

# The promise and the premise: How digital media present big data by Gastón Becerra

---

## Abstract

This paper analyzes the thematic and discursive construction of big data by the Argentine digital press. Using text mining techniques — topic modelling and enriched associative networks — together with qualitative and quantitative content analysis — in both discourse and images — over 2,026 articles, we sought to identify the topics wherein big data is treated, the promises and risks it addresses, its definition within the semantic field in which is explicitly expressed, and the pictures that illustrate it. Results herein presented compare how big data is portrayed in news about politics, business, and technological innovations, as well as in focal pieces targeted to a generic and massive audience, and critical reflections about its risks. Although in each of those thematic contexts big data is anchored differently, there is a common idea that associates big data with a socio-technological premise and an epistemic promise: because of the availability of large volumes of data, something new that will allow better decisions can be known. Our exploration contributes to a more detailed knowledge on how the news media social systems make sense of novel phenomena such as big data.

## Contents

[Introduction](#)

[Big data semantics](#)

[Methods](#)

[Results](#)

[Conclusion](#)

---

## Introduction

In this work we seek to describe how big data is portrayed in mass media communications, specifically, in the Argentinian digital press. Such objective is a particular step within a broader project aiming to compare how big data is treated and framed in different social systems, such as science, politics, commerce, economics, among others (Becerra, 2018a). Theoretically, we build on Niklas Luhmann's systemic and constructivist perspective, which holds that complex social phenomena's meaning is not univocal, but rather is constructed within each system's communications conditioned by their specific operations and coding (Becerra, 2018b; Luhmann, 1995). This is in line with the meta-theoretical advice from recent literature (Beer, 2016; Boellstorff, 2013; Kitchin, 2014), that state that big data should be understood as a complex issue, fragmented in social reality by different rationalities, and what is currently required from social sciences are case studies detailing these constructions.

Following Luhmann (2000), we understand that a mass media social system consists of communications that make use of technical means of massive reproduction to disseminate communication beyond a physical presence, which then rules out any coordination between sender and receiver [1]. Having to assume acceptance of their communications, mass media standardize messages in a way that modulates expectations, take the same products they have created as points of reference, and create their own receivers' profiles (Becerra and Arreyes, 2013; Bechmann and Stehr, 2011; Gerim, 2017).

Mass media have a self-observation function in society: they continuously update the limits between the known and the unknown, reconstructing a social memory whereby they can add marks and symbols, such as those that identify something as new, relevant, or urgent. Systems' communication self-organize into themes and topics, which "... gather contributions into complexes of elements that belong together, so that it can be discerned in the course of communication whether a topic is being retained and carried forward or whether it is being changed" [2]. Themes and topics work as narrower selective contexts where sense-making complexity is reduced and communication expectations can be set [3]. Without topics, there would be no need for more information about something since there would be no background to assess novelty or surprise, or coherence, consistency, or completeness of each piece of communication. In the case of mass media, topics subsist as long as they arouse social interest. Mass media success depends on their ability to impose acceptance of topics [4].

For our general comparison interests, topics have the advantage of being abstract designations that can cross different social systems. For the narrower and descriptive reach of this paper, topics will be treated as frames (Gamson, *et al.*, 1992; Jacobi, *et al.*, 2016) wherein big data is defined and problematized, or as we will say, thematized. The general questions guiding our explorations are: Which are the themes and topics wherein big data is being framed? How is big data portrayed and defined across topics?

In order to further specify these questions for an empirical survey, in the next section we present a brief state-of-the-art about the big data semantics. Then, we report our methodology and results from exploring topics and thematization of big data in a corpus of 2,026 news from the Argentinian digital press. We conclude by summarizing and comparing these thematizations of big data in order to elucidate the rationality of its mass media treatment.



## **Big data semantics**

We refer to “semantics” as the way in which society communicates about something (Luhmann, 2007). According to Luhmann: “... an intervening requirement mediates between language and interaction — a supply of possible themes that is available for quick and readily understandable reception in concrete communicative processes. We would like to call this supply of themes culture, and, if it is reserved specifically for the purposes of communication, semantics” [5]. Following this lead, we’ll speak of a “big data semantics” to refer to the themes and topics wherein big data is anchored, alluded, and problematized for communication. Working with semantics, we are compromised to a “second order observation”, thus an observation of what other social systems observe (communicate). Therefore, we are not interested in giving our own definition of big data but to observe what — and eventually theorize about how and why — big data is for mass media and news.

Due the fragmented nature of the semantics of social complex phenomena, comparative and detailed studies from second order observations are required. To the best of our knowledge, there are very few examples of such efforts in big data literature. Kitchin’s (2014) *The data revolution: Big data, open data, data infrastructures and their consequences* is a great one. Kitchin looks into the rationalities behind the discourses of big data in four illustrative contexts, and details the main task or promise for each: first, within political and state communication, the task is of governing people, and big data is related to more efficient and transparent administration, in addition to precise needs

such as surveillance and security; second, organizations, where the task is to manage them through efficient decision-making processes based on rich, detailed and real-time information; third, city planning and urban management, closely related to “smart city” projects, with the task of creating better living spaces; and finally, commerce, with the task of adding value and creating capital, big data is presented as the possibility to add intelligence to the whole commercial chain. Pinpointing tasks or promises is an effective way to synthesize the most powerful rationales behind a “promotional” discourse of big data. Therefore, we can extract our first research question for the survey: (*RQ1*) What are the tasks and promises, and/or risks and threats to which big data relates to in different framings of each topic?

Another element worth exploring is the meaning of the very term “big data”. One can argue that this abstract and vague designation is well suited for the complex or fragmented scenario we are describing: neither “big” nor “data” have an intrinsic meaning; they are both relational and positional concepts. Yet, together they give rise to a tacit critique on how we deal with information in the different systems of society. In fact, a common definition for “big data” identifies it with “information that can’t be processed or analyzed using traditional processes or tools” [6]. Yet its rhetoric does not only consider its manageability: big data is allegedly “bigger” and “better” to whatever we consider data — which is always a local, sociopolitical, dated, and theory-driven designation. Further, as Portmess and Tower (2015) suggest, there is something unsettling about the openness of the term “big data” that positions us in the center of a tension between promises and risks. In their own words: “Linguistically, the expression Big Data frequently seems less descriptive than rhetorical, suggesting new uses and new insights from mining massive datasets yet carrying darker intimations of manipulation and new forms of social control, ‘a linguistic cousin to the likes of Big Brother, Big Oil and Big Government’ (Lohr 2012)” [7]. This ambiguity of the big data discourse also translates to the metaphors used to convert its abstractness into images able to engage us. In this vein, Puschmann and Burgess (2014) reviewed business and technology press Web sites and identified two recurring metaphors: the first one, big data as a force of nature to be controlled; the second one, big data as nourishment or food to be consumed. In these, the images used to evoke it — among the liquid images: a flood, an ocean, a stream, a tsunami; or, among the ethereal ones: a cloud, an explosion; or even mixed ones, such as energy and fuel — are ambivalent in regards to the promises/risk tension: all of these (natural) images tend to obscure the human and social nature of big data, while also offering a risky or epic framework for its exploitation. Looking towards our survey, we can ask: (*RQ2*) How is big data defined? Which is the semantic context in which it is explicitly mentioned? Are those unspecific and abstract ideas, or on the contrary, concrete and context-specific ideas?

Since there is no reason to limit ourselves to the textual components of semantics, visual imagery used to illustrate big data communications should also be considered. Another study conducted by Pentzold, *et al.* (2019) analyzed article illustrations from the *New York Times* and *Washington Post* (2010–2016) to identify (among other goals) how is it represented and what it could mean to society. Their findings showed a preeminence (29.4 percent) of images of big data people — ranging from protagonist to IT workforce — which could be understood as the “human face” of big data; and of application contexts — *e.g.*, banks, offices, courts, etc. — (26.4 percent), wherein big data itself is not represented. Next cluster is technology and the material side of big data (19.8 percent) — including IT logos, apps screenshots, and infrastructure. The authors conclude that “... depictions of people, materialities, and application contexts serve as concrete visual surrogates for the virtuality and immateriality of big data”. Less significant clusters are the most abstract: visualizations — infographics, large numbers, and artistic renditions — (13.9 percent); and illustrations of the datafication process and datafied individuals (10.4 percent). Regarding these, the authors stated: “Our analysis could not confirm ... the primacy of metaphorical imagery of data as a natural force or nourishment/fuel that seem to dominate on the verbal level of news. Datafication was the only of the 13 image types that also uses a visual rhetoric of big data drawing on such kind of metaphors — but this only occurred in 10 cases, accounting for 2.2 percent of the total big data imagery” [8]. Cautious, the authors advise us to approach media framings holistically, considering pictures and texts as complementing modes of communication and sense-making. Thus, the next questions for our survey is: (RQ3) What images are used to illustrate news about big data, and how they complement the textual message, in different topics?

The theoretical guideline that compromises us to explore the fragmented semantics of big data should not impede us looking for common or core ideas across different framings, nor analyzing their synergy and crossover. In fact, Kitchin’s aforementioned reconstruction of the four tasks is a step towards showing how different actors, and through different channels, try to install a common “discursive regime” — in the Foucauldian sense — that justifies and naturalizes the adoption of new ideas and practices, and by generating an ambience of desirability and interest. The quid of this discourse is “its promise is to offer a radically new way of understanding and managing all aspects of human life”. The universal reach of this claim must not be undermined, as the author highlights: “What is interesting in the case of big data is that its discursive regime is being targeted at all sectors — social, political, economic, environmental”. Both the epistemic promise and the universal reach are elements that can be identified in other keen analysis, *e.g.*, Dijck’s (2014) account of the “ideology of dataism”: “... [the] widespread belief in the objective quantification and potential tracking of all kinds of human behavior and sociality through online media technologies”; boyd and Crawford’s

(2012) description of the “mythology of big data”: “the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy”; or the aforementioned analysis of Puschmann and Burgess (2014) on metaphors: “... the widely held hope that data can be effectively harnessed to better approach a wide range of societal issues, from economic growth and development to security and health care, with far-reaching implications”. We’ll return to this issue in the conclusion.



## Methods

In order to answer these questions, we pursued a mixed design that combines qualitative and quantitative discourse analysis with different text mining and natural language processing techniques. In doing so, we aim to retain the strengths of “traditional” content analysis while maximizing the reproducibility and large-scale efficiency of computational techniques.

Currently there is no canonical methodology for this type of mixed designs, although there are solidly founded previous studies (Bryant and Raja, 2014; DiMaggio, *et al.*, 2013; Evans, 2014; Grimmer and Stewart, 2013; Lewis, *et al.*, 2013; Törnberg and Törnberg, 2016; Wiedemann, 2015). Our process followed these steps:

**Corpus construction.** We started by creating a list of digital newspapers and news Web sites, by crossing different Argentinian news portals and listings. Then, using a Web search engine we queried those sources using the term “big+data” [9] (search conducted in October 2018). We listed up to the first 100 articles per source, sorting by relevance [10], and filtered out articles from sources with less than 15 results. This resulted in 89 different sources and 3,270 articles for which we retrieved their full content and some metadata (including pictures in social media, which we used to answer *RQ3*). Then, we extracted the main text of the articles and filtered out those without “big data” (thus removing related articles, indexes, and other internal search results). This resulted in the final corpus of 2,026 articles from 79 different sources, that we used to answer *RQ1*. From this set of articles, we collected the sentences that explicitly referred to the term “big data”, resulting in a second corpus of 4,100 sentences used to answer *RQ2*.

**Corpus preprocessing.** Our first pre-processing step on the corpus was to annotate the text with Part-of-Speech, using the UDPipe Package for R and its language model for Spanish text (Straka and Strakova, 2017). We used the annotated text for both articles and sentences corpora. For analysis we kept

only nouns, adjectives, pronouns, and verbs (excluding the Spanish translation of to be, to have, and to do); removed punctuation; lowered the text; did a few text replacements on specific N-grams (*e.g.*, *big\_data*, *inteligencia\_artificial*); removed infrequent words (<10); finally, we used lemmas instead of words.

**Topic modelling.** To explore the different topics, we used the latent dirichlet allocation model (Blei, *et al.*, 2003), via the topicmodels R package (Grün and Hornik, 2011), which posits different distributions of the corpus' vocabulary as topics, and calculates the proportional mix of them for each document. In this work, we are using topic modelling to automatically identify recurring topics, and then to classify the articles according to their topic composition. This is an unsupervised modelling technique, with no predefined topics nor any semantic information, that draws upon word correlations. According to one of the model's authors, the interpretability of most topics is a result of "the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of LDA" [11] The model assumptions allow both word polysemy and document heteroglossia (DiMaggio, *et al.*, 2013), which render it useful for our purposes of comparing how different semantic contexts frame big data. Since the number of topics must be introduced as a parameter, after several runs and statistical tests with a different parameters, we settled on 24 topics [12]. Then, after fitting the model, it is the researcher who selects and hypothesizes which ones could be indicative of relevant latent topics: we manually labeled them by analyzing their top terms — *e.g.*, "work, profession, career" could be indicative of a topic about jobs — and later by sampling the most relevant documents per topic. It should be noted that not all inferred topics are interpretable or relevant (Chang, *et al.*, 2009); in our case, we discarded six topics [13].

**Qualitative and quantitative content analysis.** To answer *RQ1*, we compared how big data is portrayed in the articles from different topics, following a constant comparison technique [14]. Then, to assess some of our hypotheses in a quantitative manner we performed word frequency and co-occurrence analysis. To answer *RQ2*, we built associative enriched networks that visualize the main term correlations both at first and second degree for "big data". Here, we assume that an implicit definition is provided within the semantic context of the sentences that explicitly mention it. Additionally, we used a lexicon that grades the imagery of these terms, understood as the ease with which people could "form a mental picture of the involved word", thus being an indicator of abstraction (for this, we used an adaptation into Spanish of Whissell's (2009) model, developed by Gravano and Dell'Amerlina Ríos [2014]). Finally, in order to answer *RQ3*, we manually classified over 1,500 pictures from the most representative articles, following Pentzold, *et al.* (2019) categories and examples [15].

Since we were only interested in sampling the most relevant articles from each topic, for all of the analysis we considered those articles wherein the topic in question had the highest proportion. For the qualitative analysis, we worked on a random selection of ~30–50 articles from these, following theoretical saturation criteria.



## Results

From the 24 topics solution fitted, we kept 18 for further analysis. The topics we worked on are shown in [Table 1](#), with the label we assigned to them, the number of articles wherein each topic is predominant, and their main terms.

<b>Table 1: Twenty-four-topic solutions (non-discarded topics) with distribution and main terms.</b>		
<b>Topic labels</b>	<b>Articles</b>	<b>Terms</b>
1. Social networks	88	facebook usuario dato redes_social campaña empresa twitter red
2. Biz. int.	107	nuevo digital proceso tecnología cambio negocio innovación organización
4. Society	94	libro vida mujer mundo sociedad mismo historia humano
6. Politics (int.)	48	país argentina internacional presidente china mundial desarrollo américa
7. Politics (national)	81	provincia desarrollo bigdata nacional jujuy ministerio gobierno realizar
9. Jobs	85	trabajo profesional empresa laboral sector



		demanda empleo perfil
10. Elections (national)	109	gobierno macri presidente política campana político cambiemos elección
13. Tourism/urban	92	ciudad aires buenos adsbygoogle viaje turismo primero destino
14. Data	149	dato información bigdata análisis permitir analizar decisión grande
15. Elections (int.)	46	política medio social poder estado político gran Trump
16. AI	103	tecnología internet tecnológico dispositivo inteligente AI nuevo máquina
17. Apps and platforms	46	usuario persona plataforma mismo contenido ver millón Netflix
18. Investments	115	empresa compañía servicio negocio cliente mercado año argentina
19. Privacy	76	dato público información seguridad personal gobierno internet ley
20. e-commerce	76	cliente empresa producto consumidor servicio compra banco permitir
22. Agro	95	tecnología producción productor campo agricultura herramienta permitir cultivo
23. Education/sports	119	universidad educación equipo programa

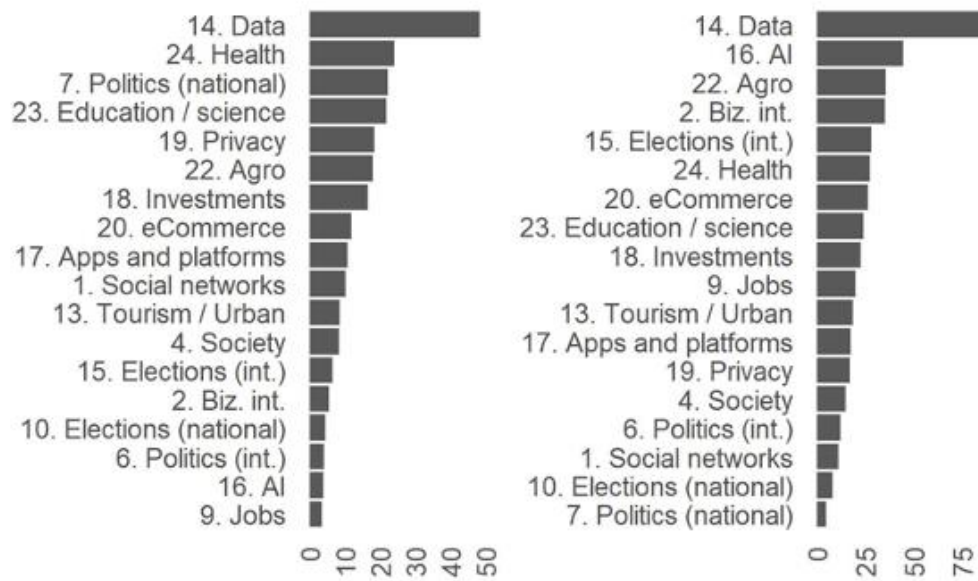
		jugador ciencia curso escuela
24. Health	58	salud médico paciente sistema enfermedad nuevo permitir tratamiento

After analysis we've grouped the topics by both theoretical affinity and empirical coherence. In this section we report our results following these groups, leaving the overall comparison and commentaries for the conclusions:

1. **Big data in focus:** topic #14 (data);
2. **The risks of big data:** topics #19 (privacy) and #4 (society);
3. **AI applications and algorithms:** topics #16 (AI), #1 (social networks), and #17 (apps and platforms);
4. **Big data in politics:** topics #6 (international politics), #7 (national politics), #10 (international elections), #15 (national elections);
5. **Big data in particular business areas:** topics #2 (business intelligence and innovation), #9 (jobs), #13 (urbanism/tourism), #18 (investments), #20 (e-commerce), #22 (agro), #23 (education and sport), #24 (health).

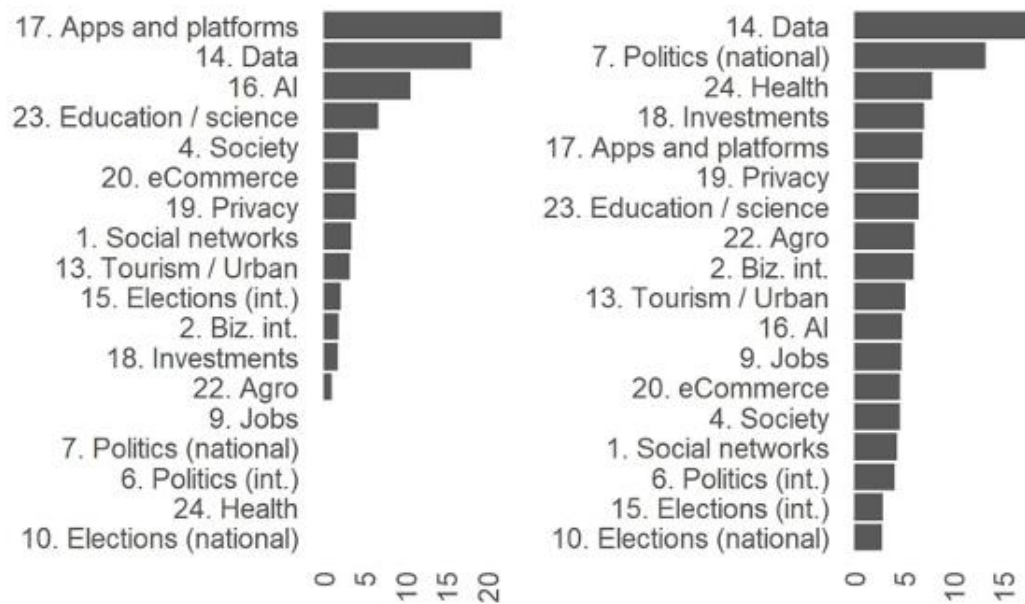
### ***Big data in focus***

The first group of articles we are interested in are those in which topic #14 ranks higher. The articles sampled include several focal pieces aiming at introducing and defining big data to a general audience. Questions like “What is big data?” or short definitions that introduce it as a valuable phenomenon, such as “Big data: a key tool for decision making” or “Big data, a one-way journey in marketing” are frequently read in headlines. In fact, as it is observed in [Figure 1](#), there are more articles including “big data” in their headlines in this topic than in others. These articles usually include: explicit definitions of big data, lists of benefits and possible uses of big data, and quotes from specialists and experts (and even in some cases, the piece itself are authored by a professional from a company that renders big data services). Although the most portrayed user of big data is the manager of a business, it is not unusual to find other professionals, such as a scientist or a medical doctor.



**Figure 1:** Percentage of articles including “big data” in their headlines (left), and the terms “volume”, “variety”, and/or “velocity” (right).

In these articles, the “promise” of big data is clear: to gain more information and support better decisions. It is a very general and abstract claim of epistemic nature, that asserts that there is something more to know. This is a problematic stance, since it is both counter-intuitive and unspecific. To make it real, something unquestionable (a “premise”) must be offered, and that is where (the availability of) huge volumes of data appears. On these focal pieces there is an insistence on numbers, measurements, and other notions related to quantification of data. As it is observed in [Figure 2](#), articles wherein this topic is preeminent tend to mention units of measurement for information, such as terabyte or petabyte. Managing, storing and querying this volume of data were the primary challenges within the IT industry in which the expression “big data” emerged (Diebold, 2012). In a known consultancy piece, Douglas Laney (2001) synthesized these challenges by referencing three v’s: volume, velocity, and variety [16]. Almost 75 percent of the top articles for this topic include these, as shown in [Figure 1](#).



**Figure 2:** Percentage of articles a measurement unit of data (\*byte) (left), and “big data” (right).

It is in the conjunction of these claims that big data raises both the problem and the solution — *e.g.*, in an article titled “What is big data?” it is said that “What makes Big Data so useful for many companies is the fact that it provides answers to many questions that companies didn’t even know they had.” In an efficiency-based competitive society, big data becomes an unavoidable challenge. In the words of IBM’s specialists:

As the amount of data available to the enterprise is on the rise, the percent of data it can process, understand, and analyze is on the decline, thereby creating the blind zone ... What’s in that blind zone? You don’t know: it might be something great, or may be nothing at all, but the “don’t know” is the problem (or opportunity, depending on how you look at it). [17]

Given the focality of these articles, it is not surprising that we find several sentences explicitly mentioning big data, as observed in [Figure 2](#). When we turn our attention to them, and see how big data relates to other terms, we notice that the most prominent relations are with words such as company, information, allow, offer, industry, intelligence, clients, decision, and tool, which are all within the semantic scope of the promise. These are all easy to

imagine words, which could be understood as the solid substratum to anchor the abstract idea of big data.

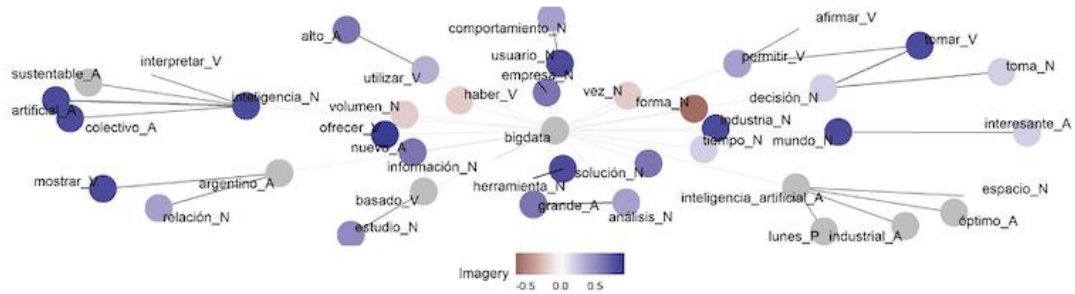
Regarding the images included in these articles, two subcategories stand out: processes of datafication (30 percent), and materiality of technology (26 percent). The most frequent type allegedly represents the flow of data in society, *e.g.*, drawing of linked symbols — or the intersection of the empirical realm and data one — *e.g.*, pictures of mobiles with data coming out. The materiality of technology is depicted by huge servers and workstations.

### ***The risks of big data***

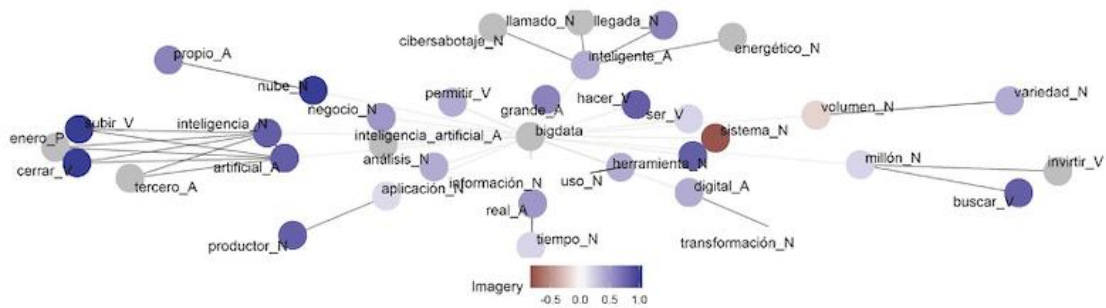
In the next set, big data is again the main theme in a large proportion of the articles. Those that rank higher on topic #19 usually focus on the privacy issues raised by big data and new technologies, ranging from pieces that report on the political and legal manifestations —with headlines such as “In the era of big data, what are the 5 changes that seek to modernize the Personal Data Protection Law”, or “Goodbye to privacy in networks: projects that seek to defend people” — to critical commentaries and awareness events — *e.g.*, “Big Data and the Internet of things, new challenges in personal privacy” or “[Activists] warn about gender violence through social networks”. Also, although not that close in terms of vocabulary, there’s another set of articles that discuss similar themes. If we go through the documents that rank higher for topic #4, we’ll find two types of articles: divulgation pieces on cultural and psychosocial issues addressed with big data, *e.g.*, “With technology, [scientists] discover the four most frequent personality types”, and critical reflections about the (mis)use of technology in current society, from a philosophical, psychoanalytic, or even literary point of view, with explicit references to Jorge Luis Borges, Sigmund Freud, or Byung-Chul Han.

If we turn our attention to the correlation of terms within the sentences that mention big data for both topics #19 and #4 (see [Figure 4](#)), we can see that there is a large overlap with the terms from the previous set (this is much higher if we consider topic #19 alone). However, qualitative analysis suggests a shift in the framing: indeed, big data still is defined as the analysis of the available data — and interestingly we can see a high correlation with the three v’s — but the goal of such effort is not pointed out. Instead, what is stressed is that there is an ill-defined limit in which there is an abuse of data, and goals such as the personalization of experience by companies or the protection by government agencies, may turn to manipulation and surveillance. In one piece, a privacy expert states: “In a legitimate job, the Police may require that information to investigate a crime which is fine if it is within a legal framework. The question is that some of these measures are disproportionate or do not register much transparency with respect to that procedure. We must distinguish between the good and the bad, and it is necessary to supervise who

watches over us”. This change is also visible in the negative correlation of articles proportions for topics #14 and #19, either news reports about the “promise” or the “risks”.



**Figure 3:** Main “big data” correlations in sentences and their imaginability (topic #14).



**Figure 4:** Main “big data” correlations in sentences and their imaginability (topics #19 and #4).

The imagery used to illustrate the articles also show a shift in focus with regards to the previous set: in both topics that compose this theme there is a prevalence of pictures of protagonists (50 percent for topic #4; 31 percent for topic #19), which involves a wide mix of business leaders, politicians, industry experts, or even thinkers; also, given the opinionated nature of these articles, we can see portraits of authors and/or those interviewed. According to

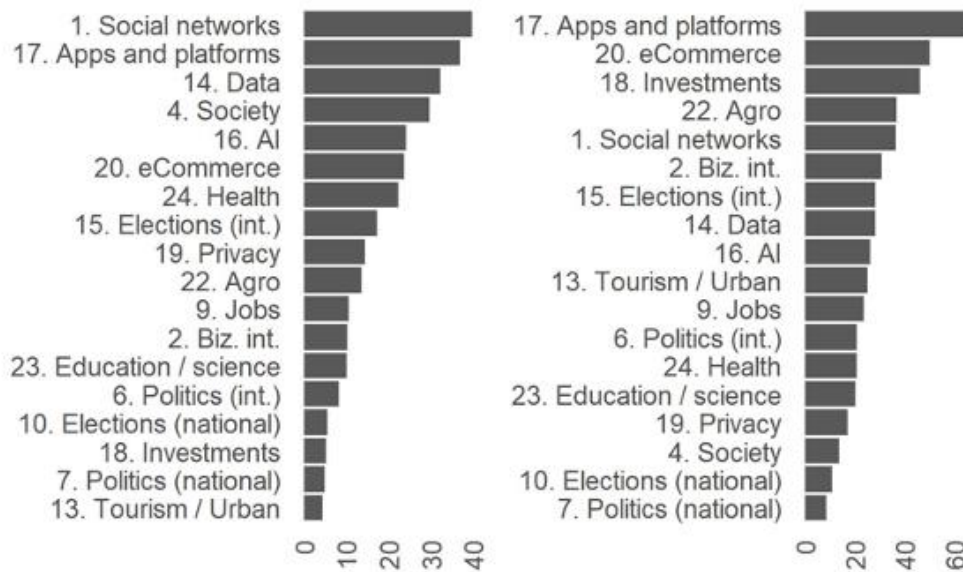
Pentzold, *et al.*, “presumably, that way big data shall be given a human face” [18]. The second most frequent picture categories are the use of applications and technology (17 percent for topic #19) and the representation of “human machines” that suggest a conversion between humans and robots (27 percent for topic #4).

### ***Artificial intelligence, applications, and algorithms***

The next cluster of articles wherein big data is thematized discuss technologies, apps, and algorithms. In topic #16 prevail words that belong to the 2.0–5.0 technologies vocabulary, such as artificial intelligence, the Internet, IOT, or 5G, while topics #1 and #17 include in their prominent words several references to social networks, such as Facebook, Twitter, and other applications and platforms, like Netflix or Youtube. They all share top terms that are characteristic of digital consumption — such as user, profile, networks, personal — as well as others associated with the quantification of data, million, data, information. By reviewing the top-ranked articles, we can see these prominently make focus on technologies that build upon or alongside big data, or applications that could be understood as paradigmatic cases of it, with headlines such as “IoT: trends and challenges” or “AI: the new ingredient in mobiles” (topic #16), and “Facebook is questioned about Cambridge Analytica” or “Everything Netflix knows about us” (topics #1 and #17).

In these articles big data is not usually in the focus; rather it is part of the social and technological background for other developments. Instead, algorithms are mentioned more frequently (as observed in [Figure 5](#)), as platforms’ hidden or backstage techniques. An article entitled “This is how the mysterious algorithms Google, Facebook, Netflix, and Twitter work” clarifies in its body that “they are like the Coca-Cola formula. We all know it exists, but only few know it. Algorithms serve as a bridge between the machines’ actions and results”. This refocusing provides an interesting comparing perspective, since algorithms are presented as agents with an active role in the creation and manipulation of data. An article under the heading “Algorithm: The new Big Brother is here to stay” states in its epigraph “They are a set of instructions with rules to get the consumers’ attention”. Another entitled “The hidden power of Facebook’s Likes” states that “The 1.7 billion people who use Facebook per month and generate billions of publications and ‘likes’ make up what is called ‘graph’, from which information can be drawn to make inferences, such as predicting user attributes and understanding their behavior”. In these cases, actual goals are explicit, and it could be hypothesized that it makes more sense to stress the idea that data must be actively collected and manipulated to make it valuable, than insisting that the data “is out there” and that it should simply be “mined” — a key idea in big data rhetoric (Portmess and Tower, 2015; Puschmann and Burgess, 2014). In

this sense, Elish and boyd (2018) have suggested that there is a tension between big data’s rhetoric and algorithms. In recent years, the corporations that were once seen as in the forefront of the former are trying to rebrand towards the latter, as a way to revalue the sophistication of algorithms and data analytics.

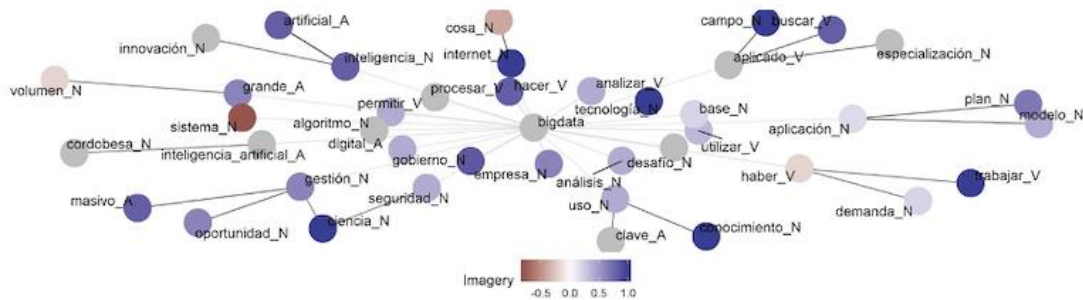


**Figure 5:** Percentage of articles including “algorithm” (left); percentage of articles including “platform” (right).

Among the first words that most correlate with big data in the sentences from these topics (see [Figure 6](#)) there is approximately a 40 percent overlap with those observed in the first set — tool, application, decision, million. This could be indicative of a common core vocabulary closely related to the “promise” of big data, no matter the larger context. As per the rest, in topic #16 (IA), appear more words related to new technologies (*e.g.*, could, intelligence, artificial), which rank slightly higher in the imagery lexicon; and in topics #1 and #17 (social networks, apps, and platforms) appear words that resemble data analytics. These semantic fields overlap in words that resemble the use of new technologies for analytic purposes — IA, algorithm, allow, digital, analyze, real, reality, company — thus framing the promise of big data to a larger ensemble of technologies. A taste of this can be glimpsed in affirmations like “AWS and Microsoft dispute the primacy in the field of artificial intelligence, which is the main instrument for creating algorithms



that serve to guide the immense mass of information (big data) produced by the 3,500 million Internet users in the world”, or “Global communications infrastructure and data storage in the cloud (big data), machine-to-machine communication technologies (the Internet of things), large-scale parallel processing, and new and more powerful hardware systems and algorithms accelerate the arrival of a form of super-intelligence, which generates ever greater and dramatic impacts on a global scale”.



**Figure 6:** Main “big data” correlations in sentences and their imaginability (topic #16).

When it comes to pictures that illustrate AI articles (topic #16), there is a preeminence of the process of datafication (37 percent) and human machines (29 percent); and for articles involving social networks and applications (topics #1 and #17), pictures of protagonists and application screenshots (40 percent and 65 percent, for topics #1 and #17). As expected, company logos are the second most frequent pictures in these topics (21 percent and 12 percent, respectively).

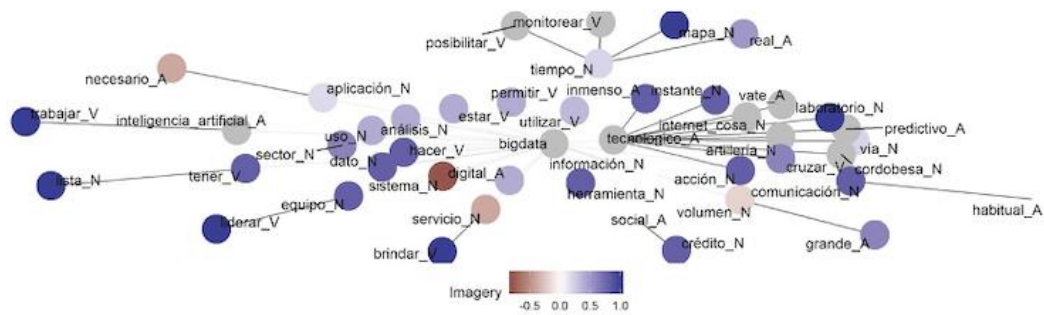
### *Big data in politics*

The next group of articles include those that rank higher for topics we’ve labelled as politically related. In them, big data is treated as part of the political agenda, with headlines such as “[Argentinian President] Macri in China: contacts with Xi Jinping and Putin”, “Macri closed the G20 summit”, or “[Province] Jujuy thinks in the future: public and private sectors are committed to the development of Big Data”; or as a key component in electoral campaigns and partisan strategy, with headlines like “Campaign’s final days” and “Trump’s triumph meaning”, or “Big data: [Political party] Cambiemos’ recipe to take advantage over [Candidate] Cristina Kirchner in

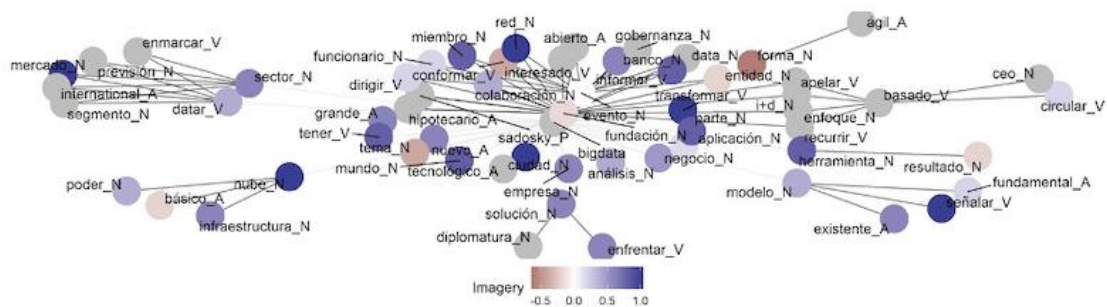
October” and “Technopolitics to change political communication and election campaigns”. In both cases, there are two topics that address them respectively in the international and national contexts (for the former, topics #6 and #7; for the latter, topics #15 and #10, although topic #15 is not that clear or univocal).

When big data is treated as an object of interest for politics, there are not usually substantial definitions. In fact, if we turn our attention to the words that correlate with big data in sentences from these articles (see [Figure 8](#)), we’ll notice two big groups of words: some technological terms, wherein big data is yet another development among others, such as artificial intelligence, cloud, or platforms; and some terms that resemble political and institutional actors, such as foundations, bank, functionary. If we intersect these words with those that correlate to big data in what we call the “big data in focus” set, we’ll notice that there are no terms that are most likely to appear in a general definition. Thus, from these articles is not clear which task, promise, or benefits big data would bring to politics. Quite the contrary, it seems to be addressed as a political challenge. One should step out of mass media communication systems and into the political communication realm to look for reasons and expectations for such interest, in a resolution from the *Argentina Official Bulletin (Boletín Oficial de la República Argentina)*, in the considerations for the creation of a national observatory of big data, it is stated:

That we are witnessing a revolution in the treatment and production of massive amounts of data, from the traffic generated on the Internet and the use of smart devices. That the challenge is the development of new forms of data processing that cannot be analyzed using traditional tools or processes. That, this set of data of great volume, high speed and/or high variety of information, generated on the Web and through the use of intelligent devices, which demands new forms of processing and that will influence in decision making and process optimization, is called Big Data. [[19](#)]

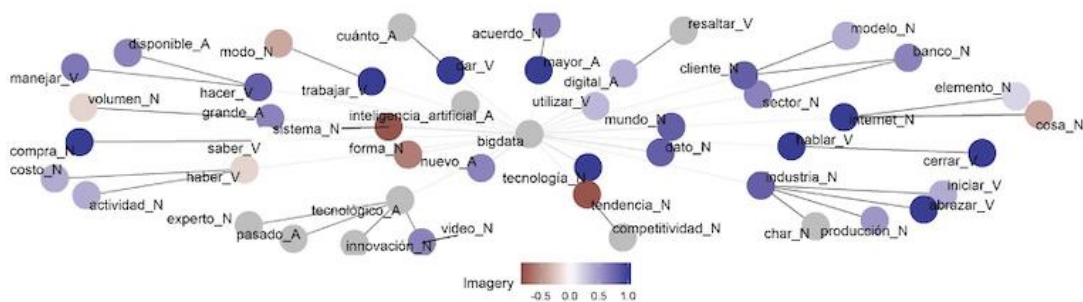


**Figure 7:** Main “big data” correlations in sentences and their imaginability (topics #1 and #17).



**Figure 8:** Main “big data” correlations in sentences and their imaginability (topic #6 and #7).

A much clearer task for big data can be inferred in articles about elections and political parties’ strategies. In these communications, big data is identified with the task of tightening the relations between parties and electorate, thus it is much closer to the challenges of persuasion and government/opposition tension, than to issues in the state/civil society, such as open government. When these articles mention big data, these sentences usually have an explicit definition that takes elements from the core definition mentioned above (see [Figure 9](#)). This could be indicative of the need to explain briefly to the general audience what big data is, when reporting about its use in political campaigns.



**Figure 9:** Main “big data” correlations in sentences and their imaginability (topic #10 and #15).

Going forward, it would be risky to extrapolate other meanings for big data, because of the timely and contextual nature of both national and international articles. Topic #10 is much more focused on the Argentinian context, with President Mauricio Macri (2015–2019) the most named political actor, followed closely by opposition leader, former President Cristina Kirchner (2007–2015). Articles that rank higher for topic #15 are a mix of local and international news with a loose interest. Macri and Kirchner’s presence is shared with Donald Trump, Hillary Clinton, and the CEOs of Facebook and Cambridge Analytica, Mark Zuckerberg and Alexander Nix (see [Table 2](#)). In the Argentinian case, the articles refer mostly about the incorporation of big data by Mauricio Macri’s presidential campaign in 2015, when the myth of being the first (local) President based in social media arose (Galup, 2019), and further campaigns while in office. In this context, “using big data” was meant as an accusation from the opposition, denouncing political surveillance, by exploiting censal and fiscal information, and political persecution in conjunction with major media groups. All of these political and business figures are put in focus in the pictures that illustrate the articles, making “protagonist people” the most frequent picture category across the four topics (93 percent for topic #10, 44 percent for #15, 91 percent for #6, and 78 percent for #7).

**Table 2: Political actors mentioned in articles (topic #10 and #15).**

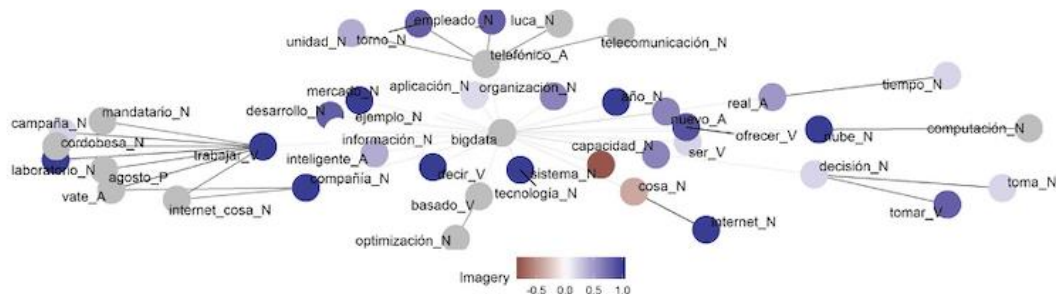
Topic	Total	Macri	Kirchner	Trump	Clinton	Nix	Zuckerberg
1. Social networks	128	18.8%	8.6%	33.6%	6.3%	13.3%	19.5%
2. Biz. int.	3	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
4. Society	18	27.8%	0.0%	44.4%	5.6%	0.0%	22.2%

6. Politics (int.)	31	61.3%	0.0%	35.5%	3.2%	0.0%	0.0%
7. Politics (national)	8	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
9. Jobs	7	42.9%	14.3%	42.9%	0.0%	0.0%	0.0%
10. Elections (national)	147	48.3%	43.5%	6.1%	2.0%	0.0%	0.0%
13. Tourism/urban	11	36.4%	36.4%	27.3%	0.0%	0.0%	0.0%
14. Data	5	20.0%	20.0%	20.0%	0.0%	0.0%	40.0%
15. Elections (int.)	46	21.7%	10.9%	28.3%	26.1%	6.5%	6.5%
16. AI	4	0.0%	25.0%	25.0%	0.0%	0.0%	50.0%
17. Apps and platforms	0	—	—	—	—	—	—
18. Investments	6	33.3%	50.0%	16.7%	0.0%	0.0%	0.0%
19. Privacy	31	45.2%	22.6%	16.1%	0.0%	3.2%	12.9%
20. eCommerce	4	50.0%	0.0%	25.0%	0.0%	0.0%	25.0%
22. Agro	4	25.0%	0.0%	50.0%	0.0%	0.0%	25.0%
23. Education/science	2	0.0%	0.0%	50.0%	0.0%	0.0%	50.0%
24. Health	1	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%

### ***Big data in business***

The last set of articles report the incorporation of big data in different business areas. Articles that ranked higher for topic #2 have a broad scope, reporting about general innovations in 4.0 industry and business intelligence, and also transformations of the job market. The main theme of articles that rank higher in topic #9 feature headlines such as “What is 4.0 industry and when will it arrive Argentina?”, “What opportunities does AI bring to enterprises?”, “Big Data, getting closer, and in great need of specialists”, and “Which professions will secure you jobs in the future?”, “Curiosity is key for tomorrow’s jobs”, respectively. There is the topic #23, whose highest ranking articles do not allow us to identify a clear cut unique theme, but a mix of reports about the use of big data in the local academic community, big data capacitation initiatives, and the quantification of sports through the exploitation of big data. In terms of the style, these three sets of articles have characteristics that resemble the “big data in focus” set, including sub headers with rhetorical questions about what is big data, addressing to a general audience, and

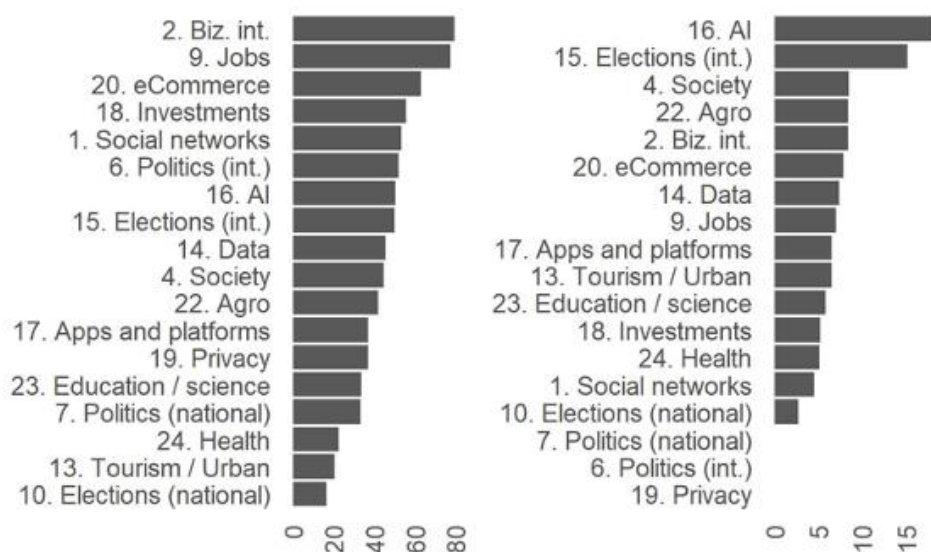
including short definitions that treat it as a tool for decision-making. In fact, if we intersect the words that correlate to big data in sentences that explicitly mention it from top articles of these topics (Figure 10) and from topic #14, we'll see a huge overlap; the semantic fields main differences are found on the second grade relations. The focus is on big data's potential value, an epistemic promise as noted earlier. In words of van Rijmenam [20]: "... data in itself is not valuable at all. ... The value is in how organizations use that data to create information-centric companies that base their decision making on insights derived from data analyses." Yet, how to create value from big data is not a trivial thing. Schmarzo (2013) proposes five business models for big data adoption, ranging from business monitoring, that is, the use of basic analytics to assess performance, and going through insights, optimization, data monetization, to business metamorphosis, that is, the analysis of customers' usage patterns, product performance behaviors, and overall market trends to create new services in new markets. However, most of the reviewed articles in this category do not mention any of these detailed ways of engaging with big data, but instead reinforces the hype, with general and unspecific assertions about data's potential value, while also claiming the urgency and acknowledging the difficulties that local industry faces, such as the lack of qualified staff and investments opportunities, or general economic uncertainty [21].



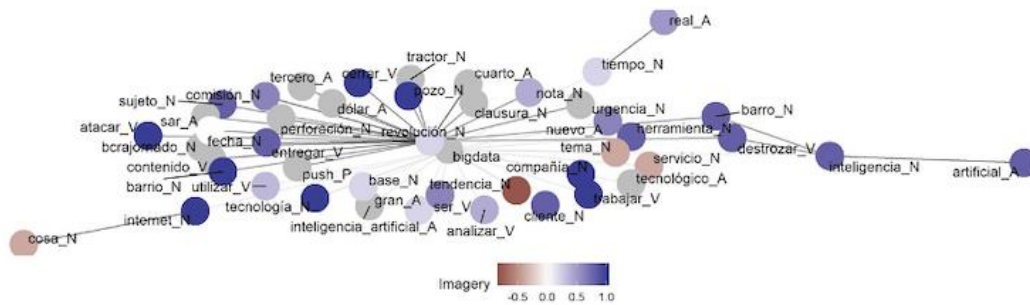
**Figure 10:** Main “big data” correlations in sentences and their imaginability (topics #2, #9 and #15).

The other articles conforming this set are those that rank higher for topic #18, which include reports about big data related investments in the local industry, entitled for example “Movistar invests US\$4M in Argentina to launch cloud services” or “IBM created a new enterprise and will open offices in Argentina, Chile, Colombia, Costa Rica and Mexico”. For topic #20, articles were more focused on the transactional side of big data, reporting applications in taxes, e-

banking, and e-commerce, including “[FISCO] will have an economic profile of each person and company” or “Mobiles as wallets”. Topic #22 focuses on the transformations on the agro sector, with headlines such as “The role of artificial intelligence in agriculture” and “Why big data will be central for the milk industry”. Topic #24 deals with health and medical related news, where the promise of big data enabling better medical attention and prevention; “Health and big data: More data, better care?” and “Big data and health: how new technologies will change medical care”. Topic #13 includes articles about “2.0 tourism” and urban planning; “Queries about national tourist destinations are increasing” or “Through ‘big data’ they seek to improve urban mobility in the city of Neuquén”. It must be noted that in this set it is common to see non-editorial pieces, and contents that could have been reproduced from corporate press releases. Here, big data is reported as part of the digital transformation of business areas, a theme that we can find in other topics but is predominant in these (Figure 11). In the sentences that mention “big data” (Figure 12), the term correlates less with those in the core definition, and much more with other technologies, maybe because of its incorporation in a larger list of technical developments. An example of this can be read in the subheader of the article “4.0 industries, a mandatory revolution for companies”: “Technologies such as artificial intelligence and virtual reality, together with big data, offer optimal conditions for a fourth industrial revolution. The industries that apply them will increase their productivity by more than 25% in the next five years. What do companies have to do not to be left out?”. Again, one can speculate that this lack of definition may be caused by big data triviality within business, or because it is meant to create an unspecified sense of urgency in the reader.



**Figure 11:** Percentage of articles including “digital” (left); percentage of articles including “revolution” (right).



**Figure 12:** Main “big data” correlations in sentences and their imaginability (topics #18, #20, #22, #13, and #24).

With the exception of topics #2 that refers to new technologies, in which prevail the pictures of the process of datafication, and of topic #9, where the dominant illustrations focus on people (protagonists, or studying or working), in all these topics the most recurrent pictures are about big data application contexts, defined “... by the conspicuous absence of indexical images of big data as they visualize the plethora of application contexts in which big data do or potentially can play a role” [22].



## Conclusion

Big data is a novel phenomenon, far more complex than what is involved by the technological problems it raises, such as those alluded with the three v’s. Above all, big data is a socio-cultural phenomenon that has been consolidating alongside the technological development, one that involves new actors and dynamics in knowledge production, and new epistemic claims about what is knowable and actionable in society (Becerra and Alurralde, 2017; Burrows and Savage, 2014; Halford and Savage, 2017). Big data is also a socio-cultural



construction, in the sense that it conforms a polyphonic and fragmented discourse where fantasies, fears, and anguishes about this transformation and its potentialities are projected. This is what we've called its "semantics", a limitation of the possible social meanings attributed to big data, which make it a topic for communication (Luhmann, 2007). To explore, systematize, and critique such semantics is a novel challenge for the social sciences.

In this work we aimed at describing how big data is framed and thematized by the mass media, focusing on the case of the Argentinean digital press. We were able to perform an automatic topic exploration along with qualitative and quantitative content analysis, and described how big data is portrayed in six types of articles: wherein big data is the main topic, wherein the focus is put on the risks of big data and new technologies, when reporting about AI and applications, politics related news, and sectoral business news. Throughout these framings we proposed three general inquiries.

First, we aimed at identifying tasks and promises, and also risks and threats, associated to big data (*RQ1*). Our analysis suggests that big data is mostly associated with an epistemic promise and a socio-technical premise: there is something else to know, something that will enable better decision-taking, and this is possible because of the availability of large volumes of data. This promise is best presented in the articles that focus on big data, but does not change substantially in articles with other topics. What indeed changes is the level of explicitness of the promise/premise pair, the focus on big data among other technological developments, and the anchoring of the cases and examples. The clearest cases of such shifts in the framing can be observed in articles that render big data as a risky phenomenon, wherein the promise is not that prominent, perhaps because of the need to focus on possible abuses of data, and the premise itself is challenged, since privacy is the limit to the availability of social data; or in articles about AI and applications, wherein the promise is mediated by the algorithms, a topic with a competing rhetoric. All these topics appeal to the general interest. On the other hand, in topics such as politics or business, there seems to be two types of articles: those close to a particular case, about a political party strategy, or when showcasing some business success story, wherein it is easy to identify big data to its promise; or those with a more general scope where big data as a novel phenomenon is reported, as part of the political agenda, or in a reports about innovations in some business area, wherein what big data offer is not so clear because either it is treated as a valuable phenomenon in itself, or it is mentioned along with several other technological developments thus becoming less specific. Interestingly, we could not find any cases wherein the epistemic capabilities of big data were questioned. Furthermore, the premise is in itself never criticized. Indeed, we live in a world of data, but its availability and accessibility is not a given, nor an unquestionable phenomenon. In fact, it is

the cause for new divides and new inequalities (Andrejevic, 2014; McCarthy, 2016; Taylor, 2016) — an issue hardly reported.

Second, we were interested in exploring the narrower semantic context of the sentences that explicitly mentioned big data, in order to assess the generality or specificity of its anchorage and measure its imagery/abstractness level (RQ2). Results show the prominence of a common set of terms that link big data to the aforementioned promise and premise, such as tool, application, information, decision, data, and millions. These widen to different topics and are linked to other terms. Word correlations indicate that big data is usually more mentioned alongside concrete and easy to imagine terms, perhaps because of the need to anchor it in a more solid ground, even in more general topics. We also calculated an imagery score per semantic context by applying an annotated lexicon, and although, as expected, a few topics involve more concrete ideas — tourism and urbanism related news, education and sports, business, health, agro — than others — society, privacy, commerce — yet we cannot affirm any significant variance across topics.

Finally, we aimed at complementing these analyses with a classification of the images that illustrate the articles (RQ3). The most general topic (#14) is the only one wherein an abstract visualization — in this case, about the datafication process — is preeminent, followed by pictures of technical infrastructure. In both cases, what is becomes visible is the “premise” of big data. Interestingly, we found no figures that resemble the “natural” metaphors of “data deluge” that the literature points out (Lupton, 2014; Portmess and Tower, 2015; Puschmann and Burgess, 2014). One can hypothesize that focusing on the availability of data and the robustness of its supporting architecture are far more beneficial for a “promotional” discursive strategy than insisting on floods or tsunamis. Then, whenever the articles are more case-oriented, relevant photographs were placed: in the case of politics, of the figures involved; in the case of businesses, of sectoral applications; in the case of AI and apps, a mix of company leaders, logos, and screenshots. This variance suggest that big data is a very versatile idea to illustrate. In these last cases, we can agree with Pentzold, *et al.* (2019) on the rather neutral stance of pictures that illustrate big data articles.


In the theoretical background of this paper we also mentioned that keen analysis of big data, from a social and cultural perspective, usually stress both its epistemic promise of a radical new way of constructing knowledge and actionable insights, and its universal reach across several social areas. But the news’ discourse on big data has some elements that could point to its pretended universal reach — although this should and will be better treated in a comparative setting. The amount and the variety of sectoral and specific topics we could fit, that proved plausible for a coherent interpretation, could be indicative of the versatility of big data as a topic, and also of its relevance

for such different communicative spaces. This also means openness in terms of the target of the communication, as Michael and Lupton state: “... The ‘public’ of big data is a constantly moving virtual artefact that has varying meanings and constituents depending on which actors are seeking to define this entity and at which point in time they are seeking to do so” [23]. In the case of mass media, this communication is even tighter, since it is also required some alleged identification between the sender and the receiver, without which it would be impossible to make a topic relevant, urgent, or provocative. In this case, the premise that we are living in a world of data could be thought of as an observation of society.

Our exploration contributes to a more detailed knowledge on how news media social systems make sense of novel and abstract phenomena, such as big data. The usefulness of reconstructing big data semantics, from an empirical corpus, in terms of a promise and a premise was unanticipated. This is interesting because it moves away from more technological definitions, such as the three v’s, that, although present in the corpus, do not seem to be within the core rationale of the message constructed by the news, perhaps because of their need to make it relevant and interesting to the general public. Furthermore, this way of analyzing the representation of big data in news articles, along with the consideration of its scope — ranging from more general and broad articles, to narrower and case-based reports — could be explored as a preliminary theoretical typology.

We also aimed at offering a new case study for mix methodologies designs that combine qualitative and quantitative discourse analysis with text mining techniques. In our case, this was done following a three-level exploration — articles, sentences, illustrations — which render coherent results, and provided support for some of the inferences that we did during the analysis, such as proposing a topic from a set of prominent words. To the best of our knowledge, similar attempts of surveying media reconstruction of big data (Elish and boyd, 2018; Lupton, 2014; Pentzold and Fischer, 2017) have relied on purely qualitative analysis.

This study had several limitations. As any case study, it relied on a limited corpus, and the specifics of the Argentine case cannot be simply generalized to a larger discourse about big data without further considerations. Arguably, because of it being an English term with no translation in Spanish, “big data” is a signifier with a much reduced interpretability than in English. Future research could expand this methodology to other regional corpora. Also, the way we implemented quantitative analysis — the scoring of imagery of semantic fields — relied on methodological decisions about natural language processing for which there are no canonical guidelines. This is also true for the mix of qualitative and quantitative analysis, and even more for their complementation with computational techniques. Since this is a novel area,

further methodological explorations are required. Another important limitation is that our analysis did not track shifts over time and between groups of articles, which is a serious limitation given the velocity at which technological developments occur. In order to make sense of these, future research could explore other models, such as dynamic topic models or structural topic models (Blei and Lafferty, 2007). 

## About the author

Gastón Becerra is an Assistant Professor in the Social Sciences Faculty at the Universidad de Buenos Aires (Argentina), and assistant researcher at CONICET (National Scientific and Technical Research Council, Argentina). He holds a doctorate in philosophy and a bachelor's degree in sociology, both from the Universidad de Buenos Aires, and a master's degree in epistemology and history of sciences from the Universidad de Tres de Febrero. His research focuses in big data, both as a social-technological complex phenomenon, and as a methodological and epistemological challenge for social sciences.  
E-mail: gastonbecerra [at] sociales [dot] uba [dot] ar

## Notes

- [1.](#) The original text was first published in 1995 before Web 2.0 technologies and social networks became more available and common. Lately these technological innovations have led to microsegmentation, user experience customization, and the introduction of interaction mechanisms, which all have been inducing changes in mass media (Lüders, 2008; Raimondo Anselmino, 2012). However, these changes could actually have caused the de-differentiation of the mass media system, or they might have subverted how this system typifies its reader; its message is yet a very difficult thesis to accept.
- [2.](#) Luhmann, 2000, p. 12.
- [3.](#) Luhmann, 1995, pp. 154–156.
- [4.](#) Luhmann, 2000, p. 12.
- [5.](#) Luhmann, 1995, p. 164.
- [6.](#) Zikopoulos, *et al.*, 2012, p. 3.
- [7.](#) Quoted in Portmess and Tower, 2015, p. 2.

[8.](#) Pentzold, *et al.*, 2019, p. 25.

[9.](#) This is a methodological sensitive decision. It is our choice to circumscribe to explicit uses of this criterion since we want to clarify its semantic boundaries. This is not against other textual criteria that could also be relevant for big data thematization (*e.g.*, “large sets of data” or “artificial intelligence”). Using alternative criteria may be useful to acquire information in contexts where the term is not being used, despite the fact that, for an observer or analyst, they are working with big data (Taylor, *et al.*, 2014).

[10.](#) This certainly delegates an important methodological criterion to Web search engines. However, this is part of a social scenario we are interested in, namely, the performative role of these algorithms in culture (Kitchin, 2017).

[11.](#) Blei, 2012, p. 79.

[12.](#) We tried the model with 9, 12, 24, 50, 100, 150, and 200 topics. Although a higher number of topics performed better in statistical analysis (*e.g.*, held-out likelihood test), when hand-coding we found overlaps that could merge on a smaller solution. We confirmed these assumptions by calculating Hellinger distances on topics vocabulary. It must be noted, also, that previous works with similar corpus size (*e.g.*, Baumer, *et al.*, 2017) suggest an even smaller number of topics to avoid overfitting.

[13.](#) One topic (#3) mostly refers to terms used in a particular event that was reported as non-editorial (commercial) content in several sources; other topics (#5, 8, 12) consisted mostly of words related to the news genre and interactivity of articles; one topic (#21) loosely resembled political economics, like financial or tax policies, but later we could not extract a coherent theme from the sampled articles and showed no clear correlation with any other relevant topics; and, finally, the last topic (#25) was very inconsistent during our tests and showed no robust statistical regularity nor semantic clarity.

[14.](#) The convergence of topic modelling with these types of techniques that were mostly developed on interpretative traditions, such as grounded theory (Glaser and Strauss, 1967), has spawned an interesting debate with good empirical tests and epistemological considerations (Baumer, *et al.*, 2017; Berente and Seidel, 2014; Nelson, 2020).

[15.](#) Each picture was classified by two independent researchers, unaware of the topic exploration, who discussed conflicting cases. The author would like to thank Lic. Laino for her assistance through the whole process. During these discussions, some categories were merged, *e.g.*, datification + datified individuals (in “processes” category); and IT workforce + nerds/geeks (in “people” category).

[16.](#) Over the years, Laney’s formulation has been expanded by other authors aiming at a much more exhaustive definition, such as in a series of publications IBM made a strong case for considering the validity and veracity of data as additional v-words (Kobielus, 2013), while van Rijmenam (2014) expanded these to seven by adding variability, visualization, and value, the last one referring to the transformation of big data into insights for the creation of profit. Others, less enthusiastic, have parodied this approach by adding words like vagueness, vogue, or vanilla, reaching a 42 v-words list (Shafer, 2017). Consensual definitions of big data, based on surveying experts from business and technology, include some of these v-words too but also require mentioning specific technologies and/or analytical approaches, such as Hadoop or machine learning (De Mauro, *et al.*, 2015; Ward and Barker, 2013; Favaretto, *et al.*, 2020). Both approaches have been criticized from a sociological point of view for focusing on characteristics of data and its handling, instead on its social significance (Uprichard, 2013; Lupton, 2014); in this vein, Lupton prefers to speak of the “thirteen ps” of big data, including words like portentous, perverse and political.

[17.](#) Zikopoulos, *et al.*, 2012, pp. 6–7.

[18.](#) Pentzold, *et al.*, 2019, p. 19.

[19.](#) *Argentina Official Bulletin (Boletín Oficial de la República Argentina, Resolution 11-E/2017*, p. 1.

[20.](#) Van Rijmenam, 2014, p. 12.

[21.](#) Interestingly, this editorial line does not reflect how experienced and mature are media organizations regarding the adoption of big data. In a local case study, Retegui (2020) shows how the digital transformation of newspapers and the use of metrics — number of reads, online interactions, digital subscription rate — are changing journalistic practices, while also raising new tensions within newsrooms. The author would like to thank an anonymous reviewer for pointing this out.

[22.](#) Pentzold, *et al.*, 2019, p. 23.

[23.](#) Michael and Lupton, 2016, p. 6.

## References

Mark Andrejevic, 2014. “The big data divide,” *International Journal of Communication* volume 8, pp. 1,673–1,689, and at <https://ijoc.org/index.php/ijoc/article/view/2161>, accessed 29 August 2021.

Eric P.S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay, 2017. “Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?” *Journal of the Association for Information Science and Technology*, volume 68, number 6, pp. 1,397–1,410.  
doi: <https://doi.org/10.1002/asi.23786>, accessed 29 August 2021.

Gastón Becerra, 2018a. “Interpelaciones entre el Big data y La Teoría de los sistemas sociales. Propuestas para un programa de investigación,” *Hipertextos*, volume 6, number 9, pp. 41–62, and at <http://revistahipertextos.org/ediciones/hipertextos-no-9/>, accessed 29 August 2021.

Gastón Becerra, 2018b. “La epistemología constructivista de Luhmann. Objetivos programáticos, contextos de discusión y supuestos filosóficos,” *Sociológica (México)*, number 95, pp. 9–38, and at <http://www.sociologicamexico.azc.uam.mx/index.php/Sociologica/article/view/1461>, accessed 29 August 2021.

Gastón Becerra and Juan Pablo López Alurralde, 2017. “Big Data y Data Mining. Un Análisis Crítico Acerca de Su Significación Para Las Ciencias Psicosociales a Partir de Un Estudio de Caso,” *{PSOCIAL} Revista de Investigación en Psicología Social*, volume 3, number 2, pp. 66–85, and at <http://publicaciones.sociales.uba.ar/index.php/psicologiasocial/article/view/2610>, accessed 29 August 2021.

Gastón Becerra and Vanessa Arreyes, 2013. “Los medios de comunicación de masas y las noticias como objeto de estudio de la sociología en la perspectiva del constructivismo operativo de Niklas Luhmann,” *Revista Mad. Revista del Magíster en Análisis Sistemico Aplicado a la Sociedad*, number 28, pp. 47–60, and at <https://www.redalyc.org/articulo.oa?id=311226876005>, accessed 29 August 2021.

Gotthard Bechmann and Nico Stehr, 2011. “Niklas Luhmann’s theory of the mass media,” *Society* volume 48, number 2, pp. 142–147.  
doi: <https://doi.org/10.1007/s12115-010-9410-7>, accessed 29 August 2021.

David Beer, 2016. “How should we do the history of Big Data?” *Big Data & Society* (4 May).  
doi: <https://doi.org/10.1177/2053951716646135>, accessed 29 August 2021.

Nicholas Berente and Stefan Seidel, 2014. “Big Data & Inductive Theory Development: Towards Computational Grounded Theory?” *Twentieth Americas Conference on Information Systems*, at <https://aisel.aisnet.org/amcis2014/ResearchMethods/GeneralPresentations/1/>, accessed 29 August 2021.

David M. Blei, 2012. “Probabilistic topic models,” *Communications of the ACM*, volume 55, number 4, pp. 77–84.

doi: <https://doi.org/10.1145/2133806.2133826>, accessed 29 August 2021.

David M. Blei and John D. Lafferty, 2007. “A correlated topic model of *Science*,” *Annals of Applied Statistics*, volume 1, number 1, pp. 17–35.

doi: <https://doi.org/10.1214/07-AOAS114>, accessed 29 August 2021.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan, 2003. “Latent dirichlet allocation,” *Journal of Machine Learning Research*, volume 3, pp. 993–1,022.

Tom Boellstorff, 2013. “Making big data, in theory,” *First Monday*, volume 18, number 10, at <https://firstmonday.org/article/view/4869/3750>, accessed 29 August 2021.

doi: <https://doi.org/10.5210/fm.v18i10.4869>, accessed 29 August 2021.

danah boyd and Kate Crawford. 2012. “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon,” *Information, Communication & Society*, volume 15, number 5, pp. 662–679.

doi: <https://doi.org/10.1080/1369118X.2012.678878>, accessed 29 August 2021.

Antony Bryant and Uzma Raja, 2014. “In the realm of Big Data...,” *First Monday*, volume 19, number 2,

at <https://firstmonday.org/article/view/4991/3822>, accessed 29 August 2021.

doi: <https://doi.org/10.5210/fm.v19i2.4991>, accessed 29 August 2021.

Roger Burrows and Mike Savage, 2014. “After the crisis? Big Data and the methodological challenges of empirical sociology,” *Big Data & Society* (1 April).

doi: <https://doi.org/10.1177/2053951714540280>, accessed 29 August 2021.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei, 2009. “Reading tea leaves: How humans interpret topic models,” *NIPS’09: Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pp. 288–296.

Andrea De Mauro, Marco Greco, and Michele Grimaldi, 2015. “What is big data? A consensual definition and a review of key research topics,” *AIP Conference Proceedings*, volume 1644, pp. 97–104.

doi: <https://doi.org/10.1063/1.4907823>, accessed 29 August 2021.

Francis X. Diebold, 2012. “On the origin(s) and development of ‘Big Data’: The phenomenon, the term, and the discipline,” *PIER Working Paper*,



number 12-037,  
at [https://www.sas.upenn.edu/~fdiebold/papers/paper112/Diebold\\_Big\\_Data.pdf](https://www.sas.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf), accessed 29 August 2021.

José van Dijck, 2014. “Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology,” *Surveillance and Society*, volume 12, number 2, pp. 197–208.  
doi: <https://doi.org/10.24908/ss.v12i2.4776>, accessed 29 August 2021.

Paul DiMaggio, Manish Nag, and David Blei. 2013. “Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding,” *Poetics*, volume 41, number 6, pp. 570–606.  
doi: <https://doi.org/10.1016/j.poetic.2013.08.004>, accessed 29 August 2021.

M.C. Elish and danah boyd, 2018. “Situating methods in the magic of Big Data and AI,” *Communication Monographs*, volume 85, number 1, pp. 57–80.  
doi: <https://doi.org/10.1080/03637751.2017.1375130>, accessed 29 August 2021.

Michael S. Evans, 2014. “A computational approach to qualitative analysis in large textual datasets,” *PLoS ONE*, volume 9, number 2, e87908 (3 February).  
doi: <https://doi.org/10.1371/journal.pone.0087908>, accessed 29 August 2021.

Maddalena Favaretto, Eva De Clercq, Christopher Schneble, and Bernice Elger, 2020. “What is your definition of Big Data? Researchers’ understanding of the phenomenon of the decade,” *PLoS ONE*, volume 15, number 2, e0228987 (25 February).  
doi: <https://doi.org/10.1371/journal.pone.0228987>, accessed 29 August 2021.

Luciano Galup, 2019. *Big data y política. De los relatos a los datos. Persuadir en la era de las redes sociales*. Buenos Aires: Ediciones B.

William A. Gamson, David Croteau, William Hoynes, and Theodore Sasson. 1992. “Media images and the social construction of reality,” *Annual Review of Sociology*, volume 18, pp. 373–393.  
doi: <https://doi.org/10.1146/annurev.so.18.080192.002105>, accessed 29 August 2021.

Giray Gerim, 2017. “A critical review of Luhmann’s social systems theory’s perspective on mass media and social media.” *İnsan & Toplum (Journal of Human & Society)* volume 7, number 2, pp. 141–154.  
doi: <https://doi.org/10.12658/human.society.7.14.M0218>, accessed 29 August 2021.

Barney G. Glaser and Anselm L. Strauss, 1967. *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Agustín Gravano and Matías Dell’Amerlina Ríos, 2014. “Spanish DAL: A Spanish dictionary of affect in language,” *Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires*, at [http://digital.bl.fcen.uba.ar/Download/technicalreport/technicalreport\\_00001.pdf](http://digital.bl.fcen.uba.ar/Download/technicalreport/technicalreport_00001.pdf), accessed 29 August 2021.

Justin Grimmer and Brandon M. Stewart. 2013. “Text as data: The promise and pitfalls of automatic content analysis methods for political texts,” *Political Analysis*, volume 21, number 3, pp. 267–297. doi: <https://doi.org/10.1093/pan/mps028>, accessed 29 August 2021.

Bettina Grün and Kurt Hornik. 2011. “Topicmodels: An R package for fitting topic models,” *Journal of Statistical Software*, volume 40, number 13, at <https://www.jstatsoft.org/article/view/v040i13>, accessed 29 August 2021. doi: <https://doi.org/10.18637/jss.v040.i13>, accessed 29 August 2021.

Susan Halford and Mike Savage, 2017. “Speaking sociologically with Big Data: Symphonic social science and the future for Big Data research,” *Sociology*, volume 51, number 6, pp. 1,132–1,148. doi: <https://doi.org/10.1177/0038038517698639>, accessed 29 August 2021.

Carina Jacobi, Wouter van Atteveldt, and Kasper Welbers. 2016. “Quantitative analysis of large amounts of journalistic texts using topic modelling,” *Digital Journalism*, volume 4, number 1, pp. 89–106. doi: <https://doi.org/10.1080/21670811.2015.1093271>, accessed 29 August 2021.

Rob Kitchin, 2017. “Thinking critically about and researching algorithms,” *Information Communication & Society*, volume 20, number 1, pp. 14–29. doi: <https://doi.org/10.1080/1369118X.2016.1154087>, accessed 29 August 2021.

Rob Kitchin, 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. Los Angeles, Calif.: Sage. doi: <https://dx.doi.org/10.4135/9781473909472>, accessed 29 August 2021.

Seth C. Lewis, Rodrigo Zamith, and Alfred Hermida. 2013. “Content analysis in an era of Big Data: A hybrid approach to computational and manual methods,” *Journal of Broadcasting & Electronic Media*, volume 57, number 1, pp. 34–52.

doi: <https://doi.org/10.1080/08838151.2012.761702>, accessed 29 August 2021.

Marika Lüders, 2008. "Conceptualizing personal media," *New Media & Society*, volume 10, number 5, pp. 683–702.  
doi: <https://doi.org/10.1177/1461444808094352>, accessed 29 August 2021.

Niklas Luhmann, 2007. *La sociedad de la sociedad*. México D.F.: Universidad Iberoamericana, Herder.

Niklas Luhmann, 2000. *The reality of the mass media*. Translated by Kathleen Cross. Stanford, Calif.: Stanford University Press.

Niklas Luhmann, 1995. *Social systems*. Translated by John Bednarz, Jr., with Dirk Baecker. Stanford, Calif.: Stanford University Press.

Deborah Lupton, 2014. *Digital sociology*. New York: Routledge.  
doi: <https://doi.org/10.4324/9781315776880>, accessed 29 August 2021.

Matthew T. McCarthy, 2016. "The big data divide and its consequences," *Sociology Compass*, volume 10, number 12, pp. 1,131–1,140.  
doi: <https://doi.org/10.1111/soc4.12436>, accessed 29 August 2021.

Mike Michael and Deborah Lupton, 2016. "Toward a manifesto for the 'public understanding of big data'," *Public Understanding of Science*, volume 25, number 1, pp. 104–116.  
doi: <https://doi.org/10.1177/0963662515609005>, accessed 29 August 2021.

Laura K. Nelson, 2020. "Computational grounded theory: A methodological approach," *Sociological Methods & Research*, volume 49, number 1, pp. 3–42.  
doi: <https://doi.org/10.1177/0049124117729703>, accessed 29 August 2021.

Christian Pentzold and Charlotte Fischer, 2017. "Framing Big Data: The discursive construction of a radio cell query in Germany," *Big Data & Society* (29 November).  
doi: <https://doi.org/10.1177/2053951717745897>, accessed 29 August 2021.

Christian Pentzold, Cornelia Brantner, and Lena Fölsche. 2019. "Imagining big data: Illustrations of 'big data' in US news articles, 2010–2016," *New Media & Society*, volume 21, number 1, pp. 139–167.  
doi: <https://doi.org/10.1177/1461444818791326>, accessed 29 August 2021.

Lisa Portmess and Sara Tower, 2015. "Data barns, ambient intelligence and cloud computing: The tacit epistemology and linguistic representation of Big

Data,” *Ethics and Information Technology*, volume 17, pp. 1–9.  
doi: <https://doi.org/10.1007/s10676-014-9357-2>, accessed 29 August 2021.

Cornelius Puschmann and Jean Burgess, 2014. “Metaphors of Big Data,” *International Journal of Communication*, volume 8, pp. 1,690–1,709, and at <https://ijoc.org/index.php/ijoc/article/view/2169>, accessed 29 August 2021.

Natalia Raimondo Anselmino, 2012. *La Prensa Online y Su Público: Un estudio de los espacios de intervención y participación del lector en Clarín y La Nación*. Buenos Aires: Editorial Teseo.

Lorena Retegui, 2020. “Métricas y cuantificación del rendimiento individual de los periodistas: Un estudio en el interior de una sala de redacción,” *Austral Comunicación*, volume 9, number 1, pp. 45–67.  
doi: <https://doi.org/10.26422/aucom.2020.0901.ret>, accessed 29 August 2021.

Tom Shafer, 2017. “The 42 V’s of big data and data science,” at <https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html>, accessed 29 August 2021.

Bill Schmarzo, 2013. *Big data: Understanding how data powers big business*. Indianapolis, Ind.: Wiley.

Milan Straka and Jana Strakova, 2017. “Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe,” *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88–99.  
doi: <https://doi.org/10.18653/v1/K17-3009>, accessed 29 August 2021.

Linnet Taylor, 2016. “The ethics of big data as a public good: Which public? Whose good?” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, volume 374, number 2083 (28 December), 20160126.  
doi: <https://doi.org/10.1098/rsta.2016.0126>, accessed 29 August 2021.

Linnet Taylor, Ralph Schroeder, and Eric Meyer, 2014. “Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?” *Big Data & Society* (1 July).  
doi: <https://doi.org/10.1177/2053951714536877>, accessed 29 August 2021.

Anton Törnberg and Petter Törnberg, 2016. “Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum,” *Discourse & Society*, volume 27, number 4,

pp. 401–422.

doi: <https://doi.org/10.1177/0957926516634546>, accessed 29 August 2021.

Emma Uprichard, 2013. “Big data, little questions?” (1 October), at <https://archive.discoversociety.org/2013/10/01/focus-big-data-little-questions/>, accessed 29 August 2021.

Mark van Rijmenam, 2014. *Think bigger: Developing a successful big data strategy for your business*. New York: Amacom.

Jonathan Stuart Ward and Adam Barker, 2013. “Undefined by data: A survey of big data definitions,” *arXiv:1309.5821* (20 September), at <http://arxiv.org/abs/1309.5821>, accessed 29 August 2021.

Cynthia Whissell, 2009. “Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Language,” *Psychological Reports*, volume 105, number 2, pp. 509–521. doi: <https://doi.org/10.2466/PRO.105.2.509-521>, accessed 29 August 2021.

Gregor Wiedemann, 2015. *Text mining for qualitative data analysis in the social sciences. A study on democratic discourse in Germany*. Wiesbaden: VS Verlag für Sozialwissenschaften. doi: <https://doi.org/10.1007/978-3-658-15309-0>, accessed 29 August 2021.

Paul C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, and George Laplis, 2012. *Understanding big data: Analytics for enterprise class Hadoop and streaming data*. New York: McGraw-Hill.

---

## Editorial history

Received 1 March 2020; revised 5 July 2020; revised 9 July 2020; accepted 10 July 2020.



To the extent possible under law, this work is dedicated to the public domain.

The promise and the premise: How digital media present big data  
by Gastón Becerra.

*First Monday*, Volume 26, Number 9 - 6 September 2021

<https://firstmonday.org/ojs/index.php/fm/article/download/10539/10220>  
doi: <https://dx.doi.org/10.5210/fm.v26i9.10539>