Mind the gap: from data to big data in the economic domain

Research Master Thesis by Virginia Ghiara

Student Number: 43784820

Supervisor: Prof. Dr. M. Boumans

Advisor: Dr. H.C.K. Heilmann

Third reader: Prof. Dr. J.Vromen

Word Count: 21291

Erasmus Institute for Philosophy and Economics,

Erasmus University Rotterdam

Content

Acknowledgements

One year ago, I came to the Erasmus Institute for Philosophy and Economics (EIPE) because I was looking for a work environment in which to discuss topics at the intersection of philosophy and economics. Needless to say, my desire has been abundantly satisfied, and I can only say that I have a great intellectual debt to the EIPE community. In particular I owe many thanks to Marcel Boumans, my supervisor, and Conrad Heilmann, my advisor. At the early stages of developing my thesis, Marcel Boumans helped me develop the ideas presented in this thesis. He has always supported my research with optimism, and his enthusiasm has given me confidence and made me believe that, despite the limited time, I would have been able to complete this work. At the same time, Conrad Heilmann helped me to restrict the topic of this thesis into the appropriate scope. Without their help, I would not have been able to write this thesis. Moreover, Conrad Heilmann has played a central role in the choice on my academic future: without his advice, I would not be at the University of Kent as a Ph.D. student. I would like to thank him for this.

My experience at EIPE has been great, and for that I have to thank everybody at EIPE, but in particular my fellow students Emanuele Di Francesco, Lukas Beck, Mario Castellano, Nicolás Berneman and Zoe Evrard. Moreover, I would like to thank Constanze Binder and Attilia Ruzzene. Constanze Binder has been an excellent tutor, always willing to listen to me and to advise me. Furthermore, I have to thank her for having closely followed the drafting of my Ph.D. proposal; her help has been invaluable. Finally, Attilia Ruzzene is the living proof that philosophers are not isolated persons locked in their room: her humanity has made last year even more enjoyable.

In conclusion, I want to thank my family, Antonio and Ana for their support. They are always willing to help me at any time, and for that I am truly grateful. In particular, I would like to thank Ana for all her advice, and Antonio for every chat that accompanied the writing of this thesis and for every moment he stood by me.

Introduction

Nowadays, the economic world is largely composed of elements called "data". Economists work every day with data on the GDP of countries, on the level of poverty of specific areas, on the prices of items and so on. Data[1] look like the raw material that allows economists to do their job. Not only, over the last decade new expressions have been added to the vocabulary of economists and, more generally, of scientists. "Data deluge" and "big data" have become pervasive expressions used to account for an emergent phenomenon: our world is experiencing a data explosion. As a consequence, more data, as well as new forms of data, can be used within the scientific fields. However, the only possible way to *really understand* what big data are and what changes they have caused, is to first investigate what the nature of traditional data is.

Of course, the nature of data can be studied at a very general level (for instance it may be questioned what all the data have in common, regardless of the discipline) as well as at a more specific level (for instance it can be studied what are the main important features of data in a particular discipline). This thesis entails the second approach of data study: it deals with the role of data and big data in economics, and in particular in macroeconomics. There are three main reasons that motivate this choice.

First, questions concerning what economic data are, how big data have "sneaked in" the economic field and what the consequence of such "data deluge" are, require a combination of philosophical and economic knowledge in order to be answered. As a student of a Research Master in Philosophy and Economics, I hence find this topic particularly interesting. Economics is involved since we are examining economic data, philosophy cannot be forgotten given that several debates about the nature of data can be found in the philosophy of science.

Second, the emphasis on the potential of big data is growing day after day. However, the analysis of the true novelty contained in such data is still superficial, especially within the economic field. This means that there is still much work to do. In the chapters of this thesis, I shall raise what I consider relevant questions with the goal to extrapolate new significant claims. My purpose is to offer new insights to the debates on this topic.

Third, among the economic disciplines, macroeconomics has caught my attention. The reason is that big data are often claimed to cast a new light on some macroeconomic issues such as unemployment and poverty. Furthermore, macroeconomic data appears pretty different if

[1] I use the term "data" as a plural term given that its usage has evolved from the Latin plural of "datum".

compared to other forms of data, like those generally employed in microeconomics or in other sciences such as biology. As I shall underline, macroeconomics often deals with facts about phenomena and, as a consequence, macroeconomic data are generally related to these abstract concepts.

Given these considerations, the vital question that will guide this work is the following: what is the difference between data and big data within the economic domain? The answer to this question will be articulated throughout three chapters.

Chapter 1 provides an account of what traditional economic data are and how they became so important for economic studies. By examining the past, I underline that the emphasis on data is relatively new since for a long time data have not been included in economic research. Moreover, two categories of data are distinguished: one is made up of data about elements of the world, the other is constituted of data on facts about phenomena. Finally, given such distinction between data and phenomena, I describe Bogen and Woodward's account (1988) and the philosophical debate emerged after the publication of their 1988 paper. I will then stress why the data/phenomena distinction should be preferred to the statistical distinction between sample statistics and population models proposed by Glymour (2000).

Chapter 2 proposes a survey of the several descriptions of big data until now proposed. By considering such descriptions, I verify whether those features already suggested are actually new characteristics detectable in big data. My examination follows partially Kitchin and McArdle's research (2016), based on the analysis of 26 data bases containing big data. Nonetheless, at the end I reject their conclusion, and I articulate a new approach to big data's innovation. According to it, the most important novelty brought by big data is the automated production by means of which big data are generated. To end with, my examination goes back to macroeconomics. With the aid of a case study (the macroprudential data cube), the way in which big data affect macroeconomic data are sketched out.

Chapter 3 is focused on the way in which big data are studied within the macroeconomic domain. After the description of the new approaches based on visual analytics, I analyse the process required by visual analytics to shed light on the fact that, despite the technological development, the passage from big data to facts still requires theory-laden human decisions. In fact, both initial decisions concerning the "transformation" of data and further decisions about the kind of visualization to perform cannot but depend on previous assumptions, as in the past. As a consequence, data used by scientists are actually "data models" at different levels (Harris 2002; Boumans 2004; Morgan 2004).

Chapter 1

Economic data

1.1 Data and observations in economics

The claim that economics deals with data is nowadays commonly accepted as an obvious statement, nevertheless a strong relation between data and economics had not been considered obvious at all for a long time. If we look back at the history of economics, the great importance given to data was one of the novelties of the twentieth century. The increasing emphasis on measurement and quantification was responsible for the intensification of the use of mathematics as well as for the development of statistics. Statistical tools started to be considered indispensable to study economic phenomena and, as a consequence, the scientific approach witnessed a methodological change in which the collection of data became a central activity. Moreover, after the diffusion of econometric techniques in the 1930s, the weight attributed to data, especially in numeric formats, dramatically increased.

Data were engaged in generating models, and models became the lens through which economists observed the world. Considering the economists' work before the nineteenth century, it involved activities of observation that did not deal with the collection of data as it is commonly understood nowadays. For instance, references to "evidence" in the reports of the British parliamentary were never references to numerical data contained in tables, but rather to reports of eyewitnesses (Morgan 2012). Observations were reported in many different forms such as sermons, letters, pamphlets, narratives or book-length tracts, but statistical reports and graphical representations were unlikely to be produced. The employment of documents like those mentioned in the development of economic studies involved (and cast light on) the idea that the task of economists was not a mere registration of given facts followed by subsequent analysis. In that period, it was easy to find the terms "observation" and "reflection" as interchangeable words in the economic tracts' titles (Morgan 2012): observing and thinking were considered strictly connected, and the claim that economic documents were almost never neutral collections of mere observations, was hardly denied by anyone.

Let us consider again the methodological change experienced by economics in the twentieth century. After the rise of statistics, standard questionnaires, taxonomies and tables became the basic materials for economic studies and they, in turn, demanded new material to be observed for economists' analyses. Economists began to work on tables and collected data to develop research, while less attention was spent on other sources of information. Many economists described

economics as an observational science (Morgan 2012), but in the last centuries the most prominent kind of observation has hardly dealt with what is considered the "real world". On the one hand, the economic world has increasingly become a world of data, statistics[2] and phenomena; on the other hand, economists have started to consider themselves as experts capable of inferring knowledge from such elements rather than from direct eyewitnesses or from personal experiences. It is still questionable, however, what economic data are, and whether all of them have the same features. In the next section, I provide an answer to this question by distinguishing two groups of economic data.

1.2 Categories of economic data

Generally, when economic data are described, they are separated into several groups by using different methods (Griliches 1986). First, they can be distinguished given the level of aggregation: indeed, data can be about individuals, families, firms, districts, industries, sectors or countries. Second, data can be divided according to the content: data might be on prices, quantities, population statistics, individual behaviours and so on. Further distinctions could be then based on the periodicity or on the level of fabrication.

Though, economists usually do not distinguish between data on visible elements and "data"[3] about abstract entities. For instance, macroeconomics facts are just considered as "data" with a high level of aggregation, from districts to states, industries and countries (Griliches 1986). Despite these distinctions between economic data should of course be taken into account, the divisions into such groups is not sufficient to cope with all the differences existing between those "data" engaged in economics. Especially for macroeconomic "data", another aspect plays a very significant role.

By examining data from the point of view of their production, I propose to differentiate two main groups of economic data. In the group composed of economic data, it is in fact possible to find data such as numbers corresponding to the amount of money spent by an institution or by a singular person, numbers related to specific indicators such as GDP, numbers denoting the age of a private bank's clients, numbers about the rate of unemployment of a country and so forth. Even if all of them are involved in economic analyses, economic data can be divided into two categories

[2] With the term "statistics" I design the outcomes of the measurements performed using statistical tools. Such elements are neither "raw data" nor facts about phenomena, and appears to be indispensable in many economic researchers aimed to discover facts about economic phenomena.

[3] From this moment I will use the term data between quotation marks to denote all the facts that are considered data by economists but that are actually far from being given.

with the aim to cast light on the ontological distinction between "visible things" and "facts about phenomena" and on the novelty contained in big data.

The first group is composed of data created through direct observations. In this case data can be for instance on observed prices and quantities in a market, or on the outcomes of specific transactions, and their production does not require a high level of conceptualization: they are visible elements of the world translated into new visible numbers, images, texts and videos. This kind of data comes closest to the original meaning of data, that is, "givens". Despite this group of data can be used by economists, it is a small group if compared to the complete amount of "data" engaged by economists.

This consideration leads to the second group of economic "data": that composed of facts. The units belonging to such group are not "given" in nature; on the contrary they require a high level of conceptualization in order to be produced. In other words, they are "data" on the invisible subjects of study within the economic field. For instance, a "datum" on the national expenditure of a country belongs to the second group: there is not a physical object in the world which is considered the "national expenditure". This "datum" is the outcome obtained by aggregating data on individual expenditures. It is a "datum" on a fact (the national expenditure) measured once this fact has been conceptualized and rules of data aggregation have been formulated. Without this conceptualization and the employment of a measurement process, such a "thing" like the national expenditure would not exist.

It can be noticed that, while the main problem related to the production of the first category of economic data is the accuracy of the observations, understood as their truthfulness with respect to the reality, the second group of economic "data" may show a lower level of accuracy due to the technique engaged in producing them. Two main issues can be identified for these "data". First, since measures are aggregate, they may lead to a loss of information. Second, often such "data" are created by imputing different values to the aggregated parts. "Data" about the national expenditure, for instance, could be produced by measuring the total marketed transactions, but in practice while some outputs are directly not taken into account (like home production), others are included using the values attributed to them, which may be consistent with the theoretical interest (Pesaran and Smith 1992).

Surely, the level of conceptualization needed for the production of the "data" of the second group is always higher than those required for the generation of the data of the first group. Nevertheless, a certain degree of heterogeneity can be found between the second group's "data" when the attention is on the reasoning from which they have been generated. To clarify this point, let us

consider a “datum” about the number of hours worked during a particular week by a specific group of individual and a “datum” about the unemployment rate of a country in a precise year. They are both “data” on facts, and they could be distinguished because of diverse levels of aggregation. However, this is not enough to account for the differences between them: the dissimilar processes entailed in their productions cannot be entirely explained by appealing just to the level of aggregation. What is more important, the levels of conceptualization can be compared, and it may be stated that measuring the unemployment rate of a country appears much more complex than measuring the working hours of a group of persons. The former measurement process, indeed, does not depend only on an observable fact or on the aggregation of observable facts. Using nothing than human eyes it is just possible to observe a more or less large amount of persons doing leisure activities during working hours, but from this observation it is not possible neither to conclude that those persons do not have a job (they could have a day off, or irregular working hours, or they might be retired) nor to appeal to the idea of unemployment and to the method to measure it.

In addition, even using surveys, no one will be capable of answering what the rate of unemployment in a country is. Unemployment is not a physical object directly visible in the world, but an abstract concept linked to a specific theoretical framework that at first needs to be considered by researchers as a fact to be subsequently investigated. Thus, the measurement is possible only because it is rooted on the conceptualization of this fact: after having determined what unemployment is, a researcher may decide to make it measurable by means of some directly observable variables in the world. By using this measurement, then, researchers can create data on this unobservable fact and can make it visible in the real world.

1.3 “Saving the phenomena”: data to phenomena reasoning and its criticisms

As I mentioned above, macroeconomics deals with concepts such as unemployment, investment, national income, consumers’ expenditure and so on (Backhouse 1995). Almost every concept considered by macroeconomics is a concepts of an abstract entity translated into a “datum” resulting from a specific measurement. Such entities are often described by economists as “phenomena”. A long philosophical tradition supports the idea that to be a phenomenon, one thing needs to be perceived by the human senses. For example, the term “phenomenon” has often been used to translate in English the German word *Erscheinung* (“appearance”), used by Kant (1781) to denote the direct object of sensory intuition. Nonetheless, the contemporary philosophy of

science lost the connection between such term and the immediacy of experience, and nowadays the term "phenomenon" is generally used to denote things invisible to human eyes.

The philosophical debate on this topic took a significant turn in the 1980s, when Bogen and Woodward's paper "Saving the phenomena" (1988) stressed the idea that hardly ever phenomena are observable in any sense of the term, but despite their invisibility they are the only *explananda* of scientific theories. The data/phenomena account contrasted with the more traditional framework according to which scientific theories should be used to directly associate theoretical and observational claims. The structure advocated by Bogen and Woodward contained not two, but three components: theoretical claims, phenomena claims and data; and the connection between data and phenomena, according to them, is not prescribed by the theory explaining the same phenomena.

To introduce such distinction, Bogen and Woodward used the example of how scientists would determine the melting point of lead. In order to do it, scientists would take many thermometer readings of samples of lead just melted; these readings, according to Bogen and Woodward, constitute data. Moreover, such data can be used as evidence for a hypothesis about the melting point of lead, which is, in this example, the phenomenon. The authors noted that in this process of determining the melting point of lead, scientists actually do not observe the melting point of lead, but the data (the thermometer readings). Therefore, at least in this example, scientists infer an unobserved phenomenon from observed data.

So far, this is pretty straightforward and uncontentious. However, Bogen and Woodward moved a step forward. They argued that scientists tend not to develop theories with the aim of explaining the specific data points they have acquired. Rather, scientists construct theories that explain those phenomena that can be directly inferred from data. After this proposal, various criticisms of Bogen and Woodward's view have been offered (McAllister 1997; Glymour 2000; Harris 2002; Schindler 2007). Some of such criticisms have cast light on the issues regarding the specific passage from data to phenomena, while other criticisms have considered the distinction between data and phenomena.

To begin with, McAllister (1997) has cast light on the fact that every data set can be understood as being composed of stable patterns and a certain level of noise (the term "noise" denotes not a margin error or factual inaccuracy in data, but the mathematical divergence between a given pattern and a data set). According to McAllister, in a single data set infinitely many patterns can be discerned, but only some of these patterns are taken by scientists as corresponding to phenomena. However, such distinction between patterns corresponding and not corresponding to

phenomena is not based on intrinsic properties that some patterns lack: the essential, qualitative properties that should define a phenomenon are in fact not always clear[4]. Rather, McAllister has claimed that specific patterns are chosen according to the investigators' thinking or theoretical commitments, and among them an important commitment is the level of noise the investigator is willing to tolerate. However, for any noise level there is an infinity of patterns exhibited by the data, therefore also other theoretical commitments are involved in such decision. In other words, the passage from data to phenomena is theory-driven.

A similar claim has been proposed ten years later by Schindler (2007):

> "Bogen and Woodward 1988 [...] have argued [...] for a bottom-up construction of scientific phenomena from data. For them, the construction of phenomena is 'theory-free' and the exclusive matter of statistical inferences, controlling confounding factors and error sources, and the reduction of data." (Schindler 2007, p. 161)

What was rejected by these scholars, is that the passage from data to phenomena does not require any kind of assumptions. Nevertheless, Woodward (2011) has replied to such criticisms to clarify that his view, as well as Bogen's view, is not that the reasoning from data to phenomena does not imply any additional substantive empirical assumptions. Further assumptions are always required, and sometimes also the very theory that explains a phenomenon is used to "mediate" the reasoning from the data to that phenomenon. The only idea excluded in the 1988 papers is that such mediating role takes the form of theory providing an explanation of the data (Woodward 2011, p. 174). According to this clarification, on the one hand, it is completely acceptable that the selection of those patterns corresponding to phenomena deals with theoretical assumptions or commitments. On the other hand, Woodward has claimed, theories are always aimed to explain phenomena rather than data.

Another criticism has been proposed by Glymour (2000): according to his account, no distinction between data and phenomena is needed since the subject of scientific theories is the group of independence relations found in a population of data. His consideration started with the analysis of a method for discovering causal relations among variables (Spirtes et al. 1993). Scientists can

[4] Woodward (1989) tried specify some property that only those patterns that correspond to phenomena have: they should be (i) stable and invariant; (ii) characterized by some simplicity and generality and (iii) they should also show recurrent features. Nevertheless, McAllister considered such properties vague and he rejected the idea that phenomena are vital constituents of the world whose identity does not depend on stipulations by investigators. According to him, Bogen and Woodward failed in providing a way to distinguish patterns that correspond to phenomena from patterns not corresponding to phenomena (1997, p. 225).

construct a sample of data by measuring the joint distribution of values among these variables. Then, the sample is treated as a sample from a population of data with a particular statistical structure, given by a probability density function on joint values for the variables. It is this function that entails particular conditional independence relations among the variables.

In other words, scientists perform a double inference: at first, they calculate various samples of data and use such sample statistics to infer a model of the population of data; then, from the conditional independencies entailed by this population model, scientists infer a causal model that accounts for (or, in other words, that *explains*) the conditional independence relations in the population model. If scientific theories account for such population model, it is questionable why, instead of using the statistical distinction between sample and population structure, the distinction between data and phenomena should be preferred. According to Glymour, given the data/phenomena distinction, we have two options. We could accept that this deals with something more than the sample/population structure distinction. Then, the data/phenomena distinction would be unnecessary for such statistical inferences procedures, since the sample/population structure distinction would be enough. We could also recognize that the data/phenomena distinction corresponds to the sample/population structure distinction: in this case we would just give a new name to a well-known distinction. As a consequence

> "no such distinction between data and phenomenon is needed, and the distinction which is needed, [theory/data distinction] is already well established in the relevant literature" (Glymour, 2000: 32).

In conclusion, the account proposed by Bogen and Woodward has generated several discussions about the possibility to find patterns between data. According to someone (McAllister 1997), many patterns can be found among a data set, therefore the selection of those patterns related to phenomena cannot but involve theory-laden decisions performed by scientists. Glymour (2000) recognized that different approaches can lead to different theory-dependent inferences from the data to causal structure (for instance, subjective and objective Bayesian methods are two examples in which theoretical accounts and subjective considerations could be involved). Also Woodward (2011) did not deny the role that theoretical commitments could play in the passage from data to phenomena.

While great attention has been spent on this kind of inference, the reason why the data/phenomena distinction is important can still be questioned. The next section aspires to provide a new answer to such question that is compatible both with Bogen and Woodward's account and with Glymour's view.

1.4 The data/phenomena distinction within the economic domain

Data, Bogen and Woodward claimed (1988), are never explained by theories. Scientific theories only cope with phenomena, inferred from patterns among data. However, Glymour added (2000), such patterns and the related phenomena are mere correlations between variables existing in a data set. As a consequence, instead of using the terms "data" and "phenomena", it is enough to consider that scientists start with the correlations existing in a sample of data to infer and explain the structure at the population level.

Even if Bogen and Woodward did not reply to Glymour, they probably would accept his view. As Woodward claimed (2011), the aim of the paper written with Bogen was to replace the idea that scientific theories should link theoretical and observational claims with the idea that theories are involved only when scientists consider phenomena, not when they consider observational claims about data. As a consequence, if patterns among the sample of data are considered local while the patterns among the population data are taken to be more stable, it would be enough to say that theories only account for those patterns at the population levels, and that the samples are just needed to find the populations' structure.

Nevertheless, I would say that the data/phenomena distinction, at least within the economics field, offers great insights that would remain almost invisible with the statistic distinction between sample statistics and population models. In fact, a particular advantage of using the conceptual dichotomy between data and phenomena is that the attention can be focused not only on the passage from data to phenomena and on the *explananda* of scientific theories, but also on the starting point of economic research: the so-called economic data. Therefore, the claim I am going to support is that the notion of data, despite data are not the subject of theoretical explanations, is however an important concept which requires further analyses and which should not be replaced by the statistical concept of sample statistics.

First, I have already underlined that in economics two categories of data can be distinguished, and that the second category of economic "data" is actually composed of data about facts, that are the outcome of specific measurements of economic phenomena. If the starting point of the scientific inquiry was considered the group of correlations detected among a sample of data, less attention would be given to the difference between economic data on visible things of the world and economic "data" on facts about phenomena. This is particularly relevant since the data/phenomena distinction allows scientists to analyze a chain of reasoning that would not be underlined if the focus was on the sample/population statistical distinction. In fact, also such "data" on facts (that are the outcome of measurements of facts about phenomena) can be engaged

to find new facts about phenomena. For instance, data about GDP might be used to make international comparisons and find new patterns. In other words, the passage from data to phenomena could actually be almost an infinite path:

Data → Patterns → Facts about phenomena→ Measurement and new "data" → Patterns → Facts about phenomena → Measurement and new "data" …

If the scientific path involved only correlations between variables among the sample of data, correlations between variables in the total population and explanations of these correlations by means of causal relations; it would be difficult to cast light on this chain. Any sample of data (of the first or of the second category) would be considered just a sample, and from it the structure of the population data would be inferred and explained.

However, some "data" on facts are not produced only by examining specific correlations; rather they are the outcome of further conceptualizations which might involve, for instance, also considerations about the relevance of data patterns. This idea has already been proposed for what concerns a "datum" about the national expenditure: this datum could be produced by measuring the total marketed transactions, but in practice while some outputs are not considered (like home production), other transactions are included by attributing a specific value to them, which may be consistent with the theoretical interest (Pesaran and Smith 1992). As a consequence, the measurement procedure may change over time, and the data produced before and after such change should be considered incomparable given their differences. This concern is particularly relevant in order to avoid wrong conclusions within economics, and it is strictly related to the concept of data. For instance, if we consider again the passage from data on GDP to facts about the international economy, it would be vital to understand whether all the examined data have been produced by means of the same procedure. Indeed, as Coyle (2014) has recognized,

> A recent study found that in the data set frequently used by economists to make international comparisons, twenty-four out of forty-five countries had no price survey data at all. Some countries are using weights that have not been changed since 1968, and only ten sub-Saharan African countries use weights less than a decade old. In each case where old weights have been used for years, there will be large upward revisions in estimated real GDP when the weights are updated. (Coyle 2014, p. 29)

As a consequence, the distinction between data and phenomena, together with the further distinction between the categories of data that I have proposed, appears to cast light on the challenges that accompany the study of certain disciplines such as economics.

Second, it has been claimed that the passage from data to phenomena could be theory-laden, but also the creation of the same data used as a starting point to investigate a phenomenon might be affected, under certain aspects, by the idea of the same phenomenon that will be discovered. Woodward has expressed this idea by claiming that:

> "Investigators attempt to design experiments and arrange measurement apparatus in such a way that data produced by these reflect features of the phenomena they are trying to detect." (Woodward 2011, p. 166)

Let me elaborate more on this point: not only experiments are designed in order to generate data that reflect the features of phenomena, but also the creation of observational data might require a certain level of awareness of the features that phenomena should have. Bogen and Woodward (1988) claimed that the process of phenomena recognition is composed of the three steps:

Data → Patterns → Phenomena

Following their proposal, "data" on facts about phenomena would be consequently created after this process, to make visible such phenomena:

Data → Patterns → Facts about phenomena→ Phenomena

According to them, the only possible path goes from data to phenomena. However, as regard economic "data" on facts, often the starting point cannot be identified with the observation of data performed by researchers. Rather, the initial steps might be reversed. This may happen because, when researchers start to think of a phenomenon, those data that can offer evidence for its existence have not been produced yet. This leads to the conclusion that studies on economic phenomena appear to be data driven as much as hypotheses driven. Available data can of course enhance the possibility to study and examine economic phenomena. Perhaps in some cases they are also sufficient. Nevertheless, practical and theoretical questions on facts about phenomena are likely to affect and shape the ways in which those data from which phenomena are discovered are produced.

The conceptualization of an economic phenomenon, in fact, might not come from the recognition of a pattern in a data set, but could be guided by practical consideration about how to cope with some aspects of the reality as well as by researchers' theoretical backgrounds. Economic data do not come from nowhere; they have to be produced by someone, however it is not obvious that their creation always precedes the moment in which economists begin to think of a phenomenon for the first time. By observing the reality, economists could have a sort of "intuition" about the

presence of a phenomenon or might feel the need for its conceptualization, consequently they may decide to collect specific data to study it. Once data are created, researchers can use them to find the pattern corresponding to the phenomenon and they can start to elaborate a concept of such phenomenon. Finally, the conceptualization of the phenomenon can lead to a decision regarding how to make it observable (thus, how to measure it).

To grasp this idea, let us consider the history of economics, and in particular the emergence of the concept of unemployment. The first concepts of unemployment were vague ideas concerning the relation between people without a job and the level of poverty: when the concept was developed the true intention was not driven by a sincere interest in unemployment as an economic phenomenon, but rather the conceptualization was guided by the practical desire to fight poverty. Economists had the intuition that an important phenomenon was involved in the condition of poverty, but at the beginning it was just an idea no further conceptualized. In fact, the aim of fighting poverty led often local authorities to organize surveys, nevertheless they were unsystematic and a clear notion of unemployment was absent. The use of this term, hence, was already present in policy and public life before that a deep conceptualization was proposed by economists. Economists and politicians knew something important was at stake, but they did not have the adequate means to fully understand what the phenomenon was, and no data were eligible to cast light on it.

The first theoretical work in economics which explicitly dealt with this concept was Arthur Cecil Pigou's *Unemployment* (1913), moreover in the following years of the twentieth century many efforts were made to clarify this concept and to make it accessible for measurement. In addition, the efforts to elaborate a clear concept of unemployment were accompanied by the attempts to create the necessary data to study such phenomenon and to understand both its causes and its effects within the economic world. Once this concept was introduced, indeed, numerous questions were proposed about the phenomenon and its characteristics. For instance, it was absolutely unclear whether temporally idleness should have been considered part of the phenomenon of unemployment given that, for a part of the population working as day labourers or agricultural employees, having a period of time out of work was accepted as a part of their occupations (Rodenburg 2008). Similar problems concerned the distinction between those individuals without a job who were looking actively for work and those who were not, or between those who did not have any job and those who were working only part time but desired to work more (Kleeck 1931).

Nevertheless, since until that moment scarce attention had been spent on these issues, the surveys already conducted were generally not aimed to distinguish all these categories and the available

data were not capable of coping with them. For this reason, to study the phenomenon and to have the possibility to elaborate explanations for it, researchers had to organize more detailed surveys and to create new data. The elaboration of these surveys had thus a vital role: statisticians had to specify what they wanted to know by posing specific questions, for this reason over the years many different surveys have been proposed to produce the required data. Without having to describe all the differences between the several questions asked in the surveys, it can be underlined that the generation of those questions has always involved at least an initial idea of which concept of unemployment was the most correct.

What follows from this consideration is that the concept of unemployment was not born in the academic domain but in the sphere of public policy and it was first observed as one of the factors causing a relevant problem for public policy: poverty. Moreover, the impossibility to recognize it for the first time in a data set was also due to the fact that those data able to provide evidence for it were not available. For this reason, in order to elaborate an explanation, researchers had to produce data, only later they had the possibility to observe them to find the pattern corresponding to the phenomenon.

The described case is not isolated. Other economic phenomena were born when they were recognized as serious economic problems, as unemployment, or when researchers started to consider them and the facts about them as important elements to be taken into account in their reasoning. It is then important to consider the concepts of data and phenomena because, in a sense, while phenomena come from data, data might come from vague ideas of phenomena. The passage from data to phenomena could in fact be the following:

Observation of the reality → Vague idea of phenomena → Data production → Patterns → Facts about phenomena→ Phenomena

Someone may claim that the same analysis could be offered if the starting point was a sample of data: in that case, the vague idea of a phenomenon would lead to the selection of some variables and not of the other variables. The sample would in fact contain all those variables considered relevant according to the scholars. However, it would remain almost invisible that part of these selected data and variables were previously created according to this vague idea of the phenomenon. This consideration leads us to the third advantage related to the data/phenomena distinction: the clarification of the fact that almost all data are not “raw”.

The previous considerations have already underlined an important point: economists often call the data they use “raw”, but the adjective “raw” can be interpreted in several ways (Harris 2002).

To begin with, the expression "raw data" could denote those data that have not been processed or manipulated by scientists. However, let us consider the "classical" economic data again: data on the prices of specific goods, data on the GDP of a nation, data on the rate of unemployment of a precise area. They obviously are not "raw" since they are the outcomes of previous measurements and activities. As a consequence, it is hardly possible to have not manipulated data. Furthermore, data could be considered "raw" because it is believed that they are not influenced by theory. However, I have tried to underline that theoretical assumptions are likely to affect the way in which data are created. A third possibility is that some data are called "raw" because they are in a unique position between the real-world system and what someone considers the "first-level data model" (Lynch 1990; Harris 2002). All the data set, it is claimed, contain data models instead of "raw" data: data must go through several processing steps before being considered usable. However, since a situation could entail different kinds of levels of representation of the data, models could be generated at different levels. Therefore, those data between the "real world" and the first-level model might be called, in a misleading way, "raw".

Only this last interpretation of the expression "raw data" could be accepted, but overall it could be claimed that, within economics, almost all the so-called "raw data" are not raw. In other words, data are far from being "given" to economists, who on the contrary are generally involved in their production. This last point, as we will see, will help to clarify what the novelty bought by big data is.

1.5 A look ahead

What follows from this Chapter is that the relations between economic studies, economic data and economic phenomena have to be taken seriously in order to understand the economic field. At first, the way in which economists look at data has changed over time: it was only after the emergence of statistics and econometrics that data started to be considered indispensable elements for economic research. Moreover, the category of economic data is constituted by two sub-groups: the former consisting of data on (visible) physical elements of the world, the latter made of "data" on (invisible) facts about phenomena. While the first group is close to the original meaning of data (givens), the second group requires a high level of conceptualization to be produced.

In addition, economic "data" on facts about phenomena, generally employed in the macroeconomic domain, can exist only if the phenomena to which they are related are discovered and measured. A lucid exposition of the way in which phenomena can be recognized from data was proposed by Bogen and Woodward (1988): data are observed, patterns are found and from them phenomena are discovered and explained by means of theories. Several discussions emerged

after the 1988 paper: some of them were on the theory-laden passage from data to phenomena, another (Glymour 2000) questioned the relevance of the data/phenomena distinction. In the last section of this Chapter, I have supported the claim that at least one advantage can be found if economic research starts from this distinction rather than from the statistical distinction between sample statistics and population models: the attention can be focused on the so-called "economic data". By examining what such data are, some important considerations about the path of phenomena discovery and about the procedures of data production can emerge.

In conclusion, the emphasis on the data is particularly important if the aim is to discover whether something has changed in those considered "economic data" and in the way of approaching them after the "data deluge". Do economists still use data similar to those employed in the twentieth century? Are data still studied in the same way? Over the last years, the expression "big data" has become pervasive in many fields of study; what are "big data" and how have they changed the economic realm? The task of the next Chapters is to answer these questions.

Chapter 2

Big data in macroeconomics

2.1 Data deluge and the emergence of big data

Chapter 1 has claimed that the rise of statistics and econometric has led to consider data and models as indispensable elements to study economic phenomena and, consequently, economists have become sophisticated data users. Let me go a step further and introduce another claim. After the first recognition of the importance of data, over the last decade the relevance attributed to them has begun to increase again consistently. Needless to say, the interesting question is: why? In short, this rise in popularity of data could be ascribed to a sort of "data deluge" that has characterized the last years. The increasing capacity to trace (generally in digital forms) the activities performed by singular individuals, offices and firms, has provided in fact the opportunity to obtain much more data than those produced through the use of surveys. Until thirty years ago, data on economic activity were scarce and difficult to be created (Einav and Levin 2014). In a relatively short period of time the situation has improved significantly. The most important changes responsible for this "data deluge" seem to be the introduction and the diffusion of the Internet, where almost every action is nowadays recorded, and the creation of new tools and devices able to produce records of many elements of the world. I will now present three examples to clarify this point.

First, let us consider the data produced by retail stores. Before the invention and the diffusion of the Internet, stores created data on daily sales, but this production was time-consuming and some details (such as the product sold or the category of such product) were often omitted. Nowadays Internet retailers have a direct access to data containing information about the exact times at which specific items have been sold. Moreover, they can trace consumers' behaviours since every time a new search is introduced or a new item is viewed, a datum is automatically generated. No wonder that such information is considered a true goldmine by retailers (the major online retailer, Amazon, has taken advantage of it by developing algorithms to recommend popular products to specific costumers). Nonetheless, retailers are not the only ones interested in them: these data can also travel within or outside the economic domain, and they can be employed for different purposes. For instance, the Economic and Social Research Council (ESRC) has recently commissioned the launch of the Consumer Data Research Centre (CDRC), a collaboration between the University College London, the universities of Leeds, Oxford, Liverpool and

industry. The purpose is to provide a data infrastructure both for the collection and the diffusion of data coming from retailers, local government and other businesses located in the UK.

Second, the diffusion of the Internet has enhanced the opportunity to observe individual activities that before were very difficult to see or to trace. Over the last decade, the Internet has become time after time an important tool in human life, and it has been associated with several daily activities. Many of the Internet users, it would seem, use search engines as intermediaries for the online world: by typing key terms into a query box, they can obtain information available online. These activities are recorded into data, like those provided by Google Trends, which offer information on the volume of queries and on the queries that users enter into Google. Many social scientists have stressed the opportunity to use the Internet as a new source of data: several algorithms, in fact, continuously and automatically digitize and store Internet activities. The outcomes are, of course, data. In the economic realm, Ettredge et al. (2005) offered one of the first examples of how web search data could be used within the macroeconomic statistics' domain. In particular, they focused on data related to the phenomenon of unemployment. After them, a growing number of economists has emphasized the potential of these data: Askitas and Zimmermann (2009) underlined that keyword searches were strongly correlated with monthly German unemployment data, D'Amuri (2009) reached similar conclusion studying the phenomenon in Italy, Fondeur and Karame (2013) claimed the usefulness of Google data for the prediction of youth unemployment in France. For a long time, it was difficult to observe the activity of searching for a job; the only available method was often to ask by mean of surveys to people. The Internet seems to extend the ability to trace it, as well as many other activities.

Third, mobile phone usage, like the Internet usage, is part of the common life of a great amount of the world population. Over the last years, the claim that such mobile devices might enhance the possibility to study both social structures and human dynamics at extraordinarily large scales, has become popular among social scientists. Moreover, from this claim, other statements about the potential of the information contained in mobile call records have been supported by many scholars. In developing countries, for instance, where social and economic data production often requires more efforts, mobile call data have been used to analyse the relations between the communications networks and the socio-economic aspects of regional economies (Mao, Shu et al. 2015). Another use of mobile phone has been proposed by Blumenstock (2013): by using call detail records, he has gathered deep insights into particular nuanced forms of migration, like seasonal migration. According to him, these data may play a vital role in testing the already developed hypothesis according to which migrant workers act like arbitrageurs whose

displacements among labor markets are fundamental for markets' equilibrium (Harris and Todaro 1970).

The moral that can be derived from all these examples is the following: the "data deluge" has generated a great attention to new forms of data, which can be engaged in different ways for several purposes. Furthermore, the creation of these data is often linked to expression like "data footprints" (Einav and Levin 2014) or "little data breadcrumbs" (Pentland 2012) suggesting the idea that many of the new available data are constructed by using all the traces that every person leaves behind her during her daily life. However, while the discussions related to the potential of these data are widespread, it is far from being clear whether it is really possible to distinguish between traditional economic data and these new data, generally named "big data" and, consequently, what has actually changed. The interest of both the academia and the industry is often related to the question: what should we do to exploit big data in the best way? However, frequently such worry maintains invisible the first question that should be asked: what is truly new in big data?

This chapter aims at first of examining the accounts until now proposed (both within economics and other disciplines) to tackle this question and, then, it aspires to offer a little contribute to the search for this answer. To begin with, an overview of the first uses of this expression and of the main important reflections on the features of big data is provided. In particular, the attention is on the "3 Vs" recognized by Laney (2001) and on the reflection offered by Kitchin and McArdle (2016). After the description of these proposals, they are claimed to be unsatisfactory and not able to clarify the great difference (and the allied effects) between traditional data and big data. I shall then offer a new proposal developed from the consideration of the velocity and exhaustivity of big data and I will use it as a starting point to analyse whether and how the new generation of data has caused changes in the economic data. As in the previous Chapter, the focus will be in particular on the macroeconomic domain and on macroeconomic data; moreover, this time the examination will be conducted with the help of a case study from the macroprudential domain, the data cube described by Sarlin (2016).

2.2 How big data have been described

The "data deluge" characterizing the last decade has induced many scientists to investigate the phenomenon and to formulate claims regarding its potential and novelty. Since the new generation of data has caught the attention not only of academia but also of the business world, often the reflections proposed by one of the two fields have had an impact on the other. Moreover, within

the academia, biologists, computer scientists, social scientists, philosophers and many others have tried to cope with this concept, generating a huge quantity of considerations.

The first economic reference to the expression "big data" in the modern sense in a title is Diebold's paper on macroeconomics and macroeconometrics "Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting" (2003), presented for the first time in 2000. As Diebold (2012) underlined, the term "big data" was created in that occasion to denote a still ambiguous phenomenon identifiable with the growth in available data. Considering the change in the approach to macroeconometric dynamic factor models, his paper stressed the difference between the old and the new environment in which dynamic factor models are created. He claimed that the latter enables the analysis of much larger data sets while nonetheless retaining a likelihood-based approach. The emphasis of this approach was moreover linked to the explosive expansion of available data: from this reflection, the expression "big data" arose to conjure a stark image.

Initially, thus, such expression did not involve any consideration of the differences detectable in new data with respect to "traditional data": the only novelty was, according to Diebold (2003), the massive quantity of data accessible to scientists. However, soon "big data" started to be associated with specific properties detectable in data. The most pervasive reflection on big data became Laney's proposal of the "Three V's" of Big Data (volume, velocity and variety) in an unpublished research note at META Group (2001). According to Laney, big data do not involve merely a growth in the produced data. Volume is of course one of the features initially taken into account to distinguish between traditional and new data, but the massive quantity of data presents also other peculiar properties. Velocity, in fact, appears to be another vital aspect of the new generation of data. This term usually denotes the speed at which data are created, for instance Arnold (2011) underlined that Google "receives" 35 hours of digital video every minute. In addition, big data can also flow at an incredible speed and often need to be handled very fast (Eaton et al. 2012).

Finally, the expression "big data" is used to identify numerous categories of data: from biology to social sciences, many disciplines have witnessed a big change caused by the expansion of the Internet, the creation of new forms of data collection and the generation of social media. As a result, new types of data sources have emerged and different forms of data have started to be produced. In general, data are categorized into three classes: structured, semi structured and unstructured data (Kitchin 2014). Traditionally the collected data were structured data, with defined length and format such as numbers and dates, but nowadays this kind of data are about 20% of the existing data (Hurwitz et al. 2013). The remaining 80% is composed of data such as

email messages, videos, photos, audio files, word documents and web pages. These data belong to the categories of semi-structured and unstructured data, called in these ways because they do not have numeric or alphanumeric values. The information contained in them requires generally more efforts in order to be extracted and structured (Soman, Diwakar, Ajay 2006).

Although Laney's proposal was not published in an academic paper, in the following years many experts from the academia as well as from the business and media have adopted his description of the 3 Vs of big data. It has become a commonly accepted basis for further analyses on the nature of big data (Eaton et al. 2012, Hitzler and Janowicz 2013, Kitchin 2014). Also within the economic domain, Einav and Levin (2014) have proposed a very similar idea, claiming that the novelty in big data is that

> "[…] data is now available faster, has greater coverage and scope, and includes new types of observations and measurements that previously were not available. Modern data sets also have much less structure, or more complex structure, than the traditional cross-sectional, time-series, or panel data models that we teach in our econometrics classes." (Einav and Levin 2014, p. 3)

Moreover, after the recognition of these three Vs, many other Vs have been proposed to adequately describe big data. Veracity, that is interpreted as the truthfulness of data, is underlined as an important element since data might be noisy and may contain uncertainty (Marr 2014); value, understood as the many insights that can be extracted, is another feature attributed to big data (Marr 2014, Khan, Uddin and Gupta 2014). Another term associated with these data is variability, which denotes that data's meaning seems to shift constantly in relation to the context in which they are generated (McNulty 2014). In addition, since from the cost of big data storage it may follow that big data need to be handled fast, the term volatility has been proposed to describe the fact that big data retention period should be as short as possible (Khan, Uddin and Gupta 2014). To end with, validity can be referred to the correct use of such data (Khan, Uddin and Gupta 2014). Although the search for other Vs to be added to the description of big data has often influenced the debates by generating a sort of "V" mania, many other features belonging to such data and not represented by "V" words have been identified. Kitchin (2013, 2014) has offered a good overview of these properties, based on the recognition of four axes in addition to the 3 Vs along which traditional data and big data can be separated: (i) exhaustivity, (ii) resolution and indexality, (iii) relationality and extensionalty and (iv) scalability.

First, the term exhaustivity is engaged to underline that big data are capable to (and seek to) capture a whole system. In other words, they do no present only a sample, as in the past, but they

aspire to obtain all the relevant data for a specific domain (Mayer-Schonberger and Cukier 2013). Second, resolution in big data is considered maximized: this claim is generated from the observation that now data can show different levels of granularity. For data about people, the most obvious level is the individual one; while for objects and spaces granularity can vary. For instance, the granularity of data about spaces can differ according to the scale: it is possible to have data on households, neighbourhoods, districts, towns and so on. According to this position, hence, big data are characterized by a fine-grained resolution. Data's resolution can be moreover related to the increasing granularity of identification, which can be described as the capacity to generate unique codes to identify massive quantities of data such as banking transactions. In addition, data are stated to be uniquely indexical since even more often codes composed of strings of numbers and letters stand for unique identifiers such as the time, date, and place of a specific transaction (Dodge and Kitchin 2005).

Third, the concept of relationality is strictly connected to the opportunity to create links and comparisons between data sets with similar fields (Boyd and Crawford, 2012), while extensionality refers to the possibility to add or change related features to the system with a minimal development cost, for instance doing large-scale migration (Marz and Warren 2012). In conclusion scalability indicates the capacity that data sets have to expand rapidly and the related ability of a system to maintain performance under growing load through the addition of more resources (Marz and Warren 2012).

Until now I have just proposed a survey on all the features attributed to big data. Now it is the time to consider the points on which this analysis has cast light. Let us start with the "V" characteristics: many of them appear strictly related to the use of big data. Veracity, for instance, puts questions about the truthfulness of used data, while volatility emphasizes the costs involved in the storage of data and the necessity of using them rapidly to reduce these costs. Validity clarifies that not every use of big data is legitimate, and that every purpose should employ the appropriate data sets; value is the ultimate purpose of every analysis involving data. In addition, among the characteristics underlined by Kitchin (2013, 2014) relationality and extensionality, as well as scalability, do not highlight inherent features of data but possible usages of data sets and human interventions on them. It is what happens with the concept of relationality, which highlights the possibility to link data generated at diverse times or locales as well as the opportunity to tie together entirely different data sets which share some common fields. Needless to say, data themselves cannot provide unambiguous relationality, and such links between data are in general made possible by human interventions. It is the case for data produced through the social network Twitter, which are often composed of text and/or an image. The identification of

the content as well as the recognition of possible relation is not automatic, but it requires human interpretation (Boyd and Crawford, 2012). Also extensionality is clearly associated with data's use given the accent on the necessary cost for the addition of features to the data sets, while scalability is referred to the ability to maintain the same performance when data sets are expanded.

The importance assigned to the employment of big data is comprehensible: scientists are in general personally involved in their use. However, this means that the question about the nature of big data often remains in the shadow[5]. Perhaps the most important question related on this "data deluge" concerns exactly the use of data, and not their properties. However, since the distinction between traditional data and big data has been proposed and discussed, a clarification seems at least necessary. Let us look again at the examined descriptions: some characteristics appear prima facie candidates for this task, those which are not just linked to the employment of big data. The 3 Vs of Lanely and exhaustivity, resolution and indexality can in fact be analyzed to find out whether they are really new features that allow to distinguish between traditional and big data.

In what follows, I shall provide this study. Given that Kitchin and McArdle (2016) have already proposed a similar work, my analysis starts with their examination. This appears particularly useful since they have investigated the nature of 26 data sets recognized in the literature as sets of big data. These data sets, selected from seven domains (mobile communication; websites; social media; sensors; cameras/lasers; transaction process generated data; and administrative) are not considered an exhaustive list of all the categories of big data, however they are used as a sample for illustrative purposes. The aim of their work is in fact to gain a higher conceptual clarity on the properties actually owned by big data.

My investigation follows the path proposed by Kitchin and McArdle, but I reach a conclusion that diverges from the one they claim: while they state that at least velocity and exhaustivity are always present in big data and, for this reason, such properties appear good candidate to describe big data; I hold the position that none of these characteristics sheds light on the true difference

[5] Also a great group of the other descriptions of big data maintains the same perspective on the way in which big data can be employed. For instance, Schroeder's position (2014) is based on the idea that big data should be considered as a research made possible through the capture, aggregation and manipulation of data regarding a specific phenomenon on an incredible scale and scope. According to such consideration, the novelty provided by big data is not in the data themselves, but rather it is identifiable in the relation between the subject of study and the available digital tools and the materials engaged in the analysis. Another example is the description of big data provided by Manovich (2011) with respect to the computational difficulties involved in their analysis or in their storage: big data are data sets large enough to require supercomputers and not standard software.

detectable from data to big data. From this conclusion, I shall try to offer a new account of what I consider the most relevant novelty of big data.

2.3 What is new in big data?

As mentioned above, to tackle the question concerning the novelty of big data (and thus the key boundary markers indispensable to distinguish between these data and traditional data), 6 candidates have to be analysed. Each of them denotes a specific property that big data shows. The first one is volume, which has been often stated as one of the main differences between traditional data and big data. Despite such term can immediately suggest the idea that big data have been recognized as diverse because of the growing quantity of collected data, the concept of volume still lacks a clarification in relation to big data. As Kitchin and McArdle's observation of 26 data sets illuminates, several data sets might in fact have very different volumes, and this feature can be interpreted and measured at least in three different ways: data may have a larger volume (i) because of the quantity of data produced; (ii) in relation to the storage needed per datum or (iii) due to the total storage required for a data set.

Let us consider the first interpretation: big data are different from traditional data since the number of data produced is higher than in the past. This consideration was one of the reasons why the expression "big data" was created and it marked the beginning of the distinction between traditional data and big data (Diebold 2012). In addition, also the expression "data deluge" seems to allude to such idea. Even if the issue that this number remains undefined is not considered, it appears that some data sets generally claimed to contain big data do not fit this interpretation. In fact, the number of data generated and stored in them is not only similar, but also smaller than the quantity of traditional data contained in some data sets. For instance, data sets containing data on pollution or deriving from a sound sensor, generally considered big data, comprehend a group of data numerical inferior to the amount of data contained in many traditional data sets like a Census.

The second interpretation is related to the measurement of data size. Again, it involves the same issues of the first: claiming that each "big" datum has a bigger size than traditional data is vague since no specific demarcation has been proposed, moreover it appears also false given that for some of such data the volume per record is lower or similar to the volume of traditional data. A part of big data really present high size, this cannot be denied. For instance, images and videos have a bigger volume (thus, a bigger size) than traditional data. Nevertheless, data on consumer behaviours, call records and data produced by means of sensors, whose size can be really small (Kitchin and McArdle 2016), shed light on the limit of this reading of the term "volume".

Finally, the last possible meaning of "volume" deals with the total storage required by a data set. This is simply the sum depending on the volume of each datum and the total number of collected data. While in some cases, even though the low volume per data, the devices involved in the generation of data produce very large storage volumes, other times big data do not show this feature. For example, transaction data produced by recording the behaviours of Walmart stores' costumers reach a volume of 2.5 petabytes per hour (Open Data Center Alliance, 2012), but a data set containing all Twitter messages about a particular topic are not likely to be as large as traditional data sets.

Let me draw the first conclusion. Given the three interpretations of big data's volume and the problems related to each of them, two possible positions can be proposed. On the one hand, someone may claim that any discussion about big data requires an initial decision: one interpretation of volume should be selected and a precise number relative to this volume should be chosen to draw a line between big data and non-big data. For instance, if the initial decision were that only very large data sets reaching *n* number of data are eligible to be considered big data, data sets on pollution would be likely to be rejected from the "big data group". On the other hand, someone may state that volume is not, after all, a fundamental characteristic of big data since also data without large volume bring new challenges often related to the concept of big data. As Dodge and Kitchin (2016), I embrace this second position. If it is true that volume is not the boundary marker we are looking for, our research has to proceed.

The next property to be questioned is variety. To begin with, it is hardly deniable that some traditional data sets show a level of variety as high as the level characterizing the new data sets. Moreover, it can be also added that variety seems to be an important trait only for a part of big data. Across the 26 data sets investigated by Kitchin and McArdle, in fact, different levels of variety can be identified and confronted to traditional databases. What the authors suggest is that the introduction of a fair range of variety should not be considered as a true novelty involved by the emergence of big data: disciplines such as social sciences and humanities have always dealt with qualitative data such as text and images, and their traditional data sets showed a high level of variety also before the generation of big data. In addition, their analysis underlines that many data sets containing big data are actually composed only of structured data, thus they are not characterized by variety. For instance, big data created through sensors such as traffic data or other categories of big data like credit card data have numerical shape. It is true that nowadays the amount of unstructured data is increasing, however this does not entail that variety is a true marker of big data.

Let us go further to consider the features known with the term "resolution" and "indexality". Again, it exists not only in big data sets, but also in traditional data sets. In fact, even though many data sets can now offer high levels of granularity, such possibility existed also in the past. By means of surveys, for instance, it was possible to produce data at the individual and at the aggregate levels. In addition, also indexality does not allow scientists to distinguish data and big data: many data were produced in the past with the exact purpose to indicate a specific object, person or element. Not only codes were not used for the first time in big data, but for a long time they have played a vital role for data within some disciplined such as biology. As an example, the work of the biologist Rheinberger on the molecular structure of the DNA sequence of bacteriophage PhiX174 illustrate this point. Once Rheinberger visualized the structure of bacteriophage PhiX174, he produced a datum on it using a chain of symbols (GATC) standing for the four nucleid acid bases. In this way, the datum composed of letters stand for the unique identifiers of this bacteriophage (Leonelli 2015). Moreover, also many data about individuals employed the same numerical method to establish unique relations between data and persons: a simple example may be the use of a numerical code to indicate a specific student of a university. In closing, both resolution and indexality are refused as markers for big data.

Until now my examination and Dodge and Kitchin's study have reached the same conclusions: volume, variety, resolution and indexality have been rejected as peculiar properties owned only by big data. However, from this point our reflections follow two diverse paths. For this reason, at first I will propose their account, then I shall develop another answer capable both to partially include their proposals and to explain what is really different in big data. I shall finally provide four reasons to prefer my proposal rather than Dodge and Kitchin's one.

For what concerns exhaustivity, this property is understood as the fact that big data seek to capture the entire elements of a system and not only a sample. For example, all the tweets made by all users are recorded into data, and all the economic transactions of a credit card are translated into data. While in the past many data sets contained only samples of the complete systems, nowadays every object within a domain can be related to its datum. By studying 26 data sets, Kitchin and McArdle arrive at the final idea that exhaustivity is detectable in each of them[6] and for this reason they recognize this tract as a specific property of big data.

Moreover, the same conclusion is reached for velocity. This feature can be interpreted in diverse ways: (i) it can be related to the speed at which data are generated, or (ii) it can be associated with

[6] The only exception was the Twitter data set which contains only a sample of tweets harvested from the total Twitter shares with some researchers.

the incredible rapidity with which they can flow and can be handled. The data sets studied by Kitchin and McArdle (2016) vary with respect to such traits. For what concerns the velocity with which data are created, some data sets contain data generated constantly and very frequently: mobile phone apps, for instance, can produce data on location every 4 minutes. Other data sets are composed of data generated in real time, but not constantly, such as those data produced only at the point of use. Furthermore, with regard to the frequency of handling, recording and publishing, some data are recorded immediately after their production (like a tweet which is recorded almost at the same moment in which is tweeted), while other data can be produced very fast but then be transmitted to central servers (and published) after a considerable period of time. According to Kitchin and McArdle, if both these interpretations are taken into account, all the 26 data sets confirm that velocity is an important feature. Some of them are created sporadically but immediately handled and recorded, others are generated continuously. Consequently, Kitchin and McArdle support the candidacy of velocity as a real distinguishing property of big data.

Before going further, a consideration is required. Someone might claim that actually Kitchin and McArdle's reasoning is a vicious circle: they try to find the specific properties of big data by observing those data that are considered for some reasons (often unknown) big data. Therefore, since the name of big data might be attributed arbitrarily, the examination of the properties that such data show is not useful in order to draw a line between traditional data and big data. Needless to say, the concept of big data is elusive, it is a human construction just as the concept of unemployment, consequently it is difficult to find what the common characteristics of big data are. Nevertheless, I suggest that a little bit of optimism should be kept: if we consider the data deluge, a particular change can be associated with this event, and it is not the increase of velocity or the achievement of exhaustivity, but a technological change that could be seen as the cause of the data deluge.

2.3.1 Toward a new marker for big data

Even though both exhaustivity and velocity seem to cast light on important characteristics of big data, my suggestion is that such traits are just two possible effects of the unique main difference between traditional data and big data: the way in which data are produced. A person who considers the debates about big data is likely to run into sentences referring to the automated forms of data generation. I suggest that not enough attention has been focused on this fact. As long as the worries of scientists about the ways to use big data are examined, it appears clear that these discussions are based on an idea much more pervasive than in the past: data are *given* to scientists, who are not involved in their creation. Not only data are not created directly by scientists, they are furthermore produced through a process that is entirely automated and that requires less

human mediation than that needed for the production of traditional data. A similar idea has already been proposed by Pietsch (2013), who has claimed that a

> "crucial characteristic concerns the automation of the entire scientific process, from data capture to processing to modelling" (Pietsch 2013, p. 2).

Let us return to the three initial examples of this Chapter: data produced within the retail stores, data generated through the use of the Internet and data on mobile phone calls. All these data are produced because human activities are performed; people buy something, use the Internet or call by using their mobile phones. However, the creation of data on these activities does not require a person performing particular tasks. In the data generation process, human activities are required just at the beginning, when the devices used to produce data are implemented. Once these tools are produced, their work does not necessitate the involvement of humans to be performed and data are created automatically. Such aspect is much more evident for data on traffic or for satellite images, in which the automated process alone is capable of generating data.

Surely, this claim may be followed by the question: why the automated data production should be accepted as a better marker than exhaustivity and velocity? My claim is that four reasons can be provided to support the distinction based on the production of data rather than the one proposed by Kitchin and McArdle.

First, let me cast light on the "Achille's heel" of Kitchin and McArdle's conclusion: the concept of exhaustivity causes a lot of confusion. When they describe this notion, they state that data sets do not have to be really exhaustive, rather big data entail that data sets *aspire* to reach exhaustivity. In other words, there is no need to capture all data, what makes big data sets different from traditional data sets is that the former are produced with the idea that it is possible to have an exhaustive collection of data, therefore the collection should not be limited merely to a sample. Once they observe the 26 data sets, they rapidly conclude that this property characterizes all the data sets because such data sets do not contain only representative samples harvested from the totality of potentially available data. The boundary between data and big data hence appears, according to this account, the use of samples. Big data sets do not contain them, but much more data, with the aim to capture all the data. The big danger that might emerge from this property is that while the use of samples is always accompanied by the awareness that only parts of the elements are recorded into data, the concept of exhaustivity may weaken this consciousness. Also Kitchin and McArdle seem to have a vague idea of the very important difference between *seeking exhaustivity* and *being exhaustive*: when they describe the 26 data sets, they state that each of

them holds the characteristic of "n=all". Nevertheless "n=all" is a tricky idea that might really compromise the quality of a research.

This point can be clarified with an example: a social scientist would like to study the behaviour of the persons who search for information on the Internet. In theory, it would be possible to obtain all the data created by all the users when they "surf" the Internet. However, there is a group of users who tries to avoid the generation of data on their searches and activities in order to protect their privacy. Some users, because of their abilities with computers, manage to "hide" their activities on the Internet. In this and other cases, the statement that "n=all" would prevent researchers to doubt and ask whether part of the relevant information is missed. On the contrary, the description of big data as data produced through automated processes does not deal with the idea that big data are *all* the data of a specific system and that the data set is exhaustive. The only idea entailed by the automated process of data generation is that the production's cost is not as high as in the past (when, because of this reason, only samples were produced), and that more data can thus be created. This may mean that a very big sample can be collected or that exhaustivity might be seen as an ideal purpose, depending on the human intention.

Second, one advantage of considering the process of production as the remarkable difference of big data deals with the fact that the more "givenness" of such data, in comparison to traditional data, is highlighted. This chapter has already stressed that a frequent question asked by scientists when they look at big data is: how can we use them? With the term "givenness", I denote the fact that scholars do not need to supervise the generation of data, but they receiv directly what has been recorded. As a consequence, the claim about the automated production of big data might aid to underline that this process of data generation is performed by involving only partly researchers' background assumptions. Furthermore, as a consequence of the "givenness" of data, scientists have started to look at data with a growing curiosity: how can be exploited? What can they reveal? This curiosity, Leonelli claims (2012), is at the origin of the great impact that data have had in science: even more scientific disciplines have become data-intensive sciences.

Third, the consideration of big data as generated by means of automated activities allows scholars to spell out that the confidence in them should not be too high. Human errors are generally taken into account when traditional data are considered, but the trust in devices is often much stronger and the experience has taught that sometimes it needs to be resized. Furthermore, it should be accepted also that instruments generating data could need for improvements. At the start, the core challenge could be that the examined big data are not the output of devices designed exactly to create valid and reliable data amenable for scientific research. For instance, it is essential to remember that data produced through the use of the Internet can be employed by several scientific

research without being generated for those purposes. Other times the crucial trouble deals with the change over time of the tools generating data, which is consequently linked to the quality of data. Social network's algorithms implemented to produce data have been changed many times (Lazer et al. 2014) and with them the data that they can create. The comparison between "old" and new social network's data should be grounded in this awareness. Finally, the automated process of data production is not free from possible human pressures, for instance data on Web searches might undergo the attempts of manipulating the data generated to reach economic or political gains. It may be the case, for instance, that political campaigns or business companies use several tactics to make their candidates or their products trending.

Fourth, the emphasis on the automated production as the marker between traditional and new big data does not imply forgetting about velocity. Rather, it is such form of data generation that causes the velocity (both as frequency and as rapidity) with which data are created. Data can be produced every *n* minutes because the process is automated: those people who implement the devices collecting data have only to decide the time interval between one datum and another. Furthermore, data are created almost real time exactly because there is not the human brokerage between what is recorded and the datum.

In closing, the reference to the automated process of data production as the marker to distinguish big data from traditional data enables to: (i) avoid the problem of considering data sets as exhaustive or just seeking exhaustivity; (ii) underline the "givenness" of big data; (iii) cast light on the faith generally demonstrated in devices and on the need for resizing it and (iv) include velocity as an effect of such automated production.

Two additional remarks have to be underlined. At first, the "giveness" of big data does not entail the *total exclusion* of human activities from the process of data production. Rather, it can be concluded that big data entails just a part of the human activities required for the production of traditional data. Let us consider the data generation process I have proposed in Chapter 1. The path I have underlined has these steps:

Observation of the reality → Vague idea of phenomena → Data production → Patterns → Facts about phenomena→ Phenomena

In this case, human activities are involved in all these passages: scholars observe the reality and create a vague idea of phenomena; then they produce data (for instance by means of experiments or surveys); and from such data facts about phenomena are inferred. If my view is accepted, the only step that excludes human activities would be the moment of data production: data from social networks, for instance, are generated every time an activity is performed.

However, even such data are not completely raw at least for one reason: the development of technological devices for the automated data generation involves decisions about the nature of those data to be generated. An example could be the choice of the physical medium to use when data are created (Leonelli 2015): should the technological tools produce photos or videos, numbers or texts? Moreover, what should be the time interval between the production of a datum and the next? The answers to these questions deal with the scientists' expectations about the data on facts that are more likely to be generated through the use of such data. In other words, the idea that big data could be used to discover and to measure particular phenomena seems to play a role in the decision about how big data should be produced. For instance, a researcher might decide to develop a tool to produce videos instead of photos because he thinks that such videos would enable to find the phenomenon X.

This reflection is not limited to the field of macroeconomics: in order to construct an adequate tool, the individuals involved in its implementation always need at least a "vague idea" of the way in which the produced data could be used (and, consequently, of the facts about phenomena that might be found and measured by using data). Let us recall other examples. When new technological devices are implemented to produce data on the sale of the Internet retailers' goods, their implementation requires an idea about the consumers' behaviours that such data may show through patterns. Moreover, let us imagine that we have the opportunity to create a device able to produce data on the Internet users' activities: our decision about the kind of data to be created would be affected by our suppositions concerning the possible facts about phenomena that may be finally discovered by using that the data set. To implement efficient tools, we need to know what data would be more useful to find facts. With big data, the steps from the reality to the knowledge about phenomena would be almost the same if compared to those exposed in Chapter 1:

Observation of the reality → Vague ideas of phenomena → Implementation of tools for automated data production → Automated data production → Patterns → Facts about Phenomena→ Phenomena

The second consideration requires to bring to mind the difference between the two categories of data that I have described in Chapter 1. Every datum is a datum about an element of the world (such as a photo of a car physically present in the world) or a datum about a fact (like the change of temperature under specific condition). With regard to this division, nor exhaustivity neither velocity can make clear whether big data are data on visible elements of the world, or whether there are the products of the measurements of facts about phenomena. However, when big data

are observed, they all appear immediate outcomes of local measurement: big data are photos like satellite images, texts, price and payment records and so forth. They cannot be measurements of facts such as unemployment.

This position is consistent with what I have already suggested. In Chapter 1, in fact, I have claimed that the former group of data is close to the etymological meaning of the term datum, that is "given"; while I have stressed that the latter category requires a high level of conceptualization to be created. Given that such level of conceptualization is in general related to the presence of human beings, it appears unlikely that this latter group can contain data generated by means of automated process without any kind of human supervision. Moreover, since big data are considered more "given" than traditional data, the former category appears to be the most adequate for them. Since this point is particularly relevant, let me try to be more explicit: big data are never data on facts about phenomena and are always data of the first group I have proposed. They involve automated forms of production, and such forms cannot be employed alone to generate data on facts. Phenomena and their facts need always human beings, who find them and decide how to measure them. However, to find phenomena it is now possible to use big data: for instance, facts about unemployment might be discovered by looking at the Internet data about the users' searches for job positions.

Overall, given both the four reasons provided for preferring the automated process of data production as the boundary between big data and data, and the observation that big data are never data on facts about phenomena, I propose to describe big data as:

All those data produced through automated processes that, for this reason, can be created very frequently and almost in real time. Such data are the outcomes of measurements of local elements and they can be used as evidence for facts about phenomena.

This description can be considered not only an abstract description of the nature of big data, but also a normative criterion in a thin sense: it helps to underline what big data is and what is not, and, more important, what challenges emerge from big data. If we accept that the automated process of data production is the novelty that has really changed something in data, it will be easier to recognize that the data we observe are more "given" than the data in the past, and that as a consequence both their use in numerous and different contests, and the faith with which they are analysed should be questioned.

An illustrative example of the risk related to this faith is the failure of Google Flu Trend. In 2009, researchers from Google proposed to use the search for flu-related information on Google to "nowcast" when people were sick with the flu (Ginsberg et al. 2009). The idea was that search

data, if properly linked to the flu tracking information from the Centres for Disease Control and Prevention (CDC), could accurately estimate flu prevalence two weeks earlier than the CDC's data. However, in 2013 Google Flu Trend spectacularly failed to underestimate the peak of the 2013 flu season by 140%. Several justifications have been proposed, and many of them cast light on this point (Lazer et al. 2014). First, data might be biased: the same person could search for the same information many times and many persons who think they have "the flu" actually do not. Second, it is impossible to reach the conclusion that we have all the relevant data on a specific topic. If data are produced without the supervision of scientists, the estimation of the reliability of data becomes difficult.

2.4 Macroeconomic data in the era of big data: the data cube

In the macroeconomic domain, what are called "data" are generally measurements of facts about phenomena: unemployment, national income and inflation are the common examples provided to explain what macroeconomic phenomena are. Have big data changed something in the production of macroeconomic "data" about facts? To tackle this question, I propose as a starting point the description of macroprudential data offered by Sarlin (2016).

In his recent paper, Sarlin (2016) makes clear that macroprudential oversight, that is the supervision of the financial system as a whole, requires the use of macroprudential data. Such data belong to a group resulted from the union of (i) macroeconomic data, (ii) banking system data, and (iii) market-based data. He represents this group as a data cube, which underlines the three dimensions of each macroprudential datum: the entity to which it is referred, the time and the indicator. For instance, a macroprudential datum might be a datum on the GDP of a particular country at a specific year. By bringing together in a cube all the macroprudential data, thus, it is possible to have data on different indicators for diverse entities at several points of time. Let us turn the attention on Figure 1. What is important, according to Sarlins, is that macroprudential data can be confronted by focusing on the way in which several entities are described by different indicators at one point in time (thus observing the red slice of the cube), on the changes of entities' values over time (described in the blue slice) or on the changes of different variables over time for just one entity (looking at the green slice). Furthermore, data can be used to observe the linkages between entities at a specific point time (represented by the black edges).

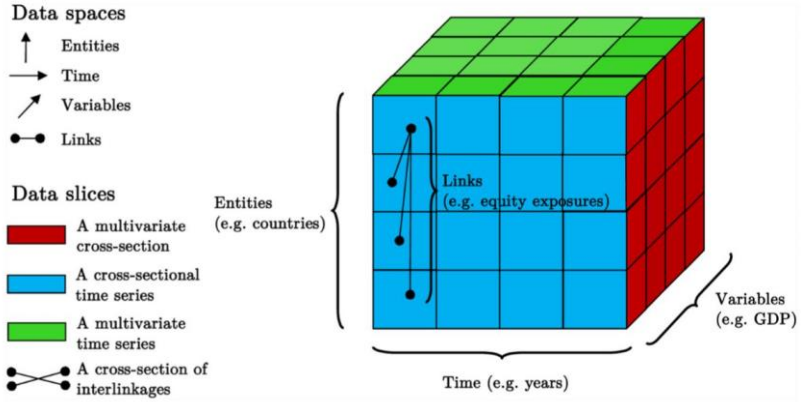

Figure 1: A macroprudential data cube

From the observation of this data cube, Sarlin claims that big data have affected all the three more standard dimensions of the cube: nowadays data (i) on more entities, (ii) at numerous time points and (iii) concerning different dimensions compose the data cube, and such changes can be attributed to the emergence of big data. I take this claim as a starting point to investigate what has really changed within the macroeconomic domain once the "data deluge" has started. In fact, even if Sarlin does not spend so much time on this point, the cube he presents and the considerations he proposes facilitate the examination of the changes that can be found within the macroeconomic field after the rise of big data. More precisely, given that the data cube is generated by grouping together numerous units with three dimensions (the entity to which the units are referred, the time and the indicators), it can be questioned whether such units are data about visible things or about facts. Furthermore, the observation of the data cube's dimensions might shed light on the way in which big data have changed them and the macroeconomic research. Of course, other examples could be chosen to conduct this analysis; nevertheless, by using the new marker I have proposed, it seems possible to underline what Sarlin leaves invisible when he says that the data cube is growing. In addition, the next considerations appear related to the idea I have already expressed about the path that goes from data of the first category, to facts and "data" of the second category, to new facts and new "data" of the second category. In Chapter 1, I have tried to represent this form of reasoning with the following steps:

Data → Patterns → Facts about phenomena→ Measurement and new "data" → Patterns → Facts about phenomena → Measurement and new "data" …

In the next pages and in Chapter 3 I propose that both the description of big data and this reasoning path can be employed to understand what has changed in the macroeconomic domain.

To begin with, I have already expressed the idea that big data can only belong to the first category of data. However, if we consider the data cube, it can be observed that one of its dimension deals with the indicators contained in data. Each indicator is the outcome of a specific measurement, and what is measured is almost always a fact about a phenomenon: indicators on inflation, unemployment, expenditure, leverage, deficit and so on are in fact obtained by aggregating several data and by conceptualizing an economic phenomenon. If the data cube is made of units containing indicators, it follows that its units are data on facts about phenomena, not data on singular visible object of the reality.

Each unit of the data cube is said to be a datum containing information on the *measure* of an *indicator* for a specific entity in a specific time point. More entities, more time points and more indicators are now found in the data cube, Sarlin claims. It seems, hence, that big data can affect macroeconomic "data" on facts. Though, if it is accepted that big data are produced through automated processes and that, as a consequence, they belong only to the first category of data, it remains to understand how big data (of the first category) can affect the data cube's units (of the second category).

Let us further examine the data cube. The presence of indicators in these data helps to underline again that the units of the data cube are all data on facts about phenomena. This claim becomes clearer if the attention is on the indicators contained in the units of the data cube: the macroeconomic data cube might, for instance, contain a unit about the GDP of the UK for the first semester of 2016. This is, of course, not a datum about an element of the world, but a datum on a fact. However, GDP, as well as all the other economic indicators, has to be formulated and the ways to measure facts have to be established in order to generate the units of the data cube. To do it, data of the first group are used: between them new patterns are discovered which might be interpreted as facts about relevant phenomena, and then these facts could be measured by using data of the first category. The number attributed to a particular indicator is the measurement of a specific fact. As suggested above, big data can be employed to perform this activity: between them patterns can be discovered and they can be involved in measurements.

Big data → Patterns → Facts about phenomena→ Measurement and new "data" (data cube's units)

Now, big data might be data on the same subjects already measured in traditional data: for instance, retail store data were collected also in the past as well as the prices of economic goods. Such data, nevertheless, were produced through surveys. For example, the Bureau of Labor Statistics generated data on goods' prices by sending its surveyors out to numerous stores to

manually collect information on the prices of approximately 80,000 selected items. In these case, hence, the automated process of data generation leads to the creation of data in less time than that needed for producing them with the traditional method. The consequence of the automated data creation would be the velocity. This would also entail that the frequency with which such data are generated might allow scholars to perform more measurement (aimed to obtain indicators) and it would consequently shorten the time interval between a measurement and the another. Let me make an example: instead of having only an annual measurement of the inflation rate of a specific country, the velocity of automated data production might lead to performing the measurement of the inflation rate more frequently. The results of such increase in the number of measurement activities would be that more data about the phenomenon of inflation could be created. In other words, the data cube would contain not only an annual datum about the inflation in the country X, but it would start to contain more data on inflation related to a particular year. For instance, data on inflation could be produced each semester, or each quarter of semester. This is a first way in which big data may affect the dimensions of the data cube. The automated process of data generation entails an increase in the velocity with which big data are created, and the possibility to have big data more timely and more frequently than in the past allows researchers to use them more frequently to find patterns about facts. Therefore, the more frequent use of big data by researchers leads to a more frequent creation of data about facts. Given that more data about facts are created, the data cube dimension grows.

Moreover, the automated production of data may also enable to obtain more easily than in the past data on different entities. Describing the macroprudential data, Sarlin underlines that they can be divided according to the level of the entity. Data can be measurements of facts regarding low-level entities such as households, firms and assets, or can be data concerning facts involving the banking system or the financial markets, considered aggregated entities. In the past, however, data on households or small firms were often not produced because of the high costs involved in the data generation through surveys. Again, the automate big data production has made such collection much more feasible: data on households' behaviours can be for instance generated from the available information on the activities they perform on the Internet. In addition, also the several activities of small and big firms can be tracked in this way. For this reason, while in the past the data cube was composed especially by units with indicators related the bigger entities (such as the most important banks), now also units containing indicators about small banks or small category of households can be included. This is the second way in which big data can cause the growth of the data cube: given that now it is possible to generate more data about low-level entities, such data can then be used by researchers to detect facts about phenomena at low-levels

and to produce data about these facts. Such data can then be added to the other data about facts at higher levels.

Finally, big data are likely to be data on elements of the world until now not measured and made observable in data. Data on the Internet searches and GPS data on the individuals' movements are perfect examples. The new big data can be used in two ways: (i) they can be observed to find out whether there are patterns suggesting a relation with facts about phenomena already studied using traditional data or (ii) they can be explored to find new patterns suggesting new facts about phenomena and new phenomena. Within the macroeconomic and macroprudential domain, it seems that such data are generally utilized in the first way. Big data, for instance can be used to find patterns linking known phenomena such as unemployment, costumer confidence and inflation (Arola and Galan 2012, Daas and Puts 2014). Both of these uses, however, lead to the formation of new indicators, thus of new process of measuring facts about phenomena.

This means that, again, big data can cause an expansion of the data cube dimension: this time the dimension to be taken into account would be that containing the indicators. If big data contain new information, then such information can be used to produce new indicators, and consequently new data about facts. Let us consider an example. In Chapter 1, I have suggested that a phenomenon such as unemployment was recognized as such because researchers "had an intuition" about its existence and its relevance. Only after its recognition, suitable data to study it were collected, and subsequently through their use the phenomenon was further conceptualized and the decision concerning how to measure it was made. The data engaged in the measurement of the phenomenon were generally created by means of surveys. However, economists and social scientists have had an intuition: big data created by social networks may show patterns actually related to this phenomenon. For instance, many studies have been conducted to verify whether Twitter data can be used to formulate a new indicator for this phenomenon. Antenucci et al. (2014), Lorente et al. (2015) have developed new indicators and tested their validity by comparing the obtained data with the data produced using the traditional indicators. Even though their research is just at the beginning, from these examinations they have concluded that such indicators work. In addition, data on price index provide another example of how big data may be used to form new indicators. The so called "Billion Prices Project", developed by Cavallo and Rigobon (2012), has used big data from online retail store to offer a new measure of retail price inflation. Since these indicators are developed to complement, not to replace, the existing indicators, once they are accepted by the researchers' community they are just added to the list of indicators. Therefore, more macroprudential data can be produced because more possible measurements exist. In other words, also the third dimension of the macroprudential data cube has been enlarged.

After this analysis on the data cube offered by Sarlin (2016), we are ready to explain how big data have changed the macroeconomic domain even though they are not directly added as its units. When the data cube is observed, it can be stated, as Sarlin does, that the three dimensions are growing.

First, the automated process of big data production allows to have data real time or in very short period of time, and this enable to use them to measure facts more frequently. As a consequence, macroeconomic facts are measured more times than in the past, and more time points can be found in the data cube.

Second, big data, because of their low production costs, offer more data on small entities than those available in the past. Such data can be exploited to obtain more measurements on facts at different levels, and in this ways also the entities of the data cube grow.

Third, big data record elements and details until now not recorded in any way. Many scientists have recognized in them a gold mine, and have tried to use them to find new patterns concerning known and unknown facts. As a result, new measurements have been developed and new indicators have been formulated. They are added to the existing indicators, enlarging the third dimension of the data cube.

2.5 Economic data reconsidered

To sum up, the debate on the importance of data has taken a significant turn over the last ten years, when a sort of "data deluge" has started to change the nature of the data until that moment produced and analysed. I have proposed to look at such big data by casting light on the novelty of the way in which they are produced, and I suggest that the key boundary marker of big data is the automated generation of these data. I have proposed a new description of big data providing a normative criterion: it casts light on what has changed from data to big data and on what are the risks related to this change. Furthermore, such automated form of data production can be used to formulate explanations both for the other changes involved in the passage from data to big data (such as the growing velocity with which data are created), and for the claim that big data are never data on facts about phenomena.

The last consideration helps to understand which role can be assigned to big data within the macroeconomic domain. Given that almost all the macroeconomic "data" are data about facts (as the units of the data cube, containing indicators, show), big data cannot be involved directly in macroeconomic studies. However, as the case of the data cube clearly describes, big data affect those data used within such field. This fact is possible because big data can be employed to find

and to measure data on facts about phenomena. As a consequence, big data can indirectly perform their influence on the macroeconomic “data”.

Chapter 3

Observing macroeconomic data

3.1 A comprehensive view of big data in macroeconomics

In Chapter 2 I have offered a new description of what big data are and I have tried, by examining the dimensions of the macroprudential data cube, to shed light on the ways in which big data have changed the macroeconomic domain. My conclusion is that, since big data are never "data" on facts about phenomena, they can have an influence on the macroeconomic realm only because they are used to produce macroeconomic "data" about facts. In other words, big data are engaged in order to *generate* new units of the data cube, but big data cannot be *directly added* as units of the data cube.

The final step of this thesis deals with the examination of how big data are actually used to produce macroeconomic data. In other words, is the passage from big data to facts different from the passage from traditional data to facts? In addition, have big data caused a shift in the way macroeconomists approach the even more numerous group of macroeconomic data? To answer such questions is the goal of this Chapter. The present Chapter follows from Chapter 1 and Chapter 2 pretty naturally: the group of reflexions proposed until now will be employed as a "toolkit" to examine how economists approach macroeconomic data. As we will see, the claim I will support is the same that many authors have claimed after the publication of Bogen and Woodward's paper (1988): the passage from data to facts cannot but be mediated by theoretical assumptions, even in the case of big data. By using the concept of "data model" (Lynch 1990; Harris 2002), I shall conclude that the only difference between the approaches used with big data and those used with traditional data is that the new approaches are likely to lead to the generation of more data models (all theory-laden) with the aim to enhance the possibility to exploit expert judgments.

3.2 Visual analytics

While in the past data were analysed especially through statistical tools (Glymour 2000), the new approaches used to examine data within the macroeconomic domain have been influenced by the rise of a new discipline called visual analytics. Visual analytics is a field developed from the field of information visualization. The main idea in which both these fields are grounded is that the increasing complexity of the data sets requires a greater aid to support the exploration of such data. Information visualization was thus developed to provide this aid by means of a new

interaction between human and computer. Information visualization has been described as "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition" (Card et al. 1999, p. 120). Over the last years, a new discipline has emerged from the combination of the typical ingredients of information visualization and analytics. This new discipline has in fact strong roots in information visualization, but it is also strongly related to data analytics. In a nutshell, visual analytics is: "the science of analytical reasoning facilitated by interactive visual interfaces" (Thomas and Cook 2005, p. 2).

The initial assumption involved in the development of visual analytics is that it can support the process of phenomena identification considerably: huge quantities of data are likely to cast light on several relevant patterns which, in turn, may lead to the recognition of new facts about phenomena and to their measurements. While often the automated process of big data production remains in the shadows, in general big data studies are associated with learning machine algorithms, aimed to infer models from data. In visual analytics these algorithms are coupled with numerous forms of visualization that allow scientists to *observe* data. The aim of visual analytics is to enhance, through the combination of analytical tools and visual devices, the possibility to capture all the important information that a data set might offer. Learning machine algorithms can quickly identify patterns between data and can produce models, while data visualizations enable scientists to identify patterns that otherwise would not be recognized by standard algorithmic means (Keim et al. 2010). The possibility of visual analytics processes depends on human interactions: analysts have to decide whether to apply exclusively visualization or automated analysis, whether to start with the automated analysis and later use visualization or whether to do the contrary, starting with visualization. Figure 2 (Keim et al. 2010, p. 10) illustrates the possible linkages (arrows) between these different stages (represented by ovals).

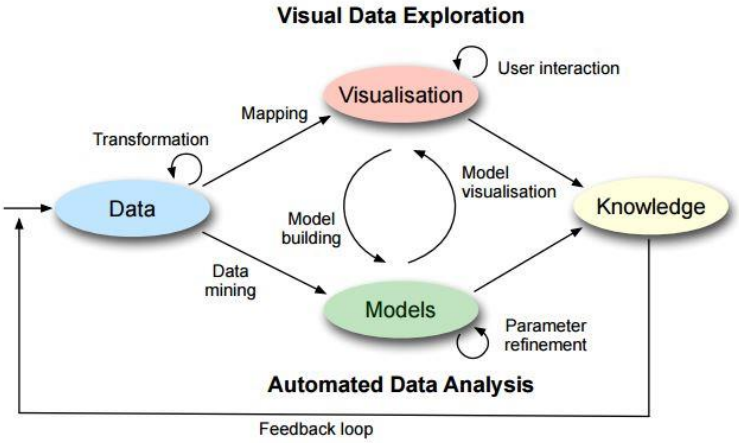

Figure 2: The visual analytics process

The figure represents the general visual analytics process, the same that Sarlin (2016) attributes to the analysis of the macroeconomic data cube's units. However, a clarification is required. The data cube's units are "data" about facts, not data of the first category. Nevertheless, Figure 2 shows the passage from data of the first category to knowledge, hence it shows what happens *before* the analysis of the data cube's units. In other words, to embrace also the use and the analysis of the data cube into the process, Figure 2 should be extended to include another step, the one from "data" about facts to knowledge. To find and measure facts about phenomena, therefore to create "data" about facts like those contained in the data cube, at first data of the first category, such as big data, are used and analysed, and by using them measurements and indicators are formulated. When "data" about facts are generated, they can be used in turn to find new patterns and phenomena, such as those studied in the macroeconomic field (Sarlin 2016). It appears, consequently, that visual analytics can play a central role in both the two steps leading to knowledge: to begin with, visual analytics can be employed to study big data and to recognize patterns between them, then from such patterns "data" on facts about phenomena can be generated and they can be analysed again by means of visual analytics to find new patterns.

Let me try to extend the process of Figure 2: Figure 3 shows what happens *before* and *after* the generation of facts about phenomena. The purple oval corresponding to ("data" about) facts is the one to which the macroeconomic data cube would be assigned. The arrow from the yellow circle to the purple circle, instead, shows what has already been underlined: this process of phenomena discovery is almost a never-ending process. When facts are discovered and "data" about facts are created, they can be used in turn to discover new patterns and new phenomena. This means that also those facts reached in the yellow circle may be added to the data cube.

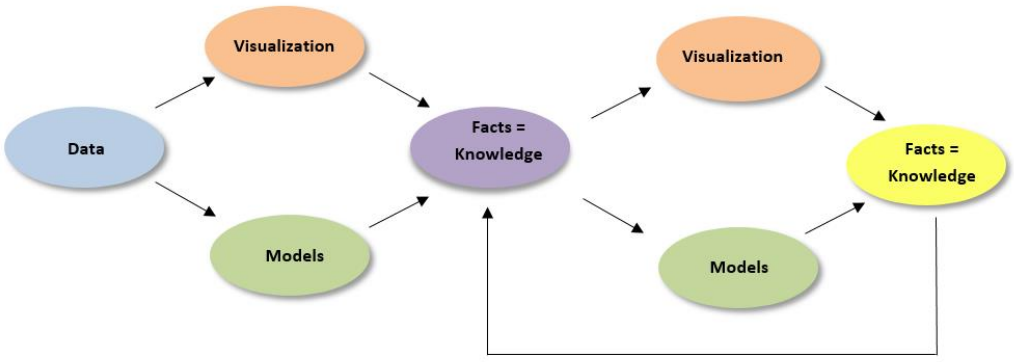

Figure 3: The visual analytics process from data to facts and from facts to knowledge

As shown in Figure 3, visual analytics can be engaged also to observe patterns between facts like the units of the data cube. Also in this case, visual analytics may enhance the possibility to find relevant unexpected patterns by allowing analysts to explore facts. For instance, let us assume that an analyst has a group of data cube's units containing quarterly observations relative to 20 macro-financial indicators for 30 economies from 1990 to 2011. Perhaps this analyst desires to observe all these units together. However, given the massive quantity of information contained in such group of units, he could decide to select only part of these units to explore possible patterns. For instance, he might decide to focus only on the change of a selected indicator for all these economies: to do it, he might use a dashboard like the one represented in Figure 4 (Sarlin 2016).

The dashboard in Figure 4 presents a time-series plot for a specific indicator and 30 economies. In this case, the chosen indicator is the real credit growth, but it could be replaced by another by using the drop-down menu. Moreover, also the indicator's transformation can be changed through the radio buttons. The analyst's exploration can also involve activities of zooming and filtering: for instance, he may decide to show only one individual economy, or to drop an economy from the visualization by deselecting it, or to visualize only a time span. Finally, a drop-down list of events can be used to select the events to be shown from the event line.

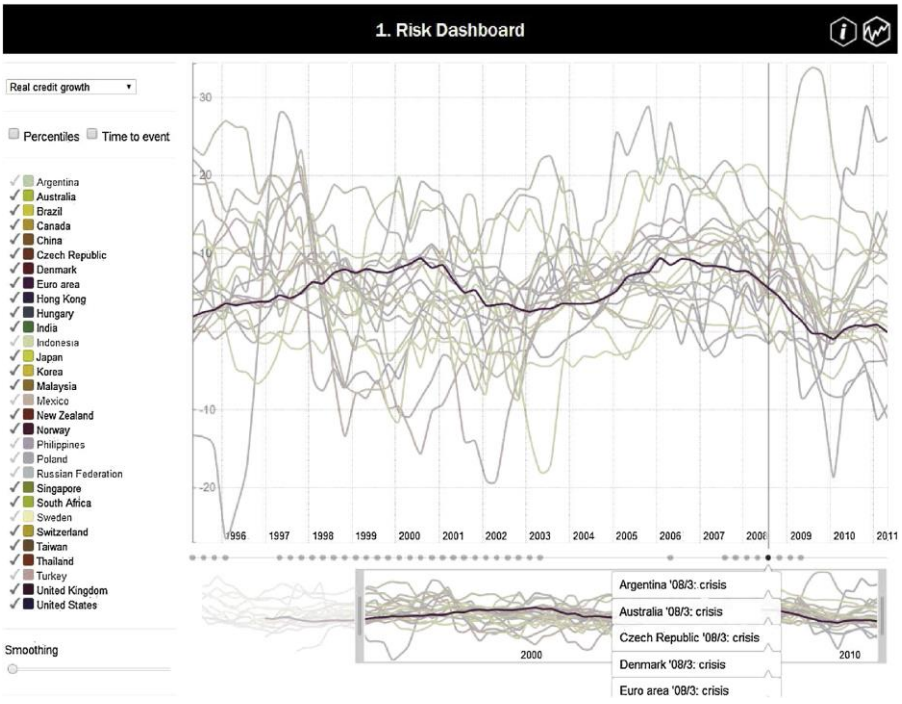

Figure 4: An interactive risk dashboard

All the points represented in the dashboard are units of the data cube, they are "data" on facts about specific phenomena that are observed together with the aim of discovering relevant patterns, patterns that could in turn reveal facts about new phenomena. One of the most important goals of visual analytics is exactly this one: it provides tools to

> "synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data" and at the same time it enables to "detect the expected and discover the unexpected" (Keim et al. 2010, p. 7).

3.3 "Big data to phenomena" reasoning

The description of visual analytics that I have proposed may lead to the idea that something has changed in the path from data to phenomena (and from "data" about facts to new phenomena). However, a further analysis of the steps involved in such "data to phenomena reasoning" clarifies that, despite the level of technology engaged, the approach has remained almost the same. In Chapter 1, I have underlined that the data/phenomena distinction proposed by Bogen and Woodward (1988) has generated several discussions. Some of them are related to the idea that, in order to recognize patterns between data, theoretical assumptions were required (McAllister 1997; Harris 2002; Schindler 2007). This statement seems consistent with the new data visualization approaches. Furthermore, such approaches cast even more light on this point. Let us consider again Figure 2.

To begin with, as Figure 2 shows, before applying visualization or algorithms to data, the same data can be "transformed". Such transformation generally involves processes such as data cleaning or data grouping (Keim et al. 2010). They are processes aimed to "clean", select or modify a bit the data set, and can be performed for different reasons. First, there are cases in which heterogeneous data sources require to be integrated before the performance of automated or visual analysis, in these situations the transformation may enhance the possibility to find both expected and unexpected patterns between homogeneous data. Second, since in general visual analytics is used to gain knowledge from massive data sets, it is inevitable that every form of visualization leaves some patterns invisible. With the purpose of not losing interesting patterns, hence, this transformation is performed to analyse data according to the specific value of interest (selected by scientists) and to enhance the probability that all the most relevant aspects of the data are visualized. This big data transformation, the first passage from data to phenomena, is therefore performed by scientists with theoretical assumptions in mind: they have to decide the level of noise they are going to accept, as well as the part of the data set they prefer to examine.

Then, once data are ready, scientists have to decide whether to apply a visual or an automated analysis. Such approach is claimed to allow the observer to explore the data set by providing several possibilities of interaction with the visualization (Keim et al. 2010). In this way, not only unexpected patterns could be found accidentally, but also every observer has the chance to adapt the visualization according to what is more suitable for him (Sarlin 2016). However, those transformed data cannot be "just observed" by using any kind of visualization. Indeed, after the transformation, visualizations have to be prepared in order to be finally observed, and only a finite number of visualizations will be available at the end. This means that visualizations as the one in Figure 4 are the outcomes of scrupulously selected preparations that can be named "modeling" (Suppes 1962; Harris 2002; Boumans 2004).

A similar idea had already been expressed by Martin Shubik (1960) with a comparison between observations in economics and in biology. He cast light on the fact that when scientists make observations through a microscope they always need a "specimen" to be observed. Scientists' observations are consequently not "just" observations, but rather they are processes that at first involves the laborious setting up of specimens and their selection. According to Shubik the same remark can be applied to social sciences where, instead of material specimens, scientists create man-made models, representation of data. Moreover, once such specimens have been constructed, scientists have to decide which one to look at and in which way.

Within philosophy of science, the language of data models has been used to provide a framework consistent with the activities performed on data (Suppes 1962; Harris 2002). The paper by Suppes titled ''Models of Data'' (1962) described a hierarchy of models connecting data to theory, and provided a description of models of data. The main idea was that, since data must undergo some degree of processing before they can be analysed, what scientists use for their research are just models of data more or less close to the reality. For what concerns big data that are studied by means of visual analytics, for instance, the first data model would be made of data "transformed" by processes such as data cleaning. Then, further data models would be particular visualizations realized by scientists. Each option about the indicators to include, the number of countries, the colour or the type of graph would lead to a new data model: in other words, the same data can be processed to produce a variety of data models.

According to Harris (2002), data models are acceptable in relation to a theoretical goal, and they cannot be evaluated independently from that goal. Even if big data visualizations are now claimed to enhance the capacity to *explore* data, the same conclusion can be drawn about visual analytics' visualizations: at the end what observers can explore is a group of specific visualizations (data models) realized by scholars in order to find new patterns within certain areas.

Overall, it seems that the passage from big data to phenomena, if mediated by data visualizations developed through visual analytics, is characterized by the same characteristics of the past: both the selection of data and the decision concerning the way to visualize them require theoretical assumptions, and the only novelty seems that more data models are now offered to observers. Let us conclude with a remark: the generation of several data models is justified by considerations about the importance of combining standard algorithmic means with the potential capacity of experts to discern in data visualizations what is relevant (Keim et al. 2010). Similar considerations about the relevance of expert judgment can be found also in the past (Gaston and Dalison 2007; Boumans 2016), nevertheless given the automated form of data generation involved in the path leading to the recognition of facts, the combination of these forms of analysis with analytical tools is highly desirable nowadays.

Not only, in fact, expert judgments may lead to discovering patterns if not undetectable by means of analytical tools, experts could also help to recognize potential errors within data analyses. Just to give an example, experts could recognize what is called the Simpson Paradox inside a data set (Spirtes, Glymour, and Scheines 2000). The Simpson's Paradox states that a trend may appear in different sub-populations of data but it could disappear or even reverse when these sub-populations are combined and the attention is on the population as a whole. For instance, in 1973, the University of California-Berkeley was accused of sex discrimination. The statistics looked pretty incriminating: the graduate schools had just accepted 44% of male applicants but only 35% of female applicants. However, by examining such numbers, scientists cast light on an important insight: on the one hand men applied more often to science departments, one the other hand women were more inclined towards humanities. An important element differentiated such departments: science departments required special technical skills but a great percentage of qualified applicants was accepted. On the contrary, humanities departments only required a standard curriculum but they had fewer slots (Bickel et al. 1975). In other words, if the attention had been on the singular department instead of on the total graduate schools, no phenomenon of sex discrimination would have emerged. Analytical tools are not capable of recognizing the presence of such paradox, therefore data visualizations and the possibility to interact with them, could be extremely useful in similar cases.

Concluding remarks

The era of big data, it is often claimed, has arrived. There is little doubt, economists add, that over the next years, big data will change the landscape of economic research. The decision to write a thesis on big data and the economic domain has emerged because of the predominance of these claims. While big data have brought some changes, in fact, it appears that many other dimensions of economics have not been modified by such data deluge.

The thesis has started with a distinction between two categories of economic data: while some data are data on things such as the price of a good or the outcome of a transaction, other data are "data" on facts about phenomena like the unemployment rate or the annual inflation of a country. From the beginning, hence, the focus has been on data. This is one of the reasons why, considering the debate on the data/phenomena distinction proposed by Bogen and Woodward (1988), I have claimed that such concepts should not be replaced by statistical notions regarding samples of data and population models: these categories of data should not be confused. The other reasons I have offered deal with the production of data and the misleading idea that economic can use "raw" data. My claim, hence, has been that if the attention was on samples, the man-made nature of data, as well as the important difference between data of the first and of the second category, would remain in the shadows.

Chapter 2 has provided another justification of the importance of the data/phenomena distinction: big data can be understood only by confronting them with traditional data. After an analysis of the most influent proposals concerning the novelty of big data, I have suggested that the marker for big data is not a quality such as the velocity of the volume, but the new method involved in the creation of those data: big data are generated by means of automated procedures. Such consideration can be used on the one hand to provide an abstract description of big data, on the other hand as a normative criterion to distinguish between data and big data and to recognize the new challenges related to them. Moreover, such marker can be employed to clarify how the emergence of big data affects the macroeconomic domain. My conclusion is that big data, that are always data of the first category and never "data" about facts, can affect macroeconomics and macroeconomic "data" such as those contained in the data cube by Sarlin (2016) only indirectly.

Chapter 3 has concluded the examination of big data in macroeconomics by analysing whether something has changed in the path from big data to macroeconomic facts. After the description of visual analytics, the conclusion is that the passage from data to phenomena still requires human (theory-laden) decisions. Therefore, despite the new technologies involved, it seems that at least

within macroeconomics, the only novelty brought by big data concerns the way in which such data are produced.

The investigation performed in this thesis has been twofold. On the one hand, my objective has been to contribute both to the debates on the data/phenomena distinction and to the portrait of the new category of data known with the name of big data. On the other hand, my purpose has been also to show that, despite big data bring important novelty to the scientific domain, the most important aspects of the economic analyses are still the same.

Chapter 2 and Chapter 3 have dispelled two myths about big data and economics. First, big data cannot *directly* affect economic research. I have claimed many times that big data cannot be "data" about facts: as a consequence, they are in general not directly employed by macroeconomics. In macroeconomics, the only way in which they can be engaged is by looking for facts about phenomena and by measuring them. Second, the new methods through which big data are analysed do not involve big novelties. Visual analytics is considered an innovative approach to big data (Sarlin 2016), however both the consideration of the importance of data observations and the possibility to interact with such observations do not deal with great transformations.

The conclusion of my thesis is therefore less enthusiastic than the great part of the papers on this topic. For sure, big data are likely to enhance our understanding of the world and the possibility to discover facts about economic phenomena. This idea would be hardly deniable. Nevertheless, together with perils, big data bring new risks: economists should not forget that data are always, potentially, misleading. The "givenness" of big data implies that data generation should be investigated in order to understand whether something is missing or whether biases exist. In addition, the trust in new algorithms and analytics should always be weighted: scholars' assumptions are still required in many passages of the reasoning that link data to phenomena.

Literature

Askitas, N. and Zimmermann, K. F. (2015). The internet as a data source for advancement in social sciences. *International Journal of Manpower*, *36*(1), 2-12.

Backhouse, R. (1995). *Interpreting macroeconomics: explorations in the history of macroeconomic thought*. Routledge, London; New York.

Blumenstock, J. and Donaldson, D., (2013). How Do Labor Markets Equilibrate? Using Mobile Phone Records to Estimate the Effect of Local Labor Demand Shocks on Internal Migration and Local Wages. *Proposal Summary for Application C2-RA4-205*.

Bogen, J. and Woodward, J. (1988). Saving the Phenomena. *The Philosophical Review* 97 (3).

Bogen, J. (2010). Theory and observation in science. In Zalta, E. (2010). *Stanford encyclopedia of philosophy*.

Boumans, M. (2004). Models in Economics. In Davis, J. B. and Marciano A. *The Elgar Companion to Economics and Philosophy*. Edward Elgar Pub.

Boumans, M. (2005) *How Economists Model the World Into Numbers.* Routledge, London and New York.

Boumans, M. (2012). Visualisations for understanding complex economic systems. In Bissell, C. and Dillon, C. *Ways of thinking, ways of seeing: Mathematical and other modelling in engineering and technology*. Springer, Berlin.

Boumans, M. (2016). Graph-Based Inductive Reasoning. Forthcoming.

Boyd, D. and Crawford, K. (2012) Critical questions for big data. *Information, Communication and Society* 15(5).

Card, S., Mackinlay, J. and Schneidermann, B. (1999). *Readings in Information Visualization, Using Vision to Think*. Academic Press Inc., San Diego.

Coyle, D. (2014) *GDP: A Brief but Affectionate History*. Princeton University Press, Princeton.

D'Amuri, F. (2009). *Predicting unemployment in short samples with internet job search query data*. University Library of Munich, Germany.

Daas, P. J. H. and Puts, M. J. H., Social media sentiment and consumer confidence, *ECB Statistics Paper Series*, No 5, 2014.

Dasgupta, A. K. (1989). *Epochs of Economic Theory*. Basil Blackwell, Oxford and New York.

Daston, L. and Galison, P. (2007). *Objectivity*. Brooklyn, NY: Zone Books

Diebold, F. X. (2003). Big Data Dynamic factor models for macroeconomic measurement and forecasting. Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society, p. 115-122.

Diebold, F. X. (2012). On the Origin(s) and Development of the Term Big Data. PIER Working Paper.

Dodge, M. and Kitchin, R. (2005) Codes of life: Identification codes and the machine-readable world. *Environment and Planning D: Society and Space* 23(6).

Eaton, C. et al. (2012). *Understanding Big Data*. McGraw-Hill, New York.

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

Einav, G. (2015) *The New World of Transitioned Media. Digital Realignment and Industry Transformation*. Springer. London.

Ettredge, M., Gerdes, J. and Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, *48*(11), 87-92.

Fondeur, Y. and Karamé, F. (2013). Can Google data help predict French youth unemployment?. *Economic Modelling*, 30, 117-125.

Ginsberg J, et al. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012-4.

Glymour, B. (2000). Data and phenomena: A distinction reconsidered. *Erkenntnis*, 52, 29–37.

Griliches Z. (1986). Economic data issues. In Griliches Z. and Inriligator M. (1986) *Handbook of econometrics*. Elsevier Science, Amsterdam.

Hacking, I. (1992). The Self-Vindication of the Laboratory Sciences. In Pickering, Andrew. 1992. *Science as Practice and Culture.* The University of Chicago Press, Chicago.

Hanson, N. R. (1958). *Patterns of discovery: an inquiry into the conceptual foundations of science*. Cambridge University Press, Cambridge.

Harris, J. and Todaro, M., (1970). Migration, Unemployment and Development: A Two Sector Analysis. *American Economic Review*,60 (1).

Keim, D., Kohlhammer, J., Ellis, G. and Mannsmann, F. (2010). Mastering the Information Age. Solving Problems with Visual Analytics. *Eurographics Association*.

Khan, M. A., Uddin M. F. and Gupta N. (2014). Seven V's of Big Data. Understanding Big Data to extract Value. *Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education*.

Kitchin, R. (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage, London.

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

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: The University of Chicago Press.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *Technical report, META Group.*

Lazer, D. et al. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343: 1203-1205.

Leonelli, S. (2014). Data interpretation in the digital age. *Perspectives on Science*,22 (3).

Leonelli, S. (2015). What Counts as Scientific Data? A Relational Framework. *Philosophy of Science* ,82 (5).

Marr, E. (2014) *Big Data: Using Smart Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. John Wiley & Sons, London.

Marz, N. and Warren, J. (2012) *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*, MEAP edition, Westhampton, NJ.

Mayer-Schonberger, V. and Cukier, K. (2013) *Big Data: A Revolution that will Change How We Live, Work and Think*. John Murray, London.

McAllister, J. (1997). Phenomena and Patterns in Data Sets. *Erkenntnis* 47 (2).

McNulty, E. (2014) Understanding Big Data: The seven V's. 22 May. Available at: http://dataconomy.com/seven-vs-big-data/

Morgan, M.S. (1997). *Searching for causal relations in economic statistics. Reflections from history*. In McKim V. R. and Turner S.P. (1997) *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences*. University of Notre Dame Press, Notre Dame.

Morgan, M.S. (2004). Simulation: The birth of a technology to create "evidence" in economics. *Revue d'Histoire des Sciences*, 57 (2).

Morgan, M. S. (2012) *The World in the Model: How Economists Work and Think*. Cambridge University Press, Cambridge; New York.

Morgenstern, O. (1954). *Economic activity analysis*. Wiley, New York.

Persons, W. M. (1910). The correlation of economic statistics. *Publications of the American Statistical Association*,12 (92).

Persons, W.M. (1919). Indices of business conditions. I. A study of statistical method. *The Review of Economic Statistics*, 1 (1).

Pesaran, M. H. and Smith, R. (1992). The Interaction Between Theory and Observation in Economics. *Cambridge Working Papers in Economics*, University of Cambridge, Cambridge.

Pigou, A. C. (1913). *Unemployment*. Williams and Norgate, London.

Popper, K. (1983). *Realism and the Aim of Science. Postscript to the Logic of Scientific* Discovery. Rowman and Littlefield, Totowa.

Reiss, J. (2013). *Philosophy of Economics. A Contemporary Introduction.* Routledge, New York.

Rodenburg, P. (2008), The Measurement of Unemployment in Dutch Official Statistics; The Operationalizing of an Elusive Concept. In Klep P., van Maarseveen J. and Stamhuis I. (2008) *The Statistical Mind. The Netherlands, 1850-1940*.

Sarlin, P. (2016). Macroprudential oversight, risk communication and visualization. *Journal of Financial Stability*.

Schindler, S. (2007). Rehabilitating theory: Refusal of the 'bottom-up' construction of scientific phenomena. *Studies in the History and Philosophy of Science*, 38, 160–184.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. *Proceedings of the IEEE Symposium on Visual Languages*, Boulder, Colorado.

Shubik, M. (1960). Simulation of the industry and the firm. *The American Economic Review*, 50 (5).

Soman L. et al. (2006) *Insight Into Data Mining: Theory and Practice*. PHI Learning Pvt.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. MIT Press, Cambridge MA.

Thomas, J. and Cook, K. (2005). *Illuminating the Path: Research and Development Agenda for Visual Analytics*. IEEE Press, New York.

Van Kleeck, M. (1931). The Federal Census of Unemployment. *Journal of the American Statistical Association* 26 (173).

Weintraub, R. (1991). *Stabilizing Dynamics*. *Historical Perspectives on Modern Economics*. Cambridge University Press, Cambridge.

Woodward, J. (1989). The causal mechanical model of explanation. *Minnesota Studies in the Philosophy of Science* 13.

Woodward, J. (2011). Data and phenomena: a restatement and defense. *Synthese*, 182:165–179.