, 19*, 1–47.

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

Fontana, M. (2012). On policy in non linear economic systems. In Heritier, P. & Silvestri, P. (Eds.), *Good goverment governance and human complexity* (pp. 221–234). Oelscki.

Fontana, M., Montobbio, F., & Racca, P. (2019). Topics and geographical diffusion of knowledge in top economic journals. *Economic Inquiry, 57*(4), 1771–1797. https://doi.org/10.1111/ecin.12815

González, I., D'Souza, G., & Ismailova, Z. (2018). Agent-based modeling: An application to natural resource management. *Journal of Environmental Protection, 9*, 991–1019.

Gould, P. (1981). Letting the data speak for themselves. *Annals of the Association of American Geographers, 71*(2), 166–176.

Grassberger, P. (1983). On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences, 63*(2), 157–172. ISSN 0025-5564. https://doi.org/10.1016/0025-5564(82)90036-0. https://www.sciencedirect.com/science/article/pii/0025556482900360

Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis, 25*(1), 77–94.

Gualdi, S., Tarzia, M., Zamponi, F., & Bouchaud, J.-P. (2015). Tipping points in macroeconomic agent-based models. *Journal of Economic Dynamics and Control, 50*, 29–61. ISSN 0165-1889. https://doi.org/10.1016/j.jedc.2014.08.003. https://www.sciencedirect.com/science/article/pii/S0165188914001924. Crises and Complexity.

Guerzoni, M., Nava, C. R., & Nuccio, M. (2021). Start-ups survival through a crisis. combining machine learning with econometrics to measure innovation. *Economics of Innovation and New Technology, 30*(5), 468–493.

Hammond, R. (2015). *Considerations and best practices in agent-based modeling to inform policy*. Wahsington, DC, USA: National Academies Press.

Hargittai, E., & Hinnant, A. (2008). Digital inequality: Differences in young adults' use of the internet. *Communication Research, 35*(5), 602–621.

Heppenstall, A., Crooks, A., Malleson, N., Manley, E., Ge, J., & Batty, M. (2021). Future developments in geographical agent-based models: Challenges and opportunities. *Geographical Analysis, 53*(1), 76–91. https://doi.org/10.1111/gean.12267. https://onlinelibrary.wiley.com/doi/abs/10.1111/gean.12267

Kenis, P., & Schneider, V. (2019). *Analyzing policy-making II: Policy network analysis* (pp. 471–491). Springer. ISBN 9783030160647. https://doi.org/10.1007/978-3-030-16065-4_27.

Kerr, C. C., Stuart, R. M., Mistry, D., Abeysuriya, R. G., Rosenfeld, K., & Hart, G. R. (2021). Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLoS Computational Biology, 17*(7), e1009149.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 2053951714528481.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78*(9), 1464–1480.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Social science. computational social science. *Science (New York, NY), 323*(5915), 721–723.

Loewenstein, G., & Chater, N. (2017). Putting nudges in perspective. *Behavioural Public Policy, 1*(1), 26–53. https://doi.org/10.1017/bpp.2016.7

Lourenço, J. S., Ciriolo, E., Rafael Almeida, S., & Troussard, X. (2016). *Behavioural insights applied to policy, european report 2016*. EUR 27726.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). Oakland, CA, USA.

Mai, J.-E. (2011). The modernity of classification. *Journal of Documentation, 67*(4), 710–730.

Moretti, F. (2013). *Distant reading*. Verso Books.

Nuccio, M., & Guerzoni, M. (2019). Big data: Hell or heaven? Digital platforms and market power in the data-driven economy. *Competition & Change, 23*(3), 312–328.

Nuccio, M., Guerzoni, M., Cappelli, R., & Geuna, A. (2020). Industrial pattern and robot adoption in European regions. *Department of Management, Università Ca'Foscari Venezia Working Paper, 1*(3), 33.

Peloquin, D., DiMaio, M., Bierer, B., & Barnes, M. (2020). Disruptive and avoidable: GDPR challenges to secondary research uses of data. *European Journal of Human Genetics, 28*(6), 697–705.

Prensky, M. (2009). H. sapiens digital: From digital immigrants and digital natives to digital wisdom. *Innovate: Journal of Online Education, 5*(3).

Ravat, F., & Zhao, Y. (2019). Data lakes: Trends and perspectives. In *International Conference on Database and Expert Systems Applications* (pp. 304–313). Springer.

Reale, G. (1985). *A History of Ancient philosophy II: Plato and Aristotle* (Vol. 2). Suny Press.

Renner, K.-H., Klee, S., & von Oertzen, T. (2020). Bringing back the person into behavioural personality science using big data. *European Journal of Personality, 34*(5), 670–686.

Richardson, E. C. (1935). *Classification*. New York: H. W. Wilson.

Runck, B., Manson, S., Shook, E., Gini, M., & Jordan, N. (2019). Using word embeddings to generate data-driven human agent decision-making from natural language. *GeoInformatica, 23*, 221–242.

Schilling, M. A., & Phelps, C. C. (2007). Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management Science, 53*(7), 1113–1126. https://doi.org/10.1287/mnsc.1060.0624.

Sorenson, O., Rivkin, J. W., & Fleming, L. (2006). Complexity, networks and knowledge flow. *Research Policy, 35*(7), 994–1017.

Suman, A. B., & Pierce, R. (2018). Challenges for citizen science and the eu open science agenda under the gdpr. *European Data Protection Law Review, 4*, 284.

Svenonius, E. (2000). *The intellectual foundation of information organization*. MIT Press.

Truszkowska, A., Behring, B., Hasanyan, J., Zino, L., Butail, S., Caroppo, E., Jiang, Z.-P., Rizzo, A., & Porfiri, M. (2021). High-resolution agent-based modeling of COVID-19 spreading in a small town. *Advanced Theory and Simulations, 4*(3), 2000277. https://doi.org/10.1002/adts.202000277

Tyron, R. C. (1939). *Cluster analysis*. Ann Arbor, MI: Edwards Brothers.

Vickery, B. C. (1975). *Classification and indexing in science* (3rd ed.).

Wang, Q., & Li, C. (2021). An evolutionary analysis of new energy and industry policy tools in china based on large-scale policy topic modeling. *Plos one, 16*(5), e0252502.

Weaver, W. (1948). There is a large literature on the subject of complexity, for example. *Science and Complexity, 36pp*, 536–544.

Wilenski, U., & Rand, W. (2015). *An introduction to agent-based modeling modeling natural, social, and engineered complex systems with NetLogo*. Massachusetts London, England,: The MIT Press Cambridge.

Wynar, B. S., Taylor, A. G., & Osborn, J. (1985). *Introduction to cataloging and classification* (Vol. 8). Libraries Unlimited Littleton.

Zubin, J. (1938). A technique for measuring like-mindedness. *The Journal of Abnormal and Social Psychology, 33*(4), 508.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.