UNIVERSIDADE CATÓLICA PORTUGUESA

JOURNALISM AND BIG DATA: THE NÓNIO CASE

Dissertation submitted to Universidade Católica Portuguesa to obtain a Master's Degree in Communication Studies in the speciality Media and Journalism

By

Vicente Maria Alexandre Lourenço

Faculty of Human Sciences

September 2020

UNIVERSIDADE CATÓLICA PORTUGUESA

JOURNALISM AND BIG DATA: THE NÓNIO CASE

Dissertation submitted to Universidade Católica Portuguesa to obtain a Master's Degree in Communication Studies in the speciality Media and Journalism

By

Vicente Maria Alexandre Lourenço

Faculty of Human Sciences

Under guidance of Prof. Dr. Nelson Ribeiro

September 2020

Abstract

As Clive Humby discribed in 2006, “Big Data is the new oil”, which is why in 2014 more than 91% of Fortune 1000 companies were investing in Big Data projects, based on the promise of higher returns. The goal of this study is to see if that is also happening with Portuguese media corporations by examining Big Data’s impact on content production and the business model of newspapers and news channels. Big Data is a term that simultaneously describes large pools of data as well as the computer techniques used to analyse those data sets. This investigation focuses on the NÓNIO project, a Big Data initiative developed by five of the largest media companies in Portugal: Cofina, Global Media, Grupo Renascença Multimédia, Impresa and Media Capital. This dissertation follows a qualitative approach. In order to gather insights on how the NÓNIO is shaping newsrooms, interviews are conducted with the people in charge of the project at each of the five media companies. Findings suggest that Big Data is having a small impact in content production, but a big one in the business model, as it is radically changing how media companies conduct advertising and how they attract and retain subscriptions.

Key words: Big Data, Media Corporations, Journalism, NÓNIO

Resumo

Tal como Clive Humby descreveu, em 2006, "Big Data é o novo petróleo". A comprovar a declaração, observou-se, em 2014, que mais de 91% das empresas da Fortune 1000 estavam a investir em projectos de Big Data. Esta investigação procura perceber se o mesmo se passa com as empresas de media portuguesas, analisando o impacto de Big Data na produção de conteúdo e no modelo de negócio dos jornais e canais de televisão noticiosos. A designação Big Data é utilizada para descrever simultaneamente grandes conjuntos de dados bem como as técnicas algorítmicas usadas para os analisar. O objeto de estudo desta investigação é o NÓNIO, um projecto de Big Data desenvolvido em conjunto por cinco dos maiores grupos de media portugueses: Cofina, Global Media, Grupo Renascença Multimédia, Impresa e Media Capital. A dissertação segue uma metodologia qualitativa. Para compreender as implicações do NÓNIO nas redacções, são feitas entrevistas aos responsáveis pelo projecto em cada uma das organizações. As conclusões revelam que a tecnologia Big Data está a ter um impacto reduzido na produção de conteúdo, mas um grande impacto no modelo de negócio das empresas de media, com as principais alterações a verificarem-se no modo como é feita a publicidade e na capacidade de retenção e atracção de subscrições.

Palavras-chave: Big Data, Empresas de media, Jornalismo, NÓNIO

Table of Contents

Introduction

"Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analysed for it to have value." This paragraph was coined by Clive Humby in 2006 (*apud* Firican, 2019). The mathematician wanted to illustrate that Data had become the most valuable resource of the 21st century and that, like it happened with oil (and gold), the world should embrace for a Data rush.

Comparing Data to oil is a good way to depict this commodity. It does bear the promise of astonishing profits, but only after refining it. In its crude form it has little value as it is merely comprised of tones of information in bytes. Valuable insights have to be carved out of the piles of Data, some of which have no interest. The carving is done though algorithms looking for patterns among the mountains of Data being generated by smartphones and the internet that, according to an article from *The Economist*:

> have made data abundant, ubiquitous and far more valuable. Whether you are going for a run, watching TV or even just sitting in traffic, virtually every activity creates a digital trace – more raw material for the data distilleries (2017).

The same article points out that in the case of Data, quantity is as important as quality. More data helps firms understand consumers better and improve their own products. A common example used to illustrate how Data abundance improves firms' products comes from Tesla with its automated cars. Tesla's cars learn how to drive by having the algorithm scanning through driving examples. The more Data there is, the more examples there are, the better and faster can the algorithm learn, and that is why Tesla was worth more than General Motors back in 2017 even though it had sold around 10% of the cars GM sold in the first quarter of that year (*The Economist*, 2017).

These large pools of Data collected by firms are so big – it is the 'Petabyte Age, said Chris Anderson in a 2008 *Wired Magazine* editorial – that such type of Data has become known as Big Data. The term also describes the techniques, such as algorithms, used to analyse the Data. Big Data has been attracting a lot of attention due to its promise of generating higher productivity and profitability. A 2014 study found that 91% of Fortune

1000 companies were investing in Big Data analytics projects (Kiron et al). Lewis and Westlund called Big Data the "buzzword *du jour*" (2014: 1). A search for the phrase "Big Data" in Google on 12th September 2020 yielded 7 260 000 000 results, proving that it is indeed a hot topic, avidly sought by scholars and companies wanting to improve their profit margins.

But how about its use in the media? Is Big Data a good investment for newsrooms? Could it revert the financial decadence of news outlets? Is there a use for Big Data in journalism?

That was the thinking that led to this dissertation, intended at examining the changing nature of journalism in an era of data deluge. Lewis and Westlund (2014), for example, point out that Big Data consists of new processes and perspectives in the way of doing journalism. Others, like Berret and Phillips (2016) see it as the natural successor of data journalism. And there are those who believe it is a path to better engagement with audiences and to improve interactivity (Pitt, 2014).

Preliminary research for this investigation showed an evident lack of studies on the impact of Big Data in Portuguese media corporations, which is why this dissertation focused on Big Data's impact on Portuguese newsrooms. It was also decided to focus on the Portuguese reality as this dissertation was written in Portugal and is part of a Master's degree at Universidade Católica Portuguesa.

Having said all this, the following research question was formed: How is Big Data changing the activity and business model of Portuguese media corporations? To serve as guiding orientation during the investigation, two other sub-questions were drawn: a) Is Big Data affecting news production? (In the sense of whether it is having an impact on what journalists write about) and b) How is Big Data affecting media corporations' revenue channels?

It was decided to analyse the ramifications of Big Data in both the business model and the activity of journalism due to the financial struggles faced by journalism, which mean that one cannot properly apprehend how newsrooms are changing if the business side is not taken into consideration. At the same time, any analysis that didn't take into account how journalism's doing is changing because of Big Data would be incomplete.

To answer these questions, this dissertation focused on the NÓNIO project, which was the object of study. NÓNIO is a Big Data initiative that results from the association of five of the largest Portuguese private media companies: Cofina, Global Media, Grupo Renascença Multimédia, Impresa and Media Capital. The project's own website describes it as a "tool for audiences' segmentation" and it works by collecting Data from users who register in the NÓNIO's platforms. NÓNIO's platforms are comprised of the websites of each of the five media groups listed above, where their own content, like news articles, is posted (in the same way *The New York Times* publishes content in its official website and platforms).

This investigation follows a four-chapter structure, plus the current introduction and a conclusion. Introduction aims at providing an overview of the investigated topic – Big Data - and why it is relevant. It presents the study object, research questions and main objectives for this dissertation. Introduction is also intended at clarifying the dissertation's structure by presenting a brief explanation of each section.

The first chapter consists of the literature review. It is an overview of a selection of academic studies on Big Data and its relation with the media and journalism. Literature review is split in two sections. The first presents an introduction to Big Data, where its history is traced and where the concept is defined, as well as it once again delves into why it is a relevant topic for the 21st century. The second half starts off by looking at journalism's relationship with Data and technology. Topics such as databased journalism, computer-assisted reporting (CAR) or precision journalism are analysed. It then proceeds to examine Big Data's influence on the media, which includes examples of Big Data projects being developed by media institutions.

Chapter two presents the methodology and the research question and sub-questions. It is then detailed the methodological approach to this dissertation. The current investigation is based on qualitative methods as it relies on interviews with field experts. By having open-ended conversations with leading experts in Big Data's fusion with the media it is possible to detect tendencies and changing behaviours among newsrooms. The methodology chapter includes as well an overview of preliminary interviews done at very beginning stages of this dissertation, and which were used to formulate the interview

questions later conducted and on which this investigation heavily relies. Each of the interview questions is also analysed at the methodology chapter.

The third chapter, entitled Findings, thoroughly examines the interviews. The section consists of an analysis of the most relevant things said by each respondent. The chapter follows the structure of the interviews.

Findings of the third chapter are then bundled in a fourth-discussion chapter. This is the section where respondents' views over Big Data and the media are set up against academic studies. The fourth chapter presents answers to the research question and sub-questions guiding this investigation. Discussion is then organized around each question, starting out with the sub-questions and ending with the main research question.

The final part of this dissertation is comprised of the conclusion where everything found out and worked on during this dissertation is put together. Literature, interviews and some of my own thoughts and impressions on Big Data's impact on the media are expressed, while the investigation questions are revisited and addressed once again. Conclusion also delves into obstacles that had to be overcome during the investigation (such as the COVID-19 pandemic) as well as limitations. On a final note, there are hints on possible future investigations regarding technology and the media.

1. Literature Review

1.1 Introduction to Big Data

1.1.1 Context: Where does Big Data come from?

The word data is derived from Latin and it means the plural of datum. Datum is the past participle of the verb dare, meaning "to give", which is generally used as "something given" in the English language. Therefore, data is the information which is given (Puschmann & Burgess, 2014). As the authors Cornelius Puschmann and Jean Burgess point out, the expression "datum", which also refers to a calendar date, was used in medieval manuscripts as "effectively timestamping the preceding text" (2014; 2). For that end, the final phrase of the manuscript would be "datum die", which translates to "given on".

According to the same authors, the word data only started being used in a theological and mathematical context in the 17th century. In fact, the first time Puschmann and Burgess found the word being used in that sense was in the issue 17 of the *Philosophical Transactions* (a scientific journal of the London Royal Society), written in 1693, where "data" twice described a mathematical variable.

Daniel Rosenberg (2013) states it was during the 18th century that the word data started being used in a more rigorous way and it stopped referring to simply given or granted information (to something generally known). From there on, it described the result of experimentation, discovery or collection of information. It was this precision of datum that allowed for the word to be used in economics.

"Data became firmly entrenched in science, business, and administration in the 19th and 20th centuries, while both its frequency and use contexts expanded significantly" (Puschmann & Burgess, 2014: 3). As for Alfred Chandler, data became an input of a firm's production function during the industrial revolution (1977). Improvements in the way it was collected, stored and analysed allowed for firms to use it in decision processes (Bajari et al., 2019).

It was in the 1940s that the term started describing information which was used, stored and analysed in the context of computing. This led to a sharp rise in the usage of the word in the 1960s (Rosenberg, 2013), which was also due to the transition from paper records to digital information. To a certain extent, Data became a synonym of digital information. The term would commonly refer to digital objects which could "be manipulated using a computer rather than generally accepted facts of the outcomes of experimentation or observation" (Puschmann & Burgess, 2014: 4). Bowker believes this new digital meaning marked a departure from previous understandings (2005).

It was in the field of geography, during the 1950s, that attempts were made to "link relatively large sets of data to the calculative power of the computer" (Barnes, 2013: 298). In fact, scientist William Warntz tried using the Newtonian potential model (a formula which allows to determine the location of an object, also known as gravitational potential) to determine "spatially variable U.S. agricultural supply and demand schedules from census data" (Warntz, 1959, *apud* Barnes, 2013: 298). It was the beginning of a quantitative revolution.

From the second half of the 20th century until the 21st century, technology changed at an incredibly fast pace, requiring faster and better market analysis. The paper "Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective" mentions this disruptive change by comparing the growing importance of the digital world. According to the authors, in 2000, 25% of the world's stored information was digital, far from the 98% that existed in 2015, when the paper was published (Xu et al., 2015).

Hidden among these massive amounts of data, stored in the cloud, are key aspects for business innovations which can provide a critical edge (Lamb, 2014). As Lewis (2015) points out there is a consensus that we are presently living in a moment of data deluge. That is mostly due to an overwhelming amount of information produced by human activity, but also due to advances in computing processing, machine learning and algorithms (Manovich, 2012; Mayer-Schönberger and Cukier, 2012; O'Neil and Schutt, 2013; Provost and Fawcett, 2013). Kauffman et al. think data is flourishing because of the internet, social networking, mobile telephones and other technologies which allow for information gathering (2012). Lewis and Westlund state that it is the ubiquity of mobile

devices, smart machines, digital trace data, as well as "fragments of social and natural activity represented by clicks, tweets, likes, GPS coordinates" that produce big amounts of data (2014: 1,2).

Sandra González-Bailón agrees with this view, pointing out that digital technologies keep track of everything people do while online. And, as the author mentions, people keep spending more and more time online. Databases keep track of people's movements and communication patterns: our lives become "quantified to unprecedented levels" (2013: 147). That not only means more information for companies and suppliers to absorb and study, but the explosion in personal tracking technologies has also opened up questions about how individuals relate with technology (Nafus & Sherman, 2014). According to González-Bailón, digital data can also help us in better understanding how to navigate social networks (2013).

Still from the consumer's perspective, the rise of social media has also changed the way people obtain information, connect with others or purchase products (Xu et al., 2015). Content such as tweets, blogs or Facebook wall postings originate click-stream data, which can be analysed by companies and used in determining a firm's strategy (Akter et al., 2016). In fact, several studies have found that most e-commerce firms such as Amazon, eBay or Alibaba rely on click-stream data to predict customer preferences and tastes (Akter et al., 2016).

How these companies collect and analyse these "big data" is different from how traditional data is handled, especially due to unique nature of Big Data (i.e volume, variety, velocity and veracity) (Akter et al., 2016). As Mayer-Schönberger and Cukier wrote, the volume of digital data and its rapid growth is staggering (2013), but more significantly is "the increasing ease with which standard computer software can manage and manipulate data sets that once required supercomputers, thereby magnifying this episode of digital data exploration" (Manovich, 2012, *apud* Lewis & Westlund, 2014: 2).

1.1.2 Defining Big Data

Chen et al. state that Big Data is a term used to describe large data sets. By large, the authors are talking about terabytes of information. So much information, in fact, that can account for exabytes (10^6 terabytes) of digital memory (2012). As an example, take Facebook, which deals with Big Data on a daily-basis and hosts 500 terabytes of data everyday (Provost & Fawcett, 2013).

Although its definition varies from paper to paper, most authors seem to agree that it is about digital data abundance, as Lewis and Westlund called it. Parasie talks about an expression referring to the "processing of massive quantities of information – government records, genetic sequences, traces left by internet users, etc… - in various domains such as scientific research, public policies or business" (2015: 1,2).

As Martha L. Stone points out, "Little" and "Big" Data have different characteristics. The digital space used to store Little Data is usually measured in gigabytes or smaller and it "can be contained on a personal computer" (2014: 1). No such thing can happen with Big Data, as it is normally required the cloud or other vast storing systems:

> To illustrate the point about the differences in storage requirements for big and little data, a seven minute high-definition video requires one gigabyte of storage. However, one petabyte, which equals one million gigabytes, could store 13.3 years of continuously running high-definition videos. Google and its video website, YouTube, processes more than 24 petabytes of Big Data per day (Stone, 2014: 1).

As Suthaharan described it, "when the volume, variety and velocity of the data are increased, the current techniques and technologies may not be able to handle storage and processing of the data" (Suthaharan, 2014: 70).

Most authors (Erevelles et al., 2016; Akter & Wamba, 2016; Russom, 2011; Schroeck et al., 2012; Gentile, 2012; Beulke, 2011) state there are 4 key characteristics that set Big Data apart from other types of data. Those 4 elements are referred to as the 4 Vs:

Volume: Big Data and its analyses rely on a large number of records. These large volumes of data are normally expressed in petabytes and exabytes. These types of data

are usually unstructured and can consist of video, image or data generated by mobile technology.

Variety: The word refers to the fact that Big Data can come from many different sources. These sources can be structured, semi-structured or unstructured. Variety is indeed a key attribute, as Big Data can originate from a wide array of formats such as text, web, tweets, audio, video, click-stream, log files, etc... As an example, take the data dealt by e-commerce firms (Biesdorf et al., 2013). Big Data used by e-commerce firms is comprised of customer profiles which include their buying patterns, regional and seasonal buying patterns, supply chain operations, advertising activities and social media information which can help predict customer's behaviour. Akter and Wamba point out that variety is a good thing for companies and that more variety can lead to better business decisions (2016).

Velocity: Big Data is also different from other types of more 'traditional' data due to its frequency. The frequency of data generation is unique to Big Data and the speed to its delivery must be accompanied by its analyses in order to provide meaning and valuable business decisions. Firms use complex techniques to capitalise on the high pace of data.

Veracity: It is important to be aware of the data quality (IBM, 2012). High data quality is necessary for a proper Big Data analyses. Only by ensuring that data is rigorous and that it complies with quality and security issues will Big Data help a firm. That is why verification is an essential part of the data management process. Assuming there are mechanisms in place to ensure the quality of data, Big Data is traditionally comprised of rigorous and high quality data.

González-Bailón suggests Big Data is not so much about the size or the volume of the data sets, but more about the quality of information it provides. It is by applying filters or by aggregating information in a certain way – by reducing these Big Data - that firms are able to extract relevant information. "Only when the data are assembled in the right way it is possible to build a story that makes substantive sense" (2013: 154). It is the level of detail provided by its analyses that makes Big Data unique.

That is why Xu et al. (2015) believe Big Data is about the insights data provide. In order to provide a thorough analyses of the data, there are specific techniques used to collect, manage and visualize data. Whenever the term Big Data is used, it is also describing the techniques used to deal with these large sets of unstructured and complex data (Chen et al., 2012). Therefore, Big Data is also the "techniques and technologies that make handling data at extreme scale affordable" (Forrester, 2011: 4). Another advantage of Big Data is how it allows firms to analyse large data sets in real time, instead of waiting several weeks or months for the information to be studied and for a conclusion to be drawn (Sathi, 2014; Mayer-Schönberger and Cukier, 2012). As Martha L. Stone claims, Big Data is "an umbrella term for a variety of strategies and tactics that involve massive data sets, and technologies that make sense out of these mindboggling reams of data" (2014: 2). That is why technologies keep being developed to automate the process of data analyses (Stone, 2014).

All in all, Big Data is as much about the size of the data as it is about the scope of data itself. Mayer-Schönberger and Cukier describe it as "our newfound ability to crunch a vast quantity of information, analyse it instantly, and draw sometimes astonishing conclusions from it" (Mayer-Schönberger and Cukier, 2013: 6).

Crawford et al. share Stone's opinion. The authors claim Big Data is a plastic concept as it assumes different meanings according to the contexts in which it is used. It can both serve to police cities as well as to predict preferences for breakfast cereals (2014). It is a social, cultural and technological phenomenon, resulting from the combination of three dynamics (Lewis & Westlund, 2014), technology, analysis and mythology:

> (…) 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 (boyd & Crawford, 2012 *apud* Lewis & Westlund, 2014: 3).

Goes points out that one of the key advantages of Big Data is helping the decision-making process. The author thus defines the term Big Data as various observational data that support different decisions (2014). In some cases, the algorithms of data analyses and processing are the ones making decisions, which evokes descriptions of utopian

cyberspaces (Puschmann & Burgess, 2014). These computational nervous systems take data from different sources and make decisions in real time (Webster, 2012). They behave like "sentient" organisms (Puschmann & Burgess, 2014). Consequently, Big Data provides firms with an edge. Akter et al. (2016) stress out that Big Data Analytics Capability is the ability of firms to set the optimal price, detect quality problems, decide the lowest possible level of inventory and identify loyal or profitable customers. Akter and Wamba (2016) share the case of Amazon. Amazon was able to use large volumes of unstructured data to develop "sophisticated recommendation engines that deliver over 35% of all sales" (pp8), as well as to program automated customer service and algorithms that calculate the best competing prices, by drawing comparisons with other websites every 15 seconds. Manyika et al. (2011) point out that this is a common practice among many other e-commerce firms such as eBay, Expedia and Travelocity.

Therefore, the ultimate goal of Big Data analysis is to generate business value, which can be translated to economically worthy insights, from the proliferation of data (Beath et al., 2012; Akter & Wamba, 2016).

That may be why the genesis of the term lies in the business world. Puschmann and Burgess (2014) argue that companies pushed for better results through technology, which allowed for the creation of predictive models – to "interrogate the stored information systematically to make predictions" (Kennedy, 2012, *apud* Puschmann & Burgess, 2014: 1693).

Predicting and personalizing

Kopp defines predictive analytics as the ability to identify events before they take place through the use of Big Data (2013). It can help firms preparing revenue budgets and recognizing future sales patterns. Predicting what customers will want in the future allows companies to better optimize their resources, for example by ordering more appropriate quantities of inventory, which can prevent product stock outs or lost customers (Mehra, 2013; Akter & Wamba, 2016). Netflix used this predictive ability to decide on whether or not to invest in a TV series called "House of Cards" (Ramaswamy, 2013). In fact,

Netflix was so confident in its predictive model that it didn't require a pilot episode, but instead immediately asked for two seasons of 26 episodes of content (Marr, 2016).

In the third chapter of his book, *Big Data in Practice – How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*, Bernard Marr states that networks, producers and streaming platforms rely heavily in their predictive ability to determine what consumers will want to watch. The preferences of consumers are assessed quantitatively, making it one of the fundamental principles of Big Data. Netflix was a pioneer in this field, says the author, which is why it can now identify what films a particular user will like watching with much more accuracy than simply identifying that he or she likes horror or spy films. At the very end of the chapter, Marr talks about how:

> Netflix have begun to build the foundations of "personalized TV", where individual viewers will have their own schedule of Netflix entertainment to consume, based on analysis of their preferences. This idea has been talked about for a long time by TV networks but now we are beginning to see it become a reality in the age of Big Data (2016: 21,22).

Personalization is another application of Big Data (Koutsabasis et al., 2008). This personalization can come in the form of special viewing content, in the case of Netflix, HBO or Amazon, but it can also comprise of other types of special content such as promotions for customers (Akter & Wamba, 2016). Personalized services help firms in separating loyal customers from new ones in order to adjust promotions accordingly (Mehra, 2013). Personalization is so important for retail firms that Liebowitz (2013) estimates that it can increase sales by more than 10% and to provide five to eight times the return on investment (ROI) on marketing expenditures.

Big Data marked a shift from traditional database management systems to platforms offering advantages in the long-run. After its genesis in the business world, the term has nowadays acquired a level of abstraction. Even though data can and is used as a count noun with singular and plural forms, Big Data is grammatically a mass noun. Somewhere along the way, there was a conceptual shift of emphasis from individual pieces of information to a homogeneous aggregate state. That is how it is perceived nowadays. Its current popularity arose in 2011. It was at that time that Big Data sparked the curiosity of

people and the media when the company Accel pledged 100 million dollars to startups involved with Big Data through the initiative Accel Partners Big Data Fund (Puschmann & Burgess, 2014).

1.1.3 Why Big Data matters

In 2013, IBM estimated that about 2.5 quintillion bytes of data were generated each day (Yakabuksi, 2013). There is an unprecedented amount of data being generated in the 21st century which can all be analysed (Barnes, 2013). Also in 2013, Mike Batty wrote a paper on how his transportation research unit at University College London was analysing the information of 7 million daily journeys taken on London's public transportation system. That amounts to a set of 15 billion data over a 5-year period. As Lewis points out, there "(…) seems to be broad agreement that, in the developed world of digital information technologies, we are situated in a moment of data deluge" (2015: 2). Manyika et al. talks about how "in a digitalized world, consumers going about their day – communicating, browsing, buying, sharing, searching – create their own enormous trails of data" (2011: 1). "This data layer is a shadow. It's part of how we live. It is always there but seldom observed" (Bell, 2012: 48).

The abundance of data has enabled firms and governments to analyse public life, much to the concern of critics worried about consumer privacy and data ethics (boyd & Crawford, 2012; Oboler et al., 2012). For better or worse, data-driven techniques of computation and quantification are crucial to understanding the intersection of media, technology and society (González-Bailón, 2013).

Media companies and marketers have been particularly welcoming of this data revolution due to Big Data's ability to measure audience ratings (Kosterich & Napoli, 2015). Traditionally, media managers and marketers have relied on representative samples of the population to assess media audiences (Webster et al., 2014). Questionnaires, diaries and meters have all been used to target respondents. These techniques assess the population sample's characteristics, their likes and dislikes, ability to recall messages and their behaviours, all of which are later projected to the larger population. Even though these metrics have been useful in the past, few rise to the level of currencies (Nelson & Webster, 2016).

According to Napoli (2012), currencies are a class of metrics used by advertisers to quantify audience attributes. Due to their unique characteristics, currencies are a coin-of-exchange that buyers and sellers of media use to conduct business. They must be cost

effective to produce and agreed by the affected parties (Nelson & Webster, 2016). One classic example of a currency is time spent with media which stakeholders believe reflects engagement or loyalty (Ingram, 2014, *apud* Nelson & Webster, 2016). Even though there is no clear definition of what engagement is, there seems to be a general agreement that it refers somewhat to content retention (Plummer, 2006, *apud* Nelson & Webster, 2016). This is important as it is believed that retention follows attention. Therefore, measuring retention is a metric for attention, for quantifying engagement (Abourezk, 2014). Retention used to be assessed through special surveys, but it is now being measured through web analytic tools like cursor movements, clicks and time spent with content (Nelson & Webster, 2016).

Online publishers favour the term "attention minutes". Attention minutes is the time users spend looking at particular programs, pages and stories (Nelson & Webster, 2016).

When it comes to media consumption, the industry and academic research are split between analysing the frequency and the duration. Frequency relates to how many times a site is visited or to how many episodes of a program are watched and it is a metric of loyalty (Ehrenberg & Wakshlag, 1987). Duration of consumption is the average time spent on a site visit or average amount of a program episode consumed and normally points to notions of attentiveness or engagement (O'Brien & Lebow, 2013). The problem with both analyses is that they are subjected to the double jeopardy effect where the highest rated television programs and movies are the ones retaining audiences episode-to-episode, while the lower rating ones have far fewer episodes of gaining attention (Ehrenberg & Wakshlag, 1987). This is where Big Data can make a difference as it can help producers to more accurately understand why a certain TV show is succeeding, in terms of frequency and duration, in a way where the double jeopardy effect has little or no impact on the analyses (Nelson & Webster, 2016). "Big Data may offer a fresh look and new insights precisely because it is unencumbered by the conventional thinking and inherent biases implicit in the theories of a specific field" (Mayer-Schönberger & Cukier, 2013: 71).

Later in the same paper, Mayer-Schönberger and Cukier point that "Big Data is poised to reshape the way we live, work and think" (2013: 190). Big Data has emerged as an exciting frontier of productivity and opportunity, and firms are keen on exploring it in

their search for a competitive advantage (Akter et al., 2016). It has changed how firms conduct business (Barton & Court, 2012). The truth is that humans are constantly generating data which can be used by companies: "Every interaction, every communication, every touchpoint creates a digital breadcrumb – a piece of data that can be analysed and manipulated" (Dwoskin, 2014, *apud* Lewis & Westlund, 2014: 1). Strawn called it the "fourth paradigm of science" (2012) and McAfee and Brynjolfsson the next "management revolution" (2012).

The field of Big Data has seen a spike in investment. According to Lunden, in 2013 around 2.1 trillion dollars were spent in stimulating Big Data analytics. Accenture and General Electric reported in 2014 that 87% of enterprises believed Big Data analytics would redefine the competitive landscape of their industries within the next 3 years. At the time, 89% believed that companies that didn't adopt a Big Data strategy risked losing their market share and momentum (Columbus, 2014).

That is because Big Data analytics enables firms to use data more efficiently, drive a higher conversion rate, improve decision making and empower customers (Miller, 2013). It also offers the perspective of improving service innovation models, which has been reflected by firms such as Rolls Royce or Google (Barret et al., 2015; Davenport and Harris, 2007). The overall existing literature implies that Big Data can improve employment and that it can increase productivity while also improving consumer surplus (Loebbecke & Picot, 2015).

Satya Ramaswamy's (2013) investigation showed that companies with huge investments in Big Data were generating excess returns and gaining competitive advantages, while threatening the position of those companies without significant investments in Big Data. Ramaswamy found a correlation between higher spending in the Big Data department and larger revenue growth. In some cases, companies got a return of more than 50% on their investments. Of the 53% of the survey respondents that had Big Data initiatives, median spending per company was 10 million dollars; a fairly small amount when compared to the median revenue of 6.9 billion dollars per company. Still, the median spending hid a great disparity in Big Data investments, as 7% of the respondents reported to have invested a minimum of 500 million dollars each on Big Data software, hardware, data scientists, consultants and other related expenses.

McAfee and Brynjolfsson calculated in 2012 that e-commerce firms using Big Data analytics in their value chain experienced 5% to 6% higher productivity than their competitors. As Jao explained Big Data allows e-commerce firms to "track each user's behaviour and connect the dots to determine the most effective ways to convert onetime customers into repeat buyers" (2013: 1). Because of this explosion of interest in Big Data and in how it can increase the productivity of a firm, Kiron et al. reported that 91% of Fortune 1000 companies were investing in Big Data projects back in 2014.

The consultancy firm Bain & Company summarized the Big Data payoff in 4 topics. According to the firm's 2013 report, *Big Data: The Organization Challenge*, businesses that use Big Data in their value chain are:

- 5x more likely to make decisions faster than market peers
- 3x more likely to execute decisions as intended
- Twice as likely to be in the top quarter of financial performances within their industries
- Twice as likely to use data very frequently when making decisions

For media companies specifically, the payoff includes engaging the audience more deeply with more targeted news and advertising, more relevant and engaging content and more compelling videos and photos (Stone, 2014). "The key point here is that Big Data in the media sector is not just an audience-centric phenomenon; it can be content-centric as well" (Napoli, 2016: 1). Apart from this, Big Data also gives firms the ability to compete with other sophisticated online media companies (Stone, 2014). Business intelligence, as Xu et al. refer to it, allows monitoring competitors' new designs, their key-product features, pricing strategies and customer feedback (Xu et al., 2015).

But above all, what changes in this digital era is the size of the samples: "now it is possible to conduct experiments with millions of people and thus identify effects that would have been impossible to grasp with smaller data sets" (González-Bailón, 2013). Large data sets drawn from mobile phone communications or location-based software have allowed researchers to extract valuable insights which used to be out of reach (Gonzalez et al., 2008; Noulas et al., 2012). Researchers can now measure the predictability of individual

mobility patters by identifying high frequented locations and by analysing how social networks emerge out of those locations. Through Big Data, scientists are now able to tackle the complexity of urban life (González-Bailón, 2013). That is only possible due to Big Data's ability to (Flew et al., 2011):

- Synthesize and give meaning to large amounts of data
- Enable humans to achieve insights from data fusion and processing
- Infer hypotheses
- Enable people to have access to the intuitions of others
- Present information in relevant ways that improve humans' knowledge about the subject matter (Klein et al., 2006).

1.1.4 Limitations of Big Data

The biggest challenge for Big Data users is finding the right information about each customer out of the large pool of data (Miller, 2013). It is why it is important to discourage that computational techniques and the avalanche of numbers become ends in themselves: "That is, techniques and numbers become fetishized, put on a pedestal, prized for what they are rather than for what they do" (Barnes, 2013: 299). This phenomenon was observed back in the 1970s, during geography's quantitative revolution, to which David Harvey wrote that:

> there is a clear disparity between the sophisticated theoretical and methodological framework which we are using and our ability to say anything really meaningful about events as they unfold around us (1972: 6).

A distinction between data and knowledge must be made: "Clearly, Big Data has information coming out of its ears, but is it generating useful knowledge?" (Barnes, 2013: 299). Barnes warns about the dangers of collecting data for data's sake. Nate Silver, a Big Data statistician who managed to successfully predict the November 2012 U.S. election results in each of the 50 states, observed that most data is just noise, "as most of the universe is filled with empty space" (2012: 50). However, as its output comes in a mathematical form – the hallmark of science (as Galileo said: "mathematics is nature's language") – this noise is regarded as knowledge (Barnes, 2013).

Facebook, Google and Twitter possess vast amounts of person-specific information, but all their data becomes useless unless they understand what is relevant and, from there, manage to create personalized offers, set dynamic prices and use the right channels to provide consumer value (Akter & Wamba, 2016). "Theory and interpretation are more necessary than ever before if we are to find the appropriate layer of information in what otherwise is an unnavigable sea" (González-Bailón, 2013: 147). As a consequence, the role played by researchers in the Big Data analytical process should not be underestimated: identifying what is important is a subjective matter, as is providing the context which will allow for meaningful connections and to disregard those that are unsubstantial (Silver, 2012).

As previously mentioned, Big Data is about data insights (Xu et al., 2015). However, the challenge lies in how to turn data into meaningful insights so that managers can use it to solve problems. As Xu et al. called it, "Big Data is just a raw material, not a solution" (2015: 1). Gitelman talks about how raw data is an "oxymoron" to explain that data requires interpretation (2013).

There are other limitations for Big Data such as not aligning with an organizational culture and capabilities (Kiron et al., 2014) and facing suspicion by employees who are traditionally reluctant to trust the data and the conclusions it generates, mostly because they can't understand it (Akter & Wamba, 2016). Bose defends that to overcome those problems managers should strive to present Big Data in a comprehensible format such as a dashboard, reports or other types of visualization systems (2009). Kiron supports an engaged discussion between managers and employees where the top management incentivizes the adoption by employees of Big Data (2014). As McAfee and Brynjolfsson argue, it is unlikely for a firm to be a top performer through Big Data unless all workers are on the same page regarding the company's strategy and goals (2012).

There are also ethical concerns surrounding the use of Big Data (a topic which will be later expanded in the Literature Review chapter). Questions about privacy or accuracy of the information keep being raised whenever the massive amount of data yield social, cultural and monetary value (boyd & Crawford, 2012).

Lewis and Westlund write that the data privacy and surveillance questions raised by Edward Sonwden's revelations, among other incidents, "make Big Data a matter of public as well as professional concern" (2014: 2). Still, despite Snowden's claims, González-Bailón makes an important point about how people are willing to disclose personal information in exchange for something of value and not that well understood, which is why the author believes the benefits of sharing data mostly offset the dangers (2013). González-Bailón concludes that the benefits of Big Data depend on how it is applied, but that the general public should also demand that it is used in a responsible way.

On a slightly different critic, academics complain that Big Data poses methodological challenges, such as trading large scale for reduced depth, even though the proponents of Big Data infer a degree of scientific authority from the abundance of information

available (Mahrt & Scharkow, 2013). It is true that Big Data can help in predicting a wide variety of natural and social phenomena, but, at the same time, it can offer very little on an explanatory level, especially when misunderstood (Bowker, 2013). The same author talks about how the obsession with Big Data is making it an instrument of control. Puschmann and Burgess (2014) agree with Bowker and warn about the dangers of focusing too much on the data presented rather than its relation to the world it is supposed to represent.

More than a technological transition, Big Data represents a sociotechnical phenomenon with cultural, economic and political origins and implications, making it as much of a mythology as a science or business (boyd & Crawford, 2012).

1.2 Big Data and journalism

1.2.1 Evolution of databased journalism

"Journalism has long been familiar with data and databases as an object of news work and journalistic evidence (…)" (Lewis & Westlund, 2014: 4). In fact, quantitative skills have always been vital to a journalist's profession (Berret & Phillips, 2016). In the beginning of the 1900s, Joseph Pulitzer wrote an essay on the importance of teaching journalism at universities. In that essay, the author outlined the skills he thought journalists needed to carry their civic role, which included learning statistics:

> Everybody says that statistics should be taught. But how? Statistics are not simply figures. It is said that nothing lies like figures - except facts. You want statistics to tell you the truth. You can find truth there if you know how to get at it, and romance, human interest, humor and fascinating revelations, as well. The journalist must know how to find all these things - truth, of course, first (Pulitzer, 1904: 53).

According to Mark Coddington (2015), numbers have always played a role in journalism, which is why newsrooms welcomed the introduction of computers in the 1950s (Meyer, 1973). Computers allowed for what Meyer called "precision journalism" by helping journalists achieve greater accuracy through the use of databases. It was Meyer himself who established computers as a fundamental tool in the work process of journalists, with his coverage of the 1967 rioting in Detroit, in which he used a computer to create a survey in order to assess how those involved perceived the event (Berret & Phillips, 2016). Meyer's coverage of the incident granted him the 1968 Pulitzer Prize for Local General Reporting. In his book, *Precision Journalism: A Reporter's Introduction to Social Science Methods*, published in 1973, Meyer wrote the following:

> The tools of sampling, computer analysis, and statistical inference increased the traditional power of the reporter without changing the nature of his or her mission to find the facts, to understand them, and to explain them without wasting time (Meyer, 1973: 3).

Meyer's work is said to be the beginning of what has been termed precision journalism or computer-assisted reporting (Berret & Phillips, 2016). The precision journalism method was modelled after social science, using empirical methods and statistical

analysis to reach more assertive and definite answers to journalistic questions (Coddington, 2015). Meyer's influence of the use of computers in journalism is such that he is dubbed the godfather of computer-assisted reporting. And, according to Coddington (2015), it is due to Meyer that computer-assisted reporting is at the core of all data-driven journalism practice. Computer-assisted reporting embodied two main ideas from Meyer's precision journalism: data gathering and statistical analysis (Miller, 1998). But it also paved the way for a broader use of computer-based information, online research and even email interviews (Yarnall et al., 2008).

According to Cox (2000), computer-assisted reporting, also known as CAR, promised a more scientific form of journalism, allowing for cross-checking information with data publicly available through the internet. Truth claims could now be tested and backed through verifiable data. As Schudson (1978) put it, facts have always been of critical importance to journalists who are set on providing a fact report in a detached, unbiased and impersonal manner. Sylvain Parasie (2015) adds that journalists, especially investigative journalists, believe that there is only one true and complete statement of what happened, which is independent of their investigation. Computers can therefore help in reaching this truth unique report.

In the 1980s and 1990s, computer-assisted reporting became the mainstream type of reports compiled in newsrooms (Coddington, 2015). This was influenced by several high-profile Pulitzer Prize-winning stories (Houston, 1996).

The replacement of typewriters by desktop personal computers in most newsrooms around the world, in the 1980s, would also be of critical importance for a more general use of CAR. In 1986, a reporter named Elliot Jaspin, who worked at the *Providence Journal-Bulletin*, used databases to match felons and bad driving records to school drivers. In 1988, Bill Dedman, a reporter at the *Atlanta Journal Constitution*, managed to prove that banks weren't providing services to African Americans, while still accounting for the needs of white people living in the poorest neighbourhoods. The series of articles published by Bill Dedman was called "The Color of Money" and it won a Pulitzer Prize in Investigative Reporting (Berret & Phillips, 2016).

Later in the 1990s, the increasing prominence of the internet meant that more and more news organizations posted stories online. As a result, journalism education focused on a more digital teaching. Universities started teaching multimedia, online video skills, HTML coding and data analysis for storytelling (Berret & Phillips, 2016).

As pointed out by the authors, despite being taught at university, most journalists still learned these skills through conferences or from other journalists. Berret and Phillips (2016) explain that working journalists were the ones who first saw the potential of analysing data through the use of spreadsheets and databases, and that they were the first to develop in the field such quantitative reporting methods.

The introduction of computers and a more mainstream use of spreadsheets, databases and the internet allowed for what Mark Coddington (2015) and Petre (2013) called a "quantitative turn" in journalism (2015). Anderson states that the story of quantitative journalism in the United States is one of rupture and of transformed techniques (2015). As more information became ones and zeros, the more journalism involved analysing and computing quantitative data as well (Coddington, 2015).

Anderson (2015) states that this "quantitative turn" manifests itself in an approach where the objects being studied by the journalist are analysed through databases. The author goes on explaining that data is embodied in objects such as databases, survey reports and paper documents.

The use of algorithms, machine learning and other computational methods to accomplish journalistic goals gave rise to what is called "computational journalism". Berret and Phillips (2016) defend that what computational journalism adds to the field of journalism is how it uses algorithms to mine unstructured data in new ways, the creation of digital platforms to better manage documents and technologies as well as the introduction of programming languages like Python, Ruby and R. Computational journalism is also credited for popularizing the application Jupyter amongst journalists, which provides a framework to successfully mix code and prose.

Hamilton and Turner defined computational journalism as the combination of algorithms and knowledge from social sciences that supplement the "accountability function of

journalism" (2009: 2). The authors claim that it builds on computer-assisted reporting as it helps reporters to explore large amounts of structured and unstructured information in their search for stories.

Flew et al. (2011) believed computational journalism worked by finding sets of data for analysis, making sense of such data and by representing it in an interesting, accessible and newsworthy way. For that to be useful to investigative journalists, it is necessary that the datasets used are rigorous and fact-checked.

Some authors are not fond of the term as they claim computers have long been a vital part of journalism since typewriters were replaced (Myers, 2009; Miller & Page, 2007). Miller and Page add that it is important to distinguish between computers as tools and computation as theory. The authors argue that the theory involves understanding the processes underlying the computations. By processes, Miller and Page refer to things that used to be done by humans but that are performed with greater speed and accuracy by computational devices such as searching for information, filtering or identifying patterns.

Despite not being averse to the term, Coddington's view on computational journalism is somewhat similar to Miller and Page's. He defined it as:

> a strand of technologically oriented journalism centred on the application of computing and computational thinking to the practices of information gathering, sense-making and information presentation, rather than the journalistic use of data or social science methods more generally (2015: 335).

Coddington adds that abstraction is at the root of computational thinking. His view comes from Wing's belief that it is a cognitive process and not simply a practice done by computer (2008). Wing concludes by stating that computing is therefore the automation of abstracted information.

As an example of computational journalism, Coddington points to *The Wall Street Journal*'s "use of simulated user profiles to determine the algorithms governing price discrimination in online commerce" (2015: 336). The platform Document Cloud, created by *ProPublica* and *The New York Times* in 2008, which compiles user-submitted and

user-annotated documents and offers computational tools to process them such as optical character recognition is another breakthrough of computational journalism (Cohen et al., 2011).

Computational journalism gained a lot of attention in the last two decades, which, according to Flew et al., is due to three main factors:

1. Increasing amounts of publicly available data, especially information from government sources, even if sometimes such information is divulged through unofficial channels such as WikiLeaks. Flew et al. point to the examples of the US Open Government Initiative or Britain's data.gov.uk as proof that governments are making information more openly accessible.
2. Declining cost, greater ease of use and increasing power of data-mining software.
3. Explosion of online participation and engagement that comes from "a plethora of Web 2.0 and social media sites" (2011: 4).

Computational journalism has proved to be of critical importance for newspapers as it helped them to adapt to a changing environment characterized by a greater speed of news circulation. In that sense, computers were vital for the survival of newsrooms in the face of decreasing profits, by decreasing costs, while maintaining the quality and accuracy of the final product (Flew et al., 2011). This opinion is shared by Anderson et al., who argued in 2012 that automation paved the way for news media to cut expenses, for example by no longer wasting resources on stories that could be written by robots, and at the same time creating value, like redeploying humans to projects where they could provide a unique contribute. As Carlson argued, algorithms are capable of prioritizing, classifying and filtering information (2014). Algorithms are also useful in journalism at several stages, such as distribution or at writing stories themselves (Carlson, 2014).

Computation journalism was also crucial in attracting online audiences as participants and not mere readers/consumers (Flew et al., 2011). It enhances engagement and enables greater interaction with news media by making interactive content available to readers.

Another advantage of mixing computational methods with journalism techniques is how it can bring together technology experts and journalists. This type of brainstorming ends

up being beneficial for both fields as it leads to the development of new ideas (Flew et al., 2011).

Computational journalism is many times confused with data journalism, which is also a successor of computer-assisted reporting (Coddington, 2015). But while computational journalism is focused on the application of the processes of abstraction and automation, data journalism is characterized "by its participatory openness and cross-field hybridity" (Coddington, 2015: 337).

Stavelin defined data journalism as working primarily through data sets analysis to produce data oriented stories (2014). It "retains CAR's emphasis on subordinating data to the professional journalistic value of narrative and the 'story'" (Coddington, 2015: 339). Howard put it as "telling stories with numbers or finding stories in them" (2014: 5). Unlike computer-assisted reporting, data journalism isn't only focused on investigative journalism (Mashall, 2011; Minkoff, 2010).

Parasie and Dagiral observe that journalists have long worked with many types of data, but what is different now is how news organizations increasingly need computer programming, sophisticated databases and data science techniques to comprehend the amount of data that abounds nowadays (2013).

Rebekah McBride (2016) agrees with this view. She says that it is hard to point to the exact moment data journalism was born. *The Manchester Guardian* (nowadays simply known as *The Guardian*) claims to have been the first newspaper to publish a data journalism story back in 1821. However, other scholars argue that Florence Nightingale's statistical reports on the Crimea War were the first evidence of data journalism. Others opt for a more conservative approach and pinpoint CBS's use of a computer to predict the outcome of the US presidential elections of 1952 as its first moment (McBride, 2016).

In Bell's opinion, data journalism was only born in the 2000s, before which most data analysis within newsrooms were in the form of CAR or in news organizations that specialized in financial information (2012).

Most authors make zero or little distinction between data journalism and data reporting. Charles Berret and Cherry Phillips (2016) describe data reporting as obtaining, cleaning and scrutinizing data for use in story-telling, particularly in telling journalistic stories. The same authors claim that it is basically the use of computer-assisted reporting for writing journalistic stories. Another way to look at what data reporting (or data journalism) is, is by drawing a comparison with Philip Meyer's precision journalism. Berret and Phillips claim that it is nothing more than mixing it with the use of social science research methods in the interest of journalism.

In their book, *Teaching Data and Computational Journalism*, Berret and Phillips argue that journalists with an understanding of data and computation are more qualified and better capable of doing their jobs in an era ever more reliant on complicated streams of information. That is why the authors believe that journalists need to learn how to visualize data through maps and charts, in order to conduct thorough analysis, as well as learning how to code in a way that it helps them to analyse data for writing journalistic stories. Apart from that, this new era of technology requires that journalists become acquainted with using web scraping tools and techniques and relational database software and understand statistical concepts and software with statistical packages such as SPSS or R (Berret & Phillips, 2016).

All these new investigative techniques and technologies are in the interest of the public (Stray, 2011). It is important to bear in mind journalism's vital role for democracy, as its ability to publicly denounce the behaviours of those in power is crucial to the well-being of any society (McNair, 2006). That is also why data journalism is characterized by another particular feature: the use and proliferation of open data and open-source tools to analyse and display data, which means journalists from different newsrooms and even from different countries (and even other types of collaborators such as readers) can cooperate in complex and intricate investigations (Gynnild, 2014). Data journalist and researcher Liliana Bounegru worte that:

> By enabling anyone to drill down into data sources and find information that is relevant to them, as well as to verify assertions and challenge commonly received assumptions, data journalism effectively represents the mass democratisation of resources, tools, techniques and methodologies that were previously used by specialists – whether

> investigative reporters, social scientists, statisticians, analysts or other experts (Gray et al., 2012: 22).

Data journalism arises from the junction between journalism and the social, computer and information sciences (Lewis, 2015). But despite being a popular field, there is little studying on the subject. As Lewis points out, the role of data in journalism and the interrelated notions of algorithms, computer code and programming in the context of news took time until it began to receive attention: "Among scholars, there is a rapidly growing body of work focused on unpacking the nature of computation and quantification in news." (Lewis, 2015: 2).

1.2.2 How Big Data is changing journalism

Lewis and Westlund (2014) believe Big Data is neither good nor bad for journalism as it simply relates to processes and perspectives on how to quantify information. The big drawback, the authors claim, is that journalism it still trying to make sense of what Big Data is and how it can be used. Because of that confusion, agents in news media organizations end up trying different approaches: there are those who resist and reject Big Data completely, others try to adapt to it and some are shaped by this new technology. Seth C. Lewis wrote that in 2015 there was a lot of experimentation going on with digital data, computational techniques and algorithmic forms of representing and interpreting the world, which made it difficult to properly comprehend what was becoming of journalism because of Big Data's influence.

Lewis and Westlund (2014) see the fusion of Big Data and journalism as the intersection of editorial, business and technological interests and practices. The authors then question how such intersection can affect newsrooms' interaction with audiences. In Lewis and Westlund's opinion, the way Big Data can improve the engagement with audiences is through its major advantage of using analytical data tools. Lowrey (2009) claims journalists have long been oblivious to readers and viewers, but, via digital metrics, that has changed and, nowadays, it has a direct impact on how news are presented to the public (Tandoc, 2014).

Big Data's implications for journalism can range from knowledge work and economic rationale to practical skills or philosophical ethics. Crawford et al. (2014) point out that one obstacle to journalism's understanding of Big Data lies with the term itself as it is too vast. Lewis and Westlund relate to Crawford et al.'s view, however, they argue that it is at the same time the "most succinct way of referring to a larger and complicated set of facts at play in technology and society as well as in technology and journalism" (2014: 3).

As it has been pointed out in the previous chapter, many authors see the interaction of Big Data and journalism as data journalism. Charles Berret and Cherryl Phillips (2016) define data journalism as a term used to describe using data for finding and telling stories in the public interest. According to the authors, through the use of algorithms, machine learning

and other emerging technologies, data journalism can take many forms such as analysing data, conveying such analysis in a written form, verifying data in reports, visualizing data, etc…

Alexander Howard sees it as "the application of data science to journalism, where data journalism is defined as the study of the extraction of knowledge from data" (2014: 4). To do it, it is crucial that an ability such as digital literacy is applied to every area of journalistic practice (Berret & Phillips, 2016).

But, as Flew et al. (2011) argue, computational journalism techniques and data journalism go beyond number crunching and search-based activities. The use of Big Data in journalism is very much about the engagement with audiences as well (Pitt, 2014). Pitt gives examples of this engagement by pointing to newsrooms that have created projects where the participation of audiences is required and encouraged. These types of new products carry certain epistemological assumptions about how audiences might acquire knowledge, as users are encouraged to interact with the data (Lewis & Westlund, 2014).

The notion of interactivity actually dates back to the emergence of online journalism in the late 1990s (Young et al., 2018). In those early days, news sites tended to focus on navigational and functional interactivity (Massey & Levy, 1999). Nowadays, it is all about social sharing, most read lists of stories and comment sections (Stroud et al., 2016). Tools such as interactive data visualizations have become popular among news websites as interactivity is considered a key aspect of online graphics as opposed to those which appear on print (Burmester et al., 2010). The inherent supposition being that online journalism has, by its very nature, to be interactive (Deuze, 2004).

In the beginning of the 2000s, most research in journalism studies focused on interactivity and on the ability for users to participate in news production work (Young et al., 2018). Domingo points out that interactivity was a very popular term in the 2000s. What followed, argued the author, was the term "participatory journalism" (2008: 680). Participatory journalism is a way for newsrooms to better engage with audiences (Singer et al., 2011). It is an opportunity for audiences to select, customize, highlight or participate in information flows, where users are framed as active participants in creating, distributing and consuming news (Singer et al., 2011).

There is an ongoing discussion about how the industry of journalism has adopted and implemented interactivity or participatory journalism (Young et al., 2018). David Domingo (2008) talks about the "myth of interactivity". He admits that online journalism pursued interactivity because it was believed to be a worthy ideal, but at the same time news organizations tried for it not to affect news production. On the other hand, Stroud, Scacco and Curry have found little evidence of newsrooms adopting and deploying interactive features (2016).

Participatory journalism is only possible because media companies are constantly collecting reams of data from every area of their organisations such as sales and advertising, readership and membership, content, accounting and many more (Stone, 2014). Therefore, it is safe to conclude that data journalism is everywhere (Fink & Anderson, 2014). The ethical consequences that arise from data journalism's omnipresence lead Lewis and Westlund (2015a) to ask for a set of perspectives which highlight the different interrelated roles of social "actants" at this emerging intersection of journalism. The authors share Lewis and Usher's opinion that there are a multitude of views over the phenomenon. Anderson wrote that one should bear in mind that algorithms and related computational tools and techniques "are neither entirely material, nor are they entirely human – they are hybrid, composed of both human intentionality and material obduracy" (2012: 1016).

The path to better comprehend journalism's potential use of Big Data comes from exploring not only the uses journalists make of computational tools, but by analysing the underlying principles and processes through which such tools are developed and for which they are employed (Flew et al., 2011). It is not "about getting journalists to think or act like computers, but enabling them to use computing devices to tackle problems beyond the scope of everyday action prior to the age of computing" (Flew et al., 2011: 2).

Humans using software to improve journalistic reports will become more and more automatized as artificial intelligence keeps improving (Flew et al., 2011). Minsky defined artificial intelligence as the "science of making machines to do things that require intelligence if done by a man" (1968: 5).

Journalism's automatization is a phenomenon long studied by Matt Carlson. In his 2014 paper "The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority" the author tries to explain what happens as Big Data's role in journalism shifts from a reporting tool to the generation of news content. Carlson claims that nothing is as disruptive to journalism as the algorithmic process converting data into news texts with limited human intervention (automated journalism) as it raises important questions regarding our understanding of what journalism is and how it ought to operate.

Mary Lynn Young and Alfred Hermida (2015) shared similar concerns after studying the emergence of computationally based crime news at *The Los Angeles Times*. The authors observed that many stories were being generated automatically through "robo-posts". Young and Hermida then concluded that one should question "how decisions of inclusion and exclusion are made, what styles of reasoning are employed, whose values are embedded into the technology, and how they affect public understanding of complex issues" (2015: 4).

Nicholas Diakopoulos (2014a) is another author concerned with interrogating the algorithm. Diakopoulos focuses on the notion of algorithmic accountability reporting, which he describes as a mechanism for explaining the power structures, biases and influences of computational journalism in society. He mentions that understanding algorithmic power means analysing the atomic decisions made by algorithms, including prioritization, classification, association and filtering.

It is by looking at what is happening with data journalism in French-speaking Belgium that de Maeyer and colleagues (2015) found that it is not only the production of journalistic artefacts that has an impact on the notion of data journalism, but also the discursive efforts of all actors involved. De Maeyer et al. point out that data journalism faces the threat of becoming nothing more than an overlap of discourses by a range of actors with origins from different social worlds. The authors argue that one should be cautious about the path of experimentation to where it may lead, as, in the case of data journalism algorithms are the end of human creations.

Lewis and Westlund (2015b) argue that developments of Big Data are changing journalism not only in ways of knowing and doing, but in economic and ethical ways as well. Take the process of producing news (Lewis & Westlund, 2014). First is necessary to observe an event, select and filter information and then comes the processing and editing. Nowadays, most newsrooms use machines and algorithms to do the first step. Editing and processing is done involving a computerized gatekeeping or "watchdogging in code" (2014: 7). This is what Crawford et al. call economic efficiency, as Big Data allows for "more observation at less cost" (2014: 1666). Powers (2012) believes it is important to see this economic disruption as catalysing labour to new types of news work, rather than simply displacing workers and putting them out of work. Still, as Mark Deuze (2008) pointed out, Big Data raises important questions about the relative status and precariousness of journalistic labour.

Another example of how the process of news production is being affected can be seen by how *The Los Angeles Times* has structured the creation of news articles regarding earthquakes (which are quite common in the west coast of the US). The LA Times has developed an algorithm to record earthquake notifications, process alerts into epistemological facts and facilitate easy editing and rapid publication (Lewis & Westlund, 2014). One can also look at *The Washington Post* and its Truth Teller prototype. Such prototype combines speech-to-text algorithms with databases to fact-check political speech in real-time (Lewis & Westlund, 2014). All these are cases of how Big Data approaches present new facets for understanding the epistemology of transforming raw information into journalistic truth.

Even though it can be a threat to the livelihood of some journalists, most researchers and academics consider the benefits of Big Data outweigh the drawbacks. Taylor (2010) for example believes that computational processes play a fundamental role in humanistic investigation and analysis through automation, algorithms and abstraction. Automation, as mentioned before, alleviates humans from data gathering, number crunching, analysis and filtering information. The algorithms help in identifying problems and finding solutions in immense sets of options, besides being useful in verifying whether certain information is reliable and consistent. Finally, abstraction enables for different and new perspectives on an idea and allows for new directions to be explored (Flew et al., 2011).

Big Data and other computational journalism techniques enables journalists to make sense of complex data much more quickly than before by revealing relationships and connections that would otherwise be missed by the sheer volume of data at hand, while at the same time producing datasets which can be easily verified (recall the Truth Teller protocol of *The Washington Post*) (Flew et al., 2011). There is no denying that software can scan databases to identify and report patterns, which can be used by journalists for story leads. As Flew et al. (2011) point out, data is not only numbers and statistical records, as it can come from a digitised format, including text, audio, photographic and video files. "The idea is that an investment in codes and algorithms can trawl data rapidly and effectively" (Flew et al., 2011: 3).

Impact in newsrooms and journalism's business model

As Martha L. Stone puts it: "the acts of analysing and making the data actionable are the new mantras for media companies" (2014: 3). The author argues that data, even though it can come in unstructured formats, has potential value in terms of high quality journalism and business insights and revenues.

Stone's research also focused on understanding what motivates companies to collect personal data, to which an inquiry revealed that thirty-six percent of the respondents said it was about stimulating greater loyalty and stickiness on digital platforms. 33 percent claimed data collection makes for more effective marketing strategies and optimal media spends. A quarter answered that it was about better predictive capabilities, while 24 percent said it was all about optimal pricing. In fact, Martha L. Stone found out that 53% of those surveyed confessed to use data collection for loyalty and affinity programmes (Stone, 2014).

Curiously, Stone's own report *Big Data for Media* mentions a survey conducted by the British newspaper *The Economist* which somewhat contradicts the author's conclusions. According to *The Economist*'s study, 70 percent of the executives surveyed claimed to look to Big Data analysis because of its predictive nature, in order to help them anticipate the market's needs and to help their company grow. The second and third most important advantages listed by the surveyed executives focused on trend data and analysis such as

sales trends (43%) and scenario-based data analysis such as data about the company's performance (41%).

An interesting conclusion drawn by *The Economist* is that companies embarking on Big Data strategies must ask better questions about their objectives in order to reach desirable outcomes. Asked about what is a desirable outcome, most executives listed the following:

- Make effective decision (59%)
- Avoid missed opportunities (44%)
- Keep up with competitive pressures (30%)
- Manage risk (30%)
- Control costs (25%)
- Empower employees (20%)

Most respondents said that the quality and reliability of the data is their top challenge when it comes to dealing with Big Data, followed by lack of effective systems to gather and analyse data and a lack of skills to interpret such data. Other difficulties found by *The Economist* when it comes to companies keen on investing in Big Data are a widespread misunderstanding of what the data can be used for, a concern over disclosure of confidential or sensitive corporate information and an inability to track the impact of data (Stone, 2014).

The Economist's study distinguished between strong performing companies and average/weak performing companies. The newspaper found that strong performing companies were much more likely to have mapped out their Big Data strategies compared to weaker performers. Furthermore, stronger performing companies were more likely to have incorporated insights from their Big Data analysis into their own business strategies (Stone, 2014).

The British newspaper's research also found out that the most popular plans for deriving insights from data include customer segmentation, external sources such as third-party data providers, surveys and census data, social media analysis and focus groups (Stone, 2014).

Going back to Stone's own survey, the author concluded that companies collect data from their customers through various ways, the most popular being tracking cookies (61% percent of inquired companies used this method). Another popular way is by monitoring how consumers navigate the companies' websites. This is known as "pathing" and 58% of companies claimed to rely on this approach. To collect the data, companies say the most useful tool is their own website as it is the portal through which most of the user's information flows. Call centres, cell-phones, social media, company cards and email are other important places where data is collected (Stone, 2014).

There is a discrepancy between the percentage of companies observing consumers' interaction and the amount of consumers who believe they are being watched by enterprises. Stone's research shows that 65 percent of business executives surveyed said they were likely or highly likely to observe website behaviours of consumers. However, 84% of the B2C (Business to Consumer type of commerce) and 53% of B2B (Business to Business type of commerce) perceive they were being observed while on a company's website. But perhaps the biggest disconnect lies in the percentage of consumers who believe their card transactions are being monitored and those companies that actually monitor it. 73% of B2C consumers say it is likely or very likely that their card transactions are being observed, while less than half of the executives surveyed admitted to observe their customers' transactions. Stone concludes that companies need to put on a better effort in disclosing if, when, where and how they are tracking customers (Stone, 2014).

Jimmy Maymann, HuffPost's (previously known as The Huffington Post, it consists on an American blog which aggregates news*)* CEO, shares some of Stone's vision on the importance of data. He has been quoted on saying that "it's all about data" (Stone, 2014: 4). It makes sense that Maymann said that, as HuffPost has been using Big Data – and small data – to improve the user experience, such as optimising content. HuffPost has algorithms which derive interest graphs in real-time to identify the user's level of interest. By looking at these graphs, someone can easily understand what is driving engagement. The new technology employed by HuffPost also informed the blog where individuals would go to access content, which helped the staff decide how to improve the user experience for the front page and other pages. Because of all this, Maymann happily reported that each user was accessing between ten to twelve articles per session, on average (Stone, 2014).

Big Data is also being employed by the North-American news aggregator to authenticate readers' comments. In 2013, HuffPost garnered 300 million readers and the high volume but sometimes low quality of the comments lead its CEO, Maymann, to seek a Big Data analysis on how to improve the comments' quality. The blog's analysis determined that the quality of the comments were higher when written by people from specific geographies and when they were not anonymous. As a consequence, HuffPost started requiring commentator registration and disallowing anonymous posting (Stone, 2014).

BuzzFeed is another site of news and entertainment that has seen major improvements after adopting Big Data practices. BuzzFeed's popularity comes from publishing viral content on its platforms. To help with that, BuzzFeed has a data science team which identifies trending stories and their unique characteristics. For example: photos of food are popular, as are photos with guns and the colour red. The site's chief data scientist says that when trying to identify characteristics with predictive relationship to "virality", they focus on quantitative and descriptive characteristics. BuzzFeed's machine learning algorithms predict social hits as they are capable of knowing what is viral before it even takes off, claims BuzzFeed's chief data scientist (Stone, 2014).

The key takeaways, says Ky Harlin, director of BuzzFeed's data science department, are that data can be used to optimise content sharing.

> Before publishing, data can help determine what to write about and how to present it. After publishing, data can help optimise the article's promotion as well as the article itself (Stone, 2014: 7).

"Big Data is an important tool for understanding what content and how to serve that content" to users, said the co-founder of Circa, an app that specializes in aggregating news for mobile users (Stone, 2014: 7). It was through the use of Big Data that Circa's executives reached the conclusion that for their business model it doesn't make sense to break the news, to try and give the news first. Circa's executives argue that "there's an escalating arms race for urgency and attention" (Stone, 2014: 7) which incentives news outlets to overuse the breaking news alerts and consequently disappoint readers and viewers.

The *Financial Times* also credits its success in the recent years to Big Data (Stone, 2014). The newspaper has a *paywall*, which means a reader cannot have access to its articles online unless it pays for it (*Financial Times* currently employs a type of *soft paywall,* where readers can have free access up to a limited amount, from which point on they have to pay). Andrea Carson talks about the benefits of *paywalls* in her 2015 paper *Behind the newspaper paywall – lessons in charging for online content: a comparative analysis of why Australian newspapers are stuck in the purgatorial space between digital and print*. According to Carson, paywalls provide newsrooms with a digital tool through which readers' behaviours can be tracked and later studied.

> Five or six years ago we started a new media model, charging for access through a metered system. When we started doing that, it was primarily to build a revenue stream online, but probably what was more important over time was the data and customer insight that that gave us. That's what transformed the business (*Financial Times* CEO John Ridding as cited in Stone, 2014: 8).

The collection of data through the *paywall* and its analysis through Big Data analytics helped the *Financial Times* grow its digital subscriptions by allowing a better understanding of the customer. Big Data techniques were crucial for improving the newspaper's communication with customers and creating personalized content. The revolution in how the *Financial Times* dealt with data was also pivotal due to the increasing number of readers who accessed articles via mobile channels such as smartphones or tablets. It was the monitoring of consumers' behaviour that led the newspaper's analysts to the conclusion that weekend content is more consumed on a smartphone or a tablet, while management or finance content is more consumed on the desktop in the office during weekdays (Stone, 2014).

Such a multichannel view matters a lot to the company as it provides a unique context on the customer's needs. Understanding customer needs is an important step in product placement: which content must be targeted at which type of customer? The head of customer analytics and research for the *Financial Times* states that the segmentation is nothing more than engagement. Not only does this have implications when it comes to content creation and marketing, but also in publishing patterns, as nowadays there is a

discrepancy between what is published and what is consumed. Therefore, a newspaper or a blog must take into consideration publishing hours. It is crucial that everything is optimised to make the most out of every opportunity. At the end of Stone's paragraphs about the *Financial Times* usage of Big Data, there are recommendations made by the newspaper to other media companies planning on using such analytics. One of them, is that they should obsess about data capture (Stone, 2014).

"If you don't engage with your best customers, they won't want to engage with you. Every decision either grows or erodes loyalty", said the head of data analytics of dunnhumby, a data science company. "See what motivates [each customer]. Build out DNA for each customer. This makes your business a better business" (Stone, 2014: 9).

This approach is also favoured by BBC and CNN. In the case of BBC, the news company has also been very focused on using Big Data as a support of its visual storytelling. BBC believes that significant visual stories, with personal relevance to users, is the best way to engage with audiences. As for CNN, the company's Big Data tools are very much centred at an early warning system for breaking news, monitoring users' behaviours and at analysing data sets to create journalistic content. For their data journalism strategy, CNN acknowledges that what works online may not be the case with TV. As it also happens with BBC, CNN realizes that data allows journalists to show something online rather than telling it. As an example, you can take CNN's coverage of the World Economic Forum Annual Meeting of Davos in 2014. CNN created a series of maps showing key panels discussed at the summit. On a different occasion, on a story about U.S. war casualties, CNN created a satellite map populated by data showing where concentrations of servicemen were from (Stone, 2014).

The final example comes from the *Sacramento Bee*. The American newspaper, concentrates the efforts of its Big Data department in figuring out the reader. The *Bee* is currently mapping where subscribers (and former subscribers) live. Audience analytics is therefore crucial in understanding what, when and how long readers engage with stories. This type of information is later used in the business decision-making process. It was the data gathering and analysis that led the Sacramento Bee to develop a path-to-purchase visualization tool which helps advertisers target demographic interest groups (Stone, 2014).

In this new data era, the head researcher at *Sacramento Bee* recommends that media companies inform their subscribers of what is being done with their data. Most customers what to know what is being done with their personal information and, as the CEO of Acxiom (a database marketing company) said, seventy-five percent of consumers are actually happy to exchange their data for added value. As long as readers are well informed they are happy. In return, the *Bee* wants its readers to visit its website frequently, to share the newspaper's content and to engage with the newspaper and its writers. To achieve their objectives, the organisations has changed to accommodate programmers, data miners and research analysts (Stone, 2014).

1.2.3 Problems that arise from the fusion of Big Data and Journalism

Rebekah McBride wrote in her 2016 paper that "when discussing the ethics of data-based journalism, it's also important to discuss the role Big Data (…) will have on the future of data journalism" (pp22). The author was concerned with something more than the mere technological revolution and its practical side. McBride wanted to know how the phenomena of Big Data and its influence in newsrooms would impact journalism's professional norms, routines and ethics. This apprehension was shared by other authors such as Seth Lewis. Lewis wondered how data's influence in journalism would not only change the organizations (newsrooms) but the institutions of society and the economy (2015). In his 2015 academic work, Lewis also expressed preoccupation as to how data could undermine journalism's role in society: "What are the implications (…) for its authority and expertise? And for the epistemology that undergirds journalism's role as knowledge-producer and sense-maker in society?" (pp1).

Lewis believed that there was not enough research on how the world was changing because of journalism's permeability to Big Data. Even though scholars were increasingly devoting more attention towards the fusion of both fields, the outcome of such studies was too much focused on the nature of computation and quantification in news (2015). C. W. Anderson had a similar view. The author felt that it was needed a more academic look into the sociological approach of data journalism (2012). Anderson pointed out that one should question the extent of the technological transformation. Like McBride, Anderson was worried about how data and technology could be the downfall of journalism despite also having the potential to save it (2015; 2016).

Both authors agreed on the lack of understanding on data journalism and how it would all play out in the future. Still, as McBride wrote, "addressing the lack of education as well as formulating clear ethical codes will be paramount if data journalism is to become a successful part of the news industry" (2016: 18).

Even though Anderson's work poses many questions, it fails to answer them. It seems Anderson was focused in calling out people's attention to the problems, rather than solving them (2015 *apud* McBride, 2016). The critic raised by McBride to Anderson's contributions to the debate can actually be applied to many authors who preferred

mentioning the challenges faced by a 21st technological century, rather than trying to solve them, perhaps because it is still early in Big Data's revolution and most academics are still trying to grasp what it means and its implications. Lewis and Westlund, for example, used Crawford et al.'s take on Big Data's ethical quandaries about user privacy, information security and data manipulation to say that the phenomenon deserves scrutiny and reflection as journalists figure out how to navigate the algorithmic innovations (2014; 2014).

Martha L. Stone was one of the first academics to actually try to gather information that could translate into numbers the privacy ethical concern. In her 2014 report, the author conducted a survey and found out that privacy was also a preoccupation of business executives. 85 percent of those inquired either agreed or strongly agreed that breaches of customer data security could do great harm to customer relationships. Stone also found a correlation between the size of the company and its worries with privacy concerns, as the larger the company the more risk was perceived to exist to the company's reputation in case of a privacy breach (2014):

> While Big Data audience analytics data collection presents an enormous opportunity for companies to better understand their audiences and consumers, the collection of consumer information can be problematic if not handled properly. Every country has different regulations applying to data collection, and the use of these data to content targeting and advertising. Corporate policy making regarding the use of these data will be key as Big Data strategies progress among media companies (Stone, 2014: 19).

Stone's also interviewed media company consumers and found out that most of them do not take measures to protect their privacy. One-third of the respondents said they rarely or never delete cookies and 31% said they rarely or never use ad-blocker software. However, consumers are in fact reluctant in providing companies with too much personal information as they frequently choose not to provide details when asked about it (for example, when creating a user profile) (Stone, 2014).

Interestingly though, Tom Evens and Kirstin Van Damme found out that people's willingness to provide personal data increases in exchange for news (2016 *apud* Napoli, 2016) (something which has already been explored in a previous chapter).

Lewis and Westlund (2014) also observed that the Big Data phenomenon posed important questions as to the legitimation of claims about knowledge and truth (epistemology) as well as new developments that could affect the norms and values guiding human decisions (ethics). In their 2014 paper, the authors wrote that technological advancements opened new ways for imagining how journalistic investigations develop epistemologically relevant revelations.

The idea of journalism reinventing itself through technology is not original to Lewis and Westlund's work. In 2013, Lewis and Usher suggested that one way to do it was by making data sets and programming code tools available to public examination. In the authors' view, the way forward for journalism was done by integrating norms such as iteration, tinkering, transparency and participation (Lewis & Usher, 2013). In Mark Coddington's (2015) perspective, data and computational journalism have arisen from the intersection of professional journalism with open-source culture.

However, some believe that transparency has only become increasingly difficult. Diakopoulos (2014b) argues that algorithmic transparency faces endemic obstacles to computational work. Such transparency involves additional labour costs for creating and making sense of an algorithm for public consumption (Diakopoulos, 2014b). Stavelin (2014) has a similar belief. He claims software to be opaque by nature and thus any transparency in computational journalism actually borrows from professional journalistic values, rather than coming from its own native framework (Stavelin, 2014).

Another aspect troubling scholars, and which has somewhat been referenced in the first chapters about Big Data, is the conceptual implications of digital datasets so large that challenge how we think about the nature of mediated communication (Driscoll & Walker, 2014). That is why a field once dominated by professionals versed in writing and editing has shifted to include nowadays software developers and computer scientists (McBride, 2016). As McBride points out, the data rush has put huge amounts of data at journalists' disposition, but accessing it and interpreting it require technical skills not usually associated with journalism. If done correctly, the author adds, it can lead to "ground-breaking stories that open the door for an innovative era in journalism" (pp13).

That is why some specialists like IBM advise companies to put into practice a data governance plan. These data governance plans are comprised of technical and security procedures which are the foundation of the Big Data strategy (Stone, 2014). In 2013, InformationWeek (an American weekly magazine mostly about lifestyle) stated that one of the biggest barriers to the successful use of Big Data was the lack of Big Data management tools. According to the magazine's survey, 40% of the respondents called their data analysis practices limited or "abacus-like" (Stone, 2014).

As the number of companies gathering data for audience analytics keeps increasing, it is important to bear in mind the regulations specific to each country. The collection of consumer information must always abide to every rule and ethical principle as it can be very problematic if not handled properly. That is why Sone argues that corporate policy making regarding the collection of data will be key as more and more companies invest in Big Data strategies (Stone, 2014). "Despite the difficulty, it is important to remember that this step serves as the backbone to the story, so understanding the technical needs and ethical challenges is key to creating a quality end product" (McBride, 2016: 4).

This, in turn, raises another important question: should journalists be accountable for the collection, analysis and presentation of the data? (McBride, 2016). Accountability is a part of many ethical codes, so much in fact, that the Society of Professional Journalists defends that accountability is at the "heart of journalism" (2014 *apud* McBride, 2016: 5). The Society of Professional Journalists adds that a "willingness to be accountable is what cements a journalist's credibility" (2014 *apud* McBride, 2016: 5).

McBride goes on to cite the opinion of Erik Litke, a reporter at the *USA Today Network*. According to an interview with this reporter, a journalist is only accountable for asking the right questions, but not for falsified data. By asking the right questions, Litke defends, a journalist should be able to question the nature of the data in order to understand whether the numbers and the information are wrong. Therefore, a journalist "must not just investigate with data, but investigate the data we are using" (McBride, 2016: 5).

The point is that there are many examples of useful data journalism methods and tools used in stories, but these stories must be backed by powerful ethical standards (McBride, 2016). *The New York Times* coverage of the Panama Papers is a good example of data

ethics at work. After the International Consortium of Investigative Journalists started releasing the story, *The New York Times*, which did not belong to the consortium, waited to verify the documents before publishing a piece of their own (McBride, 2016).

One final concern comes from Philip Napoli. The author is worried about the "dictating" power of Big Data, in the sense that Big Data is increasingly telling media consumers how to navigate their media environment, while also increasingly dictating content production decisions (2014a, 2014b). This can betray the purpose of journalism, which is to inform, as it is possible that, one day, newsrooms and media companies invest only in content production that they know is going to get the most views (Napoli, 2016a).

2. Methodology

The careful review of the literature identified some shortcomings in the existing research, such as a lack of a sociological approach to Data journalism (Anderson, 2015) and contradictions between theoretical and practical studies (take for example how Stroud et al. (2016) found little evidence that newsrooms were adopting interactive features, when other studies indicated that many media corporations claimed to be pursuing Big Data precisely because it allows for more interactivity with audiences).

It is important to point out that, as it was repeated many times in the literature review chapter, most authors claim that Big Data is still in its early stages and the fusion with journalism something even fresher (Lewis & Westlund, 2014). This means that one should be cautious when reading studies on Big Data's role in journalism as it can be difficult to draw valid scientific conclusions from a field which is brand new and, due to its beginning stages, constantly changing. That is one of the reasons why this dissertation is important. This study will look into an environment characterized by an ever-changing technology and try to assert how it is today (2020). Take for example Moore's law, which states that the number of transistors in a microchip doubles every two years, though the cost of computers is halved, which can roughly be interpreted as the speed and capability of computers doubling every couple of years; nowadays Moore's Law doubling period is considered as approximately every 18 months (Golio, 2015).

There's also a lack of studies on how Big Data is affecting the advertisement business of media corporations. Despite many authors having looked into how Big Data is impacting newsrooms' business model, the publicity side remains a blind spot. In fact, preliminary expert interviews conducted among those with deep insights into the topic supported that advertisement is one area where Big Data can substantially lead to improvements.

There is as well insufficient literature on the way journalism is adapting to Big Data in Europe, particularly in Portugal.

Building on all this, this dissertation examined Big Data's role in journalism by analysing how the Portuguese media corporations are adapting to the phenomenon. To do it, the

present study examined the NÓNIO platform: a Big Data tool developed by five Portuguese private media companies.

2.1 Methodological Approach

<u>Research question:</u> How is Big Data changing the activity and business model of Portuguese media corporations?

a) Is it changing the production of news? Does it have an impact on what journalists write about?

b) How is it affecting the revenue channels?

This set of questions builds the basis for the present dissertation. The main question started out as "how is Big Data influencing journalism?", but after the literature review, it became clear that many authors had tried to answer it already. Besides, it was too generic, and a proper analysis oriented by such question would require examining changes in journalism in a lot of countries, which would significantly increase the complexity of this Master's dissertation. Due to time constraints, it was important to narrow the scope of the study.

The research question presented above allows to understand how Big Data is being adapted to journalism and how journalism is adapting to Big Data, by looking at the Portuguese media. It was decided to focus on the Portuguese situation due to a lack of investigation and because it is a matter of interest for Portuguese news corporations. Portuguese newsrooms have been struggling with economic viability threaten by the abundance of online news. It is vital that the Portuguese media find a way to improve their financial situation and relevance in the market. Thus, this dissertation aims at analysing how technological breakthroughs are being put into place to achieve those goals.

It was decided to examine both the business model and the activity of journalism because it is hard to separate them nowadays. It is important to bear in mind that journalism is facing many economic difficulties in the 21st century and those financial problems are responsible for shaping the organization of newsrooms. For that reason, one cannot properly study how the journalistic activity is evolving without taking into consideration the economic powers responsible for those changes. Questioning the impact of Big Data on the revenue channels of media companies is also an interesting way of challenging C.

W. Anderson's view that Big Data has the potential to save journalism from its precarious economic situation (2015).

Furthermore, this study aims at assessing whether Matt Carlson's claim is true. The author observed in his 2014 paper "The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority" that Big Data was affecting the content of news. This concern was backed by authors Mary Lynn Young and Alfred Hermida (2015). If Big Data is indeed having an impact on news production, how is this impact manifested? To what extent is news production permeable to Big Data algorithms?

That is why it is necessary to "interrogate the algorithm", as Nicholas Diakopoulos put it in his 2014(a) study. Diakopoulos explained that only by examining the algorithm can one understand the power structures, biases and influences of computational journalism in society. This dissertation also looked at the construction of the Big Data algorithms used by Portuguese media companies.

In order to answer the research question, this dissertation explored the NÓNIO platform, which consists on the object of study.

NÓNIO comes from an association between five of the largest Portuguese private media companies: Cofina, Global Media, Grupo Renascença Multimédia, Impresa and Media Capital (the newspaper PÚBLICO, owned by group Sonae, was initially part of the NÓNIO initiative, but it quit in December 2019). Its official website describes it as a "tool for audiences' segmentation". It collects Data from users who register in the technological platform[1]. In exchange, users get free access to content from the five media groups listed above, by being able to freely navigate each media group's websites. Registering on the platform also allows the media groups to offer personalized content to each user, which is advertised on NÓNIO's website as one of the platform's advantages.

The need to create NÓNIO comes from the financial difficulties faced by the media. As it is written on the platform's website: "Serious and independent journalism and national

[1] At the time this dissertation was written, NÓNIO had between 1,5 and 2 million users, as reported by João Paulo Luz on 22nd May 2020

entertainment has costs. All media groups (national and international) face a big challenge in terms of sustainability (…). By registering on NÓNIO, you will be contributing to increase the competitiveness of the national media (…)".

NÓNIO works by providing access to most content offered by the five media groups listed above through one single user and registration. It is important to point out that NÓNIO is not a platform that aggregates all the content from these five media groups in one place. For the consumer, what it does is grant free access to most content from the media groups' websites through one unique[2] username and password. The registration is free and it asks users for an email address, name, gender, date of birth and to tick two boxes: one is the Terms & Conditions and the other has to do with granting permission to collect and analyse users' personal Data. The project is funded by Google's innovative fund called Digital News Initiative (DNI).

[2] This is known as an SSO: Single Sign On

2.2 Methods

Given that the study aims at understanding how Big Data is affecting the media sector, qualitative methods are more suitable than quantitative ones (Corbin & Strauss, 2014). As Lune and Berg explain (2017), a qualitative approach allows to detect a tendency towards certain behaviours and under specific circumstances. It focuses much more on the how and the why rather than the what. That is why the research question starts with a "how" instead of asking "what changes has the use of Big Data led to in newsrooms?"

Even though a qualitative study makes it hard to draw generalizations, it is possible to detect certain patterns. Besides, this study examines how five of the six major private media groups in Portugal are implementing Big Data practices. Nevertheless, it is always important to remind that this study deals with "patterns, not laws" (Lune & Berg, 2017: 13).

Therefore, this qualitative study was conducted through interviews. A conversation with open-ended questions allows for the interviewee to fully express his/her perceptions on the matter. The interviews were with the person responsible for the NÓNIO project in each of the five media groups listed above. Talking with the leading experts in the pioneer field of Big Data and journalism seemed like the best way to understand how technology and social sciences are merging and to figure out its implications for the media.

Preliminary interviews and the design of data collection instruments

As mentioned above, before conducting those interviews, it was decided to have preliminary conversations with field experts. Due to insufficient information regarding the use of Big Data by Portuguese newsrooms, such informal talks were perceived as being a good entry point to this master thesis.

It was during these preliminary interviews with field exports that it was understood that the NÓNIO project is in the beginning stages, meaning that one cannot apprehend the changing behaviours and tendencies of newsrooms through a quantitative research.

The first person to help with this investigation was Professor Nuno Conde. Nuno Conde is a professor at Universidade Católica (which means it was easy to approach him, as I am also a student at Universidade Católica) and a Chief Legal Officer at Impresa, one of the five media groups involved with the NÓNIO project. Even though Nuno Conde is not coordinating the project, he has been directly involved with it in many occasions due to the nature of his job. In his view, the NÓNIO was the beginning of "datification", a term used to describe the process of collecting and analysing data, in the Portuguese media. The project is an association of five different media groups because only then would there be enough scale to examine the data and draw valuable conclusions from it.

Not only did Nuno Conde referred me to the best person within group Impresa to talk to about NÓNIO, but he also gave me hints about possible questions to ask in future interviews, such as what each individual involved in the NÓNIO project perceives to be the datification of media and why it is important.

Nuno Conde also called out my attention to the fact that the market for data analysis is very much concentrated, as most of it is conducted by technological giants such as Apple or Google. Due to the penetration of its devices and platforms, they end up tracing every internet user's movements, says João Paulo Luz, Commercial Director of the Digital Department at Impresa, and the second person interviewed for these research. He says that is why Google invests so much in Chrome, which is free for users, so that it can have access to their navigation and their digital ID. That is also one of the reasons why NÓNIO was created: to know how users navigate the media companies websites.

It is ironic that Google is financing the project, considering NÓNIO can be seen as the development of tools for the Portuguese media to compete with companies such as Google. When asked about this, both Nuno Conde and João Paulo Luz replied that it is a publicity stunt by Google. At the same time as the company founded by Sergey Brin and Larry Page gets to know the ideas and movements of its rivals - by financing the project, Google gets to know how it works - it also projects this "good Samaritan image" of a patron willing to help companies in need. As Nuno Conde pointed out "There is a curious paradox here. The same Google that has a dominant position in the market and is responsible for some asphyxia of the media, due to its monopolistic position, is also responsible for creating a funding system, the DNI (Digital News Innitiative), designed

to benefit the media". In João Paulo Luz's words: "Google is a very competent company with very bright people (…) and at the same time as it gets more and more of our value chain, it launches initiatives in which it pretends to be our friend and that it wants to help us. It is actually helping us and taking our money all at once".

João Paulo Luz observed that the Silicon Valley companies have disrupted the traditional business model of media companies. "Traditionally, those who produced content were the ones dominating the industry. He who produced the content was the owner of what people wanted and that would attract publicity. What changed is that digital distribution is not as "agnostic" as we wished internet was. Internet is dominated, at least in our side of the world, by Californian companies from Silicon Valley (…). What do I mean with this? Most information brands don't reach their readers directly. They reach their customers through platforms such as Google or Facebook, which means that we, in the media business, are very much conditioned by these intermediaries."

According to the Commercial Director of the Digital Department at Impresa, such disruption comes with a loss of publicity for the media business and the NÓNIO project is also an attempt at defining readers profiles so that it can later be sold to advertisers: "Advertisers want to know who's looking to buy a car or a house (…) this means that the advertising business no longer lies with the content producer but with whoever possesses the data".

In João Paulo Luz's opinion, the NÓNIO will help media companies in capturing more advertising revenue, to improve subscriptions and the relationship with customers. João Paulo Luz believes that the datification of users will improve companies' content supply so that it matches users' interests. But how exactly will this happen? And is this the reason why the other four media groups decided to join the NÓNIO project?

To answer these questions and to have a more general overview, Nuno Conde suggested investigating the commercial practice associated to tracing someone's profile. In Nuno Conde's view, such a question would immediately allow to understand whether or not NÓNIO is helping to the economic survival of the media, which is directly linked to this thesis' research question.

The preliminary interview with João Paulo Luz shed light on possible questions for future interviews but it partly answered the investigation question as well. He confirmed that Big Data is changing the business model of media corporations at least in the way they deal with advertising.

After the input from the preliminary interviews, an interview guide to conduct conversations with the five people in charge of the NÓNIO project (one for each of the five media groups that are part of the initiative) was drawn.

All questions were open-ended to encourage the interviewee to fully express his or her own view on the topic without being pressured to answer in any specific way. The interviews relied on the participants' professional experience, their perceptions and beliefs on Big Data and their thoughts on the Big Data-journalism relationship.

The first question interviewees were faced with was what they perceived to be the datification of the media and why it matters nowadays. This question was suggested by Professor Nuno Conde as a good entry point to the interview by establishing the topic of the conversation. It was also deemed as a good way to get a definition from each expert on the fusion of Big Data and media. One thing that became evident from the literature review is that this is a brand new area and there are different definitions and understandings of what it is and how it works.

The second question was a simple one and it served to make the interviewee feel comfortable: what is NÓNIO? Not only does this ease the interviewee in the conversation, but it also gives a more practical insight into the expert's view on the connection between Big Data and NÓNIO.

The third question was a follow-up of the second question: why did corporation X decide to be part of the NÓNIO initiative? The X was used here as a variable that could take the form of any five media groups involved with the project. The question was tailor-made to each interviewee, as each one represents one of the five media corporations. The goal here was to see what each interviewee considered to be the added value of being part of the NÓNIO. This question ends up being similar to the first one as it explores why Big Data, in this case Big Data techniques used through the NÓNIO platform, is relevant to

media companies. It also gives insight on how these five newsrooms are employing Big Data techniques.

The fourth question was also a follow-up of the previous one: what type of data treatment is done with NÓNIO? This question intended to find out what type of profiles are traced with NÓNIO and what other information is being collected. This is a very important question as it is hard to understand the impact of Big Data in the journalism industry without properly comprehending how newsrooms are employing Big Data.

Interviewees were then asked to explain the commercial practice associated with NÓNIO. This is once again a very important question, as it is the best way to understand how the information gathered with NÓNIO translates into more revenue for the media. It was expected that respondents focused on four crucial aspects for this investigation:

a) The impact on publicity
b) The impact on the type of news being produced
c) The impact on the type of news readers have access to via email or when they first enter a media group's website
d) The impact on the way new and current customers are attracted and retained (changes to the digital subscription model)

The sixth question confronted those in charge of the project with whether or not they believe NÓNIO and Big Data practices can lead to the economic survival of the media. This is a question of short answer intended to understand how optimistic (or pessimistic) the industry is about the technological disruption, never forgetting that it was also a technological disruption that led to the financial difficulties of the media.

Next came a question that arose from the lack of information regarding the use of Big Data in Portuguese media: does corporation X have other types of Big Data practices besides NÓNIO? The X was once again a variable which could take the form of any of the five media groups involved with the project (the question was tailor-made to each interviewee, as each one represents one of the five media corporations). This was an interesting question as it allowed to go beyond NÓNIO and to look into Big Data practices in newsrooms besides NÓNIO. It also served as a backup in case the NÓNIO project

wasn't as fruitful as expected. Thus, this seventh question was a plan B to gather insights on the role of Big Data in journalism.

Then came three questions meant to explore the ethics side of employing Big Data techniques. The eighth question was directly linked to Martha L. Stone's 2014 work, in which the author cited the CEO of Acxiom saying that seventy-five percent of consumers are happy to exchange their data for added value. Therefore, it was thought that it could be interesting to see if the people responsible for the NÓNIO project shared this view.

The ninth question of the interview explored possible ethical challenges posed by NÓNIO. As Lewis and Westlund (2015b) argued, Big Data is not only changing journalism in ways of knowing and doing, but in economic and ethical ways as well.

Still regarding ethical concerns, the tenth question was inspired by Diakopoulos (2014a) remark that it is important to interrogate the algorithm. Like Diakopoulos, De Maeyer et al. (2015) were also focused on the notion of algorithmic accountability. As De Maeyer and colleagues expressed, one should be cautious about the path of experimentation as algorithms are the end of human creations. Based on this, interviewees were asked to express their thoughts on the possibility of human biases manifesting themselves on NÓNIO's algorithms.

The last question was whether NÓNIO is a first step in news automation. The goal here was to understand whether news media group X is considering developing algorithms that can generate news automatically like some of the examples mentioned in the literature review (e.g. the algorithms used by *The Los Angeles Times* to edit news regarding earthquakes). If the Portuguese media invests in the so called "robot reporters", the composition and day-to-day life of newsrooms will be affected, as that means some journalists will be replaced by "robots" and can spend time and energy doing something else. This question also hints at possible future investigations into the relation between Big Data and the use of artificial intelligence in journalism.

All these questions were designed to provide answers to the dissertation' research question: How is Big Data changing the activity and business model of Portuguese media corporations?

In order to schedule the interviews, all respondents were contacted via email. Because of the coronavirus pandemic and due to Portugal being in lockdown, the interviews were done via computer (either by Skype or Zoom). Even though it was a disadvantage not to be face to face when the conversations took place, this was the best way to avoid major delays with the investigation, as at the point in which the interviews took place it was still unclear when restrictions would be lifted and "normality" resumed.

Respondents were not previously informed of the questions. Prior to each interview they were only informed that conversations were about the NÓNIO project. Considering these interviews dealt with the respondents' professional activities, it was assumed that respondents would be fairly at ease with the topic.

All interviews were conducted in Portuguese and through video-conference (because of the coronavirus pandemic). Interviews were of voluntary participation and respondents were able to decline answering certain questions or to opt out at any point of the interview.

Each interview took around 30 min, with the exception of Marcelo Leite's, the Digital and Big Data Director at Global Media, who didn't have that much time to spare. Because of that, not all questions mentioned above were asked in detriment of those considered more vital to this investigation.

Interviews were recorded and later transcribed (see Appendices section) in order for the conversations to be analysed.

3. Findings

This section presents the discourse analysis of the interviews. The findings are presented following each question asked and in a full text composed of the most relevant answers provided by each respondent.

To facilitate reading and avoid lengthy repetitions, respondents (listed below in alphabetical order) are identified by their initials:
JF – José Frade, Digital Business Director at Cofina
JPL – João Paulo Luz, Commercial Director of the Digital Department at Impresa
ML – Marcelo Leite, Digital and Big Data Director at Global Media
MM – Mário Matos, Commercial and Marketing Director at Media Capital
TC – Tiago Cruz, Digital Advertisement Sales Coordinator at Renascença

Advertisement segmentation

Regarding the datification of the media, ML credited the development of new analytical tools as responsible for providing better objective audience measures. Meaning that the datification of the media is comprised of tools that give a more efficient and detailed feedback on how users consume media content, which is similar to Kosterich and Napoli's conclusions (2015). "Datification comes from the fact that the internet is a highly measurable medium", stated ML, adding that it is also about the need to quantify audiences. This is aligned with Sandra González-Bailón 2013's paper, where the author states that digital technologies keep track of everything people do while online. Our lives, she adds, have become "quantified to unprecedented levels" (2013: 147).

A word constantly used by respondents when addressing the matter of media datification was "relevance". In the experts' view, datification allows for the media to be more relevant for consumers. This is done through customization of content for each user. Datification allows for tracing users' profiles, as JF explained, which leads to a better understanding of consumers' needs and interests. Datification is a way for the media to "reach the right people at the right moment" (TC), which comes down to providing specific content to specific consumers. "We will witness the construction of certain

products to certain targets", said JF. This segmentation matters not only in terms of content production and consumption, but also to the advertisement business which is many times conducted through media platforms. Overall, all respondents said that advertisers are not willing to work in the same way as in the past, where ads would consist of buying some television time or a newspaper page that would later be seen by everyone watching television at that time or reading that newspaper. MM observed that advertisers want less waste: they want the ad campaign to be only seen by those who have the means or the desire to buy the advertised product (this topic will be covered with more depth in upcoming paragraphs). Therefore, the datification of the media also comes from the need to please advertisers, who are responsible for large chunks of media companies' revenues.

"Either things like NÓNIO are done by European publishers or it is the end. Everything [the advertising market] will be in the hands of companies like Google or Facebook", says ML, when questioned as to why Global Media decided to be part of the project. "Most of our revenue comes from selling publicity and the market for digital publicity has been monopolised in the last few years by Facebook and Google", states MM, when discussing the lack of revenue the media sector has faced in the last years and why NÓNIO has the potential to invert the scenario. "We need these data driven solutions (…) to compete with the [technological] world giants", adds TC.

NÓNIO is also about gaining independence from such Silicon Valley companies, says MM. Due to the monopolistic position of Google in the advertising online market, Portuguese publishers (and most other international publishers) are forced to sell online publicity[3] through Google. JPL explains that Google has built two platforms that match those wanting to sell online publicity and those willing to buy it: there is one for the demand side, DSP – Demand Side Platform, and another for the supply side, SSP – Supply Side Platform. Publishers use this Supply Side Platform to manage their incoming publicity and to sell their ad slots. Advertisers then use the DSP to buy publicity slots in publishers. Google then matches demand and supply (it matches the requests in each platform) while receiving a commission, which, according to the Digital Director at Impresa, is too much and deprives the media sector of critical revenue to its economic

[3] In the case of the Portuguese publishers, selling online publicity is the act of selling publicity slots to companies and advertisers. This can take the form, for example, of space on an online article where a company can advertise its product.

survival. Google is only able to do this because of its Big Data tools and the way its products are consumed worldwide, giving it a unique advantage in knowing who the potential consumers of the advertised product are and how to reach them. By employing Big Data tools, Portuguese publishers don't need to rely on Google's analytics.

> "We decided to gain independence in the distribution and selling of publicity, by creating a content market that allows to gather data and information regarding users' preferences, through one single login with NÓNIO", MM.

Recalling on Martha L. Stone's (2014) research on what motivates companies to collect personal data, 33 percent of respondents claimed it led to more effective marketing strategies and optimal media spends. In the case of the present investigation, all five respondents pointed to this being one of the added values of NÓNIO.

Another advantage of being part of the NÓNIO is scale. The five interviewees believe the market quota of each individual publisher they work for is insufficient to provide data sets large enough to reach valuable conclusions, however, the same thing cannot be said if all publishers share the information. The need to have large data sets is embedded in Big Data's definition, where the only way to extract meaningful information is by having a large pool of data (Chen et al, 2012; Lewis & Westlund, 2015a).

According to TC, NÓNIO asks users to register from the moment they want to read a second article on one of NÓNIO's websites (it is possible to read an article without being registered in the platform). NÓNIO then gathers users' information through two ways. The first one is known as declarative first party data. According to MM, it is data provided directly by the audience, such as age (birth date) and gender. The second one, is defined by JF as:

> "behavioural [first party] data, when the user visits a [NÓNIO's] website and sees a certain type of articles, content, news or entertainment. Data is being collected based on that behaviour, which is normally done through semantic algorithms capable of understanding what type of content is the user consuming. This then leads to the creation of interest profiles."

By analysing a user's navigation, Big Data algorithms classify that user as belonging to a particular segment and this is all first party data, declarative and behavioural, because it is information that the NÓNIO is able to collect through itself, without needing outside sources to gather that information.

The semantic algorithm mentioned by JF which allows for the classification of the users' behaviours is what JPL calls a TAG:

> "We put this TAG, this piece of code, into every [NÓNIO's] websites page. All pages have in its hardcode that TAG. (…) In a very simplified way, it is called a learning TAG. Why learning TAG? Because it collects every information it can. What can that TAG collect? All the information provided by the browser[4] and the browser provides lots of information - geo-localization of the network through the IP[5], it provides the type of device being used [to access the internet and to navigate the website], the type of internet access, the time, etc… - and then, those pieces of code also have the power to crawl[6] the text. (…) it consists of evaluating the tittle [of the article] and if the tittle has certain words belonging to a certain thematic, then, in theory, that text can be classified as one particular segment, for example sports or economics; secondly, it checks the lead of the text (…) and finally the text itself. (…) Obviously, one thing is for the tool [the algorithm in the piece of code] to do this in English, another is for it to do in European Portuguese."

The collected data is then used to create interest profiles. These interest profiles are what companies are looking for when wanting to advertise a certain product, in the sense that the advertising companies want their product being added and seen by the people belonging to those interest profiles.

Consider, for example, that a user is constantly reading articles about cars. Then he and all consumers with a similar behaviour are clustered into the segment of "Auto Lovers".

---

[4] A browser is a computer program used to navigate the World Wide Web (https://www.mozilla.org/en-US/firefox/browsers/what-is-a-browser/)

[5] IP is short for Internet Protocol and is a numerical label assigned to each device connected to the internet (the same way every house has one given postal code) (https://www.investopedia.com/terms/i/ip-address.asp)

[6] In the case of the NÓNIO, crawling is used to describe an algorithm which "scans" or "reads" a webpage. Web crawling is used among search engines such as Google to discover new webpages (https://support.google.com/webmasters/answer/70897?hl=en)

MM says there are 18 segments, which were defined by IAB – Interactive Advertising Bureau:

-Active Outdoor and Nature
-Animal Lovers
-Auto Lovers
-Business and Career
-Celebrities and Entertaining
-Children and Parenting
-Culture and Arts
-Food, Wine and Dining
-Gaming
-Health and Fitness
-Lifestyle
-Mommy
-Music Lovers
-Plants and Gardening
-Science and Discovery
-Sport Followers
-Technology and Gadgets
-Training and Exercise

MM explains that publishers then take these segments and sell them to companies wanting to advertise products. By "selling the segments" what publishers are doing is selling the people, meaning that publishers will make sure, through algorithms, that the advertised product will reach the people that fit into the desired segment. ML gives the example of BMW advertising a new car. For BMW, it doesn't make sense to pay for an advertisement in the newspaper Expresso (owned by Impresa) which is seen by elderly people above 80 years old. BMW wants less "waste", as described by MM, and is only interested in buying advertisement that is seen only by potential customers, so, when it buys publicity slots in Expresso online, it buys slots that will only appear to readers that fit the segment of Auto Lovers and maybe those fitting the segment of Business and Career (because of purchasing power).

The segments sold by publishers can be intertwined with other characteristics such as age, for example. It is possible for BMW to buy advertising slots with the segment of auto lovers for people between the age of 40 and 55. However, these types of combination are not available yet, as the respondents explained. JPL and ML point out that it is not economically viable yet: "It is up to us, the media, to find segments that are appealing to advertisers (…) vs what is viable for our own business". According to them, it is still early to sell too specific segments to advertisers, as otherwise advertisers will only demand such specific segments, which would ruin the business potential.

TC points out that such *modus operandi* is exactly what companies like Google or Facebook have been doing for a while: segmenting their users and selling the segments to advertisers. That is one of the reasons why, after checking an article about cars through Facebook's platform, ads for a Volkswagen car will appear on someone's wallpaper (also known as feed): Facebook identifies the user who looked at the car's article as someone who is potentially interested in buying a Volkswagen.

MM (and as JPL and Nuno Conde explained during the preliminary interviews) argues that a lot of how Google and Facebook determine which person belongs to which segment is done through media content. Media content consumers resort to Google and Facebook's platforms to access media content. When someone searches for information regarding a new topic on Google's search engine and then decides to read one of the articles Google presents, the company's algorithms "see" which article is read. It is why media companies have criticized Google, Facebook and other technological firms in the last few years, particularly in the European Union, for not redistributing their revenue among news corporations. During preliminary interviews, JPL and Nuno Conde pointed out that companies like the one founded by Sergey Brin and Larry Page use the articles written by media firms to attract customers to their search engine. In return, media corporations feel that Google is using them without giving them a monetary compensation, which is why Google has repeatedly faced lawsuits in the European Union and one of the main reasons for Google's launch of the DNI.

Editorial impact

It is therefore easy to see why publishers want to develop tools such as NÓNIO that help them gain independence from the Googles and Facebooks.

However, MM, believes publishers have an advantage over the Googles and Facebooks. The Commercial and Marketing Director at Media Capital reasons that those technological firms are only interested in exhibiting ads without paying attention to how or where the ads are shown. MM points out that YouTube (owned by Google) shows the same ad to someone whether that ad comes before a video about cats or before a video clip from a news by TVI24 (owned by the Portuguese publisher Media Capital):

> "I can show one advertisement to a particular audience with a certain type of profile, but there will be an added value to the ad being exhibited next to a brand such as TVI24. Because it is a brand with credibility, the brand adds value to the advertisement. What Google and Facebook have been saying is that content doesn't matter as long as the audience is reached. [Google and Facebook say that] if I reach John, whether through a blogger's content or through content on TVI24, the value is the same, as long as it reaches the audience. Unfortunately, there is a deficit worldwide, because publishers have not been able to show with analytical KPI's that there is indeed an added value in exhibiting publicity next to a premium brand vs exhibiting it next to a content which merely consists of someone with a camera recording a video for YouTube."

MM hopes that NÓNIO will lead to the creation of a Marketplace where premium publishers can sell advertisement directly to advertisers and companies, boycotting the need for Google or Facebook.

Regarding NÓNIO's impact content wise, respondents have mixed feelings. Both MM and TC believe NÓNIO can really influence what journalists write about. JF admits that, hypothetically, if NÓNIO finds out that readers are interested in Cristiano Ronaldo and chocolate, then it is possible for future articles to combine both those things. Even though, as JF points out, it would make more sense to write separate articles about chocolate and separate articles about Cristiano Ronaldo:

> "Data actually allows for many combinations. There can be people who like Cristiano Ronaldo but dislike chocolate. I must be able to make that distinction and to not give irrelevant content to users".

JPL thinks that NÓNIO itself will not be responsible for any editorial direction followed by journalists. He does, however, credit Big Data analytics as being responsible for that. When given the same example about chocolate and Cristiano Ronaldo, JPL answered that Impresa's existing analytical tools were already capable of finding out those users' preferences. JPL says that the analytics that accompany the publisher's articles already provided feedback on which articles were attracting the most readers. ML shares a similar opinion:

> "In the whole history of the media, no publisher survived without listening to its readers. In the history of journalism and the media there has always been some part of the production of content which (…) hasn't been demanded, but there is also another part where it is needed to listen to the people, to the consumers, to understand which content they want to see. (…) But that has nothing to do with NÓNIO. Journalists know, every day, which content was the most consumed from the previous day because there is Google analytics. In television there are ratings."

JF states there is another way Big Data can affect content production, even though it is not a direct result of NÓNIO's analytics:

> "(…) all support data that can enrich an article is an important area. (…) the data journalism you can create by inserting data in the articles, by linking many different sources and platforms which can contribute to better articles is an area where I foresee a bright future".

JF is talking about how data can be used for finding and telling better stories (Berret & Phillips, 2016).

Content personalization: changing the subscription model

Even though there seems to be little enthusiasm regarding how NÓNIO can affect content production, the conversation is radically different when mentioning content personalization (Koutsabasis et al., 2008), or, as TC puts it, "the way we arrange content [on the publishers' platforms] being different depending on who sees it".

All respondents were excited about this advantage that comes with NÓNIO, including TC who warns that content personalization based on NÓNIO's input is still at beginning stages and is lacking some maturity.

As previously explained, NÓNIO gathers behavioural Data regarding the users' navigation on the websites associated to the initiative as well as socio-demographic information. JF describes how that information helps the publishers supplying the most relevant content to each user based on his/her profile:

> "We can better segment our supply and our products (…) by better knowing who our users are. (…) If I have the profile of an executive who likes information about economics or the financial markets, I can sell him a subscription for *Jornal de Negócios* (owned by Cofina). (…) For a user who doesn't subscribe, I can also improve the relevance of the non-premium or non-paid content [seen by that user]. By better knowing the user's profile, I can improve the relevance of the articles supplied through newsletters, my marketing and other things"

The basic principle is that by knowing the user better, publishers are more equipped to supply the user's needs, which translates into him/her appreciating and valuing more the publisher's content, consequently visiting the publishers' platforms more often and, ultimately, subscribing and maintaining that subscription (a similar principle set in motion by Netflix, for example): "All we can do is keep recommending more relevant content", says MM.

Content recommendation and a better knowledge of the consumer are two ways NÓNIO is helping the publishers improve their subscription model and their ability to retain

readers. Another way NÓNIO helps at "capturing a subscriber", as TC says, is due to the fact that it asks users for a registration:

> "There is no subscription system in the world that doesn't require the identification of the person who is subscribing. In order to access the content they are paying for, people have to subscribe, have to login, and what NÓNIO does in a big way (…) is asking people to register, like Facebook, YouTube or Gmail require users to register. (…) So NÓNIO, through its Single Sign On, is nothing more than a first step in the process of selling content".

JPL shares a similar view:

> "(…) the first grand barrier to subscription is registration. For a while, *The New York Times* and *The Wall Street Journal* would require registration without asking for payment because it was a first step they were taking. Everyone knows that requiring registration and payment are two big walls. Having the registration already done is a tremendous advantage, because after someone had the impulse to register, the task becomes increasingly easier. And it also allows us to make [marketing] campaigns after those frequent users."

By registering, by creating a username and password that grants users a login, users are overcoming the first major step in subscribing to a media platform. From there on, respondents believe it will be easier to convince that registered user to start paying for a monthly/weekly/annual subscription.

When asked about this topic, JF stated that NÓNIO's contribution was indeed noticeable, however he also pointed out that publishers more or less already knew how to attract new subscribers, but that the problem is maintaining them:

> "I think that is the challenge faced by all subscription models and the challenge is much harder when talking about news platforms. Users who subscribe to entertainment services such as Netflix or Spotify tend to do it for longer periods. (…) I'll give an example related to this pandemic crisis that has been happening. Especially in the beginning of the crisis, users realized that trust-worthy news laid with the traditional publishers and we saw an increase in subscriptions. But right now [14th May, 2020] we are understanding that those people subscribed for a particular period of time where they thought it was going to be

> very important and rapidly cancelled their subscriptions. Clearly, the challenge of keeping those subscriptions is very big. (…) I think there is a big challenge for all publishers which consists on remaking the products, meaning that it can't only be the breaking-news product or the news of the moment. We have to invest in the premium offer. The paid offer has to have added value, a tangible value which makes a difference from the traditional information which is free."

Because of all this, ML and MM believe NÓNIO and Big Data practices can lead to the economic survival of media companies. TC, JF and JPL are less enthusiastic about the future, but they all agree that it is important to do something and that Big Data is probably the way to do it. JF argues that "it is not possible to survive in the future without data", but that doesn't necessarily mean that Big Data practices imply the economic survival of the media.

JF's belief in a future data-world is accompanied by the company he works for: Cofina. JF was one of the few respondents to admit working in a company that employs other Big Data practices besides NÓNIO, despite not wanting to expand much on the subject. JF also implied that all other publishers participating in the NÓNIO initiative employ some level of Big Data analytics apart from what is attached to NÓNIO: "Each publisher also has its own tools to analyse all the data that it generates individually. (…) there is a lot of individual data that goes beyond NÓNIO".

JF is contradicted by TC, who replies that group Renascença is still in early stages when it comes to employing Big Data techniques: "(…) we are the most recent NÓNIO members and we have had a digital operation for 2 years (…). Big Data is maybe too much for where we are. Our Big Data is NÓNIO".

MM was the other respondent apart from JF who confirmed the publisher he works for has an investment in Big Data besides NÓNIO. However, MM is talking about second or third party data, meaning that the data is not directly collected by Media Capital, but rather provided by other sources such as Google. MM says that when it comes to audience analytics and even audience segmentation for advertising purposes, Media Capital has been relying on data from other sources and it is that "third party data that we want to stop using. We want to rely only on our own data and not on data from Google."

Ethical concerns

Despite Lewis and Westlund (2014) and boyd and Crawford's concerns (2012) about data privacy, all respondents argued that consumers are willing to provide the data, all that information regarding their preferences and behaviours, in exchange for something of value, which in this case is free trustworthy news, an idea presented by Sandra González-Bailón in her 2013 paper. There are those, like ML, who reason that consumers are prepared to make such an exchange, and points out the similarities between NÓNIO's *modus operandi* and Facebook's or Google's.

> "It happens the same to the 6 million Portuguese registered on Facebook and all of them did it without blinking. It is the same way with the cell-phone market where 80% of cell-phones in Portugal are androids and in order to activate the android it is required to have a Google account and everyone has one. (…) the trade-off is people registering so that we [publishers] can continue operating. Otherwise they can read the news from free fake news websites. Journalism has to be paid and people [journalists] need to earn a salary."

MM thinks that the trade-off cited by ML and Sandra González-Bailón (2013) hasn't been properly explained to customers. In his view, the consumer perceives the interest of publishers to gather user-data mostly as an abusive attitude. This opinion is shared by JF:

> "It is always difficult for the consumer to make that exchange. From our experience, consumers always resist to providing data, however, (…) we expect that when consumers do understand that they are getting free content in exchange for some very basic and simple information, they end up providing private data."

A term of comparison with Google and Facebook was always drawn when discussing consumers' privacy concerns with the five interviewees.

> "Users nowadays have no problem in providing personal information", sustains MM. "When we are on Facebook, we're saying what our preferences are and Facebook is monitoring that (…). Probably, after this videoconference, I will be your new Facebook friend suggestion. Take notice of the amount of data Facebook and Google are collecting

in order for them to make their recommendations. It is much more than NÓNIO wants to know. All I want to know is what type of content the user sees when he/she is in our own platforms. I don't care to which websites he/she goes (…). I think that users don't have a positive reaction when they see companies like Media Capital or Impresa or Global Media wanting to record their personal data. But the truth is that those same users give all their data freely and willing to giant corporations".

"Europe got carried away with GDPR[7] (…)", says JPL. "(…) Apple and Google keep collecting all the information they want and violate every single legislation line in GDPR, all of them, not just a few, all (…). At the moment, there are a few Apps whose terms & conditions state they can have the microphone on by default to listen to our conversations and to collect data which can be used for advertising purposes. The invasion levels are extreme (…) but people complain about NÓNIO, which is child's play next to everything there is. So, this discussion about whether people are willing to accept the tracking is non-sense. (…) When Google buys Waze it is because it generates revenue selling advertising through the data collected. But we love to use Waze because it is free and we close our eyes to the fact that the App has to be paid somehow. Who pays for it? Us, with our privacy."

"We always allow people to pay, but it has been proved that the vast majority of people would rather watch publicity so that they don't have to pay [for the content]. We always give the option of paying, so that there is no tracking (…). All we are trying to do is to keep the revenue here in Portugal. This is a problem newspapers all over the world are facing. People have to realize that either they are willing to pay for content or otherwise decide not to pay and we [publishers] have to live-off advertising, and advertising wants to know who the consumers are. That is the trade-off we have to explain better", defends ML.

TC was more succinct when discussing this topic, and when asked the same question, he replied with: "that is the one million dollar question".

Despite existing concerns particularly from consumers regarding data privacy, the suppliers, meaning the five respondents who work for the publishers, all deny the

[7] The General Data Protection Regulation was created in 2018. It legislates the digital privacy and Data protection of all individuals in the European Union and in the European Economic Area (https://eur-lex.europa.eu/eli/reg/2016/679/oj)

existence of ethical issues associated with the NÓNIO project. However, MM does warn about the dangers of letting algorithms recommend content to users. By programming algorithms to tailor the experience to a particular consumer whenever he/she accesses a publisher's platform, there is the possibility that the user will be kept in a bubble of information, where he/she will only see content that is directly aligned with his/her interests without being exposed to other type of articles. MM raises awareness to what Philip Napoli defined as the "dictating" power of Big Data when it comes to content consumption (2014a, 2014b).

"These machines (…) they have limits, they are not smart enough. (…) there has to be an input from the editor along with recommendations based on [the user's] navigation." To better explain the flaws of recommendation software, MM shares the example of what happens with YouTube, where the recommendation software keeps recommending the same songs to the user in a music loop.

> "I think this can affect the real journalism. Journalism is based on having a curator (…) or a journalist that recommends news and content which are relevant. I feel that Big Data, for now, is incapable of recommending anything other than a trend. (…) [recommendation] has to do with my personal taste and past navigation. It has to do with the past and not with a current relevant event to the world or to a country. The role of the curator and the human seizes to exist. Information starts being delivered by a machine which is agnostic (…). The problem for mankind is that we face the risk of being kept in a loop of recommended content and lose touch with reality. Look at North Korea which is closed and has no clue of what is happening with the outside world."

Thus, MM agrees with Lewis and Westlund (2015b) that Big Data is changing the ethical way of conducting journalism. ML, though, is more sceptical about the dangers of software recommendation. He does recognize the ongoing discussion in the journalistic field, but he summarizes it all as "philosophical questions": "Google and Facebook do it (content recommendation) with algorithms. We [Global Media] do it manually and with an editorial criteria. I think a mix of both might be a good idea."

Overall, apart from MM, respondents don't seem as preoccupied about how Big Data will affect the future of journalism in an ethical way as Rebekah McBride (2016) who is far

more concerned with something other than the technological side of the data revolution (McBride worries about how Big Data will shape the professional norms, routines and ethics of journalists). At the same time, there still seems to be little evidence on the sociological impact of employing Big Data in Portuguese journalism, something which was already stated by C. W. Anderson back in 2012.

Other ethical issues such as Diakopoulos's (2014a) and De Maeyer et al.'s (2015) algorithmic accountability seem to have been considered when creating NÓNIO. None of the respondents is worried about the possibility of human biases manifesting themselves on the NÓNIO algorithms. JF observes that most of what NÓNIO's algorithms do is classification. The algorithms are analysing the users' behaviours and classifying them as belonging to the segments described above. That classification is made using precise criteria which are defined not by one single person, ads JPL, but by a team comprised of several people, thereby ensuring that no preferences from one particular programmer can be implemented on the software. "We define someone's profile as belonging to the segment Sport Followers (…) if he read X sports articles, 10 times a week, 3 times a day, whatever (…)" explains MM, reinforcing that all thresholds are defined as a group. "There is no one saying that someone likes sports because he has a beard and an orange hair. (…) The programmers won't have a big influence on the final output".

Part of the reason why the Portuguese publishers decided to create NÓNIO was to "interrogate the algorithm", as Diakopoulos (2014a) describes it. According to the respondents, the publishers weren't happy with receiving the data, or, more concretely, the segments from other sources such as Google, without being able to question how those segments were defined. They call it the black box of Google and Facebook, in the sense that no one (except for Google and Facebook's employees) knows how it operates:

> "In reality, that is defined by people and programmers who decide what the pattern is and that is the biggest black box of Google and Facebook. I have no idea how the Data profiles sold by Google are created. (…) our pitch is that they [the Data profiles sold by the publishers] is completely transparent and I can tell the advertisers what were exactly the thresholds used in defining those users as part of one particular segment", observes MM.

By creating a software that generates Data for them and segments the consumers according to their own stipulations, publishers are more in control of the algorithms and no longer need to use second or third party data.

The last question of the interviews was whether NÓNIO could be a first step in news automation, to which the answers varied from "not really" to "hope not":

> "I do hope not, but I see in humanity that tendency (…). When you work in a capitalist world where the goal is for the economy to grow year after year and you see demographically that the number of consumers doesn't increase, the solution is to force people to consume more. This means markets will have to produce more with less and less has usually to do with less costs, which translates into human resources", says MM.

> "I don't think anyone wants that, no one will fight for that, because automation… are we going to be left without journalists? That doesn't make sense. (…) if we think properly about it, we will see that we are working to become irrelevant", believes TC.

JF and JPL are more sceptical about NÓNIO being a first step in news automation, but keep an open door to that eventually happening, which means Portuguese publishers can one day become like *The Los Angeles Times* and its automation regarding earthquake news or like *The Washington Post* with its Truth Teller prototype (Lewis & Westlund, 2014). JF believes the rise of the robotic reporter (Carlson, 2014) will happen, but not necessarily because of NÓNIO. He argues that Big Data might play a role in helping the algorithm getting smarter because of all the data it can feed it, and NÓNIO can contribute to that, but only in that sense, he says.

JPL looks at the problem in a more economical way, weighing the costs of such an initiative. In his opinion, the NÓNIO can lead to news automation in Portugal because it is proof the Portuguese publishers can cooperate to achieve innovation, something which could happen again in order to develop news automation software. JPL talks about news automation as an inevitable thing and shares the example of news sports:

> "It is the most obvious one. I am not underestimating the work of sports journalists, but think about sports articles. The pre-match and the post-match interview are always the same. (…) These type of news, mixed with some data… the outcome is impressive

because there is little difference from the ones written by humans. (…) Machines will replace us in those behaviours where we act like machines. (…) When journalists simply follow a rigid structure, (…) obviously machines will do it better than humans. (…) Machines have an advantage over anything that is just repetition, just training. (…) But when we want creativity, then we are talking about artificial intelligence. (…) It is not just pure repetition and, if that is the case, we cannot claim to be dealing with artificial intelligence just because we are employing Big Data and repetition. All computers were capable of playing chess and were already as good as the player who had played against them, the problem was that they couldn't beat that player. And in those cases we didn't talk about artificial intelligence. (…) Automation is not linked to NÓNIO. Automation will appear in areas where people ad little value."

4. Discussion

It is now possible to address the research questions presented on the methodology chapter. To help understand whether Big Data is changing the activity and business model of the Portuguese media corporations, which was the main research question, two other sub-questions were drawn.

a) Is Big Data changing news production? Does it have an impact on what journalists write about?

The first sub-question regards the impact of Big Data on the production of news and whether tools such as NÓNIO were influencing and changing what journalists wrote about. Based on the conducted interviews, the answer is yes, though small. Big Data is having an impact on journalistic content production, but not through NÓNIO. There were, however, respondents who considered hypothetic scenarios where the information collected through NÓNIO could lead to particular articles being constructed. Take for example JF and the discussion about the possibility of articles including both chocolate and Cristiano Ronaldo if found out that readers were interested in those two things. On the other hand, JPL and ML claim that the Portuguese publishers are already aware of which articles are most consumed due to other data analytical tools employed:

> "(…) that has nothing to do with NÓNIO. Journalists know, every day, which content was the most consumed on the previous day because there is Google analytics. In television there are ratings", says ML.

Therefore, it is safe to say that respondents weren't interested in how NÓNIO could affect content production, as it wasn't one of the reasons why the project was created. JPL has indeed stated that out of all the fields where NÓNIO and its Big Data practices can have an impact, content production is the one where he expects to see the least.

It is true that both MM and TC stated NÓNIO could have an impact on what journalists write about, but when discussing this topic, there was the feeling that, once again, this was not the publishers' top priority, which is surprising considering this dissertation's

literature review where the international community seems to have devoted a lot of attention to the matter.

According to Lewis and Westlund (2014), Big Data has changed how the filtering and processing information is done, as well as the editing part. Take for example CNN. The North-American broadcaster's Big Data tools are focused on providing breaking news, monitoring users' behaviours and analysing data sets for journalistic content (Stone, 2014). Even the *Sacramento Bee* newspaper, which didn't clearly state their Big Data tools were leading to major changes in content production, stated to be using them to figuring out consumers. The American newspaper is not only interested in finding out what consumers are reading, but also the level of engagement with the story (Stone, 2014). Comprehending such things as for how long do customers engage with the product is crucial in improving that same product, which is a big indicator that the Sacramento Bee is using Big Data to refine its content production.

There is another way NÓNIO can influence journalistic content production by providing data to enrich articles, as better source material (more and better information) correlates highly with more quality articles. JF suggested this when discussing the topic. This type of data journalism was coined by Berret and Phillips (2016) as computer-assisted reporting, whereas Meyer (1973) called it precision journalism:

> The tools of sampling, computer analysis, and statistical inference increased the traditional power of the reporter without changing the nature of his or her mission to find the facts, to understand them, and to explain them without wasting time (pp3).

This dissertation's investigation found little evidence of NÓNIO having an impact on content production, though there is a connection between data analytics and the articles produced by journalists, which also extends to news broadcasted in television (recall ML's remarks about ratings in television).

b) How is Big Data affecting the revenue channel?

The second question is designed to understand how Big Data is affecting the media's business model and whether it can lead to the economic survival of the media.

When asked directly about their opinions on Big Data assuring the economic viability of the sector, some respondents answered affirmatively, as they believe indeed in the media's financial sustainability through Big Data. Others, despite refusing to state so in such a straightforward way, agree that Big Data is the path most likely to lead to the economic survival of media corporations.

Very much like expected, respondents were extremely interested in Big Data's impact on publishers' revenue streams. There was a big enthusiasm when discussing, for example, content recommendation through NÓNIO and how it translates into more and longer subscriptions. This conclusion about longer subscriptions is based on JF's statement regarding how publishers more or less already know how to attract new subscriptions, but that the problem is maintaining them.

Big Data leads to a better understanding of the consumer which makes it easier for publishers to supply content that best fits the interests of that consumer. Such content recommendation is no different from what many e-commerce firms do, as pointed out during literature review (Manyika et al., 2011). Take for example the case of Amazon, which relied on Big Data to create a recommendation engine that accounts for over 35% of all sales (Akter and Wamba, 2016). The ability to recommend personalized content greatly improves any company's sales and for publishers, it ends up improving the number of subscriptions.

Respondents also observed that NÓNIO can help with subscriptions through personalized marketing campaigns. Consumers can access content through NÓNIO without paying, but that doesn't mean those same consumers aren't generating value to the publishers. While browsing through NÓNIO's websites, users are providing data about their interests, which can then be used by newsrooms to better target those same consumers. As JF explained, if someone has the profile of a business man, then Cofina (the group JF works for) will target that consumer and see if he/she is interested in paying for a subscription of *Jornal de Negócios* (a newspaper focused on finance, economics and business).

Another way Big Data is improving the revenue streams of publishers has to do with advertising. Wanting to change how publishers conduct advertising was one of the main reasons for creating the NÓNIO project. Publishers were tired of depending on Silicon Valley companies such as Facebook or Google for advertising. Newsrooms would sell advertising slots through Facebook's and Google's platforms, which would then be matched by companies wanting to advertise their products, while the tech companies would take a percentage of the revenue, which, according to respondents, was too high. By creating the NÓNIO project, publishers will be able to sell advertising slots directly to companies. Why? Because publishers themselves will be able to sell the segments demanded by advertisers, which so far only companies like Facebook and Google were capable of providing and which is why advertising in the digital world has heavily relied on those companies for the last few years.

Respondents' excitement with Big Data's impact on advertising is understandable. According to Visual Capitalist, an online publisher, 98% of Facebook's revenue comes from advertising (see figure 1 in Appendices section). The tech giant founded by Mark Zuckerberg excels at attracting advertisers due to its Big Data techniques, which allow for greater efficiency every time an advertising campaign is done through Facebook.

It was also the *Sacramento Bee* newspaper that claimed to be using data analysis to improve its advertising and to help advertisers target demographic interest groups (Stone, 2014). Portuguese publishers have the same goal, according to the interviewees' responses. Their belief is that advertisers are willing to spend huge amounts of money in improving their own advertising, which comes down to better targeting potential consumers. These potential consumers can be identified by publishers and other content producers as they can easily figure out users' interests based on the content consumed. Users are then grouped into segments, or interest profiles, which match the target audience for advertisers.

It is important to point out that in the case of Portuguese publishers, the "advertising revolution" is still in early stages. It seems, from the interviews, that publishers are still waiting for the NÓNIO to gather more users, before delving deep into selling the segments to advertisers: "It will come on a second phase when we have a larger database with more users." said JF. However, the advertising business model of publishers is

indeed already changing, making it possible to conclude that Big Data is substantially altering the revenue channels of publishers.

So, how is Big Data changing the activity and business model of Portuguese media corporations?

Lewis and Westlund (2015b) argued that Big Data is not only changing the media in ways of knowing and doing, but in economic and ethical ways as well.

As already discussed in this chapter, Big Data is revolutionizing the media's business model. Nowadays, the media heavily relies on digital platforms to conduct its business, which is not a consequence of Big Data. However, reality is that Big Data is enhancing how publishers and other types of media can explore their digital streams and revenue.

First, Big Data drastically improves the knowledge of the consumer, which is an advantage for all types of businesses. By knowing the consumer better, suppliers are able to better match advertisers' needs and consumers' needs. Truth is all consumers have needs and that is why they consume, why they buy products. In the case of the media, consumers want content, and in the case of journalism, it comes down to information. But what kind of information do consumers want? Do consumers only want the immediate breaking-news information? Or are they interested in more in depth articles?

JF stated that:

> "there is a big challenge for all publishers which consists on remaking the products, meaning that it can't only be the breaking-news product or the news of the moment. We have to invest in the premium offer, the paid offer has to have added value, a tangible value which makes a difference from the traditional information which is free."

Suppliers already realized that to convince consumers to pay for products, they can't simply sell 'breaking news'. Perhaps because in the age of internet it is difficult to persuade people to pay for breaking news, as it is too hard to compete with platforms like Google, Facebook and many others, where such type of content abounds freely.

As the investigated Portuguese publishers believe that consumers are willing to pay for desired content, all that is needed is to find out what consumers want, and for that Big Data is a real game changer. From that point on, publishers must show the consumer that they are producing the desired content.

Thus, Big Data is changing how Portuguese publishers display the information to the user. Displayed content will be much more personalized, like Netflix does with its platform's users.

Big Data is changing the display of information, but not the information produced by Portuguese publishers, as respondents rejected the idea of producing tailor-made information. This was quite surprising, considering what is done internationally and how Netflix produces content that matches users' needs. However, publishers might reject the idea of tailor-made information due to the fact that they are information providers and not content providers. Journalists see themselves as the guardians of democracy and for that to happen, readers must be exposed to all sorts of news.

Otherwise, there is the risk that consumers will be kept in a bubble, as warned by MM. MM's warning poses an important question regarding the ethical behaviour of journalists and all those associated with the field. Should the media fulfil its duty of informing society or should they thrive for economic profits by delivering only what consumers want? This question is particularly relevant nowadays, as the media face increasing economic struggles. So far, Portuguese publishers seem willing to maintain its integrity, but will it be the case in the future? NÓNIO is still in the early stages and money may end up dissuading good intentions.

Conclusion and Future Research

This qualitative study examined how the Portuguese media, and particularly Portuguese journalism, is using Big Data. As seen in the literature review, Big Data is an innovative concept, mostly explored in the second decade of the 21st century. It both describes large data sets and the techniques used to analyse them. It has proved a strategic ally for many businesses, but how about the media? Is there a use for Big Data in journalism? This dissertation's literature review does suggest so indeed, with many companies having already invested and presently collecting the returns on Big Data projects. One objective of this dissertation was to understand if that was the case in Portugal, where newsrooms have long been facing financial struggles and fighting irrelevance caused by the internet as well as international newspapers and broadcasters finding their way into the Portuguese market.

To answer such questions, this dissertation focused on a new Big Data project being implemented by five of the biggest media corporations in Portugal. The goal was to study the NÓNIO project and assess Big Data's influence in newsrooms: what were its implications on the daily life of journalists and on the business side of running newspapers and information television channels?

To better guide the investigation, such doubts were split into two smaller questions: a) what was the impact of Big Data in news production? and b) was it having an impact on revenues?

Both questions can be summed up by how Portuguese journalism is adapting to new technology. It is fair to say that journalism has traditionally been slow in applying technological breakthroughs to improve its field and businesses. Journalists tend to see technology as an enemy capable of depriving them of a profession and as an intruder that takes away the romantic side of the "best job in the world" as described by the Nobel literature laureate and former journalist, Gabriel García Márquez, in his 1997 essay where it is evident a disdain for technology.

To answer the research questions, this research was conducted through half-hour interviews, consisting on open-ended questions that allowed for interviewees to fully

express their perceptions on the matter. Interviewees were the people in charge of NÓNIO at each of the media institutions participating in the project.

Contrary to what was expected, this research found Portuguese newsrooms technologically equipped and prepared to face the digital reality of the 21st century. Take Impresa's (one of the media groups joining the NÓNIO initiative) recent decision to move SIC's newsroom, its television broadcaster, to a new building (January 2019), in an effort to modernize its infrastructure and enhance its technological power. Findings do suggest that newsrooms are already jumping on "the Big Data bandwagon" (FSR Magazine, 2018).

Though Big Data is being employed by Portuguese media corporations, there is little interest in how it can be used in content production to better match readers' preferences, once again contrary to what could be expected. It is evident that readers' preferences are at the top of priorities for the companies adopting Big Data techniques. That is what Big Data is all about: a tool to get to know consumers and potential consumers better, to figure out who they are so that businesses are better able to match their needs. Which is why it is surprising to see some apathy in how NÓNIO can lead to changes in content production. As previously explained, this finding goes against what other companies dealing with content production and employing Big Data are doing. Take the case of Netflix, which is famous for axing series after the first or second seasons in case of little engagement with audiences. Netflix's production decisions are solely based on collected Data (Tryon, 2015). Based on the conducted interviews, consumers' interests and feedback on articles seems to be taken into consideration, however, don't dictate content production, unlike some of the examples found in the literature review (take the cases of BuzzFeed, the *Financial Times* or the *Sacramento Bee*). Portuguese newsrooms still very much value editors' picks on what is news information and journalists' ways of writing and editing. Even though content production worldwide is influenced by data, it can be that Portuguese journalists still deem themselves the gatekeepers of information, a role they refuse to trust to machines.

The same cannot be said about content recommendation, one of the new Big Data tools used to extract the most out of each produced content. As stated, Big Data is a way of getting to know the consumer better. Its algorithms are looking for patterns in consumer's

behaviour that match given interest profiles. For example, if a consumer is a constant reader of sports articles, NÓNIO will determine that such consumer is a sports enthusiast. After knowing this, NÓNIO's algorithms will be able to recommend to that consumer sports articles whenever he/she accesses one of NÓNIO's websites. Through content recommendation, newsrooms decrease the risk of generating sports articles – which cost money to produce - that are not seen by consumers, while at the same time making sure that readers who are not interested in sports do not come across sports articles. Therefore, content recommendation is a way for newsrooms to generate less waste.

However, as MM warns, it is important to be aware of information bubbles. MM is worried about only exposing people to information that matches their interests. He believes such *modus operandi* is a disservice to mankind and contrary to journalism's principles of safeguarding democracy. Former American president Barack Obama agrees on the dangers of information loops. In an interview that aired on January 12, 2018, on Netflix's show *David Letterman – My Next Guest Needs no Introduction,* Obama said the following:

> "if you are getting all your information of algorithms being sent through a phone and it is just reinforcing whatever biases you have, which is the pattern that develops, (…) at a certain point you just live in a bubble, and that is why our politics is so polarized right now."

To fight this bubble, MM believes that content recommendation cannot be entirely trusted to algorithms and that human journalists should also play a part in telling consumers what is important and what should be read.

Even though content recommendation might inadvertently be doing a disservice to journalism, there is the conviction within the sector that it will increase its revenues. The thinking is as follows: through content recommendation, consumers will be more satisfied with the product, which in this case is the newspaper or news broadcaster, and will be willing to pay more for that product. It relies on the economic assumption that greater satisfaction leads to greater demand and consequently to more money from the client (the principle guiding all firms employing Big Data techniques). In the case of newsrooms, it very much comes down to longer and/or more expensive subscriptions.

The same applies to potential consumers. NÓINO will improve newsrooms' marketing campaigns, meaning that potential readers will be "bombarded" with tailor-made content in order to spike their interest and stimulate demand. This will improve subscription ratings.

Still regarding subscriptions, truth is that NÓNIO also leads to improvements in this area in a way not linked to its Big Data algorithms. By asking consumers to register, NÓNIO is overcoming the first major obstacle in any subscription. JPL describes it as a "tremendous advantage" in the whole process of convincing consumers to pay for news.

Research also shed light on how Big Data can have an impact on publishers' advertising business. It relies on the same way Facebook and Google use algorithms to generate higher revenues. By understanding the consumer better and finding out what his/her interests are, it is then possible to sell those interests to advertisers. Advertisers want to market a product among those who are interested and have the means to buy it. It no longer makes sense to advertise sports items among people who have no inclination to practice sports. The fact is that content producers have an advantage over other types of businesses in knowing what interests their consumers, as content consumption is very much aligned with "life interests", unlike, for example, food consumption. Take the example of a supermarket. Even though food chains can employ Big Data techniques to find out a consumer's food tastes, that doesn't tell anything on whether that person likes to play sports, videogames or if he/she likes cars or travelling. Knowing consumers' preferences not only improves the business side of content creation, but the side of selling advertising slots as well.

Respondents were particularly excited with NÓNIO's potential in changing how Portuguese publishers' advertising business was conducted. Since the advent of Big Data, publishers have been paying big techs such as Google to sell their own advertising slots to publishers. Google, through its browser and search engine, would figure out consumers' interests by seeing what articles they read. These were articles created by Portuguese publishers which could be found on Google (though not necessarily available for reading; most often Google shows the link for the article and when trying to access it, it appears locked requiring subscription or payment or both). Google would then sell these

interests to advertisers who would want to buy advertising slots from the publishers and the publishers would sell them, while Google would be acting as a kind of broker in this process. Before NÓNIO, Portuguese publishers had no way of figuring out consumers' interests and therefore were forced to resort to Google for better advertising. From now on, they will be able to run advertising without the interference of big techs.

After analysing the interviews and answering individually to the two sub-research questions, it was finally possible to address how Big Data is changing the activity and business model of Portuguese media corporations.

The implications on the day to day life of journalists are small, considering the reluctance of respondents in using the insights gathered from NÓNIO when writing and editing articles. Respondents answered that newsrooms already employ techniques established prior to NÓNIO that tell them how readers react to articles. Even though newsroom have such knowledge, that doesn't seem to have much impact on content production, both in newspapers and in television news, at least according to respondents. Unfortunately, there was no time to properly assess if these claims are in fact true. Could it be that respondents refuse to acknowledge that insights on consumers' preferences guide news content creation? Not being able to test this hypotheses is certainly a limitation of this study. This hypothesis is reinforced by what was found during literature review. Data can enrich an article by providing more information and also by functioning as a fact-checker, and that may have an impact on the journalistic profession, but apart from that, according to this research, Big Data is expected to have a small impact on journalists' content production day-to-day life. Truth be told, this is a disappointing conclusion.

There is however an ongoing revolution on its business model, which is why respondents believe Big Data has the potential to financially save the media from its precarious situation. The way subscriptions are attracted and sustained and the way advertising is conducted is drastically changing due to Big Data. The change is evident when looking at it from the producer's side, but if one examines the consumer side, it is possible to find modifications as well. The way consumers "receive" news whether through email or when accessing a publisher's platform will heavily rely on content recommendation (in a similar way to Facebook's or Twitter's methods) and that is due to Big Data.

It has been pointed out that the inability to actually test some of the respondents' claims such as the one about disregarding readers' preferences when writing articles is a limitation of this study. That limitation has to do with time constraints.

Another factor that also had an impact in shaping this research was the COVID-19 pandemic. Most of this investigation was written in 2020, the year that saw the rise (and so far the continuum) of the pandemic. Portugal was in a state of lockdown (state of emergency) from March to the beginning of May. During that period, people were advised to stay at home and to avoid unnecessary contact. For that reason, the interviews for this investigation were not conducted face-to-face as originally planned. In some cases, as with ML and JF, interviews were done through phone calls. The inability to see the faces of respondents creates a barrier (which also exists when conducting interviews trough skype, zoom or Microsoft teams) and ultimately it leads to less information that can be extracted from the conversation. Take the case of respondents' unease when discussing ethic problems regarding NÓNIO. That discomfort could be seen in their faces. The same thing did not happen when talking with JF and ML and more attention had to be paid to things like the tone of voice or how long it took them to answer the question.

Besides the obstacle of conducting interviews through the computer or through the phone, it was also very hard to find a suitable time for JF and JPL. This is not a limitation, as both of them ended up being available to talk for half an hour, but it was a hardship worth mentioning.

Another limitation was the fact that ML's interview was limited to 15 minutes. ML informed that he only had 15 minutes to spare for the interview. As a result, a selection of the questions was done and the ones deemed more important were asked.

It is also important to point out that focusing solely on NÓNIO was a reflection of lack of other materials that could be investigated so as to provide valuable insights for this dissertation. Not finding other Big Data projects in the Portuguese media besides NÓNIO is a limitation of this study, as this research draws conclusions from one single study object, though observed through five different perspectives. Time constraints, together with it being a recent technology in Portuguese media corporations, worked as an impediment in trying to find other Big Data projects in Portuguese newsrooms.

Big Data is still a new concept in Portugal, particularly in the field of journalism and there are no terms of comparison. Because Big Data and NÓNIO are still in beginning stages, some conclusions are based on respondents' expectations of what will happen in the future and not on past events. Therefore, it would be interesting to investigate whether those conclusions still hold in the future.

One more limitation worth pointing out is that there was no actual scrutiny over the algorithms developed in the NÓNIO project, apart from questioning interviewees on whether they felt the algorithms could manifest human biases. The actual programmers who developed the algorithms were never interviewed and the people who answered the question lacked the technical skills to answer it. Therefore, Diakopoulos (2014b) and Stavelin's (2014) claim that algorithms are opaque proved to be true in this investigation as they were never fully and comprehensively examined.

Thus, the algorithm should be interrogated (Diakopoulos, 2014a) in future research on NÓNIO or on Big Data in journalism. Important questions still remain regarding the algorithm: how was it developed? Who programmed it? Why were the thresholds on classifying someone as sports enthusiasts or an auto lover defined as they were?

Future research should also investigate the development of news automation in Portugal. So far, it seems like a distant reality. Yet, literature review showed that it is already in progress in many parts of the world (recall *The Los Angeles Times* and *The Washington Post*'s examples). Will Portuguese publishers delve into the reins of artificial intelligence and news automation? And if so, what will happen to journalists? How will newsrooms be shaped by that decision?

The economic situation of Portuguese publishers should also be looked at in the future. Will Big Data actually lead to the economic survival of the media?

That publishers have for a long time struggled with economic viability is an undeniable truth, evident in the Gabriel García Márquez essay referenced earlier in this chapter (1997). Journalism is a profession known for low salaries and high unemployment, but, who knows, data might be about to change it in a big way.

References

Abourezk, L. (2014), "How time-based measurement is grabbing digital publishers' attention", Retrieved from https://digitalcontentnext.org/wp-content/uploads/2014/10/DCN-Report_Time-BasedMeasurement_10.22.14.pdf

Akter, S. & Wamba, F. (2016), "Big data analytics in E-commerce: a systematic review and agenda for future research", *Electron Markets*, 26, 173–194.

Akter, S., Wamba, F. S., Gunasekaran, A., Dubey, R., Childe, S, J. (2016), "How to improve firm performance using big data analytics capability and business strategy alignment?", *International Journal of Production Economics,* 182, 113-131.

Anderson, C. (2008), "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", Wired Magazine, Retrieved from https://www.wired.com/2008/06/pb-theory/.

Anderson, C. W. (2012), "Towards a sociology of computational and algorithmic journalism", *New Media & Society* 15 (7), 1005-1021.

Anderson, C. W. (2015), "Between the unique and the pattern: Historical tensions in our understanding of quantitative journalism." *Digital Journalism*, 3 (3), 349-363.

Anderson, C. W. Bell, E. & Shirky, C. (2012), "Post-industrial Journalism: Adapting to the Present", *Geopolitics, History, and International Relations*, 7, 2, 32-123.

Bain & Company. (2013), "The Organizational Challenge" Retrieved from http://www.bain.com/Images/

BAIN_BRIEF_Big_Data_The_organizational_challenge.pdf, 2013

Bajari, P., Chernozhukov, V., Hortaçsu, A., & Suzuki, J (2019),"Big Data in the Modern Economy – The Impact of Big Data on Firm Performance: An Empirical Investigation", *AEA Papers and Proceedings*, 109, 33 – 37.

Barnes, T. J. (2013), "Big data, little history", *Dialogues in Human Geography,* 3(3) 297-302.

Barrett, M. Davidson, E. Prabhu, J. & Vargo, S. (2015), "Service innovation in the digital age: Key Contributions and Future Directions" *MIS Quarterly*, 39, 135-154.

Barton, D. Court, D. (2012), "Making advanced analytics work for you", *Harvard Business Review* 90, 78.

Batty, M. (2013), "Big data, big issues", *Geographical,* 85(1), 75.

Beath, C. Becerra-Fernandez, I. Ross, J. & Short, J. (2012), "Finding value in the information explosion", *MIT Sloan Management Review*, 53, 18–20.

Bell, E. (2012), "Journalism by Numbers." *Columbia Journalism Review,* Retrieved from http://www.cjr.org/cover_story/journalism_by_numbers.php?page=all.

Berg, B. L., & Lune, H. (2017), *Qualitative research methods for the social sciences*. 9th edition, Boston: Pearson.

Berret, C., Phillips, C. (2016), *Teaching Data and Computational Journalism*, Columbia Journalism School/Knight Foundation, New York: Rosemont Press.

Beulke, D. (2011), "Big Data Impacts Data Management: The 5 Vs of big data", Retrieved from URL:http://davebeulke.com/big-data-impacts-datamanagement-the-five-vs-of-big-data.

Biesdorf, S., Court, D., & Willmott, P. (2013), "Big data: What's your plan?", *McKinsey Quarterly*, 40-41

Bose, R. (2009), "Advanced analytics: opportunities and challenges" *Industrial Management & Data Systems*, 109, 155 –172.

Bowker, G. C. (2005), *Memory practices in the sciences*, Cambridge, MA: MIT Press.

Bowker, G. C. (2013), "Data flakes: An afterword to "raw data" is an oxymoron" In L. Gitelman (Ed.), "Raw data" is an oxymoron (pp. 167–171). Cambridge, MA: MIT Press.

Boyd, D. & Crawford, K. (2012), "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon", *Information, Communication & Society* 15 (5), 662-679.

Burmester, M. Mast, M. Tille, R. & Weber, W. (2010), "How Users Perceive and Use Interactive Information Graphics: An Exploratory Study", *IEEE Proceedings of the 14th International Conference Information Visualization* (IV 10), London, 361–368.

Carlson, M. (2014), "The Robotic Reporter: Automated Journalism and the Redefinition of Labor, Compositional Forms, and Journalistic Authority", *Digital Journalism,* 3 (3), 416-431.

Carson, A. (2015), "Behind the newspaper paywall – lessons in charging for online content: a comparative analysis of why Australian newspapers are stuck in the purgatorial space between digital and print", *Media, Culture & Society,* 33(8): 1202–1219.

Chandler, A. D. Jr. (1977), *The Visible Hand: The Managerial Revolution in American Business,* Cambridge, MA: Harvard University Press.

Chen, H., Chang, R. H., & Storey, V.C. (2012), "Business intelligence and analytics: From big data to big impact", *MIS Quarterly*, 36(4), 1165-1188.

Coddington, M. (2015), "Clarifying Journalism's Quantitative Turn", *Digital Journalism*, 3 (3), 331-348.

Cohen, S. Li, C. Yang, J. & Yu, C. (2011), "Computational Journalism: A Call to Arms to Database Researchers." Paper presented at the 5th Biennial Conference on Innovative

Data Systems Research (CIDR '11), Asilomar, CA. Retrieved from http://ranger.uta.edu/~cli/pubs/2011/cjdb-cidr11-clyy-nov10.pdf.

Columbus, L. (2014), "84% Of Enterprises See Big Data Analytics Changing Their Industries' Competitive Landscapes In The Next Year", *Forbes Magazine*, Retrieved from https://www.forbes.com/sites/louiscolumbus/2014/10/19/84-of-enterprises-see-big-data-analytics-changing-their-industries-competitive-landscapes-in-the-next-year/#37bdb9a117de.

Corbin, J. & Strauss, A. (2014), *Basics of qualitative research: Techniques and procedures for developing grounded theory*. 4th Edition, London: SAGE.

Cox, M. (2000), "The Development of Computer-Assisted Reporting". Paper presented to the Association for Education in Journalism and Mass Communication, Chapel Hill, NJ: University of North Carolina.

Crawford, K. Miltner, K. & Gray, M. (2014), "Critiquing Big Data: Politics, Ethics, Epistemology", *International Journal of Communication* 8, 1663–1672.

Davenport, T. H., & Harris, J. G. (2007), *Competing on analytics: The new science of winning*. Boston: Harvard Business School Press.

de Maeyer, J. Libert, M. Domingo, D. Heinderyckx, F. & Le Cam, F. (2015), "Waiting for data journalism: A qualitative assessment of the anecdotal take-up of data journalism in French-speaking Belgium." *Digital Journalism*, 3 (3), 432-446.

Deuze, M. (2004), "What is Multimedia Journalism?" *Journalism Studies* 5 (2), 139–52.

Deuze, M. (2008), "Understanding Journalism as Newswork: How it Changes, and How it Remains the Same." *Westminster Papers* in Communication and Culture 5 (2), 4–23.

Diakopoulos, N. (2014a), "Algorithmic Accountability: Journalistic Investigation of Computational Power Structures", *Digital Journalism*, 3:3, 398-415.

Diakopoulos, N. (2014b), "Algorithmic Accountability Reporting: On the Investigation of Black Boxes", New York: Tow Center for Digital Journalism

Domingo, D. (2008), "Interactivity in the Daily Routines of Online Newsrooms: Dealing with an Uncomfortable Myth." *Journal of Computer-Mediated Communication* 13 (3), 680–704.

Driscoll, K. & Walker, S. (2014), "Working within a black box: Transparency in the collection and production of big Twitter data." *International Journal of Communication,* 8th edition, Boston: Pearson, Retrieved from http://ijoc.org/index.php/ijoc/article/view/2171.

Ehrenberg, A. S. C., & Wakshlag, J. (1987), "Repeat-viewing with people-meters", *Journal of Advertising Research*, 27(1), 9.

Erevelles, S., Fukawa, N., Swayne, L. (2015), "Big Data consumer analytics and the transformation of marketing", *Journal of Business Research,* 60, 897-904.

Fink, K. & Anderson, C. W. (2014), “Data journalism in the United States: Beyond the 'usual suspects'”, *Journalism Studies*, 16 (4), 467-481

Firican, B. G. (2019), "How Data Is (And Isn't) Like Oil". *Transforming Data with Intelligence*, Retrieved from https://tdwi.org/articles/2019/04/22/data-all-how-data-is-like-oil.aspx.

Flew, T., Spurgeon, C., Daniels, A., & Swift, A. (2011), “The Promise of Computational Journalism”, *Journalism Practice*, 6 (2), 157-171

Forrester (2011), “Expand your digital horizon with big data”, Retrieved from https://www.forrester.com/report/Expand+Your+Digital+Horizon+With+Big+Data/-/E-RES60751-

García Márquez, G. (1997). “The best job in the world”, Retrieved from https://journals.sagepub.com/doi/pdf/10.1177/030642209702600314.

Gentile, B., (2012), “Top 5 myths about big data” Retrieved from http://mashable.com/2012/06/19/big-data-myths/#MwZnjjrOR8qV.

Gitelman, L. (2013), *Raw Data” is an Oxymoron*, Cambridge, MA: MIT Press

Goes, P.B. (2014), “Big Data and IS Research”, *MIS Quarterly* 38, iii-viii.

Golio, M. (2015), “Fifty Years on Moore’s Law (Scanning our Past)”, Vol. 103, No. 10, Proceedings of the IEEE, 1932 – 1937.

Gonzalez, M.C., Hidalgo, C.A. & Baraba, A.L. (2008), “Understanding Individual Human Mobility Patterns”, *Nature,* 453, 779–82.

González-Bailón, S. (2013), “Social science in the era of big data”, *Policy & Internet,* 5 (2), 147-160.

Gray, J., Bounegru, L., Chambers, L. (2012), *The Data Journalism Handbook*, Sebastopol, USA: O’Reilly Media.

Gynnild, A. (2014), “Journalism Innovation Leads to Innovation Journalism: The Impact of Computational Exploration on Changing Mindsets”, *Journalism*, 15, 713–730.

Hamilton, J. T. & Turner, F. (2009). “Accountability through Algorithm: Developing the field of computational journalism”, Report from Center For Advanced Study in the Behavioural Sciences, Summer Workshop Retrieved from http://dewitt.sanford.duke.edu/images/uploads/About%203%20Research%20B%20cj%201%20finalreport.pdf.

Harvey, D. (1972), “Revolutionary and counter-revolutionary theory in geography and the problem of ghetto formation”, Antipode 4 (2), 1–13.

Henessy, M. (2018), “Why It’s Time to Jump on the Big-Data Bandwagon”, Retrieved from https://www.fsrmagazine.com/finance/why-its-time-jump-big-data-bandwagon.

Houston, B. (1996), *Computer-assisted Reporting: A Practical Guide*, New York: St. Martin's.

Howard, A. B. (2014), "The Art and Science of Data-driven Journalism", New York: Tow Center for Digital Journalism.

IBM (2012), "What is big data?" IBM Corporate Website, Retrieved from http://www-01.ibm.com/software/data/bigdata/.

Ingram, M. (2014), "Chartbeat gets certified to measure attention, tries to move advertising away from clicks and pageviews". Retrieved from https://gigaom.com/2014/09/29/chartbeat-gets-certified-to-measure-attention-tries-to-move-advertising-away-fromclicks-and-pageviews/.

Jao, J. (2013), "Why big data is a must in e-commerce". Retrieved from http://www.bigdatalandscape.com/news/why-big-data-is-a-must-inecommerce.

Kiron, D., Prentice, P. K., & Ferguson, R. B. (2014), "The analytics mandate" *MIT Sloan Management Review*, 55,1.

Klein, G., Moon, B. & Hoffman, R. R. (2006), "Making Sense of Sense-Making", *IEEE Intelligent Systems,* 21 (4), 70-73.

Kopp, M. (2013), "Seizing the big data opportunity, Ecommerce Times". Retrieved from http://www.ecommercetimes.com/story/78390.html.

Kosterich, A., & Napoli, P. M. (2015), "Reconfiguring the audience commodity: The institutionalization of social TV analytics as market information regime", *Television & New Media*, 17(3), 254–271.

Koutsabasis, P., Stavrakis, M., Viorres, N., Darzentas, J. S., Spyrou, T., & Darzentas, J. (2008), "A descriptive reference framework for the personalisation of e-business applications", *Electronic Commerce Research*, 8, 173-192.

Lamb, J. (2014), "Need for speed in data analysis", *Raconteur*, Retrieved from http://raconteur.net/technology/need-for-speed-in-data-analysis.

Lewis, S. C. (2015), "Journalism in an Era of Big Data", *Digital Journalism*, 3 (3), 321-330.

Lewis, S. C., & Usher, N. (2013), "Open Source and Journalism: Toward New Frameworks for Imagining News Innovation", *Media, Culture and Society,* 35 (5), 602–619.

Lewis, S. & Westlund, O. (2014), "Agents of Media Innovations: Actors, Actants, and Audiences", *The Journal of Media Innovations*, 1(2), 10-35

Lewis, S. & Westlund, O. (2015a), "Big Data and Journalism: Epistemology, Expertise, Economics, and Ethics." *Digital Journalism*, 3(3), 447-466

Lewis, S. C., & Westlund, O. (2015b), "Actors, actants, audiences, and activities in cross-media news work: A matrix and a research agenda", *Digital Journalism,* 3(1), 19-37

Liebowitz, J. (2013), *Big data and business analytics*, Boca Raton: CRC Press.

Lowrey, W. (2009), "Institutional Roadblocks: Assessing Journalism's Response to Changing Audiences", *Journalism and Citizenship*, New Agendas, edited by Zizi Papacharissi, 44–67. Mahwah, NJ: Lawrence Erlbaum Associates.

Lunden, I. (2013), "Forrester: $2.1 Trillion Will Go Into IT Spend In 2013; Apps And The U.S. Lead The Charge". Retrieved from https://techcrunch.com/2013/07/15/forrester-2-1-trillion-will-go-into-it-spend-in-2013-apps-and-the-u-s-lead-the-charge/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAHoX6Vx786sXLrBcsU5QQGnxCMwAXUGrg3pwu24oOVMzLVCZesfHVIG4rp5hMGctyTf1jgXNo1QqwFZdeNYS7aXoYsXbFJLc6TAl6obBVLjJtOuR7ZXE40N3uiTAxMulpGSaTTh3X2WGwxSWE6xSBD2tnBGneYNM79-kMW-qzUrp.

Mahrt, M. & Scharkow, M. (2013), "The value of big data in digital media research", *Journal of Broadcasting & Electronic Media,* 57(1), 20–33.

Manovich, L. (2012), "Trending: The Promises and the Challenges of Big Social Data." *Debates in the Digital Humanities*, edited by M. K. Gold, 460–475. Minneapolis, MN: The University of Minnesota Press.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., (2011), "Big data: The next frontier for innovation, competition, and productivity", *McKinsey Global Institute.*

Marr, B. (2016), *Big Data in Practice – How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*, TJ International Ltd.

Marshall, S. (2011), "10 Things Every Journalist Should Know about Data", *News: Rewired.* Retrieved from http://www.newsrewired.com/2011/04/26/10-things-every-journalist-shouldknow-about-data/.

Massey, B. L., & Levy, M. R. (1999), "`Interactive' Online Journalism at English-Language Web Newspapers in Asia: A Dependency Theory Analysis", *International Communication Gazette,* 61 (6), 523–38

Mayer-Schönberger, V. & Cukier, K. (2013), *Big Data: A Revolution that Will Transform How We Live, Work, and Think*, Boston, MA: Houghton Mifflin Harcourt.

McAfee, A., Brynjolfsson, E., (2012), "Big data: the management revolution", *Harvard business review*, 60-66, 68, 128.

McBride, R. E.D., (2016), "The Ethics of Data Journalism", *Professional Projects from the College of Journalism and Mass Communications*. 9. http://digitalcommons.unl.edu/journalismprojects/9.

McNair, B. (2006), *Cultural chaos: journalism, news and power in a globalised world,* London: Routledge.

Mehra, G., (2013), "6 uses of big data for online retailers", *Practical Ecommerce*. Retrieved from http://www.practicalecommerce.com/articles/3960-6-Uses-of-Big-Data-for-Online-Retailers.

Meyer, P. (1973), *Precision Journalism: A Reporter's Introduction to Social Science Methods*, Bloomington, Indiana University Press.

Miller, G., (2013), "6 ways to use "big data" to increase operating margins by 60 %". Retrieved from http://upstreamcommerce.com/blog/2012/04/11/6-ways-big-data-increase-operating-margins-60-part-2.

Miller, J. H. & Page, S. E. (2007*), Complex Adaptive Systems: An Introduction to Computational Models of Social Life*, Princeton, NJ: Princeton University Press.

Miller, L. C. (1998), *Power Journalism: Computer-assisted Reporting,* Fort Worth, TX: Harcourt Brace and Co.

Minkoff, M. (2010), "Bringing Data Journalism into Curricula", Retrieved from http://michelleminkoff.com/2010/03/24/bringing-data-journalism-into-curricula/.

Minsky, M. (1968), *Semantic information processing*, Cambridge: MIT Press.

Myers, Steve (2009), "Using Data Visualization as a Reporting Tool Can Reveal Story's Shape", Poynter. Retrieved from http://www.poynter.org/column.asp?id=101&aid=161675.

Nafus, D., and Sherman, J. (2014), "This One Does Not Go Up To 11: The Quantified Self Movement as an Alternative Big Data Practice." *International Journal of Communication* 8.

Napoli, P. M. (2012), "Audience evolution and the future of audience research", *International Journal on Media Management*, 14, 79 –97.

Napoli, P. M. (2014a), "Automated media: An institutional theory perspective on algorithmic media production and consumption", *Communication Theory*, 24(3), 340–360.

Napoli, P. M. (2014b), "On automation in media industries: Integrating algorithmic media production into media industries scholarship", *Media Industries*, 1(1), 33–38.

Napoli, P. M. (2016), "Special Issue Introduction: Big Data and Media Management", *International Journal on Media Management*, 18, 1, 1-7.

Netflix (2018, January 12) *David Letterman – My Next Guest Needs no Introduction.*

Noulas, A. Scellato, S. Reanud, L. Massimiliano, P. & Mascolo, C. (2012), "A Tale of Many Cities: Universal Patterns in Human Urban Mobility", *PloS ONE* 7 (5), e37027.

O'Brien, H. L., & Lebow, M. (2013), "Mixed-methods approach to measuring user experience in online news interactions", *Journal of the American Society for Information Science and Technology*, 64 (8), 1543–1556.

Oboler, A. Welsh, K. & Cruz, L. (2012), "The danger of big data: Social media as computational social science", *First Monday,* 17 (7-2).

Parasie, S. (2015), "DATA-DRIVEN REVELATION? Epistemological tensions in investigative journalism in the age of 'big data'", *Digital Journalism*, Routledge, 3 (3), pp.364-380. DOI: 10.1080/21670811.2014.976408.hal-01284731.

Parasie, S. & Dagiral, E. (2013), "Data-Driven Journalism and The Public Good: 'Computer-assisted-reporters' and 'Programmer-journalists' in Chicago", *New Media & Society*, 15 (6), 853–871.

Petre, C. (2013), "A Quantitative Turn in Journalism?", *Digital Journalism*, 3(3), 331-348

Pitt, F. (2014), "Sensors and Journalism", *Digital Journalism*, New York: Tow Center, Columbia University.

Powers, M. (2012), "In Forms that are Familiar and Yet-to-be Invented': American Journalism and the Discourse of Technologically Specific Work", *Journal of Communication Inquiry* 36 (1), 24–43.

Provost, F., & Fawcett, T. (2013), "Data science and its relationship to big data and data-driven decision making", *Big Data*, 1(1), 51-59.

Pulitzer, J. (1904), "Planning a School of Journalism – The Basic Concept in 1904." *The North American Review,* 178, 5.

Puschmann, C. & Burgess, J. (2014), "Metaphors of Big Data", *Big Data, Big Questions*, *International Journal of Communication*, 8, 1690-1709.

Ramaswamy, S. (2013), "What the Companies Winning at Big Data Do Differently", Bloomberg. Retrieved from http://www.bloomberg.com/news/2013–06-25/what-the-companies-winning-at-big-data-do-differently.

Rosenberg, D. (2013), "Data before the fact", in Gitelman, L. *"Raw data" is an oxymoron* (pp. 15–40). Cambridge, MA: MIT Press.

Russom, P. (2011), "The three Vs of big data analytics", *TDWI Best Practices Report, Fourth Quarter*, 18, 1–35.

Sathi, A. (2014), *Engaging customers using big data: how Marketing analytics are transforming business*, New York: Palgrave Macmillan.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012), *Analytics: The real-world use of big data,* NY, USA: IBM Institute for Business Value.

Schudson, M. (1978), *Discovering the News. A Social History of American Newspapers*, New York: Basic Books.

Silver, N. (2012), *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't*, New York, NY: Penguin.

Singer, J. B., Domingo, D. Heinonen, A., Hermida, A., Paulussen, S. Quandt, T., Reich, Z. & Vujnovic, M. (2011), *Participatory Journalism: Guarding Open Gates at Online Newspapers*, Malden, MA: John Wiley & Sons.

Stavelin, E. (2014), *Computational Journalism: When Journalism Meets Programming,* PhD diss., Norway: University of Bergen.

Stone, M. L. (2014), "Big Data for Media", **Reuters** *Institute for the Study of Journalism*, Oxford: Oxford University Press.

Strawn, G.O. (2012), "Scientific Research: How Many Paradigms?", *EDUCAUSE Review* 47, 26.

Stray, J. (2011), "A Computational Journalism Reading List". Retrieved from http://jonathanstray.com/a-computational-journalism-reading-list.

Stroud, N. J., Scacco, J. M., & Curry, A. L. (2016), "The Presence and use of Interactive Features on News Websites", *Digital Journalism* 4 (3), 339–358.

Suthaharan, S. (2014), "Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning", *SIGMETRICS Perform, Eval. Rev.* 41 (4): 70–73.

Tandoc, E. C. (2014), "Journalism is Twerking? How Web Analytics is Changing the Process of Gatekeeping", *New Media & Society*, 12, 1085–1102.

Taylor, M (2010), "How Journalists Can Incorporate Computational Thinking into Their Work", *Poynter*. Retrieved from http://www.poynter.org/column.asp?id=31&aid=187439.

The Economist, (May 6th, 2017), "Regulating the internet giants – The world's most valuable resource is no longer oil, but data". Retrieved from https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

Tryon, C. (2015), "TV Got Better: Netflix's Original Programming Strategies and Binge Viewing", *Media Industries Journal,* 2 (2), 104-116.

Webster, J. (2012), "Big data deserves IT's attention", *Computerworld*. Retrieved from http://www.computerworld.com/s/article/357189/Big_Data_Deserves_IT_s_Attention?axonomyId=11

Webster, J. G., Phalen, P. F., & Licthy, L. W. (2014), *Ratings analysis: Audience measurement and analytics*, 4th edition, New York: Routledge

Wing, J. M. (2008), "Computational Thinking and Thinking about Computing", *Philosophical Transactions*, Series A, Mathematical, Physical, and Engineering Sciences 366, 3717–3725.

Xu, Z., Frankwick, G. L., Ramirez, E. (2015), "Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective", *Journal of Business Research,* 69, 1562 – 1566.

Yakabuski, K. (2013), "Big Data should inspire humility, not hype", *The Globe and Mail*, Retrieved from https://www.theglobeandmail.com/opinion/big-data-should-inspire-humility-not-hype/article9234569/.

Yarnall, L., Johnson, J. T., Rinne, L. & Ranney, M. A. (2008), "How Post-Secondary Journalism Educators Teach Advanced CAR Data Analysis Skills in the Digital Age", *Journalism and Mass Communication Educator,* 63, 146–164.

Young, M. L., & Hermida, A. (2015), "From Mr. and Mrs. Outlier to Central Tendencies: Computational journalism and crime reporting at The Los Angeles Times", *Digital Journalism*, 3(3), 381-397.

Young, M. L., Hermida, A., & Fulda, J. (2018), "What Makes for Great Data Journalism?", *Journalism Practice*, 12, 1, 115-135.

Appendices

1. Transcript of ML's interview, recorded on April 29th 2020

VL: O que é que entende pela datificação dos media e porque é que é importante?

ML: Os media sempre trabalharam, sempre foram meios muito pouco quantitativos até haver medições. Desde que há medições na Internet e estando os meios dependentes de receitas publicitárias, os anunciantes enquanto investidores que são querem medir, querem , são adversos ao risco e portanto funcionam com base em decisões analíticas. E portanto, sendo a internet uma coisa altamente mensurável e analítica, segue um caminho, segue com toda a velocidade que caracteriza a internet, o caminho que a televisão já seguiu, que a rádio tenta seguir, mas não é assim tão mensurável, tão bem como a televisão que tem audímetros, e a tecnologia em Portugal não é baseada em audímetros, e a imprensa ainda é mais mal-medida, com um outdoor ainda mais mal-medida, e portanto a datificação da internet advém do meio ser muito mensurável e de os anunciantes e do mercado publicitário só trabalharem com medições objectivas.

VL: Portanto é uma forma de mensurar melhor aquilo que é feito nos media…

ML: E havendo capacidade de uma internet saber quem está do lado de lá, o produto pode ser mais afinado à pessoa que está do lado de lá, antigamente e havia jornais nacionais com uma capa que era igual para toda a gente, depois começou a haver canais de tv, agora há milhentos canais de tv e vamos ao youtube e vemos lá milhões de canais de youtube. Há uma capacidade através do meio digital de customizar o conteúdo, a publicidade, e de adaptar aquilo que cada consumidor vê que é diferente do consumidor do lado e aí a internet e o digital são um meio muito mais revolucionário do que todos os outros, mas todos eles foram sempre neste sentido.

VL: E porque é que a Global Media decidiu fazer parte do NÓNIO?

ML: Mandei-lhe uma apresentação por email do economista chefe do IAB (centro internacional de publicidade) e vai perceber lá… Ou este tipo de coisas como o NÓNIO

é feito por publishers europeus ou então acabou. Fica tudo para o Google e para o Facebook. Mas aí no Power Point que lhe enviei está tudo explicadinho.

VL: Não sei se estará também neste power point… gostava de saber que tipo de perfilagem é que é feita através do NÓNIO e que tipo de tratamento de dados é que o NÓNIO permite fazer.

ML: Sabe que existe um site do NÓNIO? As coisas que medimos e perguntamos estão lá nos termos e condições e na política de privacidade e pode ver no site, portanto, a segmentação que formos fazendo é uma segmentação em função daquilo que os consumidores permitem vs aquilo que o mercado publicitário procura vs aquilo que nós enquanto detentores da ligação com o consumidor faz sentido comercializar. O que é que quero dizer com isto? Se for perguntar à BMW, a BMW claro, só quer um segmento híper-fino de gente muito rica da Quinta da Marinha. Mas sei que só tenho 2 ou 3 pessoas, portanto não me compensa vender o segmento de hiper-ricos da Quinta da Marinha, por abstracção. Obviamente que à BMW também não interessam pessoas com mais de 80 anos que nem têm carta de condução. Cabe-nos a nós media arranjar segmentos que sejam mais apelativos para os anunciantes do que propriamente o target total e os alvos totais vs uma dimensão que a nós nos permita viabilizar o nosso negócio e ter um mínimo de volume, porque os anunciantes não fazem campanhas para 2 pessoas, tipicamente são para milhares ou centenas de milhares de pessoas. Portanto, há sempre um equilíbrio entre o que o mercado nos dá, a frequência com que as pessoas vêm cá consumir os nossos conteúdos e o que os anunciantes querem comprar e estão dispostos… e o preço que estão dispostos a pagar.

VL: Pegando nesse caso dos hiper-ricos da Quinta da Marinha, o NÓNIO permite descobrir que parte das pessoas ouvem a TSF ou vão ao site da TSF e se são da Quinta da Marinha?

ML: Nesta altura, não… "Nim". Se eu sei que uma pessoa que só consome conteúdos de um determinado perfil que são conotados com classes médias altas e com um grande poder económico eu consigo, seja perguntando às pessoas se são ricas ou não, e o rendimento é sempre uma coisa muito delicada de se perguntar às pessoas, mas sim, poderíamos, mas o NÓNIO não está preparado para isso agora, mas, sim, no limite até

podemos pôr um pop-up a perguntar qual é o seu rendimento. Está estudado em research que as pessoas não declaram... é a variável que menos têm disposição para responder, é o rendimento, não gostam. Agora, outro tipo de questões que são perguntadas directamente, a first party data. Que educação é que tem, etc…, ou onde é que habita… nós conseguimos chegar lá. Há outra coisa que nos permite fazer isso que é pelo consumo de conteúdos das pessoas, é muito fácil perceber que tipo de pessoa é que está por trás. Vou exagerar para que perceba, só vai ao Expresso ler artigos de macroeconomia e de mobiliário de alta gama um determinado tipo de perfil. Não vão as tais senhoras de Trás-os-Montes de 80 anos. Há uma série de inferências que nós conseguimos saber. Trata-se de saber depois se queremos fazer essas segmentações ou não. Mas aqui para nós, aquilo que temos planeado no NÓNIO no curto prazo são segmentos muito menos sexys, são tranches de segmentos de idade combinadas com género e se calhar, e mesmo isso não está definido, alguma outra variável. Por exemplo, uma potencial apetência por desporto, ou por carros novos, ou por life style. São segmentações muito mais básicas do que aquela coisa de chegar quase ao quarteirão da pessoa e de saber quem ela é.

VL: A malha ainda é muito larga?

ML: O Facebook permite esse tipo de cosias. Se o Vicente for fazer uma campanha, pegar no seu cartão de crédito e gastar 1€ ou 5€, vá fazer uma campanha ao Facebook, o Vicente consegue chegar à pessoa naquele quarteirão. Nós não temos necessidade de o fazer. Nem há mercado em Portugal para chegar lá.

VL: Como é que o NÓNIO está a alterar o modo como a Global Media faz publicidade, mas se calhar não é bem estar a alterar, é mais simplesmente para poder fazer concorrência com as Googles e Facebooks deste mundo. É por aí, não é?

ML: Os anunciantes querem comprar segmentos de pessoas, mais do que capas ou primeiras páginas. E portanto esta é a forma dos media tradicionais se adaptarem aos segmentos que são vendidos pelo Google, pelo Facebook, e que os anunciantes pretendem no dia-a-dia. Eu quero vender um carro que foi feito para mulheres, só quero chegar a mulheres, não quero que mostrar o carro a homens, por exemplo. Portanto, é isso que os grandes meios digitais já dão e é isso que nós também temos de dar.

VL: Ainda na prática comercial associada ao NÓNIO, mas afastando-nos da publicidade, como é que o NÓNIO pode vir a afectar… no caso da TSF não é bem o modelo de subscrições, não é? Mas como é que pode influenciar as pessoas que vão ver o site da TSF e que continuam a ir… como é que do NÓNIO pode…

ML: O NÓNIO é óptimo no processo de captar um subscritor. Porquê? Porque não há nenhum sistema de subscrição no mundo que não obrigue à identificação da pessoa subscritora. As pessoas têm de pagar, têm de dizer quem são. As pessoas para aceder ao conteúdo que pagaram, as pessoas têm de se registar, têm de se logar, e portanto o que o NÓNIO faz e faz de uma forma massificada porque tem outros esquis para além das assinaturas é pedir às pessoas que se registem, tal como o Faceook e o Youtube e o Gmail pedem às pessoas para se registarem. Tal como acontece se comprar um iPhone, a primeira coisa que vai acontecer é ser-lhe pedido que se registe numa conta Gmail. Também nós pedimos às pessoas que se registem. Portanto, o NÓNIO como um Single Sign On que é, não é mais nem menos do que o primeiro passo no processo de venda de conteúdos. Portanto, o NÓNIO é uma coisa boa para se conseguir identificar os consumidores e se eu os chatear para eles comprarem quando a gente percebe que vêm cá algumas vezes e que são pessoas que são interessadas. Um tipo que vem do Facebook e cai cá porque encontrou uma notícia ou porque vem de uma pesquisa no Google sobre bolo de chocolate tem muito menos propensão para ser assinante do que alguém que está registado e que nós sentimos que vem cá quase todos os dias à mesma hora. Esses são o tipo de pessoas que nos interessam convidar mais para que subscrevam os nossos conteúdos, porque sabemos que têm alguma apetência e preferência pelas nossas marcas. Portanto, o registo único e o Single Sign On do NÓNIO ajudam a que esse entendimento de quem são os consumidores seja melhor feito por nós. Podemos submeter as propostas de assinatura às pessoas certas nos momentos certos.

VL: E em termos de produção de conteúdos, acha que pode ter impacto? Imaginando que descobrem um grupo de leitores com muito interesse no Ronaldo, acha que isto pode levar a uma maior aposta na produção de conteúdos do Ronaldo?

ML: O NÓNIO per si não. Em toda a história dos meios, nunca nenhum meio sobreviveu sem escutar os seus leitores, na história do jornalismo e dos media sempre houve muita produção de conteúdo que as pessoas não estão à espera e não pedem, mas também há

uma parte em que é preciso estudar quem são as pessoas, quem são os consumidores, para perceber quais são os conteúdos que as pessoas procuram e querem ver. É sempre um misto das duas coisas. Mas o NÓNIO não tem nada a ver com isso. Os jornalistas todos os dias sabem quais foram os conteúdos mais ou menos consumidos no dia anterior, porque temos Google analytics. Na tv todos os dias há ratings. Isso é uma coisa absolutamente banal que não tem nada a ver com o NÓNIO.

VL: Mas e se descobrir por exemplo que o João gosta de conteúdos do Cristiano Ronaldo e que ele faz uma subscrição convosco. Acha que os conteúdos a que o João tem acesso quando entra no vosso site ou nos conteúdos que recebe por email já são mais dirigidos? Acha que o NÓNIO pode ajudar a dirigir conteúdo para determinadas audiências?

ML: Está a falar de uma teoria da personalização do conteúdo para as pessoas em função de quem elas são. Sendo certo que essa questão é uma questão muito volátil porque tem muito a ver com regras de privacidade e com aquilo que aceitam ou não aceitam, todos os meios digitais fazem de alguma forma a personalização de conteúdos independentemente de terem NÓNIO ou não. Todos os meios portugueses têm software de recomendação de conteúdo e de personalização de conteúdo feito por empresas em Israel ou em Portugal, como a Priberam, SIC, Cofina, todos eles independentemente do NÓNIO têm ferramentas de personalização de conteúdo. Sim, se souber quem é a pessoa, e o grande mérito disto, sem entrar em questões de privacidade, se eu souber que o senhor se registou e já viu 5 conteúdos, abstenho-me de lhe mostrar 5 conteúdos que já viu, mostro-lhe fruta fresca, para ver se de cada vez que cá vem não está sempre a ver coisas que já viu. Sem entrar em questões de privacidade, se soubermos quem é a pessoa, sim, podemos personalizar conteúdos. Há sempre uma grande discussão no mundo jornalístico: até que ponto é que a personalização tem de ser manual ou automática, tem de ser dado a um algoritmo ou a um editor humano, mas isso são grandes questões filosóficas que há uns que tentam mais ou menos responder… O Google e o Facebook fazem isso com algoritmos. Nós fazemos de uma forma algo manual e com algum critério editorial. Eu diria que não pode ser nem tanto ao mar nem tanto à terra, mas essa tecnologia de recomendação existe e a gente usa-a. Já usamos todos independentemente do NÓNIO.

VL: Acredita que o NÓNIO e práticas de Big Data podem levar à sobrevivência económica dos media?

ML: Pode, pode. Sim, sim. É tão fácil de explicar como isto: se eu sei quem é a pessoa, os anunciantes pagam muito mais por ela. Porque estão dispostos a não desperdiçar dinheiro a impactar pessoas que não lhes interessa. É o tal exemplo da Quinta da Marinha e dos idosos de Bragança. Se sei quem é a pessoa, o anunciante, que é adverso ao risco, tem mais confiança em fazer investimentos comigo. Hoje faz esses investimentos com o Facebook que lhe dá esses segmentos e eu tenho que lhe dar esses segmentos para conseguir que ele tenha confiança em mim e que invista comigo.

VL: Se acredita, acha que os consumidores estão dispostos a trocar dados pessoais, a trocar informação privada sobre quem são, em troca de algo que lhes acrescente valor, neste caso notícias?

ML: Sim. Da mesma maneira que há 6 milhões de portugueses registados no Facebook e todos registaram-se no Facebook sem pestanejar. Da mesma forma que 80% dos telemóveis em Portugal são androids e para se activar um android é preciso ter uma conta no Google e toda a gente activou. O que nós temos de fazer e provavelmente não temos explicado isso bem, é esclarecer que para conseguir ter conteúdos portugueses feitos por jornalistas em Portugal, temos de pagar ordenados, e portanto o trade-off é as pessoas registarem-se para conseguirmos estar abertos. Se não, podem ir ver notícias a sites de fake news que são de borla. O jornalismo paga-se, as pessoas têm de receber ordenado. Não é a receber umas migalhas do Facebook que se conseguem pagar ordenados.

VL: Mas quando falou do Facebook, usou uma expressão interessante: se calhar não estamos a explicar bem às pessoas. Diria eu, quando uma pessoa se regista no Facebook ou no Gmail não tem consciência que o Facebook ou a Google estão a monitorizar tudo o que fazem quando usam essas plataformas. Mas acha mesmo assim que, se souberem que isso está a acontecer, estão dispostas a dar essas informações em troca de…

ML: Nós damos sempre a hipótese de as pessoas pagarem, mas está provado que a enormíssima maioria das pessoas prefere ver publicidade para não ter de pagar. Nós damos sempre a opção da pessoa poder pagar, não ver nada, não haver tracking sobre a

pessoa, mas a experiência o que nos diz é que isso é tudo muito bonito. Os maiores detractores do NÓNIO são os jornalistas. Quando viram o New York Times ou o Washignton Post a fazerem exactamente a mesma coisa que o NÓNIO está a fazer, já ficaram mais calmos. Mas eram os grandes inimigos do NÓNIO, porque a privacidade isto e aquilo, e nós fartámos-lhes de explicar que o NÓNIO não fazia mais do que o Facebook e o Google faziam. Não percebemos qual é o grande problema sendo que o que estamos a fazer é conseguir receitas para manter as coisas locais em Portugal. Este é um problema que os jornais no mundo inteiro estão a ter. As pessoas têm de perceber que ou querem pagar conteúdos ou aceitam não os pagar e a gente tem de viver de publicidade. A publicidade quer saber quem são as pessoas que estão do lado de lá. É esse trade-off que temos de explicar. Estamos num país onde toda a gente pirateia jornais e não faltam sites de tudo e mais alguma coisa a piratear o jornal de ontem e de ontem. O português rouba conteúdos, pirateia conteúdos, portanto esta conversa do lirismo de vender conteúdos, a gente gostava que isto fosse como a Suécia e como os países escandinavos e que houvesse histórias de sucesso como o Washington Post ou o New York Times, mas estamos em Portugal. Com o nível de pirataria que a gente tem cá, nós vamos ter sempre de ser capazes de explicar que ou as pessoas dão os seus dados para a publicidade permitir que se pague ordenados ou vão ver as notícias ao New York Times ou a sites de Fake News. Nós não conseguimos pagar ordenados. Mas nós temos alguma dificuldade em explicar isto mesmo aos nossos jornalistas, que são sempre uns idealistas e acham que o dinheiro cai das árvores. Neste momento é preciso pagar ordenados e neste momento nos grupos de media em Portugal ninguém sabe muito bem como é que se vão pagar ordenados. As questões da privacidade do NÓNIO, isto é tão simples como isto, se estivéssemos na Escandinávia… as pessoas aí estão dispostas a pagar por conteúdos. Em Portugal não estamos. Em Portugal, os meios, as rádios e as televisões são gratuitas. O que temos de fazer aqui é arranjar maneira de juntar a oferta à procura.

2. Transcript of TC's interview, recorded on May 5th, 2020

VL: O que é que entende pela datificação dos media e porque é que é importante?

TC: A datificação dos media… Ou seja, estamos a falar de utilização da data nos media, não é? Hoje em dia, nós entramos cada vez mais num mundo mais globalizado e que cada vez mais procura encontrar ganchos de relevância para chegar às audiências e para comunicarmos com os nossos públicos-alvo. Já passámos há muitos anos aquele sistema do broadcasting em que difundimos uma coisa para toda a gente e quem quer quer, quem não quer é o que tem. Hoje, temos muita informação ao nosso lado e temos de aumentar essa elevância para as pessoas. Quanto mais os media souberem a quem é que podem chegar, como é que têm de formatar os conteúdos que distribuem, para serem mais relevantes para as pessoas, isso faz parte da sua estratégia de competitividade uns com os outros. Quando começamos a ver um mundo cada vez mais tecnológico começam a abrir-se essas ferramentas para encontrarmos esses, essa, para afinarmos as estratégias de comunicação e relevância para chegarmos aos públicos-alvo, não é nada assim de muito sofisticado. A tecnologia veio permitir-nos isso e isso acaba por estar ao serviço dos media e não só, mas do negócio em geral, para chegarmos às pessoas certas no momento certo.

VL: O que é que é o NÓNIO?

TC: O NÓNIO, de uma forma resumida, porque acaba por ser uma coisa mais complexa, é uma aliança entre grupos de media nacionais que como individualmente têm mais limitações pelo facto da sua escala e conseguir combater com soluções um bocadinho data driven que os players globais têm, formar uma aliança para que com essas sinergias, mantendo as suas identidades e a sua concorrência interna e as suas estratégias comerciais individualizadas, conseguem ter sinergias e alguma união que como um todo consigam criar essas vantagens competitivas, ou pelo menos não perdê-las, em relação àquilo que são os gigantes globais. Porque hoje em dia é o Facebook e a Google parece que são os mais emblemáticos que têm muita informação nossa. A questão do data está na ordem do dia por causa disso. Só que o Google e o Facebook têm uma vantagem que nós fornecemos toda a informação sem qualquer tipo de prurido e estamos confortáveis com isso… E o NÓNIO tem de recorrer a outras estratégias porque as pessoas não têm uma pré-disposição tão óbvia para partilhar informação com grupos de media nacionais, quer dizer porquê, não estão habituadas, é um choque cultural, demora algum tempo a interiorizar. Como é que o NÓNIO materializa isto? Cria um sistema de login único para todos os publishers e nesse sistema de login acaba por cadastrar de alguma forma e

tipificar aquilo que são as audiências. Não está em causa o que é a privacidade porque não sabemos se é o Manuel se é o João, isso nem queremos saber. Queremos é saber se estamos a falar com um homem ou com uma mulher, a faixa etária e depois num segundo momento dá para inferir até comportamentos típicos que é aquilo que as redes de advertising do Google e do facebook fazem. Se nós consumimos muita informação sobre automóveis, esse data acaba por ser processado e passamos a fazer parte de uma pool que é uma audiência qualitativa para uma campanha de comunicação do lançamento de um automóvel. É afinar interesses e de relevância para as próprias pessoas porque nós temos de viver com publicidade. Embora esta não seja a realidade que represento directamente, a lógica da imprensa, nós enquanto consumidores, durante toda a vida íamos à banca comprávamos o jornal, pagávamos o preço do jornal e tínhamos publicidade lá dentro. Publicidade faz parte dos conteúdos. Vemos televisão e temos publicidade. Vamos à internet e temos publicidade. Publicidade é um trade-off necessário porque é uma forma de suportarmos os conteúdos e quem produz esses conteúdos sem que tenhamos de pagar por eles ou pelo menos na sua totalidade. O modelo da imprensa era um modelo claro, nós pagamos um jornal e depois temos a publicidade porque se não tivéssemos publicidade o jornal teria de ser mais caro. Não pagando o jornal, a publicidade tem de gerar mais receita. Quando nós começamos a consumir as coisas na internet, e ainda utilizando o exemplo da imprensa porque é o exemplo mais óbvio, quando nós passamos isto para o online nós habituamo-nos enquanto consumidores a que tudo no online seja gratuito, mas na prática, se queremos produzir conteúdo profissional e partindo da premissa que o nosso trabalho é de borla, temos de ter fontes de receita para esses conteúdos.

VL: Acha que os consumidores estão dispostos a trocar informação, dados pessoais, por algo que lhes acrescente valor, neste caso informação, notícias. Se acha que aos olhos do consumidor esta é uma troca justa?

TC: Essa é a one million dollar question. Podemos ter informações diferentes. Eu acho que o caminho pode ser esse. Não é um trade-off fácil de aceitar se calhar culturalmente. Pelo menos no primeiro momento, mas nós sabemos que depois destes modelos ganharem alguma maturidade, as coisas acabam por poder vir a ser mais bem-aceites. Numa primeira fase há sempre resistências para esses modelos.

VL: NÓNIO consiste essencialmente numa forma nova de vender publicidade aos anunciantes. Permite conhecer melhor os clientes, permite segmentos mais finos, quando é feita essa publicidade. O que queria perceber é qual é a prática comercial associada ao NÓNIO, além da publicidade. Como é que pode afectar o modelo de subscrições?

TC: Neste momento partilhamos daquilo que é a estratégia do NÓNIO, que é a partir do segundo artigo… ou seja, entramos, se queremos continuar a consumir artigos somos convidados a registarmo-nos, é um modelo comum a todos os sites do NÓNIO. Mas não temos subscrições, ou seja, as nossas plataformas digitas são representações daquilo que as marcas representam on air. Difundimos conteúdos em canal aberto e no NÓNIO, no caso das nossas marcas. Estarmos a pedir registos no online já é um passo à frente daquilo que á a natureza das marcas do grupo. Quem tem um título ou uma revista ou um jornal pelo qual a pessoa paga, por exemplo, se quiser pagar o Expresso vai pagar o Expresso, sabe que quando vai para o online tem de fazer um registo, mas também não tem de pagar o Expresso, ou pode pagar para ter outros conteúdos. Mas é um modelo mais fácil de entender. Uma televisão em sinal aberto ou uma rádio, pois, não há grande margem para haver subscrições no online, ou seja, no limite pedimos apenas o registo para podermos ter uma experiência melhor par ao utilizador, obviamente que isto…

VL: O que é que isso quer dizer? Ter uma melhor experiência? Quer dizer que pode afectar a produção de conteúdos ou mudar aquilo que o consumidor vê quando entra no site?

TC: Em teoria sim. Mas na prática é muito difícil, ou seja, requer alguma maturidade. Ou seja, algures no tempo isso poderá ser um caminho. Estamos a começar agora. Mas isso numa fase inicial não vejo que haja muita viabilidade, sobretudo porque nós por vezes temos a iniciativa e não somos piores do que o que vemos lá fora, de todo, só que temos sempre aquele problema de escala. Ou seja, isto só é viável darmos esse passo quando as cosias assumem uma escala que, mesmo no seu pleno, Portugal por si só já é pouca. Mas sim, numa lógica de comunicação, benefício e trade-off para o utilizador, essa podia ser, a personalização de conteúdos, a forma como arrumamos os conteúdos para si ser diferente da forma como arrumamos para outra pessoa, se tivermos comportamentos diferentes, isso podia ser um bom argumento. Está implícito, mas ainda não está em prática.

VL: Acredita que o NÓNIO ou as práticas de Big Data no geral podem levar à sobrevivência dos media? Sabemos que o sector é precário e que enfrenta grandes dificuldades. Acha que isso pode ajudar os meios de comunicação social?

TC: Pode ajudar todos os negócios. É uma questão relativamente genérica. Ao dia de hoje, nós temos muitas convicções, mas também não temos bolas de cristal. Sabemos que, de forma muito pragmática, se todas as coisas que vamos fazendo nos vão permitir fazer crescer o negócio, ter um impacto muito significativo nos negócios sejam eles quais forem, não sabemos em rigor. Sabemos é que se não fizermos nada e continuarmos no passado e sem este, sem acompanhar estas tendências, de certeza que morremos. Agora o digital e o Data e todo este novo mundo ligado aos media não é diferente das empresas que têm CRM. De conhecerem o cliente. Das bases de dados. O digital não veio trazer… em termos de conceitos, tudo o que existe no digital são coisas que já tínhamos, mesmo empresas tradicionais que têm um modo de operar completamente offline, as que trabalham bem já têm sistemas de conhecer os clientes, saber os comportamentos, saber uma série de cosias, esse data não é uma cosia exclusiva do digital. O digital veio é trazer isto para um patamar completamente diferente. Os bits e bytes começaram, permitiram que isto tivesse uma escala completamente diferente, mas conceptualmente aquilo que as coisas representam e os impactos que têm nos nossos negócios, conceptualmente é o mesmo, o digital só veio acelerar muito e facilitar e veio permitir fazer-se muito mais sobre os mesmos conceitos.

VL: Que outras práticas de Big Data é que a Renascença tem para além do NÓNIO?

TC: Que outras práticas… não sei se… nós, a Renascença dentro daquilo que é o grupo do NÓNIO, e não digo isto no sentido pejorativo, mas se calhar é o menos sofisticado. Temos um modelo muito mais simples. Até há pouco tempo, e somos os membros mais recentes do NÓNIO, temos uma operação digital com para aí 2 anos, que é há quanto tempo estou na Renascença, antes disso os sites estavam entregues à comercialização da rede SAPO, e o foco era exclusivamente produzir conteúdos na net, com alguns eventos, com algumas interacções integradas, mas o digital aparecia ali, não havia uma estratégia específica para o digital. Ou seja, o nosso grupo tem uma estratégia mais assertiva desde há 2 anos. Ficou a representar comercialmente os próprios sites e eu que estava no SAPO

fui para a Renascença e entretanto tempos também um director digital em termos que trata de outra área que é a transformação do digital, mas isso é um processo que começou há 2 anos. Big Data… é se calhar demasiado para o estado/ponto em que estamos. O nosso Big Data é NÓNIO. É o nosso envolvimento no projecto.

VL: Que problemas éticos é que podem advir do NÓNIO e como é que se resolvem?

TC: Aquilo que o NÓNIO representa, e se for ao twitter pesquisar NÓNIO pode ver uma série de coisas engraçadas, mas aquilo que o NÓNIO recolhe de informação é uma ínfima parte do que aquilo que Googles e Facebooks e por aí fora sabem de nós. Problemas éticos, nós usarmos a informação que temos quase no limite das possibilidades, se calhar não íamos sequer chegar perto daquilo que os gigantes tecnológicos sabem sobre nós. Acho que isso nem sequer é tema. É tema na cabeça das pessoas efectivamente. É um receio legítimo, é um receio também pouco reflectido porque muitas vezes mesmo as pessoas e às vezes vemos isso nos nossos círculos de amigos "ah o registo no NÓNIO, dar aqui a minha informação, mas porquê", mas são as mesmas pessoas que pegam num Facebook e no Instagram e expõem lá a sua vida toda. Não tem mal. Não há mal nenhum em fazerem isso. A questão é que a excessiva resistência de um lado e depois a utilização das redes sociais em todo o seu esplendor do outro é que não combina. Porque na prática nada daquilo que seja a informação e dos nossos comportamentos de consumo que são tratados por algoritmos e em larga escala não são restringidos ao indivíduo A ou indivíduo B… problemas éticos… nem sequer temos plataformas para incorrer em plataformas éticos.

VL: Os algoritmos são programados por humanos.

TC: Os algoritmos são feitos de interesse para quem os programa. Ou seja, são feitos por pessoa. O algoritmo é sempre aquela meia entidade, parece que uma entidade externa que valida tudo. Os algoritmos são feitos por pessoas e são feitos em favor de quem os faz. Os algoritmos não são feitos para prejudicar quem os faz.

VL: Precisamente no seguimento do seu raciocínio. Acha que as tendências humanas, os "human biases" como se diz em inglês, se podem de alguma forma manifestar nos

algoritmos do NÓNIO? Se acha que os meus próprios preconceitos podem aparecer no algoritmo? Acha que isto é possível ou que não faz sentido nenhum?

TC: Não é aplicável porque não temos, somos empresas que queremos acreditar que somos tecnológicas, mas nós não temos a capacidade tecnológica de desenvolver algoritmos… nós basicamente trabalhamos com tecnologia e com plataformas tecnológicas de referência e não estamos no campeonato de desenvolver algoritmos. Podemos fazer set-up de critérios, mas esse exemplo… podíamos comercialmente apontar para determinadas pessoas com o pressuposto de aquelas pessoas faziam parte de uma determinada tipologia social, como por exemplo a raça. Isso poderia acontecer em termos de segmentação, de segmentações de perfis, mas a cor da pele nem sequer é critério… nem sequer temos essa informação. Mas se o tivéssemos, poderia ser feito, na criação de segmentos. Os media têm sempre esses preconceitos. Por isso é que nós cada vez mais, e a parte do comportamental ganhou um bocadinho ao sócio-demográfico, porque é aquela coisa do anunciante quer promover detergentes para a roupa quer atingir mulheres. Está a assumir que quem usa detergentes para a roupa são mulheres da idade tal para a idade tal. Isso vai sempre de encontro aos preconceitos. Se formos muito rigorosos em relação aos preconceitos, o público-alvo da maior parte das campanhas de determinados produtos envolve sempre preconceitos. Os detergentes querem comunicar com mulheres, os carros querem comunicar com homens, e isso cada vez mais é discutível. Por isso, essa questão, faz parte dos media. Não é uma questão de, não tem necessariamente a ver com o NÓNIO. Obviamente, se o NÓNIO der mais ferramentas na segmentação das audiências e definição dos targets da campanha, se houver preconceitos à nascença mais tecnologia podia aumentar a utilização desses preconceitos. Mas esses preconceitos estão na base dos media.

VL: Acredita que o NÓNIO pode ser um primeiro passo na automação dos media? Se isto levar ao desenvolvimento de algoritmos que gerem notícias de forma autónoma como o New York Times ou o Wall Street Journal têm? Ou seja, acha que Big Data e o NÓNIO podem levar a este futuro em Portugal?

TC: Se o Big Data pode, pode. Eu tenho muitas reservas. O Big Data… tenho muitas reservas em Portugal. O NÓNIO… porque é uma questão de escala. Por muito que a tecnologia evolua, nós não temos escala para que as coisas depois façam… tem sempre

de haver um equilíbrio entre esforço-resultado. O NÓNIO nem sequer tem essa ambição de automação. Acho que ninguém vai querer… ninguém vai lutar por isso, porque a automação, ou seja, vamos ficar sem jornalistas? Isso não faz sentido. Aliás a automação que temos nas nossas actividades, se pensarmos bem, estamos quase a trabalhar para ser irrelevantes. Estamos a substituir-nos a nós próprios. Mas acho que isso também não é uma estratégia de sucesso porque também não sei se do outro lado é isso que queremos enquanto consumidores.

3. Transcript of MM's interview, recorded on May 11th 2020

MM: As práticas de Big Data estão a influenciar o jornalismo pela negativa em Portugal. Em muitas partes. Qual é o tema da Big Data e o que é que pode ser um engodo em relação ao Big Data. Big Data assenta numa tecnologia que são cookies, ou seja, ficam nos browsers dos computadores… uma tecnologia que foi criada essencialmente para perceber a navegação que é feita dentro dos sites e no fundo serve para recomendar conteúdos que sejam mais relevantes conforme a navegação de determinado utilizador. Utilizador que na verdade é um browser. O NÓNIO tem muito a ver com isso que é tentar perceber qual é o ID de uma pessoa, dentro do limite que é a privacidade de utilização desta informação que pode ser pessoal ou não, comparado com estas tecnologias de cookies, os IDs do Google e do Facebook têm muita informação e na realidade recomendam conteúdos com base na navegação. Sei que muitas vezes aquilo é meio esquizofrénico. O que é que eu acho que isto pode afectar o que é o verdadeiro jornalismo? Jornalismo assenta num princípio que temos um curador, um emissor ou um jornalista que recomenda as notícias e os conteúdos que são relevantes. Ora o tema da Big Data para já… sinto-o como incapaz de recomendar o que quer que seja… a não ser que seja uma trend, ou seja, o conteúdo que está a ser muito consumido e então é mostrado no meu portal ou no meu site ou no meu feed no Facebook ou no Instagram. Mas na realidade, a nossa experiência o que nos mostra é que o nosso feed do Facebook ou Instagram, se eu não seguir e não consumir muito dos conteúdos que estão na TVI24 ou na SIC ou no Jornal de Notícias ou no Expresso, na realidade vou estar a receber conteúdos que são ou recomendados por uma black box que é de um Facebook e propriedade de um Google ou de um Instagram ou do que quer que seja, de um DMP

(Data Management Platform) vai recomendar determinado conteúdo que tem que ver com os meus gostos pessoais e com a minha experiência antiga de navegação. Tem a ver com o passado e não com um evento que seja relevante para um jornalista que está a investigar um determinado evento no mundo ou no país. O papel do curador e do humano deixa de existir aqui. Tu vais passar a receber a informação que será relevante para uma máquina que é agnóstica ou que é relevante para a humanidade. Ou seja, temos aqui um problema face ao que é que é percepcionado como relevante e qual é o papel do jornalismo quando temos toda a humanidade a produzir conteúdo e a dizer que… e na realidade a produzir notícias que não são conteúdos jornalísticos. Digo que não são conteúdos jornalísticos porque não verificaram o contraditório, não foram verificadas fontes e podem estar a publicar uma fake news. O problema da Big Data assenta muito nisto. Por isso é que comecei por dizer que este é um risco para o jornalismo mais sério. Na realidade é jornalismo vs o que pode ser feito com isso (isso = Big Data).

VL: O que é que é o NÓNIO e porque é que a Media Capital decidiu fazer parte deste projecto?

MM: Há 2 temas principais. 90% da receita do nosso negócio… a Media Capital tem uma parte substancial da receita na distribuição aérea do sinal da TVI, depois a TVI24 e depois tem uma componente digital de produção própria de conteúdos de jornalismo e de entretenimento e tem um conjunto de sites. A generalidade da receita depende da venda de publicidade. E a venda de publicidade digital tem sido dominada nos últimos tempos pela venda do Facebook e da Google. A Google e o Facebook nos últimos anos conseguiram junto do mercado de anunciantes e das marcas que compram publicidade, junto dos vários stakeholders da compra de publicidade e da escolha de onde deve ser servida minha peça criativa, quero vender um automóvel. Durante uns tempos, o que as marcas de media venderam aos anunciantes é para além da audiência, o impacto que eu tenho, eu vou mostrar este teu anúncio a determinada audiência o que corresponde a determinado perfil, mas acrescenta valor por exibir esse anúncio junto de uma marca como a TVI24, ele acrescenta valor ao próprio anúncio pelo facto do anúncio ser exposto junto de uma marca com credibilidade. O que o Google e o Facebook dizem é que o conteúdo não interessa, interessa única e exclusivamente a audiência. Se eu atingir o João, esteja ele a ver o conteúdo de um blogger ou na TVI24, o valor é o mesmo desde que atinja a audiência. Onde tem sido, e reconheço este défice a nível global onde as marcas

de media não têm conseguido fazer demonstrar com KPIs analíticos qual é o valor que acresce pelo facto de a publicidade ser exibida numa marca premium vs um conteúdo que é feito por um utilizador a filmar um vídeo do Youtube. Porque é que quando vendo um anúncio exibo um anúncio da Volkswagen antes de um anúncio da TVI24? O que é que ofereço à marca que é vista pela audiência que vê a TVI24 vs a audiência estar a ver um vídeo de gatinhos no Youtube que leva a este mesmo vídeo da Volkswagen. Este é o grande dilema das marcas de media porque na realidade o conteúdo custou muito mais a produzir. Paguei a jornalistas, tenho um carro de exteriores, fui filmar uma peça, é caro vs alguém que filmou com o seu telemóvel e distribuiu no Youtube. O Youtube gastou zero a produzir o conteúdo e o utilizador zero. Ou seja, o Youtube gastou em tecnologia a criação da plataforma de distribuição de conteúdos gratuitos e exibe publicidade, como neste caso da Volkswagen, e vende esta publicidade mais barata do que aquilo que eu vendo. Como é que eu consigo convencer as marcas de que tem mais valor… e a porta valor é a minha marca… pelo facto de exibir este conteúdo junto dos meus conteúdos. Esta é a grande dificuldade. É a partir daí que decidimos querer ganhar independência na distribuição e na venda de publicidade e criar mercado de conteúdo que permita com o login do NÓNIO que eu consiga recolher data e informação das preferências dos utilizadores que pode ser obtida através de 2 coisas: 1) first party data declarativa que é dos utilizadores que me dizem qual é o seu género e idade e eu tenho a capacidade de segmentar a publicidade para que tenha menos desperdício. Ou seja, vou entregar publicidade aos utilizadores que são do género masculino, entrego-lhes anúncios que são especificamente par ao seu género e não entregar para toda a minha audiência. Ou seja, ter capacidade de ter alguma data para segmentar publicidade e desperdiçar menos desperdício com essa publicidade. Ao mesmo tempo criar um Marketplace em que vendo publicidade aos stakeholders dentro do universo dos publishers premium, mas que tem um objectivo que é termos mais independência e mais receita publicitária para produzirmos melhores conteúdos.

VL: O NÓNIO permite-vos fazer melhor publicidade porque permite-vos criar segmentos mais finos que são depois vendidos aos anunciantes. Que outros tipos de perfilagem é que o NÓNIO permite fazer?

MM: Declarativos, só género e idade. Depois tem dados que são recolhidos com base na navegação. Ou seja, existem algoritmos que são criados para esta plataforma… em termos

tecnológicos existem 2 vectores aqui: o SSO (Single Sing On) que faz o registo em cada um do sites e depois ele é reconhecido pelos 70 sites que pertencem ao NÓNIO e existe outro vector tecnológico que é o DMP que recolhe a navegação, que tipo de conteúdos é que o utilizador por clusters (grupos), não o utilizador individualmente porque eles são colocados em segmentos, em clusters, ou seja, não quero saber exactamente do que é que o Manuel gosta ou deixa de gostar, eu coloco lá aquele IP dentro de um segmento que é sport lovers, por exemplo. São segmentos comportamentais ou de interesses. Os interesses são Sport Lovers, Notícias de economia ou finanças; Mommy; Ciências e Discovery; Lifestyle; Trainning and Exercise; Health and Fitness; Food, Wine and Dining; Tecnologia and Gadgets; Plants and Gardenning; Active Outdor and Nature; Animal Lovers; Business and Career; Gaming; Cultura e Artes; Music Lovers; Sport Followers; Celebrities and Entretaining; Children and Parenting, Auto Lovers. Ou seja, são um conjunto de segmentos que são definidos pelo IAB (Interactive Advertising Bureau) que definiu para o mercado de publicidade quais o tipo de segmentos de interesse mais relevantes para a indústria. O que nós temos é um conjunto de algoritmos que recolhem a navegação dos nossos utilizadores e os colocam dentro desses perfis para quando alguém de uma determinada marca quiser comprar publicidade para o segmento de Auto Lovers ela vai servir nos sites àquela audiência que já foi perfilada como sendo correspondente a esta, a este tipo de audiência. É por interesses ou comportamentos e por dados demográficos, género e idade.

VL: O que é que fazem com esses dados? Qual é a prática comercial associada ao NÓNIO?

MM: A prática comercial é, existem com base nestes segmentos… o output desta recolha de data é a capacidade de entrega em determinado produto de impressões. O que é transacionado com as marcas é volume de impressões, ou seja, 5 milhões de impressões para este segmento. Ou seja, para o segmento auto lovers homens, posso cruzar o segmento demográfico com interesse, homens que gostam de carros… tenho a capacidade de entregar dentro da rede NÓNIO 5 milhões de impressões no formato X de publicidade, ou seja, para este formato que é um MR, que é um Medium Rectangle – é aquele mais tradicional que é um quadradozinho do lado direito em que aparece uma marca – para este segmento consigo entregar 5 milhões de impressões neste segmento, porque nós vendemos não as pessoas, mas sim, para este segmento, nós vendemos X impressões de

publicidade (vender 5 milhões de impressões, deve ser 5 milhões de vezes em que o rectângulo aparece aos vários utilizadores do NÓNIO).

VL: Como é que acha que o NÓNIO pode afectar o modo como vocês, Media Capital, retêm leitores e fazem leitores voltar à página da TVI24, como é que acha que isso pode acontecer?

MM: É com base na recomendação dos artigos, ou seja, um utilizador que está a navegar, a única forma que temos de fazer isso é recomendar-lhe conteúdo que seja mais relevante. Tem de ser feito com parcimónia. Mais uma vez, estas máquinas, este machine learning, têm limites na realidade, não são inteligentes o suficiente… é aquela base inicial de que estava a falar, tem de haver sempre a curadoria do editor depois de existir um conjunto de recomendações com base na navegação. Mas na realidade, o que vai acontecer é que, alguém que consome sempre os artigos do Benfica, o consumo que lhe vai ser recomendado, o que tem de ser feito é deixar para as minhas recomendações um terço das recomendações serem, isto ainda não está definido, serem um terço recomendadas pela máquina e outras tantas por um editor. E a máquina o que vai fazer sempre é este IP gosta, este utilizador vê sempre ou 70% do seu consumo são artigos do Benfica em Andebol, eu vou-lhe recomendar artigos com Andebol, ou seja, posso dar, quem vê este artigo passa a ver este tipo de artigos, ou então, com base na sua experiência tem mais artigos aqui de Andebol ou mais histórias sobre Andebol, tenho aqui mais histórias interessantes sobre Andebol e, para não repetir durante mais tempo, vou-lhe dar mais artigos que tenham a ver com esta experiência que ele tenha porque começo a perceber que este utilizador gosta de Andebol e vou buscar ao meu portefólio de histórias de fundo de Andebol ou artigos que passaram na televisão sobre Andebol e vou-lhe dar esses artigos.

VL: Por causa da repetição, acha que podemos cair aqui num loop de recomendar sempre a mesma coisa, um bocado como o Youtube faz com a música?

MM: Acho! Acho, acho. Daí achar que a recomendação do editor tem sempre mais relevância.

VL: E sobre a produção de conteúdos, acha por exemplo que se a TVI perceber que 90% dos leitores só quer artigos sobre o Cristiano Ronaldo, isso faz com que a produção de artigos sobre o Cristiano Ronaldo aumente?

MM: Sim. Acho que pode ser estupidificante nesse sentido. Acho que vai ter sempre de… se deixarmos totalmente o escrutínio das marcas, seria um projecto engraçado… pegue num site que faz um clip automático de conteúdos copiados de outros sites… imagine, o Notícias ao Minuto, que na realidade tem só jornalistas juniores que não saem da redacção, que no fundo vêem os outros sites, não copiam o conteúdo, mas dão-lhe tratamento editorial diferente, mas na verdade não produzem notícias. Mas repare, nós também temos muitos jornalistas que estão nas redes sociais à procura de conteúdo. Mas na realidade é, até que ponto é que se fosse uma marca, ou uma máquina só a dizer… a trend que está a dar é conteúdos sobre o Cristiano Ronaldo, de maneira que vou continuamente buscar na internet todo o conteúdo que exista disponível e vou disponibilizar… mais isso já é o que faz o Facebook e o Google. O que fazem na realidade é dar a alguém que tem muito interesse no Cristiano Ronaldo conteúdos sobre o Cristiano Ronaldo, tudo o que é publicado mundialmente sobre o Cristiano Ronaldo. O problema para a nossa humanidade é até que ponto é que vamos ficar numa redoma dos conteúdos que são recomendados e perdemos a noção do que é a realidade. Olhe para a Coreia do Norte que está fechada sem noção do que é a realidade exterior, a Síria que não fazemos ideia do que por lá se passa porque não temos um jornalista que vá lá para saber o que é que se passa. Na realidade, estamos a proteger determinado tipo de regimes que podem fazer o que quiserem às populações porque ninguém sabe o que é que se passa lá. Há aqui grandes riscos para a informação e para a liberdade até. Mas sim, poderá haver uma tendência, mas não o fazemos ainda. Existem, sim, recomendações que temos do Google Trends, do Facebook, vê-se o que é que tem melhores resultados, as pessoas têm mais tempo para ver aquele conteúdo, elas dão recomendações aos nossos editores para perseguirem mais este tipo de conteúdo. Ou seja, existe, ele tem relevância e até o próprio NÓNIO vai ter esse tipo de utilidade. Ou seja, começo a ver quais são os artigos e conteúdo que tem mais interesse e dão dicas aos editores, mas nunca pode passar por ser única e exclusivamente uma recomendação… ou seja, não substituem os editores. Caso contrário passamos a ter informação em loop e ficamos sempre no mesmo conteúdo. Há aqui um limite difícil de definir sobre qual é que é a utilização que temos de dar a esta

informação que recolhes porque ela é estupidificante. As pessoas têm interesses próprios, mas os interesses próprios vão sendo manipulados pela informação que lhe dás.

VL: Quando uma pessoa entra no NÓNIO, o NÓNIO vai imediatamente oferecer-lhe conteúdo… ou seja, uma pessoa que é cliente e que utiliza frequentemente o NÓNIO, ao voltar a entrar no NÓNIO, vai ser bombardeada logo com conteúdos que se adequam ao perfil dessa pessoa? Lá está, se eu gosto do Cristiano Ronaldo, mal acedo ao NÓNIO tenho logo conteúdo do Cristiano Ronaldo à disposição quando entro no NÓNIO?

MM: Não porque os sites ainda não estão preparados dessa maneira. O que acontece é que no final de um artigo que o utilizador vê aparece uma recomendação de conteúdos. Essa é que já pode ser feita com base no DMP do NÓNIO. Mas os nossos sites ainda são todos feitos com base na curadoria editorial. A recomendação de conteúdo desaparece… por exemplo tens o IOL que é um portal de conteúdos, para já, exclusivamente da Media Capital, mas a recomendação ainda é feita manualmente. Mas esse que é um site que define só conteúdos que estão na nossa rede, esse é um site que podia ser construído só à medida que um utilizador do NÓNIO que esteja lá logado e que seja recolhida informação sobre essa preferência, o primeiro conteúdo que aparece destacado já pode ser feito só com base na navegação que ele tem dos conteúdos consumidos da Media Capital. Ou seja, acredito que possam surgir projectos que sejam, da mesma maneira que um utilizador tem na Netflix recomendações feitas com base nos gostos e até a evolução do que possam ser as boxes da MEO, NOS e Vodafone, deixar de ter a lógica de distribuição de canais, mas recomendar conteúdos, o que passa a ser esquizofrénico, porque se há uma senhora que só vê novelas, assim que entra a única coisa que vê são as novelas que estão a dar. Acredito que possam surgir projectos que sejam com base num automatismo só de recomendação de conteúdos, mas para já não, para já só há recomendados depois, no final, de se consumirem conteúdos.

VL: Acredita que o NÓNIO e as práticas de Big Data podem levar à sobrevivência económica dos media?

MM: Acho que podem ajudar à sobrevivência dos meios e por isso é que temos o projecto. Se continuarmos a saber distinguir os meios, mercado e todos os stakeholders da indústria de marketing, reconhecer que existe valor acrescentado no que estas marcas (media) dão

à publicidade nos seus sites. Nós nunca vamos conseguir competir em termos de volume, dimensão de audiência, com distribuidores globais como o Google ou o Facebook, mas é essencial para perceber que temos uma missão diferente e que tem mais valor a exibição de publicidade dentro destes nosso conteúdos. O Big Data ajuda a diminuir o desperdício das campanhas de publicidade no mercado digital. Que elas sejam entregues com mais eficácia. É compatibilizar o mercado de data com o contexto de publishers profissionais ou de conteúdo premium.

VL: Além do NÓNIO, que outras práticas de Big Data é que a Media Capital tem?

MM: Existem mecanismos ad hoc que são fornecidos por tecnologias terceiras, o Google, de avaliação de audiência. Ou seja, Big Data só na perspectiva de medir a audiência e de recomendação de conteúdos que não seja ainda com base no NÓNIO. Ou seja, no fundo é a recomendação de conteúdos e de utilização de… na compra de publicidade existem intermediários nossos que utilizam data para segmentar publicidade, ou seja, existe ainda a 3 níveis, mas que é o que gostaríamos de deixar de ter. Compradores de publicidade que utilizam data terceira para segmentar publicidade nos nossos sites. Como é que isto funciona? A agência que tem a campanha da Voklkswagen compra data ao Google que está ligado aos nossos sites e quando faz match do cookie do utilizador que está nos nossos sites ou no cookie que é utilizado naquela campanha, que foi comprada ao Google, ou seja determinado utilizador passou no site do Google, o Google ficou com essa informação sobre o utilizador e essa informação sobre o utilizador é injectada na campanha da Volkswagen, e esse utilizador quando está no site da Media Capital, se corresponder àquele perfil, a publicidade que é exibida é a dessa Volkswagen. É essa utilização, essa third party data, que queremos que deixe de ser usada e que ela seja só com base na nossa data e não com data da Google. A compra programática é feita desta forma, pu seja, na realidade todos nós já usamos, por terceiros, Big Data. Já existe Big Data na segmentação de publicidade, mas é uma Big Data de terceiros.

VL: Acha que os consumidores e as pessoas que acedem aos vossos sites estão dispostos a dar informação privada, os dados sobre quem eles são, que fazem e o que consomem, em troca de uma coisa que lhes acrescente valor, neste caso notícias fidedignas. Acha que aos olhos do consumidor é uma troca justa?

MM: É uma mensagem que é muito difícil de passar. Quando fizemos até uma campanha. Há muitos utilizadores, ou seja, o nosso rácio relacionado com reclamações técnicas, de usabilidade, da não permanência do login e ele não funcionar muito bem porque está a ser pedido muitas vezes quando a nossa promessa é de que basta fazer o login uma vez e pode-se navegar pelos sites, sendo que o login não é pedido novamente, é certo que muitas vezes existem erros aqui. Isto é uma dificuldade técnica que temos de implementação, de maneira que existem muitas reclamações nesse sentido. Esta é uma altura boa até porque se vêem muito mais campanhas de subscrição à procura de utilizadores que sejam pagantes e na verdade a única coisa que pedimos não é mais do que qualquer utilizador dá ao Facebok e a o Google. Os utilizadores hoje em dia não têm problema em dar todos os seus dados e repare, quando estamos no Facebook estamos a dizer que gostamos de determinada coisa e que o Facebook está a utilizar isso, está a chegar à nossa caixa de email, se reparar, depois de fazer esta chamada, provavelmente se for ao Facebook eu vou-lhe aparecer como recomendado para sermos amigos. Veja bem o nível de data a que o Facebook e a Google conseguem chegar e que lhes permite fazer determinadas recomendações. Muito mais do que o NÓNIO quer saber. Eu só quero saber o tipo de conteúdos que o utilizador vê nos nossos sites. Eu não quero saber se o utilizador vai a um site de prostitutas, eu só quero saber quais são os conteúdos que o utilizador vê, se é homem ou mulher e qual é o intervalo de idades onde se encontra. Apesar de isto dos GDPR declarar isto como sendo dados pessoais, o utilizador tem de aceitar partilhar esta informação quando acede partilhar ao NÓNIO, na verdade o tipo de informação… acho que os utilizadores quando vêem uma empresa como a Media Capital ou a Impresa ou a Global Media a querer ficar com esta informação não têm uma reacção muito positiva, achando que tem de dar dados pessoais. Mas o que é certo é que estes mesmos utilizadores dão de livre vontade a uma entidade gigante e terceira os dados todos. Mas há uma reacção negativa à utilização destes dados.

VL: Os algoritmos são programados por pessoas. Acha que quaisquer tendências ou preconceitos que o programador possa ter podem manifestar-se no modo como os dados são recolhidos?

MM: É um tema muito interessante e pouco abordado. Na realidade, aquilo é definido por pessoas e programadores que definem qual é o padrão e essa é a maior black box do Google e do Facebook. Não faço a mínima ideia sobre como são construídos os perfis de

data que o Google vende. Ou seja, os perfis típicos de interesse, a nossa pitch é que eles são totalmente transparentes e eu posso dar exactamente às marcas quais foram as definições usadas para que determinado utilizador fosse considerado para determinado perfil de interesse. Sendo que os sócio-demográficos são básicos, ou seja, o próprio utilizador põe lá um ticker a dizer que corresponde a determinado perfil de idade ou se é homem ou mulher, na realidade, se calhar devia existir uma terceira hipótese que é o "outro" ou "não quero declarar o meu género", admito que possa haver aqui um bias, o Bloco de Esquerda já pode ter aqui, já há aqui um preconceito, na realidade. Mas repare, todo o negócio da publicidade está assente em temas que hoje em dia podiam ser considerados preconceitos. Que é, estar a segmentar publicidade só para determinado tipo de indivíduos. Porque é que na realidade estou a servir publicidade que é só para homem ou mulher? Agora que falamos neste tema, ao abrigo do que têm sido temáticas que têm sido fracturantes na sociedade, podemos achar que todo este negócio da publicidade digital está assente em preconceitos.

VL: Mas é a única forma de fazer publicidade, não é? É preciso critérios? É preciso partir de pressupostos?

MM: Senão, o que se fazia era fazer publicidade como antigamente. Publicidade cobrada ao site que é entregue a toda a gente, mas isso destruía todo o negócio. Isso seria o ideal para nós. Comprava-se ao site da TVI e a TVI atingia toda a gente. Mas todo o planeamento, toda a indústria de publicidade aceita na ideai de que há um alvo, há uma audiência alvo que quero atingir. Quem compra este Volkswagen tem determinado perfil. São mais mulheres que homens. Senão negamos a nossa própria existência. Vamos negar que somos homens ou mulheres, somos todos electrões e protões e somos todos iguais.

VL: Os preconceitos do programador podem então manifestar-se nos algoritmos?

MM: Sim. Qual é então a comissão ética ou moral que está a definir o algoritmo? Essa pergunta também devia ser feita ao Google, mas a nossa é transparente. O nosso pitch é que quando chegarmos à fase de definir um algoritmo que define a que tipo de perfil se enquadra o segmento Sport Followers, ele vai definir como, a condição é se este IP consumiu X artigos que tinham uma TAG que diz desporto, aquilo é um artigo sobre o Andebol do Benfica, se viu X vezes este tipo de artigos que têm a TAG de desporto…

atenção que é preciso colocar a TAG de desporto nos artigos que se publica para a máquina ir lá buscar… se ele viu X vezes artigos que têm a TAG desporto, 10 vezes por semana, 3 x por dia, ou o que quer que seja, ele vai enquadrar-se neste tipo de segmento. Ou seja, se aquilo estiver mal TAGgado toda a audiência vai corresponder a todos os perfis. E isto é já chegar ao fundo da questão. Só que a uma certa altura, os próprios marketeers vão perceber que as campanhas passam a ter menos eficácia. Se usei uma campanha com data e tive certos resultados e outra sem data nenhuma sem segmentação, e os resultados foram iguais é porque esta Big Data não presta. O próprio mercado que compra esta data vai perceber se a data tem valor ou se foi manipulada.

VL: Além dos problemas que já mencionou, que outros problemas éticos estão associados ao NÓNIO?

MM: Além dos que mencionei, nenhum, sendo que acho que não há problema com a questão da privacidade. Também não acho que haja problemas de concorrência.

VL: O NÓNIO e o Big Data podem ser um primeiro passo na automação dos media? Se os media portugueses podem copiar o exemplo dos media norte-americanos que têm algoritmos a produzir conteúdos de forma automática?

MM: Espero bem que não, mas na realidade eu vejo na própria humanidade que nós temos esta tendência. Ou seja, para criar mecanismos…. Primeiro, quando se trabalha num mundo capitalista onde o objectivo é que a economia cresça ano após ano e vê-se demograficamente que o número de consumidores não aumenta, a solução é obrigar as pessoas a consumirem mais. Isto quer dizer que todos os mercados vão ter de produzir mais com menos e o menos normalmente é menos custos o que, nestas empresas de media, traduz-se em recursos humanos. Se houver mecanismos que no mundo digital possam ter receita na publicidade em conteúdos que são gerados sem custos como pessoas, com custos mais baixos, acredito que possa haver essa tendência. Ou seja, esse risco é real, existe, de produção autónoma de conteúdos e existe até o risco de criar uma comissão de ética que seja automática. Seja gerada por um computador. Imagine que se cria um computador tão potente que consegue cobrir de forma semântica todos os temas que são de ética. Isto é tramado. Nesse caso, o computador vai substituir, na realidade, uma pessoa, ou conjunto de pessoas.

4. Transcript of JF's interview, recorded on May 14th 2020

VL: O que é que entende por datificação dos media e porque é que é importante?

JF: É uma pergunta pouco específica no sentido que Data hoje em dia tem muitas aplicações dentro de uma empresa de media. Eu trato muito em especial sobre dados que se reflectem mais no ponto de vista comercial e dos utilizadores e de que tipo de mais-valias é que podemos retirar neste Data. Não estou tão envolvido naquilo que o Data pode contribuir para as equipas editoriais ou para os jornalistas construírem os seus artigos e as notícias… Eu acredito que o Data para aproveitamento jornalístico e editorial é cada vez mais importante não só porque toda essa informação permite estruturar e desenvolver produtos que sejam muito mais segmentados e relevantes para cada utilizador e portanto pode-se vir a assistir não só a uma escrita mais genérica mas também na construção de produtos específicos para determinados targets e utilizadores, dependendo dos seus interesses, e isso pode constituir de facto um excelente activo para o desenvolvimento jornalístico. Depois, todo o Data de suporte que existe à volta para enriquecer a própria peça, o próprio artigo, é outra área importante. Ou seja, a interactividade que se pode criar, o data journalism que se pode criar, inserindo Data nos próprios artigos para ir buscar vários tipos de fontes ou plataformas que podem contribuir para o enriquecimento dos artigos com data é uma outra área que eu acho que vai evoluir muito nos próximos tempos. Se falarmos de como é que este data nos media também contribui para o seu negócio e para as suas receitas, é obviamente também muito importante recolher cada vez mais data sobre os utilizadores, aqui estamos a falar muito mais de perfil quer socio-demográfico, quer de interesses ou digamos o comportamental dos utilizadores, de forma a poder servir melhor e ser mais relevantes aqui nas mensagens de publicidade que podemos servir a cada utilizador. E portanto, isto terá mais relevância para os compradores, para os anunciantes e para as marcas, e portanto este aqui é o factor em que o Data vai ser mais importante, quer por estas características, quer por questões de geo-localização do utilizador, tudo isto combinado com o Data da marca do anunciante vai

também obviamente trazer mais conversão, mais notoriedade junto das marcas de acordo com cada perfil de utilizador que chega aos nossos sites.

VL: Em termos de perfis, que tipo de perfilagem é que é feito através do NÓNIO? Que tratamento de dados é que o NÓNIO permite?

JF: Neste momento temos uma metodologia de recolha de dados a que nós chamamos first party data declarativa. Portanto, quando nós implementámos o registo do utilizador, estamos a recolher geralmente dados socio-demográficos. Estamos nomeadamente a falar de género e idade. Esses são os primeiros dados que recolhemos e que são declarativos, ou seja, cada utilizador ele próprio disponibiliza esses dados. Depois, são recolhidos também dados comportamentais. Quando o utilizador faz uma visita ao site e visita determinado tipo de artigos, conteúdos, notícias ou entretenimento, também é recolhido com base nesses comportamentos mais dados e isso é geralmente feito com um algoritmo semântico que permite saber a que tipo de conteúdos é que o utilizador está exposto ou que consome e com isso são construídos geralmente perfis de interesses, por exemplo, desporto, automóvel, área financeira…. Há várias tipologias de perfis de interesses que também são construídos a partir desse comportamento. Tudo isto nós consideramos que é first party data, porque no caso do NÓNIO temos cerca de 70 sites onde o utilizador pode visitar vários destes sites e com a recolha em todos eles nós conseguimos construir então estes perfis para além dos que são declarativos, também estes que são comportamentais.

VL: É essa a vantagem do NÓNIO? Dar escala? É por isso que são 70 sites?

JF: Claramente a escala aqui é… existem 2 coisas fundamentais no pensamento do projecto e no objctivo do projecto NÓNIO: um é a escala, e podemos ter mais oportunidades de fazer esta recolha em maior volume dos utilizadores, e por outro lado é o ser declarativo e first party, que é um dado recolhido por cada um destes 70 sites e não aquilo a que se chama o third party data que é, digamos, data proveniente de outras plataformas ou de outras origens mas que não é controlado por nós, ou seja, tem menos valor. Poderá ser um Data recolhido em milhões de eventos ou noutros tipos de situações que não estão controladas nem dentro do nosso eco-sistema e portanto que não sabemos valorizar se o critério usado foi ou não um critério relevante. Portanto, aqui, quando

alguém visita o Jornal de Negócios e que sabemos que visita todos os dias, 3x por dia, e que depois vai ao Expresso e que tem determinado tipo de comportamentos, podemos perfeitamente criar os nossos critérios para construir depois segmentos e data a partir desses critérios que nós conhecemos. E portanto é essa first party data e a escala que fazem do NÓNIO um projecto único.

VL: E depois pegam nesses dados e fazem o quê? Ou seja, como é que a informação reunida se traduz em maiores receitas para a Cofina?

JF: Não é só para a Cofina, é para todos os publishers do NÓNIO. O que nós estamos a fazer para monetizar isso é oferecer essas audiências segmentadas já com estes critérios de que falei às marcas. Geralmente isso tem um valor superior a quando se oferece uma audiência sem qualquer tipo de caracterização.

VL: Em termos do modelo de subscrições. Como é que acha que o NÓNIO pode melhorar o modo como a Cofina retém subscritores e leitores?

JF: O NÓNIO, recolhendo este data comportamental dos utilizadores e também o próprio sócio-demográfico, vai permitir-nos chegar com maior relevância àqueles que conhecemos e que têm determinado tipo de preferências ou por conteúdos ou por determinado tipo de informação, portanto, podemos também segmentar melhor as nossas ofertas e os produtos que temos de assinaturas e de subscrições, conhecendo melhor quem são esses utilizadores.

VL: Isso traduz-se de que forma? É o conteúdo que recebo no email ou quando acedo ao vosso site aquilo que vejo?

JF: Tem duas fases. Uma que me perguntou e de que lhe estava a falar. Se tenho aqui um perfil de um executivo que gosta muito de informação sobre a área financeira ou económica, eu provavelmente posso oferecer uma assinatura do Jornal de Negócios. Essa é uma forma de criar conversões, de criar novos negócios a partir de assinaturas. Para os utilizadores que não chegam à assinatura, então também posso, para aquilo que é a parte não-premium ou não-paga dos conteúdos e que é o acesso geral, eu também posso melhorar a relevância dos artigos que ofereço seja de newsletters ou o meu marketing ou

outras coisas conhecendo melhor o perfil do utilizador. Não é pago, mas poderei fornecer produtos e artigos de maior relevância.

VL: Quando falei com as pessoas da Impresa, foi-me dito que já sabiam captar subscrições, mas que a dificuldade era mantê-las. Concorda com esta análise? É este o caso para a Cofina?

JF: Acho que esse é o desafio de todas as plataformas de subscrição porque… e muito maior é o desafio quando se trata de serviços de notícias porque os utilizadores que subscrevem um serviço de entretenimento do estilo Netflix ou de música como o Spotify talvez fidelizem durante mais tempo. O utilizador de notícias tenderá muitas vezes a fazer subscrições no tempo. Dou-lhe um exemplo: neste caso da crise pandémica, isso tem vindo a acontecer. Especialmente no início da crise, os utilizadores realmente perceberam que as notícias credíveis estavam nos publishers mais tradicionais e nós notámos um grande crescimento de subscrições. Mas, neste momento, já sentimos que subscreveram durante um determinado período de tempo em que achavam que isso era muito importante para eles e rapidamente cancelaram as subscrições. Claramente, o desafio de os manter é muito grande. E traduzir o valor que isso representa para eles… Acho que há um desafio grande em todos os publishers que é refazer os produtos em termos de media e artigos, ou seja, não pode ser só o produto breaking-news ou a notícia do momento, mas tem de passar pela oferta premium, a oferta paga, ser uma oferta com valor que se possa traduzir, um valor tangível, e que faça a diferença em relação à tradicional informação que é gratuita.

VL: Em termos de produção de conteúdos, como é que o NÓNIO pode afectar ou está a afectar a produção de conteúdos noticiosos?

JF: Penso que neste momento ainda não está a afectar a produção de conteúdos. Virá numa segunda fase em que tivermos uma base de dados ainda mais alargada sobre os utilizadores. Nós contamos neste momento com um milhão e qualquer coisa de utilizadores, o nosso objectivo é chegar perto dos 4/5 milhões.

VL: Um milhão no total? Dos 70 sites?

JF: É mais de um milhão… à volta de um milhão e meio no total, nos sites todos. Isso está em crescendo e acredito que a primeira fase do NÓNIO até para que este projecto possa continuar. Nós temos de monetizar isto junto dos mercados de publicidade e marketing. Mas vamos continuar depois a desenvolver outros produtos dentro daquilo que são as plataformas de Data que estivemos a construir aqui para que possam também dar maior suporte à produção. Mas é uma segunda fase.

VL: Quando diz que pode vir a afectar a produção de conteúdos, pode dar-me um exemplo de como é que isso funcionaria?

JF: O que pode existir é mesmo uma personalização na forma como servimos no site as notícias ao utilizador. Imagine, por exemplo, que hoje nós temos se calhar num site, numa homepage de um site, o mesmo conteúdo para todos os utilizadores, no futuro verá conteúdos que serão servidos só para determinado tipo de utilizadores. Haverá uma customização do conteúdo servido em função do perfil, dos interesses do utilizador. Haverá uma personalização daquilo que é mostrado.

VL: O que não é só na oferta, mas também na produção. Escrever determinados tipos de conteúdos só para aquele tipo de pessoas?

JF: Sim, conhecendo melhor os perfis permite que não só na escrita possa haver esse cuidado, como também ao servir a cada perfil, isso também será diferenciado.

VL: Ou seja, vamos supor que há um segmento de pessoas que gosta de chocolates e de Cristiano Ronaldo, então o artigo poderá incluir as duas coisas?

JF: Não iria tão longe, mas sim. Poderá haver artigos sobre o Cristiano Ronaldo e chocolate… pode acontecer no mesmo artigo, mas se calhar faria mais sentido que houvesse artigos sobre o Cristiano Ronaldo e outros sobre chocolate e ambos serão mostrados a utilizadores com esse tipo de perfil. O Data acaba por pedir muitas combinações. Pode haver pessoas que gostam do Ronaldo, mas que não gostam de chocolate. Tenho de ter a capacidade de servir aquilo que é, digamos, separar coisas que não tenham interesse nem relevância para alguns utilizadores.

VL: Acredita que o NÓNIO e as práticas de Big Data podem levar à sobrevivência económica dos media?

JF: Também eu gostava de saber isso. Eu acredito que não há forma de sobrevivência… não sei se vai ser um sucesso. Acho é que não há forma de sobreviver sem Data no futuro. Portanto, quem não desenvolver estas suas capacidades tecnológicas para utilizar o Data quer para a produção, quer para a distribuição, quer para a identificação dos perfis dos utilizadores não vai conseguir sobreviver, claramente. Agora, até que ponto isto sozinho é decisivo… a qualidade do jornalismo, a credibilidade dos jornalistas, das marcas que estamos habituados a consumir, etc… isso também vai ditar o futuro. Mas acho que sem Data será muito mais difícil sobreviver num futuro muito competitivo e em que temos plataformas globais que inclusive, muitas vezes, servem-se de capacidades tecnológicas muito superiores às dos publishers, porque as plataformas globais nasceram tecnologicamente, nasceram na tecnologia, enquanto os publishers tradicionais não, estão sim a adaptar-se à tecnologia, e portanto vai ser sempre mais lento o nosso processo de actualização tecnológica. Mas que temos de ter bem em mente que isso é fulcral para o futuro dos nossos negócios, é.

VL: Para além do NÓNIO, a Cofina tem outras práticas de Big Data?

JF: Sim, nós temos outras práticas de Big Data, claramente temos outras ferramentas. O NÓNIO é uma plataforma e uma aliança entre publishers para determinado tipo de objectivos, muito estes de que falei de ter um registo de utilizadores de escala, mas cada grupo tem também ferramentas próprias para analisar todo o seu próprio Data que produz de forma individual. Há estratégias também depois individuais que são aplicadas às equipas, não só equipas editoriais, mas de marketing, equipas comerciais, há muito outro tipo de Data individual que ultrapassa o âmbito NÓNIO.

VL: Acredita que os consumidores estão dispostos a trocar informação privada, ou seja dados pessoais, por algo que lhes acrescente valor? Acha que esta é uma troca justa para o consumidor e aos olhos do consumidor?

JF: É sempre difícil o consumidor fazer essa troca, pelo menos daquilo que nós temos visto. De acordo com a nossa experiência, os consumidores resistem a dar os seus dados,

mas com o milhão e meio de utilizadores que já se registaram no NÓNIO, também ficamos de alguma forma com a esperança de que, quando os utilizadores percebem e entendem que estão de forma gratuita a ter acesso a conteúdos e que unicamente lhes são pedidos alguns dados muito simples e básicos, eles acabam por facultar esses dados. Mesmo a experiência que todos nós temos não com os media mas com plataformas como o Facebook… nesses casos os utilizadores fornecem muito mais dados pessoais do que aqueles que nós pedimos e recolhemos. Nesses casos, os utilizadores não se importam de os fornecer. No entanto, há ali um valor percepcionado por trás. Ou seja, estou disposto a fornecer os dados porque tenho acesso a determinado tipo de serviço ou discussões ou outro tipo de vantagens. Acho que a minha geração tinha a ideia de que tudo na internet era gratuito, mas parece-me que isso também está a mudar. As gerações mais novas já percebem que o fornecimento de dados ou o pagamento de subscrições faz sentido. É natural que cada vez mais isso aconteça.

VL: Que problemas éticos advêm do NÓNIO?

JF: Problemas éticos, acho que não temos. Eventualmente teríamos questões relacionadas com privacidade dos dados, mas todo o nosso projecto está validado de acordo com as leis portuguesas e da comunidade europeia. Quer no que respeita aos termos e condições e utilizado dos dados quer à privacidade dos dados, nós seguimos todas as regras. Está tudo de acordo com os standards europeus.

VL: Os algoritmos são programados por pessoas. Acredita que as tendências humanas, ou os preconceitos do programador, podem de alguma forma manifestar-se nos algoritmos do NÓNIO ou nos algoritmos de Big Data?

JF: Depende daquilo que o algoritmo for preparado para fazer. No caso do NÓNIO, os algoritmos que temos aqui nesta fase do projecto, são algoritmos que fazem classificação, fazem a recolha e depois a classificação dos utilizadores em coisas muito objectivas. Quando dizemos que um utilizador pode ser classificado como alguém com um perfil que gosta de desporto, o algoritmo foi construído com base em factores muito factuais e mensuráveis. É porque o utilizador fez determinado tipo de acções várias vezes ao longo do tempo. Aqui não há ninguém a dizer "ah acho que este tipo gosta de desporto porque usa barba e tem um cabelo cor-de-laranja". Sei que existe outro tipo de experiências com

dados… até estive envolvido num projecto que tenta fazer previsão daquilo que os consumidores e utilizadores vão fazer nos próximos dias - estamos a entrar no campo da neurociência – são coisas que têm um algoritmo um bocadinho mais nessa linha de que me estava a falar. Agora, no nosso caso, e nesta fase, o que tenho para dizer é que os algoritmos funcionam com bases factuais. Acho que os programadores não terão uma grande influência assim tão grande naquilo que é o output final.

VL: O NÓNIO pode ser um primeiro passo na automação dos media? A Cofina está a pensar desenvolver algoritmos que gerem notícias de forma autónoma, à semelhança do que acontece noutros publishers internacionais?

JF: Acredito que isso poderá ser interessante e está já a ser experimentado e de alguma forma feito por outras organizações. Não sei se o NÓNIO em si tem algum lugar nessa construção, porque parece-me que esse é um tipo de tecnologia ligeiramente diferente, embora possa depois recorrer a Data que vem do perfil dos utilizadores para produzir ou automatizar a escrita de acordo com um determinado perfil. O que nós estamos a dizer é que são algoritmos que vão buscar Data a vários tipos de plataformas, o NÓNIO pode ser uma delas. Mas esse caso de que falou, de quase inteligência artificial, em que se procura criar coisas, está ligado a eventualmente a esta plataforma onde temos os perfis dos utilizadores e poderá ir lá buscar mais informação relevante para a construção de notícias. De qualquer forma, para as construir, terá de estar ligado a outras plataformas para saber quais são as notícias relevantes. O NÓNIO só tem um papel de despejar algum tipo de informação, como o perfil ou o comportamento dos utilizadores.

5. Transcript of JPL's interview, recorded on May 22nd 2020

VL: Como é que o NÓNIO funciona especificamente? Como é que o algoritmo reconhece que uma pessoa esteve a ler um artigo por exemplo sobre desporto? Isto é, é o próprio jornalista que, ao escrever o artigo, coloca lá uma TAG a dizer que é um artigo de desporto ou o algoritmo reconhece a semântica do artigo?

JPL: Nós estamos a falar de venda de publicidade programática. Programática significa que temos sempre aqui 3 intervenientes. Temos o DSP, do lado da procura, o demand side platform; no outro extremo temos o SSP, suply side platform, onde o publisher, a arquitectura que o publisher tem para falar com o DSP. Portanto, o publisher tem este SSP e os compradores estão todos do lado do DSP; e no meio há uma coisa que é o DMP, Data Management Platform. Esse DMP é que nos passa uma TAG que nós colocamos, é um pedaço de código, colocamos em todas as páginas do site. Ou seja, todas as páginas do site têm no seu hardcode essa TAG. Quando os jornalistas fazem uma peça, o editor de conteúdos vai publicar mas numa página onde já existia, onde já lá estava, o resto do código e nesse resto de código que já existia está lá essa TAG (esse algoritmo que lê TAGs – palavras minhas, do Vicente). Essa TAG, de uma forma muito abrangente, chama-se learning TAG. Learning TAG porquê? Porque ela recolhe tudo o que consegue recolher. E o que é que ela consegue recolher? Todas as informações que o browser lhe passa, e o browser passa imensas informações – geo-localização da rede através do IP, passa-lhe o tipo de device, o tipo de acesso à internet, a hora, etc... - e depois, esses pedaços de código, muitas vezes, têm capacidade de dar indicação de poder fazer crawling ao texto. Exactamente como o Google faz crawling às páginas da web para as indexar, também estes DMPs têm a capacidade de fazer crawling ao texto. O crawling tem, nuns casos regras muitas sofisticadas como as do Google neste momento serão, outras têm regras mais simples, mas que até pessoas que não são engenheiras conseguem perceber quais são: em primeiro lugar consistem em avaliar o título, se o título tiver determinadas palavras que pertencem a determinada temática, em teoria, esse texto pode ser catalogado debaixo desse segmento; vamos imaginar que é um segmento de desporto ou de economia; a seguir, já com menos peso, vai ver o lead, que é o pedaço de texto que constroem e que antecede a notícia; e por fim a notícia. Na notícia, tem sempre, e isto é algo que para nós é um bocadinho opaco, mas tem sempre aqui capacidades de categorizar sobretudo se forem ferramentas mais ricas do ponto de vista de trabalhar essa língua. Obviamente uma coisa é uma ferramenta habilitada para fazer isto em inglês, outra coisa é estar habilitada para fazer isto em português de Portugal. Portanto, há aqui algumas nuances, mas no limite os DMPs mais sofisticados também fazem isto. Quando não são tão sofisticados utilizam a categorização que previamente foi dada aos sites e às suas grandes áreas respectivas. Expresso, economia, Trbiuna é desporto, etc... Nesses casos poderão ser menos sofisticados e utilizam apenas pelo URL. Temos é de ter em conta que a maioria dos advertisers querem atingir o perfil de um tipo de pessoa. Não querem atingir

as pessoas que neste momento estão a ver desporto. Para isso não precisam de Data. Para isso bastava ir à Bola ou ao Record e já tinham essa segmentação, que é a segmentação que no passado todos usávamos. O que os advertisers agora querem é, "ok há muita gente que vai à Bola que tem este perfil, mas há uns quantos que lá vão e que têm um perfil completamente distinto". Portanto, o que quero é segurança no perfil e não apenas o contexto temático. E portanto, o que o DMP constrói são grandes segmentos de perfis de pessoas que estão eventualmente nas mesmas ranges etárias, ou são do mesmo sexo, ou gostam das mesmas coisas porque têm uma frequência elevada de - e aqui são as tais questões dos tresholds e aí é que digo que para nós são um pouco opacos, mas quando dizem que é alguém que gosta de conteúdo desportivo, normalmente têm um tapete estatístico do universo e a partir desse tapete conseguem perceber onde é que estão os tresholds, os pontuais e os regulares, e portanto estabelecem... é esta a segmentação que é feita.

VL: Quando diz que o treshold é opaco, agora com o NÓNIO já não é tanto, certo?

JPL: Mais ou menos. Em teoria teríamos, mas teria de ser uma arquitectura desenvolvida por nós, teríamos de ser nós a desenvolver o nosso próprio DMP. Nós optámos por fazer outsourcing disso e, ao fazer esse outsourcing, os tresholds ficam mais opacos. Nós discutimos critérios, mas não a esse nível de detalhe. Decidimos critérios de: neste segmento vamos juntar quem gosta de desporto com quem gosta de fitness. Gosta de praticar de desporto além de gostar de acompanhar o fenómeno desportivo. Esse género de customização temos a capacidade de o fazer. Agora, definir o que é que é o critério que habilita uma pessoa a ser um fitness enthusiast ou sports lover, isso já foge ao nosso controlo. Até mesmo as pessoas do DMP que falam connosco, não foram elas que definiram isso. Estamos a falar de organizações que têm por trás equipas técnicas com data scientists que tomam decisões e as pessoas do negócio não estão conscientes disso nem têm a percepção.

VL: Como é que o NÓNIO está a alterar o modo como o grupo Impresa faz publicidade?

JPL: Para já, não está. Naquilo que tenho a expectativa que venha a alterar... hoje somos vistos pelos anunciantes apenas como a cereja no topo do bolo, ou seja, quando eu quero enriquecer a minha campanha com contextos premium penso na Impresa, eu anunciante.

E estamos mais afastados da ideia e das campanhas e do dinheiro que são dirigidas para as pessoas, qualquer que seja o contexto. Com o NÓNIO conseguimos, acreditamos nós, uma oferta atractiva de preço competitivo, porque misturamos tudo dentro da mesma oferta, em que temos misturado o Correio da Manhã, o Record, o Expresso, o DN, a TSF, a TVI… temos muita coisa misturada e é aquilo que nos vai permitir ser competitivos anda para clientes que, a um preço interessante, preferem estar numa zona de publishers mesmo que não seja toda ela premium, mas que é numa zona controlada. Controlada significa que não estou naquelas zonas não só onde o contexto pode ter algumas sensibilidades como a própria forma como as páginas estão feitas... basta clicar nalguns artigos de alguns sites sugeridos pelo Google e entra naquelas páginas em que vê logo três banners no mesmo ecrã todos a piscarem... Porque essas páginas estão feitas só para essa receita publicitária. Depois é evidente que quando está no Excel a avaliar o número de impressões, uma coisa é a impressão ter sido feita sozinha numa página, outra é estar no meio de uma árvore de Natal, e portanto há alguma consciência por parte dos compradores disso e por isso para a Impresa pode haver essa vantagem de ter um complemento de competitividade.

VL: Como é que o NÓNIO pode vir a afectar, ou está a afectar, o modelo de subscrições e o modo como retêm os subscritores e os leitores?

JPL: Aí afecta de forma muito positiva porque a primeira grande barreira à subscrição é conseguir o registo. Durante muito tempo, o New York Times e o Wall Street Journal obrigavam ao registo e não obrigavam ao pagamento, porque era um passo que eles estavam a dar nesse sentido. Toda a gente sabe que obrigar ao registo e ao pagamento são dois muros altos. Ter já o registo feito é uma vantagem tremenda, porque quando a pessoa tem o impulso de registar a tarefa fica muito mais simples. E depois, também nos permite fazer campanhas atrás das pessoas que nos visitam de forma mais frequente. Portanto, aí há um impacto muito positivo.

VL: Quando está a falar do registo, é o registo que temos de fazer quando aderimos ao NÓNIO, certo?

JPL: Sim.

VL: Em termos de produção de contéudos, qual é o impacto esperado do NÓNIO? Refiro-me aos aritgos que são produzidos pelos próprios jornalistas.

JPL: Creio que é onde se notará menos. Os jornalistas são muito condicionadas pela linha editorial da redacção em que se inserem e pelo perfil de público a que querem chegar, ou por aquilo que acham que é o perfil do público a que querem chegar. Dito de outra forma, temos, dentro das nossas vastas audiências, nas nossas e nas dos outros... olhando para os nossos números, temos cerca de 40 mil assinantes, temos registados no NÓNIO (global) entre 1,5 e 2 milhões de pessoas, mas temos a vir aos nossos sites cerca de 4 milhões de pessoas, no total dos sites Expresso e SIC, portanto, há aqui muita gente diferente; agora, os jornalistas sabem sempre, em qualquer momento, que há um core target e há uma linha editorial a seguir, o que não depende do número de pessoas que estão registadas. Portanto, aí não terá impacto.

VL: Mas e se através do NÓNIO se descobrir que os leitores gostam muito de notícias sobre o Cristiano Ronaldo e chocolate. Hipoteticamente, isso poderia levar a que fossem feitos artigos com as duas coisas: Cristiano Ronaldo e chocolate?

JPL: Já temos esse know-how, mesmo antes do NÓNIO. Os nossos sistemas analíticos já nos dão isso. Depois, aqui mais uma vez, há sempre um trade-off. Se nós estivermos à procura daquilo que os consumidores mais querem, ou seja naquilo em que eles mais clicam, naquilo que lhes desperta mais curiosidade, nós aproximamo-nos de uma linha editorial de tabuleiro, porque estamos em cima da sensação. É sem dúvida aquilo que no curto prazo gera mais audiências. Mas, no longo prazo, as pessoas às vezes não gostam. As pessoas muitas vezes não gostam delas próprias. Ou seja, as pessoas não resistem a parar na estrada para ver o acidente, mas depois não gostam do resultado de ter visto o acidente. Depois ficam incomodadas, mas é a natureza humana e nas notícias acontece exactamente a mesma coisa: a pessoa não resiste a ler todo o folhetim da Valentina e o Observador, por exemplo, aí cedeu, eles acompanharam se calhar a linha de um Correio da Manhã e a cada 10 minutos davam as notícias da Valentina. O Expresso não fez isso e, na minha opinião, agiu bem. São cedências que na busca por audiências às vezes se fazem. Mas isso afasta-nos de linhas editoriais. No final do dia há aqui uma linha editorial. Hoje, já todas as redacções conhecem perfeitamente aquilo que faz explodir as audiências, mas sabem que às vezes isso retira-lhes alguma consistência relativamente àquilo de que

as pessoas estão à espera. Se nós damos aquilo que todos estão a dar, entramos na zona da indiferenciação. Tornamo-nos uma commodity completa e as pessoas passam a informar-se em qualquer lado.

VL: Acredita que o NÓNIO e as práticas de Big Data podem levar à sobrevivência económica dos media?

JPL: Em princípio, a Big Data acrescenta valor. A ameaça que vem com ela é a indústria, globalmente, ter perdido uma cumulação de intervenção na cadeia valor de alguns players, sobretudo de um player que é o Google, que tomou uma posição absolutamente dominante na cadeia de valor. E quem tem uma posição dominante na cadeia de valor tem a capacidade de capturar o valor dessa cadeia, portanto, neste momento, a maior ameaça que os media têm no mundo é essa posição dominante que o Google tem, porque o Google intervém em todos os passos. Não vale a pena estarmos agora aqui a detalhar, mas ao intervir em todos os passos (do modo como os conteúdos noticiosos são oferecidos e do modo como a publicidade é oferecida) todos ficam dependentes da sua actuação. Como uma empresa constituída por pessoas inteligentes que é sabem retirar valor accionista dessa posição privilegiada que têm e essa é a parte que mais ameaça no desenvolvimento que a indústria está a tomar, mas não está relacionado com existir ou não existir Big Data. Bastava que houvesse uma lei anti-trust que não permitisse a concentração vertical que foi permitida ao Google, em que acumula sistemas operativos, tem o Android que é o sistema mais dominante do mundo, tem um browser que é o mais dominante do mundo que é o Chrome, tem o email que é o mais dominante, tem o search que é o mais dominante, comprou o Youtube que é a plataforma de vídeo dominante, comprou a double click que faz todo o ad-serving de toda a publicidade, quer dizer, cercou por completo a indústria e isso foi permitido. Geralmente, em muitas indústrias, sobretudo em indústrias que não são globais, ou seja, localmente isto nunca seria permitido. Localmente, há uma autoridade da concorrência que permite este tipo de concentração. Como não há uma autoridade da concorrência para o modo - e aqui já discutimos geoestratégias de eixos americanos vs eixos chineses – até agora o mundo ocidental foi muito condescendente com isto e estamos onde estamos.

VL: Acha que os consumidores estão dispostos a trocar informação privada, os seus dados pessoais, por algo que lhes acrescente valor, neste caso notícias e notícias de confiança?

JPL: Acho que essa questão é uma questão ultrapassada, para ser sincero. Acho que nós prestamos maus serviços a nós próprios porque, por exemplo, nós os dois temos um telemóvel no bolso o dia todo. A Europa entusiasmou-se com o RGPD e foi das coisas mais estúpidas que podíamos ter feito porque levantámos uma série de condicionalismos aos agentes europeus e os dois norte-americanos, Apple e Google, continuam a recolher toda a informação que querem e violam todas as linhas de legislação do RGPD, todas. Não são só algumas, são todas. Os terms & conditions não são explícitos, são demasiado longos, estão feitos para que ninguém os leia. As pessoas que criticam o rastreio dos dados, mas fazem-no no Twitter (pessoas que vão para o Twitter criticar o rastreio dos dados) através de um Android é de um ridículo total. Porque primeiro ligaram um Android que lhes vigia a vida toda. Neste momento até algumas Apps nativas permitem que o microfone esteja ligado por default e ouvem as nossas conversas e conseguem recolher informação que depois pode ser utilizada para efeitos publicitários. Os níveis de invasão são extremos ao nível daquilo que hoje os telemóveis nos fazem. E as plataformas onde essas pessoas estão registadas impõem terms & conditions que as pessoas nunca lêem e que dizem lá que vão ler todas as mensagens emitidas e depois usadas para a construção de um perfil, seja num Facebook, num Instagram, num Twitter, onde for. Mas as pessoas estão lá a criticar o NÓNIO, que ao pé de tudo aquilo que as pessoas aceitaram de forma pouco consciente é uma brincadeira de crianças. Portanto, essa discussão sobre se as pessoas estão disponíveis para aceitar o tracking com os cookies é uma perfeita tontice. Só na Europa é que se discute isso porque houve uma inabilidade total dos media, protagonizada pelos seus próprios jornalistas a darem tiros nos seus próprios pés. Em vez de vermos o mundo em que nós vivemos, falamos das pequenas intromissões comparadas com estas levadas a cabo pela indústria de publicidade dos publishers para que tenha alguma capacidade competitiva.

VL: Mas acha que a indústria já tentou explicar ao consumidor que a única forma de ter notícias gratuitas é dando alguma coisa em troca?

JPL: Mais uma vez, acho que essa discussão está ultrapassada. As pessoas já deram. Quando vamos por aí estamos a colocar mal as questões. Dá a ideia que somos nos que lhes estamos a fazer algo intrusivo, quando nós somos, hoje, os que fazemos a menor intrusão. As pessoas já se expuseram por completo. A não ser que as pessoas se sintam

mais confortáveis a expor toda a sua vida, toda a sua data, em empresas norte-americanas que têm uma jurisdição diferente da europeia… e mesmo assim, porque é distante sentem-se mais confortáveis, só pode ser por causa disso. Porque tudo o resto, essa conversa é completamente descabida ao dia de hoje. Quando nós insistimos nessa conversa, nós estamos a ajudar à percepção de que somos nós que estamos a ser intrusivos quando, comparado com o que elas já aceitaram… Os únicos sites, em toda a navegação que se faz dentro do Chrome, que não captam muita informação são os nossos. Todos os outros captam. Portanto a conversa é completamente absurda. Nós colocamo-la mal.

VL: Que problemas éticos advêm do NÓNIO?

JPL: Já percebeu pelas minhas palavras que não há quaisquer problemas éticos. Se nós conhecermos o mundo em que vivemos, e tivermos consciência do nível de vigilância a que a nossa privacidade está hoje sujeita… é que nós estamos disponíveis para esses trade-offs porque queremos ter o trade-off, depois queremos que a aplicação seja óptima a funcionar e para a aplicação ser óptima a funcionar tem centenas de engenheiros por trás. Quando o Google compra a Waze é porque acumula dinheiro a vender publicidade com base na Data que recolhe. Mas depois adoramos usar o Waze de forma gratuita, como se fechássemos os olhos ao facto daquilo ser pago de alguma forma. Quem é que paga? Nós, com a nossa privacidade. Por isso, o Google sabe, pelos mapas e pelo Waze, mais uma vez nunca devia ter sido permitido que o Google comprasse o Waze porque são os dois maiores players da mobilidade, o Google Maps e o Waze que o estava a desafiar, é mais uma vez aquele movimento de concentração, mas este é um bom exemplo. Toda a gente acha fantástico que o Google é o máximo porque me fornece um Waze gratuito que me ajuda imenso, mas depois ninguém discute este tarde-off e, na minha opinião, neste mundo, se tivermos consciência disso… o NÓNIO é explícito, avisa ao que vem e avisa quais são os segmentos e não detalhe capilar, ou seja, enquanto o Google recolhe informação ao nível individual de grande intrusão, nós estamos completamente distantes disso, nós estamos em segmentos, é completamente distinto. Do ponto de vista ético, zero.

VL: Uma vez que os algoritmos são programados por pessoas, acredita que quaisquer tendências do programador podem vir a manifestar-se nos algoritmos do NÓNIO?

JPL: Não, no NÓNIO não de certeza porque não temos sequer visibilidade sobre os tresholds do output do algoritmo e o treshold é sempre o importante. Ao final do dia, decisões complexas resumem-se a um sim ou não. É tudo binário na nossa vida. É tudo muito complexo, mas depois é faço ou não faço e aqui os algoritmos funcionam exactamente da mesma forma. Tenho esta informação toda, mas ele pertence ou não pertence ao segmento? Qualifica ou não qualifica? É um positivo ou um negativo? Por causa disso, o que acontece é que essas regras têm de ser discutidas em equipas, senão depois o algoritmo não é consistente. O algoritmo tem de ter uma lógica consistente entre a recolha e o resultado que vai produzir que é conseguir agrupar segmentos consistentes. Tem de haver discussões alargadas com as pessoas envolvidas na construção desse mesmo algoritmo, para que ele depois seja válido e produza resultados. Se não for feito dessa forma, se tiver inconsistências na sua construção, não vai agrupar pessoas de comportamento semelhante e essas pessoas não vão reagir de forma semelhante à publicidade exposta e a eficácia não se percebe. Estaríamos a "afinar" sem ver melhorias nos resultados das campanhas. Porque de facto não estamos a juntar as pessoas que pensávamos estar a juntar. Para que isto corra bem, tem de haver uma visão muito mais partilhada de equipas na construção disso e não depende de uma pessoa individual que está ali a fazer uma poção mágica sozinha.

VL: O NÓNIO pode ser um primeiro passo na automação dos media portugueses?

JPL: Acho que isso não está ligado ao NÓNIO. Só pode estar ligado ao NÓNIO se considerarmos que o NÓNIO, ou neste caso o DMP, permite que haja uma discussão conjunta que… essas coisas para serem bem-feitas exigem um grande investimento, pelo que se houver sinergias e houver vários interessados que queiram ir por aí, poderia eventualmente avançar-se com isso. Ou seja, tem muito mais a ver com a colaboração entre empresas que perseguem sinergias e custos partilhados do que com o NÓNIO em si. Não é a Data que o NÓNIO recolhe que pode ajudar a construir melhor essas notícias. Hoje, já vi algumas a funcionar, algumas ferramentas dessas a funcionar, sobretudo para construir conteúdos à volta de desporto. É o mais óbvio. Não estou a dizer mal dos jornalistas desportivos, mas pense nos artigos de desporto. A entrevista pré-match e pós-match é quase sempre a mesma coisa, tal como com as declarações dos jogadores na zona mista. O que perdeu diz que "temos de sair de cabeça levantada, é jogo a jogo, estamos ansiosos para o próximo jogo, ganhamos o próximo, queremos é ultrapassar esta derrota

rapidamente" e o que ganhou, "isto significa que a equipa está a trabalhar bem…" é sempre a mesma coisa, e as notícias são sempre mais ou menos a mesma coisa. Estas notícias salpicadas com a Data, o resultado, é impressionante porque têm muito pouca diferença face àquelas que foram escritas por humanos. Mas estamos no território do mais fácil. Diria que aquilo que as máquinas vão fazer é a commodity, é aquilo onde nós… ou seja, as máquinas vão sempre substituir-nos naqueles comportamentos em que nós nos aproximamos das máquinas. Quando fazemos apenas tarefas repetitivas, é evidente que podemos ser substituídos por máquinas e se calhar até devemos vir a sê-lo no futuro. Quando os jornalistas se limitam a dar os factos seguindo uma estrutura rígida estabelecida e é apenas isso, é evidente que uma máquina vai fazer isso melhor do que os humanos. Mas as máquinas também vão fazer operações ao coração melhor do que os cirurgiões, ninguém tem dúvidas disso, não estou aqui a menosprezar o trabalho de ninguém. Tudo o que for repetitivo, tudo o que for apenas treino, as máquinas terão sempre vantagem. Pode é não haver máquinas que tenham as capacidades para executar determinadas tarefas, mas quando as tiverem, vão executá-las melhor que nós. Quando metemos criatividade, aí já entramos nos temas da inteligência artificial. E em 98% dos casos em que se fala de inteligência artificial não há inteligência artificial nenhuma. É apenas automação. Inteligência artificial no seu estado mais puro pressupõe que haja alguma criação de algo novo com base em coisas anteriores, tal como o nosso cérebro funciona, mas só aí é que entramos na IA. Não é na repetição pura, aí é forçado dizer que se trata de inteligência artificial só porque estamos a utilizar Big Data e repetição. Os computadores já jogavam todos xadrez e eram tão bons quanto o jogador que tivesse treinado o programa, não conseguiam era batê-lo. Não dizíamos que esses computadores recorriam a inteligência artificial. No limite, a automação não está ligada ao NÓNIO, e a automação, vir a aparecer, será em zonas onde as pessoas acrescentam pouco valor.

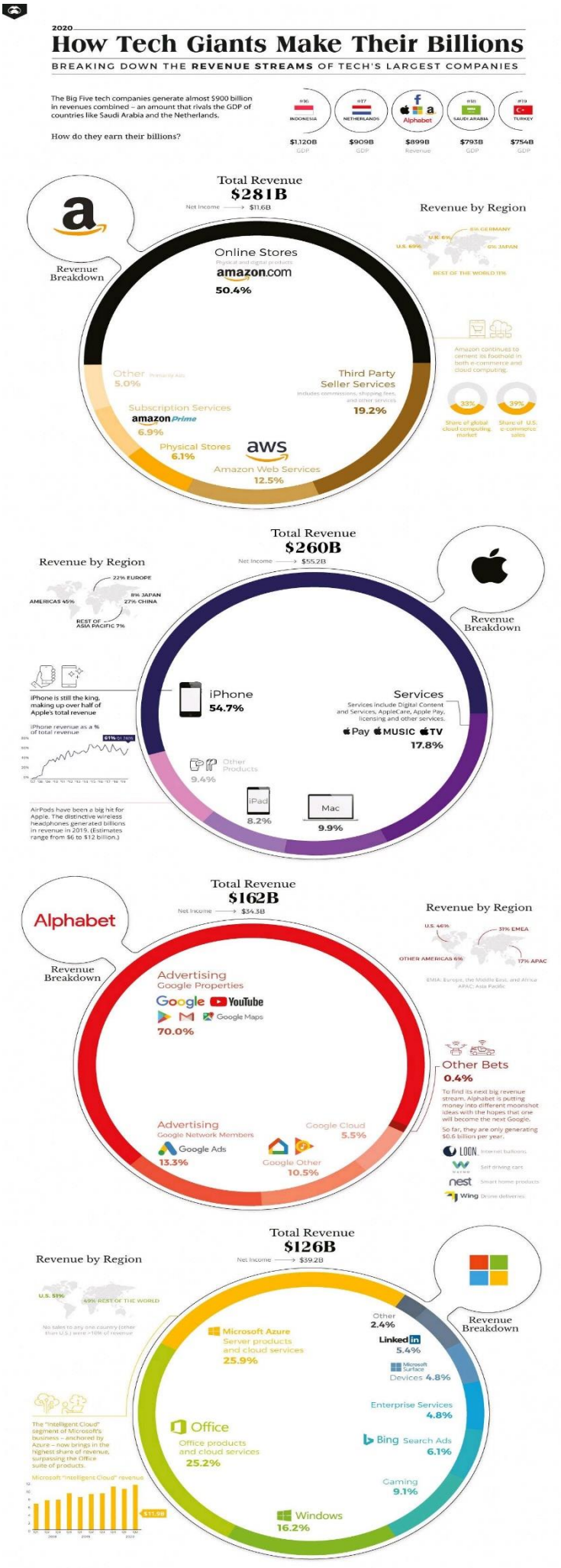

Figure 1 – How Tech Giants Make Their Billions
Source: Visual Capitalist