Technological Forecasting & Social Change 130 (2018) 99–113

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Big Data sources and methods for social and economic analyses

Desamparados Blazquez, Josep Domenech*

Department of Economics and Social Sciences, Universitat Politècnica de València, Camí de Vera s/n., Valencia 46022, Spain

ARTICLE INFO

Keywords:
Big Data architecture
Forecasting
Nowcasting
Data lifecycle
Socio-economic data
Non-traditional data sources
Non-traditional analysis methods

ABSTRACT

The Data Big Bang that the development of the ICTs has raised is providing us with a stream of fresh and digitized data related to how people, companies and other organizations interact. To turn these data into knowledge about the underlying behavior of the social and economic agents, organizations and researchers must deal with such amount of unstructured and heterogeneous data. Succeeding in this task requires to carefully plan and organize the whole process of data analysis taking into account the particularities of the social and economic analyses, which include the wide variety of heterogeneous sources of information and a strict governance policy. Grounded on the data lifecycle approach, this paper develops a Big Data architecture that properly integrates most of the non-traditional information sources and data analysis methods in order to provide a specifically designed system for forecasting social and economic behaviors, trends and changes.

1. Introduction

What comes to your mind when talking about “The Digital Era”? For sure, concepts as the “Internet”, “Smartphones” or “Smart sensors” arise. These technologies are progressively being used in most of the everyday activities of companies and individuals. For instance, many companies conduct marketing campaigns through social networks, sell their products online, monitor the routes followed by sales representatives with smartphones or register the performance of machinery with specific sensors. At the other side, individuals make use of computers, smartphones and tablets in order to buy products online, share their opinions, chat with friends or check the way to some place. Moreover, citizens' movements and activities are daily registered by sensors placed in any part of cities or roads and in public places such as supermarkets.

Therefore, all of these technologies are generating tons of digitized and fresh data about people and firms' activities that properly analyzed, could help reveal trends and monitor economic, industrial and social behaviors or magnitudes. These data are not only updated, but also massive, given that daily data generation has been recently estimated in 2.5 Exabytes (IBM, 2016). For this reason, they are commonly referred to as “Big Data”, concept which first appeared in the late 90s (Cox and Ellsworth, 1997) and was defined in the early 2000s in terms of the 3Vs model (Laney, 2001), which refers to: Volume (size of data), Velocity (speed of data transfers), and Variety (different types of data, ranging from video to data logs for instance, and with different structures). This model evolved to adapt to the changing digital reality, so that it was extended to 4Vs, adding the “Value” dimension (process to extract valuable information from data, known as Big Data Analytics). Currently, the “Big Data” concept is starting to be defined in terms of the 5Vs model (Bello-Orgaz et al., 2016), which added the “Veracity” dimension (related to proper data governance and privacy concerns).

This new data paradigm is called to transform the landscape for socio-economic policy and research (Einav and Levin, 2014; Varian, 2014) as well as for business management and decision-making. Thus, identifying which data sources are available, what type of data they provide, and how to treat these data is basic to generate as much value as possible for the company or organization. In this context, a Big Data architecture adapted to the specific domain and purpose of the organization contributes to systematize the process of generating value. This architecture should be capable of managing the complete data lifecycle in the organization, including data ingestion, analysis and storage, among others.

Furthermore, the design of a Big Data architecture should consider the numerous challenges that this paradigm implies. These include: scalability, data availability, data integrity, data transformation, data quality, data provenance (related to generation of right metadata that identify the origin of data as well as the processes applied to them during the data lifecycle, to assure traceability), management of huge volumes of information, data heterogeneity (structured and unstructured, with different time frequencies), integration of data from different sources, data matching, bias, availability of tools for properly analyzing such kind of data, processing complexity, privacy and legal issues, and data governance (Fan et al., 2014; Jagadish et al., 2014;

* Corresponding author.
E-mail addresses: mdeblzso@upvnet.upv.es (D. Blazquez), jdomenech@upvnet.upv.es (J. Domenech).

http://dx.doi.org/10.1016/j.techfore.2017.07.027
Received 1 March 2017; Received in revised form 7 July 2017; Accepted 25 July 2017
Available online 08 September 2017

Hashem et al., 2015).

The Big Data paradigm also offers many advantages and benefits for the companies, governments, and the society. Jin et al. (2015) highlight its potential contribution to national and industrial development, as it enforces to change and upgrade research methods, promotes and makes it easy to conduct interdisciplinary research, helps to nowcast the present and to forecast the future more precisely. In this vein, first Big Data architectures designed for specific fields are being proposed in order to surpass the previously mentioned challenges and make the most of the data available with the aim of nowcasting and forecasting variables of interest.

However, no specific architecture for social and economic forecasting has been proposed yet. This emerges as a necessity, in the one hand, because of the particular nature of socio-economic data, which have important components of uncertainty and human behavior that are particularly complex to model; and, in the other hand, because of the great benefits that can be derived from the use of Big Data to forecast economic and social changes. For instance, Big Data approaches have been proved to improve predictions of economic indicators such as the unemployment level (Vicente et al., 2015), help managers detect market trends so that they can anticipate opportunities, and also help policy-makers monitor faster and more precisely the effects of a wide range of policies and public grants (Blazquez and Domenech, 2017).

In this context, this paper aims to i) establish a framework about the new and potentially useful available sources of socio-economic data and new methods devoted to deal with these data, ii) propose a new data lifecycle model that encompasses all the processes related to working with Big Data, and iii) propose an architecture for a Big Data system able to integrate, process and analyze data from different sources with the objective to forecast economic and social changes.

The remainder of the paper is organized as follows: Section 2 reviews the Big Data architectures proposed in the literature; Section 3 compiles the new socio-economic data sources emerged in the Digital Era and proposes a classification of them; Section 4 reviews the new methods and analytics designed to deal with Big Data and establishes a taxonomy of these methods; Section 5 depicts the data lifecycle on which the proposed Big Data architecture is based; Section 6 proposes a Big Data architecture for nowcasting social and economic variables, explaining its different modules; finally, Section 7 draws some concluding remarks.

2. Related work

Since the advent of the concept of “Big Data” two decades ago, some architectures to manage and analyze such data in different fields have been proposed, having their technical roots in distributed computing paradigms such as grid computing (Berman et al., 2003). However, the current data explosion also referred to as “Data Big Bang” (Pesenson et al., 2010) in which there is a daily generation of vast quantities of data from a variety of formats and sources, is revealing the fullest meaning of “Big Data”.

The particular properties and challenges that the current Big Data context opens require specific architectures for information systems particularly designed to retrieve, process, analyze and store such volume and variety of data. Therefore, we are living the constant births of new technologies conceived to be useful in this context such as, to mention some, cloud and exascale computing (Bahrami and Singhal, 2014; Reed and Dongarra, 2015). Given this recent technological and data revolution, research in this topic is in its early stage (Chen et al., 2014). In this section, we review the novel and incipient research works that develop general frameworks and specific architectures for adopting the Big Data approach in different fields from the point of view of data analytics applications.

Pääkkönen and Pakkala (2015) proposed a reference architecture for Big Data systems based on the analysis of some implementation cases. This work describes a number of functionalities expected to be considered when designing a Big Data architecture for a specific knowledge field, business or industrial process. These include: Data sources, data extraction, data loading and preprocessing, data processing, data analysis, data transformation, interfacing and visualization, data storage and model specification. Besides that, Assunção et al. (2015) reflected on some components that should be present in any Big Data architecture by depicting the four most common phases within a Big Data analytics workflow: Data sources, data management (including tasks such as preprocessing and filtering), modelling, and result analysis and visualization. This scheme was put in relation to cloud computing, whose potential and benefits for storing huge amounts of data and performing powerful calculus are positioning it as a desirable technology to be included in the design of a Big Data architecture. Concretely, the role of cloud computing as part of a Big Data system has been explored by Hashem et al. (2015).

About architectures for specific domains, Zhang et al. (2017) proposed a Big Data analytics architecture with the aim of exploiting industrial data to achieve cleaner production processes and optimize the product lifecycle management. This architecture works in four main stages: in stage 1, services of product lifecycle management, such as design improvement, are applied; in stage 2, the architecture acquires and integrates Big Data from different industrial sources, such as sensors; in stage 3, Big Data is processed and stored depending on their structure; finally, in stage 4, Big Data mining and knowledge discovery is conducted by means of four layers: the data layer (mixing data), the method layer (data extraction), the result layer (data mining) and the application layer (meeting the demands of the enterprise). Results from last stage fill the ERP systems and are used along with decision support systems to improve product-related services and give feedback in all product lifecycle stages.

In the domain of healthcare, a complete and specific Big Data analytics architecture was developed by Wang et al. (2016a). This architecture was based on the experiences about best practices in implementing Big Data systems in the industry, and was composed of five major layers: first, the data layer, which includes the data sources to be used for supporting operations and problem solving; second, the data aggregation layer, which is in charge of acquiring, transforming and storing data; third, the analytics layer, which is in charge of processing and analyzing data; fourth, the information exploration layer, which works by generating outputs for clinical decision support, such as real-time monitoring of potential medical risks; last, the data governance layer, which is in charge of managing business data throughout its entire lifecycle by applying the proper standards and policies of security and privacy. This layer is particularly necessary in this case given the sensibility of clinical data.

The review of these architectures evidenced some common modules or functionalities. After homogenizing the different names for modules very similar responsibilities, and considering their sequence in the process, they can be summarized as follows: first, a data module, which includes different sources of data with different formats; second, a data preprocessing module, which includes data extraction, integration and transformation; third, a data analytics module, which includes modelling and analysis techniques for knowledge discovery; and fourth, a results and visualization module, which includes tools for representing the results in a way useful for the firm or organization.

However, there are other functionalities whose location within the Big Data architecture is not homogeneous across the different proposals. For instance, the data storage responsibilities, which are basic for enabling data reuse and bringing access to previous results, have been included in a variety of places, ranging from being included in the data module (Assunção et al., 2015) or the preprocessing module (Wang et al., 2016a; Zhang et al., 2017), to being a macro-functionality present in each module of the architecture (Pääkkönen and Pakkala, 2015). The last approach is better reflecting the nature and complexity of Big Data analysis, given that not only the original data requires storage, but also

the integrated data, processed data, and the results derived from data analytics.

Other functionalities whose consideration in the literature has been divergent are those related to data governance, which is concerned to preserve privacy, security and assure the accomplishment of data-related regulations. Despite its importance, data governance was only considered by Wang et al. (2016a). As long as the 5Vs model expands, data governance is expected to gain relevance and become a requirement in the design of any Big Data architecture.

For the case of Big Data for social or economic domains, no specific architecture has been proposed yet in the literature. Given their particular characteristics and increasing potential for detecting and monitoring behaviors and trends, which is basic to anticipate events, design better action plans and make more informed decisions, an architecture specifically devoted to treat these data emerges as necessary. Thus, this work proposes a Big Data architecture designed for nowcasting and forecasting social and economic changes. This proposal aims to help business implement the most appropriate architecture for their decision making needs, make the most of the data available and assure that it is treated according to the ethic and legal standards.

3. Non-traditional sources of social and economic data

The digital footprint left by individuals has caused an exponential growth of the data sources available for social and economic analyses, which broadens the possibilities for conducting socio-economic studies beyond traditional data sources, namely surveys and official records. Although the reasons why these new data are generated are numerous, the way they are generated has important ethical and legal implications. For instance, personal data in a purchase order cannot be used for the same purposes as the data from a public profile in Twitter. To some extent, the usage of the data is limited by how they are generated. This fact motivated us to review and classify the newborn non-traditional sources of social and economic data according to the purpose of the user generating the data, as Fig. 1 shows.

The first level in the taxonomy includes five categories: i) purpose of searching for information; ii) purpose of conducting a transaction, which could be of a financial or non-financial nature; iii) purpose of disseminating information; iv) purpose of doing a social interaction; and v) not a deliberate purpose. The first four categories correspond to an active generation of data, while the last correspond to an inactive generation: that is, data is not intentionally generated as a result of a particular purpose, but just derived from the own use of any device (PC, smartphone, tablet…) with any of the purposes explained above. Data that fall in this category have been divided in three types: usage data, location data and personal data. A brief description of each purpose from which data is generated and examples of sources involved in each data generation process is shown in Table 1.

The majority of non-traditional sources of social and economic data mentioned above needs the Internet for working. Indeed, the increasing penetration and importance of the Internet in almost every social and economic activity has positioned it as a basic means for the generation of such kind of data.

3.1. The Internet as basic means for generating socio-economic data

The “Data Big Bang” originated in the Internet, which unstoppable expands, is transforming the way of interacting in the economic and social framework. Myriad individuals, companies and public organisms search, post and generate tons of information daily through the Internet. These online activities leave behind a digital footprint that can be tracked and, if treated with the proper Big Data architecture, could help to describe their behavior, decisions and intentions, and thus, to monitor key economic and social changes and trends. Indeed, recent research highlighted the increasing role of the Internet as a provider of data for explaining, modelling, nowcasting and forecasting social behaviors (Askitas and Zimmermann, 2015).

3.1.1. Google Trends: the power of search engines

Google Trends (GT) is an Internet-based facility, released on May 2006, which provides up-to-date reports on the volume of search queries on a specific keyword or text, with historic searches available since January 2004. It captures how the demand of information under certain topics varies over time, providing useful data to detect emerging trends and underlying interests and concerns of society. The use of GT data to nowcast social and economic (particularly macroeconomic) variables was introduced by Choi and Varian (2009a,b), who showed that some search categories in the Google search engine helped to predict car and home sales, incoming tourists or unemployment claims. Afterwards, various studies in different countries focused on improving unemployment-related variables' forecasts by using GT data, obtaining successful results (Askitas and Zimmermann, 2009; McLaren and Shanbhogue, 2011; Fondeur and Karamé, 2013; Vicente et al., 2015).

The aggregate consumer behavior in different sectors has also been successfully predicted with GT data. For instance, using GT data as predictors has been proved to improve forecasts of tourist inflows (Artola et al., 2015; Bangwayo-Skeete and Skeete, 2015), of trading decisions and transaction volumes on the stock market (Preis et al., 2013; Moat et al., 2014), of private purchases of different goods and services (Vosen and Schmidt, 2011) or of cinema admissions (Hand and Judge, 2012). Recently, GT data have proven to be useful for forecasting political inquiries' results (Mavragani and Tsagarakis, 2016). However, elections results and topics with such components of opinion and ideology have been particularly studied through data from sites focused on social interaction, as are Social Networking Sites (SNS) such as Facebook and Twitter and opinion platforms such as Ciao.

3.1.2. Social Networking Sites and blogs

SNS are online places specifically addressed to encourage users express their feelings and opinions about any kind of topic. Therefore, the information they contain is to some extent a reflection of what happens in society. Indeed, the term “Social Big Data” is becoming popular to refer to data generated by SNS and blogs (Bello-Orgaz et al., 2016). For that reason, more attention is being paid to SNS as sources of data potentially useful in forecasting social variables.

Among SNS, the microblogging service Twitter is one of the most popular, with 332 million users who are active monthly and send on average more than 500 million tweets per day. This huge amount of “user-generated” information, though implies some issues, weaknesses and challenges that require further research (Gayo-Avello, 2013; Schoen et al., 2013), could help to predict both present and future social and economic events, as verified in different works. For instance, tweets' contents have helped to describe political preferences and forecast elections results (Tumasjan et al., 2011; Kim and Park, 2012; Ceron et al., 2014), to predict stock market movements (Bollen et al., 2011), to forecast box office in the motion pictures industry (Kim et al., 2015; Gaikar et al., 2015) or to monitor the public opinion on new policies (Ceron and Negri, 2016).

Facebook, which is the third most visited site worldwide[1] with 1,650 million active users, doubtlessly also represents a source of powerful data for analyzing social and economic behaviors. However, given that its contents are more heterogeneous and user-adjustable, they are also more difficult to retrieve and analyze. Notwithstanding this, incipient studies have shown the ability of Facebook data to determine consuming profiles, which are useful for marketing purposes (Arrigo et al., 2016), and to predict election results and the population's political orientation (Cameron et al., 2016; David et al., 2016).

Other principal SNS are LinkedIn, Youtube, Instagram, Google+, Tumblr and Flickr (Bello-Orgaz et al., 2016). They are also rich sources

[1] alexa.com.

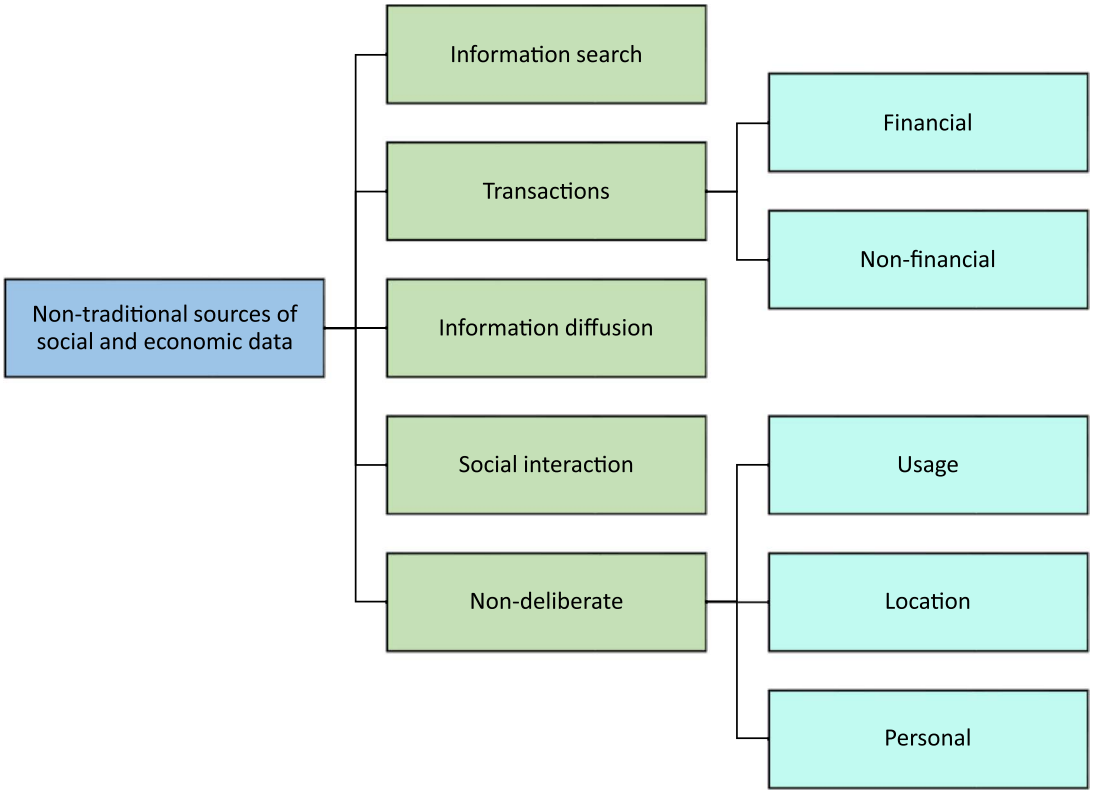

Fig. 1. Taxonomy of non-traditional sources of social and economic data.

Table 1
Classification of sources of socio-economic Big Data.

User's purpose	Description	Examples of sources
Information search	The user aims to find information about a topic of his interest. Data is actively generated	Search engines, Google Trends
Transactions	The user interacts with an individual and/or machine to achieve an agreement in which the user demands and obtains a product or service in exchange for a financial or non-financial compensation. Data is actively generated.	
• Financial transactions	Event in which the user makes a payment to obtain a product or service	E-banking, e-commerce, urban sensors (tolls, credit card readers, retail scanners, public transport card readers)
• Non-financial transactions	Event in which the user provides the counterpart with required information to obtain a product or service	E-government, e-recruiting
Information diffusion	The user aims to spread information or knowledge. This includes marketing purposes, in order to establish a public image of the user or the agent he represents. Data is actively generated.	Corporate websites, apps, Wiki pages
Social interaction	The user wants to share information, opinions and ideas with other users. Data is actively generated.	Social Networking Sites, opinion platforms, blogs
Non-deliberate	The user does not pursue to generate data with his/her action, but data are generated by the use of some means. Data is passively generated as a result of any other user action.	
• Usage	The simple fact of using any device generates data related to how, when and where an action has been done.	Web cookies, Internet Protocol, Sensors for self-tracking
• Location	The use of mobile phones generates data particularly related to the position of the user.	GPS, GSM, Call Detail Records, Bluetooth, WiFi Points
• Personal	Personal data (age, sex, etc.) is generated consciously (e.g. filling a form to complete a purchase) or unconsciously (e.g. data about the type of information we look for is used to infer our incomes) as a consequence of using any device or tool to achieve a purpose.	Forms, profiles, type of searches or purchases

of social and economic data, which could eventually be used to find changes in the unemployment patterns or detect what entertainment activities people prefer, among other topics (Russell, 2013). However, the diverse and complex formats of the information provided, along with the newness in some of these SNS, makes them remained almost unexplored. It should be noted that blogs are also important generators of "Social Big Data", though research in relating blogs' data to forecasting is also in its early stage. The pioneer work of Liu et al. (2007) examined the usefulness of opinions and sentiments extracted from blogs to forecast sales performance, while more recently Saleiro et al. (2015) combined data from news, blogs and SNS to track political opinion in real-time. Nevertheless, these sources are not without limitations. It is common that they are biased towards one segment of the population, e.g., young people, and English language, e.g., blogs in non-English language link more frequently English content than the other way round (Thelwall, 2007). Thus, some correcting measures should be considered before generalization (Gayo-Avello, 2012).

3.1.3. *Websites and apps: transactional, opinion platforms and information diffusion*

In the Digital Era, firms generally establish their official public image on the Internet by implementing corporate websites. Through

these sites, companies inform about their products, services, organizational structure and intentions, such as exporting and opening a branch office abroad. Corporate websites encompass all kind of websites implemented by firms in relation to their economic activity, ranging from websites used only to give information about the firm, to transactional websites devoted not only to provide information but also to offer online services (e-commerce, e-banking…), about which users are sometimes allowed to give their opinion in the website itself. That is, corporate websites may present three different functionalities: spreading information about firms (related to establishing a public image), conducting transactions (e-business processes) and facilitating opinion sharing (electronic word-of-mouth (eWOM) booster).

It is remarkable that websites have a complex structure which differ from one case to another, so that standardizing the retrieval and analysis of their information requires from a specific Big Data architecture. That difficulty has contributed to corporate websites being an almost unexplored source of data. However, their public, updated and "business generated" nature makes them potential sources of economic data. Moreover, business characteristics could emerge on the web and be monitored by massively analyzing corporate websites, as recent research shows.

Applying Big Data approaches (particularly web data mining and machine learning) to the "spreading information" functionality of corporate websites, firms' sales growth and business activities such as strategies of technology adoption, innovation and R & D have been successfully detected (Arora et al., 2016, 2013; Gök et al., 2015; Li et al., 2016). In addition, by using a specifically designed web data mining system for analyzing corporate websites (Domenech et al., 2012) the export orientation of firms has also been successfully detected (Blazquez and Domenech, 2017). In addition, there exist other type of websites created with the specific aim of spreading information, such as are Wiki pages, from which Wikipedia is the most important representative nowadays with more than 730 million unique visitors monthly (Wikimedia Foundation, 2017). Its penetration in the society along with its collaborative nature have positioned it as a potential source of social and behavioral data. Concretely, Wikipedia page views, edits and contents have already proven to be useful for socio-economic forecasting. Incipient research works have successfully used Wikipedia data to better forecast stock market movements (Moat et al., 2014) and tourism demand (Alis et al., 2015; Khadivi and Ramakrishnan, 2016). This kind of studies aim to create new indicators in advance or to complement those used in current official statistics.

The prominent role of the Internet in today's economy and society has promoted the emergence of e-business services, which firms can use to sell their products and do transactions in an online base with customers (E-commerce), recruit candidates (E-recruiting) or offer their services online (e.g. E-banking). E-business may even go a step further and represent not only a complementary tool for firms (e.g., selling locally and online), but a new type of business model characterized by operating just online. Many of these sites offer users the chance to post their opinions and do reviews on the product or service acquired, which may range from any manufacture to a hotel stay or an experience in a restaurant. This web feature is generally known as opinion platform (even a website can be designed just to act as opinion platform), which is used to bring together online communities of users, whose opinions are basic information for social science research.

One of the most important e-commerce and opinion platform worldwide is Amazon. It is one of the biggest online retailers, with more than 300 million active customers' accounts. This website provides customers' reviews and opinions on millions of products and services, being therefore a source of data potentially useful to detect consumer preferences or predict sales. For instance, the forecast of consumer product demands in Amazon.com has been significantly improved by using the textual contents of consumer reviews (Chong et al., 2015; Schneider and Gupta, 2016). Another noteworthy topic for managers is to detect the so-called "influencers" in consumer-opinion platforms, given that their comments may influence other consumers' purchase behavior and, thus, detecting and monitoring them is essential. For instance, Arenas-Márquez et al. (2014) successfully identified influencers in Ciao.com by retrieving and analyzing characteristics of the product reviews such as the rating received by other users.

Other sites that provide potentially useful data for detecting social and economic trends are, for instance, eBay.com, whose information has been helpful to explain price differences among remanufactured, used and new items (Frota Neto et al., 2016), TripAdvisor.com, which has been successfully used to detect tourist preferences thus helping hotel managers to adapt their offers (Li et al., 2015), and Monster.com, which organizes the available job offers and helps to track changes in job search (Edelman, 2012).

When using opinion platforms as sources for social and economic analyses, limitations related to the veracity of the contents must be considered. Sellers and marketers may have the temptation to generate fake consumer reviews to influence in the consumer decision (Malbon, 2013). In this context, some techniques for detecting such manipulations could be applied to alleviate this limitation (Hu et al., 2012).

Apps provide access to information and services that may or may not be offered by other means, such as websites. Since the use of apps is becoming widespread in the daily activities of individuals and organizations, they have become a source of data with great potential for forecasting social and economic topics. Although accessing data generated by them is currently a difficult task, some incipient research works are appearing. To date, apps data logs have been proved to be successful for forecasting users' intentions to use a specific app, automatically forecasting depression (Wang et al., 2016b; Suhara et al., 2017) or helping to detect mobility patterns as reviewed by Pan and Yang (2016).

3.2. Urban and mobile sensors

Ubiquitous computing is one of the technological areas that has experimented the greatest development in the context of the Digital Era. Its advances have resulted in the generation of wireless, inconspicuous and inexpensive sensors to gather information on our everyday life activities (Krishnan and Cook, 2014). Specifically, urban sensors and mobile embedded sensors are potential generators of social and economic data.

Among urban sensors, one of the most widespread and used worldwide is the credit card reader. Credit card transactions are recorded and provide data potentially useful for firms to detect and predict, for instance, personal bankruptcy (Xiong et al., 2013), fraudulent purchases in online stores (Van Vlasselaer et al., 2015) and default and repayment, which in the context of credit card companies is useful for defining marketing strategies (Einav and Levin, 2014).

Retail scanners are also very extended, and their function is to record the characteristics of customers' everyday purchases. These data has proven to be useful for forecasting consumer behaviors, sales and prices, as recent research shows. For instance, Dey et al. (2014) successfully used retail level scanner data to model market trends, prices and sales in the industry of catfish products, suggesting a particular competition strategy based on the results obtained. Another study, focused on explaining human behavior, employed weekly scanner data to detect consumer boycotts in response to an international conflict (Pandya and Venkatesan, 2016).

A pioneer study by Askitas and Zimmermann (2013) successfully used data from tolls to nowcast business cycles, creating a Toll Index that represents a technological, innovation-driven economic telemetry. Other sensor networks that provide useful data for forecasting a manifold of socio-economic variables are smart grid, WiFi access points and public transport card readers, among others (Kitchin, 2014; Chou and Ngo, 2016).

Some sensors embedded in mobile phones are also potential sources of social data: GSM, GPS, Bluetooth, accelerometer or sensors for

connecting to the telephonic network through Base Transceiver Stations (which produce the so-called “Call Detail Records”, with information regarding all call-related activities, such as sending SMS and phoning, conducted by mobile phone users in the network). These sensors generate data related to the user location that have been successfully used to study social behaviors, preferences and mobility patterns. Properly treated, these data can contribute to better understand in which way human mobility affects well-being and human behaviors at the micro level, and social organization and change at the macro level (Williams et al., 2015).

Concretely, data from such sensors have been useful for detecting places of interest, that is, places where people go and stay for a while (Montoliu et al., 2013), and for detecting personality traits, which companies may use to personalize their services (Chittaranjan et al., 2013). Moreover, Laurila et al. (2013) summarized different human behaviors analyzed to date with such mobile embedded sensors data, including: mobility patterns and their relation with the weather, the perceived level of safeness and intimacy of a given location, the relation among moves from individuals and from their friends and acquaintances, and the transition between spatial habitats. Other recent applications of mobile phones' data in relation to mobility are recreating and drawing maps of population distribution (Deville et al., 2014; Graells-Garrido et al., 2016) and detecting anomalous behavioral patterns associated to emergency (e.g. earthquakes) and non-emergency (e.g. holidays) events (Dobra et al., 2015).

4. Non-traditional methods for processing social and economic data

Data obtained from non-traditional socio-economic sources are generally large, heterogeneous and unstructured or semi-structured. These characteristics imply a number of challenges when it comes to retrieving, processing, analyzing and storing data. Accordingly, methods and techniques related to machine learning and Big Data are being developed. Many of such methods have been widely applied in other knowledge fields such as engineering, medicine and biostatistics. Despite their potential for treating socio-economic data, their application in this field is still at an early stage (Varian, 2014).

This section enumerates and describes the most relevant methods for treating socio-economic data from a Big Data approach, with the objective of providing a framework. The reviewed techniques are summarized and classified in a taxonomy illustrated in Fig. 2.

4.1. Methods for structuring data

Big Data sources can be classified as structured (tabular data), semi-structured (data with machine-readable tags that do not follow a strict standard) or unstructured (data that lacks from any scheme allowing machines to understand them, e.g. a video). Since analysis algorithms require some structure to interpret the data and given that about 95% of Big Data is unstructured (Gandomi and Haider, 2015), the process of structuring the information is basic. This includes transforming the data into an organized set, with clearly defined variables and the relations among them identified. Below, some of the most common methods for structuring data with applications to social and economic analyses are surveyed.

Almost any source of data, and particularly the Internet, is plenty of human generated text that requires proper retrieval and processing. To exploit the full potential of text in databases, specific techniques for processing natural language are required. Natural Language Processing (NLP) is a research area focused on exploring how computers can be used to understand and shape natural language text so that it can be useful for different applications (Chowdhury, 2005). NLP is in itself a computational method that comprehends a series of techniques that provide an easy interface for information retrieval systems and, at the same time, to structure texts in different ways so that the underlying information can be more easily extracted. Some interesting NLP techniques for social analysis are Sentiment Analysis (also referred to as Opinion Mining), Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), TF-IDF (Term Frequency - Inverse Document Frequency) and Word Embeddings. Liu (2012), Evangelopoulos et al. (2012), Blei et al. (2003), Moro et al. (2015), Armentano et al. (2014) and Rudolph et al. (2016), among others, provide some reference for these methods.

Linking records from the same user (or entity) across different data sources is also an important challenge for analyzing social and economic information. Data Matching (which is also commonly known as Record Linkage or Entity Resolution) is a computational process used to identify, match and merge records from several databases that correspond to the same entities. A special case of data matching is deduplication, which consists in the identification and matching of records about the same entities within just one database (this step is crucial in data cleaning). Matched data are becoming more important because they may contain information impossible to obtain by means of other sources or processes. This technique is a complex process encompassing five steps, from data cleaning and standardization to data quality and completeness measuring. For a detailed description, see the work by Vatsalan et al. (2013).

4.2. Methods for modelling data

Modelling data (and their relationships) is the main process in a Big Data analysis. This includes reducing the dimensionality of data sets, applying modelling techniques to data and obtaining outcomes. Depending on the type of data available and the objective of the analysis, two different paradigms for modelling data may be applied: Supervised Learning and Unsupervised Learning (Hastie et al., 2013).

On the one hand, Supervised Learning refers to problems in which each observation in a data set has inputs (also referred to as independent variables, features or predictors) and outputs (also referred to as targets, responses or dependent variables), and the main goal is to use inputs in order to infer the values of outputs. These problems can be further categorized as classification problems, in which outputs are expressed as categories, or as regression problems, in which outputs are expressed in a continuous space. On the other hand, Unsupervised Learning refers to problems in which each observation has some inputs but no outputs, and the main goal is to find the relationships or structure among inputs. These problems can be further categorized into clustering problems, in which the goal is to discover groupings in the data, and association problems, in which the objective is to find rules that describe the behavior of part of the data.

Depending on the learning paradigm, different machine learning techniques can be applied. For nowcasting and forecasting applications, supervised methods are generally employed. The most common supervised machine learning techniques successfully applied in other disciplines, such as medicine and engineering, and that are potentially useful for the social sciences, are enumerated below.

Linear and logistic regressions are two useful machine learning techniques widely applied by economists and social scientists. However, alternative methods to regressions have been developed and demonstrated to perform as well as or better when using big data sets (Varian, 2014). For instance, Decision Trees, which are a type of predictive models that can be used to represent both classifiers and regression models; Support Vector Network (Cortes and Vapnik, 1995), more commonly known as Support Vector Machine (SVM), which is a learning machine for two-group classification; Artificial Neural Networks (ANN), which are two-stage regression or classification models able to identify non-linear relations among a set of input variables, and generate forecasts about the variable under study by modelling and weighting those relations (Hastie et al., 2013); and Deep Learning methods, which develop a layered and hierarchical architecture where higher-level (more abstract) features are obtained by transforming

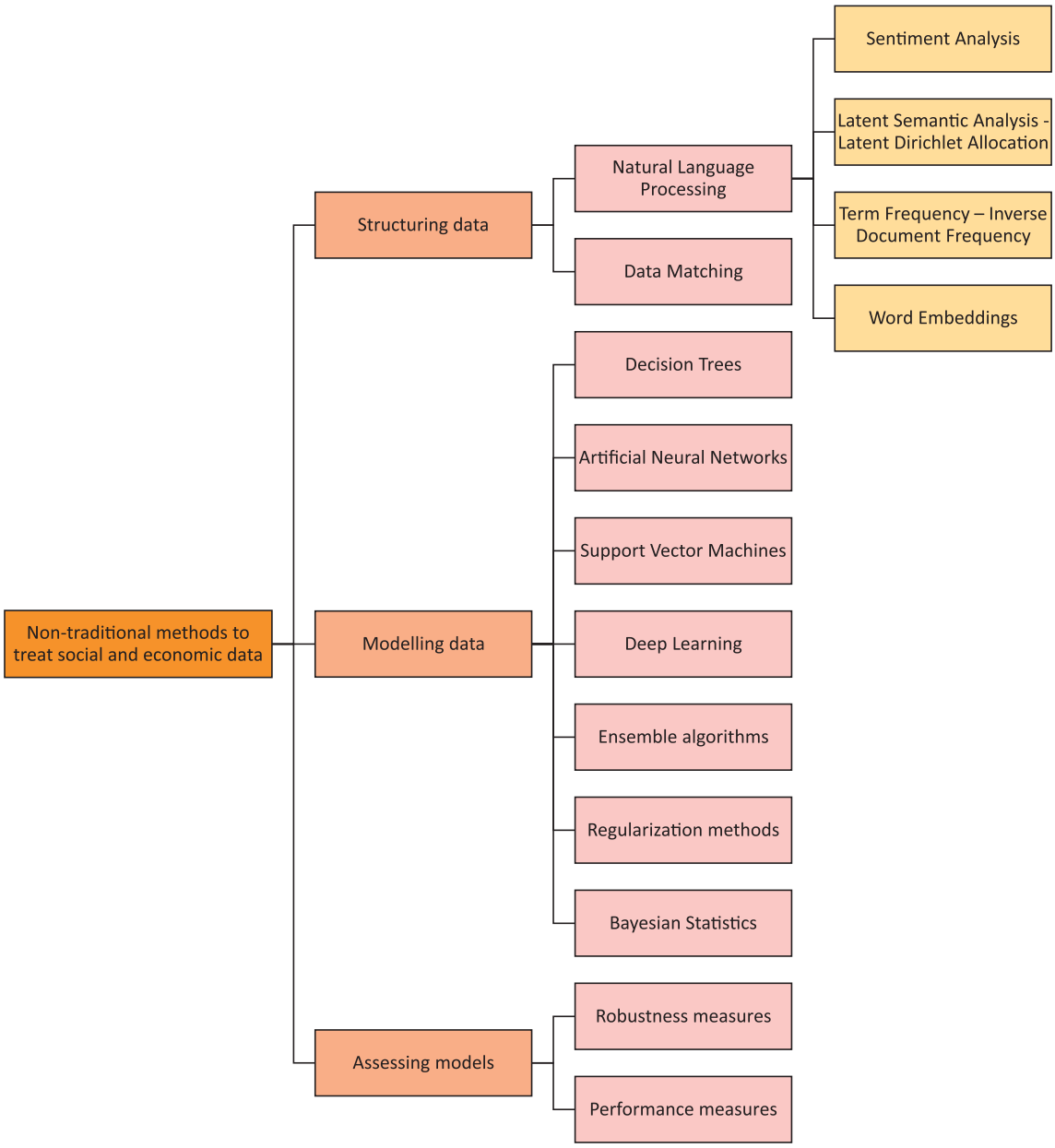

Fig. 2. Taxonomy of non-traditional methods to treat social and economic data.

lower-level (less abstract) features. For classification, higher-level features highlight aspects of the raw input that are relevant for discrimination. Deep Learning methods can deal with huge quantities of unstructured data, reason why they are positioning as a promising tool in Big Data analysis (LeCun et al., 2015; Najafabadi et al., 2015). ANN and Deep Learning are special cases, given that their learning algorithms can be either supervised or unsupervised.

In addition, there exist a group of techniques which are focused on improving the performance of the previously described ones, and that are starting to be known as “Ensemble algorithms”. Some of these algorithms work by adding randomness to data, which is a useful procedure to deal with overfitting. These techniques include the Bootstrap, Bagging, Boosting and Random Forests (Varian, 2014).

Regularization methods are another group of supervised learning techniques, whose objective is to obtain sparse solutions and that, due to the increased amount of information available, have been increasingly studied in recent years by the scientific community (Friedman et al., 2010). These methods can be applied to a number of supervised learning techniques, from regressions to support vector machines, and include, to mention some examples: the Least Absolute Shrinkage and Selection Operator (LASSO), which was one of the first regularization methods (Tibshirani, 1996); the regularization for support vector machines (Hastie et al., 2004); the Elastic Net, which is a mixture of the LASSO and Ridge Regression (Zou and Hastie, 2005); and a regularization scheme for neural networks, aimed at improving the classification margin (Ludwig et al., 2014).

Finally, Bayesian Statistics constitute an alternative approach to frequentist statistics (as are the methods describe above) in both terms of decision theory and inference. Though their potential in the social sciences and economy was pointed out almost 40 years ago (Harsanyi, 1978), the complex numerical integrations needed made them remain unused. However, the recent advances in computation methods have made it possible to easily apply Bayesian methods (Congdon, 2007).

To mention some, Bayesian Model Averaging (BMA) is a multi-modelling method that is starting to be applied to linear regression and generalized linear models (Ley and Steel, 2012). Naive Bayes, which is

a machine learning tool for classification whose popularity is starting to increase due to its simplicity for being implemented, being fast and computationally efficient, and obtaining high classification accuracy, especially for Big Data (Wu et al., 2015). Also, the Spike-and-Slab Regression, which is a variable selection method for linear regression models (Varian, 2014). Besides this, the Bayesian Structural Time Series (BSTS) technique is devoted to treating panel or longitudinal data, which are very common in the social sciences. This is a method for variable selection and time series forecasting and nowcasting, used as an alternative to traditional time series analysis methods such as Autoregressive (AR) and Moving Average (MA) models.

4.3. Methods for assessing models' performance and robustness

A basic objective in any data analysis focused on forecasting is to obtain a robust model with the best out-of-sample predictive precision possible. In this subsection, a brief review on techniques for improving the performance of forecasting and nowcasting models is provided.

Assessing the performance and robustness of predictive models is essential to determine their validity and applicability, and the quality of the predictions. In this case, performance refers to how well a model fits the data and how accurate it is, while robustness refers to how well a model works on alternate data, that is, on data which is different from that used to build the model. If a model has a good performance, then it is capable of detecting the characteristics in a data set and providing highly accurate predictions. Moreover, if it is robust, then the predictions obtained could generalize and so the model is valid and useful with new data. The goal in any Big Data analysis is to build models that simultaneously are robust and provide accurate outputs: this is the only path to use them as reliable tools for forecasting and nowcasting whose results can be used for decision-making.

To compare and select different kind of models depending on how well they fit to data and how complex they are, there exist uncountable classically applied tests such as Nagelkerke's R^2, Hosmer-Lemeshow, Mallows' Cp, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Deviance and Log-Likelihood, among others. Although these tests and indices provide useful information about model performance, they were not conceived for treating the huge amount of complex data with which we work nowadays. The particular characteristics and issues of Big Data (size, bias, imbalanced sets, complex relations…) make necessary to complement classical tests with more recently developed techniques that are capable to better deal with these issues (Varian, 2014).

First of all, to ensure that the predictions obtained are robust it is recommended to build the models by conducting a holdout process in which the initial sample is split into two subsets: the training set and the test set. The former is used to train the model, and generally includes about 75% to 90% of the initial observations, while the latter is used to evaluate its predictive performance and includes the remaining percentage of observations. Even if data is large enough, it may be divided in three sets: a train set (the largest), a validation set and a test set. This method ensures that the predictions obtained are robust, so that they can be generalized to an independent data set. Another approach with the same objective is K-Fold Cross-Validation. In this method, data are split into K parts of equal size and the model is fitted K times, where K-1 parts are used to train the model and the remaining is used to test its predictive performance. Finally, the K estimates of the prediction error are combined. In case each part includes just one observation, then the process is called Leave-one-out Cross Validation (Hastie et al., 2013). For big data sets, the first method is recommended.

In addition, for properly training classifiers, at least the train set should be balanced, because this way the model is built to successfully detect each of the categories equally. Otherwise the learning process could be endangered (Menardi and Torelli, 2014). A sample is balanced when each of the categories of the response variable is present in the same proportion. To balance an unbalanced data set, solutions such as oversampling, undersampling, synthetic sampling and kernel methods can be applied (He and Garcia, 2009). Unbalanced data sets are common in social sciences, so it is expected that the use of these procedures in socio-economic research will expand in the near future.

Moreover, think if what we are trying to predict is if someone is infected with a disease; then, the best situation would be to obtain a true negative (individual not infected). That is, not only false positives, but also false negatives, imply costs. It is important to assign a monetary value to these costs in order to influence the decision making of a model. This process is known as “Cost-sensitive analysis” (Sun et al., 2007). It makes use of the Cost Matrix, which reflects the costs and benefits associated to each of the four possible outcomes of a classifier. Providing this information to a classifier, it can be influenced to minimize the most costly errors or to maximize beneficial classifications, so that we obtain a “weighted accuracy”. Similarly, by using Loss Functions, classifiers are forced to give preference to predictors that help to predict true probabilities accurately (Witten et al., 2016).

To check the predictive accuracy of classifiers, methods such as the Lift analysis, Precision-Recall Curves, ROC Curves and the Confusion Matrix are pertinent, whose fundamentals and applications regarding the social sciences can be looked at (Fawcett, 2006; Witten et al., 2016). When the output variable is not categorical, but numerical, other measures are available, such as the Root Mean Squared Error (RMSE), the Percentage Error (PE), the Fractional Bias and the Index of Agreement (IA), whose popularity is starting to increase.

5. The data lifecycle

Digital data have many advantages, such as being easy to share, replicate and recombine, which make them potentially reusable. Business and researchers can take advantage of this to boost research in progress and leverage past investments, for instance. However, to exploit all the benefits of digital data, they must be properly collected, processed and preserved. Data loss or damage may imply economic costs as well as lost chances, reason why funder agents (public or private) are increasingly demanding institutions to document and run data-management plans taking into account the whole lifecycle of data (Lynch, 2008). For this reason, it is basic to define what phases and processes form this lifecycle in order to implement robust and flexible architectures to manage data in the context of the Digital Era.

The data lifecycle is the sequence of stages that data follow from the moment they enter a system to the moment they are erased from the system or stored (Simonet et al., 2015). Between the data entrance and exit or storage, data go through different stages, which may differ depending on the type of data and purpose to achieve as documented in the compilation of classic data lifecycles (Committee on Earth Observation Satellites — Working Group on Information Systems and Services, 2012). The Knowledge Discovery in Databases (KDD) process was the first proposal of a model to manage digital data (Fayyad et al., 1996). It refers to the complete (non-cyclical) process of extracting knowledge from data, and includes five main stages: data selection, data preprocessing, data transformation, data mining and data interpretation. As databases started to exponentially grow in size and complexity, the necessity of a wider scheme to appropriately manage these data was highlighted, especially by the industry. This derived into the development of the Cross-Industry Standard Process for Data Mining (CRISP-DM process), which is an expanded implementation of the KDD process that introduced the management of digital data as a cycle (Chapman et al., 2000). It comprises six stages: business understanding, data understanding, data preparation, modelling, evaluation and deployment. If both are compared, the first and last stages of CRISP-DM process are new with respect to the KDD process, while the “data understanding” stage of CRISP-DM is similar to the “data preprocessing ” and “data transformation” stages of KDD.

The next approach to data management within a digital environment that the scientific and industrial community focused on, and to

which most research efforts have been paid since these days, was called itself the “data lifecycle”. The Data Documentation Initiative Alliance (DDI Alliance) was one of the first voices to focus their efforts on this idea (DDI Alliance, 2008). It proposed a data lifecycle including the following five stages: first, discovery and planning; second, initial data collection; third, final data preparation and analysis; fourth, publication and sharing; and last, long-term management. This departing point considers from planning the project (what is being studied, what data are needed and how they are going to be treated, etc.) to determining how to store and preserve data in the long-term. With respect to KDD and CRISP-DM processes, this is a more extensive approach that includes important concepts within digital data such as sharing and long-term management.

Afterwards, Corti et al. (2014) described the phases and activities typically undertaken within the research data lifecycle. These phases, each of which included a number of specific activities, are the following: discovery and planning, data collection, data processing and analysis, publishing and sharing, long-term management and reusing data. This proposal extends the initial one by DDI Alliance to include an additional stage at the end of the cycle devoted to data reuse. A more exhaustive data lifecycle to date was proposed by Rüegg et al. (2014), who included up to eight stages: the first four stages (planning, data collection, data quality control, and analysis) correspond to managing data in a traditional project which is new (no previously results or data exist). If the project relies on existing data (referred to as “data reuse”), then it follows the third first stages and continues with additional data discovery, data integration, and finally, the analysis.

While Corti et al. (2014) consider data reuse as a step itself, Rüegg et al. (2014) reference data reuse as a type of project in which existing data are used, including some steps within this lifecycle. The context of economic and social analyses makes it more appropriate to consider data reuse as a step itself, given that as a project that started from scratch develops and data are obtained and exploited, these data may be reused many times in the same project with different purposes. That is, the view that a project is new or departs from data seems excessively static for economic and social nowcasting purposes. Additionally, to complete each of the data lifecycles, this work includes two more steps: data documentation and data archiving in a public repository, which we consider basic for preserving and publishing data.

The review of these works of reference allowed us to integrate and fully describe the different stages of a full data lifecycle in the context of economic and social analyses. Its aim is to standardize the concept of data lifecycle and serve as framework when it comes to designing a proper data management architecture in this context. Our proposal for a data lifecycle includes nine stages, as reflected in Fig. 3. These stages are described as follows:

1. Study and planning: This first stage consists in designing the research or business project to achieve the desired goals of funders or managers. Once each phase of the study is defined, it is necessary to plan what procedures to treat data (collected or generated throughout the research) will be applied. For instance, this includes planning what type of data are going to be collected, how and from which sources, which methods will be used for their processing and analysis, where will they be stored, and to find out what legal regulations and privacy issues affect the type of data that is going to be analyzed, in order to adapt the operating procedures.
2. Data collection: This stage consists in accessing the sources, which can be internal or external, and collecting initial or raw data. Depending on the field of knowledge and the data required for developing the project, activities such as phenomena observation, experimentation, recording, simulating, scraping and negotiating with third-party data providers will be part of this stage.
3. Data documentation and quality assurance: This stage consists in documenting the acquired data and checking their quality. First, the data acquisition process should be documented by associating data to metadata. The metadata include information related to the source of origin, data format, technical details on the retrieval process and accessing dates, among others, thus enabling their reuse and correct referencing. Second, data quality and validity should be assured. It is required to verify the trustworthiness of the data sources as well as of the own data, to control for any data inconsistencies, such as unexpected values and typing errors, and to clean and anonymize data if necessary.
4. Data integration: This stage consists in fusing data obtained from different data sources with a coherent and homogeneous structure, which helps to make data traceable and easier to access and manipulate in successive projects. This includes activities such as establishing relations among variables of different data sources, adapting units, translating, and creating a single database with all the acquired data. Data integration should also incorporate privacy constraints to avoid disclosing some private information in the integrated data. This is a major concern because rich integrated data may facilitate discovering some personal details otherwise anonymous.
5. Data preparation: This stage consists in transforming data so that they meet the format requirements of the analysis tools and techniques that are going to be applied. This includes activities such as transcribing, digitizing, interpolating, establishing a tabular format in the data set and deriving new data by operating with the existing data.
6. Data analysis: This stage consists in analyzing data, obtaining and interpreting results, and achieving conclusions. A huge range of statistical techniques and computational tools are called to be used in this stage. The final selection of the most appropriate techniques will depend on the type of data analyzed and research objectives. The interpretation of the results and conclusions achieved, as well as the results themselves, are basic inputs for the next stage.
7. Publishing and sharing: This stage consists in publishing results and conclusions derived from data analysis, or the generated data sets themselves. The outputs of this stage aim to facilitate the decision-making process of managers or policy-makers (when data is presented in reports, for instance), to spread knowledge (if a research article is published, for instance) and to feed automatic systems of companies with information of relevance to help the staff make decisions such as ordering supplies, among many others. Other related activities in this stage are establishing copyright of data and results, authoring publications, citing data sources, distributing data and controlling data access.
8. Data storage and maintenance: This stage consists in archiving and registering all the data gathered, processed and analyzed, for allowing long-term data preservation, curation and reuse. Actions to be done may include storing data in specific repositories or computational systems, migrating them to other platforms or mediums, regularly backing up the data, producing associated metadata, preserving the documentation generated during the whole process, controlling data security and privacy and erasing data if required by legal regulations, for instance.
9. Data reuse: This stage consists in reusing data that have been previously gathered, processed, analyzed and stored. This action can be originated in a variety of different purposes such as testing new hypotheses related to the same project for which data were collected, sharing or selling data to companies, conducting new projects for which existing data can be useful and using data with instructive purposes.

6. A Big Data architecture for nowcasting and forecasting social and economic changes

Our proposal of a Big Data architecture for nowcasting social and economic changes is presented in Fig. 4. Departing from the approach of the data lifecycle in the organization, it includes layers and modules

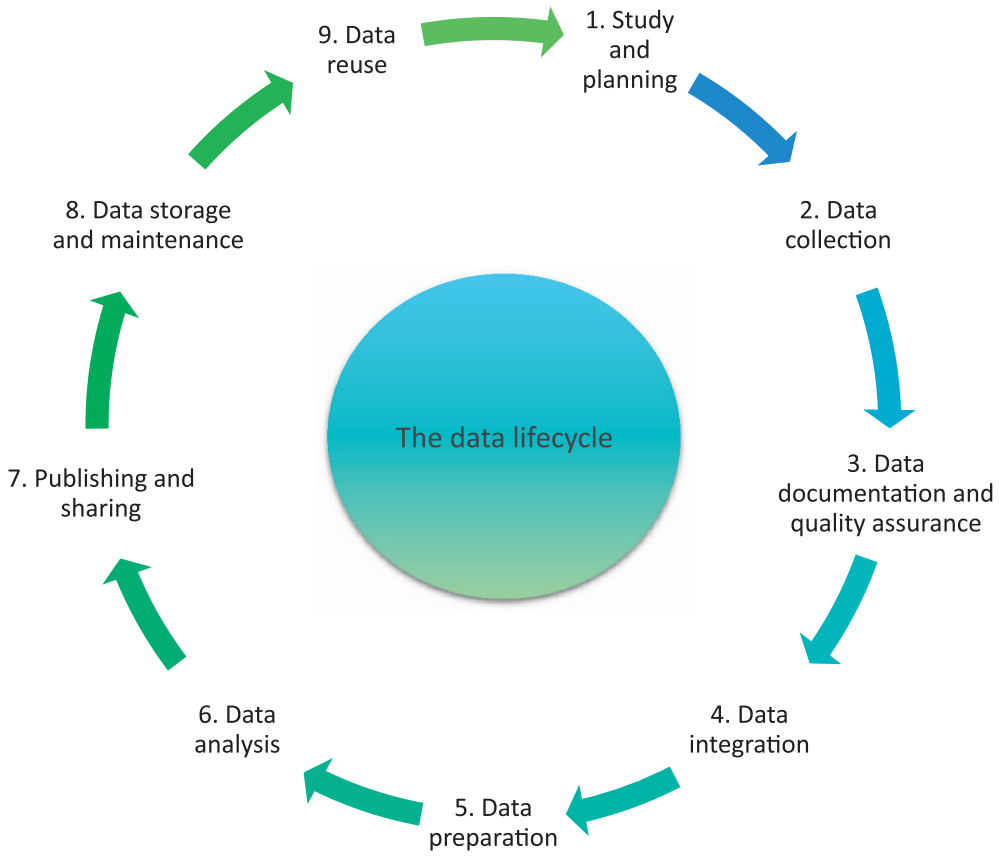

Fig. 3. The data lifecycle within a Big Data paradigm.

to manage the processing and integration of social and economic data, including the storage, processing policies and publication of results.

This architecture is organized in three layers. The data analysis layer contains the main processes of generating knowledge from the input data: from the ingestion of data from multiple sources to the publication of reports. Together with this layer, there are two other layers that work as support to the data analysis: The governance layer is in charge of applying policies and regulations to the whole data lifecycle, as well as managing the licenses related to the data sets. The persistence layer deals with the storage and management of data to make them available to the different modules in the data analysis layer.

6.1. Data analysis layer

The data analysis layer is the part of the architecture that implements the main processes required to generate knowledge, in form of reports or predictions, from the different data sources to which the organization has access. It is composed of six modules that work sequentially, from the data reception to the publishing of results.

6.1.1. Data receiving module

This module constitutes the data ingestion point in the architecture, so that data external to the system are made accessible to the other modules of the architecture. This way, external data are connected to the processing stream of the nowcasting system. It is composed of different elements, as many as different data sources are used as input.

Connected data sources can be classified into two main groups: First, those sources owned by the organization implementing the architecture. These sources may include relational databases managing the daily operation of the company, that is, business-specific information, such as sales, customers, purchases, website analytics, and so on. Own sources also involve data not directly generated by the business operation, but collected or requested by the organization at variable frequencies. This includes surveys, market research data, and non-periodic report data.

The second group of data sources are those external, that is, those sources which are not controlled by the organization, though they may contain information relevant to its operation. A wide variety of sources may be considered as relevant for the organization purposes. For instance, some open data offered by public institutions might provide some information on the context of the company customers. Similarly, social and economic data published by the official statistics institutions have also potential for explaining the context in which individuals make decisions. Google Trends and social media platforms, such as Twitter and Facebook, are useful sources for detecting trends and relations between relevant topics. Furthermore, many other websites or RSS providing product opinions, political comments, product releases, etc. might be explored to find some other contextual variables that could complement own data sources.

Since the access to these sources is widely heterogeneous, the elements of this data receiving module must hide the complexity for accessing the sources. The access to these sources by the different elements in the module may be done by means of an Application Programming Interface (API) when available from the data provider, or by means of specific software developed for this purpose, e.g., web scraper.

All elements in this module will receive the data with the format and structure provided by the origin, which could be incompatible among them. According to their structure, data can be classified as structured, semi-structured or unstructured. Structured data includes information organized at high level, such as in relational databases, which apart from data, contains a schema with restrictions and relations. Semi-structured data also have some organization of the information, although the schema is embedded in the data. That is, its structure is self-describing, as in XML documents. Unstructured data provide no structure at all, and can be considered as a collection of

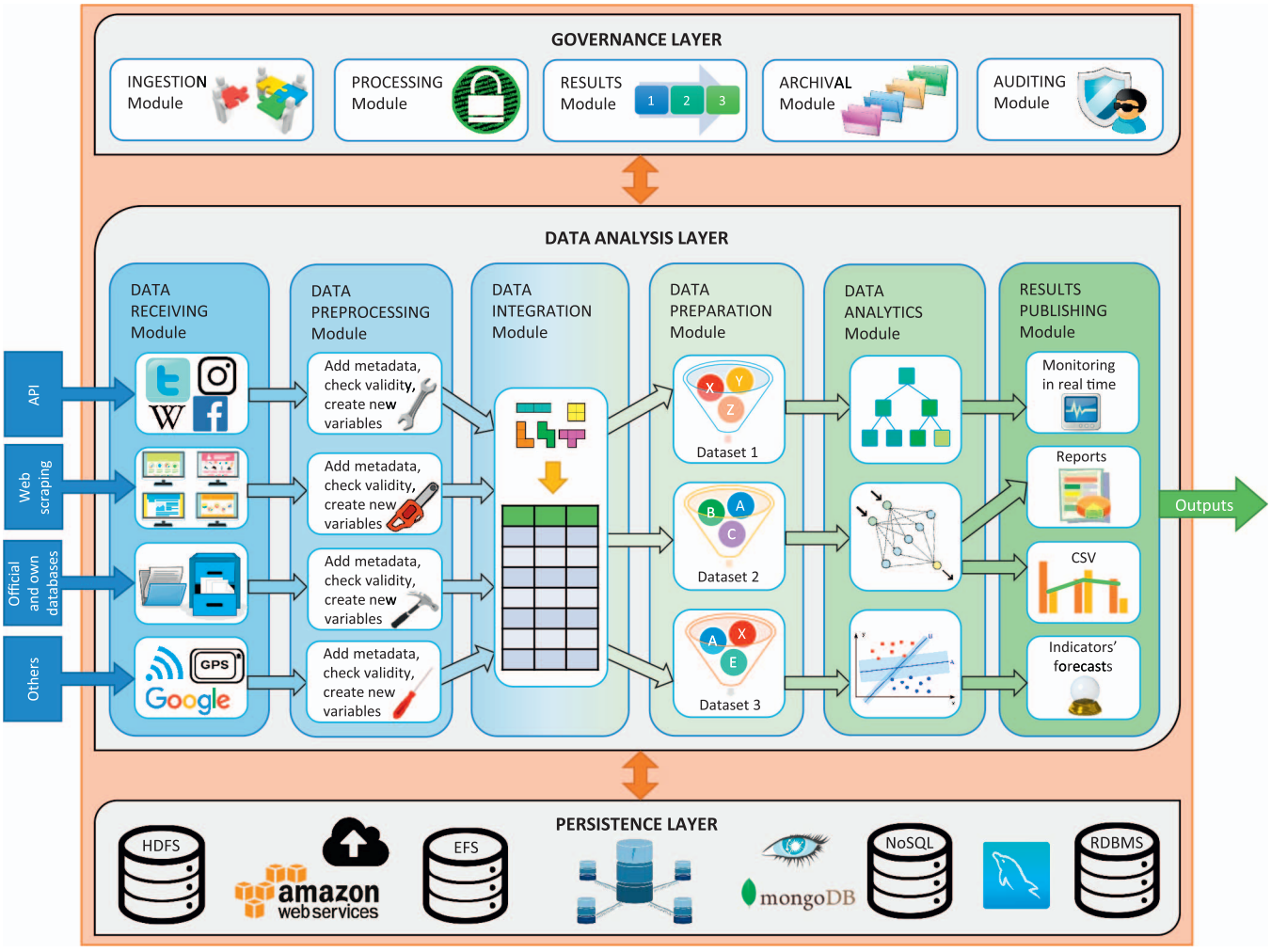

Fig. 4. Big Data architecture for nowcasting and forecasting social and economic changes.

elements. This does not mean that each element does not have a structure, but that the schema has not been described, so additional techniques to infer some structure should be applied. Text documents are typical unstructured sources. Data structure is a key factor to succeed integrating data from different sources, as it is the first step to establish the links between them. Structured data are usually related to SQL databases, while NoSQL are more suitable for storing unstructured and semi-structured data.

The elements in this module could access origins in batch or in stream. Stream processing is possible when the source allows access with high bandwidth and low latency conditions, e.g., when accessing an own relational database. However, when access conditions are not so favourable, the elements should work in batch, thus requiring persistent storage for the accessed data. In such event, the type of storage (SQL or NoSQL) must be consistent with the data source type. In any case, the data access that this module provides to the remaining modules of the architecture should be given as in stream processing.

6.1.2. *Data preprocessing module*

This module departs from the data connections prepared in the data receiving module. It aims to validate and preprocess data to leave them ready for integrating different sources. This preprocessing is divided in three steps.

The first step is to record and document the procedure of data acquisition by attaching metadata to the original source. These metadata should include information about the data source, the version of the collector (in the receiving module) used for the retrieval, the schema with the data structure (if any) and other technical details such as the codification, format and so on.

The second step is to check the internal validity of each source. Although structured sources usually keep all observations in the right format, other sources may be internally inconsistent. Thus, this step involves checking observations for anomalous values (e.g., text when a number is expected) and dealing with them, for instance, by marking them as missing or wrong. This may result in a number of useless observations, that is, those with an excessive number of missing features, which may be cleaned up to avoid including noise in the data analysis process.

The third step is related to the extraction of features and the generation of new data derived from the original source. At this step, only derived data at entity level should be created. That is, if the origin provides rows, only data from each row may be used to generate new features. If the origin provides documents, only document contents may be used to generate variables describing the document, for instance, by applying natural language processing techniques. Examples of entity-level derived data may include counting the number of words of a comment, detecting the language and computing the term frequency. Derived data whose computation requires analyzing several entities (e.g., computing an average) should be generated in the data analytics module. When the computational effort to generate new data is high, the resulting features should be persisted to allow for reusing them in subsequent data integrations. This involves using a database consistent with the structure type of the origin.

6.1.3. Data integration module

The objective of this module is to merge the different data sources and provide homogeneous access to all data available to the organization. To do so, data integration must deal with five challenges: i) heterogeneous sources, whose access was homogenized by the data receiving module; ii) related data structures whose relation has not been explicitly established by the sources; iii) a variety of data sizes and probably inconsistent formats; iv) heterogeneous time frequencies, ranging from milliseconds to years; and v) heterogeneous geographic groupings, ranging from detailed GPS coordinates to state or country level.

To relate data from different sources, it is required to define schemes that establish the relation among them. For instance, establishing the relation of a commercial establishment to a region, it will be possible to relate its sales to the average income of the area in which it is located. These data usually come from different sources: sales are accessed through internal sources, while the average income could be provided by an official statistics institute.

To establish such relations, some linkage techniques and hierarchical groupings might be applied. Geographic hierarchies are useful to link records to the most appropriate geographic level, which is not necessarily the lowest one. For instance, street-level economic situation may be useful for analyzing housing prices, but it is too specific for a business whose influence area is wider, e.g., an airline office. Linkage techniques are required when the same entity does not receive the same identifier across the different sources. This could happen simply because it is written in a different language (e.g., the country name), situation which can be solved with a simple translation; but also because of lacking of a standardized or public id. In such cases, some analysis to match the record could help find relations and provide new insights on the data.

Adapting time frequencies is also included in this category. It is required to adopt some criteria to generate high frequency data from lower frequencies and vice versa. Reducing time frequencies may involve computing some summarizing statistics (e.g., average and maximum), while increasing time frequencies may involve interpolating data or selecting the closest value in time.

Once several data sources are integrated and their relations are established, they could be stored in the persistence layer and feed the data analysis layer again as a new element in the data receiving module. This way, it is possible to use these sources as a single one when integrating with additional sources.

6.1.4. Data preparation module

The organization in which data are stored after the integration may not be suitable to perform the analysis. This module takes the data as prepared by the data integration module and transforms them to match the format expected by the data analytics module. Since each element in the analytics module may expect data in a different format, data preparation is also specific to each analytics element.

These transformations may involve grouping some elements or joining data from different tables to enrich the information about each entity or individual. This is also the most suitable module to alleviate the missing data, which may be estimated or interpolated to avoid losing cases.

A common operation in this module is the pivot transformation. Storing information as key-value pairs, in which entity features are spread among many rows, is quite convenient in Big Data environments. However, this may not be the table format expected by the analysis software. By applying the pivot transformation, all features regarding the same entity are arranged in the same row, which is the data organization commonly required to feed the analysis.

The resulting data after the preparation process should be stored to provide consistent input to the analysis. After conducting data integration and depending on the purpose of each particular study, it is possible to obtain small data sets derived from the initial big data set, which could be treated with traditional statistical techniques. The storage in this module must be analysis-driven, unlike the previous modules, whose storage is source-driven.

6.1.5. Data analytics module

This module applies statistical and machine learning methods to extract knowledge and make predictions from the data prepared by the previous module. To do so, descriptive and predictive techniques are applied. The descriptive analysis could provide some insights on the characteristics and evolution of the socio-economic variables under study. Its results will be used in the results publishing module to create tables and graphics representing the relationship among variables.

Predictive techniques are based on models that help explain, classify, forecast or nowcast the socio-economic variables under study. To do so, the models are estimated or trained by using learning methods and relying on any of the methods described in Section 4 for selecting the most meaningful variables and improving predictions. The computing-intensive nature of these techniques makes it more challenging to deal with large data sets, since they may not properly scale when data size grows.

Before using the models, they must be validated with a different set of data than that used for estimation or training. The validation provides an estimation on the robustness of the models and the quality of the predictions, so that the risk related to an inaccurate prediction can be taken into account.

The methods used in this module may be applied in stream (i.e., the models are continuously being trained with new data), scheduled (i.e., the models are trained periodically), or on demand (i.e, the user manually requests to train again the models). Choosing one or other approach depends mainly on the computational resources available for this module.

The main output of the predictive techniques are the trained models, whose application can guide the operative and the strategy of the organization. They are made available to the rest of the organization by means of the results publishing module.

6.1.6. Results publishing module

The purpose of this module is to provide the organization with a decision-making tool. To do so, it makes the results of the analysis conducted in the data analytics module available to the organization, which includes the people that make decisions, but also other information systems that could benefit from the data analysis. For this reason, this module should offer the results in different formats, adapted to the different consumers of information in the organization.

The publication of results for decision-makers should be done in the form of reports, including tables, graphics and other visual elements that help understand the social and economic behavior behind the data. The main objective of these reports is to support decisions at strategic or tactical levels.

Making the analysis results available to other information systems in the organization contributes to support the decision-making at operational level. There is a wide variety of options to do so. For instance, a trained model can be stored in a database or in any other storage for being applied by different business units. The model could also be offered as a service (under the SaaS paradigm) so that when a new event occurs, the service offers the prediction as a result of applying the model. This way, the trained models can be successfully applied to some operational actions such as purchases and financial resources management.

6.2. Governance layer

This layer is horizontal to the rest of the system and applies the organization policies and regulations to the whole data lifecycle: from the data ingestion to the disposal. It is composed of five modules, four of them related to the data lifecycle, plus one for auditing purposes.

- Ingestion module: It deals with the management of the sources, including the licenses and allowed uses, credentials for accessing them, internal user permissions, completeness of metadata, and so on.
- Processing module: It manages the privacy and anonymization policies, controls processing for ethical principles, keeps track of the transformations, as well as of the permissions for accessing the data and computing resources.
- Results module: It is concerned with the traceability of the results (from the sources to the final report), the permissions for accessing the reports and results, along with the privacy aspects that may affect the reports.
- Archival and disposal module: It implements the policy for archiving and disposing the information related to data sources, processing procedures and generated reports.
- Auditing module: It inspects that the implementation of the architecture is consistent with the current regulations, as well as with the security and privacy policies. It may also include checking the overall performance of the architecture in order to ensure that the system has an acceptable response time.

6.3. Persistence layer

The persistence layer supports the other layers by managing all issues related to the storage needs. Its main function is associated with the storage of the data used as input in the data analysis layer, including the schema for describing the relations among sources and other metadata. Not only the data itself is covered, but also the storage of the procedures followed in the different modules to access and transform the data.

Furthermore, this layer serves the data analytics and results publishing modules by providing storage for the results. This includes storing the models and providing them with the inputs required for computing new predictions or estimations as part of the publication of the results.

It is in the persistence layer where the storage infrastructure is controlled and managed. This layer will typically use distributed storage systems, combining local storage with cloud solutions that allow elastic storage and large volume data. The decision on whether to use local or cloud storage mainly depends on where (on- or off-premises) the modules intensive in computing power (e.g., data analytics) are implemented.

7. Conclusions

In the Digital Era, most economic and social behaviors leaves behind a huge digital footprint, which is incipiently being used with nowcasting and forecasting purposes. Despite the enormous potential of these data, integrating and analyzing the wide variety of heterogeneous sources cannot be tackled with the traditional methods used in economics and social sciences. To succeed in this purpose, it is mandatory to carefully plan and implement the whole process of data extraction, transformation and analysis. This is the point in which the Big Data and data lifecycle paradigms arise as helpful perspectives on how to deal with this process.

This paper has proposed a novel Big Data architecture that accounts for the particularities of the economic and social behavior analyses in the Digital Era. The first particularity is related to the variety of sources that could provide information about economic and social topics. Our first contribution addresses this issue by reviewing the multiple data sources and proposing a taxonomy to classify them according to the purpose of the agent generating the data.

Following the Big Data paradigm, this wide variety of heterogeneous sources requires specific methods for processing them. The second contribution of the paper addresses this issue by reviewing those methods not so commonly used in the social sciences, and classifying them according to the phase of the data analysis they operate.

In order to frame the data analysis in an organizational perspective and allow its management in a robust and flexible architecture, the data lifecycle approach has been taken. Different perspectives on this approach have been reviewed and synthesized to establish and define all the involved phases and processes.

Finally, the main contribution of the paper is the proposal of a Big Data architecture adapted to the particularities of the economic and social analyses, and grounded on the data lifecycle approach for the management of data in the organization. At the same time, the proposal aimed to be general enough to be implemented with different technologies, computing paradigms and analytical software depending on the requirements and purposes of each particular case. By implementing this architecture, an organization will be able to make the most of all social and economic sources of information to which it has access. Not only the organization of sources is advantageous, but also their integration and connection to Big Data analytics tools able to run the models for nowcasting and forecasting socio-economic variables. The wide variety of data sources and techniques considered in the architecture results in potentially more accurate and granular predictions.

Governments and official statistics institutions may also benefit from the implementation of an information system with the proposed architecture. Integrating the multiple sources to which they have access may result in improved predictions about key economic indicators and planning economic policies accordingly.

Although the proposed architecture is general enough to be implemented with any technology, its adoption is not without obstacles. To mention some of them, the integration of the architecture in the existing organizational information systems is a critical process to ensure the smooth generation of forecasts and nowcasts. The implementation of the modules in a proper cloud computing environment so that the system can scale easily is also crucial. As future work, we plan to implement the proposed Big Data architecture, in order to generate and publish real-time nowcasts and forecasts of some socio-economic variables using Internet data.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness under Grant TIN2013-43913-R; and by the Spanish Ministry of Education under Grant FPU14/02386.

References

Alis, C.M., Letchford, A., Moat, H.S., Preis, T., 2015. Estimating tourism statistics with Wikipedia page views. In: WebSci-15 Proceedings of the ACM Web Science Conference, Oxford, United Kingdom, http://dx.doi.org/10.1145/2786451.2786925.

Arenas-Márquez, F., Martínez-Torres, M., Toral, S., 2014. Electronic word-of-mouth communities from the perspective of social network analysis. Tech. Anal. Strat. Manag. 26 (8), 927–942. http://dx.doi.org/10.1080/09537325.2014.923565.

Armentano, M.G., Godoy, D., Campo, M., Amandi, A., 2014. NLP-based faceted search: experience in the development of a science and technology search engine. Expert Syst. Appl. 41 (6), 2886–2896. http://dx.doi.org/10.1016/j.eswa.2013.10.023.

Arora, S.K., Li, Y., Youtie, J., Shapira, P., 2016. Using the Wayback Machine to mine websites in the social sciences: a methodological resource. J. Assoc. Inf. Sci. Technol. 67 (8), 1904–1915. http://dx.doi.org/10.1002/asi.23503.

Arora, S.K., Youtie, J., Shapira, P., Gao, L., Ma, T., 2013. Entry strategies in an emerging technology: a pilot web-based study of graphene firms. Scientometrics 95 (3), 1189–1207. http://dx.doi.org/10.1007/s11192-013-0950-7.

Arrigo, E., Liberati, C., Mariani, P., 2016. A multivariate approach to Facebook data for marketing communication. In: Proceedings of the 1st International Conference on Advanced Research Methods and Analytics (CARMA 2016). UPV Press, Universitat Politécnica de Valéncia, Valencia, Spain. http://dx.doi.org/10.4995/CARMA2016.2016.2974.

Artola, C., Pinto, F., de Pedraza García, P., 2015. Can internet searches forecast tourism inflows? Int. J. Manpow. 36 (1), 103–116. http://dx.doi.org/10.1108/IJM-12-2014-0259.

Askitas, N., Zimmermann, K.F., 2009. Google econometrics and unemployment forecasting. Appl. Econ. Q. 55 (2), 107–120. http://dx.doi.org/10.3790/aeq.55.2.107.

Askitas, N., Zimmermann, K.F., 2013. Nowcasting business cycles using toll data. J. Forecast. 32 (4), 299–306. http://dx.doi.org/10.1002/for.1262.

Askitas, N., Zimmermann, K.F., 2015. The internet as a data source for advancement in

social sciences. Int. J. Manpow. 36 (1), 2–12. http://dx.doi.org/10.1108/IJM-02-2015-0029.
Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A., Buyya, R., 2015. Big Data computing and clouds: trends and future directions. J. Parallel Distrib. Comput. 79-80, 3–15. http://dx.doi.org/10.1016/j.jpdc.2014.08.003.
Bahrami, M., Singhal, M., 2014. The Role of Cloud Computing Architecture in Big Data. vol. 8. Springer International Publishing, Cham, pp. 275–295. http://dx.doi.org/10.1007/978-3-319-08254-7_13.
Bangwayo-Skeete, P.F., Skeete, R.W., 2015. Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. Tour. Manag. 46, 454–464. http://dx.doi.org/10.1016/j.tourman.2014.07.014.
Bello-Orgaz, G., Jung, J.J., Camacho, D., 2016. Social big data: recent achievements and new challenges. Inf. Fusion 28, 45–59. http://dx.doi.org/10.1016/j.inffus.2015.08.005.
Berman, F., Fox, G., Hey, A.J., 2003. Grid Computing: Making the Global Infrastructure a Reality. Communications Networking & Distributed Systems John Wiley and Sons.
Blazquez, D., Domenech, J., 2017. Web data mining for monitoring business export orientation. Technol. Econ. Dev. Econ. Online, 1–23. http://dx.doi.org/10.3846/20294913.2016.1213193.
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. J. Comput. Sci. 2 (1), 1–8. http://dx.doi.org/10.1016/j.jocs.2010.12.007.
Cameron, M.P., Barrett, P., Stewardson, B., 2016. Can social media predict election results? Evidence from New Zealand. J. Polit. Mark. 15 (4), 416–432. http://dx.doi.org/10.1080/15377857.2014.959690.
Ceron, A., Curini, L., Iacus, S.M., Porro, G., 2014. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media Soc. 16 (2), 340–358. http://dx.doi.org/10.1177/1461444813480466.
Ceron, A., Negri, F., 2016. The “social side” of public policy: monitoring online public opinion and its mobilization during the policy cycle. Policy Internet 8 (2), 131–147. http://dx.doi.org/10.1002/poi3.117.
Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. CRISP-DM 1.0 - Step-by-Step Data Mining Guide. https://www.the-modeling-agency.com/crisp-dm.pdf (accessed 1st June, 2017).
Chen, M., Mao, S., Liu, Y., 2014. Big Data: a survey. Mob. Netw. Appl. 19 (2), 171–209. http://dx.doi.org/10.1007/s11036-013-0489-0.
Chittaranjan, G., Blom, J., Gatica-Perez, D., 2013. Mining large-scale smartphone data for personality studies. Pers. Ubiquit. Comput. 17 (3), 433–450. http://dx.doi.org/10.1007/s00779-011-0490-1.
Choi, H., Varian, H., 2009a. Predicting Initial Claims for Unemployment Benefits. http://research.google.com/archive/papers/initialclaimsUS.pdf (accessed 10th October, 2016).
Choi, H., Varian, H., 2009b. Predicting the Present with Google Trends. http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/en//googleblogs/pdfs/google_predicting_the_present.pdf (accessed 10th October, 2016).
Chong, A.Y.L., Ch'ng, E., Liu, M.J., Li, B., 2015. Predicting consumer product demands via Big Data: the roles of online promotional marketing and online reviews. Int. J. Prod. Res. Online, 1–15. http://dx.doi.org/10.1080/00207543.2015.1066519.
Chou, J.-S., Ngo, N.-T., 2016. Smart grid data analytics framework for increasing energy savings in residential buildings. Autom. Constr. 72 (3), 247–257. http://dx.doi.org/10.1016/j.autcon.2016.01.002.
Chowdhury, G.G., 2005. Natural language processing. Annu. Rev. Inf. Sci. Technol. 37 (1), 51–89. http://dx.doi.org/10.1002/aris.1440370103.
Committee on Earth Observation Satellites — Working Group on Information Systems and Services, 2012. CEOS Data Life Cycle Models and Concepts. https://my.usgs.gov/confluence/download/attachments/82935852/Data%20Lifecycle%20Models%20and%20Concepts% (accessed 27th September, 2016).
Congdon, P., 2007. Bayesian Statistical Modelling. Wiley Series in Probability and Statistics, 2nd. John Wiley & Sons.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. http://dx.doi.org/10.1007/BF00994018.
Corti, L., Van den Eynden, V., Bishop, L., Woollard, M., 2014. Managing and sharing research data: a guide to good practice, 1st. Sage Publications.
Cox, M., Ellsworth, D., 1997. Managing Big Data for scientific visualization. ACM Siggraph, MRJ/NASA Ames Res. Cent. 5, 1–17.
David, E., Zhitomirsky-Geffet, M., Koppel, M., Uzan, H., 2016. Utilizing Facebook pages of the political parties to automatically predict the political orientation of Facebook users. Online Inf. Rev. 40 (5), 610–623. http://dx.doi.org/10.1108/OIR-09-2015-0308.
Alliance, D.D.I., 2008. DDI Lifecycle 3.0. http://www.ddialliance.org/ (accessed 29th September, 2016).
Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J., 2014. Dynamic population mapping using mobile phone data. Proc. Natl. Acad. Sci. 111 (45), 15888–15893. http://dx.doi.org/10.1073/pnas.1408439111.
Dey, M.M., Rabbani, A.G., Singh, K., Engle, C.R., 2014. Determinants of retail price and sales volume of catfish products in the United States: an application of retail scanner data. Aquac. Econ. Manag. 18 (2), 120–148. http://dx.doi.org/10.1080/13657305.2014.903312.
Dobra, A., Williams, N.E., Eagle, N., 2015. Spatiotemporal detection of unusual human population behavior using mobile phone data. PLOS ONE 10 (3), 1–20. http://dx.doi.org/10.1371/journal.pone.0120449.s001.
Domenech, J., de la Ossa, B., Pont, A., Gil, J.A., Martinez, M., Rubio, A., 2012. An intelligent system for retrieving economic information from corporate websites. In: IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Macau, China, pp. 573–578. http://dx.doi.org/10.1109/WI-IAT.2012.92.
Edelman, B., 2012. Using Internet data for economic research. J. Econ. Perspect. 26 (2), 189–206. http://dx.doi.org/10.1257/jep.26.2.189.
Einav, L., Levin, J., 2014. The data revolution and economic analysis. Innov. Policy Econ. 14 (1), 1–24. http://dx.doi.org/10.1086/674019.
Evangelopoulos, N., Zhang, X., Prybutok, V.R., 2012. Latent semantic analysis: five methodological recommendations. Eur. J. Inf. Syst. 21 (1), 70–86. http://dx.doi.org/10.1057/ejis.2010.61.
Fan, J., Han, F., Liu, H., 2014. Challenges of big data analysis. Nat. Sci. Rev. 1 (2), 293–314. http://dx.doi.org/10.1093/nsr/nwt032.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27 (8), 861–874. http://dx.doi.org/10.1016/j.patrec.2005.10.010.
Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. The KDD process for extracting useful knowledge from volumes of data. Commun. ACM 39 (11), 27–34.
Fondeur, Y., Karamé, F., 2013. Can Google data help predict French youth unemployment? Econ. Model. 30, 117–125. http://dx.doi.org/10.1016/j.econmod.2012.07.017.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for Generalized Linear Models via coordinate descent. J. Stat. Softw. 33 (1), 1–22.
Frota Neto, J.Q., Bloemhof, J., Corbett, C., 2016. Market prices of remanufactured, used and new items: evidence from eBay. Int. J. Prod. Econ. 171 (3), 371–380. http://dx.doi.org/10.1016/j.ijpe.2015.02.006.
Gaikar, D.D., Marakarkandy, B., Dasgupta, C., 2015. Using Twitter data to predict the performance of Bollywood movies. Ind. Manag. Data Syst. 115 (9), 1604–1621. http://dx.doi.org/10.1108/IMDS-04-2015-0145.
Gandomi, A., Haider, M., 2015. Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manag. 35 (2), 137–144. http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007.
Gayo-Avello, D., 2012, Nov, Nov. No, you cannot predict elections with twitter. IEEE Internet Comput. 16 (6), 91–94. http://dx.doi.org/10.1109/MIC.2012.137.
Gayo-Avello, D., 2013. A meta-analysis of state-of-the-art electoral prediction from Twitter data. Soc. Sci. Comput. Rev. 31 (6), 649–679. http://dx.doi.org/10.1177/0894439313493979.
Gök, A., Waterworth, A., Shapira, P., 2015. Use of web mining in studying innovation. Scientometrics 102 (1), 653–671. http://dx.doi.org/10.1007/s11192-014-1434-0.
Graells-Garrido, E., Peredo, O., García, J., 2016. Sensing urban patterns with antenna mappings: the case of Santiago, Chile. Sensors 16 (7), 1098–1123. http://dx.doi.org/10.3390/s16071098.
Hand, C., Judge, G., 2012. Searching for the picture: forecasting UK cinema admissions using Google Trends data. Appl. Econ. Lett. 19 (11), 1051–1055. http://dx.doi.org/10.1080/13504851.2011.613744.
Harsanyi, J.C., 1978. Bayesian decision theory and utilitarian ethics. Am. Econ. Rev. 68 (2), 223–228.
Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Ullah Khan, S., 2015. The rise of “big data” on cloud computing: review and open research issues. Inf. Syst. 47, 98–115. http://dx.doi.org/10.1016/j.is.2014.07.006.
Hastie, T., Rosset, S., Tibshirani, R., Zhu, J., 2004. The entire regularization path for the support vector machine. J. Mach. Learn. Res. 5, 1391–1415.
Hastie, T., Tibshirani, R., Friedman, J., 2013. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer Series in Statistics, 3rd. Springer.
He, H., Garcia, E., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21 (9), 1263–1284. http://dx.doi.org/10.1109/TKDE.2008.239.
Hu, N., Bose, I., Koh, N.S., Liu, L., 2012. Manipulation of online reviews: an analysis of ratings, readability, and sentiments. Decis. Support. Syst. 52 (3), 674–684. http://dx.doi.org/10.1016/j.dss.2011.11.002.
IBM, 2016. Big Data and Analytics. http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html (accessed 21st December, 2016).
Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., Shahabi, C., 2014. Big data and its technical challenges. Commun. ACM 57 (7), 86–94. http://dx.doi.org/10.1145/2611567.
Jin, X., Wah, B.W., Cheng, X., Wang, Y., 2015. Significance and challenges of big data research. Big Data Res. 2 (2), 59–64. http://dx.doi.org/10.1016/j.bdr.2015.01.006.
Khadivi, P., Ramakrishnan, N., 2016. Wikipedia in the tourism industry: forecasting demand and modeling usage behavior. In: Thirtieth AAAI Conference on Artificial Intelligence. February 12–17, 2016, Phoenix, Arizona, pp. 4016–4021.
Kim, M., Park, H.W., 2012. Measuring Twitter-based political participation and deliberation in the South Korean context by using social network and Triple Helix indicators. Scientometrics 90 (1), 121–140. http://dx.doi.org/10.1007/s11192-011-0508-5.
Kim, T., Hong, J., Kang, P., 2015. Box office forecasting using machine learning algorithms based on SNS data. Int. J. Forecast. 31 (2), 364–390. http://dx.doi.org/10.1016/j.ijforecast.2014.05.006.
Kitchin, R., 2014. The real-time city? Big data and smart urbanism. GeoJournal 79 (1), 1–14. http://dx.doi.org/10.1007/s10708-013-9516-8.
Krishnan, N.C., Cook, D.J., 2014. Activity recognition on streaming sensor data. Pervasive Mob. Comput. 10, 138–154. http://dx.doi.org/10.1016/j.pmcj.2012.07.003.
Laney, D., 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Application Delivery Strategies. pp. 949. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed 21st December, 2016).
Laurila, J.K., Gatica-Perez, D., Aad, I., Blom, J., Bornet, O., Do, T.M.T., Dousse, O., Eberle, J., Miettinen, M., 2013. From big smartphone data to worldwide research: the mobile data challenge. Pervasive Mob. Comput. 9, 752–771. http://dx.doi.org/10.1016/j.pmcj.2013.07.014.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

http://dx.doi.org/10.1038/nature14539.
Ley, E., Steel, M.F., 2012. Mixtures of g-priors for Bayesian model averaging with economic applications. J. Econ. 171 (2), 251–266. http://dx.doi.org/10.1016/j.jeconom.2012.06.009.
Li, G., Law, R., Vu, H.Q., Rong, J., Zhao, X.R., 2015. Identifying emerging hotel preferences using Emerging Pattern Mining technique. Tour. Manag. 46, 311–321. http://dx.doi.org/10.1016/j.tourman.2014.06.015.
Li, Y., Arora, S., Youtie, J., Shapira, P., 2016. Using web mining to explore Triple Helix influences on growth in small and mid-size firms. Technovation Online, 1–12. http://dx.doi.org/10.1016/j.technovation.2016.01.002.
Liu, B., 2012. Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. 5 (1), 1–167. http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016.
Liu, Y., Huang, X., An, A., Yu, X., 2007. ARSA: sentiment-aware model for predicting sales performance using blogs. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 607–614.
Ludwig, O., Nunes, U., Araujo, R., 2014. Eigenvalue decay: a new method for neural network regularization. Neurocomputing 124, 33–42. http://dx.doi.org/10.1016/j.neucom.2013.08.005.
Lynch, C., 2008. Big data: how do your data grow? Nature 455, 28–29. http://dx.doi.org/10.1038/455028a.
Malbon, J., 2013. Taking fake online consumer reviews seriously. J. Consum. Policy 36 (2), 139–157. http://dx.doi.org/10.1007/s10603-012-9216-7.
Mavragani, A., Tsagarakis, K.P., 2016. YES or NO: predicting the 2015 GReferendum results using Google Trends. Technol. Forecast. Soc. Chang. 109, 1–5. http://dx.doi.org/10.1016/j.techfore.2016.04.028.
McLaren, N., Shanbhogue, R., 2011. Using internet search data as economic indicators. Bank Engl. Q. Bull. 2011 Q2, 134–140.
Menardi, G., Torelli, N., 2014. Training and assessing classification rules with imbalanced data. Data Min. Knowl. Disc. 28, 92–122. http://dx.doi.org/10.1007/s10618-012-0295-5.
Moat, H.S., Curme, C., Stanley, H.E., Preis, T., 2014. Anticipating Stock Market Movements with Google and Wikipedia. NATO Science for Peace and Security Series C: Environmental Security Springer Science, pp. 47–59. http://dx.doi.org/10.1007/978-94-017-8704-8_4.
Montoliu, R., Blom, J., Gatica-Perez, D., 2013. Discovering places of interest in everyday life from smartphone data. Multimedia Tools Appl. 62, 179–207. http://dx.doi.org/10.1007/s11042-011-0982-z.
Moro, S., Cortez, P., Rita, P., 2015. Business intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Syst. Appl. 42, 1314–1324. http://dx.doi.org/10.1016/j.eswa.2014.09.024.
Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E., 2015. Deep learning applications and challenges in big data analytics. J. Big Data 2 (1), 1–21. http://dx.doi.org/10.1186/s40537-014-0007-7.
Pääkkönen, P., Pakkala, D., 2015. Reference architecture and classification of technologies, products and services for big data systems. Big Data Res. 2 (4), 166–186. http://dx.doi.org/10.1016/j.bdr.2015.01.001.
Pan, B., Yang, Y., 2016. Monitoring and Forecasting Tourist Activities with Big Data. Apple Academic Press, pp. 43–62 chap. 3.
Pandya, S.S., Venkatesan, R., 2016. French roast: consumer response to international conflict — evidence from supermarket scanner data. Rev. Econ. Stat. 98 (1), 42–56. http://dx.doi.org/10.1162/REST_a_00526.
Pesenson, M.Z., Pesenson, I.Z., McCollum, B., 2010. The data big bang and the expanding digital universe: high-dimensional, complex and massive data sets in an inflationary epoch. Adv. Astron. 2010, 1–16. http://dx.doi.org/10.1155/2010/350891.
Preis, T., Moat, H.S., Stanley, H.E., 2013. Quantifying trading behavior in financial markets using Google Trends. Sci Rep 3, 1–6. http://dx.doi.org/10.1038/srep01684.
Reed, D.A., Dongarra, J., 2015. Exascale computing and big data. Commun. ACM 58, 56–68. http://dx.doi.org/10.1145/2699414.
Rudolph, M., Ruiz, F., Mandt, S., Blei, D., 2016. Exponential family embeddings. In: Advances in Neural Information Processing Systems, pp. 478–486.
Rüegg, J., Gries, C., Bond-Lamberty, B., Bowen, G.J., Felzer, B.S., McIntyre, N.E., Soranno, P.A., Vanderbilt, K.L., Weathers, K.C., 2014. Completing the data life cycle: using information management in macrosystems ecology research. Front. Ecol. Environ. 12, 24–30. http://dx.doi.org/10.1890/120375.
Russell, M.A., 2013. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub and More, 2nd. O'Reilly Media, pp. 448.
Saleiro, P., Amir, S., Silva, M., Soares, C., 2015. Popmine: tracking political opinion on the web. In: 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM). IEEE, Liverpool, United Kingdom, pp. 1521–1526.
Schneider, M.J., Gupta, S., 2016. Forecasting sales of new and existing products using consumer reviews: a random projections approach. Int. J. Forecast. 32, 243–256. http://dx.doi.org/10.1016/j.ijforecast.2015.08.005.
Schoen, H., Gayo-Avello, Panagiotis Takis Metax, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., 2013. The power of prediction with social media. Internet Res. 23, 528–543. http://dx.doi.org/10.1108/IntR-06-2013-0115.
Simonet, A., Fedak, G., Ripeanu, M., 2015. Active data: a programming model to manage data life cycle across heterogeneous systems and infrastructures. Futur. Gener. Comput. Syst. 53, 25–42. http://dx.doi.org/10.1016/j.future.2015.05.015.
Suhara, Y., Xu, Y., Pentland, A., 2017. Deepmood: forecasting depressed mood based on self-reported histories via recurrent neural networks. In: Proceedings of the 26th International Conference on World Wide Web - WWW '17. ACM Press, pp. 715–724.
Sun, Y., Kamel, M.S., Wong, A.K., Wang, Y., 2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recogn. 40 (12), 3358–3378. http://dx.doi.org/10.1016/j.patcog.2007.04.009.
Thelwall, M., 2007. Blog searching: the first general-purpose source of retrospective public opinion in the social sciences? Online Inf. Rev. 31 (3), 277–289. http://dx.doi.org/10.1108/14684520710764069.
Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B Methodol. 58, 267–288.
Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M., 2011. Election forecasts with Twitter: how 140 characters reflect the political landscape. Soc. Sci. Comput. Rev. 29 (4), 402–418. http://dx.doi.org/10.1177/0894439310386557.
Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., Baesens, B., 2015. APATE: a novel approach for automated credit card transaction fraud detection using network-based extensions. Decis. Support. Syst. 75, 38–48. http://dx.doi.org/10.1016/j.dss.2015.04.013.
Varian, H.R., 2014. Big Data: new tricks for econometrics. J. Econ. Perspect. 28, 3–28. http://dx.doi.org/10.1257/jep.28.2.3.
Vatsalan, D., Christen, P., Verykios, V.S., 2013. A taxonomy of privacy-preserving record linkage techniques. Inf. Syst. 38, 946–969. http://dx.doi.org/10.1016/j.is.2012.11.005.
Vicente, M.R., López-Menéndez, A.J., Pérez, R., 2015. Forecasting unemployment with internet search data: does it help to improve predictions when job destruction is skyrocketing? Technol. Forecast. Soc. Chang. 92, 132–139. http://dx.doi.org/10.1016/j.techfore.2014.12.005.
Vosen, S., Schmidt, T., 2011. Forecasting private consumption: survey-based indicators vs. Google Trends. J. Forecast. 30 (6), 565–578. http://dx.doi.org/10.1002/for.1213.
Wang, Y., Kung, L., Byrd, T.A., 2016a. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol. Forecast. Soc. Chang. Online, 1–11. http://dx.doi.org/10.1016/j.techfore.2015.12.019.
Wang, Y., Yuan, N.J., Sun, Y., Zhang, F., Xie, X., Li, Q., Chen, E., 2016b. A contextual collaborative approach for app usage forecasting. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing — UbiComp '16. ACM Press, pp. 1247–1258.
Wikimedia Foundation, 2017. Dashboards and Data Downloads for Wikimedia Projects. https://analytics.wikimedia.org/ (accessed 7th July, 2017).
Williams, N.E., Thomas, T.A., Dunbar, M., Eagle, N., Dobra, A., 2015. Measures of human mobility using mobile phone records enhanced with GIS data. PLOS ONE 10, 1–16. http://dx.doi.org/10.1371/journal.pone.0133630.
Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques, 4th. Morgan Kaufmann - Elsevier.
Wu, J., Pan, S., Zhu, X., Cai, Z., Zhang, P., Zhang, C., 2015. Self-adaptive attribute weighting for Naive Bayes classification. Expert Syst. Appl. 42 (3), 1487–1502. http://dx.doi.org/10.1016/j.eswa.2014.09.019.
Xiong, T., Wang, S., Mayers, A., Monga, E., 2013. Personal bankruptcy prediction by mining credit card data. Expert Syst. Appl. 40, 665–676. http://dx.doi.org/10.1016/j.eswa.2012.07.072.
Zhang, Y., Ren, S., Liu, Y., Si, S., 2017. A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. J. Clean. Prod. 142 (2), 626–641. http://dx.doi.org/10.1016/j.jclepro.2016.07.123.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.

Desamparados Blazquez received a B.S. in Business Administration and Management and a M.S. in Data Analytics Engineering from the Universitat Politècnica de València (Spain). She is currently a Ph.D. student and a predoctoral research fellow under the Programme for the Training of University Lecturers (FPU) from the Spanish Ministry of Education. She develops her work at the Department of Economics and Social Sciences of the Universitat Politècnica de València. Her research interests include web economic indicators and internet economics.

Josep Domenech received a B.S., M.S. and Ph.D. in Computer Science from the Universitat Politècnica de València, and a B.S. and M.S. in Business Administration and Economics from the Universitat de València. Since 2009, he is an associate professor at the Department of Economics and Social Sciences of the Universitat Politècnica de València. He is currently leading a research project on planning and using cloud computing platforms efficiently, funded by the Ministry of Economy and Competitiveness of Spain. His research interests are focused on multidisciplinary approaches to internet systems and digital economics, including web economic indicators, internet economics and web performance characterization.