Agent-Based Modelling as a Foundation for Big Data

Shu-Heng Chen*

Ragupathy Venkatachalam †

Abstract

In this article we propose a *process-based* definition of big data, as opposed to the size - and technology-based definitions. We argue that big data should be perceived as a continuous, unstructured and unprocessed dynamics of primitives, rather than as points (snapshots) or summaries (aggregates) of an underlying phenomenon. Given this, we show that big data can be generated through agent-based models but not by equation-based models. Though statistical and machine learning tools can be used to analyse big data, they do not constitute a big data-generation mechanism. Furthermore, agent-based models can aid in evaluating the quality (interpreted as information aggregation efficiency) of big data. Based on this, we argue that agent-based modelling can serve as a possible foundation for big data. We substantiate this interpretation through some pioneering studies from the 1980s on swarm intelligence and several prototypical agent-based models developed around the 2000s.

Keywords: Big Data, Swarm, Prediction Markets, Information Aggregation, Agent-based Models, Abduction

1 Introduction

Few developments have garnered the imagination and attention of researchers and the public alike in recent years in the way big data has. The scale of data becoming available for academic and commercial research has been growing exponentially since the beginning of the 2000s and this has opened many possibilities, concerns and speculations. This *big data phenomenon* has enjoyed a sweeping reach across different disciplines ranging from particle physics to history, and from epidemiology (Howe et al., 2008) to economics (Einav and Levin, 2014; Varian, 2014).

*AI-ECON Research Center, Department of Economics, National Chengchi University, Taipei, Taiwan 11605, E-mail: chen.shuheng@gmail.com

†Institute of Management Studies, Goldsmiths, University of London, UK. Email: rpathy@gmail.com

Despite the brouhaha, however, there is relatively little research devoted to the ontological nature of big data, methodological aspects and its epistemological possibilities (and the associated boundaries). These questions need to be addressed in order to gain a deeper understanding of big data and to harness its potential.

The term *big data* is often used alongside the term 'computational social sciences', possibly due to the important role that data mining and computational tools play in organizing, processing and analysing high-dimensional, often unstructured data. The popularity of the term big data is however more recent in comparison to the inception of computational social science (CSS) as an independent research field. For instance, consider the authoritative four volume collection of articles (Gilbert, 2010). We find 66 influential papers in the field of CSS, published all the way from the 1960s to the beginning of the twenty-first century. These articles include some of the pioneering papers that exemplify the field of CSS. However, in these four volumes, one fails to find explicit references to what is now known as the big data phenomenon. The articles in Gilbert (2010) begin with agent-based models and end with a methodological part and much of the attention is devoted to addressing various issues related to model-building, such as verification, validation and principles related to agent-based modelling. However, there is a noticeable change that occurred in the literature developing since the late 2000s, (i.e., roughly in the years after the survey by Nigel Gilbert), where big data is explicitly included as an essential part of CSS (Cioffi-Revilla, 2013). Some scholars even identify CSS predominantly with big data (Lazer et al., 2009; Alvarez, 2016).[1] This shift in focus also places little emphasis on agent-based modelling and its role in CSS.[2]

Some early and well-cited examples (Epstein and Axtell, 1996; Epstein, 2007) and other classic contributions in the area of CSS are collected in Gilbert (2010).[3] In this collection, CSS is mainly characterized as the employment of a new kind of model, known as the *agent-based model*, for simulating and understanding a variety of social phenomena.[4] The theoretical underpinnings of these models are traced back to John von Neumann's work on the theory of automata and the later contributions by John Conway and Stephen Wolfram. These initial efforts and later studies (Schelling, 1971) attempt to better understand the nature and functioning of the physical universe and social systems through a class of computational models. These models can be seen as providing a fundamentally different and novel way to view how various aggregate patterns can be generated bottom up from individual level phenomena. More importantly, they attempt to distinguish their generation process from the conventional equation-based models. Hence,

the essence of agent-based modelling can thus be seen as a new way of thinking about the 'data-generation process', instead of data per se, regardless of its size or granularity.

It may be argued that the nature of the data-generation process in social systems has remained largely unchanged. However, since the late 2000s, mainly due to the developments in information and communications technology (ICT), there have been remarkable advances in the areas of data-collection and data-archiving processes. A variety of data which can only be imagined or artificially generated and analysed in CSS have now become viable and available in reality for social scientists to work with. These developments in CSS can be seen as providing a new momentum and possibly enable it to broach new frontiers which were previously not possible. As mentioned earlier, this development has, intentionally or otherwise, shifted the attention of the scholars from data generation to that of data and its practical uses. This trend is best exemplified in the use and abuse of the term 'data-driven science', with little focus on underlying theories or even being 'theory-free'. This unfortunate shift has also managed to confine CSS to a subject that merely engages in data mining or knowledge discovery. Given this, we can ask whether big data is a mere neologism or *is there something genuinely novel about it*?

Despite the exponential growth in the number of research articles on this subject, there is little research to understand this big data phenomenon from an ontological, epistemological and methodological perspectives. Neither is there a systematic effort to search for possible theoretical foundations, especially from the viewpoint of the social sciences. This paper is an attempt to fill this gap. First, the paper highlights that the existing definitions of big data in the literature do not provide sharp ontological boundaries to distinguish it from other forms of data. We propose and adopt a *process-based* definition and argue that big data should be perceived as a continuous, unstructured and unprocessed dynamics of primitives, rather than points (snapshots) or summaries (aggregates) of an underlying phenomenon. Second, by focusing on the data-generation aspect of agent-based models and based on our definition, we demonstrate that they are capable of generating big data. In fact, it is the only model that can generate big data to the best of our knowledge. Third, in so far as social phenomena can be viewed as resulting from the actions of individuals and their interactions, we argue that the quality of big data generated in the social sciences needs to be assessed based on its information aggregation efficiency (along the lines of the Hayek hypothesis (Hayek, 1945; Smith, 1982)). To this end, we show that agent-based models provide a way to achieve this goal through two different examples. Finally, based on its ability to generate big data, its power to be able to

'speak' to big data, and its ability to serve as an abductive tool for knowledge discovery, we propose that agent-based models can provide a possible foundation for big data.

There are important relevant issues concerning this theme, and the social sciences in particular, that we do not address in this paper due to space limitations. These include some of the frontier work on causal inference in the context of machine learning (Athey and Imbens, 2016) and the ethical aspects associated with big data. For similar reasons, we do not discuss engineering-related concerns or provide a detailed review of any specific statistical method or machine learning tools that are used in the analysis given that ours is a more general argument about method, i.e., how explanations are made. The rest of this paper is organized as follows. Section 2 proposes a process-based definition of big data. Section 3 shows how agent-based modelling fits this operational definition of big data using two pioneering studies on swarm intelligence. In Section 4.2, we consider the notion of information aggregation in this context and its relation to addressing the quality of big data. We use prediction markets as an example, and show how the quality of big data can be evaluated through an agent-based model, viz., the agent-based prediction market.

2 What *is* Big Data?

In order to understand what is novel about the big data phenomenon, we need to have a clear idea about how it differs from our prior understanding of data. To do so, we break this query into three sub-questions, much in the Kantian mode.[5] We ask (a) what *is* big data, (b) what can we *know* from big data, and (c) *how* can we know from big data? Essentially, these are questions of the ontology, epistemology and methodology of big data, respectively. Concerning the ontological question, we need to identify characteristics that define big data and features, if any, that clearly distinguish big data from other forms of data; in other words, we need a clear definition of big data. At the risk of being flippant, in this article we refer to other forms of data as 'small data' for the sake of simplicity. The origins of the term Big Data have been traced to the period roughly around the turn of the twenty-first century (Diebold, 2012). Despite rapid developments over the years, a clear and comprehensive definition of big data still remains elusive.

Let us start with a prominent definition of big data in the literature in terms of the 3Vs: *volume*, *velocity* and *variety*. The *volume* aspect calls attention to the magnitude of the data, which

has been continuously increasing over time. *Velocity* refers to the frequency with which data is generated, the time taken to harness this data and the speed at which we can analyse data. The rapid growth of hand-held devices, internet users and increased computing power have resulted in the availability of real-time or near-real-time data which is ready for analysis. *Variety*, on the other hand, indicates the diversity in the data types that are available. In addition to the structured, tabular or rectangular data, big data is said to include non-traditional, semi-structured and unstructured data in the form of text, images, video, audio and different types of metadata.[6] Other attempts to distinguish the novelty of the big data phenomenon focus on the *technologies* and the compendium of fast developing *techniques* that enable the storage and processing of high magnitudes of data. The relative focus of the technology-based definition is not on *magnitude*, but on *methods* that facilitate the extraction of insights that can help in decision making and prediction.[7] In particular, its novelty is attributed to the use of statistical and machine learning techniques for data mining (or sometimes referred to as big data analytics). These include text analytics (for example, natural language processing and sentiment analysis), audio, video analytics, social media analytics and predictive analytics.[8]

However, neither the definition based on magnitude (3Vs) nor the one based on technologies provides a clear ontological boundary to distinguish big data from small data. For instance, the exponential growth of data has been a continually occurring phenomenon through the years since systematic data collection started. What is considered as big today is almost surely on its way to becoming relatively small in the years to come. Similarly, high-frequency data has been a familiar term among scholars studying financial markets at least since the mid-1980s (Taylor, 1987; Bansal et al., 1995). Similarly, variety also fails to account for the novelty of big data given that non-conventional sources of data such as text and videos, as we will see in section 3, have long been employed for analysing various phenomena much before the term big data became popular. Attempts to distinguish characteristics like volume, velocity and variety in such definitions invariably suffer from *context specificity*: they are relative to a certain time and the state of technology, and therefore lack the ontological clarity that we seek.[9] What about the technological and tools-oriented view of novelty associated with big data? They too suffer from the above-mentioned shortcomings. More specifically, data mining tools ranging from statistical learning, regression analysis, tools inspired from machine learning, natural language processing, computational intelligence tools such as fuzzy logic, neural networks, genetic algorithms, instance-based learning, reinforcement learning and so on had been developed long

before the mid-1990s or even before.[10] It is also worth recalling that the idea and techniques associated with data mining had already been developed during the days of data warehousing (Kantardzic, 2003). Therefore, the techniques and tools utilized by big data analytics alone fail to explain what constitutes big data and the associated novelty.[11]

While we do not intend to add a new definition of big data in this paper, we adopt a relatively neglected definition in the literature. The definition which we follow is motivated by the director of the United Nations Global Pulse project, Robert Kirkpatrick (2014):

> Global Pulse is interested in trends that reveal something about human well-being, which can be revealed from data produced by people as they go about their daily lives (sometimes known as "data exhaust"). Broadly speaking, we have been exploring two types of data in the Pulse Labs. The first is data that reflects "*what people say*," which includes publicly available content from the open web, such as tweets, blog posts, news stories, and so forth. The second is data that reflects "*what people do*," which can include information routinely generated for business intelligence and to optimize sales in a private sector company. An example of "what people do" data is anonymised mobile phone traffic information, which can reveal everything from footfall in a shopping district during rush hour to how a population migrates after a natural catastrophe. (Ibid, p. 4)

Our concern in this section is to define and understand the nature of big data and its usage in the realm of the social sciences. Given this focus, we define big data as follows: big data is, in a given spatio-temporal domain, the continuous archive of whatever *people said*, *did* and even *thought*. In other words, it is a microscopic and dynamic view of human activities. Though this definition may seem narrow, it provides a clear ontology for big data.[12] What follows from this definition is the realization that big data is nothing new; so long as there are human activities in a society, there is big data, whether or not we record, store and analyse it. However, it can be argued that, in the past, one did not have adequate technology to archive a substantial proportion of this data. With the remarkable advancements in recent years in information, communication and digital technology (ICDT), data has become 'big' as our communication activities concerning leisure and commerce have moved to the Internet. These in turn have moved into our phones, cars and even our glasses. These activities and a substantial portion (by no means complete) of social life can now be recorded and quantified in a way that would have been hard to imagine just a decade ago. Given this, it is understandable that some

definitions place a great emphasis on the technological aspects, which makes the realization of big data possible. The shortcomings of these definitions lie in the fact that they blur the line between the intended phenomenon that they strive to capture (or target) and the tools which allow us to reach this target or capture the phenomenon.[13] To make this point clear, in Section 3, we qualify our definition by bolstering this claim with some pioneering studies on swarm intelligence long before the current state of ICDT.

Following this definition, what actually makes big data novel for us is not its ontological aspect, but rather its epistemological aspect. Alternatively, the novelty is not its existence but our awareness of its existence. This awareness of big data certainly broadens our perception of what we traditionally understand as data along the lines of Euclidean geometry (rectangular data), to the possible non-Euclidean manifestations (images and voices). A number of prominent definitions of big data attempt to try to capture the novelty of this generalization. Novelties certainly imply several non-trivial extensions to our understanding and concepts to embody them.[14] Big data in this framework can be simply seen as a continuous, structured or unstructured, unprocessed dynamics of primitives rather than as points (snapshots) or summaries (aggregates) of an underlying phenomenon. Note that our definition of big data eschews physical size constraints and the volume criterion that appears in the 3V or 5V characterization is not required. It is more general since it circumvents any temporal and context specificities that plague other definitions. However, if we instead consider how much people said, did and thought, these are a function of their actions and interactions with the number of objects in their environment. In this case, our definition of "big" data is naturally big if a large group of people are being considered, vis-à-vis a smaller group. Formally, if the number of objects in a specific spatio-temporal domain is N, then we may expect that the size of "big" data will grow in $O(N)$.[15] Since in most contexts to which big data is applied come with an N that is large enough, our definition automatically implies a non-trivial size.

In a nutshell, by emptying out technological and physical essentials in various definitions, we hope to have characterized big data through a concise definition. This, however, does not mean that the technologies which help access big data are unimportant or that they should be taken for granted. It is quite the contrary. Neither does it imply that we should assume away the psychological and cognitive effects associated with being exposed to big data as a decision maker. We simply distinguish them as being causes for or consequences of big data, and not big data per se. In the following section (Section 3), we give two illustrations, both of which fit

the big data research paradigm in the view of our definition. They address methods of big data research, how they produce explanations and outline how we can engage with big data. It is important to note that these studies were conducted in the 1980s, i.e., two decades before the advent of the big data era. This further bolsters the claim that our definition is not bound by technological and physical (size) concerns.

3 Two studies on swarms

In this section, we look at two pioneering studies which investigated collective behaviour exhibited by different species. First, we look at the contributions by Brian Partridge and his colleagues (Partridge, 1981) in the field of marine biology, who studied schooling behaviour among different fish species. Outlining their novel methodological approach, we argue that the data collected from their experiments is in fact big data, along the lines in which it is defined in Section 2. Second, we examine another strand of research initiated by Craig Reynolds in his Boid project (Reynolds, 1987) to understand collective motion among birds, animals and fish through computer simulations.[16] Based on these studies, we outline and argue how agent-based modelling can be seen as a possible foundation for big data.

3.1 *Schools of fish*

Collective behaviour exists in various animal societies, for instance, swarming and flocking among birds, synchronized flashing among fireflies, schooling and shoaling among fish, and herding behaviour among various animals, to cite a few. The possible motivations and the mechanics behind such behaviour have fascinated researchers for long and still continue to do so (Partridge, 1981; Bonabeau, Dorigo, and Theraulaz, 1999; Ballerini et al., 2008a,b; Katz et al., 2011; DeLellis et al., 2014). Brian Partridge made one of the pioneering contributions on how groups of animals can sometimes move, act, and make decisions as though they were a single super unit. He was studying the Atlantic Pollock (*Pollachius virens*, also known as *Saithe*), a species of fish that exhibit schooling behaviour, with the aim of shedding light on the structure and function of schools (Partridge, 1981, 1982). For this, it was critical to know *how* synchronization and co-ordination is achieved among different individuals in a school. To develop a comprehensive understanding of how the school stayed together and moved as a unit, Partridge and his team came up with a new technique, which later turned out to have

a profound impact in the study of swarming behaviour among various animals. In contrast to the earlier attempts, they realized the importance of following *every* individual member of the school, continuously observing them, and more crucially, '*[t]he importance of looking at the interactions between members of schools instead of merely sampling school structure at discrete intervals*' (Partridge, 1981, p. 314).

Subsequently, they ventured to record the three-dimensional positions and identify interactions among each and *every* individual fish in a freely swimming school (of around 20-30 members) which was continuously observed over a period of time. The fish were allowed to swim freely in a large doughnut-shaped tank at the University of Aberdeen in Scotland and their movements were filmed. While the fish were swimming, the experimenter lay on a rotating gantry above the tank and followed the movements of the school, providing commentary on the relative positions of the individuals.[17] After filming, Partridge and his team painstakingly measured the relative positions of the individual fish in more than 12,000 frames of film. They carefully studied various correlation measures between different fish, their relative positions, velocities and so on. Through a detailed analysis of this dynamic stream of data, his team discovered key empirical rules that permitted the school to move as a unit. These rules were that the Pollock: (a) "match changes in both swimming direction and speed of their neighbours but correlations are greater for swimming speed' and (b) 'simultaneously match the headings and swimming speeds of at least their first two nearest neighbours within the school. (Ibid, p. 313)".

Their mode of analysis may seem rudimentary by modern standards. However, it is worth noting that these behavioural rules discovered among Atlantic Pollocks, expressed in slight variations, are believed to underlie many complex group movements, from schools of fish to flocks of birds, swarms of insects, and crowds of humans. Although their method of data collection is quite labour-intensive, the innovation in terms of the method introduced by Partridge has been followed even in the recent past (Katz et al., 2011). It has also been adopted to study flocking behaviour among birds (Hayes, 2011; Ballerini et al., 2008a,b), for instance, using stereoscopic photography. Similarly, long distance navigation patterns among homing pigeons are obtained in a similar manner by attaching lightweight GPS devices to collect high-resolution spatio-temporal data (Nagy et al., 2010).

It is fairly easy to see the connection between the nature of the data generated in these studies and what we refer to as 'big data' in the social sciences. The decentralized, unstructured and continuous nature of the data stream which is generated is, in fact, the hallmark of what we have

come to understand as big data. It is squarely within the scope of the process-based definition that was presented in the previous section. In the context of human societies (as opposed to fish in the case of Partridge), big data can be thought of as an archive of whatever people did, said or even thought. As we have seen, Partridge's data allows us to get as close as possible to see how agents (fish) interact and how collective dynamic patterns emerge at the macroscopic level. Even though the rules they found are simple, the data analysis involved in selecting among competing behavioural assumptions may be non-trivial and extremely time-consuming.

A data driven approach to unearth behavioural rules from high volumes of data has been argued as the essence of the big data approach [18] There have been remarkable improvements in the methods that are used to learn statistical patterns from the data, which has advanced the correlation methods that Partridge had employed. Machine learning algorithms such as the Isometric mapping algorithm (ISOMAP) have been employed to study various aspects of collective behaviour (DeLellis et al., 2014) from video data. However, these statistical and machine learning algorithms do not in themselves delineate the ontological aspect of big data, but rather only help facilitate the analysis as tools often do. We argue that although Partridge carried out his work almost 35 years ago, the data that he obtained can be regarded as big data from an ontological standpoint. Attempts to discover empirical patterns from continuous observations of unprocessed dynamics of primitives, however, can surely be increased in sophistication (often referred to as big data analytics) using machine learning techniques and tools from computational intelligence. However, the approach as a whole remains conceptually and qualitatively similar to the one employed by Partridge.

3.2 *Flocks of birds*

Craig Reynolds made another important innovation in terms of the method used to understand swarming behaviour. Partridge's seminal work and other studies in that tradition uncovered important behavioural rules underpinning schooling behaviour by careful experimentation, collecting big data and by employing data mining. By contrast, Reynolds resorted to simulations for understanding the aggregate motion among birds (Reynolds, 1987). He programmed artificial birds - *bird-oids* or *boids*, which were then simulated and *each* boid could be tracked. Reynolds assumed that flocks emerged as a result of *interaction* among the individual boids. He employed a distributed behavioural model, comprising the following behavioural rules for the boids (Reynolds, 1987, p. 28): (i) collision avoidance, (ii) velocity matching with the neigh-

bouring boids (in the same direction), and (iii) flock centering or cohesion. Based on these simple rules, Reynolds was able to demonstrate that flocking behaviour resulted even when boids were moving randomly to begin with. In other words, the local interaction among boids which followed the above mentioned behavioural rules was shown to spontaneously generate flocks.

Although Reynolds simulated flocks of birds (boids), the same model can be readily applied to simulating schools fish and other forms of collective behaviour. This marked an important breakthrough in thinking about bottom-up, local, interaction-based computational models for studying collective behaviour and the eventual structure of the flocks and the schools started to be viewed as an 'emergent' behaviour. His animation approach laid the foundation for almost all the subsequent advances in understanding complex, collective behaviour through simulations. Reynold's Boids project can even be considered as a primer on agent-based modelling since the design of agents and their behavioural rules constitute the starting point of the latter. In fact, Reynolds (1987) has been regarded as one of the precursors of what has come to be known as computational social sciences (Gilbert, 2010).

Each boid can be tracked and data on the behaviour of each boid (what it "did") is continuously *generated* throughout the simulation. In other words, there is agent-level data which provides a microscopic view of the simulated system. It is worth further examining these two methods to deal with big data put forward by these two pioneering studies on swarms. In the case of Partridge, big data was obtained through observing a phenomenon and the behavioural rules were derived using statistical methods. By contrast, Reynolds' study, which is a precursor to the modern day agent-based models, generated big data based on *given* behavioural rules that were assigned to individual boids. This is shown in Figure 1.

Agent-based models can thus generate big data as defined in section 2, and can take into account local interactions. Although behavioural patterns were inferred through statistical methods in the study by Partridge, it should be noted that there was no guarantee that these exact behavioural rules, when followed by the agents in a system, would lead to the observed collective behaviour at the macroscopic level. In addition, not all observed statistical patterns may be relevant, but only a selective subset of them need to be incorporated into an explanatory hypothesis. It is here that Reynolds' simulation approach employing agent-based models becomes important. It acts as a means to validate or corroborate the insights obtained via experiments, surveys or detailed observations. Similarly, alternative behavioural rules can be easily tested

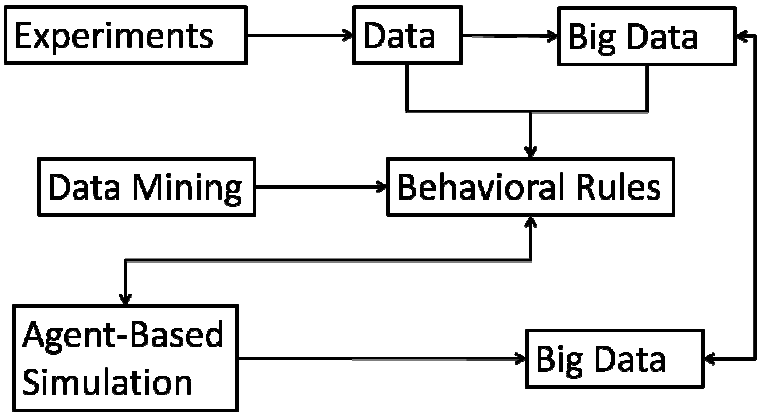

Figure 1: Agent-Based Simulation of Big Data

as potential explanations for the phenomenon. Thus they outline a scope for detailed computational experiments to evaluate different theories. The logic behind this combined activity of hypothesis generation (for explanatory purposes) and its validation (through simulations) can be seen as *abductive* reasoning.

3.3 Abduction and Agent-based models

Abduction, or retroduction, is a form of inferential reasoning in the philosophy of science that is distinct from deduction and induction. Abductive reasoning was championed by Charles Sanders Pierce, who in turn was influenced by Kant. Deductive reasoning involves inference about a particular object based on the knowledge of characteristics of the population to which that object belongs, in other words, a *top-down* logic. Inductive reasoning, on the other hand, proceeds to infer characteristics of a population based on the characteristics of a sample, a *bottom-up* logic. Peirce distinguishes abduction from these two modes of reasoning:

> Abduction is the process of forming an explanatory hypothesis. It is only logical operation which introduces any new idea; for induction does nothing but determine a value, and deduction merely evolves the necessary consequences of a pure hypothesis.
>
> Deduction proves that something *must* be; Induction shows that something *actually is* operative; Abduction merely suggests that something *may be*. (Peirce, 1903, 5.172)

Although induction and abduction are both ampliative forms of reasoning (unlike deduction), they cannot be reduced to each other and they also differ in their targets or aims. While the target of induction is inference about a future course of events (beyond the observations at hand),

abduction aims to infer *explanations* of observed events (Schurz, 2008). Thus, explanatory concerns are at the heart of abduction. In recent times, abduction has been largely identified with the notion of *Inference to the Best Explanation* (IBE). However, there is an important distinction between the Peircean view of abduction and the views of those like Peter Lipton who identify abduction as IBE: while the Peircean focus is on the process of generating an explanatory hypothesis, IBE involves both hypothesis generation *and* its evaluation (Campos, 2011).

Abductive reasoning has also garnered the attention of economic methodologists and other social scientists in the past (Lawson, 2003, 2006; Mabsout, 2015; Simon, 1977).[19] Agent-based modelling can be understood as a form of abductive inference if we take a bird's-eye view of the underlying process[20]: one starts with a surprising observed fact or phenomena that is to be explained; explanatory hypotheses about behavioural rules of individual agents and their rules of interaction are abductively selected among competing explanations; this hypothesis is then evaluated through multi-agent simulations to see whether it may be *sufficient* to explain the phenomenon in question.

A few points are worth noting here: (i) Even if behavioural rules may be based purely on statistical facts, eventual hypothesis selection is not trivial. The precise form that a hypothesis takes involves selecting a subset or a specific combination among these statistical facts (usually with some error-prone additions to them) that a modeller considers relevant to constitute an explanation. (ii) Agent-based modelling is more akin to abduction understood as "inference to the best explanation" where hypothesis selection and its evaluation are both combined. However, the notion of *best* explanation ought to be seen as the 'lovelist' explanation and not strictly as the 'likeliest' explanation (Campos, 2011). (iii) Discovery through agent-based simulations showcase how researchers often go back and forth using different tools and techniques (method) to produce explanatory insights (epistemology). This can be clearly seen in the example discussed above to explain how schools and flocks emerge. (iv) The implicit process of arriving at a plausible (not necessarily true) explanation to the *explanandum* through trial and error simulations involves computationally experimenting with plausible hypotheses via heuristic search. The choice among several such potential explanations is in turn often disciplined by the criteria of parsimony. This process of discovery as problem solving itself can be seen as abductive reasoning.

Returning to Partridge's approach, it should be noted that it has an obvious counterpart in the field of behavioural finance. It is similar to the approach employed in the early findings concerning fundamentalists and chartists in the financial markets, which was based on the analysis of the data from different kinds of surveys, such as questionnaires and telephone interviews, with financial specialists, bankers, currency traders, and dealers, etc. (Frankel and Froot, 1990; Allen and Taylor, 1990). These behavioural rules are obtained as statistical abstractions, which since then have been utilized and extensively experimented with in the literature on agent-based financial markets.

In conclusion, given their ability to generate and *speak to big data*, agent-based models can serve as a foundation for big data analysis. Since agent-based simulation can generate big data, it is interesting to see whether and how the big data thus generated can be meaningfully related back to the big data obtained from laboratory, natural and field experiments. The two approaches are in fact complementary: experimental and statistical methods can offer us a window into underlying behavioural rules, and the use of agent-based models can regenerate the same phenomena at various levels and help identify the sufficient (and possibly necessary) behavioural rules.

4 Quality, not just quantity

Big data is often characterized in terms of its quantity (volume). However, the feature of critical importance is its *quality*. In this section, we examine the issue of identifying the *quality* of information embodied in big data, in other words, the reliability of knowledge about social phenomena that is inductively inferred through statistical aggregation. In a variety of situations ranging from healthcare and education to migration and elections, data is pooled together and often aggregated or *fused* from diverse sources (Einav and Levin, 2014). One of the advantages of big data and machine learning tools allegedly stems from its efficiency in terms of information aggregation. In this context, several questions arise: How efficient is information aggregation resulting from big data methods compared to other aggregation mechanisms such as markets? Are there limits to such information aggregation? Does bigger data necessarily imply better information or knowledge? Do vast amounts of micro-level information pooled from diverse sources necessarily guarantee a superior collective opinion, i.e., the 'wisdom of crowds'?

The issue of information aggregation is discussed in detail in the literature on the *Hayek hypoth-*

esis (Smith, 1982). In the spirit of Hayek (1945), we ask whether there exist mechanisms that can help pool and aggregate knowledge that is dispersed among the individuals in a society, better than markets do. Big data tools that employ developments in computational intelligence can be viewed as a potential candidate for such a mechanism. This essentially boils down to addressing whether developments in computational intelligence can overcome the problems that might plague a central planner in theory. Such a possibility is intimately related to the central issue of the *socialist calculation debate*. Massive amounts of data from social media platforms and blogs are now easily available for analysis. This, in conjunction with increased computational power, developments in natural language processing, sentiment analysis and computational intelligence, is perceived as a promising avenue to effectively aggregate information, for instance, in the case of elections.

Cass Sunstein addresses this issue in response to the claim by Richard Posner that blogs had the 'potential to reveal dispersed bits of information' and rightly cautions about any misplaced celebrations (Sunstein, 2008). Opinions shared in social media, blogs, and twitter are often subjective and the tacit dimension of knowledge still persists, these being both examples of limitations which are harder to overcome. More importantly, there is no equivalent of a coordination device here like prices in the markets as in Hayek's example. The lack of an 'effective coordinating device would mean that there is no guarantee that the incentives and actions of the players can be appropriately aligned in a dynamic context, and these are thus prone to both noise and the possibility of manipulation. Chen and Venkatachalam (2016) argue that despite notable improvements in computational intelligence, there are inherent limitations to price discovery, and more broadly to information aggregation.[21]

Using the example of prediction markets and on-line reviews, we further illustrate the potential issues concerning the quality of big data. We also show how agent-based models can help answer questions regarding the quality of big data, wherein quality is interpreted as the efficiency of information aggregation. Based on the underlying characteristics of an agent-based model, such as the number of agents, learning or meta-learning behaviour, network structure, personal traits, preferences, and cultures (social norms), it is possible to capture the information aggregation process. It is also possible to examine the extent to which micro-level properties are inherited by 'big data' at the aggregate level.

4.1 *On-line reviews*

The accumulation of massive volumes of data can be seen as an inevitable consequence of the developments in communications technology, the proliferation of hand-held devices and the world wide web. This phase of evolution of the world wide web that enables users to generate data and for others to use this data is popularly referred to as Web 2.0. We will refer to the economy that operates in the background of these developments as the Web 2.0 economy. In order to understand the behaviour of agents in the Web 2.0 economy, it is necessary to pay attention to the structure of their social networks, nature of their local interactions and learning within those networks. Similarly, psychological factors influencing an agent's attention given the vast amounts of information are important. Traditionally, equation-based models have proved to be inadequate in dealing with such challenges given the dynamic nature of learning and the presence of recursive feedback effects, but more importantly, due to their weakness in modelling local interactions in social networks. We argue that agent-based models can provide the necessary theoretical tools to understand such user-generated big data.

Let us consider some familiar, everyday situations like choosing a restaurant or buying earphones. We often turn to user-initiated, user-supplied data that is aggregated in the form of on-line customer reviews which reveal information concerning the quality of the product in question. We review a contribution made by Nick Vriend (Vriend, 2002) to demonstrate the possibility of assessing the quality of information aggregation through agent-based models. Though customer review reports can help consumers acquire more information regarding the quality of a product, an increasing accumulation of reviews can result in information overload for those analysing them. To understand how consumers utilize information and the consequences of their behaviour which feed back into the system, a model that uncovers the underlying data generation process is necessary. Agent-based modelling can cater to this need and Vriend (2002) is one of the earliest examples in this direction. In fact, Vriend (2002) can be viewed as a contribution to understanding the Web 2.0 economy and as an agent-based study of big data.

While exploring to make an informed choice, a digital-savvy consumer in the Web 2.0 economy is often bombarded with a large pile of diverse on-line reviews. These are often uploaded through established platforms by users who may have experienced the product. When this size of such information is overwhelmingly large, as it often is, consumers may need to resort to a set of behavioural rules. These rules can guide them in the process of search, the decision to stop

searching, processing the collected information, and finally making a decision. By focussing on the decision mechanisms, Vriend studied a diverse set of rules such as: following the others, trying to replicate or to avoid the experiences of others and also random behaviour. While these rules are fast and frugal (Gigerenzer, 2007), they do involve different intensities of information processing. In investigating the information aggregation efficiency and the behavioural rules employed by heterogeneous agents in a population, Vriend asks: (a) would this decentralized information processing be able to distinguish products with superior quality from those that are inferior? and (b) regardless of the efficiency achieved, how well did these rules generally perform?

Vriend considers a society in which a finite number of consumers are repeatedly offered a binary choice. Each time the choices given are different (novel) so that there are no past experiences upon which individual learning can be based. In other words, each consumer is 'forced' to learn from others, i.e., the experiences of the previous users. In each period, of these two choices, there is always one that is relatively superior to the other. To analyse whether and how path dependence may occur in the presence of such social learning, Vriend also set up a series of check points.[22] At these checkpoints, the two choices are essentially identical. Therefore, it is expected that the society as a whole should show no excessive preference for either product. However, it turns out that such unbiasedness is merely one possible outcome. His simulations show that there are other outcomes, for instance, one product gets kicked out or almost kicked out by the other. This provides evidence of a possible path-dependence and lock-in effects. Hence, dynamic learning processes can result in self-organization towards a specific set of rules and that will not be unbiased when the society returns to the checking periods.

However, does a society distinguish and adopt a superior-quality product (service) during regular (i.e., non-checking) periods?[23] In Vriend's simulation model, the classifier system performs well by constantly increasing the correct discrimination frequency from 50% to 80%, which can be attributed to the wisdom of the crowds. Under this system, 80% of consumers make a right pick. But the improvements plateau out at 80% and the remaining 20% of consumers suffer from inferior decisions in each period. This result is based on the long-window moving average and hence it does not give finer details concerning the frequency of wrong decisions in each individual period. While the majority may not make mistakes in the long term, they do make mistakes in many single periods.[24] This is a good example where increasing the volume of data does not automatically guarantee an increased aggregation efficiency. The quality of these choices in the

aggregate can therefore be assessed only by having a window into the data generation process and a platform to (computationally) experiment with diverse behavioural rules. Agent-based models can thus be used to understand the behaviour of a decision maker situated in the era of user-generated Big Data. This enables us to assess the quality of information aggregation under various behavioural rules. In these cases, agent-based models can be seen as employing a form of *abductive* reasoning, aiming to generate new hypotheses that provide explanations in ways distinct from deduction and induction (Chen and Kao, 2016, see pp.130-134).) It is also fairly straightforward to see parallels of this approach with the studies on swarms described earlier, concerning inference in the presence of dynamic behaviour and experimenting with counterfactuals.

4.2 *Prediction markets*

We now return to the skeptical observations made by Sunstein (2008) about the ability of blogs, more generally, non-market platforms, to effectively aggregate dispersed information in the absence of effective co-ordination devices like prices. What if we could construct mechanisms in which such coordination devices are built in? Prediction markets are one such example, where deliberations among groups are disciplined by the incentive structure (Sunstein, 2006, ch. 4) to reveal the true information or expectation that they hold. In these markets, agents trade contracts whose payoffs depend on the future outcomes that are yet unknown. Private information is reflected in terms of the trading prices. When these uncertain events in question are a function of expectations or private information held by diverse agents, as in the case of elections for instance, mechanisms relying on predictions made by groups (rather than on individuals with partial knowledge) can prove to be effective.[25] They are believed to increase the accuracy of forecasts of various social phenomena compared to more traditional forecasting methods (Arrow et al., 2008; Wolfers and Zitzewitz, 2004).

For these prediction markets to be superior, their relative information aggregation efficiency ought to be greater than that of the traditional channels. How can we determine this? Predictions arising from different aggregation platforms can be contrasted with actual outcomes to assess the relative information aggregation efficiency of the former. For example, in the case of elections, outcomes based on both market (prediction markets, political futures markets) and non-market (polls and social media) sources can be compared with actual outcomes of the elections (Berg et al., 2008; Huberty, 2015; Erikson and Wlezien, 2008; Rothschild, 2015).

The underlying assumption behind the promise of prediction markets is linked to the efficient market hypothesis.[26] However, as in real financial markets, there is evidence to suspect the 'efficiency' of the prediction market in practice as well. As noted by Wolfers and Zitzewitz (2004, pp.117-119), agents do behave in ways that violate rationality assumptions in these markets. We also observe that prediction markets are not always accurate. More recently, they failed to predict the outcome of U.S. elections by a wide margin. In such situations, we need models that can help us understand the reasons behind the shortcomings of prediction markets and the quality of the data that they generate by pooling in information from many participants. The underlying factors associated with the success or failure of information aggregation need to be understood. In other words, *why* and *how* they succeed or fail, not just that they do. Note that these markets are capable of generating not just high-frequency data, but big data according to our earlier definition in section 2.

Let us consider an agent-based model of an information platform, specifically, an *agent-based prediction market* (Othman, 2008; Bothos, Apostolou, and Mentzas, 2010; Jumadinova and Dasgupta, 2010; Klingert and Meyer, 2010; Othman and Sandholm, 2010; Yu and Chen, 2011; Chie and Chen, 2015a,b). These models show a potential way to assess the quality of predictions emanating from actual prediction markets. They are populated by virtual agents who interact in the market place. Their behaviour and topology of network interactions are based on a number of economic and social variables inspired from the literature. For instance, studies have considered utility functions (Wolfers and Zitzewitz, 2006), beliefs, learning behaviour, wealth distribution (Manski, 2006) and also social networks (Cowgill, Wolfers and Zitzewitz, 2009) in their model construction.[27] Such a set-up provides a versatile environment for experimenting with various behavioural assumptions concerning learning and how they lead to speculative bubbles and ultimately examine the quality of actual predictions. The dynamics of such agent-based prediction models can be matched to the dynamics of the actual prediction markets. Although the current models are relatively simple, they do show potential to help us understand the success and failures of these predictions markets. Some early models interpret prediction markets in terms of the *distribution of beliefs of traders* and, in doing so, they rest upon the equilibrium price or the Walrasian equilibrium price (Manski, 2006; Wolfers and Zitzewitz, 2006).[28] This interpretation offers a *static* view.[29] However, agent-based models can potentially promise more: they can relate the dynamics of spread, volatility or depth observed in the prediction market to the plausible economic and social characteristics of the markets.

4.2.1 *Big data and financial markets*

Another research area in which large amounts of data are available is the study of financial markets. The ability of financial markets to aggregate information, scarcity and expectations among different individuals has been analysed in great depth. In the mainstream literature on finance, markets are perceived to be efficient in aggregating dispersed information concerning different assets. Furthermore, the asset prices are believed to provide a complete reflection of all necessary and relevant information. The trading prices are seen as reflecting the fundamental value associated with an asset. This view has been challenged in the past, both on theoretical and empirical grounds. Many puzzles concerning financial markets still remain open and the connection between rationality, efficiency and predictability in the broader context has been hotly debated to the present time. The structural analogy between the example of on-line reviews and financial markets and the possibilities for a breakdown in efficiency is relatively straightforward. In so far as social behaviour is perceived as a result of interaction among heterogeneous individuals, agent-based models provide a powerful tool to understand the dynamic processes underpinning such phenomena. Agent-based computational finance is a relatively mature field compared to other areas in which agent-based modelling has been employed. Agent-based financial models view financial markets as being comprised of interacting heterogeneous groups of learning and boundedly-rational agents.[30]

Challenges to the efficient market hypothesis in the past have been mounted both from behavioural finance and agent-based financial markets.[31] Studies have argued that there can be persistent deviations from the so-called fundamental values in the presence of news. These can be seen as external shocks to expectations held by agents. Under certain assumptions concerning learning behaviour (imitation and herding, for instance) and psychological biases, it has been shown that there can be price changes resulting in bubbles, despite any changes to the fundamentals (Shiller, 1981; De Grauwe and Grimaldi, 2006). Agent-based modelling of financial markets has also shown that price movements can occur even in the absence of news (Hommes, 2006; LeBaron, 2006), provided that there is constant interaction among agents. The research on financial market efficiency can be viewed as an attempt to understand the causes of potential changes in investor sentiment in the market, and how these changes in turn affect different variables like price, volatility, stock-market returns, and trading volume. What is the significance of this mode of behaviour in the age of big data?

Financial markets generate high-frequency, dynamic data and this can, in principle, be obtained at the level of an individual and hence naturally is a candidate for big data. News announcements, learning and interaction among agents are both shown to be capable of having an impact on prices in the financial markets. Investor sentiment can be shaped as a result of interaction among agents in social networks, based on sharing actual texts. With the advent of social media, researchers now have a window to understand how the sharing of news and discussion among agents on social media platforms can possibly shape investor sentiment and its evolution. In particular, perception of news and discussion among agents can be seen as processes to develop a context, which in turn impacts our expectations. Consequently, there is an impact on the price of assets. For instance, the growth of social medium networks and, consequently, the increased interaction among agents, may enhance or deteriorate market efficiency, for example, due to herding. Recent empirical studies have attempted to measure investor sentiment from textual data on social media (known as *sentiment analysis*) and analyzed their relation to stock market variables and their predictability (Kim and Kim, 2014; Siganos, Vagenas-Nanos and Verwijmeren, 2014; Bukovina, 2016). However, in addition to empirical evidence, models describing *how* the investor sentiments emerge from such textual data are needed. Although this area of research is still in its infancy, agent-based models can be extended to incorporate textual data through modifications to agent-design to equip them with text-analysing capabilities. Thus, agent-based models have the potential to incorporate evolving modes of social interactions and non-standard forms of data.

5 Concluding remarks

Existing definitions in the literature view big data in terms of its size or based on technological concerns. In this article, we departed from this and adopted a process view that distinguishes its ontology from other forms of data. The importance of this distinction lies in the fact that data is no longer just about points, snapshots, or sampling segments of a social phenomenon; it is about the entire dynamic process of individuals in a specific spatio-temporal domain. We argued that although engineering, and technological concerns are important in harnessing the potential of big data, they do not capture its novelty. Similarly, tools used for data mining and statistical analysis also do not sufficiently distinguish big data from other forms of data. Based on this view, we argued that agent-based modelling provides a possible foundation for big data by being able to act as a data-generation process for the observed data, and as an abduction

tool for examining or evaluating the features which we observed from big data, such as the quality of big data, as we show in the agent-based prediction markets (Section 4.2).

If we acknowledge this process-based definition of big data, the need for a joint research program between agent-based modelling and big data is quite evident. Although agent-based models often generate big data, researchers in this community do not tend to store, retrieve and reuse the data, partly because of the associated demands on memory, but also due to the relative lack of awareness concerning the use of machine learning to handle it. The data which can potentially keep track of the movements of each and every individual agent in the model, on what the agent did, "said", or "thought", are often removed before the next iteration, leaving behind summary statistics such as prices and volumes alone. This is rather unfortunate. One may argue that individual-level data may not be important in all contexts, which may seem reasonable. If individual-level data or more generally phenomena are not important, what is the *raison d'être* of agent-based models in such a context? However, research in the area of agent-based modelling has shown that changes at the individual level can have a non-trivial impact at the aggregate level. Even though alternatives to agent-based models that are implicitly underpinned by individual interaction are available - such as the mean-field and other statistical mechanics models, or random particle systems - they are attractive only if our concern is restricted exclusively to the mesoscopic or macroscopic part of the society in situations where aggregation irons out the effects of local interaction and heterogeneity.

The real issue, however, is not about when individuals gain importance and when they do not. Instead it is the general lack of the awareness that agent-based models do generate big data and using tools from data mining to exploit the data generated by these models. Earlier, Chen and Yu (2010) suggested that agent-based modelling can serve as a knowledge discovery and data mining tool for formulating economic theory. They use an agent-based model of double auction markets, modified from the original Santa Fe Double Auction Tournament, as an illustration to show how a theory of optimal procrastination as a strategic behaviour of auctioning can be discovered by markets purely composed of artificial agents. They note:

> It may not surprise us to see that these autonomous agents eventually beat all these programmed agents, but it is difficult to perform an *in-depth analysis* of the discovered bargaining strategies given these complex surroundings. Maybe a challenging task for the future would be to introduce novel data mining or text mining techniques to this large database so as to know more of the "mental process" of these

autonomous agents. (Ibid, p. 170; Italics added)

As we have tried to argue, there is a potential synergy between big data and agent-based models. The latter can act as a foundation to understand the knowledge embodied by big data and the analytic tools developed to harness large data sets. This, in turn, can enhance the analysis of big data from the agent-based models themselves. In our opinion, it is only a matter of time before it becomes necessary to exploit this synergy.

Acknowledgements

A primitive version of this paper was presented at the *Duke Forest Conference on the Era of Natural Computationalism and Big Data*, held in Durham, North Carolina, November 11-13, 2016. We thank John Davis and Wade Hands for their encouragement and support. We thank Paul Wang, Leigh Tesfatsion, John Duffy, Vikas Kumar, Jagannath Iyer and Sunil Mitra Kumar for providing helpful comments on the paper at different stages. We also thank the anonymous referees for valuable suggestions to improve this paper. Support for this research in the form of Ministry of Science and Technology (MOST) Grants, MOST 106-240-2410-H-004-006-MY2 and MOST 103-2410-H-004-009-MY3, is gratefully acknowledged.

Notes

[1]Some of these works have also included humanities (digital humanities and computational humanities) under the larger umbrella of CSS. Cioffi-Revilla (2013) introduces the work of the pioneers in what came to be known as CSS and points to the work by Charles Osgood (1916-1991) on the semantic differential model of human mind and the work by David Heise on affective control and the project *Magellan*. Their pioneering work laid the foundations for corpus linguistics and facilitates the task of automatic information extraction.

[2]Among many recent publications, Cioffi-Revilla (2013) is a notable exception, which attempts to give a more balanced treatment on both subjects, agent-based modelling and big data, while only a chapter (his Chapter 3) is given on the latter, and is mainly from the viewpoint of data mining and automatic information extraction.

[3]For a detailed survey on agent-based models, see Tesfatsion and Judd (2006); Gilbert (2008).

[4]For a discussion on the different varieties in which agents are modelled in this tradition, see Chen (2012).

[5]In his monumental work *Critique of Pure Reason*, Immanuel Kant breaks down his inquiry into human reason in the form of three questions.

> "All interest of my reason (the speculative as well as the practical) is united in the following three questions: 1. What can I know? 2. What should I do? 3. What may I hope?" [A805/B833]

[6]Some argue for including more Vs in broadening this characterization such as veracity, vincularity, value, etc. For a more comprehensive survey and discussion see Kitchin and McArdle (2016).

[7]For surveys of the definitions of big data, the reader is referred to Ward and Barker (2013); Gandomi and Haider (2015); Kitchin and McArdle (2016).

[8]See Gandomi and Haider (2015, Sec.3) for a detailed description of each of these categories.

[9]Note that the 3V characteristics used to describe big data can also be applied to the description of different forms of conventional data since each V listed here is just a matter of degree. Kitchin and McArdle (2016), who undertake a study of 26 datasets from different areas, note that there is notable diversity among them and that not all of them fulfil all the generally perceived criteria. They conclude that velocity and exhaustivity are more important.

[10]For instance, neural networks have been around at least since the seminal work by McCulloch and Pitts (1943) and instance-based learning methods such as K-nearest neighbours (KNN) can be traced back to the 1960s. See Chen, Kao and Venkatachalam (2017, pp. 301-305).

[11]Perhaps one could argue that the combination of these two elements could possibly provide the necessary ontology. However, it would be unsatisfactory since the problems of context specificity still remain.

[12]A similar view is put forward by (Kitchin, 2014, p.2): "Big Data is characterized by being generated continuously, seeking to be exhaustive and fine-grained in scope, and flexible and scalable in its production. Examples of the production of such data include: digital CCTV; the recording of retail purchases; digital devices that record and communicate the history of their own use (e.g., mobile phones)... clickstream data that record navigation

through a website or app; measurements from sensors embedded into objects or environments; the scanning of machine-readable objects such as travel passes or barcodes; and social media postings. These are producing massive, dynamic flows of diverse, fine-grained, relational data."

[13]An analogy would be the existence of bacteria or viruses independent of microscopes, gold being independent of alchemies, or a moon being independent of spaceships. In a similar vein, big data should be conceptualised and defined as independent of ICDT.

[14]Using big data, one may hope (even if unsupported by evidence concerning possibility) to understand what people need, think, intend to do, and, in the extreme case, hope for a *complete* or a comprehensive understanding of the dynamics of streets, markets, families or even society. Regardless of our scepticism about such a possibility, we do not need to augment the above definition of big data with such possible emotional or affective consequences.

[15]$O(N)$ is read as "big order of N", for example, a multiple of N. However, it can also grow even faster, depending on how interactions happen and depending on the physical or mental constraints of the agents, such as attention (Simon, 1971). Therefore, in general, the size of "big" data can grow in $O(f(N))$, where f is determined by various constraints imposed on interactions.

[16]At the time of carrying out the Boid project, Reynolds did not know of Partridge's work.

[17]A more detailed description of this arrangement can be found in Partridge (1981, 1982).

[18]There is often an implicit assumption that there is no need or role for a priori theory. However, Partridge cannot be judged guilty of such a naive methodology since his choice of observations and analysis are very much informed by the existing theories.

[19]In particular, Simon (1977) in which his disagreement with Popper on whether there is a logic behind scientific discovery is relevant. Simon argues that hypothesis generation, and not testing, is the most valuable scientific activity and in his view scientific discoveries are problem solving situations. This act of hypothesis generation (or law discovery) is abductive according to Simon.

[20](Peirce, 1903, 5.189) describes the form of abductive inferences as follows: "The surprising fact, C, is observed; But if A were true, C would be a matter of course; Hence, there is a reason to suspect that A is true." Halas (2011) provides a reformulation of this definition in the context of agent-based modelling as follows: "An unexplored emergent phenomenon of some complex social system is observed and agent-based model of corresponding complex system is then constructed. If multi-agent simulations lead to growing of the emergent phenomena, then there is a reason to suspect that assumptions of the model are correct."

[21]See Ruths and Pfeffer (2014) for the limitations of using social media data to understand behaviour.

[22]Specifically, along his 25,000-period simulation, Vriend used multiples of 500 periods as the check points.

[23]This is a simplified case in which one product is *objectively* superior to the other. This is different from determining the quality of the restaurant, which may involve *subjective* factors and a proportion of agents with specific tastes.

[24]This can be viewed as the simultaneous existence of the *wisdom of crowds* and the *stupidity of herds*.

[25]Although prediction markets can operate for sporting events like football, the outcomes of those events are strictly not a function of the expectations held by the agents alone. Election outcomes are an example of the latter.

[26](Wolfers and Zitzewitz, 2004, p.108): 'Much of the enthusiasm for prediction markets derives from the efficient markets hypothesis. In a truly efficient prediction market, the market price will be the best predictor of the event, and no combination of available polls or other information can be used to improve on the market-generated forecasts.'

[27]In the context of prediction markets, Cowgill, Wolfers and Zitzewitz (2009) illustrated that the embedded social network topology can cause a wave of optimism and hence a potential bias from the information aggregation. See also Chie and Chen (2015b).

[28]Wolfers and Zitzewitz (2006) provides one of the most significant results in the prediction market literature: prediction market prices approximate the central tendency of the distribution of beliefs of traders, the so-called *the market's beliefs*. However, discovering the central tendency does not directly imply forecasting accuracy. On the one hand, it depends on how well-informed the traders are (the proportion of informed traders or marginal traders). The issue becomes harder when the probability of the occurrence of the event is not exogenous, but endogenously determined by beliefs held by the agents themselves. So, even though the market does aggregate opinion most of the time, it may fail to foretell what will happen in future.

[29]However, some recent progress has been made toward the dynamic interpretation of prediction markets (Frongillo, Della Penna and Reid, 2012).

[30]A comprehensive survey of this approach, even if it is dated, can be found in LeBaron (2006).

[31]Behavioural finance attributes limits to arbitrage and cognitive biases among agents as reasons to doubt the efficiency of the market. Agent-based financial models have largely stuck to more conventional representations of agents, although this gap has narrowed in the past few years.

References

Allen H, Taylor M (1990), Charts, noise and fundamentals in the London foreign exchange market. Economic Journal 100: 49-59

Alvarez R (Ed.). (2016). Computational Social Science: Prediction and Discovery. Cambridge University Press.

Arrow et al. (2008), The Promise of Prediction Markets, Science, 320, 5878, pp. 877-878

Athey S, Imbens G (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353-7360.

Ballerini M, Cabibbo N, Candelier R, Cavagna A, Cisbani E, Giardina I, Orlandi A, Parisi G, Procaccini A, Viale M and Zdravkovic V (2008). Empirical investigation of starling flocks: A benchmark study in collective animal behaviour. Animal Behaviour, 76(1), 201-215.

Ballerini M, Cabibbo N, Candelier R, Cavagna A, Cisbani E, Giardina I, Lecomte V, Orlandi A, Parisi G, Procaccini A and Viale M (2008). Interaction ruling animal collective behaviour depends on topological rather than metric distance: Evidence from a field study. Proceedings of the National Academy of Sciences, 105(4), 1232-1237.

Bansal R, Gallant, AR, Hussey R, Tauchen G (1995). Nonparametric estimation of structural models for high-frequency currency market data. Journal of Econometrics, 66(1), 251-287.

Berg J, Forsythe R, Nelson F,& Rietz T (2008). Results from a dozen years of election futures markets research. Handbook of Experimental Economics Results, 1, 742-751.

Bonabeau E, Dorigo M, Theraulaz G. (1999). Swarm intelligence: from natural to artificial systems (No. 1). Oxford University Press.

Bothos E, Apostolou D, Mentzas G (2010) Agent based information aggregation markets. In: van der Hoek W, Kaminka G, Lesperance Y, Luck M, Sen S (eds.), Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'2010), May 10-14, 2010, Toronto, Canada, 449-454.

Bukovina J (2016). Social media big data and capital markets An overview. Journal of Behavioral and Experimental Finance, 11, 18-26.

Campos, D. G. (2011). On the distinction between Peirce's abduction and Lipton's inference to the best explanation. Synthese, 180(3), 419-442

Chen SH (2012). Varieties of agents in agent-based computational economics: A historical and an interdisciplinary perspective. Journal of Economic Dynamics and Control, 36(1), 1-25.

Chen SH, Yu T (2010) Agents learned, but do we? Knowledge discovery using the agent-based double auction markets. Frontiers of Electrical and Electronic Engineering (FEE) in China, 6(1): 159-170.

Chen SH, Kao YF, Venkatachalam R (2017). Computational Behavioural Economics. In: Roger Frantz; Shu-Heng Chen; Kurt Dopfer; Floris Heukelom and Shabnam Mousavi, eds. Routledge Handbook of Behavioral Economics. Routledge. pp.297-319.

Chen SH, Kao YF (2016). Herbert Simon and agent-based computational economics. In Minds, Models and Milieux. Palgrave Macmillan UK. pp. 113-144

Chen SH, Venkatachalam R. (2016). Information aggregation and computational intelligence. Evolutionary and Institutional Economics Review, 1-22.

Chie B-T, Chen S-H (2015a) Spatial modelling of agent-based prediction markets: Role of individuals. In: Grimaldo F, Norling E (eds.), Multi-Agent-Based Simulation XV, Springer, pp. 197-212.

Chie B-T, Chen S-H (2015b) The use of knowledge in prediction markets: How much of them need he know? Journal of Information Science and Engineering 31(1):1-22.

Cioffi-Revilla C (2013) Introduction to Computational Social Science: Principles and Applications. Springer Science & Business Media.

Cowgill B, Wolfers J, Zitzewitz E (2009) Using prediction markets to track information flows: Evidence from Google. In: Das S, Ostrovsky M, Pennock D, Szymanski B (eds), Auctions, Market Mechanisms and Their Applications, p. 3. Springer.

De Grauwe P, Grimaldi M (2006) The exchange rate in a behavioural finance framework. Princeton University Press, Princeton.

DeLellis P, Polverino G, Ustuner G, Abaid N, Macr S, Bollt EM, Porfiri M (2014). Collective behaviour across animal species. Scientific Reports, 4, 3723.

Diebold FX (2012). On the Origin (s) and Development of the Term 'Big Data', Penn Institute of Economic Research Working Paper 12-037

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

Epstein J (2007) Generative Social Science: Studies in Agent-Based Computational Modeling. Princeton University Press.

Epstein J, Axtell R (1996) Growing Artificial Societies, MIT.

Erikson RS, Wlezien C (2008). Are political markets really superior to polls as election predictors? Public Opinion Quarterly, 72(2), 190-215.

Frankel J, Froot K (1990) Chartists, fundamentalists, and trading in the foreign exchange market. American Economic Review 80:181-186

Frongillo R, Della Penna N, Reid M (2012) Interpreting prediction markets: A stochastic approach. In: Pereira F, Burges C, Bottou L, Weinberger K (eds.), Advances in Neural Information Processing Systems 25, pp. 3266-3274.

Gandomi A, Haider M (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.

Gigerenzer G (2007) Gut Feelings: The Intelligence of the Unconsciousness. Penguin Books.

Gilbert N (2008). Agent-based models (Quantitative Applications in the Social Sciences). SAGE Publications, Inc., Thousand Oaks.

Gilbert N (2010) (ed.) Computational Social Science. Volumes 1-4. SAGE Publications, Inc., Thousand Oaks.

Halas M (2011). Abductive reasoning as the logic of agent-based modelling. In: Proceedings of the 25th European Conference on Modelling and Simulation, eds. T. Burczynski, J. Kolodziej, A. Byrski, and M. Carvalho . European Council for Modelling and Simulation.

Hayek F (1945) The uses of knowledge in society. American Economic Review 35(4):519–530.

Hayes B (2011). Flights of fancy. American Scientist, 99(1), 10-14.

Hommes C (2006) Heterogeneous agent models in economics and finance. In: Tesfatsion L. Kenneth J (eds.) Handbook of Computational Economics, 2-23:1109-1186, Elsevier.

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S (2008). Big data: The future of biocuration. Nature, 455(7209), pp. 47-50.

Huberty M (2015) Can we vote with our tweet? On the perennial difficulty of election forecasting with social media, International Journal of Forecasting, 31, 3, 2015, pp. 992-1007.

Jumadinova J, Dasgupta P (2010) Stochastic game-based multi-agent prediction markets. Department of Computer Science, University of Nebraska at Omaha, Technical Report No. cst-2010-1.

Kantardzic M (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons.

Katz Y, Tunstrøm K, Ioannou C, Huepe C, Couzin, I.D (2011). Inferring the structure and

dynamics of interactions in schooling fish. Proceedings of the National Academy of Sciences, 108(46), 18720-18725.

Kim S, Kim D (2014) Investor sentiment from internet message postings and the predictability of stock returns. Journal of Economic Behavior and Organization 107:708-729.

Kirkpatrick R (2014) A conversation with Robert Kirkpatrick, Director of United Nations Global Pulse. SAIS Review of International Affairs 34(1): 3-8.

Kitchin R (2014) Big Data, new epistemologies and paradigm shifts, Big Data & Society, 1(1), 1-12.

Kitchin R, McArdle G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1), 1-10

Klingert F, Meyer M (2010) Multi-agent-simulation of prediction markets: Does manipulation matter using zero-intelligence traders? 3rd World Congress on Social Simulation (WCSS 2010), September 6-9, 2010, Kassel University, Kassel, Germany.

Lawson T (2003). Reorienting economics. Milton Park, Abingdon, Oxon: Routledge.

Lawson T (2006). Economics and reality. Milton Park, Abingdon, Oxon: Routledge.

Lazer D, Pentland A, Adamic L, Aral S, Barabasi A, Brewer D, Christakis N, Contractor N, Fowleret J, Gutmann M, Jebar T, King G, Macy M, Roy D, Van Alstynea M (2009) Life in the network: the coming age of computational social science. Science 323 (5915): 721-723.

LeBaron B (2006) Agent-based computational finance. In: Tesfatsion L. Judd K (eds.) Handbook of Computational Economics: Agent-based Computational Economics, 2-24: 1187-1233, Elsevier.

Mabsout, R. (2015). Abduction and economics: the contributions of Charles Peirce and Herbert Simon. Journal of Economic Methodology, 22(4), 491-516.

Manski C (2006) Interpreting the predictions of prediction markets. Economics Letters 91(3):425-429.

McCulloch WS, Pitts W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115-133.

Nagy M, kos Z, Biro D, & Vicsek, T. (2010). Hierarchical group dynamics in pigeon flocks. Nature, 464(7290), pp. 890-893.

Othman A (2008) Zero-intelligence agents in prediction markets. In: Padgham L, Parkes D, Muller J, Parsons S (eds.), Proceedings of 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), May, 12-16., 2008, Estoril, Portugal, pp. 879-886.

Othman A, Sandholm T (2010) When do markets with simple agents fail? In: Van der Hoek W, Kaminka G, Lesperance Y, Luck M, Sen S (eds.) Proceedings of 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010), May, 10-14, 2010, Toronto, pp. 865-872.

Partridge B (1981) Internal dynamics and the interrelations of fish in schools. Journal of Comparative Physiology 144(3): 313-325.

Partridge B (1982). The structure and function of fish schools. Scientific American, 246(6), 114-123.

Peirce C (1903 [1931–1958]). Collected papers of Charles Sanders Peirce (Vols. I–VI). (C. Hartshorne & P. Weiss Eds.). Cambridge: Harvard University Press.

Reynolds C (1987) Flocks, herds, and schools: A distributed behavioural model. Computer Graphics 21(4): 25-34.

Rothschild D (2015). Combining forecasts for elections: Accurate, relevant, and timely. International Journal of Forecasting, 31(3), 952-964.

Ruths D, Pfeffer J. (2014). Social media for large studies of behaviour. Science, 346(6213), 1063-1064.

Schelling T (1971) Dynamic models of segregation. Journal of Mathematical Sociology 1:143-186.

Schurz G (2008) Patterns of abduction. Synthese, 164(2), 201-234

Siganos A, Vagenas-Nanos E, Verwijmeren P. (2014). Facebook's daily sentiment and international stock markets. Journal of Economic Behavior & Organization, 107, 730-743.

Shiller R (1981) Do stock prices move too much to be justified by subsequent changes in dividends? American Economic Review 71: 421-436.

Simon H (1971) Designing organizations for an information-rich world. In Greenberger M (ed.), Computers, Communications and the Public Interest. Johns Hopkins University Press, pp. 37-72.

Simon H (1977). Models of discovery. Pallas Paperbacks Series. Dordrecht: Reidel.

Smith V (1982) Markets as economizers of information: Experimental examination of the "Hayek Hypothesis". Economic Inquiry 20(2): 165-179.

Sunstein CR (2006). Infotopia: How many minds produce knowledge. Oxford University Press.

Sunstein CR (2008). Neither Hayek nor Habermas. Public Choice, 134(1-2), 87-95.

Taylor MP. (1987). Covered interest parity: a high-frequency, high-quality data study. Economica, 429-438.

Tesfatsion L, & Judd KL (Eds.). (2006). Handbook of computational economics: Agent-based computational economics (Vol. 2). Elsevier.

Varian HR (2014). Big data: New tricks for econometrics. The Journal of Economic Perspectives, 28(2), 3-27.

Vriend N (2002) Was Hayek an ACE? Southern Economic Journal 68(4): 811-840.

Ward J, Barker A (2013) Undefined by data: A survey of big data definitions. arXiv preprint arXiv:1309.5821.

Wolfers J, Zitzewitz E (2004). Prediction markets. The Journal of Economic Perspectives, 18(2), 107-126.

Wolfers J, Zitzewitz E (2006) Interpreting prediction market prices as probabilities. National Bureau of Economic Research, No. w12200.

Yu T, Chen SH (2011) Agent-based model of the political election prediction market. In: Sabater J, Sichman J, Villatoror D (eds.), Proceedings on Twelfth International Workshop on Multi-Agent-Based Simulation, Taipei, Taiwan, May 2, 2011, pp. 117-128.