van Altena *et al. J Big Data (2016) 3:23*
DOI 10.1186/s40537-016-0057-0

Journal of Big Data

RESEARCH

Open Access

Understanding big data themes from scientific biomedical literature through topic modeling

Allard J. van Altena*, Perry D. Moerland, Aeilko H. Zwinderman and Sílvia D. Olabarriaga

*Correspondence: a.j.vanaltena@amc.uva.nl
Department of Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center of the University of Amsterdam, 1105 AZ Amsterdam, The Netherlands

Abstract
Nowadays, big data is a key component in (bio)medical research. However, the meaning of the term is subject to a wide array of opinions, without a formal definition. This hampers communication and leads to missed opportunities. For example, in the (bio)medical field we have observed many different interpretations, some of which have a negative connotation, impeding exploitation of big data approaches. In this paper we pursue a better understanding of the term big data through a data-driven systematic approach using text analysis of scientific (bio)medical literature. We attempt to find how existing big data definitions are expressed within the chosen application domain. We build upon findings of previous qualitative research by De Mauro et al. (Lib Rev 65: 122–135, 14), which analysed fifteen definitions and identified four key big data themes (i.e., information, methods, technology, and impact). We have revisited these and other definitions of big data, and consolidated them into eight additional themes, resulting in a total of twelve themes. The corpus was composed of paper abstracts extracted from (bio)medical literature databases, searching for 'big data'. After text pre-processing and parameter selection, topic modelling was applied with 25 topics. The resulting top-20 words per topic were annotated with the twelve big data themes by seven observers. The analysis of these annotations show that the themes proposed by De Mauro et al. are strongly expressed in the corpus. Furthermore, several of the most popular big data V's (i.e., volume, velocity, and value) also have a relatively high presence. Other V's introduced more recently (e.g. variability) were however hardly found in the 25 topics. These findings show that the current understanding of big data within the (bio)medical domain is in agreement with more general definitions of the term.

Keywords: Text mining, Topic modelling, Big data, Biomedical research

Background

The usage of the term 'big data' has picked up since 2011. This was the year that Gartner introduced "Big Data and Extreme Information Processing and Management" in its hype cycle [1]. Furthermore, increased interest is visible in the ever growing search traffic shown by Google Trends [2]. Scientific publications in (bio)medicine, which are our main interest in this study, also show a massive increase in the number of papers published yearly that mention big data [3].

Still, in spite of the popularity of this term, there is much debate about the definition of big data. In 2001 Gartner (called "META Group" at the time [4]) published a report which in hindsight is often referred to as the first description of big data. It defines the term through information technology (IT) challenges described by three V's: volume, velocity, and variety [5].

Over the years this has evolved into many interpretations. Mostly, companies define big data in the light of their prime business, meaning that Google will mention analysis (e.g., Google Flu), while Oracle emphasises volume and storage [6], and IBM or Microsoft focus on computation and usability [7]. In a web-blog, posted on the data science sub-domain of the Berkeley school of information, 43 'thought leaders' from the industry were asked for their definition of big data [8]. Not many of these leaders agreed with each other and definitions range from "data that cannot fit easily into a standard relational database" to "big data is not all about volume, it is more about combining different data sets and to analyse it in real-time to get insights for your organisation". On a governmental level, the US National Institute of Standards and Technology (NIST) defined big data in 2014 as the need for scalable technology and four V's: volume, velocity, variety, and variability. Finally, in the scientific domain, big data is mostly understood as the challenges of working with large volumes of data [9–11].

Possibly due to this great variety of definitions, in practice we have observed many different interpretations of the term big data among (bio)medical scientists. Some understand big data as a positive development, and actively pursue usage of new methods and technology associated with the term [3]. Others, however, view it as a harmful influence on, for example, the strength of research evidence, preferring classical statistical methods [12]. A better understanding of big data would facilitate communication and clarify expectations regarding this overloaded term [13].

Some researchers have attempted to capture comprehensive definitions of big data, such as De Mauro et al. [14], Ward and Barker [15], and Andreu-Perez et al. [3]. The first two focus on no domain in particular, whereas Andreu-Perez et al. [3] focuses on health-oriented applications. Of particular interest is the work by De Mauro et al. which analysis various big data definitions and from these distil their own. Their proposed definition is based on four themes found in the underlying definitions that were gathered, namely information, methods, technology, and impact. Note that all the cases mentioned above are based on qualitative literature studies. Hansmann and Niemeyer [16], however, used text mining to understand the themes included in big data literature. They combined automatic and manual approaches to identify three themes: IT infrastructure, methods, and data. While these efforts have been valuable for a better understanding of the term big data, they do not present systematic evidence of the actual themes used in the scientific literature, in particular for the (bio)medical research domain.

In this paper we present our efforts to answer the following research question: Which themes from various existing big data definitions are expressed in (bio)medical scientific publications? For this purpose, we adopted a data-driven systematic approach. First, big data definitions were revised and 12 themes were identified. Then, (bio)medical literature was systematically gathered from two scientific databases (i.e., PubMed and PubMed Central) and analysed automatically with text mining. While there are many text mining and clustering methods, we chose topic modelling (TM, [17, 18]) because this method captures two aspects that are important for this dataset: words may have multiple meanings or interpretations and documents may contain one or more topics. The topics identified through TM were annotated with the 12 themes by a small group of observers. In the following sections we detail the methods, present the results and discuss our findings.

Methods

In this section the construction of the corpus is described, followed by an explanation of the concepts behind TM. Then the application of TM to the corpus is presented in three steps: pre-processing, model fitting, and post-processing. Finally we present the gathering and summary of existing big data definitions, and the process used to identify them in the topics determined by TM.

Corpus

The corpus of documents was created by querying two literature databases focused on (bio)medical publications: PubMed and PubMed Central (PMC). The search queries were as follows:

- **PubMed**: "big data"[TIAB] OR (big[TIAB] AND "health data"[TIAB]) OR "large data" [TI];
- **PMC**: "big data"[TI] OR "big data"[AB] OR (big[TI] AND "health data"[TI]) OR (big[AB] AND "health data"[AB]) OR "large data" [TI].

Each query was built to search for literal use of the term 'big data', therefore selecting documents that were self-identified with big data. No word spacing was allowed to minimise the amount of irrelevant results. The terms 'big health data' and 'large data' were added because they also retrieved relevant literature, especially for publications before 2011, when the term big data was not popular yet.

Titles and abstracts were exported from the databases and merged into a local repository for further processing. Based on the title (stripped of all special characters and spaces) or the digital object identifier (DOI, if available), duplicates were removed from the corpus. Lastly, any record with an empty abstract (i.e., not provided in the database) was also removed from the corpus.

Topic modelling concepts

A specific type of TM was chosen, namely Latent Dirichlet Allocation (LDA) [17]. Throughout this paper the abbreviations TM and LDA are used interchangeably to indicate topic modelling through the application of LDA. The concept of TM is captured in Fig. 1 using the plate notation [17–19]. Plate D denotes the set of documents, while $\theta^{(d)}$ is the multinomial distribution over topics for document d. Plate $N_{(d)}$ denotes the set of words w for a specific document d, while z is the topic to which word w is assigned. Lastly, plate T denotes the set of topics where $\phi^{(z)}$ is the multinomial distribution over words for topic z.

In TM, θ, ϕ, and z are the latent variables that have to be estimated. Together with the Dirichlet distributed hyperparameters α and β, the model is called Latent Dirichlet

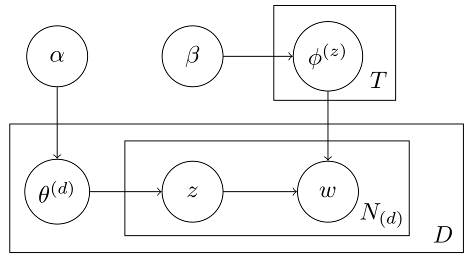

Fig. 1 Plate notation of topic modelling, plates are shown as *rectangles* and the *arrows* are conditional dependencies. Shows the relations between known variables (documents D, number of words $N_{(d)}$, and words w), latent variables (multinomial distributions $\theta^{(d)}$ and $\phi^{(z)}$, and word to topic assignment z), and hyperparameters (α and β)

Allocation [17, 19]. The hyperparameters α and β should be interpreted as smoothing factors for respectively topic-to-document (θ) and word-to-topic (ϕ) assignments.

Topic modelling implementation

The statistical software R [20] was used to implement the pre-processing, TM fitting, model selection, and post-processing steps.

Pre-processing was executed using the R **tm** and **quanteda** packages [21, 22]. Processing consisted of removing stop words taken from the SMART list [23, 24] (e.g., about, the, which).[1] Extra stop words were added, which were either junk words resulting from processing steps, or terms that appeared very often and diluted the TM outcome, such as 'big data', 'introduction' and 'discussion'.[2] From the remaining words, bi-grams were created with function **dfm**: two words that occur next to each other at least 15 times in the whole corpus are joined by an underscore (e.g., health_care). Furthermore, words were stemmed with function **stemDocument**; e.g., 'develop', 'developed', and 'development' were all stemmed to 'develop'. Lastly, words longer than 26 characters were removed.

Fitting the model consisted of estimating the latent variables θ, ϕ and z, which was done with the R **topicmodels** package [26]. Directly calculating θ and ϕ was shown to be suboptimal [19], therefore we used a Bayesian approach from the **topicmodels** package using Gibbs iterative sampling to approximate the distribution z. In this sampling process the probability of a word occurring in a topic is estimated. This probability of a given word-to-topic assignment is calculated from how often the word already occurs in the topic and how dominant the topic is for the document from which the word was sampled. Once the model fitting converges, θ and ϕ can be derived from the approximated distribution z with the **posterior** function.

Multiple models were fitted to determine the best TM parameters. We first conducted experiments to find adequate values for α and β. These influence the model as follows: with a small α (i.e., with many topics $\alpha = 50/T$ becomes smaller) it is likely for documents to contain only a few topics, whereas a bigger α (i.e., few topics) results in more

[1] The full list can be found at [25].

[2] The complete list is: big, data, ieee, discussion, conclusion, introduction, methods, psycinfo database, rights reserved, record apa, journal abstract, apa rights, psycinfo, reserved journal.

topics per document. A small β similarly makes it likely for a topic to contain a mixture of a few words, thereby pushing the model to select highly specific words per topic. A range of values was fitted for both α and β and model outcomes were compared. Within a reasonable range (i.e., $0.1 < \alpha < 1$) we observed only minor differences between topics. Ultimately, fixed values were chosen for α and β, respectively $50/T$ and 0.01 as suggested in the literature [19, 27].

For *model selection* we analysed the likelihood for varying numbered of topics in the range $T \in \{5, 10, 15, \ldots, 100, 150, 200, \ldots, 500\}$. However, likelihood alone cannot be used to find the best model. A penalising factor has to be added for the model's complexity (i.e., the number of variables that have to be estimated). Two information criteria were considered, namely the Bayesian Information Criterion (BIC) [28] and the Akaike Information Criterion (AIC) [29]. When increasing the number of topics in a model, each topic becomes more specific and, therefore, easier to interpret. BIC puts more emphasis on the simplicity (in terms of the number of free parameters) of the model, resulting in a smaller number of topics as compared to AIC. We therefore chose to perform model selection using the AIC. In the case of TM, the variables to be estimated are the latent variables ϕ and θ, which grow with the number of topics. The model where the AIC reached its minimum was considered the optimal model. Equation (1) defines the AIC, where T is the number of topics in model M_T, L is the likelihood of model M_T, and W is the number of unique words in the corpus:

$$AIC(M_T) = -2\log(L) + 2((T-1) + T(W-1)) \tag{1}$$

Post-processing consisted of retrieving θ and ϕ for the optimal model, and calculating the relevance of words within a topic according to the method described by Sievert et al. [30]. Equation (2) defines how relevance r was calculated for word w in topic t given λ:

$$r(t, w|\lambda) = \lambda \log\left(\phi_{tw}\right) + (1-\lambda)\log\left(\frac{\phi_{tw}}{p_w}\right) \tag{2}$$

The relevance is a convex combination of two measures: the topic-specific distribution (ϕ_{tw}) and 'lift' (ϕ_{tw}/p_w), which is a ratio between topic-specific and corpus-wide distributions. These measures can be balanced with $0 \le \lambda \le 1$, by giving more weight to ϕ ($\lambda = 1$) or to the lift ($\lambda = 0$). In our experiments a value of 0.6 was chosen for λ, as suggested in Sievert et al. [30]. $T \times W$ relevancies were calculated (i.e., each word had one relevance score per topic) and used to sort the most relevant words per topic.

Big data definitions

The definition proposed by De Mauro et al. was used as a starting point for this study. Furthermore, the underlying definitions gathered in De Mauro et al. were reassessed and where necessary updated (e.g., updates in white papers published by industry). Lastly, a publication by Andreu-Perez et al. [3] was added because it defined six big data V's in the context of (bio)medical research.

All the definitions were analysed. If the definition was given in free text, the major themes were extracted. Themes were then grouped on similarity, for example, volume and size were merged into one theme. For various reasons a few definitions were discarded, as discussed in the "Big data definitions" section.

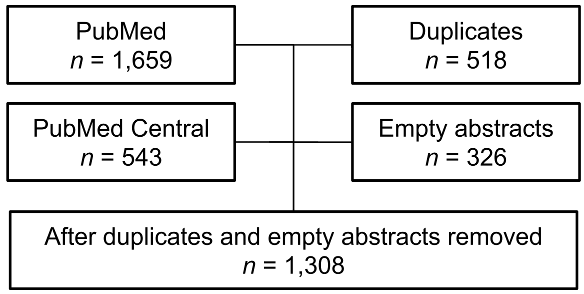

Fig. 2 Corpus generation: documents extracted per literature database, documents removed from the corpus, and total number of included documents

Topic analysis

Topic model results were analysed manually by inspecting the top relevant words (i.e., 20 per topic). The observers received a list of topics and a description of each theme. They were instructed to read all the words in each topic, then consult the big data definition themes, and finally provide their opinion about which themes are associated with that set of words. Each of the topics was assigned zero, one, or more themes by each observer individually. In total seven persons performed the analysis independently: each of the authors and three external health data scientists.

Results

This section reports the results of corpus extraction, TM model fitting and selection, gathering and consolitation of big data definitions, and annotation of topics with the themes.

Corpus

A total of 1659 documents were extracted from Pubmed and 543 from PubMed Central. After removing duplicates and records with an empty abstract, 1308 documents were included in the corpus as shown in Fig. 2.

After pre-processing 136,339 words remained in the corpus, of which 7849 were unique. A large portion (7081 words) had a low frequency (<40 occurrences). Figures 3 and 4 give an impression of the corpus's contents, showing a frequency plot of the top 2000 words, which seems to be in accordance with Zipf's law [31]. To create the word cloud the top 100 most frequent words were extracted (as marked with the vertical line in the frequency plot).

Topic modelling and model selection

In total 49 models M_T were fitted with T ranging between 5 and 500. The AIC curve for all fitted models M is shown in Fig. 5. The minimum of the AIC curve lies at $T = 14$, however the differences are small until $T = 25$. We also calculated the distances between topics from diverse models ($T \in \{14 - 25\}$), which showed that topics are fairly stable (data not shown). When increasing the number of topics, changes observed include one topic splitting into two topics or a new topic appearing. We saw no major reorganisation of topics or words within topics. We also observed that increasing the number of

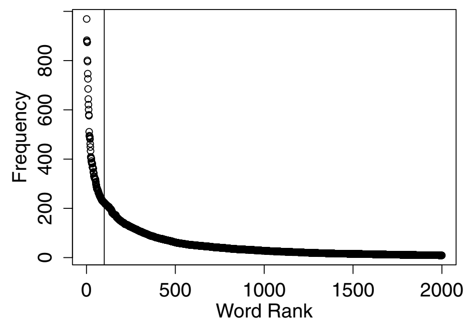

Fig. 3 Frequency of the top 2000 unique words in the corpus. The *vertical line* is the cut-off point ($n = 100$) used for the word cloud

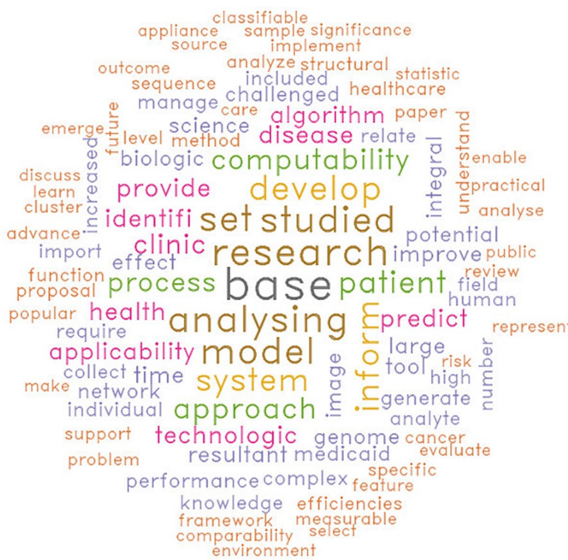

Fig. 4 Word cloud of the top-100 unique words in the corpus

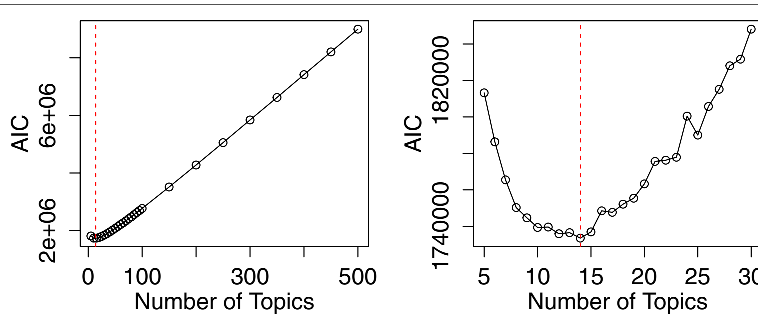

Fig. 5 *Left:* AIC curve of the 49 fitted TM models (20 models between $T = 5$ and $T = 30$ not plotted, see *right*). The minimum is marked by the *dotted line* ($T = 14$). *Right:* Close-up of the AIC curve between $T = 5$ and $T = 30$, showing 26 fitted TM models. The minimum is marked by the *dotted line* ($T = 14$)

topics in the model makes the terms in each individual topic more specific. For example, one topic covering both application and big data themes might be split into two separate topics in a larger model. We therefore selected M_{25} for annotation, as this model has a

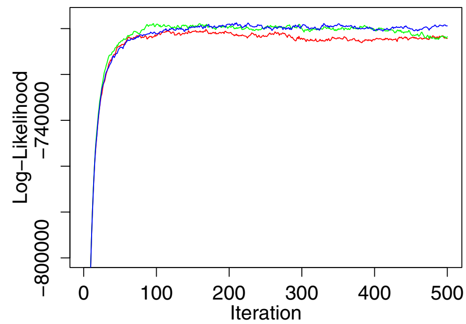

Fig. 6 Convergence of the log-likelihood for the chosen optimal model M_{25} for three runs starting from different seeds

better interpretability compared to M_{14} (more specific topics), with comparable quality of model fit (similar AIC).

To assess the robustness of the model M_{25}, the log-likelihood was tracked for each iteration of Gibbs sampling. This model was fitted three times with fixed input, but with different starting seeds for the sampling. The outcome of these fits is presented in Fig. 6. It shows that the log-likelihood reaches its approximate maximum after 100–150 iterations. Models run with a higher number of iterations (up to 4000, data not shown) showed no major difference in log-likelihood convergence, therefore, final models such as M_{14} and M_{25} were run for 500 iterations. The top-20 most relevant words per topic of the M_{25} model are shown in Table 4.

Big data definitions

In total 17 definitions of big data were considered from the following sources [3, 5, 6, 14, 15, 32–43]. Table 1 presents the results of our analysis listing the found themes, their description, and respective sources. Note that we have not attempted to consolidate the names of the themes, leaving the complete description as found in the sources. The definitions can be divided into three groups, with each group containing multiple themes.

The first group (I) corresponds to the big data V's, which occur in various forms in many of the analysed definitions. Some words were merged into one theme because they are essentially pseudonyms of each other. For example: volume, size, voluminous, and cardinality were found in ten of the definitions and, from their descriptions, refer to the amount of data. Also note that velocity and continuity, and complexity and variety were combined.

The second group (II) corresponds to the aggregated themes proposed by De Mauro et al., which represent concepts of a higher level of abstraction than the previous group.

The third group (III) includes a theme identified in three definitions, which describe big data as data that is *beyond conventional* processing and analysis. The V's describe data by many different aspects, but none of those define a hard limit beyond which data becomes big. The theme 'beyond conventional' therefore describes big data as something that needs novel specialised and scalable solutions. This also means that the types of problems and applications that are assigned to the scope of big data change over time, as technology and methods evolve and improve.

Table 1 Description of themes identified in big data definitions from literature

	Theme name	Theme description	Definition sources
I	Volume, size, voluminous, cardinality	Large quantities of data in number of bytes; size of available data (e.g. all records instead of a sample); beyond conventional storage techniques; number of records at a particular instance	[3, 5, 6, 15, 32–34, 36, 37, 39]
	Velocity, continuity	Flow rate at which data is created, stored, analysed, and visualised; increased through invention of new data streams such as social media; beyond conventional means of processing, needing new techniques such as streaming; growth of data over time	[3, 5, 6, 32–34, 37]
	Variety, complexity	Many different types of data; not bound to a traditional data format; format changes over time; heterogeneous and unstructured data	[3, 5, 6, 15, 32–34, 36, 37, 39]
	Veracity	Trustworthiness of data; reliability of data quality and gathering environment	[3, 32]
	Value	Worth/relevancy of data (e.g. economic, individual/privacy, societal, humanity value)	[3, 6, 38]
	Variability	Consistency of data over time; influences which systematically change data measures over time	[3, 34]
II	Information	Where signals are turned into data (e.g. book digitalisation, or gathering from personal device measurements)	[14]
	Technology	Tools, systems, and software (e.g. scalable processing and transmission systems such as Hadoop)	[14, 15, 34–36, 38]
	Methods	Procedures and their application (e.g. clustering, natural language processing, machine learning, neural networks, visualisation)	[14, 35, 38]
	Impact	Ethical, business, societal	[14]
III	Beyond conventional	Data whose size call for methods beyond the tried-and-true; necessity of scalable systems for storage, processing, manipulation, analysis, visualisation	[35–37]
IV	Application	About the application domain treated in the papers	–

The fourth group (IV) was not found in the studied definitions, but was added to cope with the reality of our data. Because the body of literature used in this study was obtained from (bio)medical literature databases, we expected to see application-related themes to be strongly represented in the resulting topics. We therefore included the Application theme to classify those topics that do not fall under big data.

Note that some definitions considered by De Mauro et al. were not used here:

- The definition by Microsoft [40] was a web-blogpost from 2013, therefore possibly outdated;

- Shneiderman et al. [41] does not specifically mention big data, as it was a publication from 2008 when this term was not in use yet;
- The definition by Manyika et al. [43] was only described in the executive summary;
- Mayer-Schönberger et al. [42] propose an abstract definition that was considered too difficult to convert into interpretable themes for topic analysis.

Topic analysis

The list of topics and words and big data themes were analysed by the seven observers. The observers all worked at the local department of epidemiology, biostatistics and bioinformatics, therefore they were extremely suitable for the annotation task. The big data themes (Table 1) and topic words (Table 4) were well understood and the task could be finished without further help in a reasonable amount of time (30 min to an hour).

The raw annotation results are displayed per observer and per topic in Table 2. Note that some observers did not assign any theme to some topics, and that in many cases more than one theme was assigned to the topics. Table 3 presents the frequency of themes assigned per topic, highlighting high or unanimous agreement among the observers (shown underlined and bold). It also shows the *overall* themes, i.e., those that were assigned to a topic by at least four observers.

In four topics less than four observers assigned the same theme to it (i.e., 3, 17, 19 and 25). Out of the remaining 21 topics, five had unanimous agreement between the observers for some theme (i.e., 6, 7, 8, 20 and 21). The remaining 16 topics could be split into topics with a single overall theme (i.e., 2, 4, 9, 10, 11, 13, 14, 15, 16, 18, 22, 24) and topics with two overall themes (i.e., 1, 5, 12, 23).

Note that the most frequently assigned theme was Application (66 times), followed by the themes in the second group, proposed by de Mauro et al. From the themes in the first group, volume and velocity occurred more often than the others. Notably, variability was hardly identified among these topics.

Figure 7 presents the distribution of topics over documents based on the probability of each topic to each document (i.e., θ). The large majority of topics (in black) have a strong presence in only a few hundred documents. However, there are four topics (in red and blue) that deviate from this pattern. The two red topics (topic 1 and 2, see Table 4) have a stronger presence in more documents as compared to the topics pictured in black. The blue topics (topic 3 and 5, see Table 4) have a stronger presence in nearly all documents.

Discussion

In this paper we attempted to identify themes related to big data definitions in a large corpus of (bio)medical literature through topic modelling. We have followed a structured and objective approach as much as possible. This process delivered novel and interesting results, which however need to be carefully interpreted due to remaining limitations in our study.

Identification of themes in big data definitions

Due to the lack of a consolidated and widely accepted definition of big data, it was necessary to consult a large number of scientific papers. This work is limited to scientific literature, but obviously there are many other definitions of big data that have not been

Table 2 Raw annotation results per observer

Topic	Theme assignment grouped by observer						
	A	**B**	**C**	**D**	**E**	**F**	**G**
1	Imp, value		Value	App, imp, value	Vera, value	imp, app, vera	Imp, value
2	Vera, app		Imp, app	Info, app	Vera, velo	App	Tech, variety, vera
3					Imp, app	App	App
4	Met	Met	Vol, met	Met	Tech, met	Tech, velo	Met
5	Vol, velo, beyond	Tech	Vol, tech, beyond	Beyond, vol, velo	Tech, complex, beyond	Vol	Vol
6	Tech	Tech	Tech, velo	Tech, beyond	Tech, beyond	Tech	Tech, variety, vera
7	Met	Met	Vera, met	Met	Tech, met, info, app	Met	Met
8	App	App	Info, app	App, info	App	App	Variety, app
9	App			Imp	Imp	Imp	Value, imp, app
10	App	Met, tech	Variety, info, met	App, met	App	App, variety, info	Vol, beyond
11	App	App	App	App, Imp	App	App	Imp, value
12	Tech, vol, velo	Vol	Vol, velo	Vol, velo, beyond	Tech, vol, velo	Vol, velo	Met, vol
13	Variability, vera	Met	Met	Met	App, info	Met	Met
14	Info	Info	Tech, app	App, info	Imp	Info	Value, imp, app
15	Imp	App	Imp	App	Info, app	App, imp	Value, vera
16	App	Met	App	Info, app	Info, app	App	Beyond, vol
17	Value	Info	Tech, beyond	Info	Continuity, variability	Tech	Value, tech
18	App	Met	Info	App, info	Met, app, tech, info	App	Vol, vera
19	Value	App	Met, app	Info	Continuity, app	Variety	Tech, imp
20	Met	Met	Met	Met	Met, info	Met	Met
21	App	App	App	App, imp	Info, app	App	Variety, app, vera
22	Info, velo	Info	Info, app	Info, vera	Velo, continuity, app	App, info	Info
23	Info, app	App	Info, app	Info	Info	App, info	Beyond, vol, vera, info
24	Value	App	Info, app	Info, app	Continuity, info, imp	App	Vol, variety
25	Met	Met	Info		Info, met, tech	Vol, velo	Velo
Total	33	22	39	40	53	35	49

The following coding is used to represent the themes described in Table 1: *vol* volume, *velo* velocity, *vera* veracity, *info* information, *met* methods, *tech* technology, *imp* impact, *app* application, *beyond* beyond conventional

considered in our work, such as the Berkeley blog mentioned in the introduction [8]. Nevertheless, most of the definitions in [8] can be mapped to the themes identified in this study. Interestingly, the word cloud in [8] highlights words such as size, complex, and techniques, which are also found in the descriptions of the themes consolidated in Table 1. Furthermore, there are qualitative approaches to describing the big data field in

Table 3 Summed annotations per topic and theme, and overall theme per topic (≥4 counts)

Topic	Themes												Overall
	Volume	Velocity	Variety	Veracity	Value	Variability	Information	Technology	Methods	Impact	Beyond con.	Application	
1				2	**5**					4		2	Value, Impact
2		1	1	3			1	1		1		4	Application
3										1		3	–
4	1	1						2	**6**				Methods
5	**5**	2	1					3			4		Volume, Beyond conventional
6		1						**7**			2		Technology
7				1			1	1	**7**			1	Methods
8			1				2					**7**	Application
9					1					4		2	Impact
10	1		2				2	1	3		1	4	Application
11					1					2		**6**	Application
12	**6**	**5**			1			2	1		1		Volume, Velocity
13				1		1	1		**5**			1	Methods
14					1		4	1		1		2	Information
15				1	1		1			3		4	Application
16	1						2		1		1	**5**	Application
17		1			1	1	2	3			1		–
18				1	1		3	1	2			4	Application
19		1	1		1		1	1	1	1		3	–
20							1		**7**				Methods
21			1	1			1			1		**7**	Application
22		2		1			**6**					3	Information
23	1			1			**6**				1	4	Application, Information
24	1	1	1		1		3			1		4	Application
25	1	2					2	1	3				–
total	17	17	8	12	14	2	39	24	36	19	11	66	

Table 4 Top 20 words for the 25-topic model identified with TM

Topics

1	2	3	4	5
Health	Patient	Article	Algorithm	Challenged
Research	Clinic	Review	Cluster	Analyte
Healthcare	Hospital	Discuss	Learn	Tool
Policies	Electron	Field	Method	Amount
Health_care	Care	Recent	Feature	Technologic
Privacies	Outcome	Issue	Efficiencies	Computability
Nation	Medicaid	Aspect	Approximate	Analysing
Ethic	Record	Focus	Tree	Require
Protect	Ehr	Emerge	Represent	Advance
Govern	Clinical_research	Future	Fast	Varieties
Inform	Health_record	Highlight	Matrix	Solution
Secure	Clinician	Current	Accuracies	Growth
Challenged	Treatment	Context	Problem	Large_amount
Share	Improve	Overview	Distance	Massive
Concern	Assess	Paper	Hierarchical	Generate
Access	Healthcare	Paradigm	Computability	Dataset
Communities	Qualities	Confer	Faster	Vast
Fund	Potential	Natural	Calculate	Process
Health_informatics	Patient_care	Technologic	Graph	Handle
Health_system	Routine	Literature	Outperform	Infrastructural

6	7	8	9	10
System	Model	Age	Change	Network
Process	Predict	Risk	Nurse	Molecular
Device	Infer	Influenza	Innovated	Structural
Framework	Statistic	Indicating	Science	Biomarker
Cloud	Regress	Exposure	Social	Complex
Architectural	Simulate	Cohort	Question	Heterogeneities
Hadoop	Predictor	Rate	Historian	Integral
Applicability	Bayesian	Symptom	Influence	Systems_biology
Service	Fit	Month	Practical	Mechanical
Manage	Good	Yearbook	Insight	Omic
Platform	Optimal	Variable	Cultural	Approach
Design	Prior	Life	Turn	Character
Mapreducable	Base	Death	Product	Dynameomics
Computability	Variable	Diabetes	Food	Function
Base	Machine_learning	Adjust	Societies	Biologic
Support	High_dimensional	Geographic	Understand	Transit
Implement	Tradition	Condition	Drive	Rdge
Task	Rank	Factor	Evolution	Topological
Deploy	Parameter	Demographic	Scientific	Protein
Cloud_computing	Feature	Incidence	Principle	Organ

11	12	13	14	15
Disease	Dataset	Effect	Search	Biomedical
Prevent	Time	Group	Social_media	Informatic
Epidemiologic	Sample	Measurable	Language	Science
Vaccination	Large_scale	Testable	Google	Medicinal
Progress	Computability	Estimate	Word	Medicaid

Table 4 continued

11	12	13	14	15
Immune	Speed	Analysing	Public	Educate
Leverage	Performance	Studied	Relate	Research
Popular	Increased	Statistic	Psychological	Learn
Initial	Approach	Bias	Trend	Personalized_medicine
Develop	Thousand	Large	Emoticon	Era
Heart	Step	Eandom	Twitter	Ontological
Administration	Rate	Valuable	Message	Disciplinary
Intervention	Implement	Power	Online	Translate
Generate	Full	Method	Relationship	Student
Blood	Memorial	Sample_size	Social	Scientist
Advance	Scale	Marker	Visit	Train
Public_health	Hundred	Find	Content	Impact
Reported	Block	Large_set	Caseness	Workshop
Consensus	Applicability	Import	Posit	Discoveries
Earlier	Multiple	Error	Investigacin	Knowledge
16	**17**	**18**	**19**	**20**
Genet	Web	Sequence	Mine	Classifiable
Gene	Resource	Genome	Knowledge	Set
Associating	Code	Bioinformatic	Extract	Object
Phenotype	File	Proteome	Inform	Large_set
Pathway	Laboratories	High_throughput	Chemical	Class
Disease	Public	DNA	Specialised	Noise
Genotype	Compress	Transcriptome	Plant	General
Factor	Semantic	Protein	Biologic	Pair
Enrich	Software	Composite	Concept	Performance
Trait	Retrievable	Ngs	Develop	Abilities
Genome_wide	Access	Metagenome	Toxic	Neural_network
Metabolic	Share	Virus	Construct	Similar
Genome	Format	Analysing	Note	Train
Mutated	Inform	Host	Curate	Dimension
Number	Interface	Biologic	Rich	Machine
Identifi	Source	Assemble	Gap	Categorical
Polymorphism	Platform	Cell	Preservation	Appliance
Individual	Metadata	Microbiome	Ecological	Formula
Regular	Storage	Align	diverse	Encounter
Unification	Exchange	Human	Abstract	Coefficient
21	**22**	**23**	**24**	**25**
Drug	Visual	Image	Cancer	Low
Target	Activated	Brain	Studied	Reduce
Cell	Human	Disorder	Tumor	Time
Event	Behavior	Signal	Valid	Base
Screen	Mobile	Subject	Research	Reduction
Response	Environment	Resolution	Registries	Digital
Experiment	Interact	Neuroimaging	Therapeutic	Node
Detected	Exploration	Function	Database	Energies
Analyse	User	Neuron	Injuries	Deep
Adversary	Collect	Segment	Oncologist	Small
Multiple	Sensor	Psychiatric	Clinical_trials	Cost

Table 4 continued

21	22	23	24	25
Compound	Tool	Connectome	Claim	Size
Profile	Wearable	Neuroscience	Therapies	Numerator
Miss	Quantifiable	Mode	Efficacies	Operability
Type	Track	Mri	Diagnostic	Combina
Potential	Movement	Scan	Heterogeneities	Peak
Combina	Physical	Quantitation	Set	Spectral
Meta	Display	Analysing	Specific	Structural
Complete	Smartphone	Microscopic	Ongoing	Locate
Point	Interest	Multi	Consortium	Qualities

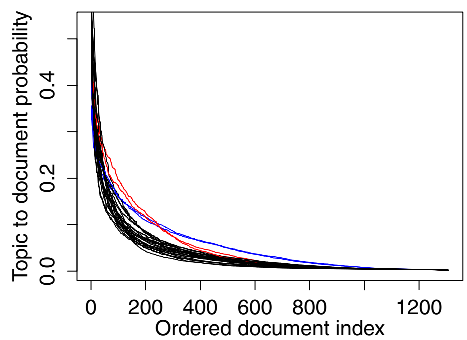

Fig. 7 Distribution of topics over documents (i.e., θ, y-axis). The documents are sorted on topic-to-document relevance within each topic. The *x-axis* represents the order of the sorted documents. Each *line* represents one topic, in *black*. Exceptions are topics 1 and 2, plotted in *red*, and topic 3 and 5, plotted in *blue*

publications such as Chen et al. [13] and Tsai et al. [44]. Note that, although these works do not strive to deliver a formal definition, the description of the big data field in both these publications include the same aspects found in the definition themes.

We have observed a large overlap among the big data definition literature considered in this study, nevertheless with variations in the focus applied by each author. Furthermore, certain themes occur more often than others in the definitions (Table 1). The original three V's (volume, velocity, variety) occur in many definitions compared to the relatively 'newer' V's (veracity, value, variability), which are present in only a few. This is also the case with Technology and Methods which are found in definitions more often than Information and Impact.

Finally, as the corpus was gathered from (bio)medical literature databases, we expected to find topics describing this domain. Therefore the theme 'Application' has been introduced, which is obviously not found in the published big data definitions. Indeed, the annotation results presented in Table 3 show that 10 out of 25 topics have been annotated with Application by the majority of the observers. Note that the large fraction of application-related words might have overshadowed others that are related to big data themes. Scrubbing the corpus of application-related words could be used to circumvent this problem. This opens the possibility for fitting highly granular models that would be more easily interpretable and better reflect big data instead of the research field topics.

Corpus gathering

By design, in this study we only considered papers that were self-annotated with big data, whatever definition the authors might have used. This led to an interesting observation by one observer who could not find his research domain in any of the topics. However, the searched databases certainly included this domain and many of the big data themes could potentially be assigned to its papers. The domain could be missing due to various reasons, such as a low frequency of this research domain in the corpus. However, this observer acknowledged to consider his domain as 'conventional', therefore, papers published about this research domain most likely do not mention big data and were therefore not captured in the search performed in this study.

Note also that we only considered two databases, whereas many others could be included as well (e.g., Scopus or Ovid). Nevertheless, PubMed and PMC are important sources in medical research and therefore have been considered sufficiently representative for the purposes of our study.

Finally, a potential limitation of our study is that only abstracts were included in the corpus instead of full-text papers. Our assumption is that the abstracts contain the essence of a paper and are therefore representative of the actual themes found in a full paper. Moreover, it is currently still difficult to retrieve and parse full papers in an automated fashion, which would have severely limited the number of papers considered in our study.

Automatic identification of topics

In the progress of this research various text mining approaches were attempted to identify relevant topics to characterise the publications. First, we attempted to use AlchemyAPI [45], a natural language processing service that is accessible through the web. However, in a pilot experiment of 100 documents we observed that the number of results produced would be too big for effective analysis (i.e., 3774 results, of which 3006 were unique). Moreover, AlchemyAPI's method is implemented by proprietary code, so relations between documents and results were difficult to interpret.

We continued searching for a text mining method and considered document clustering to find the definition themes in literature. In principle, document clustering could capture themes but results are often limited to one theme per document. Furthermore, analysing document clusters to find definition themes would be a non-trivial (if not impossible) task.

A seemingly more suitable method was topic modelling, a method that can discover latent semantics in text. The main purpose of topic models is described as "discovering main themes that pervade large unstructured collections of documents" [18]. Furthermore, TM captures multiple meanings of words, but most importantly, it can identify multiple topics for each observed document. The LDA approach is perhaps the most popular and common topic model. The R package implementing the algorithm `topicmodels` had 22,576 downloads in 2015.[3] Moreover, the paper describing the underlying model by Blei et al. [17] has been cited over 16,000 times.[4] We therefore chose to use the

[3] http://cran-logs.rstudio.com/ on 9 June 2016.

[4] https://scholar.google.com/ on 20 October 2016.

LDA implementation of TM because of its appropriateness for our data, the relative ease of use of this approach (i.e., ready to use implementations in R), and extensive use in the literature by our peers.

Various TM approaches were tried to find a model with a manageable number of topics which allowed for manual annotation. The largest challenges were encountered during model selection. Two model evaluation methods (i.e., perplexity and harmonic mean) are often used in TM literature [16, 19, 46, 47]. The harmonic mean method calculates an approximation of the marginal likelihood of a fitted model, while perplexity measures how well a fitted model can predict unseen data. These criteria were calculated for multiple models with varying parameters expecting that the model decision boundary lay at some optimum of the response curve. For both criteria we were looking for a sudden decrease in marginal difference between two consecutive data points (i.e., models). Unfortunately, in our case, even when fitting models with up to 1,500 topics (data not shown), the curves did not show an optimum.

Finally we opted for TM with model selection through AIC, a method based on likelihood and model complexity. The AIC curve shows an optimum at M_{14}, however M_{25} was chosen for further analysis. While experimenting with the parameter T we noticed that quantitatively measuring model fit did not relate to the interpretability of the topics, as also noted in [30, 48]. Comparison between models showed that there was no major reorganisation of topics (data not shown), but increasing the number of topics made them more specific and therefore more interpretable.

Manual annotation of topics

Subjectivity of the manual annotation is one of the limitations of this study. Some research has been done in objectifying the analysis of TM results [27, 30, 49, 50]. However, so far, the results of TM cannot be quantitatively evaluated [16, 48]. For the purpose of this study, a group of seven observers was deemed enough for the topic analysis. We also present all the data in the paper, such that the reader can assess the topics themselves to confirm or dispute our results.

We took great effort to objectify the interpretation of TM results, but seven is a small number of observers. Ideally more persons should be involved in the assessment of theme assignment. For example, crowd sourcing services such as Mechanical Turk could be used [51]. However, this particular annotation task requires sufficient background knowledge in health data science, which significantly reduces the pool of suitable observers.

All the observers in this study were trained in health data science, therefore they are familiar with the terms and concepts that appeared in the topics and the big data themes. Nevertheless, no baseline assessment was performed to more precisely understand their own interpretations, which might have introduced some noise in our results.

In general, the observers reported some difficulty to associate words with a theme. They also noted that their annotation decisions were mostly based on words that stood out in the topic, which means that not all words were considered equally. This possibly led to the discrepancy between annotators displayed by the results (Tables 2, 3). For example, when asked, annotator F noted that he chose Technology for topic 4 because of the specific word 'cluster', while all others chose Methods. Note that cluster could be interpreted as a computer cluster (i.e., Technology) or a cluster used in unsupervised machine learning (i.e., Methods). Furthermore, note that Information is often co-annotated or interchanged with Application. For example, neuroimaging, neuroscience, image, and signal are present in topic 23. The first two words can be associated with Application, and the latter with Information. Also, topics containing words referring to data (e.g., images and age) have been annotated as Information and/or Application by some observers. For such reasons some observers said that it was possible that their annotation might change slightly if they would analyse the topics again.

Big data themes in biomedical literature

Despite annotation subjectivity we consider to have found sufficient agreement between the observers to support our findings, which show how big data themes are identified in biomedical literature (see Table 3).

Technology and methods are found fairly often in topics. Note that the identification of these themes is facilitated because they can be associated to concrete terms such as device, cloud, and platform for Technology, or model, infer, and simulate for Methods. From the V's, volume and velocity were the most identified themes, which are also easily associated with terms such as large scale, performance, and computability. These terms are frequently used in practice, explaining why they have been so strongly identified in topics 4, 5, 6, 7, 12, 13 and 20.

Impact, variety, veracity, value, and beyond conventional were annotated less often. Because these are more abstract concepts it is likely that they are more difficult to discover within topics. For example, Value was annotated to topic 1, containing words such as secure, challenged, and protect. Compared to concrete themes (e.g., technology and volume), it was more difficult for the annotators to find a fitting theme. Variability was annotated only twice, however we do believe that it is an integral part of big data. Variability not being recognised could mean that the observers could not identify the theme properly (due to poor theme description or understanding), or that the topics in the selected model could not capture this theme (due to insufficient representation in the corpus).

Each of the themes from the definition by De Mauro et al. (information, methods, technology, impact) was annotated more often than any other (apart from Application). Note that by design these themes are defined in a broader manner, which means that they include the others. For example, Methods includes a few V's such as volume and velocity as well as beyond conventional. Perhaps due to their broadness, the themes from De Mauro et al. were chosen more easily, indicating that their definition covers the understanding of big data in a better way. However, one might wonder whether these themes are exclusively related to big data or whether they will also pop-out in other types of papers. The set-up of our study is not able to answer this question.

Related work

Other studies have been performed to discern a definition of big data [3, 14, 15]. These have provided an overview of big data research in different research fields [3]; a literature analysis to discover big data themes and a proposal for their consolidation into one definition [14]; and an analysis of industry statements on big data [15]. Each of these studies used qualitative methods, whereas our work builds upon their findings with a quantitative method. In particular, our study provides evidence that supports the definition proposed by De Mauro et al. [14] and an aggregation of its underlying definitions (see Table 1).

Many researchers have applied TM for text analysis in various fields [52]. Most similar to our approach is a study by Hansmann and Niemeyer [16], which applied TM to a big data corpus to discover its characteristics. Their research identified three themes, namely IT infrastructure, methods, and data, and applied TM in two stages. The first stage separated the corpus of 248 manually selected papers into the three themes mentioned above. Then, in the second stage, TM was applied to the papers which had been grouped by theme. An in-depth word-by-word analysis of big data characteristics was performed on the second stage TM results. The meaning of each word was assessed, finding the important concepts for each of the themes and where research focus lies in the corpus. Our work differs from [16] in three ways. First, their analysis was based on only three big data themes, whereas we used multiple definitions which led to twelve themes. Secondly, we collected a larger corpus resulting from a systematic review of the literature. Lastly, the research goals differ: instead of finding the defining concepts for each of the themes, our approach identifies existing definitions in a biomedical big data corpus.

There are also more sophisticated (and complex) text analysis approaches such as the method described by Hurtado et al. [53]. Whereas we applied a bag-of-words principle, where each word is considered independently, the method by Hurtado et al. processes whole sentences and preserves context information. In [53] text mining was applied to find trends in topics over time and predict topic popularity in the future. While this is not applicable in our current case it might be interesting for further research (e.g., finding trends of big data over time within scientific literature). Lastly, their method to generate topics also gives them a concise label built from the topic's keywords. This would partially remove subjectivity from annotation, however interpretation of the results is still bound to human interpretation.

Conclusion

In this work we describe a systematic study that attempted to answer the question: 'Which themes from various existing big data definitions are expressed in (bio)medical scientific publications?'. A large number of existing definitions were analysed and consolidated into twelve themes. A large corpus of representative biomedical scientific publications was collected and automatically analysed with text mining to identify the 25 most relevant topics based on title and abstract. Manual annotation was performed by seven observers to identify big data themes in the topics. In spite of the limitations of our study, the results show that these themes can be identified in this corpus. Volume, Velocity and Value are recognized frequently, but in particular results show strong presence of the themes defined by De Mauro et al. (i.e., Information, Methods, Technology, and Impact). This finding indicates that their definition of big data is supported by the current understanding expressed by authors when they use the term big data in their own (bio)medical publications in this corpus. To our knowledge this is the first time that this is shown in a systematic manner for literature in an application field.

Abbreviations

IT: information technology; NIST: National Institute of Standards and Technology; TM: topic modelling; DOI: digital object identifier; LDA: latent dirichlet allocation; V's: big data aspects i.e., volume, velocity, variety, veracity, value, variability.

Authors' contributions

SDO and AJvA conceived the study and together with PDM and AHZ created the study design. AJvA performed the study execution, SDO and AJvA analysed and interpreted the results. AJvA drafted the manuscript which was proofread and edited by SDO, the final manuscript was also proofread by PDM and AHZ. All authors read and approved the final manuscript.

Acknowledgements

This work was carried out on the High Performance Computing Cloud resources of the Dutch national e-infrastructure with the support of SURF Foundation. Furthermore, we would like to thank the observers for their work on annotating the results.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The original corpus data will not be published due to copyright concerns. However, the search can be repeated with the same results, see Methods section. The search was performed on 29 March 2016 and therefore includes publications up to this date. Our R implementation of TM can be found on GitHub, see [54].

Funding

This publication was supported by the Dutch national program COMMIT/ which is funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

Received: 24 September 2016 Accepted: 1 November 2016
Published online: 15 November 2016

References

1. Fenn J, LeHong H. Hype cycle for emerging technologies, 2011. Stamford: Gartner; 2011.
2. Google: Google Trends. https://www.google.com/trends/explore#q=big+data. Accessed 28 Mar 2016.
3. Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang G-Z. Big data for health. IEEE J Biomed Health Inform. 2015;19(4):1193–208. doi:10.1109/JBHI.2015.2450362.
4. Gartner: Gartner Acquisitions. http://www.gartner.com/technology/about/acquisition_history.jsp. Accessed 27 Mar 2016.
5. Laney D. 3D data management: controlling data volume, velocity and variety. META Group Res Note. 2001;6:70.
6. Dijcks JP. Oracle: Big data for the enterprise. Redwood City: Oracle; 2012.
7. IBM: IBM - What Is big data? Accessed through Google cache. https://www.ibm.com/software/data/bigdata/what-is-big-data.html. Accessed 17 Dec 2015.
8. Dutcher J. What is big data? https://datascience.berkeley.edu/what-is-big-data/. Accessed 12 Sept 2016.
9. Jacobs A. The pathologies of big data. Commun ACM. 2009;52(8):36–44. doi:10.1145/1536616.1536632.
10. DeRouen T. Promises and pitfalls in the use of "Big Data" for clinical research. J Dent Res. 2015;94(9):107–9. doi:10.1177/0022034515587863.
11. Zikopoulos P, Eaton C. Understanding Big data: analytics for enterprise class hadoop and streaming data. New York: McGraw-Hill Osborne Media; 2011.
12. Levi M. Kleren van de keizer [The emperor's clothes]. Medisch Contact; 2015.
13. Chen M, Mao S, Liu Y. Big data: a survey. Mobile Netw Appl. 2014;19(2):171–209. doi:10.1007/s11036-013-0489-0.
14. De Mauro A, Greco M, Grimaldi M. A formal definition of big data based on its essential features. Lib Rev. 2016;65(3):122–35. doi:10.1108/LR-06-2015-0061.
15. Ward JS, Barker A. Undefined by data: a survey of big data definitions; 2013.
16. Hansmann T, Niemeyer P. Big data - characterizing an emerging research field using topic models. In: Proceedings of the 2014 IEEE/WIC/ACM International joint conferences on web intelligence (WI) and Intelligent Agent Technologies (IAT). Vol 1. WI-IAT '14. Washington, DC: IEEE Computer Society; 2014. p. 43–51. doi:10.1109/WI-IAT.2014.15
17. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
18. Blei DM. Probabilistic topic models. Commun ACM. 2012;55(4):77–84. doi:10.1145/2133806.2133826.
19. Steyvers M, Griffiths T. Probabilistic topic models. Handbook Latent Semant Anal. 2007;427(7):424–40.
20. R Core Team R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2015. https://www.R-project.org/.
21. Feinerer I, Hornik K, Meyer D. Text mining infrastructure in r. J Stat Softw. 2008;25(5):1–54.
22. Benoit K, Nulty P. Quanteda: quantitative analysis of textual data. 2015. R package version 0.8.5-10. http://github.com/kbenoit/quanteda.
23. Lewis DD, Yang Y, Rose TG, Li F. Rcv1: A new benchmark collection for text categorization research. J Mach Learn Res. 2004;5:361–97.
24. Salton G. The SMART retrieval system-experiments in automatic document processing. Upper Saddle River: Prentice-Hall Inc; 1971.

25. Lewis DD, Yang Y, Rose TG, Li F. http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop. Accessed 2015-11-20
26. Grün B, Hornik K. Topicmodels: an R package for fitting topic models. J Stat Softw. 2011;13(40):1–30.
27. Chuang J, Gupta S, Manning C, Heer J. Topic model diagnostics: assessing domain relevance via topical alignment. In: Proceedings of the 30th International Conference on machine learning (ICML-13); 2013. p. 612–20.
28. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4. doi:10.1214/aos/1176344136.
29. Akaike H. In: Parzen E, Tanabe K, Kitagawa G, editors. Information theory and an extension of the maximum likelihood principle. New York: Springer; 1998. p. 199–213. doi:10.1007/978-1-4612-1694-0_15
30. Sievert C, Shirley KE. LDAvis: a method for visualizing and interpreting topics. In: Proceedings of the Workshop on interactive language learning, visualization, and interfaces; 2014. p. 63–70.
31. Zipf GK. Human behavior and the principle of least effort: an introduction to human ecology. Indianapolis: Addison-Wesley Press; 1949.
32. Schroeck M, Shockley R, Smart J, Romero-Morales D, Tufano P. Analytics: the real-world use of big data. IBM Global Business Services. 2012: 1–20.
33. Suthaharan S. Big data classification: problems and challenges in network intrusion prediction with machine learning. SIGMETRICS Perform Eval Rev. 2014;41(4):70–3. doi:10.1145/2627534.2627557.
34. Chang L. NIST big data interoperability framework. vol 1. Definitions. doi:10.6028/NIST.SP.1500-1
35. Fisher D, DeLine R, Czerwinski M, Drucker S. Interactions with big data analytics. Interactions. 2012;19(3):50–9. doi:10.1145/2168931.2168943.
36. Chen H, Chiang RH, Storey VC. Business intelligence and analytics: from big data to big impact. MIS Q. 2012;36(4):1165–88.
37. Dumbill E. Making sense of big data. Big Data. 2013;1(1):1–2.
38. 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. doi:10.1080/1369118X.2012.678878.
39. Center I. Big data analytics. Intel IT Center; 2012.
40. Microsoft: the big bang: how the big data explosion is changing the world. https://news.microsoft.com/2013/02/11/the-big-bang-how-the-big-data-explosion-is-changing-the-world/. Accessed 11 Feb 2013.
41. Shneiderman B. Extreme visualization: Squeezing a billion records into a million pixels. In: Proceedings of the 2008 ACM SIGMOD international conference on management of data. SIGMOD '08. New York: ACM. p. 3–12; 2008. doi:10.1145/1376616.1376618
42. Mayer-Schönberger V, Cukier K. Big data: a revolution that will transform how we live. London: John Murray Publishers; 2013.
43. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big data: the next frontier for innovation, competition, and productivity. 2011.
44. Tsai C-W, Lai C-F, Chao H-C, Vasilakos AV. Big data analytics: a survey. J Big Data. 2015;2(1):21. doi:10.1186/s40537-015-0030-3.
45. Alchemy API: Alchemy. http://www.alchemyapi.com. Accessed 15 Dec 2015.
46. Wallach HM, Murray I, Salakhutdinov R, Mimno D. Evaluation methods for topic models. In: Proceedings of the 26th Annual international conference on machine learning. ICML '09. New York: ACM; 2009. p. 1105–1112. doi:10.1145/1553374.1553515.
47. Sievert C. Finding structure in xkcd comics with latent dirichlet allocation. https://cpsievert.github.io/xkcd/. Accessed 20 Nov 2015.
48. Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM. Reading tea leaves: how humans interpret topic models. In: Bengio Y, Schuurmans D, Lafferty J, Williams C, Culotta A, editors. Advances in neural information processing systems 22. Red Hook: Curran Associates Inc; 2009. p. 288–96.
49. Lau JH, Grieser K, Newman D, Baldwin T. Automatic labelling of topic models. Proceedings of the 49th Annual Meeting of the association for computational linguistics: human language technologies, vol 1. HLT '11. Stroudsburg: Association for Computational Linguistics; 2011. p. 1536–45.
50. Mei Q, Shen X, Zhai C. Automatic labeling of multinomial topic models. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '07. New York: ACM; 2007. p. 490–499. doi:10.1145/1281192.1281246
51. Amazon: Amazon Mechanical Turk. https://www.mturk.com. Accessed 27 Feb 2016.
52. Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X. Comparing twitter and traditional media using topic models. In: Proceedings of the 33rd European Conference on advances in information retrieval. ECIR'11. Berlin: Springer; 2011. p. 338–349. http://dl.acm.org/citation.cfm?id=1996889.1996934.
53. Hurtado JL, Agarwal A, Zhu X. Topic discovery and future trend forecasting for texts. J Big Data. 2016;3(1):1–21. doi:10.1186/s40537-016-0039-2.
54. Altena, van AJ. AMCeScience/R-topicmodelling at Submission. https://github.com/AMCeScience/R-topicmodelling/releases/tag/Submission.