Journal of Big Data

SURVEY PAPER

Open Access

Conceptualizing Big Social Data

Ekaterina Olshannikova[1]*, Thomas Olsson[1], Jukka Huhtamäki[2] and Hannu Kärkkäinen[3]

*Correspondence: ekaterina.olshannikova@tut.fi
[1] Department of Pervasive Computing, Tampere University of Technology, Korkeakoulunkatu 10, 33720 Tampere, Finland
Full list of author information is available at the end of the article

Abstract

The popularity of social media and computer-mediated communication has resulted in high-volume and highly semantic data about digital social interactions. This constantly accumulating data has been termed as Big Social Data or Social Big Data, and various visions about how to utilize that have been presented. However, as relatively new concepts, there are no solid and commonly agreed definitions of them. We argue that the emerging research field around these concepts would benefit from understanding about the very substance of the concept and the different viewpoints to it. With our review of earlier research, we highlight various perspectives to this multi-disciplinary field and point out conceptual gaps, the diversity of perspectives and lack of consensus in what Big Social Data means. Based on detailed analysis of related work and earlier conceptualizations, we propose a synthesized definition of the term, as well as outline the types of data that Big Social Data covers. With this, we aim to foster future research activities around this intriguing, yet untapped type of Big Data.

Keywords: Big Social Data, Social Big Data, Digital human, Conceptualization, Social Data, Social media, Computational social science, Social computing, Classification, Big Social Data analysis

Introduction

Background

We live in an "always-on society" [1–3], meaning that people constantly interact with each other. Due to the rapid development of social computing and mushrooming of social media services, much of social interaction is nowadays mediated by information technology and takes place in the digital realm. An average Internet user consumes and shares large amounts of digital content every day through popular social online services, such as Facebook, Twitter, YouTube, Instagram and SnapChat.

From data perspective, this has led to emergence of extensive amounts of human-generated data [4, 5] with diverse social uses and rich meanings (for example, communication text, videos for entertainment and self-representation, sharing of news and other 3rd party content in social media). Such unstructured/semi-structured, yet semantically rich data has been argued to constitute 95% of all Big Data [6]. This Social Data explosion has resulted in theorizations and studies about the emerging topic of Big Social Data (BSD).

Broadly speaking, BSD refers to large data volumes that relate to people or describe their behavior and technology-mediated social interactions in the digital realm. The sheer volume and semantic richness of such data opens enormous possibilities for utilizing and analyzing it for personal [7, 8], commercial [9, 10] as well as societal purposes [11–13]. For example, the scattered social media would benefit from meta-services that bring together all the content from a user. Commercial use could include even more targeted advertising, matchmaking services, or many unimaginable data-centered business models [14, 15]. The search for beneficial applications and services in regard to BSD has only just begun.

Central concepts and goals of the research

In the research literature, the concept of Big Social Data has been defined and interpreted in many ways for various purposes; for example, the viewpoints from which it has been explored include social media, online social networks, social computing, and computational social science (CSS). The role of these fields in the scope of BSD is discussed in detail in the following sections.

As a rule, BSD is mainly utilized to extract insights from social media data and online social interactions of people for descriptive or predictive purposes to influence human decision-making in various application domains [16–18]. In general, researchers have focused on the analytics and utilization, having paid little attention to clarifying the very concept of BSD and understanding the related phenomena (for example, [19–21]).

In fact, there seems to be lack of consensus about the definition of BSD and the related terms, as we will analyze in the upcoming sections. Inconsideration of proper conceptualization may bring researchers methodological challenges in their studies, especially in such inherently broad and multi-disciplinary field as BSD.

Therefore, we argue for conceptual and theoretical work about the concept of BSD in order to inform future research activities as well as to foster the practical utilization of the data, which may signify social insight. There is a timely need to describe, review, and reflect on BSD literature in order to bring clarity to the concept and understanding about its beneficial opportunities for the practitioners of computational social science and other related research fields.

The potential value of this paper for the readers is presented as follows:

1. Firstly, by the literature review we aim to bring clarity on various existing BSD concepts and its definitions. We discuss relations between BSD and related fields of science in order to inform readers about the domains where this concept is currently applied. We consider these aspects will help researchers to properly identify scope and directions for their investigation on the topic;
2. Secondly, by providing a synthesized concept and definition of BSD we want to motivate researchers to develop better conceptualizations and clarifications of the BSD meaning in regard to their research. Currently, the majority of papers related to the topic are focused on analytical tasks and methods missing the explanation about what researchers consider as BSD and why. As an improvement step towards a holistic approach to this emerging field, BSD practitioners can utilize the definition presented in this work by revising it according to their research objectives;
3. By providing a comprehensive list of BSD types we aim to inform researchers about categories of data that is currently available for research and analysis. This serves as a starting point to identify research opportunities and practical means towards

data-driven research. It is worth noting that there is no extensive taxonomy of BSD in related literature and we neither aim to design one; however, our classification of such data serves as an inducement to the research community for collaboratively creating this taxonomy;

4. Moreover, by describing the key characteristics of BSD we differentiate it from the concept of Big Data. By doing so, we anticipate the emphasis on its unique qualities to open new opportunities for multi-disciplinary research ventures.

In general, we assume this work will attract researchers' attention to explore the holistic view on BSD concept and help them to identify relevant sources of data to utilize in BSD studies.

Related concepts and literature

Due to rapid development of online social services and tremendous growth of data therein, various concepts have emerged in different research fields to help understanding digital environments and their social effects. This section reviews related concepts relevant to BSD and their correlations, as well as outlines existing literature on the topic (see Fig. 1).

There are many interpretations and terms to refer to the "social" aspect in Big Data. The most widespread terms so far are Social Big Data (SBD) and Big Social Data (BSD). Various definitions and approaches are presented and compared in the following, in order to outline the existing research directions.

Big Social Data as science: Ishikawa's and Pentland's concepts

Hiroshi Ishikawa is a central adherent of Social Big Data concept, which he described and defined in his book as science of analyzing interconnections between physical world data and social data for the good of public:

"Analyzing both physical real world data (heterogeneous data with implicit semantics such as science data, event data, and transportation data) and social data (social media data with explicit semantics) by relating them to each other, is called Social Big Data science or Social Big Data for short" [22].

It is worth noting that Ishikawa is one among few who provide a proper conceptualization of his ideas and views on the social phenomenon in Big Data. Accordingly, he clarified and supported by arguments relevant related terms, data sources and analytical approaches.

Thus, he defines *social data* as *social media data*, which, in his opinion, is one kind of Big Data with four V's characteristics—*volume*, *variety*, *velocity* and *vague*. While the first three and *veracity* characteristics are already discussed in multiple studies on Big Data [23–26], the *vagueness* first appears in this book as essential characteristic of social data. It should not be mixed with *vagueness* proposed by Venkat Krishnamurthy on Big Data Innovation Summit in Silicon Valley in 2014, which refers to the confusion over the meaning of Big Data [27–29]. According to Ishikawa, vagueness characteristic is a result of a combination of various types of data to be analyzed, which lead to inconsistency and deficiency. It also relates to the issues of privacy and data management as social data involves individuals' personal information.

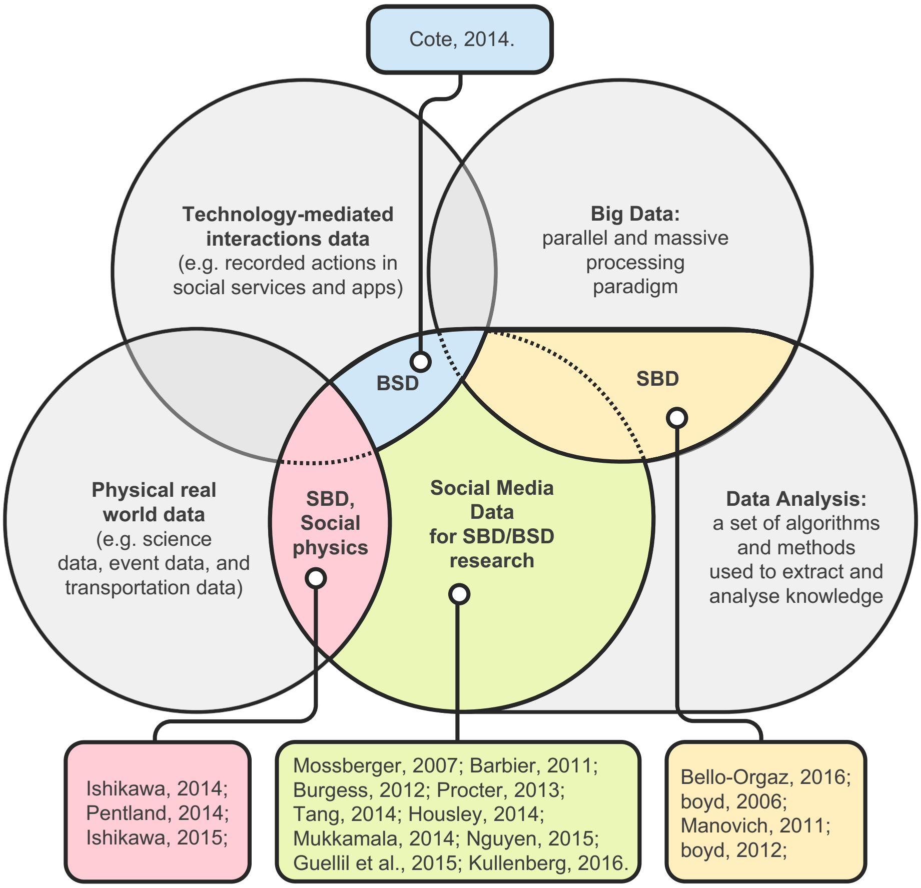
Cote, 2014.
Technology-mediated interactions data (e.g. recorded actions in social services and apps)
Big Data: parallel and massive processing paradigm
BSD
SBD
Physical real world data (e.g. science data, event data, and transportation data)
SBD, Social physics
Social Media Data for SBD/BSD research
Data Analysis: a set of algorithms and methods used to extract and analyse knowledge
Ishikawa, 2014; Pentland, 2014; Ishikawa, 2015;
Mossberger, 2007; Barbier, 2011; Burgess, 2012; Procter, 2013; Tang, 2014; Housley, 2014; Mukkamala, 2014; Nguyen, 2015; Guellil et al., 2015; Kullenberg, 2016.
Bello-Orgaz, 2016; boyd, 2006; Manovich, 2011; boyd, 2012;

Fig. 1 Conceptual map of various BSD/SBD interpretations in the related literature. This illustration depicts four main domains, which were studied by different researchers from various perspectives and intersections of science field/data types

Additionally, Ishikawa classifies the sources of social media data accordingly: *blogging, micro blogging, social network services, sharing* and *video communication services, social news* and *gaming, social search* and *crowd sourcing services*, and *collaboration services*. All data in such services would therefore be regarded as Big Social Data.

Ishikawa is interested in relationships between physical and cyber worlds. He considers SBD should follow the bidirectional analysis that includes influences from the physical real world on social media, and vice versa, in order to develop a complete model (theory). Such theory may explain interactions between both realms and enable potential prediction, recommendation and problem solving. In other words, he suggests tracking social media data and physical world data in order to reveal mutual interdependencies that in turn would result in actual insight. Ishikawa provides an example of traffic authorities predicting public transportation issues in context of massive social events that are actively discussed in social networks, blogs, news, etc. Thus, the data from social media could be analyzed to prevent traffic jams or to increase the amount of public transportation next to the event location.

Ishikawa's thinking is in line with Pentland's concept of social physics [30]. According to Pentland, social physics is the *"quantitative social science that describes reliable,* *mathematical connections between information and idea flow on the one hand and people's behavior on the other".* While Ishikawa aims to bring clarity about analytical techniques for SBD (for example, modeling, data mining, multivariate analysis), Pentland envisions a data-driven society. Even though Pentland does not utilize SBD or BSD terms directly in the conceptualization, he defines Big Data as the engine of social physics. The author refers to the data about human behavior, which consists of both human-generated content (from social media platforms) and data from the physical world (for instance, transactions, locations, call records), which is similar to Ishikawa's vision about social data sources. The main goal of Petland's research is to show how this data together with social science theories could be applied in practical settings.

Data-driven approaches to Big Social Data

Guellil and Boukhalfa consider SBD as a part of social computing [31]. To differentiate their view on SBD from general Big Data, authors provide certain characteristics referring to the research of Tang et al. [32]: *"the set of links (due to relationships between users), a nonstructural nature (due to the length of messages required by some microblogging, the presence of spelling mistakes or other) and the lack of completeness (due to certain user requirements for data privacy)".* Authors provide a classification of the research works on SBD and discuss various analytical approaches and related challenges.

Guellil and Boukhalfa compile their vision of SBD based on the works of Barbier [33], Mukkamala [34] and Nguyen [35]. Notably, Mukkamala and Nguyen utilize SBD and BSD terms interchangeably and mention only social media data as a major data source. Even though Guellil and Boukhalfa point out the inconsistent use of terms in related literature, they do not provide clear conceptualization of the SBD in their own research. In fact, SBD term from the perspective of Guellil and Boukhalfa might be interpreted as a synonym of social media data with qualities such as *large volume*, *noisiness* and *dynamism* that were already revealed earlier in Barbier's work.

From another perspective, Mark Coté makes the attempt to distinguish BSD concept from the broader category of Big Data [36]. In his viewpoint, Big Data is any data produced as the result of the quantification of the world that may include data from sensors, multiple industrial and domestic networks as well as financial markets, whereas BSD *"comes from the mediated communicative practices of our everyday lives, whenever we go online, use our smartphone, use an app or make a purchase."* Moreover, Cote provides reasoning for the importance of BSD. According to him, the concept is not novel, but may significantly affect the media theory. Among those reasons are: the enormous size of data generated by humans that enables endless future analysis; the symbolic nature of social data that is challenging to process even though it is produced in the structured platform spaces; the infrastructure of BSD is very distributed that require scalable computer architecture and network capacity; challenges related to processing, storing, costs and data regulations.

Purpose-driven approaches: Big Social Data for society

Jean Burgess and Axel Bruns discuss Big Data in terms of social media and use the BSD term to refer to this research area [37]. Their vision is based on Manovich's ideology [38], which is focused on bringing the potential of social or cultural data into *humanities* and *social sciences*. Thus, Jean Burgess and Axel Bruns present the BSD concept by mentioning the shift of Big Data towards media, communication, cultural and computational social science, which has led to the wave of research on digital humanities [39–41]. According to Burgess and Bruns, such changes "...provoked in large part by the dramatic quantitative growth and apparently increased cultural importance of social media—hence, "big social data". Their research is aimed to clarify the role of social media in context of the contemporary media ecology with focus on communication, societal events and the nature of human's engagement by applying computational methods towards Twitter archives. Inspired by the Manovich's concept of BSD they trialled the feasibility of research on the phenomenon in order to reveal potential technical, political and epistemological issues. They identified ethical concerns as well as data accessibility, authenticity and reliability challenges. Based on the results, they stated that research on BSD requires the elaboration of mature conceptual models and methodological priorities.

Housley et al. [42] also take a society-oriented view to discuss Big Data. The authors have been conducting observatory research on the opportunities and challenges of open source social media data in the context of social sciences. They seek for the governance and organization improvements through the sense of civil society by means of 'big and broad' social data. According to authors, the term "big and broad" social data refers to three V's (*volume*, *variety*, *velocity*)—already well-known dimensions of related data, which also might be real-time and dynamic. Accordingly, social media could be used to empower people engagement in civil society through a methodological approach to generate sociological insight as proposed in the paper. William Housley et al. characterize digital innovations with qualities such as interaction, participation and "social" that affect complicated relationships between data and analytical capacity, thus enabling participatory infrastructure for public sociology. Consequently, in this regard, the authors point to "citizen social science", which is aimed to assist social scientists by decreasing the challenges of social media data with the help of volunteers among citizens [43]. Such members of public may contribute with research by recording their knowledge, opinions and beliefs, thus connecting the social science academy and society [44, 45].

Big Social Data as method

Bello-Orgaz et al. [46] consider SBD is a combination of Big Data and social media. According to the authors, SBD is needed for analysis of large amount of data from diverse social media sources. They theorize the concept as follows: "Those processes and methods that are designed to provide sensitive and relevant knowledge to any user or company from social media data sources when data sources can be characterized by their different formats and contents, their very large size, and the online or streamed generation of information".

Thus, the conceptual map of SBD from Gema Bello-Orgaz et al. incorporates *Big Data* as processing paradigm, *social media* as the main source of data, and *Data Analysis* as method gaining and analyzing knowledge. Authors revise analytical methodologies for social media as well as new related applications and frameworks.

Summary of the related literature

Even though not all in the above-mentioned papers explicitly use BSD as a term, we consider these works are relevant to the topic. Researchers try to clarify the phenomenon of rapidly growing amount of human-related social data and seek for ways to apply it for the good of the society, data analytics and various fields of science. The key content of the approaches under discussion and theorizations about BSD is summarized in Table 1.

One central commonality among existing research directions is the presence of social media as major data source and orientation towards analytics. The conceptualizations in these scientific articles vary from fundamentally broad (e.g. Ishikawa [22] and Pentland [30]) to vaguely described (e.g., Guelil and Boukhalfa [31]). Additionally, there are only a few attempts to distinguish the concepts from mere Big Data. What is also important, there is lack of clarity regarding the relations between researchers' concepts and related fields: it is hard to outline how other sciences affect the scope of BSD/SBD and directions of studies. Moreover, it is often confusing what data types are considered relevant and valuable for research, and it is hard to understand which data was utilized in the reported research.

We conclude that there are research gaps that researchers of BSD should bridge in order to achieve holistic understanding about the concept of BSD and its characteristics. For example, it is essential to identify the data types that can be explored and studied in this domain. Sophisticated conceptualization and definition of BSD would help researchers build proper methods to process and analyze it. This is essential also because the growth in human-generated data engenders new challenges to solve, requiring novel tools, frameworks and methodological approaches as well as multidisciplinary expertise.

Theoretical foundations of Big Social Data

Based on the literature overview we perceive the concept of BSD as a combination of four fields of science: social computing (including social media and social networks), Big Data science and data analytics as fields that enable and contribute to the existence of the data, and CSS as a field that primarily utilizes the data to gain insight and conduct research (see Fig. 2).

We emphasize that the concept should be understood in an interdisciplinary way in order to open new research avenues. The current and possible roles of each field of science in the context of BSD are discussed in the following.

Social computing

Social computing is a research and application field that integrates social and computational sciences [47]. According to Wang, the theoretical foundations of social computing incorporate Social Psychology, Sociology, Social Network Analysis, Anthropology as well as theories of organization, communication, human–computer interaction and computing theory. In his work, Kling [48] addresses the idea of a mutual interference between communication technologies and society. Therefore, social computing favorably affects both society and technology development: on the one hand enabling smooth socialization and social interactions through various computational systems, and on the other hand, introducing social practices and theories in the development of computational systems and applications. In terms of BSD, *social computing enables services*

Table 1 Summary of BSD-related concepts, types of data they cover as well as challenges and research goals related to the area

Authors	Proposed conceptualization	Data sources	Data challenges/ characteristics	Research goals
Big Social Data as science				
Ishikawa [22]	SBD—science about relationships between physical world data and social media data	Social media data (explicit semantics); physical real world data (implicit semantics)	Volume; variety; velocity; vague	To clarify fundamental conceptualization of SBD and its applications
Pentland [30]	Social physics—quantitative social science about connections between idea/information flow and human behavior	Physical world and social media data that reveal human behavior	Large size; data is ubiquitous and real-time	To conceptualize social physics; To reveal applications of social physics in real-world settings
Data-driven approaches to Big Social Data				
Guelil and Boukhalfa [31]	Concepts from Barbier [33], Mukkamala [34] and Nguyen [35]. SBD is a part of social computing	Social media data	Lack of completeness; large size; dynamic and unstructured	To review research on SBD and classify related literature; to study analytical approaches to SBD
Coté [36]	Concept of data motility under the age of BSD	Data from mediated humans' practices	Volume; symbolic nature of data; distributed infrastructure; lack of regulations	To conceptualize the data motility; to bring conceptual boundaries of BSD
Big Social Data for society				
Burges and Bruns [37]	BSD is social media data	Social media data	Data authenticity, reliability and accessibility issues	To assess feasibility of research in BSD; to reveal potential issues while working with BSD in academia context
Housley et al. [42]	"Big and broad" social data	Social media data	Volume; variety; velocity; real-time; dynamic	To conduct observatory research of "big and broad" social data opportunities and challenges
Big Social Data as method				
Bello-Orgaz et al. [46]	BSD —processes and methods to extract knowledge from social media to users or companies	Social media data	Volume; velocity; variety; value; veracity	To review technologies and applications for processing Big Data from social media

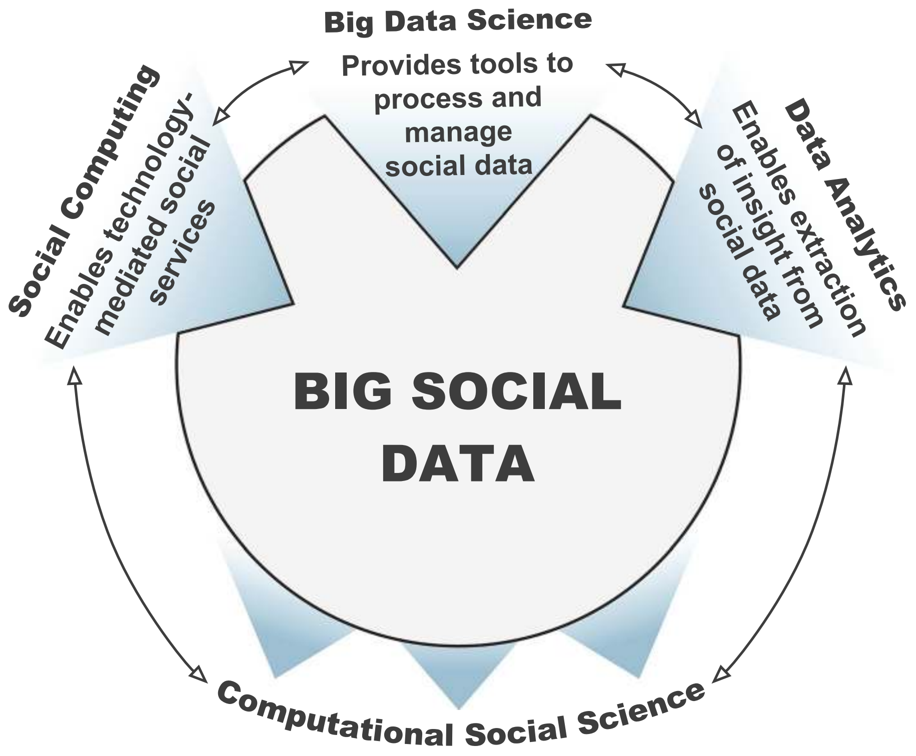
Big Data Science
Provides tools to process and manage social data
Social Computing
Enables technology-mediated social services
Data Analytics
Enables extraction of insight from social data
BIG SOCIAL DATA
Computational Social Science

Fig. 2 Different fields of science contributing to and utilizing Big Social Data as a field. There are four main science fields that contribute to and utilize Big Social Data as a research field—social computing, Big Data science, data analytics and CSS

for technology-mediated self-representation [49] *and communication and supports the building and maintaining of digital relationships through multiple technological infrastructures (for example, Web, database, multimedia and wireless technologies).* In summary, social computing approaches the topic from the perspectives of applications, communication and business.

Big data science

Big data science refers to a field that processes and manages *high-volume*, *high-velocity* and *high-variety* data in order to extract reliable and valuable insights [50]. Big Data is aimed to serve large-scale digital applications and computational systems. Therefore, from BSD perspective, *Big Data science provides solutions to process and manage data originated from technology-mediated social interactions in the context of numerous social services and applications in the digital environment.* There are both optimistic and realistic approaches in regard to recent interest to Big Data technology. One group of researchers (as a rule business-oriented) discusses potential benefits of utilizing Big Data [51, 52] to study massive data about people, things and interactions, while other researchers appeal to critical questions, assumptions and issues that may occur when accessing such data [53–55]. It is crucial to consider a critical view on BSD concept, because data that is primarily related to digital human interactions would definitely cause controversial challenges (for example, data availability, regulations on accessing data, ethics issues, and privacy). In summary, originating from computer science and information systems Big Data is a broader category than BSD, and has mostly data and infrastructure-centric perspective, for instance, with focus on Hadoop, Spark, clusters, and related infrastructural work.

Data analytics

Data analytics allows the extraction of insight or conclusions from existing massive data sets. Generally, it includes *descriptive* (describes data), *exploratory* (discovering unknown correlations in data), *predictive* (predict events and trends) and *prescriptive* (suggest actions) methods to gain meaningful insight for different domains [56, 57]. *Social Network Analysis* (SNA) is one of the most established fields of data analytics [58, 59], providing tools, methods and theories for the research of social networks in the digital realm. Other central areas that can be relevant for BSD include *Business Analytics* [60, 61] and *Sentiment Analytics* [62, 63]. Regardless of the intention and application area of the analysis, data analytics can be said to approach BSD from the perspective of utilization of data (for example, service development, gaining insight, decision making).

Computational social science

Definition of the concept is only one step towards proper understanding of BSD. Duncan Watts claimed the potential of Big Data in social domain—“we finally have our telescope” [64]. However, Macy challenges this statement [65] by referring to Gintis and Helbing [66] who point out that just having a telescope is not enough. “We also need to know where to point it, and for that we need the core analytical toolkit... Big data needs big theory” [65]. In terms of BSD such a pointer or a guide toward the theory and meaningful applications is *CSS* [67]. This multidisciplinary field seeks for theory-grounded models of the social phenomena within the intersection of social and computational sciences [68]. CSS determines a joint collaboration between social, behavioral, cognitive and computer scientists with agent theorists, mathematicians and physicists [69]. According to Conte, CSS is going beyond the traditional social science tools to unravel social complexity from new perspectives more deeply [70]. Author highlights that CSS is not only about variables and equations; the major elements of this science are “people, ideas, human-made artifacts, and their relations within ecosystems”. The theorization and modeling of society by means of computational approaches is aimed to bring comprehension of social complexity and the way social systems operate [71]. Thus, we argue that CSS utilizes BSD in order to “serve the public good and examine the public agenda” [72]. In other words, CSS can reveal the meaningful and relevant areas in utilization of BSD, thus pointing directions for the analysis, making sense of the findings and enabling predictions as well as sensible explanations.

In summary, the aforementioned areas are the central conceptual and theoretical foundations of BSD that contribute to this inter-disciplinary concept. Social computing enables and serves technology-mediated social services and applications that in turn generate vast amount of complex social data; such data are managed and processed through Big Data tools; then insights and prescriptions are derived from data analytics methods and algorithms. CSS is one of the key fields to define targets and reasons for the analysis and explanations for the analysis results.

Our synthesis and definition of Big Social Data

Drawing from our overview of the related literature and observation of contributing science fields we provide a meta-level definition of the synthesized BSD concept as follows:

Big Social Data is any high-volume, high-velocity, high-variety and/or highly semantic data that is generated from technology-mediated social interactions and actions in digital realm, and which can be collected and analyzed to model social interactions and behavior.

This definition approaches the concept from the synthesized perspective including the description of social data characteristics, its sources and origins as well as purpose of use:

Characteristics Shortly speaking, in this context, *volume* refers to the exponential growth of social data. *Variety* relates to various types and forms of social data sources: it might be structured, semi-structured or unstructured. Variety can also mean the difference of formats (for instance, text, image, video). Velocity refers to the fact that social data is generated and distributed with tremendous speed. One can simply count his/her activity in online services per hour to imagine the frequency, with which billions of people right at this moment create or share something online. These characteristics define the size of social data available for the analysis as well as real-time and dynamic nature of BSD. The volume, velocity and variety are traditional characteristics in any Big Data, while *semantic* is a more unique characteristic of BSD. It refers to the fact that all content manually created is highly symbolic with various often-subjective meanings, which require intelligent solutions to be analyzed. There have been studies on mining and analyzing such multimedia data [73–76], however we are still far from the degree of the intelligence, which may turn immense pools of user-generated content into meaningful insights.

Data sources and origins In context of BSD, we consider technology-mediated social interactions as origins of social data types. It refers to *digital self-representation*, *technology-mediated communication* and *digital relationships data* that may appear not only in social networks services but in variety of discussion forums, blogs, web and mobile chat applications, multi-player games as well as different web sites that are not for social purposes per se.

Purpose Analyzing and modeling social interactions and behavior means that researchers may use the data to describe, understand, and build models of digital interactions taking place between people and how people act (online) around these interactions (for example, profile building, self-expression and other activities that are not directly seen as interaction but, rather, necessary prerequisites for it). The knowledge, which is gained from analysis, may then be utilized in variety of applications, meaning that BSD practitioners are free to choose which domain or research question to address. For instance, researchers may aim to solve fundamental societal issues or just explore tweets for the sake of testing new semantic algorithms.

The definition is further explicated in the following subsection with the classification of data types that relate to technology-mediated social interactions.

Types of Big Social Data

We emphasize that a central element of the BSD concept is "digital human", who uses Information and Communications Technology (ICT) for digital social interactions. The rapid evolution of ICT has shifted the role of a user from a consumer to the active producer and mediator of information, thus allowing people to control, personalize and apply the digital realm according to their values, social needs and preferences [70]. We incorporate the term of "digital human" to underline the shift towards new sociality that lives in hybrid reality [77], where the dynamism and constant availability of technology-mediated communication blurs the boundaries between reality and virtuality. Thus, people do not distinct their activity in online and physical environments, because of "always-on" social networking. Similarly, Wooglar suggests the term of "virtual community" and states that it is just the matter of choosing words: *"In this usage, 'virtual,' like 'interactive,' 'information,' 'global,' 'remote,' 'distance,' 'digital,' 'electronic' (or 'e-'), 'cyber-,' 'network,' 'tele-,' and so on, appears as an epithet applied to various existing activities and social institutions".* [78].

Around digital human interactions, there are both machine-generated and human-generated data that potentially might turn into the social insight. However, in this paper we argue that exactly human-generated data makes BSD concept unique and distinguishes it from general field of Big Data. While machine-generated data could be analyzed through mere Big Data tools and applications, human-generated content requires more intelligent solutions to decode the semantics of people's beliefs, opinions and behavior. Undoubtedly, Big Data may show what and how is changing in social interactions, however it does not answer the question of why those changes and processes are happening. Therefore, we consider BSD is the solution to properly investigate the semantics of human-generated content. From our perspective, it may provide to practitioners of many research fields both facts and reasoning.

While discussing human-generated data we mean content that is produced through social technology-mediated interactions of people in social media platforms. This category may contain digital-self representation data, technology-mediated communication data and digital relationships data (see Fig. 3). These three categories define the types of data that could be interpreted and utilized as social data in the current digital environment (see Table 2). In other words, Table 2 serves as a simplified taxonomy of BSD; however, it is not meant as an extensive index of what data is BSD but, rather, as a list of currently existing BSD examples that could be available for research and analysis.

Digital self-representation

The first category to be discussed is digital self-representation. This is the initial step for "digital humans" to socialize and communicate themselves in the digital realm. These data types relate to numerous virtual profiles that have functions of identity depiction and communicative body [49]. In other words, the data is meant to reveal some information (a "face") for other users in the particular digital service. Albrechstlund proposes a concept of "sharing yourself", which is related to the way constructed identity is participating in social networks creating relations with others [79]. In digital environment people are limited in verbal and non-verbal impressions compensating it by means of text, pictures, videos and music that could be placed in the following data categories:

1. *Profile data* It includes login data (usually a name/nickname/e-mail address with which other people identify the user); identity data (depends on the digital environment, i.e. for some services one should provide real first name and last name, mobile

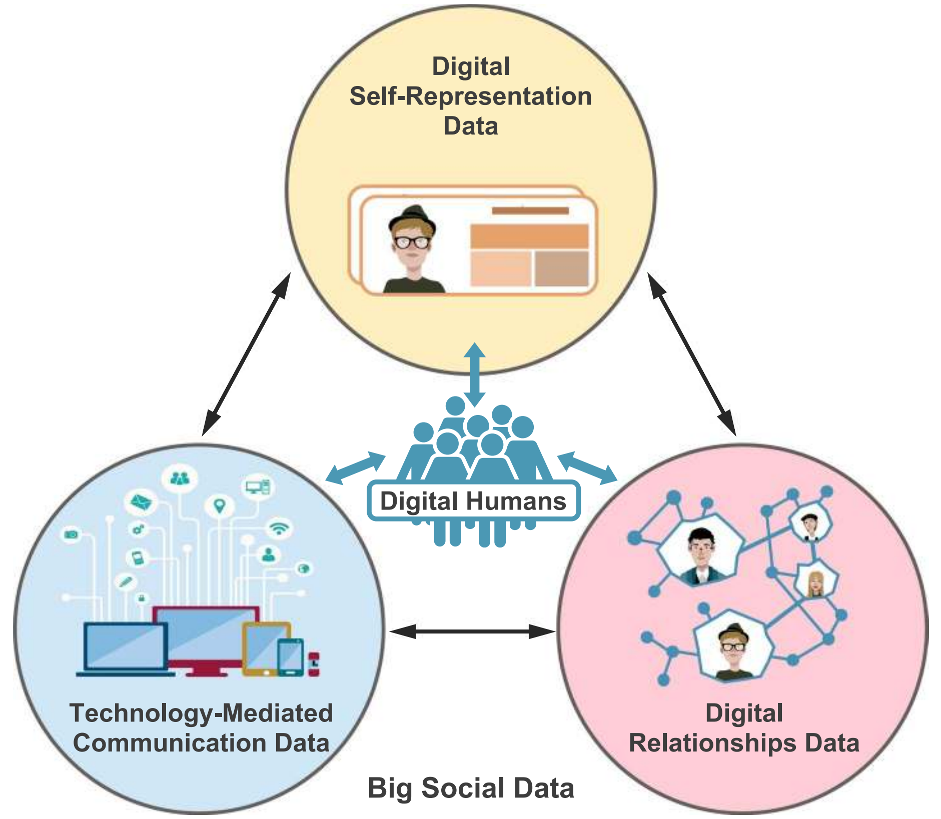
Digital Self-Representation Data
Digital Humans
Technology-Mediated Communication Data
Digital Relationships Data
Big Social Data

Fig. 3 Overview of BSD types and sources. There are three major data types of Big Social Data—technology-mediated communication data, digital self-representation data and digital relationships data

phone number, country, education, birthday); and personality data (e.g., profile pictures, tags of interest, slogan, personal signature in discussion forums) In many social media services, it is the personality data that the other users particularly focus on and analyze to assess the interestingness of the user.

2. *Self-published content* It incorporates publicly disclosed or socially restricted data (to trusted users or specific communities), such as most status updates in social media, pictures, videos, and other content that people add to services to represent themselves.
3. *Data published by the community* Self-representation could be complemented through person-related content shared by other users. This refers to collaboratively created pictures, narrations, videos, etc.

Technology-mediated communication data

Technology-mediated communication data refers to the data generated in two-way communication, collaborative knowledge creation and information distribution in the context of digital environment—the content and subjects of the communication. Technology mediates the constructed digital self-representation to contribute information, edit existing contributions, comment on entries and discuss related matters. From the fundamental perspective digital environments allow people to contribute to knowledge

Table 2 Classification of Big Social Data types

Category	Definition	Types of data
Digital self-representation data	Data related to *identity depiction* and *communicative body* in digital environment	*Profile data* (i) Login data (name/nickname/e-mail address and password); (ii) Mandatory data (services and application required data, for example, full name, citizenship, birthday); (iii) Extended data (profile pictures, education, tags of interests) *Self-published content* (e.g., personal documents, pictures, videos, interests): (i) Disclosed data (to the public); (ii) Entrusted data (content sharing within trusted digital community) *Data published by the community* (e.g., pictures, narrations, videos, posts): Relates to content shared by other users, which contribute to the digital identity creation
Technology-mediated communication data	Data related to two-way communication, knowledge creation and distribution through technology	*Private communication data* instant 1-to-1 messaging and content sharing; *Public communication data* 1-to-many messaging, commenting, information contribution and editing of existing entries; *Collaborative communication data* many-to-many participatory content sharing, chats, video-conferences
Digital relationships data	Data that reveal digital social relationships patterns	*Explicit data* Friendship data–Followee/Follower data; *Implicit data* Data, which is revealed through technology-mediated communication data (e.g., tweets could be analyzed to infer connections between people)

creation and distribution through various digital devices [80]. Digital environment facilitates physical communication channels resulting in *private communication* (i.e., one-to-one), *public communication* (one-to-many) and *collaborative communication* (many-to-many) data. Depending on the context, public and collaborative communication could also be private within the group of participants, i.e. in case it is a private channel of the organization.

Digital relationships data

Digital Relationships data describes the explicit connections and ties between users in the services. Analysis of this data can reveal social relationship patterns, social network structures and various other sociological and network level phenomena in the digital realm. Digital Representation category firstly contains explicit data, which refers to digital friendships and followership that a user has intentionally and explicitly defined. Technology-mediated social services provide the possibility to build virtual communities based on both physical and online activities (to create networks based on existing connections in physical world and/or create new networks with people from digital realm). There are two roles for users of such services—to be followee and follower. One could have followers or friends on various social platforms (Facebook, LinkedIn, Twitter, Instagram, and many others), and in turn could follow someone to maintain friendships, business relationships or track important content of another relevant user. An interesting factor to be researched is the motivation of people adding someone to the friend's lists. Obviously such lists incorporate friends and colleagues, but also there could be public figures, interesting strangers or people with weak ties [81–83]. There is also implicit data, which could be revealed through analysis of technology-mediated communication data. For instance, tweets can be analyzed to infer individual connections between people. And from these individual connections, we can build network representations of communities in system level. As another example, two users having multiple common contacts (e.g., friend-of-a-friend) can be predicted to become explicit contacts in the future. When a user has, for example, liked or otherwise interacted with a non-contact user's content or profile, there can be seen to be an implicit tie between the users [82]. However, such implicit data normally requires network analysis to be created, and there are few tools or methods to provide such data automatically.

To summarize, we consider this list of BSD types could be valuable for researchers to outline the scope of their interests and will guide them to achieve successful outcomes. Nevertheless, research community has to remember that the accessibility of such data is a crucial challenge of BSD. Lack of access to the data often held by various service providers hinders the utilization of and research opportunities related to this emerging concept. Thus, researches should search for ways of collaboration with social media platforms.

Future work

The holistic overview of related concepts, research fields as well as research communities provide ideas regarding methodological steps that should be taken to enable further research and utilization activities around BSD. This is a combination of three activities that should be primarily focused on in order to open new avenues for the utilization.

1. *Collecting data* The initial step for all researchers who work with BSD is to collect needed datasets for analysis. This step brings up the ethical issues and challenges of data accessibility. Indeed, there are challenges in terms of accessing the data as it is often held by various service providers, which hinders the utilization of the data. Manovich notes this by stating 'only social media companies have access to really large social data' [38]. Fortunately, recently we have seen various movements and joint efforts for bringing together data that, in theory, is public but very challenging to collect in high volume enough for research purposes (for example, the OSoMe[1] project to help analyzing Twitter data). One of the most troubling issues is related to ethics: majority of people are not aware about their data being collected and analyzed by different organizations (including government and social media companies). Moreover, the regulations on accessing and usage of such data are not clear and not completely unified. There are also challenges that may cause privacy violation: collecting more private data than allowed; accessing data without permissions; utilizing data for purposes, which are different from the initial purpose of collecting the data; misinterpreting the data; and changing the content. To make collecting phase feasible we need to fulfill the next step of our framework.
2. *Collaboration* BSD is multidisciplinary area that will require practitioners to build a proper team for work. Our suggestion is to build collaboration with social media platforms or companies that have access to actually large data sets. For instance, the

[1] Observatory on social media (OSoMe) project to study diffusion of information online and discriminate among mechanisms that drive the spread of memes on social media—http://truthy.indiana.edu/about/.

research outcomes from thousands of twits would be questionable in comparison with research under billions of human-generated content from multiple channels. Collaboration with people or companies with various expertize and advantages in terms of social data availability will potentially reduce challenges with collecting data for one's own study, extend the scale and scope of the work in a positive way as well as provide access to multidisciplinary expertise.

3. *Manipulating data* We argue that for gaining meaningful insights from BSD, researchers should design virtual environments where they would be able to access multiple data types, to compare and control them. It may bring new opportunities for authentic and reliable research outcomes. In this regard we agree with Watts [68] that we need *'social supercollider'*, which will obtain diverse social data streams thus opening access to knowledge about people's behavior on the massive scale. BSD artificial environments also could give opportunity to run virtual experiments and validate results with members of related research community.

This paper was aimed to bring clarity on BSD topic in general for any application area. As for our intended future work, we aim to utilize BSD to foster serendipity and, thus, innovativeness in knowledge work organizations. Our objective is to obtain empirical evidence that analysis of BSD can help identify relevant new people to collaborate with.

Conclusion

The multidisciplinary and multi-dimensional nature of Big Social Data brings challenges to the development of a useful conceptualization and definition of the concept. Our literature overview shows that majority of related work on BSD is focused on the analysis of social data, giving less attention to describing what BSD actually is. This can lead to lack of consensus, inconsistency, and vague understanding of what such data could be used for. To bring clarity and sophisticated understanding of BSD we propose a synthesized conceptualization and definition of the concept and this growing field. We reviewed existing literature that demonstrates a variety of applications and approaches to study the phenomena around social data. Based on this we outlined the fields of science that determine the scope of BSD (social computing, Big Data science, data analytics and CSS). We assume the knowledge about the involvement of each field would provide researches with the understanding of the expertise that is demanded for conducting research in this field. Additionally, we proposed the classification of BSD types that, from our perspective, well cover the spectrum of data that BSD consists of. In summary, with this paper, we aim to make researchers more informed about what is BSD, on what data to focus as well as motivate them to elaborate better conceptualization, in order to reach clear desirable research outcomes.

Authors' contributions

EO performed the primary literature review and analysis for this work as well as designed graphics. Manuscript was drafted by EO, TO and JH. EO introduced this topic to other authors and coordinate the work process to complete the manuscript. EO, TO, JH and HK worked together to develop the article's framework and focus. All authors read and approved the final manuscript.

Author details

[1] Department of Pervasive Computing, Tampere University of Technology, Korkeakoulunkatu 10, 33720 Tampere, Finland. [2] Department of Mathematics, Tampere University of Technology, Korkeakoulunkatu 10, 33720 Tampere, Finland. [3] NOVI research group, Department of Information Management and Logistics, Tampere University of Technology, Korkeakoulunkatu 10, 33720 Tampere, Finland.

Acknowledgements
We thank all members of the COBWEB project.

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported by the Academy of Finland project 295893, 295894, 295895— "Enhancing Knowledge Work and Co-creation with Analysis of Weak Ties in Online Services (COBWEB)".

Received: 1 November 2016 Accepted: 12 January 2017
Published online: 25 January 2017

References
1. Belsey B. Cyberbullying: an emerging threat to the "always on" generation. Recuperado el. 2005; 5. Retrieved from http://www.cyberbullying.ca/pdf/Cyberbullying_Article_by_Bill_Belsey.pdf. Accessed 15 Oct 2016.
2. Katz JE. Handbook of mobile communication studies. London: The MIT Press; 2008.
3. Mandiberg M. The social media reader. New York: NYU Press, New York University; 2012.
4. Monash C. Three broad categories of data. 2010. http://www.dbms2.com/2010/01/17/three-broad-categories-of-data/. Accessed 15 Oct 2016.
5. Chen W. How to tame big bad data. 2010. http://blog.magnitudesoftware.com/2010/08/25/tame-big-bad-data/. Accessed 15 Oct 2016.
6. Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manag. 2015;35(2):137–44.
7. Marwick AE. Status update: celebrity, publicity, and branding in the social media age. New Haven, USA: Yale University Press; 2015.
8. Freire FC. Online digital social tools for professional self-promotion. A state of the art review. Revista Latina de Comunicación Social. 2015;70:288–99.
9. Shih C. The facebook era: tapping online social networks to build better products, reach new audiences, and sell more stuff. Upper Saddle River: Prentice Hall; 2009.
10. Stephen AT, Toubia O. Deriving value from social commerce networks. J Mark Res. 2010;47(2):215–28.
11. Musacchio M, Panizzon R, Zhang X, Zorzi V. A linguistically-driven methodology for detecting impending disasters and un-folding emergencies from social media messages. In: proceedings of LREC 2016 workshop. EMOT: emotions, metaphors, ontology and terminology during disasters; 2016. p. 26–33.
12. Aradau C, Blanke T. Politics of prediction: security and the time/space of governmentality in the age of big data. European Journal of Social Theory. 2016:1–19. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/1368431016667623. Accessed 15 Oct 2016.
13. Saldana-Perez AMM, Moreno-Ibarra M. Traffic analysis based on short texts from social media. Int J Knowl Soc Res. 2016;7(1):63–79.
14. Qualman E. Socialnomics: how social media transforms the way we live and do business. Hoboken: Wiley; 2010.
15. Kennedy H. Commercial mediations of social media data. London: Springer; 2016. p. 99–127.
16. Golbeck J, Robles C, Turner K. Predicting personality with social media. In: CHI'11 Extended abstracts on human factors in computing systems. Vancouver: ACM; 2011. p. 253–62.
17. Power DJ, Phillips-Wren G. Impact of social media and Web 2.0 on decision-making. J Decis Syst. 2011;20(3):249–61.
18. Golbeck J. Big social data predicting the future of you. Executive Tallent Mag. 2014;5:12–4.
19. Cambria E, Rajagopal D, Olsher D, Das D. Big social data analysis. In: Akerkar R, editor. Big Data Computing. Boca Raton, Florida: Chapman and Hall/CRC; 2013. p. 401–14.
20. Bravo-Marquez F, Mendoza M, Poblete B. Meta-level sentiment models for big social data analysis. Knowl Based Syst. 2014;69:86–99.
21. Pandarachalil R, Sendhilkumar S, Mahalakshmi G. Twitter sentiment analysis for large-scale data: an unsupervised approach. Cogn Comput. 2015;7(2):254–62.
22. Ishikawa H. Social big data mining. Boca Raton: Taylor & Francis Group, CRC Press; 2015.
23. Sicular S. Gartner's big data definition consists of three parts, not to be confused with three "V's", vol. 27. Stanford: Gartner, Inc; 2013.
24. Kaisler S, Armour F, Espinosa JA, Money W. Big data: issues and challenges moving forward. In: 2013 46th Hawaii international conference on system sciences (HICSS). New York: IEEE; 2013. p. 995–1004.
25. Tole AA, et al. Big data challenges. Database Syst J. 2013;4(3):31–40.
26. Chen M, Mao S, Zhang Y, Leung VC. Big data: related technologies, challenges and future prospects. In: Springer-briefs in computer science. Cham: Springer; 2014.
27. Borne K. Top 10 big data challenges—a serious Look at 10 big data V's. 2014. https://www.mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs. Accessed 15 Oct 2016.
28. Kehoe M. What does it take to qualify as 'big data'? 2014. http://www.enterprisesearchblog.com/2014/07/hadoop-salvation-or-hype.html. Accessed 15 Oct 2016.
29. Moorthy J, Lahiri R, Biswas N, Sanyal D, Ranjan J, Nanath K, Ghosh P. Big data: prospects and challenges. J Decis Makers. 2015;40:74–96.
30. Pentland A. Social physics: how good ideas spread-the lessons from a new science. New York: The Penguin Press, Penguin Group; 2014.

31. Guellil I, Boukhalfa K. Social big data mining: a survey focused on opinion mining and sentiments analysis. In: 2015 12th international symposium on programming and systems (ISPS). New York: IEEE; 2015. p. 1–10.
32. Tang J, Chang Y, Liu H. Mining social media with social theories: a survey. ACM SIGKDD Explor Newsl. 2014;15(2):20–9.
33. Barbier G, Liu H. Data mining in social media. Berlin: Springer; 2011. p. 327–52.
34. Mukkamala RR, Hussain A, Vatrapu R. Fuzzy-set based sentiment analysis of big social data. In: Enterprise distributed object computing conference (EDOC), 2014 IEEE 18th international. New York: IEEE; 2014. p. 71–80.
35. Nguyen DT, Hwang D, Jung JJ. Time–frequency social data analytics for understanding social big data. Cham: Springer; 2015.
36. Coté M. Data motility: the materiality of big social data. Cult Stud Rev. 2014;20(1):121.
37. Burgess J, Bruns A. Twitter archives and the challenges of "big social data" for media and communication research. M/C J. 2012;15(5):1–7.
38. Manovich L. Trending: the promises and the challenges of big social data. Debates Digit Humanit. 2011;2:460–75.
39. Berry D. Understanding digital humanities. London: Palgrave Macmillan, Springer Nature; 2012.
40. Kaplan F. A map for big data research in digital humanities. Front Digit Humanit. 2015;2:1.
41. Svensson P. Big digital humanities: imagining a meeting place for the humanities and the digital. Ann Arbor: University of Michigan Press; 2016.
42. Housley W, Procter R, Edwards A, Burnap P, Williams M, Sloan L, Rana O, Morgan J, Voss A, Greenhill A. Big and broad social data and the sociological imagination: a collaborative response. Big Data Soc. 2014;1(2):2053951714545135.
43. Procter R, Housley W, Williams M, Edwards A, Burnap P, Morgan J, Rana O, Klein E, Taylor M, Voss A, Choi C, Mavros P, Hudson Smith A, Thelwall M, Ferne T, greenhill A. Enabling social media research through citizen social science. In: Korn M, Colomnbino T, Lewkowicz M (eds) ECSCW 2013 Adjunct Proceedings, 13th european conference on computer supported cooperative work, 21–25 September 2013, Paphos, Cyprus
44. Mossberger K, Tolbert CJ, McNeal RS. Digital citizenship: the internet, society, and participation. London: MIt Press; 2007.
45. Kullenberg C, Kasperowski D. What is citizen science? A scientometric meta-analysis. PLoS One. 2016;11(1):0147152.
46. Bello-Orgaz G, Jung JJ, Camacho D. Social big data: recent achievements and new challenges. Inf Fusion. 2016;28:45–59.
47. Wang F-Y, Carley KM, Zeng D, Mao W. Social computing: from social informatics to social intelligence. IEEE Intell Syst. 2007;22(2):79–83.
48. Kling R. What is social informatics and why does it matter? Inf Soc. 2007;23(4):205–20.
49. Boyd D, Heer J. Profiles as conversation: networked identity performance on friendster. In: Proceedings of the 39th annual Hawaii international conference on system sciences (HICSS'06), vol. 3. New York: IEEE; 2006. p. 59.
50. Demchenko Y, De Laat C, Membrey P. Defining architecture components of the big data ecosystem. In: 2014 international conference on collaboration technologies and systems (CTS). New York: IEEE. 2014. p. 104–12.
51. Beyer MA, Laney D. The importance of 'big data': a definition. Stamford: Gartner; 2012. p. 2014–8.
52. James M, Michael C, Brad B, Jacques B, Richard D, Charles R, Angela H. Big data: the next frontier for innovation, competition, and productivity. New York: The McKinsey Global Institute; 2011.
53. Boyd D, Crawford K. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf Commun Soc. 2012;15(5):662–79.
54. Akerkar R. Big data computing. Boca Raton: CRC Press, Taylor & Francis Group; 2013.
55. Vis F. A critical reflection on big data: considering APIs, researchers and tools as data makers. First Monday. 2013;18(10). Retrieved from http://ojs-prod-lib.cc.uic.edu/ojs/index.php/fm/article/view/4878/3755. Accessed 15 Oct 2016.
56. Davenport T. Analytics 3.0: In the new era, big data will power consumers products and services. Brighton, MA: Harvard Business Review. Retrieved from https://hbr.org/2013/12/analytics-30. 2013. Accessed 15 Oct 2016.
57. Bendoly E. Fit, bias, and enacted sensemaking in data visualization: frameworks for continuous development in operations and supply chain management analytics. J Bus Logist. 2016;37(1):6–17.
58. Wasserman S, Faust K. Social network analysis: methods and applications, vol. 8. Cambridge: Cambridge University Press; 1994.
59. Easley D, Kleinberg J. Networks, crowds, and markets: reasoning about a highly connected world. Cambridge: Cambridge University Press, University of Cambridge; 2010.
60. Phillips-Wren G, Iyer LS, Kulkarni U, Ariyachandra T. Business analytics in the context of big data. Commun Assoc Inf Syst. 2015;37:448–72.
61. Duan L, Xiong Y. Big data analytics and business analytics. J Manag Anal. 2015;2(1):1–21.
62. Chen C, Chen F, Cao D, Ji R. A cross-media sentiment analytics platform for microblog. In: Proceedings of the 23rd ACM international conference on multimedia. New York City: ACM; 2015. p. 767–9.
63. Boumaiza AD. A survey on sentiment analysis and visualization. In: Qatar foundation annual research conference proceedings, vol. 2016. Doha: HBKU Press Qatar; 2016. p. 1203.
64. Watts DJ. Everything is obvious: how common sense fails us. New York: Crown Business, Crown Publishing group; 2011.
65. Macy MW, et al. Big theory: a trojan horse for economics? Rev Behav Econ. 2015;2(1–2):161–6.
66. Gintis H, Helbing D, Durkheim E, King ML, Smith A. Homo socialis: an analytical core for sociological theory. Rev Behav Econ. 2015;2(1–2):1–59.
67. Lazer D, Friedman A. The network structure of exploration and exploitation. Adm Sci Q. 2007;52(4):667–94.
68. Watts DJ. Computational social science: exciting progress and future directions. Bridge Front Eng. 2013;43(4):5–10.
69. Wallach H. Computational social science: Toward a collaborative future. In: Alvarez RM, editor. Computational social science: Discovery and prediction. USA: Cambridge Universisty Press; 2016. p. 307–16.
70. Conte R, Gilbert N, Bonelli G, Cioffi-Revilla C, Deffuant G, Kertesz J, Loreto V, Moat S, Nadal J-P, Sanchez A, et al. Manifesto of computational social science. Eur Phys J Spec Top. 2012;214(1):325–46.
71. Cioffi-Revilla C. Introduction to computational social science: principles and applications. London: Springer; 2013.

72. Shah DV, Cappella JN, Neuman WR. Big data, digital media, and computational social science possibilities and perils. Ann Am Acad Political Soc Sci. 2015;659(1):6–13.
73. Zhu X, Wu X, Elmagarmid AK, Feng Z, Wu L. Video data mining: semantic indexing and event detection from the association perspective. IEEE Trans Knowl Data Eng. 2005;17(5):665–77.
74. Wu P, Hoi SCH, Zhao P, He Y. Mining social images with distance metric learning for automated image tagging. In: Proceedings of the fourth ACM international conference on web search and data mining. New York City: ACM; 2011. p. 197–206.
75. Hu X, Liu H. Text analytics in social media. New York: Springer; 2012. p. 385–414.
76. Naaman M. Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications. Multimed Tools Appl. 2012;56(1):9–34.
77. e Silva ADS. From cyber to hybrid mobile technologies as interfaces of hybrid spaces. Space Cult. 2006;9(3):261–78.
78. Woolgar S. Virtual society? Technology, cyberbole reality. New York: Oxford University Press; 2002.
79. Albrechtslund A. Online social networking as participatory surveillance. First Monday. 2008;13(3). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/2142/1949.. Accessed 15 Oct 2016.
80. Ruppert E, Law J, Savage M. Reassembling social science methods: the challenge of digital devices. Theory Cult Soc. 2013;30(4):22–46.
81. Granovetter MS. The strength of weak ties. Am J Sociology. 1973;78(6):1360–80.
82. Gilbert E, Karahalios K. Predicting tie strength with social media. In: Proceedings of the SIGCHI conference on human factors in computing systems. New York City: ACM; 2009. p. 211–20.
83. Haythornthwaite C. Strong, weak, and latent ties and the impact of new media. Inf Soc. 2002;18(5):385–401.