Sara Mariottini, Wenceslao Arroyo-Machado
and Daniel Torres-Salinas

A Brief Introduction to Big Data for Humanists

1 Brief Introduction

We usually associate big data with its cruder, more conventional and perhaps more obscure applications; those of the interconnected, data-driven world, where every interaction and every 'like' leaves a trace, every click is recorded and privacy is supervised by third parties as millions of data records are gathered from millions of users 24 hours a day. Here are some of the figures regarding this phenomenon: according to Eric Schmidt, every day we generate as much data as all the data produced by the whole of humanity in 2003 (Siegler 2010). These data are generated by the 4.66 billion active internet users who, to give an example, can publish 3.3 million posts on Facebook or perform 3.3 million searches on Google every minute (Alonso 2020). According to the predictions for 2025, there will be 163 zettabytes[1] of data in the world (Zgurovsky & Zaychenko 2020). In this context, one of the most common uses of big data is in digital marketing, although we can find it everywhere, whether politics (Pascual & Peinado 2018; Rands, 2018), finance, with its algorithms for surveillance and decision-making (Hasan, Popp & Oláh, 2020), health monitoring (Sun et al. 2020), the recommendation systems of entertainment platforms (Fayyaz et al, 2020) or sports (Torgler 2020).

There is also talk of a new research paradigm in the academic realm. Big data is changing the way we generate and analyze scientific results due to the massive generation of data, heavy reliance on technology and the widespread use of mathematical models, algorithms and artificial intelligence (AI). The era of big data is here to stay and will accelerate learning in all scientific fields. Universities and research institutes already promote interdisciplinary collaboration and stimulate "cross-fertilization" between different fields which have data science as a common axis (Galeano & Peña 2019). The increased capacity of acquisition, processing and analysis of data with the potential to reveal patterns has contributed to the connection of different scientific disciplines. Some of the most outstanding examples include the Large Hadron Collider (Ortíz 2019), radio telescopes such as

[1] Various current estimates indicate that the volume of data in 2021 stands at 44 zettabytes (van der Aalst 2016; Kugler 2018). A zettabyte is equivalent to one billion terabytes.

https://doi.org/10.1515/9783110753523-002

the Square Kilometer Array (Scaife 2020) and the NASA Center for Climate Simulation (NCCS) (Schnase et al. 2011) and the application of big data in education to analyze students (Fischer et al. 2020).

However, this sudden intrusion in many areas has caused some bewilderment. The term 'big data' is still somewhat confusing for researchers, as most associate it with its most basic objectives such as data collection and processing of operations and do not have a clear overview of its scope and implications (Favaretto et al. 2020). Moreover, there is a certain sense of uneasiness towards big data as it is a cultural phenomenon in a state of constant change and evolution and the use of this concept as a buzzword further aggravates its conceptual vagueness. Therefore, the aim of this chapter is to offer a synthetic vision of what is understood as big data to serve as a starting point for researchers in the field of humanities.

2 Characteristics and Definition of Big Data

The raw material of big data is obviously the data, which is understood as a symbolic representation of an attribute which may be qualitative or quantitative. In the case of big data they have been translated into a digital format allowing their use and processing and are catalogued to facilitate their processing and analysis using multiple techniques. The magnitudes of big data require the use of significant computational resources. Another fundamental aspect of big data is that it may be collected effortlessly through all kinds of devices such as smartphones, social media, sensors, etc. These gadgets determine the essential aspects of big data, namely its exaggerated volume, the speed of its collection and its variety (Laney 2001; Ward & Barker 2013). The EU[2] defines big data as:

> large amounts of different types of data produced from various types of sources, such as people, machines or sensors. This data includes climate information, satellite imagery, digital pictures and videos, transition records or GPS signals. Big Data may involve personal data: that is, any information relating to an individual, and can be anything from a name, a photo, an email address, bank details, posts on social networking websites, medical information, or a computer IP address.

However, although the characteristics of data are clear to some authors, there is no univocal definition of big data. As a result, new characteristics are added such

[2] European Commission, Directorate-General for Justice and Consumers, The EU Data Protection Reform and Big Data, Publications Office, 2018, https://data.europa.eu/doi/10.2838/190200.

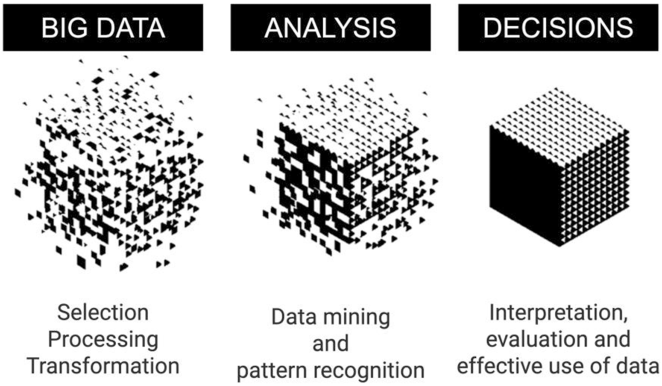


Figure 1: Metaphor of the essence of big data and its main processes.

as its capacity not only to be captured but also to be stored for permanent updating and continuous exploitation. The latter analysis processes are related to data visualization and prediction and involve the use of methods that extract value and meaning from the data (Figure 1). These analysis techniques are oriented towards three main objectives: the search for patterns, the identification of associations and the development of models that allow us to make forecasts. As can be seen, big data is a complex discipline. A simple definition that captures the above concepts is provided by the Gartner IT Glossary,[3] which defines big data as:

> high-volume, high-velocity and/or high-variety information assets that demand cost effective, innovative forms of information processing that enable enhanced insight, decision-making, and process automation.

This definition offers a framework consisting of three facets: volume, velocity and variety (physical characteristics of the data), to which we can also add veracity and value, i.e., the data must be of good quality, relevant and reliable and must allow us to achieve our objectives, and the data must provide added value to help us decide or understand a phenomenon holistically. These five characteristics make up what in big data literature has come to be known as the 5 Vs (Favaretto et al, 2020). Some authors go even further and talk about the 7 Vs, adding volatility and validity to the above (Khan, Uddin & Gupta 2014). These latter two attributes refer to the need to consider the feasibility of a big data project and the form of data presentation. While the view of the 7 Vs is somewhat Manichean and synthetic, it effectively introduces the attributes, processes and actions needed in any

[3] https://www.gartner.com/en/information-technology/glossary.

big data project. Nonetheless, certain sectors of the social sciences consider this definition to be too technological, with a certain utopian character (Kitchin & McArdle 2016; Gandomi & Haider 2015).

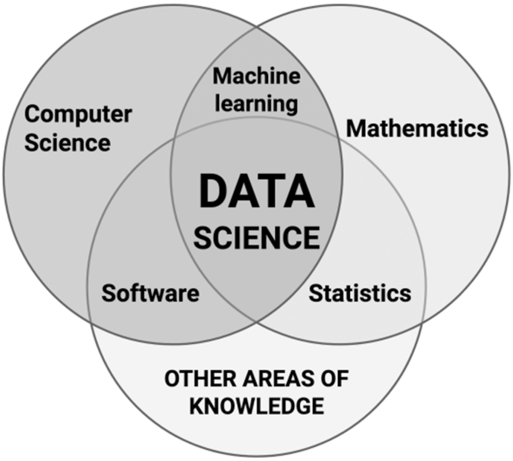


Figure 2: Disciplinary relationships of data or data science.

Precisely because not all big data share the same characteristics, it makes sense to use a purely 'technological' definition. However, from a humanistic viewpoint this definition can be improved by emphasizing the human side of data and promoting re-humanization of the digitized social product that we have become for big data. A very significant percentage of big data is devoted to studying the hyper-connected population of the so-called turbo-capitalism (Luttwak 2000), excluding from its discourse all individuals alien to big data flows. At the same time, algorithms are social products and can also reflect the prejudices, social stigmas and ineptitudes of the developer (Mac 2021; Jiménez de Luis 2021). Therefore, a merely technological definition of big data provides us with a framework, but at the same time it obviates an ethical and humanized approach, overlooking the fact that data are generated by people and algorithms are sometimes simply a mere aggregate of emotions.

As we can see, when we talk about big data we are faced with a complex phenomenon that has given rise to a new multidisciplinary field called data science (Figure 2). Data science has its origins in computer science and maintains a close relationship with AI and the internet of things, that increasingly palpable world where every activity, every click and every step is recorded and stored and even the most unassuming and irrelevant gadget (a light bulb, a refrigerator, etc.) can generate data and be connected to the internet. Data science is therefore an intimate combination of technology and mathematics aimed at understanding human behavior and making it increasingly predictable. From our standpoint, with the advent of big data we are faced with an epistemological and ontological problem that opens up a world of opportunities for social sciences and humanities in terms of definition, methodology, deconstruction and new integrations. The following basic introduction outlines some of the basic elements of working with big data.

3 Methodological Elements of Big Data

3.1 Main Types of Data and Formats

Until the first definition of big data appeared in the 1990s, all data was, in effect, small data and therefore it did not need to be labeled as such (Faraway & Augustin, 2018). Due to the difficulties of generating, processing, analyzing and storing data, it was produced in a very controlled manner using samples that limited its life cycle and size. Today, big data is generated continuously and is intended to be flexible in scope and scalable in its production. Although big data may claim to be exhaustive, it is nevertheless a representation and a sample of the social reality limited to a specific moment in time (Mayer-Schönberger & Cukier 2013). For this very reason, the data captured are conditioned by the following aspects (Li 2015):

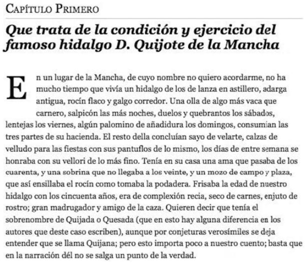
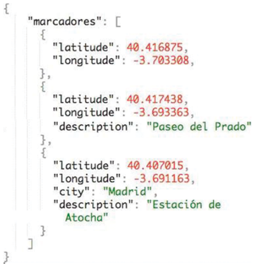
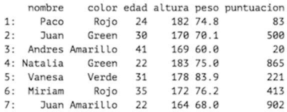
- The data collection framework (data collection devices and/or sensors, the parameters used, etc.)
- The technology/platform used (which can produce variations and biases in the data generated)
- The context in which the data are generated (data are always considered in relation to the circumstances)
- The data ontology used (how they are calibrated and classified)
- The regulatory environment governing privacy, data protection and security

Once we know what conditions the data, we can move on to consider the different types of data. Big data can also be classified into three classes according to the structure (Table 1):

- Structured data: data stored in tables with a well-defined length and format which can be easily sorted and processed by any data management tool. Examples of structured data include dates, data sheets and databases.
- Semi-structured data: information that is not regular and therefore cannot be managed in a standard way. Examples of semi-structured data include HTML, JSON and XML.

– Unstructured data: binary data that has no identifiable internal structure. This is a massive, disorganized conglomerate of data that has no value until it is organized and stored. Examples of unstructured data include images, videos, audio files and PDFs.

Table 1: Classification of big data.

Unstructured data	Semi-structured data	Structured data
		

Capítulo Primero

Que trata de la condición y ejercicio del famoso hidalgo D. Quijote de la Mancha

En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicón las más noches, duelos y quebrantos los sábados, lentejas los viernes, algún palomino de añadidura los domingos, consumian las tres partes de su hacienda. El resto della concluían sayo de velarte, calzas de velludo para las fiestas con sus pantuflos de lo mismo, los días de entre semana se honraba con su vellori de lo más fino. Tenía en su casa una ama que pasaba de los cuarenta, y una sobrina que no llegaba a los veinte, y un mozo de campo y plaza, que así ensillaba el rocín como tomaba la podadera. Frisaba la edad de nuestro hidalgo con los cincuenta años, era de complexión recia, seco de carnes, enjuto de rostro; gran madrugador y amigo de la caza. Quieren decir que tenía el sobrenombre de Quijada o Quesada (que en esto hay alguna diferencia en los autores que deste caso escriben), aunque por conjeturas verosímiles se deja entender que se llama Quijana; pero esto importa poco a nuestro cuento; basta que en la narración dél no se salga un punto de la verdad.

```
{
    "marcadores": [
      {
        "latitude": 40.416875,
        "longitude": -3.703308,
      },
      {
        "latitude": 40.417438,
        "longitude": -3.693363,
        "description": "Paseo del Prado"
      },
      {
        "latitude": 40.407015,
        "longitude": -3.691163,
        "city": "Madrid",
        "description": "Estación de
          Atocha"
      }
    ]
}
```

```
     nombre     color edad altura peso puntuacion
1:     Paco      Rojo   24    182 74.8         83
2:     Juan     Green   30    170 70.1        500
3:   Andres  Amarillo   41    169 60.0         20
4:  Natalia     Green   22    183 75.0        865
5:   Vanesa     Verde   31    178 83.9        221
6:   Miriam      Rojo   35    172 76.2        413
7:     Juan  Amarillo   22    164 68.0        902
```

Another classification may be made based on the format of the data. Below are examples of the main data formats and their description (Table 2). The following section offers an explanation of a selection of the main formats that allow data analysis.

Table 2: Typical data and file formats.

Format	Description
.xlsx/xls Microsoft Excel spreadsheet	Proprietary file format for the storage of structured data in tables. Microsoft Excel allows data display and analysis, although it is of limited use with large volumes of data due to its inefficiency.
.txt Plain text	Plain text files are the universal free format for storing information. Their content may be structured in different formats.
.csv/tsv Comma/Tab separated values	Text file format made up of structured data in tables with comma-separated (csv) or tab-separated (tsv) fields. This is the most basic and efficient format for storing structured data.
.xml Extensible Markup Language	Text file format for semi-structured data storage and data exchange between applications.
.json JavaScript Object Notation	Standard text file format for semi-structured data storage and data exchange between applications, which is lighter and more legible than XML format.

3.2 Main Types of Data and Formats

3.2.1 Basic Big Data Techniques: Basic Classification of Methods

This section will deal with machine learning techniques, a branch of AI for big data processing which essentially aims to identify patterns in the data in order to make inferences. Machine learning algorithms may be classified into three main paradigms:

- Supervised learning (SL): the algorithm learns from several examples how given inputs generate specific outputs (they are labeled) and is thus able to make inferences for new cases. Classic examples include linear regressions and decision trees (DT).
- Unsupervised learning (UL): unlike supervised learning, the algorithm does not have labeled outputs and instead of learning which combination of attributes generates them it searches for patterns in the input data. A classic example is clustering algorithms such as k-Means.
- Reinforcement learning (RL): the algorithm learns from the experience developed in a dynamic environment where it receives rewards. It also does not need to know the labeled output. One example is deep neural networks (DNN).

3.2.2 Examples of Popular Techniques

Some of the most popular machine learning techniques are outlined below. These should be understood as mere examples as there is a host of different techniques that can be used. Firstly, a decision tree (DT) is a hierarchical supervised learning model. It can be seen as a flowchart starting from a root and branching out along different nodes until it reaches a leaf. Each node tests the data, and the branches represent the concrete result of the test. Ultimately rules are generated indicating each of the paths from the root to the leaf. Decision tree models are one of the most common machine learning models because of their recursive ‘divide and conquer’ nature and the fact they are descriptive and easy to understand (Flach 2012).

Other techniques are concerned with deep learning, a subfield of machine learning that bases its high-level learning process on artificial neural networks. Generally speaking, a simple neural network is composed of an input layer, a hidden layer and an output layer. Inspired by the architectural depth of the brain, neural network researchers have for decades sought to develop and train deep multi-layer neural networks so that the model can learn increasingly complex levels of abstraction (Bengio 2009). The goal of each layer is to extract relevant features from the incoming data and after training all the layers one by one, they are all put together and the whole network is refined (Alpaydin 2014). A good example is the generative adversarial networks (GAN) that are known to be commonly used for the generation of 'deepfake' images (Figure 3). Its applications are endless in fields such as advertising and arts and crafts. For example, it can help to create new shoe designs or generate a painting inspired by a great artist from the past using a photo.

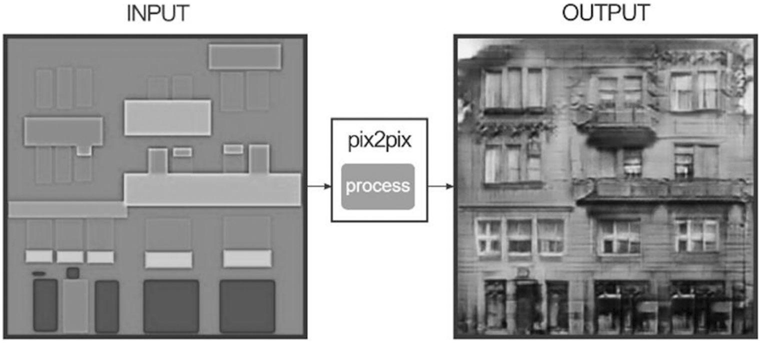


Figure 3: Deepfake of a building.

Thirdly, we should mention clustering techniques. These are machine learning approaches that attempt to find similar patterns and relationships between data points in order to group them. Each cluster is composed of data points which due to their attributes are similar to each other rather than those of another cluster (Sarkar et al. 2018). For example, this technique is commonly used in social network analysis to detect communities of users based on their social relationships and/or interests (Arroyo-Machado, Torres-Salinas & Robinson-Garcia 2021).

3.3 Big Data Tools

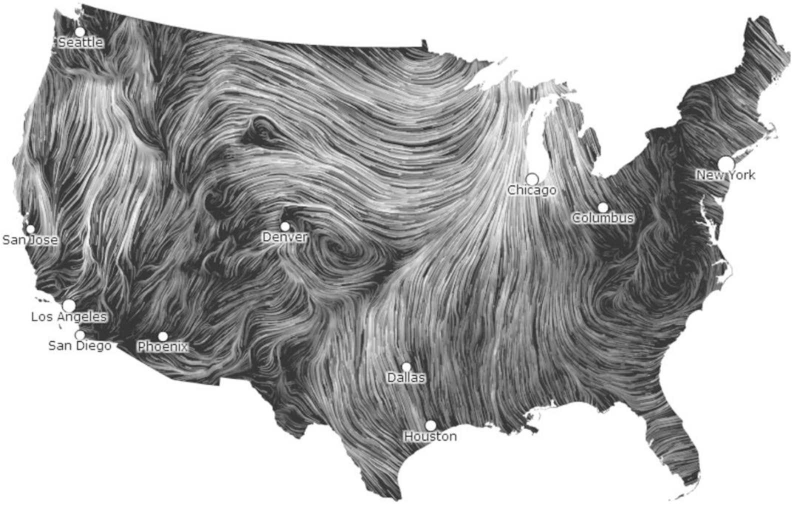
We will cap off this methodological section by describing some useful tools for data analysis and processing at different levels (Table 3). The most powerful and most directly used tools in the analysis of big datasets are data processing frameworks, of which Apache Hadoop and Apache Spark are practically the standard. However, other tools are very popular due to their versatility and power, allowing their application for anything from small data through to large volumes of data, such as the programming languages Python and R. Data mining software also exists that allows the development of models in a visual environment without requiring use of a programming language, such as KNIME and Weka. Finally, also worth mentioning are cloud computing applications, which allow the possibility of contracting a complete customized and scalable work environment that can be accessed via the internet, thus dispensing with the purchase, installation and configuration of equipment and so reducing both costs and time. One of the most popular options is Google Cloud.

Table 3: Main tools for big data analysis.

BIG DATA FRAMEWORKS	
Apache Hadoop	Framework for fast data processing
Apache Spark	Framework for data storage and querying
Apache Hive	Framework for data storage and querying
VERSATILE PROGRAMMING LANGUAGES	
Python	Programming language widely used for data science
R	Programming language widely used for statistical analysis
Scala	Programming language useful for big data processing
INTERACTIVE ENVIRONMENT TOOLS	
KNIME	Tool focused on data mining processes
Weka	Data mining tool that includes a collection of machine learning algorithms
RapidMiner	Tool that includes data mining and machine learning processes
CLOUD COMPUTING	
Google Cloud	Cloud computing services by Google
AWS	Cloud computing services by Amazon
Azure	Cloud computing services by Microsoft

4 Big Data Applied to Humanities

This section outlines some of the recent interactions between data science and humanities and social sciences. Given its multidisciplinary nature, we will see how this epistemological hybridization is taking place in several specialty areas. The art world was one of the first to pay attention to this phenomenon, giving rise to what has come to be known as art data. Within this field, one of the most frequently cited projects is the Wind Map (Viégas & Wattenberg 2012), which has several unique features; first of all, the work is exhibited in the MoMa and was created by a computer scientist and a scientific journalist (Figure 4). The project consists of a living map of the winds that sweep across the United States based on data from the National Digital Forecast Database, represented with trails reminiscent of the brushstrokes of Renaissance painters which endow the meteorological data with beauty.



Fernanda Viégas and Martin Wattenberg. Wind Map. 2012. Interactive software

Figure 4: Image from the Wind Map project that combines art with U.S. meteorological data.

Another area where data science has proven effective is heritage and archaeology, where simple information systems are being replaced by systems that integrate multiple sources (sensors, digital libraries, social networks, etc.) (Amato et al. 2017). Projects in this area are often complex, but we will begin with a small example to illustrate its possibilities. In Bogota (De Urbina 2021), digital photographs of urban scenes from Panoramio were characterized based on the collective perception of the population using semi-structured data (photographer, date, coordinates, event or tags). In a European context, the European project ATHENA (Nisantzi et al. 2018) integrates remote sensing technologies applied to cultural heritage and centralizes the data in a single point.[4] ATHENA collects data using active and passive remote sensing systems which are mainly used in archaeological contexts. Meanwhile, the SCRABS project is a combination of the two previous proposals. Described by the authors as a “Smart Context-awaRe Browsing assistant for cultural EnvironmentS”, it is a paradigmatic example of the collaboration between computer scientists, archaeologists, architects and cultural managers (Amato et al. 2017).

The researchers Zgurovsky and Zaychenko (2020) sought to identify the regularity of systemic global conflicts based on analysis of historical big data. So far,

[4] https://cordis.europa.eu/project/id/691936/es.

an analysis of the complete list of global conflicts occurring since 2500 BC shows that up until the 7th century BC these conflicts did not follow any regular pattern. However, a periodic pattern was revealed in the series of global conflicts following the emergence of higher forms of organization, with the authors relying on analysis of historical data relating to global conflicts that have taken place from 705 BC through to the present day. Using a range of primary sources, they attempted to foresee the next global conflict which they called "the conflict of the 21st century".

Some of Google's projects could also be seen as examples of big data applied to humanities. For example, in 2004 it began the ambitious mass digitization of more than 100 million books through Google Books, generating one of the largest masses of unstructured data. Some of its applications can be found in Google Books Ngram Viewer, an online search engine that charts the frequencies of any set of search strings using a yearly count of n-grams found in printed sources published between 1500 and 2019. Figure 5 shows the frequency of searches for two eighteenth-century poets in the English corpus of Google Books, revealing the interest in their work at different chronological points in time. These techniques fall under what has come to be called Text Corpus Visualizations (Hai-Jew 2015).

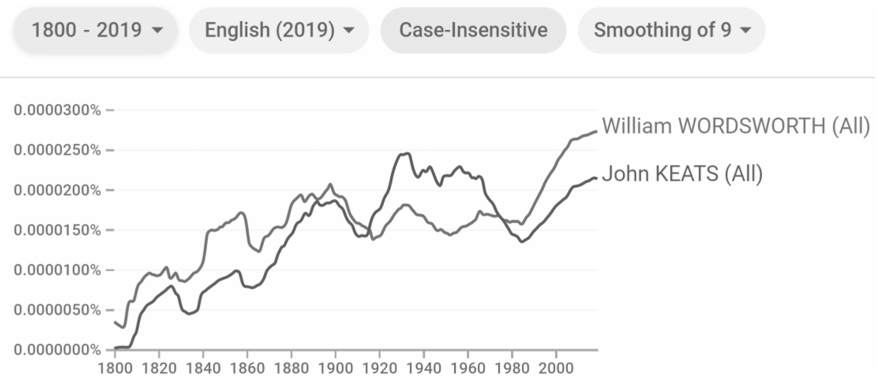

1800 - 2019 ▾ English (2019) ▾ Case-Insensitive Smoothing of 9 ▾

Figure 5: Text Corpus Visualizations using Ngram.

5 The Magister Ludi of Data

To conclude, we will briefly discuss the risks of the misuse of big data. These risks have originated in the current information society due to its dependence on ICT, which has given rise to a context of vulnerability driven by the most accelerated form of capitalism, also known as turbo-capitalism (Luttwak 2000). The dangers of this new environment are evident in the case of entertainment applications and services offered as a free service but which turn the user into the product by accessing, processing and making an economic profit from the data they generate. This is a risk that often goes unnoticed, with the need to access and consume information inevitably overcoming the privacy and rights of the consumer. All the interactions produced on the internet end up feeding algorithms, which use them to filter and catch our attention with whatever the companies that program them want. However, these tools overlook many relevant issues by converting human beings into numbers (Dodson 2008), a risky simplification that could potentially have catastrophic consequences. An example of this is the Black-Scholes equation and other similar models, which some authors point to as being complicit in the culture of excessive risk and unbridled speculation that eventually led to the 2008 financial crisis (Stewart 2012; Harford 2012; O'Donnell 2015).

Other scandals also stand out in this context, such as the company Cambridge Analytica which made improper use of Facebook data in 2016 to influence voters during the Brexit referendum (Hern 2019) and the elections of Donald Trump (Rosenberg, Confessore & Cadwalladr 2018). In both cases, personal data were unlawfully collected by creating political profiles of users in order to send personalized information (García Fernández 2018). The consequences were incalculable and it triggered a legal storm that caused Facebook to lose billions in stock market value, as well as suffering social rejection (Hindman 2018). Apart from the influence these algorithmic models have on our daily activity, there is also the added risk of learning biased or prejudiced behaviors. Social media are precisely one of the main vehicles for tracking and monitoring activity, but it is precisely in these same spaces where we are witnessing an increase in hate speech (Müller & Schwarz 2020) and sexist discourses (Rodriguez-Sanchez, Carrillo-de-Albornoz & Plaza 2020).

Finding a way to avoid falling into bias traps or negative behaviors learned from humans is but one aspect of an even greater challenge: codification of the innumerable differences and nuances of humanity in areas such as culture, politics, religion, sexuality and morality (Webb 2021). This is a major problem because AI as it is currently conceived cannot be attributed intelligence because it is closed to the world in which it has been programmed and cannot see beyond it, being insensitive to and ignoring the dynamics of a world in constant change (Masís 2009). Ultimately, AI learns patterns from the past and provides us with an approximation of the reality of that moment in a specific context.

Therefore, it does not seem entirely clear that the solution to this problem lies in the indiscriminate increase of data. Indeed, the level of knowledge is often confused with the volume of data, when in many cases it is the smaller and better curated collections that allow us to find useful solutions in an efficient way (Olson, Wyner & Berk 2018). Smart data is thus proposed as the transformation of big data into quality data after its cleansing (Triguero et al. 2019). In relation to all this, in the same way that human beings can see their critical capacity being limited in the face of information overload (Marta-Lazo 2018), big data algorithms can also end up leading to other kinds of problems when data are processed without paying any prior attention to them. That is why it is risky to directly point to the data with the highest number of instances and/or properties as being more relevant. In fact, there are already visible signs of this limited view in the academic realm, where the existence of a gap between the so-called ‘data-rich’ and ‘data-poor’ research fields has been identified (Sawyer 2008).

The last novel by Herman Hesse tells the story of Joseph Knecht, the Magister Ludi of Castalia or highest authority of the Glass Bead Game, a kind of high-level humanistic entertainment which is essentially an abstract synthesis of all arts and sciences (Hesse 2012). Players aim to establish relationships between all knowledge based on a given topic. What Hesse seemed to anticipate here is a metaphor for the fate of knowledge, encoded in data and highly connected, although in Castalia the game is controlled not by technocrats but by humanists. As Byung-Chul Han points out, big data is a rudimentary source of knowledge and AI is incapable of thinking (Han 2021), so now the role of humanists as Magister Ludi in this game of data becomes essential and immediate. As in Castalia, someone must oversee data science and the universe of non-things to establish their associations; in short, to forge a more human interpretation of the data we generate in our world.

Bibliography

Aalst, Wil van der (2016): “Data Science in Action”, en Wil van der Aalst (Hrsg.), *Process Mining: Data Science in Action*. Berlin: Springer, pp. 3–23.

Alonso, Rodrigo (2020): “Qué es Big Data o Macrodatos y qué tiene que ver con el hardware”, en *HardZone* https://hardzone.es/reportajes/que-es/big-data-macrodatos/ (letzter Zugriff 01.10.2021).

Alpaydin, Ethem (2014): *Introduction to machine learning (Adaptive computation and machine learning)*. Cambridge: The MIT Press.

Amato, Flora, Vincenzo Moscato, Antonio Picariello, Francesco Colace, Massimo De Santo, Fabio A. Schreiber & Letizia Tanca (2017): “Big Data Meets Digital Cultural Heritage: Design and Implementation of SCRABS, A Smart Context-awaRe Browsing Assistant for Cultural Environments”, in *Journal on Computing and Cultural Heritage*, 10, pp. 1–23.

Arroyo-Machado, Wenceslao, Daniel Torres-Salinas & Nicolas Robinson-Garcia (2021): “Identifying and characterizing social media communities: a socio-semantic network approach to altmetrics”, in *Scientometrics*, 126, pp. 9267–9289.

Bengio, Y. (2009): “Learning Deep Architectures for AI”, in *Foundations and Trends® in Machine Learning*, 2, pp. 1–127.

De Urbina, Amparo (2021): “Una propuesta de valoración de patrimonio urbano a la luz de Panoramio, una fuente Big-Data. El caso del centro de Bogotá”, en *Territorios* (45).

Dodson, Sean (2008): “Was software responsible for the financial crisis?”, in *The Guardian*, Abschn. Technology.

Faraway, Julian J. & Nicole H. Augustin (2018): “When small data beats big data”, in *Statistics & Probability Letters*, 136, pp. 142–145.

Favaretto, Maddalena, Eva De Clercq, Christophe Olivier Schneble & Bernice Simone Elger (2020): “What is your definition of Big Data? Researchers’ understanding of the phenomenon of the decade”, in Florian Fischer (ed.) *PLOS ONE*, 15.

Fayyaz, Zeshan, Mahsa Ebrahimian, Dina Nawara, Ahmed Ibrahim & Rasha Kashef (2020): “Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities”, in *Applied Sciences*, 10.

Fischer, Christian, Zachary A. Pardos, Ryan Shaun Baker, Joseph Jay Williams, Padhraic Smyth, Renzhe Yu, Stefan Slater, Rachel Baker & Mark Warschauer (2020): “Mining Big Data in Education: Affordances and Challenges”, in *Review of Research in Education*, 44, pp. 130–160.

Flach, Peter (2012): *Machine learning: the art and science of algorithms that make sense of data*. Cambridge: Cambridge university press.

Galeano, Pedro & Daniel Peña (2019): “Data science, big data and statistics”, in *TEST*, 28, pp. 289–329.

Gandomi, Amir & Murtaza Haider (2015): “Beyond the hype: Big data concepts, methods, and analytics”, in *International Journal of Information Management*, 35, pp. 137–144.

García Fernández, Aníbal (2018): “Cambridge Analytica, el big data y su influencia en las elecciones”, in *CELAG* https://www.celag.org/cambridge-analytica-el-big-data-y-su-influencia-en-las-elecciones/ (letzter Zugriff 15.12.2021).

Hai-Jew, Shalin (Hrsg.) (2015): *Enhancing Qualitative and Mixed Methods Research with Technology*: (Advances in Knowledge Acquisition, Transfer, and Management). IGI Global. doi:10.4018/978-1-4666-6493-7.

Han, Byung-Chul (2021): *No-cosas: quiebras del mundo de hoy*. Barcelona: Taurus.

Harford, Tim (2012): “Black-Scholes: The maths formula linked to the financial crash”, in *BBC News*, Abschn. Magazine.

Hasan, Md. Morshadul, József Popp & Judit Oláh (2020): “Current landscape and influence of big data on finance”, in *Journal of Big Data* 7, p. 21.

Hern, Alex (2019): “Cambridge Analytica did work for Leave. EU, emails confirm”. in *The Guardian*, Abschn. UK news.

Hesse, Hermann (2012): *El juego de los abalorios*. Madrid: Alianza Editorial.

Hindman, Matthew (2018): “Cómo funcionaba el modelo de Cambridge Analytica, según la persona que lo construyó”, in *El País*.

Jiménez de Luis, Ángel (2021): “Racista por defecto: la discriminación de los algoritmos que Silicon Valley no soluciona | Tecnología”, in *El Mundo*.

Khan, M. Ali-ud-din, Muhammad Fahim Uddin & Navarun Gupta (2014): “Seven V’s of Big Data understanding Big Data to extract value”, in *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*, pp. 1–5.

Kitchin, Rob & Gavin McArdle (2016): “What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets”, in *Big Data & Society*, 3.

Kugler, Logan (2018): “The war over the value of personal data”, in *Communications of the ACM*, 61, pp. 17–19.

Laney, Doug (2001): “3D data management: Controlling data volume, velocity and variety”, in *META group research note*, 70, p. 1.

Li, Gang (2015): “Big data related technologies, challenges and future prospects”, in *Information Technology & Tourism*, 15, pp. 283–285.

Luttwak, Edward (2000): *Turbocapitalismo. Quiénes ganan y quiénes pierden en la globalización*. Barcelona: Crítica.

Mac, Ryan (2021): “Facebook Apologizes After A.I. Puts ‘Primates’ Label on Video of Black Men”, in *The New York Times*.

Marta-Lazo, Carmen (2018): “Calidad informativa en la era de la digitalización: Fundamentos profesionales vs. Infopolución”, in *Calidad informativa en la era de la digitalización*. Madrid: Dykinson, pp. 1–208.

Masís, Jethro. “Fenomenología Hermenéutica e Inteligencia Artificial: Otra Urbanización de La Provincia Heideggeriana.” *Actas de Las Primeras Jornadas Internacionales de Hermenéutica*, 2009.

Mayer-Schönberger, Viktor & Kenneth Cukier (2013): *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

Müller, Karsten & Carlo Schwarz (2020): From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment. New York: SSRN Scholarly Paper Rochester.

Nisantzi, Argyro, Diofantos G. Hadjimitsis, Athos Agapiou, Vasiliki Lysandrou, Andreas Christofe, Marios Tzouvaras, Christiana Papoutsa, et al. (2018): “Remote sensing archaeology knowledge transfer: examples from the ATHENA twinning project”, in Nektarios Chrysoulakis, Thilo Erbertseder & Ying Zhang (Hrsg.), *Remote Sensing Technologies and Applications in Urban Environments III*. Berlin: SPIE.

O’Donnell, Benedict (2015): “Cracking the maths behind the economy”, in *Horizon: the EU Research & Innovation magazine* https://ec.europa.eu/research-and-innovation/en/horizon-magazine/cracking-maths-behind-economy (letzter Zugriff 09.12.2021).

Olson, Matthew, Abraham Wyner & Richard Berk (2018): “Modern Neural Networks Generalize on Small Data Sets”, in S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Hrsg.), *Advances in Neural Information Processing Systems*, *31*.

Ortíz, Georgina (2019): “Toshiba ayuda en los experimentos del Gran Colisionador de Hadrones”, in *Big Data Magazine* https://bigdatamagazine.es/toshiba-ayuda-en-los-experimentos-del-gran-colisionador-de-hadrones-lhc (letzter Zugriff 01.07.2021).

Pascual, Manuel G. & Fernando Peinado (2018):” Cambridge Analytica ofreció sus servicios en España”, in *El País*.

Rands, Kevin. “How Big Data Has Changed Politics.” *CIO*, 2018, https://www.cio.com/article/221882/how-big-data-has-changed-politics.html.

Rodriguez-Sanchez, Francisco, Jorge Carrillo-de-Albornoz & Laura Plaza (2020): “Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data”, in *IEEE Access* 8. 219563–219576. doi:10.1109/ACCESS.2020.3042604.

Rosenberg, Matthew, Nicholas Confessore & Carole Cadwalladr (2018): “How Trump Consultants Exploited the Facebook Data of Millions”, in *The New York Times*.

Sarkar, Dipanjan, et al. *Practical Machine Learning with Python*. Apress, 2018, https://doi.org/10.1007/978-1-4842-3207-1.

Sawyer, Steve (2008): “Data Wealth, Data Poverty, Science and Cyberinfrastructure”, in *Prometheus*. Routledge, 26, pp. 355–371.

Scaife, A. M. M. (2020): “Big telescope, big data: towards exascale with the Square Kilometre Array”, in *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 378, p. 2166.

Schnase, John L., William P. Webster, Lynn A. Parnell & Daniel Q. Duffy (2011): “The NASA Center for Climate Simulation Data Management System”, in *2011 IEEE 27th Symposium on Mass Storage Systems and Technologies (MSST)*. Denver: CO, pp. 1–6.

Siegler, MG (2010):” Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003”, in *TechCrunch* https://techcrunch.com/2010/08/04/schmidt-data/ (letzter Zugriff 25.03.2023).

Stewart, Ian (2012): “The mathematical equation that caused the banks to crash”, in *The Observer*.

Sun, Limin, Zhiqiang Shang, Ye Xia, Sutanu Bhowmick & Satish Nagarajaiah (2020): “Review of Bridge Structural Health Monitoring Aided by Big Data and Artificial Intelligence: From Condition Assessment to Damage Detection”, in *Journal of Structural Engineering*, 146, 04020073. doi:10.1061/(ASCE)ST.1943-541X.0002535.

Torgler, Benno (2020): “Big Data, Artificial Intelligence, and Quantum Computing in Sports”, in Sascha L. Schmidt (Hrsg.), *21st Century Sports (Future of Business and Finance)*. Berlin: Springer, pp. 153–173.

Triguero, Isaac, Diego García-Gil, Jesús Maillo, Julián Luengo, Salvador García & Francisco Herrera (2019): “Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data”, in *WIREs Data Mining and Knowledge Discovery*, 9, e1289. doi:10.1002/widm.1289.

Viégas, Fernanda Bertini, and Martin Wattenberg. *Wind Map*. Interactive software, 2012, https://www.moma.org/collection/works/163892.

Ward, Jonathan Stuart & Adam Barker (2013): *Undefined By Data: A Survey of Big Data Definitions. arXiv*. doi:10.48550/arXiv.1309.5821.

Webb, Amy (2021): “Los nueve gigantes”, in *Cómo las grandes tecnológicas amenazan el futuro de la humanidad*. Barcelona: Península.

Zgurovsky, Michael Z. & Yuriy P. Zaychenko (2020): *Big Data: Conceptual Analysis and Applications* (Studies in Big Data). Berlin: Springer.