Big Data and Educational Research

Roser Beneito-Montagut
School of Social Sciences, Cardiff University


Abstract

Big data and data analytics offer the promise to enhance teaching and learning, improve educational research and progress education governance. This chapter aims to contribute to the conceptual and methodological understanding of big data and analytics within educational research. It describes the opportunities and challenges that big data and analytics bring to education as well as critically explore the perils of applying a data driven approach to education. Despite the claimed value of the increasing amounts of large and complex data sets and the growing interest in making sense of them there is still limited knowledge on big data and educational research.

Over the last decades, the developments on information and communication technologies are reshaping teaching and learning and the governance of education. A broad variety of online behaviours and transactional data is (or can be) now stored and tracked. Its analysis could provide meaningful insights to enhance teaching and learning processes, to make better management decisions and to evaluate progresses –of individuals and education systems.

This chapter starts by defining big data and the sources and artefacts collect, generate and display data. In doing so it explores aspects related to data ownership and researchers’ access to big data. It then assesses the value of big data for educational research by critically considering the stages involved in the use of big data, providing examples of recent educational research using big data.
1. Introduction

Big data and data analytics bring the promise to enhance teaching and learning, improve educational research and progress education governance. Despite the alleged value of the increasing volume of education related large and complex data sets and a growing interest in making sense of them there is still limited research on big data and educational research and, thus, limited understanding of its value for education. Over the last decades, developments in the area of information and communication technologies (ICTs) -such as the social and semantic web or mobile technologies- are reshaping not only teaching and learning practices but also the governance of education. A huge variety of online behaviours and transactional data is (or can be) now stored and tracked. This adds to the open data movement that is providing considerable amounts of data about education and to the digitalisation of education and research resources -such as library repositories or data repositories.

The analysis of these developments could provide meaningful insights to enhance teaching and learning, make better management and policy decisions and evaluate progresses. Consequently, big data entails a knowledge system that potentially can remodel education and research. But many questions remain to be answered: are the big data promises too optimistic? What is the “real” value of big data? Where and how can educational researchers access big data? What do we know about the value of big data for education?
While not able to fully answer these questions, this chapter aims to contribute to the conceptual and methodological understanding of big data processes and analytics within educational research. It describes the opportunities and challenges that big data and analytics bring to education and critically explores the perils of applying a data driven approach to education decision-making processes. Big data is used synonymously as a substantive topic for education research and as a kind of data and research method. This chapter addresses both.

The chapter is organised as follows: it starts by defining big data, and “not-so-big data”, in the context of education. Next, it describes the main sources of big data for educational research, reviews some digital education related artefacts, and explores issues related to data ownership and researchers’ access to big data. It then assesses the value of big data for educational research by critically considering the stages involved in the use of big data: (1) data curation processes; (2) learning, research, academic and labour market analytics and (3) display and visualisation of educational big data. The three steps are intrinsically interwoven but by reviewing them separate the chapter discusses the methodological, ethical and sociological tensions emerging between better, quicker and smarter decisions on the one hand and increasing accountability and standardisation on the other. In doing so, the chapter also showcases recent educational research using big data.

Big data will probably become more common practice in educational research in the future. This demands greater awareness about its contexts and its consequences, within a domain specific perspective. This chapter aims to add conceptual clarity to the limited knowledge of the role of big data in education and to highlight its potential value for education research. It aims to describe the opportunities that big data brings to education research and the challenges and critical issues related with its adoption.

2. Concepts and terminology

The definition of big data is highly contested. Big data is a trend across many computational and social areas (including education, business, health and
government among others). Big data is not only data, a technique or an innovation. It also entails a process and a way of thinking about what can be known and how (boyd and Crawford, 2012; Housley et al, 2014; Mayer-Schoberger and Cukier, 2013; Kitchin, 2013) and a mode of informing decision-making processes.

Big data has been initially defined through their three dimensions, known as the three Vs of big data—volume, velocity and variety (Douglas 2001). Volume refers to the huge quantity of data in “big data”—from terabytes to petabytes or exabytes of data. Velocity relates to the fact that data streams are not only generated but available for analysis and display in real-time or nearly real-time. And variety refers to the multimedia character of big data, which is not only numbers, dates, text or strings, but also video, sound, 3D data and images, and is often unstructured (for example this is often the case with social media data). Computationally, big data is too big to be retrieved, collected, stored, managed, analysed and displayed using conventional software and hardware (Manyika et al 2011). In other words, in computer sciences big data describes data sets that are too large and complex to be managed by traditional means. However, from a social sciences perspective big data is not always that big. Social scientists now talk about “big data” when they get more data, quicker and richer than before, adding a contextual and relative dimension to its definition (Schroeder 2014a).

Kitchin (2014), based on an exhaustive review of literature, adds to the three Vs of big data four other elements that are significant to the social sciences in general and to educational research in particular. These elements raise new epistemological questions. Big data is “exhaustive in scope” (Kitchin 2014); data may be collected from entire populations, challenging the need for social theory as the data can reveal patterns and relationships that researchers were not even looking for (Anderson 2008, Prensky 2009). It is “fine-grained in resolution”, which means that it is as detailed as it could possibly be, and “uniquely indexical in identification”, enhancing the descriptive capacity of the data. It is “relational” and so “flexible” that can be expanded easily (boyd & Krawford 2012; Kitchin 2013; Mayer-Schonberger and Cukier 2013).
A research agenda is being shaped that aims to unlock the value of big data in education.

3. Educational big data

Over the last decades the generalisation of digital processes associated with the use of information and communication technology (ICT) – such as mobile devices, teaching and learning technologies, social media, and so on – is reshaping the form of learning and teaching from the student, teacher and institutional perspectives.

In what follows, the chapter illustrates the ways in which big data relevant to education research is being generated. In doing so, the chapter considers the arenas in which data is produced and the artefacts through which data is generated. Its purpose is not to provide an exhaustive mapping of those but give an idea of the multiplicity and variety of “big data” sources for education research. It then provides an overview of different ‘stages’ in the use of big data in educational research.

3.1. Arenas and artefacts for the production of big data in education

Big data in education is mainly produced through three broad ways that are described as follow. They have been organise into three categories. These categories are not mutually exclusive as digital artefacts generate data that may belong to more than one category.

**Teaching and learning activities**

Big data that are produced as a consequence of teaching and/or learning activities can be analysed through learning analytics – for a definition of learning analytics see section 3.2. Over the last two decades, a digital revolution associated with developments in information and communication technologies – such as the creation of ubiquitous and mobile devices –, flexible and technology-enhanced classroom designs and Massive Open Online Courses (MOOCs) is reshaping and broadening the modes of and accessibility to learning and teaching. In addition, many institutions are embracing new class formats and technologies designed to meet either
evolving student needs or as mechanisms to reduce operational costs. “Flipped classroom” is a pedagogical model that reverses traditional teaching models delivering instruction online outside the classroom (or non-contact time) and moving homework or exercises into the class (or face-to-face time). MOOCs are usually provided by higher education institutions that aim to target large numbers of students globally. Well-known examples of platforms that offer MOOCs are Coursera\(^1\), Udacity\(^2\) or FutureLearn\(^3\). MOOCs include videos (which allows tracing and monitoring students’ behaviours –observing if students watch the same lesson several times, if there is a specific part that is watched repeatedly) as well as quizzes, readings, discussion fora, and so on. Institutions, companies and organisations delivering pre-university or higher education are increasingly developing and delivering learning resources online. IXL\(^4\), Knewton Platform\(^5\) or Dreambox Learning\(^6\), among others, offer teaching and learning resources online that progressively more schools and parents are adopting as a way to enhance or complement formal education. Another relevant source of data derives from Open Universities and from higher education institutions increasing their online presence via learning management systems (LMS) owned by companies such as Moodle\(^7\) or Blackboard\(^8\). There is an additional source of students’ data as higher education and other educational institutions increase their online repositories, educational digital libraries and associated tools.

Researchers and instructors can now gain considerable more detailed insights into student’ learning activities. Interactions with educational digital artefacts can be traced and this data can be collected and analysed to imprint changes in research and practice (Borgman et al 2008; Brown 2012, Xu and Recker 2012).

User-generated “education” data

\(^1\) https://www.coursera.org/
\(^2\) https://www.udacity.com/
\(^3\) https://www.futurelearn.com/
\(^4\) https://uk.ixl.com
\(^5\) https://www.knewton.com
\(^6\) http://www.dreambox.com/
\(^7\) https://moodle.org/
\(^8\) http://uki.blackboard.com/sites/international/globalmaster/
Another source of big data in education is user-generated data -or volunteered data generation- including transactions\(^9\), social media, crowdsourcing and citizen science\(^{10}\). User-generated content in education is rapidly increasing its volume as individuals (students, teachers, parents and other social actors in the education area); public sector institutions (schools, governments, and other public sector organisations) and private companies are augmenting their online presence via social media. *Edmodo*\(^{11}\), for example, is a social network site (SNS), for teachers, students and parents which combines Web 2.0 functionalities with educational content. A particular type of social media platforms that target scholars and researchers has emerged that aims to promote social networking and resource sharing among them. The best-established examples are *ResearchGate*\(^{12}\) and *Academia.edu*\(^{13}\) both of which offer analytics on the impact of research and information on how much use is made of specific research outputs available from the website.

A further example of user-generated data comes from the employment websites dedicated to connect people to jobs (i.e. *Jobseekers*, *Monster.com*, etc.) or vice versa. These webpages contain data regarding job offers and demands that is mined and analysed under the labour market analytics label. *Island recruiting*\(^{14}\) and *Labour Analytics*\(^{15}\) in Canada deliver job market, analytics aiming at matching people with jobs. Wearable technologies and devices are another potential source of data in and for education research that are on the hype (Lima 2015; Nield 2015). Possibilities range from quick question and reply systems between students and teachers, through smart phones or smart watches, to facial recognition and virtual reality to assist

\(^9\) Transactional data refers to any form of digital footprint left “voluntarily” as a result of quotidian interactions with digital devices or artefacts, for instance university cards that are used to access university facilities, to borrow books, and even to control lectures attendance, in and outs of dormitories or exams.

\(^{10}\) Citizen science refers to the public engagement in scientific research processes in particular and democratic and policy processes more broadly.

\(^{11}\) https://www.edmodo.com/

\(^{12}\) https://www.researchgate.net

\(^{13}\) https://www.academia.edu/

\(^{14}\) http://islandrecruiting.com/

\(^{15}\) http://www.labouranalytics.com/
teachers to recognise students and get their data on the go (Google Glass and Oculus).

Finally, initiatives such as Wikipedia—and many other similar digital artefacts now introduced in the area of education—have utilized crowdsourcing capabilities. Following Wikipedia success, platforms such as Duolingo\(^\text{16}\) or Viki\(^\text{17}\) have been created, that build communities around language learning and translation.

**Academic/management processes**

Much data are now collected manually or automatically about education-related processes besides teaching and learning—for example in the areas of enrolment, curricula, outputs, performances, and so on. For instance, governmental and international organisations have created data-dashboards to publish and open their data to educational researchers and other stakeholders. The OECD has developed an *Education GPS*\(^\text{18}\) (Williamson 2015a), a data-dashboard combining comparable international data and analysis on education policies and practices, opportunities and outcomes. The European Union (EU) through its Eurostat website and related services such as the GESIS makes large volumes of data freely available.

Private companies generate data that informs the management of higher education institutions and academic processes. *i-Graduate*\(^\text{19}\) undertakes online surveys of students and other relevant actors such as education institutions, governments and private companies in Asia, Europe and North America. It claims capacity to provide “the global benchmark for the student experience”. *Schoolzilla*\(^\text{20}\) provides a data platform for schools.

These are only a limited selection of the digital artefacts available to generate and collect educational data. It is clear even from this limited sample that the mains challenges for educational research do no longer relate primarily to data generation, but also to data access, management, retrieval, curation,

\(^{16}\) [https://www.duolingo.com/](https://www.duolingo.com/)

\(^{17}\) [https://www.duolingo.com/](https://www.duolingo.com/)


\(^{19}\) [http://www.i-graduate.org/](http://www.i-graduate.org/)

\(^{20}\) [https://schoolzilla.com](https://schoolzilla.com)
analysis and visualisation or display (Souto-Otero & Beneito-Montagut 2016, Williamson 2015a).

The next section reviews different stages to big data collection, analysis and display. Most of the literature has mainly explored the “analysis” stage. Only recently educational research has started to look at a wider set of aspects of the “big data” phenomenon, looking in particular at data governance and its social and policy implications (Eynon 2013, Selwyn 2015, Williamson 2015b).

3.2. Stages and challenges in the use of big data in education research

This section reviews big data collection, analysis and display in educational research, and the main challenges associated with it. The section includes research cases to illustrate use of different modes of big data analysis.

Data collection and curation: from open data to data brokers

Big data are usually computationally collected with the assistance of algorithms and then, it is automatically organised in a database. The term data curation should be introduced in order to highlight that the data collection process is not neutral. Principled and controlled forms of data curation add a procedural dimension to the processes that form part of data collection and preparation. It denotes all the processes of extraction, mining, management and validating data. This requires, first, the creation and identification of data that can potentially reveal useful and valuable information. Second, the identification of how these data are organised or—in the case of unstructured data—how the data need to be organised. Third, it requires researchers and users to understand that data collection does not occur in a vacuum. It happens in a specific context that is affected by how we shape technology (Mackenzie 2006). This three points bring ontological aspects to the data collection and curation stage that educational researcher should reflect upon.

The collection and curation of big data carries certain problematic issues. As already mentioned, information about the methods by which data were produced is needed. The algorithm employed may need to be unpacked and curated by researchers and other experts—otherwise the data may be wrongly analysed and interpreted in a vacuum. This information is essential for the
subsequent analysis and interpretation of the data. This is a key aspect to which educational researchers can contribute in the use of big data for educational research: curation should warrant that the representations of objects/subjects of study functions effectively as data from which meaning can be extracted.

Interoperability and access generate two further challenges to the expansion of the use of big data in education research. Much data exists in disconnected silos, and some of these may not necessarily be accessible to education researchers. An increasing number of private companies are starting to collect and curate education related data for its aggregation into “analytics tools” that can be sold back to education stakeholders. The data brokers are emerging as key education data curators.

On the other side, in spite of governments’ and international organisations’ push towards open data in education (i.e. LinkedUp21 project or the education data-dashboards mentioned earlier in this chapter) data owned by private companies (such as Coursera, Facebook and Edmodo to mention a few) remains generally inaccessible to education researchers.

Research undertaken in recent years has paid close attention to analytics and to the challenges and limitations related to big data, as well as to its policy implications (i.e. Williamson 2015a, Souto-Otero & Beneito-Montagut 2013, 2016). The well-known “Facebook experiment” (Kramer, Guillory and Hancock 2014) raised ethical questions –that are relevant in education as well– about the strategies to collect big data for education research (Reich and Stevens 2014) and about the mislabelling of students according to imperfect algorithms in learning analytics. Facebook experiment intended to study whether the emotional state of its users could be altered. It also illustrated how relevant is the data curation process. While education researchers need to address contested issues related to data curation, there seems to have been limited reflection so far on this aspect aside from the field of digital humanities.

Analysis: Analytics, analytics and more analytics.

---

21 http://linkedup-project.eu/
Given the volume, variety, velocity and complexity that big data relevant to education present, traditional methods of analysis — which have been designed to extract insights from limited and static data — are not generally considered to be fit for purpose. Data analytics generally refers to the process -assisted by computers and software- for capturing, analysing and reporting digital data to support decision-making, but these have been kept separate for presentational purposes in the organisation of this sub-section.

With the recent increases in computational power, software availability and data accessibility analytics can be performed automatically in real-time or nearly real-time. Private and public organisations and institutions have increasingly augmented the use of big data to inform interventions and quickly monitor the impact of the interventions. There are several “analytics” labels that have emerged in the education arena: learning analytics, academic analytics, research analytics and labour market analytics.

**Learning analytics** applies the data analytics to the teaching and learning area (see i.e. Long & Siemmens 2011, Johnson, Adams and Cummins 2012; Sharples et al 2012, -see also the work of organisations such as Educause [22], JISC [23] or the Society for Learning Analytics [24]). Learning analytics, then, is the automatic collection, measurement, analysis, displaying and reporting of educational big data with two main aims: understand and optimise learning and the education system (see Romero and Ventura 2010; Ferguson 2012). Some scholars differentiate between learning analytics and academic analytics. This section deals, first, with learning analytics, which centers on the learning process, and then with academic analytics. However, it is necessary to recognise that this distinction poses a division between data produced while learning and data produced by the education institutions, which does not fully account for the relationship between learner, content, institutions and education systems.

In the education field a broad range of data can be generated about students and learning, and the learning analytics literature claims its power based on

---

22 [http://www.educause.edu/library/analytics](http://www.educause.edu/library/analytics)
23 [http://jisc.cetis.ac.uk/topic/analytics](http://jisc.cetis.ac.uk/topic/analytics)
its capacity to analyse “readily evident data and feedback” (Siemens and Long 2011). Learning analytics promises new avenues to design personalized learning experiences and quickly respond to learners’ needs in order to improve students’ success (i.e. Olmos and Corrin 2012; Smith, Lange and Huston 2012). These promises are potentially transformative as the rich data now available to researchers and policy makers could provide the basis for the design of new models of education, the improvement of teaching and learning, organisational management, and decision making processes and, thereby, serve as a foundation for systemic change.

Learning analytics practice

Hung, Hsu, and Rice, researchers in an Educational Technology Department, applied predictive learning analytics to evaluate an online educational programme, supplementing K-12 education, through analyses of student learning logs, demographic data, and end-of-course evaluation surveys (Hung, Hsu and Rice 2012). They applied cluster analysis—an exploratory data analysis technique— and decision tree analyses—predictive analytics—to predict student performance and satisfaction levels towards course and instructors. A total of 23,854,527 activity logs were collected from 7,539 students in 883 courses, using the Blackboard activity accumulator tool. The demographic data collected included gender, age, graduation year, city, school district, number of online courses taken, number of online courses passed and/or failed and final grade average. Finally, data were collected at the end of each course via an online questionnaire that included 24 questions about their satisfaction with the course and teacher. They also collected data to measure the level of engagement (such as average of discussion board entries per course, average frequency of logins, etc.). The results of their research suggested a set of effective indicators to identify students who are more likely to be successful in completing the course. In this application of learning analytics a framework and a model to evaluate an educational program is advanced that can be used to develop an early warning system to detect at-risk students. This would provide educators with a decision-making support tool to analyse the courses offered in addition to the early warning system.

The project RETAIN (http://retain.open.ac.uk/) undertaken at the UK Open University is another example of the use of predictive learning analytics. It found that the student level of activity was not predictive of course completion, but a decrease of their activity in the LMS was an indicator of trouble (Wolff and Zdrahal 2012). The students could be successful without much online activity, but if a usually active student stopped being so, they were unlikely to complete.

Learning analytics use several methods and techniques including web analytics (analysis of web logs), social network analysis (SNA), predictive analytics, machine learning, knowledge discovery and education data mining (EDM) (Clow 2013). Research in this area has so far been mainly conducted in STEM research areas (computing, mathematics and engineering). The main advancements have been made in the modelling algorithms to assist in
learning analytics and Business Intelligence, whereas research on their applications in educational contexts is still quite limited. Research providing evidence on the enhancement of education processes through the use of big data is still scarce. Most of the published learning analytics’ research is exploratory or small-scale experimental studies and is oriented towards the development of new tools. The learning-focused perspectives adopted tend to revolve around social-learning analytics, situating learning analytics within the constructivist paradigm (i.e. Dawson 2009, Haythornthwaite et al 2013). Educational research from a social science perspective—pedagogy and cognition—has barely engaged with the systematic study of the impact on teaching and learning of learning analytics.

**Academic analytics** connect the outcomes of the data analysis with policy and economic factors rather than teaching and learning. It refers to the analytics used to data driven-decision practices for operational and managerial purposes at higher education level. Some scholars also consider academic analytics the teaching and learning data analysis when it refers to higher education.

---

**Academic analytics**

As documented by Campbell, DeBlois and Oblinger (2007) some universities have been using enrolment predictive modelling techniques. An example is Baylor University that collect and analyses large amounts of data on prospective students and has developed a sophisticated strategy to admissions. They, first, have identified the best predictive variables for Texas residents (including number of self-initiated contacts, a mail qualifying score, etc.). University managers are then able to identify from the university data-base those students most likely to be admitted and then trigger different kinds of institutional responses or follow-ups. The effects of the creation of this analytic strategy have been “the creation of information that lets Baylor segment its prospect pool, target the most likely enrolees, and more efficiently use human and financial resources to deliver the desired freshman class.” (Campbell et al. 2007) After the implementation of this business intelligence strategy there was a significant increase in student applications from 2005 to 2006.

A different application of academics analytics can be observed in Shields's article (Shields 2015) investigating the network of social media communication—Twitter—between globally ranked universities. It examines the relationship between ranking status and network status for the selected universities.

---

Entering into discussions on what we mean by “better education” there are emerging tensions between the framing of education as an economic activity
and the idea of an education system for citizenship, social cohesion or social justice. With learning and academic analytics and big data this question can be researched as an empirical endeavour, and the impact and concrete consequences for students and teachers—and for the education systems—of education interventions around resource management, class sizes and workloads, can be empirically evaluated. But it comes with the dangers as well of enhancing the accountability power, formulating ethical dilemmas and posing difficult challenges (Slade and Prinsloo 2013).

Educational research needs to enter into both: the pedagogical implications of learning analytics through the use of big data and the study of the social and ethical implications of using big data (for instance, could the data collected and analysed be used to inform access to higher education for social justice?). It is still uncertain whether algorithms will succeed in adjusting to students’ needs or if students will adapt their behaviour to succeed (and game) the algorithm (Souto-Otero & Beneito-Montagut 2016), students will develop an ‘algorithmic skin’ (Williamson 2015c), or an ‘algorithmic self’ (Pasquale 2015). For instance, learning analytics programmes that grade essays rely on measures—such word sophistication or length of the sentences—, which tend to correlate with high grades. But once these criteria are known or deducted from experience they can be gamed: students can start writing for the algorithm (with sophisticated words and long sentences), instead of focusing on writing a coherent argument (Marcus and Davis 2014). Research also needs to engage in longitudinal studies that look into the long term effects of big data and analytics in education and across different education sectors.

Another analytic strategy with potential to improve educational practice and research is labour market analytics. Labour market analytics uses ‘big data’ mined from millions of job adverts posted online and CVs, to provide real-time labour market data and “advice employers on strategies to optimise the salaries and benefits they offer through competitor benchmarking, guide educational institutions on curriculum development or help users of public employment services to identify kinds of jobs for which they are qualified but have never thought about” (Souto-Otero & Beneito-Montagut, 2016:22). It is
emerging as an area of research and some centers, such as the Center of Job Knowledge Research at the University of Amsterdam (Kobayaski et al. 2014) or the School of Social Sciences at Cardiff University are working on projects on labour market analytics and their social impact.

**Research analytics**

Zhang et al. (2015) present a research analytics framework for matching doctoral students and supervisors. The analytics use multiple measurements on three dimensions (relevance, connectivity and quality) to do the matching. This application collects data from web 2.0 technologies and uses the same sort of technology as online dating websites. The framework represents two stages: filtering and ranking.

This analytical framework, known as people-to-people recommendation techniques and models, has also been used in labour market analytics (Malinowski et al. 2006, Yu et al 2011).

Finally, research analytics aims to analyse the way research happens and succeeds or not in getting attention (van Hamerlen 2012). Research analytics are also used for analysing research collaboration, to –for example- inform future processes of identification of patterns. More specifically, co-authorship patterns have been explored using SNA and citation patterns through the use of bibliometrics. Evidence-based and impact agendas are increasingly permeating research –in the education area and in other research areas- and enhance the motivation to adopt research analytics strategies to measure impact and reach of the research outputs. For instance, there is an ongoing debate around the possible use of research analytics in the Research and Excellence Framework (REF) in the UK (Jump 2014, Mryglod et al 2014).

From a critical perspective we should not forget that there is a risk that the systemic change implied in the use of analytics shifts the power from the institutions to the algorithm (Lash 2007). The suggestion here is that software algorithms are increasingly making decisions for organizations and institutions (Beer 2009). These are decisions taken on the bases of ‘raw sense data’ and the ‘empiricism of the thing, of the event’ (Lash, 2007: 64; Lash and Lury, 2007). The consequences for learning, academic, labour market and research analytics are far from being understood.

**Data display and data visualisation: making the data analyses**
Data display is the last stage in making sense of big education related data. Given the huge volumes, velocity and complexity of big data, visualisation is a way to make sense of data and to communicate that “sense” in an accessible manner. Visualisation strategies aim to reveal the structure, patterns and trends of the data and the relationships between variables. Thousands of data points can be plotted to reveal a structure, data flows (for instance, mapping geographically trends or hashtags across millions of tweets) or the real-time dynamics of a given phenomenon can be monitored using graphic and spatial interfaces (such as the COSMOS\(^{25}\) platform). Consequently, it is again necessary to understand that visualisation strategies, and the artefacts to perform them, are not neutral. Rather the opposite: they have a profound impact on the interpretation of the data.

The potential of “big data” visualisation artefacts in educational research comes with similar contested issues (regarding the neutrality of visualisation processes, the knowledge on how visualisation algorithm work, and so on) to those faced by data curation, as education research has barely started to address the potential and challenges of visualisation. Using visualisation strategies the data is provided to different social actors – researcher, stakeholders, organisations, institutions- in a form that is easily understandable and interpretable (Gitelman and Jackson 2013). And the effects of visual displays could amplify the use of the data to create descriptions and powerful interpretations of education matters. The visual displays have the ability to easily give meaning to the data and sustain discourses

Data displays are used to channel decision-making. Private companies have a central role in the displaying of the data, as they do regarding the curation and analysis of big data. This is because data visualisation is intrinsically linked to data curation and analysis. For instance, digital artefacts such as School Finder, curate big volumes of data to support parents to search and find schools in predetermined geographical areas (Williamson 2015a). Rightmove\(^{26}\) –a UK-based FTSE listed real estate search tool– provides a tool

\(^{25}\) COSMOS: https://www.cs.cf.ac.uk/cosmos/
\(^{26}\) http://www.rightmove.co.uk
in its search engine called *School Checker* that visualises the probabilities of a property to be within the catchment area of nearby schools based on the information collected by *SchoolGuide*.27

**FIGURE 1**

On the other hand, there have been numerous attempts to facilitate data access for students and learners through visual interfaces such as data dashboards that display multiple visualisations (Verbert et al 2013). MOOC companies and Open Universities also provide data visualisation tools for practitioners and researchers. For example, *FutureLearn* MOOCs’ platform offers several metrics (learners, social learners, etc.) to instructors and universities offering the MOOCs.

**FIGURE 2**

---

27 [http://www.schoolguide.co.uk](http://www.schoolguide.co.uk)
Displaying big data

An Economic and Social Research Council (ESRC) and Joint Information Systems Committee (JISC) investment brought together social, computer, political, health and mathematical scientists to develop a 'social computational tool kit' that captures, analyses and displays user-generated Big Data to answer social questions. This is an example of a research initiative that aims to facilitate the access of big data to social scientists. The COSMOS platform not only allows the researcher to take control over the data curation process but also provides data accessible visualisations (Housley et al 2014, Burnap et al 2014).

Figure 3: COSMOS platform

Making displays of information accessible to and adaptable by educational researchers is an important challenge to the use of big data in educational research. These displays endeavours embed the big data promise of transcending context or domain specific knowledge (Kitchin 2014). But educational researchers does not have the skills to adapt the tools and then, they only can use generic tools that cannot be tailored.

Research on big data and education has touched upon display techniques underestimating the effect on the possible interpretations that could be extracted from them. Furthermore, conventional methods for visualizing data are not appropriate for big data. It is necessary to introduce new strategies and digital artefacts to display big data but this process comes with its own challenges. First, it requires us to think about these artefacts as non-neutral. Second, educational researchers do not have the skills to tailor the tools and techniques to their needs and sometimes they need to rely in private
companies. Finally, given the limited research on the social shaping of digital artefacts displaying big education related data, researchers should engage in a critical analysis of the use of display strategies to research and decision-making processes.

4. New (critical) horizons

Once the sources of big data and the steps to employ it have been discussed the key question for educational research is to determine in which areas big data could be valued and how it can augment (Edwards et al 2013) current research. In the field of educational research the use and application of big data is still scarce and mostly concentrated in educational technology research, but it is starting to develop and it will continue doing so over the coming years. Big data has arrived, and whether we like it or not, the education area is being affected by it. In this new context educational research needs to embrace “digital sociology” (i.e. Beer and Burrows 2013, Halford, et al 2013, Housley et al 2014, Rupert 2015).

This chapter adds to the literature advocating the problematisation of the use of “big data” in education (Eynon 2013, Selwyn 2015); this needs to go hand in hand with greater efforts on evidencing the real impact of big data in education. Educational research should move beyond getting “big” evidence of better students’ achievements, better management strategies and better decisions. It needs to, at the same time, interrogate the capacity of big data in getting better societies, as the role of education in achieving better societies is of no question (Esping-Andersen 2002). Another “digital sociology” issue that educational research needs to address is the governing through data turn (Souto-Otero & Beneito-Montagut 2016, Kitchin and Dodge 2011, Williamson 2015a). The argument is that social actors’ relationships are being reshaped and changed by big data and digitalization, an aspect that has been neglected in extant literature on the government of education through data.

Second, big data is questioning established research methodologies, epistemologies and ontologies. Educational research could benefit from revisiting of epistemological and ontological issues in light of the advent of big data. Although these issues have been extensively discussed (i.e. boyd and
Crawford 2012, Kitchin 2013, Mayer-Schoberger and Cukier 2014) in various responses to overly optimistic big data claims, this is an aspect that the education research literature has largely ignored. This also implies the fostering of new research agendas that aim to unpack and understand what code is and what code does (Mackenzie 2006, Williamson 2015d) in curation, analysis and display processes.

Although big data has been characterised by a quantitative turn, we suggest (with other scholars such as boyd and Crawford 2012) a pragmatic and question-driven approach that also considers qualitative analyses of big data in other to avoid reductionist approaches. Big data is still unable to capture, for instance, complex emotions, thoughts, values and so on. It is not always able to automatically adapt to human behaviour changes or “reactive” strategies (Souto-Otero and Beneito-Montagut 2016). The challenge for educational research, then, is to push back naïve and reductionist claims about the value of big data by means of domain specific empirical research on big education data, and to embrace research on “the algorithmic self”.

The third aspect refers to ethics (Schroeder 2014). Issues around data protection, privacy, informed consent, and what information should or not be shared have been raised regarding big education related data (i.e. Eynon 2013). Others have underlined the critical importance of discussing what big data can and cannot be used for (Willis and Pistilli 2014), but such question has been less often debated in the educational sphere.

There is little doubt that educational research needs to continue working on exploring the complexities of big data use in terms of teaching and learning. In order to avoid a pure technical, mechanistic and mathematical approach to analytics educational research also should engage more intensively and systematically –especially pedagogy and cognition but also in relation to social and policy research– with the different types of analytics available. Big data analytics asks questions related to its power to shape education that go beyond the development and trial of algorithms and digital artefacts. It relates to traditional education questions of education for who, for what and how. So far, the empirical research undertaken is this area, although growing over the
last years, is limited and consigned to “small” trials and experiments that rarely are replicated, scaled up or applied to different contexts. There is thus a lack of big evidences and a lack of implication in its generation by educational researchers. This provides a broad range of opportunities for educational research.

Finally, the educational research agenda should address more centrally questions around how to foster co-design processes (Sanders and Stappers 2008) –involving several stakeholders– to develop big data education artefacts. Co-design processes have succeed in education (Pennuel et al 2017), but have not been generally applied to big data. Making sense of big data is complex yet a participatory perspective to the design of digital artefacts for the use of big data in education would provide an opportunity to solve some of the problems and the challenges addressed before and somewhere else. This also would avoid technological determinist approaches to the design of big data educational artefacts. It would mean that informed, networked, empowered and active social actors would be co-creating the value of big data analytics in education. If we put together the power of big data and the promises of co-design there will be new opportunities to get better education digital artefacts.

The need for new methods, new ethics and new approaches as described before, also generates a need for educational researchers to be better equipped with skills to understand big data (Eynon 2013; Selwyn 2015). This does not mean that social scientists need to be coders or programmers, but they should, first, learn computational thinking –undergraduate programmes and research methods modules could include basic coding, modelling and simulation– in order to be able to understand the digital turn. Second, to engage in truly interdisciplinary research. Higher education curricula in the area of education are currently underprepared to tackle this challenge. They are not designed to address the digital turn and there are not many initiatives fostering this in terms of the development of computing skills, interdisciplinary work that entails the use of big data or the development of co-participatory research and design methods around big data. Skilled educational researchers should also facilitate access to education related data, as
paradoxically, big data is there but access to it in education is not as easy as one could imagine. Regarding access the problem of lack of skills goes hand in hand with the problem of the ownership of the data –usually private companies, as reviewed in section 3. MOOCs companies, Open Universities and online course suppliers provide restricted access to their data.

To sum up, as Rupert and colleagues (2015) highlight, big data is generated through social and technical practices that need to be understood. Data have a ‘social life’ and to understand education by looking at these data educational research should engage into a research agenda that more centrally considers digital sociology, new methodologies, new approaches, co-design and participatory research processes and the redesign of undergraduate and postgraduate course to equip educational researchers with the skills needed. These changes would allow us, as a community of social science researchers, to make use of the big education related data in creative ways.

5. Conclusions
This chapter has outlined the growing significance of big data not only as a topic of interest for educational research but as a social and political issue.

The chapter has explored the broad array of sources of big educational related data and the steps to make sense of it. In doing so, the chapter reviews the valuable opportunities that big data brings for educational research. It can provide new insights to help us understand a great range of issues that are at the centre of teaching and learning processes; it can enhance and improve management and policy-making process. It also open up new (self-reflective) research areas to augment our understanding of the role of big data in education.

The chapter has also explored the challenges and perils that “data centric” educational research raises, such as the methodological questions raised by big data and analytics, and those related to epistemology and ethics. The chapter offered a pragmatic perspective to methodology and highlighted the necessity for educational researchers to acquire computational and statistical skills and to engage with interdisciplinary work in order to deal with the big data phenomenon, and, at the same time, abide a sociological mind. Finally,
the chapter argues in favour of the adoption of co-design and participatory approaches to the use of big data in educational research.

References


Edwards, A., Housley, W., Williams, M., Sloan, L., and Williams, M. (2013). Digital social research, social media and the sociological imagination:


