Universitat Oberta de Catalunya

artnodes
E-JOURNAL ON ART, SCIENCE AND TECHNOLOGY

https://artnodes.uoc.edu

ARTICLE

NODE "POSSIBLES"

A critical approach to Machine Learning forecast capabilities: creating a predictive biography in the age of the Internet of Behaviour (IoB)

Diego Díaz
Universitat Jaume I (UJI)

Clara Boj
Universitat Politècnica de València (UPV)

Date of submission: October 2022
Accepted in: December 2022
Published in: January 2023

Recommended citation

Díaz, Diego; Boj, Clara. 2023. "A critical approach to Machine Learning forecast capabilities: creating a predictive biography in the age of the Internet of Behavious (IoB)". In: Pau Alsina & Andrés Burbano (coords.). «Possibles». *Artnodes*, no. 31. UOC. [Accessed: dd/mm/yy]. https://doi.org/10.7238/artnodes.v0i31.405249

Abstract

Based on the notion of the Datacene, understood as the time when data directly affects the social, cultural, economic, political, and even affective structures of the present, in this article we propose how Big Data and Artificial Intelligence give rise to the Internet of Behaviour: a new technological paradigm that has incredible potential to forecast and induce human behaviour. Since ancient times, humans have wanted to predict and alter the future, but in the last ten years, this wish has begun to become a reality due to great advances in the field of social engineering, raising serious doubts regarding social control and the loss of freedom. In this context of analysis, we present two projects developed within the framework of Art, Science, Technology and Society. *Data Biography* shows

the enormous number of digital traces that we generate daily and uses them to compose a person's biography, composed of 365 printed books. *Machine Biography*, for its part, investigates how current artificial intelligence techniques can predict and induce future human behaviour, for which we have used various forecast and generative models trained with data from our own digital activity, in order to generate another set of books with our foreseeable activity for the year 2050. Both projects invite us to consider from a critical perspective the present and future of the social transformations produced by Big Data and AI.

Keywords

prediction; Artificial Intelligence; new media art; Internet of Behaviour; Datacene; Big Data

Un enfoque crítico para las capacidades de pronóstico del aprendizaje automático: crear una biografía predictiva en la era del Internet of Behaviour *(IoB)*

Resumen

Basándonos en la noción de dataceno, entendido como la era en la que los datos afectan directamente a las estructuras sociales, culturales, económicas, políticas e incluso afectivas del presente, en este artículo proponemos cómo el big data *y la inteligencia artificial dan lugar al* Internet of Behaviors *(IoB), un nuevo paradigma tecnológico con un potencial increíble para pronosticar y moldear la conducta humana. Desde la Antigüedad, los seres humanos siempre han querido predecir y alterar el futuro, pero en los últimos diez años este deseo ha empezado a hacerse realidad tras los grandes avances experimentados en el campo de la ingeniería social, lo que plantea serias dudas sobre el control social y la pérdida de libertad. En este contexto de análisis, presentamos dos proyectos desarrollados en el marco del arte, la ciencia, la tecnología y la sociedad.* Data Biography *muestra la enorme cantidad de huellas digitales que generamos diariamente y escribe, a partir de ellas, la biografía de una persona, compuesta por 365 libros impresos. Por otro lado,* Machine Biography *investiga cómo las técnicas actuales de inteligencia artificial pueden predecir y moldear el futuro comportamiento humano, para el que hemos utilizado varios modelos generadores y de previsión entrenados con los datos de nuestra propia actividad digital, para producir otro conjunto de libros con nuestra actividad previsible para el año 2050. Ambos proyectos nos invitan a considerar desde una perspectiva crítica el presente y el futuro de las transformaciones sociales producidas por el* big data *y la IA.*

Palabras clave

predicción; inteligencia artificial; arte de los nuevos medios; internet del comportamiento; dataceno; big data

Introduction

Prediction has always been one of the main objectives of the human species. The ability to anticipate was a key factor for human survival in primitive times. The constant observation of atmospheric conditions, with the intention of finding certain behaviours that could serve as a guide to know the best time of year to sow, was one of the great advances in knowledge of the weather that allowed for the development of agriculture in the Neolithic period, about 10,000 years ago. At that time, prediction was probably based on analytical methods, and some specialists in the arts of divination would soon begin to develop the ability to modify future behaviour, such as affecting the weather and thereby benefitting crops. Some of these so-called magicians, the most intelligent and wise, knew their limitations but took advantage of deception in order to obtain significant social privileges (Frazer 1922).

This operating structure has been adapted to different civilizations throughout history. In each of them, we find different divinatory arts: geomancy, palmistry, Tarot, scrying, divination with corn kernels and so on, until we reach such highly developed social systems as the Oracle of Delphi.

A particularly interesting case is the evolution in the prediction of atmospheric phenomena, where we can see a progressive improvement in the understanding of the phenomena of nature and consequently in its prognosis. This process begins in the year 340 BC, when Aristotle defined the term "meteorology" in his Meteorological book (Aristotle n.d.), which described, for the first time, observations and speculations about the origin of atmospheric and celestial phenomena. However, it was particularly from the 16th century onwards, with the inventions of

the anemoscope by Leonardo Da Vinci and the thermometer by Galileo, that the science of meteorology began to develop progressively up to the present day. The progressive increase in data collection on the behaviour of nature due to the invention of certain scientific data collection technologies (thermometer, barometer, anemometer, etc.) made it possible to improve the prediction of atmospheric conditions using probabilistic methods. Through this process, we can see how the ancient divinatory methods have evolved to the science of current prediction with a good level of precision. Currently, it is considered that atmospheric predictions are quite reliable up to three days in advance.

In the last ten years, the social sciences, such as meteorology, have undergone an incredible evolution in their predictive methods. At the same time, the probabilistic analysis of human behaviour is becoming more accurate thanks to the increasing numbers of digital traces provided by Big Data. In addition, for some years now we have observed in amazement how Big Data, together with computer psychology, has been able to affect some key decisions in our recent history, such as the 2016 United States elections, which resulted in the presidency of Donald Trump, or the "Yes" in the Brexit referendum held that same year. In both cases, Cambridge Analytica (Anderson 2017) used a powerful, never-before-seen social engineering system that had the ability to modify human behaviour. From this moment on, we ask ourselves: could it be that those magicians who, since the Neolithic era, aspired to control the future of our civilization have finally found a system by which they can really achieve their aspirations? Is it possible that machines can predict our future? And, most importantly, what are the social implications of these new predictive possibilities?

1. Scientific prediction in the context of the Internet of Behaviour (IoB)

The scientific predictive method tells us that if A is true then B will also be true, as long as this rational structure is formed by a logical system based on the truth test, meaning that any experiment has to be reproducible in order to ensure its veracity. Obviously, all experiments must limit the field of study, since it is impossible to carry out an analysis that encompasses all of nature as a whole. This partial sense of reality works with bounded models and hidden variables in order to simplify reality and be able to carry out multiple partial tests of it. We are, from a philosophical point of view, looking at a model that aims to establish knowledge by rendering reality real. Understanding the real as that to be studied and, therefore, that which is not yet experimentally proven, such as some concrete phenomenon of nature. Thanks to empirical procedure, as we analyse this, we can contrast some of its elements or dynamics and render them a reality.

If we look at atmospheric predictions, the models are always an idealization: a simplification, given the impossibility of covering all the variables in the environment. However, as we mentioned above, these predictions are correct only with short notice, currently up to three days; as Edward Lorenz (Lewis 2005) showed, a minimal change in the adjustment of the initial variables of a system completely changes the result of the same depending on how far in advance this is done. This occurs particularly in chaotic systems, which, according to chaos theory (Cattani *et al.* 2017), are strictly deterministic, since their behaviour can be entirely determined by knowing their initial conditions, but their long-term prediction is impossible. In addition, we must mention here, random is an accidental encounter (Aristotle n.d.), understood as that which is not predictable, since it establishes a chance generated by phenomena characterized by complex, non-linear causes and unpredictability. The processes that coincide are independent of each other and share no causal relationship, although each has a cause that acts in a necessary and independent way. Thus, a flowerpot falls due to a necessary cause – gravity – but it is by chance if, in its trajectory, it collides with a pedestrian.

Social sciences, on the other hand, are an area of study where scientists have traditionally had more difficulties in predicting human behaviour, because even in the case of being able to know all the variables of the system, traditionally it cannot be determined with accuracy regarding the relationship between them, since they are highly complex dynamics and are composed of systems with high dependence on the initial variables, as in the aforementioned case of meteorological prediction, and the high probability of the aforementioned "random as accidental encounter".

However, some of these limitations have changed in recent years with the emergence of the Internet of Behaviour (IoB) (Elayan, Aloqaily, & Guizani 2021) understood as the evolution of the Internet of Things (IoT) that reveals and processes significant information about the user. The IoB interprets data extracted from the user from behavioural psychology viewpoint, giving way to knowledge of a user's behaviour, which helps to know the consumer in more detail and better understand their needs and desires. The fundamental objective of the IoB can be summarized with the phrase "if we know the behavior of the user, we can more easily influence their decisions, thus erasing the fine line that separates the ability to predict and induce human behavior" (ibid). For example, if we know that a user participates in a certain type of activity two days a week, but also know what specific days these are, where, what type of activity, the start and end time, with which other users this coincides, and so on, we can select, design and offer information (whether truthful or not) focused on this specific user profile. In this way, we could reinforce their confirmation bias (Schumm 2021), understood as people's tendency to seek information that supports the points of view they already hold and generate a powerful dynamic that can ultimately modify their behaviour based on our predictive interests. Broadly speaking, this was the mechanism used by Cambridge Analytica (Brown 2020) to alter the behaviour of voters in the 2016 US elections and the "Yes" in the Brexit referendum. Thanks to the use of confirmation bias and other persuasive techniques of human behaviour, we can observe how certain predictions come true due to the high capacity

to induce behaviour that the IoB provides. Behavioural psychology has an incredible power in social control by using emerging technological innovations and developments in machine learning algorithms. These new technologies are both descriptive and proactive, which means that they help to analyse as well as to detect which variables to influence in order to achieve a certain result in the end user.

Concerned about the potential dangers of behavioural psychology for contemporary democracies, a group of 17 internationally renowned scientists from different disciplines recently published an article in the prestigious journal PNAS (Bak-Coleman *et al*. 2021), in which they shared their concerns about the social dangers that the rapid development of the digital society and the emergence of social networks may bring about, having accelerated changes in our social system with poorly understood functional consequences. This "gap in our knowledge represents a principal challenge to scientific progress, democracy, and actions to address global crises". Such is their concern that they "argue that the study of collective behavior must rise to a "crisis discipline" just as medicine, conservation, and climate science have, with a focus on providing actionable insight to policymakers and regulators for the stewardship of social systems".

2. Datacene: the data era

We can understand the Datacene (Díaz & Boj 2019) as a metaphor for the society mediated by data, in which immense quantities of heterogeneous data are recorded, stored and analysed. Any activity carried out by a user is digitally recorded and stored forever in the Big Data cloud. These digital traces continue to feed the insatiable voracity of the Datacene, where not only is human activity recorded, but also an increasing number of "intelligent" devices are connected to the network (IoT), communicating their status and sharing data from a multitude of sensors that digitize the physical world to the extreme, giving rise to a kind of hybrid space where literally any entity in the physical world has a digital double in virtual space.

In just 30 years, our world has gone from being completely analogue to being almost completely digitized. This fascinating evolution has forever changed the way humans live. Currently, we can always know, with the accuracy of GPS, where a person is and communicate with them through instant messaging, voice, or video, altering our perception of physical distance in an increasingly smaller and more interconnected world.

The evolution of wireless communications has been essential for the development of Datacene 4G networks, and 5G in particular, along with the implementation of protocols such as IPv6, has made it possible to create a solid infrastructure in which devices can receive and send information. Each packet of information is encapsulated in different protocols inheriting the TCPI/IP model developed in 1973 (Cerf & Kahn 1974), which contains, at a minimum, a timestamp at which the event occurred, the IP address of the device where it was created, the IP address of the recipient, and the content of the message itself. At present, other usual additions to the communications between devices include their geolocation, information related to the user (username, phone number, email, etc.), the status of the device, and so on. The study of communication protocols shows us, along with the content of the message, a series of metadata that can reveal, after analysing and crossing data from other devices and users, very valuable information about ourselves. We can deduce, for example, the typology of the most visited places, preference of activities carried out, means of transport used, consumption habits, and so on, which added to the sophisticated analysis of use and communication in social networks can provide a very descriptive image of any network user. This is highly significant if we consider that in the year 2022, the number of internet users reached almost 5 billion, representing 62.5% of the world population. There are two clearly differentiated regions: the highest internet penetration today is Northern Europe (98%) and Western Europe (94%), followed by North America (92%). The regions with the lowest penetration are Central Africa (24%) and East Africa (26%) (Kemp 2022).

The increase in the generation of digital data that feeds the Datacene seems endless; in the year 2020, the total volume of data created, captured, copied and consumed was 64 zettabytes, and by 2025 it is estimated to be 181 zettabytes (Holst 2021). Although only a small amount of this data is kept – two percent of the data produced and consumed in 2020 were still saved by 2021 – in line with the strong growth in data volume, storage capacity is also estimated to increase. In 2020, storage capacity reached 6.7 zettabytes, and in 2025 it is projected to be almost 8 zettabytes (Idem).

In this society mediated by data, a fundamental aspect that we must consider is the ownership of the infrastructures that support it and the volume of business that drives this industry. The main company that controls this huge business is Amazon with 33%, earning more from the cloud than from e-commerce, followed by Microsoft (21%), Google (8%) and, further afield, Alibaba (4%) and IBM (Statista n.d.). When we grant consent for our files and documents to be hosted on remote servers, we accept the rules proposed by the provider companies and, although the European personal data protection regulations are stricter than ever (European Commision n.d.), the dilemma of privacy and intellectual property of files in the cloud remains unresolved. To whom do files stored in the cloud belong? For example, in its data policy, Google warns that the user gives it permission to upload, send, store, send, receive or share content (Google n.d.). With regard to privacy, to further complicate things, servers are not usually in the territory of the European Union, as most of the companies providing these services are American, where the Cloud Act (Callaway & Determann 2018) approved in 2018 allows US authorities, for national security reasons, to request any type of data and information stored on the servers or cloud of a US company or provider, regardless of the country in which it is located and whether they are natural persons or legal entities. Therefore, practices such as the recently uncovered NSC scandal with Pegasus spyware are increasingly common and even enjoy the legal backing of US law (Corera n.d.)

We must not forget that the physical infrastructure that sustains the Datacene is made up of a distributed and interconnected network on a

planetary scale of physical servers located in data centres, in addition to the multitude of electronic devices that collect and send data and that support the Internet of Things. It is estimated that today there are one hundred million servers housed in gigantic data centres, connected by almost a million kilometres of fibre optics. To have a reference magnitude on the energy impact of these Data Centres and the infrastructure that provides access to the Internet, it is estimated that they currently consume around 5% of global energy. It is estimated that by the year 2025, due to the constant increase in the number and power of electronic devices, this network will use 20% of all the electricity in the world and will emit 5.5% of all carbon emissions (Andrae & Edler 2015), not to mention the enormous requirements in energy and raw materials necessary for the manufacture of electronic devices.

Another key aspect to consider in the era of the Datacene revolution is the knowledge turn associated with it, where the Datacene challenges established epistemologies in the sciences, social sciences and humanities and generates new paradigm shifts in multiple disciplines. The true protagonist of this hybrid space that makes up the Datacene is the massively tirelessly generated data that fuels Big Data. Hand in hand with Big Data are emerging new forms of empiricism that declare "the end of theory", the creation of science based on data instead of knowledge, and the development of digital humanities and computational social sciences that propose radically different forms of making sense of culture, history, economy and society (Kitchin 2014).

3. Data Biography case study: writing biographies in the Datacene era

At the time of the Datacene, we could affirm a priori that we are at the perfect time to develop the realistic biographical model, due to the immense personal information accumulated by Big Data. This apparent objectivity, where the information in principle is shown to us as an optimal resource for tracing a person's biographical account, from which we can later, using predictive algorithms, artificial intelligence and semantic analysis (among other computational methods), extract and generate results relating to psychological profiles, desires, behaviour, and so on, in addition to having precise information on the location of the biographer and places that they have visited.

That said, this vision is at the very least simplistic, since the quantitative record of our actions can be biased and partial and induce erroneous interpretations about the life of the subject to be biographed (Glauner, Valtchev & State 2018). In the same way that the criticized slogan "know your numbers to know yourself" of the Quantified Self (Reigeluth 2014) promises us an improvement in personal health and well-being through intensive monitoring of our biometric data, data biographies can show us a reductionist and biased interpretation of the biographed subject.

The artistic project *Data Biography* (Díaz & Boj 2019) endeavours to compose a global physical biography, with 365 books printed in a single edition (one per day) as a compilation of the trace of the digital data we generated throughout 2017, when we hacked our mobile phones with a spyware software (mSpy n.d.), capturing all our digital traces in order to generate the Data Biography project. This biography is made up of 365 printed books (one per day, with a total of more than 40 thousand pages) and a 24-hour film narrative (showing the 8760 hours that make up the entire year).

Figure 1. *Data Biography*, Díaz & Boj 2019
Source: own creation

If a biography is a person's life narrated by another and consigns those aspects of his life most relevant and everything that, in the eyes of the rapporteur, may be of interest, this research attempts to make a speculative design proposal on how the biography located at the current moment could be traced using the data captured from our digital trail on social networks, WhatsApp, emails, visited websites, google searches, images, GPS location, and so on.

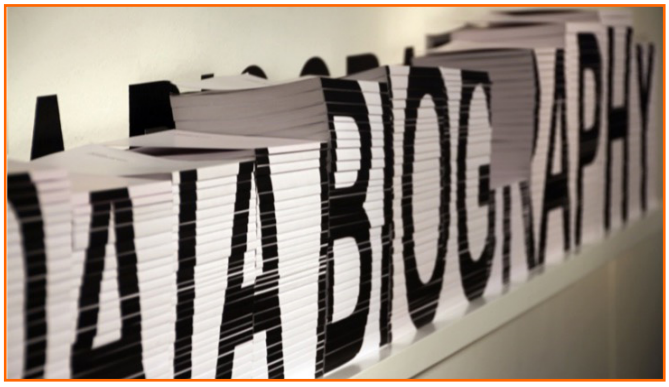

Figure 2. *Data Biography*, detailed view, Díaz & Boj 2019
Source: own creation

Each of the 365 books is numbered correlatively with a total of about forty thousand pages; this immense volume of books shows the enormous amount of data that we generate and donate to Big Data daily. Located on a shelf where the title of the work can be read, each column details a month, each book a day and each line an action carried out, where the temporary record and the different data of this are shown: the type of action (GPS location, email, WhatsApp, Facebook, Skype,

etc.) and its content. The videos and photographs are shown in image format, which generates a list of information that allows the visitor to read all the actions carried out by Díaz and Boj over the course of 2017.

For a better understanding of how we organize and manage this large volume of information, Figure 3 shows a complete list of the types of registered actions and the number of entries registered for each of these types throughout 2017. For a more descriptive visualization, we have excluded the GPS Location records with 351382 different entries.

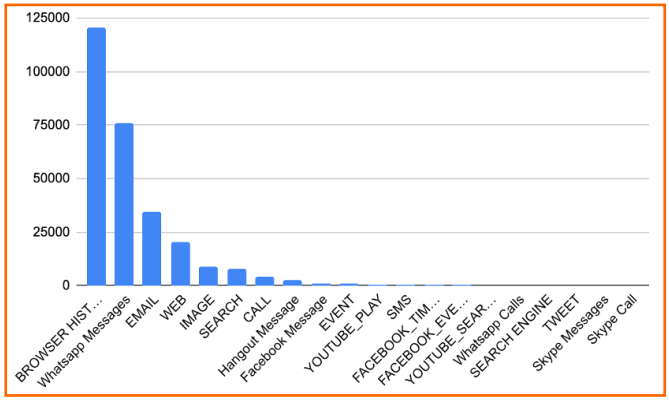

Figure 3. *Data Biography*, event types and numbers of records, 2017 (except GPS location with 351382 records)

Source: own creation

All these types of actions are registered under a standardized system with the structure shown in Table 1:

DATA	FORMAT - CONTENTS
Date	year-month-day hours:min:sec
Type	GPS Location, BROWSER HISTORY, WhatsApp Messages, EMAIL, WEB, IMAGE, SEARCH, CALL, Hangout Message, Facebook Message, EVENT, YOUTUBE_PLAY, SMS, FACEBOOK_TIMELINE, FACEBOOK_EVENTS, YOUTUBE_SEARCH, WhatsApp Calls, SEARCH ENGINE, TWEET, Skype Messages, Skype Call
Author	Clara, Diego
end Date	year-month-day hours:min:sec
Who	tlf number, username
remoteWho	tlf number, username
Duration	hours:min:sec
Direction	Outgoing, Ingoing
Latitude	gpsLat
Longitude	gpsLong
Text	text
URL	url

Table 1. *Data Biography*, data structure of each event

Source: own creation

Thanks to the use of this normalized model, the list of registered data is saved in a CSV file where, depending on the type of data, some fields are empty, since, for example, the registration of a GPS position only fills in the fields of date, type, author, Latitude and Longitude while the rest is left blank. The final appearance of the records printed in the books can be seen in Figure 4.

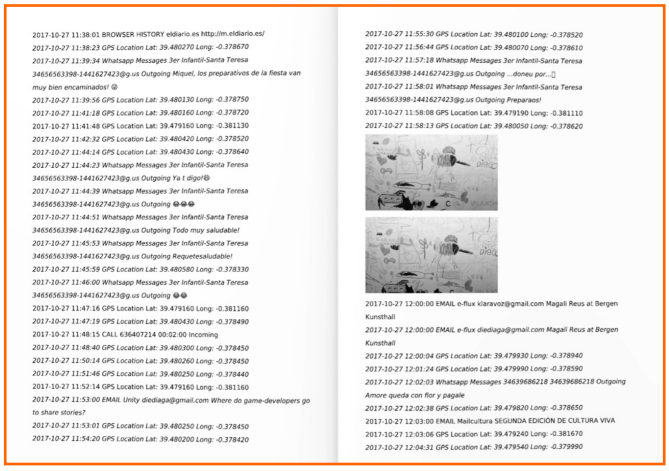

Figure 4. *Data Biography*, example of data printed in a book

Source: own creation

Finally, in this section, we would like to show the following graph (Figure 5), which displays the number of entries per day classified according to their type. A clear seasonality and a significant decrease in digital activity can be observed during the month of August.

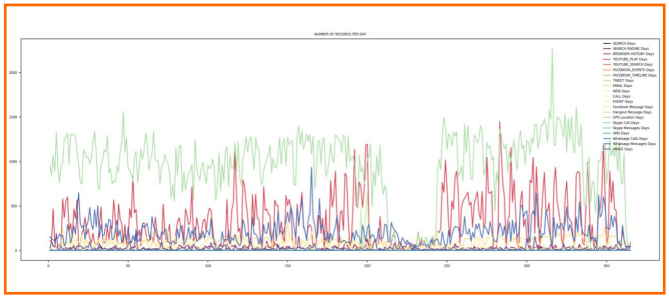

Figure 5. *Data Biography*, graph with the number of entries registered per day ordered according to their type

Source: own creation

4. Scientific predictions and determinism in the Datacene era

With the emergence of the Datacene, some scientific theories have gained popularity as Laplace's demon. Published in 1814, this was the first articulation of causal or scientific determinism. According to the author, if someone (the demon) knew the precise location and moment of each atom in the universe, the past and future values of it for any given time would be deductible from those data and could be calculated using the laws of classical mechanics (Laplace 1902). Followers of Big Data have recognized the figure of Laplace and raised this deterministic possibility from the data, where we have a noisy vision of reality provided by the immense amount of data generated by and stored in Big Data. Although it is true that both quantum mechanics and chaos theory have seriously criticized Laplace's mechanistic model (Jishi 2013), but the development of Big Data and Artificial Intelligence technologies, as well as neuroscientific advances in our knowledge of the brain and how "the human decision" is generated, are revitalizing a certain deterministic model that strongly reminds us of Laplace's theory.

From a deterministic point of view, a good part of the data is noise (Shannon 1949), which prevents us from accessing the signal, so the job consists of cleaning this data to be able to access the information it contains and the patterns it conceals. The limitation of this theory resides in the fact that in practice, mathematics falls short in the analysis of society, and from propositional logic it is impossible to demonstrate certain statements since there are more propositions than demonstrations, so not everything can be demonstrated, even if the initial hypothesis is reasonable. We are faced with a problem that is undecidable and cannot be validated due to the impossibility of proving or refuting a certain predicate based on others. An example of this is the Super Bowl impossibility theorem (Skiena 2017), by which it is stated that the true effects of advertising on the sales of products advertised during this famous event cannot be accurately measured. The effect of advertising is subject to a lot of noise, so it is impossible to deduce its direct consequence; the effect of this action can be neither affirmed nor denied, since in real life the parameters that can affect it are so many that it is impossible to create a mathematical model that can consider them all.

Although it is true that it is almost impossible to create a mathematical model of a complex real social system, current models are increasingly broad and powerful, with a greater number of parameters and input variables. These new models benefit from machine learning algorithms and their ability to autonomously learn from the analysis of massive data obtained from the real world. Examples include recurrent neural networks (RNNs), which are an excellent system for finding the signal hidden in the data, as they are great at finding patterns and correlations of behaviour (Mohajerin & Waslander 2019).

On the other hand, as our computational tools have become more advanced, they have also become opaquer. Their internal functioning, in the case of deep neural networks, is hidden and no one can be a true connoisseur of it; they behave like a black box that can lead to unforeseen results. In this sense, some data scientists believe that we are entering an age where alchemy has relegated science to second place. In the words of Ali Rahimi: "We are building systems that govern healthcare, mediate our civic dialogue and influence elections. I would like to live in a society whose systems are built on verifiable, rigorous and exhaustive knowledge, and not on alchemy" (Synced n.d.). For Elon Musk, the biggest concern relates to the AI that lives in the network and that could autonomously harm humans. "[AI] could start a war by making fake news and falsifying email accounts and fake press releases, and simply manipulating the information" (Alex Hern 2017). In a sense, this sci-fi scenario has already occurred, albeit without transcending the digital world, when in November 2017 the artificial intelligence of the video game EVE Online spiralled out of control and unleashed a battle between three cosmic flotillas with no human participation (Mildenberger 2013). In that sense, as Danaher 2016 points out, the use of algorithm-based decisions in the public and political spheres can be problematic, because current machine learning algorithms do not respond to a sequenced and predictable structure of orders. These algorithms need large datasets to train their operation, and if these data have a certain bias, it is known that the AI model generated from them will also reproduce the same bias. Therefore, the value of these algorithms falls largely on the datasets used for their training, since they are responsible for their proper functioning.

5. *Machine Biography* case study: forecasting human behaviours with machine learning in the age of the Internet of Behaviours (IoB)

In recent years, we have witnessed a true revolution in the field of Artificial Intelligence thanks to advances in Machine Learning, where machines have managed to learn on their own after analysing large amounts of data. The rapid evolution of these algorithms, their versatility and their incredible potential are affecting our society in many different areas, one of the most controversial of which is their ability to generate time-based forecasts. As mentioned above, a deterministic approach to Big Data forecasting models based on Laplace's theory is gaining in popularity and, in our opinion, a more in-depth investigation of its effects at the social level is needed. With a view to critically exploring the potential of that approach, in this research, we have produced a predictive biography for the year 2050 of Díaz and Boj using Machine Learning.

To create this predictive biography, we developed several Artificial Intelligence models that were trained with all the digital data of Díaz and Boj (GPS locations, digital conversations, photographs, videos, etc.) collected during 2017 for developing the Data Biography project described in section 3. With these datasets, we trained different Machine

Learning models to generate our foreseeable digital activity for the year 2050 and reprinted the resulting data to create another set of 365 books that constitute the artwork *Machine Biography*: a predictive biography created by artificial intelligence with the objective of analysing the forecasting capacity of algorithms and inviting consideration of their social impact in our society.

Figure 6. *Machine Biography*, Díaz & Boj 2020
Source: own creation

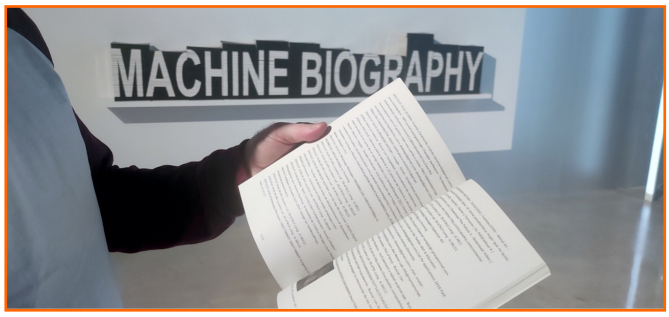

Figure 7. *Machine Biography*, Díaz & Boj 2020, book detailed view
Source: own creation

Before addressing the description of the models that we have used in our project, we believe it is necessary to clarify the difference between the terms "forecast" and "prediction" in the context of Machine Learning models. A predictive model finds and connects patterns in data relationships in order to obtain certain results - for example, a typical application of this model is for estimating and labelling the elements of an image. So, prediction in Machine Learning is not always necessarily about the future, but it can be, especially if we are using forecasting models where the objective is to estimate future events through the analysis of past data. A typical forecast model is time-based, where a value is associated with a timestamp and the dataset is adjusted, at a given time and value, with a future time and value. The training dataset has a forecast structure where a paired example with multiple variables of current data and future data allows the model to discover correlations between the data, as it is typically used in stock markets and energy consumption.

Our dataset is classified into seven different types of events: search, browser history, tweet, email, GPS location, WhatsApp message, and image. All these events share the structure used in the *Data Biography* project, described in section 3, Table 1. We have reduced the event typology from that of the previous project, from 21 to 7, because, as Figure 3 shows, if we order the events by the number of records that each one contains, beyond the first 7, the rest is mostly testimonial.

Our first step in forecasting when and what time events will happen was to train the main forecasting model responsible for generating the temporary events associated with a specific typology. The output data consists of two variables: the time when the event will be generated, and one of the 7 typologies of the event (Table 2). To generate that prediction, we used a model based on Time Series Forecasting with Long Short-Time Memory Convolutional Network (CNN-LSTM) multivariate and multi-steps (Wan *et al*. 2019). This model also uses embedding to group similar events into homogeneous groups: for example, to embed weekdays, months, day-night periods, and so on, to help the model learn the correlations between data. By running this first model, we obtained a list of events only with time and typology.

EVENT TYPE	TYPE OF INFORMATION	AI MODEL USED	AI MODEL TAXONOMY
Event	Time and text	Time Series Forecasting CNN-LSTM	Forecasting
GPS Location	text	Time Series Forecasting CNN-LSTM	Forecasting
Search	text		Generative text
Browser history	text		Generative text
Tweet	text		Generative text
Email	text		Generative text
WhatsApp message	text		Generative text
Image	image		Generative image

Table 2. *Machine Biography*, description of event typologies and AI models
Source: Own creation

Next, to generate the specific contents for each of these events according to the 7 event typologies, as seen in Table 2, we used two general model taxonomies: generative and forecasting. Generative models, which are popular nowadays, include text-to-image models, text-to-text, and the like; they are a revolution in image and text generation and editing methodologies. However, due to the focus of our research, we are focusing on the analysis of human behaviour predictive models,

especially in time series forecast models, since their application could generate a radical and alarming change in our contemporary societies, as we discussed in section 1. For example, we found it could be very alarming that we could create a GPS location forecasting model where we could obtain the foreseeable latitude and longitude of a person at a given time. To create this model, we used a modification of the one described above (Ibid), where the neural networks learned to forecast our possible location according to the future time of the event.

On the other hand, to generate data for the rest of the event typologies, we used other approaches based on generative models. For example, to generate texts, we finetuned (Gatti & Mathov n.d.) the GPT-2 model trained by the Spanish Flax community (Flax Community n.d.). The dataset used for training and finetuning has been created from different sources: namely, more than 200,000 tweets downloaded from the Twint library (Twint n.d.), selected under different concepts related to the most important concerns associated with the year 2050 (OECD n.d.). We have modified the model interface code in such a way that it carries out a conversation with itself, concretely reintroducing in the model the last sentence of the generated statements. To analyse the sentences and automatically subdivide and extract the last phrase, we used the NLTK (NLTK n.d.) library.

For creating and editing images, we used different generative models. First, we used Dall-E Mini (Dalle-Mini n.d.), a text-to-image generative model that creates images from a selection of the texts generated with the finetuning GPT-2 model. These images were inserted in the final dataset together with the texts that generated them. On the other hand, for editing the images appearing in the 2017 dataset and giving them a retro-futuristic touch, a style transfer model was used: specifically, a CUT model (Jay *et al*. 2017). This model follows a GAN structure with a generator and a discriminator, with the peculiarity that the discriminator is responsible for comparing patches of the image rather than all of it. Our custom CUT model was trained with more than 20,000 images: half of them images from the video game Cyberpunk (cd projekt red n.d.) set in the year 2077, and the other images extracted from videos of people walking through real cities. Furthermore, for the aging of faces that appear in the pictures from 2017, we used the model Only a Matter of Style: Age Transformation Using a Style-Based Regression Model (Alaluf, Patashnik, & Cohen-Or 2021). Finally, the images were upscaled with the Fast-SRGAN super-resolution model for upsampling (Raza n.d.).

Machine Biography is a speculative exercise in the predictive possibilities of artificial intelligence. The focus of this artistic research is not to assess the level of accuracy of the models developed. The research is framed within the artistic context, so we use open-source forecast and generative models and adapt them to our purposes. The models to which we had access were very basic, since the truly powerful models and large datasets are owned by a limited number of companies, the so-called GAFA: Google, Amazon, Facebook, and Apple, which support their model of business and therefore are jealously guarded. The ultimate purpose of this artistic project is to create a predictive biography that materializes in 365 books and that, as an exhibit, invites visitors to reflect on the use of these technologies and their enormous transformative capacity on a social scale.

Conclusion

In this publication, we have presented two research projects developed in the framework of Art, Science, Technology and Society that invite us to reflect on the social transformations exercised by the Internet of Behaviour at the time of the Datacene. In the case of *Data Biography*, the work challenges us: if we had to write a biography of a person today, what would be the most appropriate means of gathering information about them? And, more specifically, what form could such a biography take? We used this question to structure the entire research methodology that gave rise, after the study of Big Data, to the compilation of our digital traces and the formal concretion of the resulting artistic work: a Data Biography, composed of the record of all our digital activity during 2017.

This same structure of thought was used in 2019 for the *Machine Biography* research. In this case, the fictitious starting hypothesis was: can machines predict our future? From this perspective, we constructed the final work, composed of a predictive biography of our activity for the year 2050. The technology used for the construction of this biography was various forecast and generative models of Machine Learning trained with our digital traces collected during 2017. The objective of this research was to reflect critically on the paradigm shift brought about by recent improvements in the predictive capacity of human behaviour in social engineering. This new and powerful weapon of social control is still in a primitive phase of development, but its enormous potential means it is necessary to study and discuss its possible transformative actions. Based on these premises, we can ask ourselves if we are free or if we are subject to a destiny, to a chain of causes and laws (physical, psychological, sociological). We are growing ever closer to the superhuman intelligence of which Laplace spoke. Artificial Intelligence and Big Data have set the course for total knowledge, with the ability to have total vision and predict all kinds of events. What use, then, would be our efforts to alter the Delphic predictions of mathematical calculation, our attempts to correct that which is bound to happen?

Acknowledgements

This research was financed with the support of a 2019 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation and by the research project of the Ministry of Science and Technology PID2019-106426RB-C32 / AEI / 10.13039/501100011033 and PDC2021-120997-C31 / AEI / 10.13039/50110001103.

References

Alaluf, Yuval, Or Patashnik, Daniel Cohen-Or. "Only a Matter of Style: Age Transformation Using a Style-Based Regression Model". *ACM Transactions on Graphics*, vol. 40, no. 4 (2021). DOI: https://doi.org/10.1145/3450626.3459805.

Anderson, Berit, and Brett Horvath. "The Rise of the Weaponized AI Propaganda Machine". *Medium* (February 13, 2017). https://medium.com/join-scout/the-rise-of-the-weaponized-ai-propaganda-machine-86dac61668b.

Andrae, Anders, and Tomas Edler. "On Global Electricity Usage of Communication Technology: Trends to 2030". *Challenges*, vol. 6, no. 1 (2015). DOI: https://doi.org/10.3390/challe6010117.

Aristotle. Meteorology. (350 B.C.). *The Internet Classics Archive*. [Accessed: October 27, 2018]. http://classics.mit.edu/Aristotle/meteorology.1.i.html.

Bak-Coleman, Joseph B., Mark Alfano, Wolfram Barfuss, Carl T. Bergstrom, Miguel A. Centeno, Iain D. Couzin, Jonathan F. Donges, *et al*. "Stewardship of Global Collective Behavior". *Proceedings of the National Academy of Sciences of the United States of America*, vol. 118, no. 27 (2021). DOI: https://doi.org/10.1073/pnas.2025764118.

Brown, Allison J. "'Should I Stay or Should I Leave?': Exploring (Dis) Continued Facebook Use After the Cambridge Analytica Scandal". *Social Media and Society*, vol. 6, no. 1 (2020). DOI: https://doi.org/10.1177/2056305120913884.

Callaway, David and Lothar Determann. "The New US Cloud Act - History, Rules, and Effects". *The Computer and Internet Lawyer*, vol. 35, no. 8 (2018).

Cattani, Mauro, Iberê Luiz Caldas, Silvio Luiz de Souza, and Kelly Cristiane Iarosz. "Deterministic Chaos Theory: Basic Concepts". *Revista Brasileira de Ensino de Fisica*, vol. 39, no. 1 (2017). DOI: https://doi.org/10.1590/1806-9126-RBEF-2016-0185.

cd projekt red. "Cyberpunk 2077 — from the Creators of The Witcher 3: Wild Hunt". (n.d.) [Accessed: October 18, 2018]. https://www.cyberpunk.net/us/en/.

Cerf, Vinton G., and Robert E. Kahn. "A Protocol for Packet Network Intercommunication". *IEEE Transactions on Communications*, vol. 22, no. 5 (1974). DOI: https://doi.org/10.1109/TCOM.1974.1092259.

Corera, Gordon. "Pegasus Scandal: Are We All Becoming Unknowing Spies?". *BBC News*, (2021, July 21). [Accessed: October 17, 2018]. https://www.bbc.com/news/technology-57910355.

Dayma, Boris. "DALLE Mini". *Hugging Face* (n.d.). [Accessed: October 18, 2018]. https://huggingface.co/dalle-mini.

Danaher, John. "The Threat of Algocracy: Reality, Resistance and Accommodation". *Philosophy & Technology*, vol. 29, no. 3 (2016): 245-268. DOI: https://doi.org/10.1007/s13347-015-0211-1.

Díaz, Diego, and Clara Boj. "Artistic practices in the age of the datacene. Data Biography: digital traces to biographically explore personal identity". *Artnodes*, no. 24 (2019): 121-133. DOI: https://doi.org/10.7238/a.v0i24.3293.

Elayan, Haya, Moayad Aloqaily, and Mohsen Guizani. "Internet of Behavior (IoB) and Explainable AI Systems for Influencing IoT Behavior". *EEE Network* (2021, September). DOI: https://doi.org/10.1109/MNET.009.2100500.

European Commision. "Data Protection in the EU". *European Commission* (n.d.). [Accessed: October 17, 2018]. https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en.

Flax Community. "Flax-Community/Gpt-2-Spanish". *Hugging Face* (n.d.). [Accessed: October 17, 2018]. https://huggingface.co/flax-community/gpt-2-spanish.

Frazer, James George. *The Golden Bough: A Study of Magic and Religion*. London: Palgrave Macmillan, 1922. https://www.gutenberg.org/ebooks/3623. DOI: https://doi.org/10.1007/978-1-349-00400-3.

Gatti, Mathias, and Tamara Mathov. "¿Pueden Escribir Poesía Las Computadoras?". *CCE Buenos Aires* (n.d.). [Accessed: October 17, 2018]. https://www.cceba.org.ar/medialab/pueden-escribir-poesia-las-computadoras.

Glauner, Patrick, Petko Valtchev, and Radu State. "Impact of Biases in Big Data". In: *Google Terms of Service – Privacy & Terms – Google* (n.d.). [Accessed: October 17, 2018]. https://policies.google.com/terms?hl=en#toc-permission.

Hern, Alex. "Elon Musk Says AI Could Lead to Third World War". *The Guardian* (2017, 4 September). https://www.theguardian.com/technology/2017/sep/04/elon-musk-ai-third-world-war-vladimir-putin.

Holst, Arne. "Amount of Information Globally 2010-2024". *Statista* (2021).

Jay, Florence, Jean-Pierre Renou, Olivier Voinnet, and Lionel Navarro. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks Jun-Yan". *Proceedings of the IEEE International Conference on Computer Vision* (2017).

Jishi, Radi A. "A Brief Review of Statistical Mechanics". In: *Feynman Diagram Techniques in Condensed Matter Physics*. Cambridge: Cambridge University Press, 2013. DOI: https://doi.org/10.1017/cbo9781139177771.006.

Kitchin, Rob. "Big Data, New Epistemologies and Paradigm Shifts". *Big Data & Society*, vol. 1, no. 1 (2014). DOI: https://doi.org/10.1177/2053951714528481.

Laplace, Pierre-Simon. "A Philosophical Essay on Probabilities". *Journal of the Franklin Institute*, vol. 154, no. 4 (1902). DOI: https://doi.org/10.1016/s0016-0032(02)90322-4.

Lewis, John M. "Roots of Ensemble Forecasting". *Monthly Weather Review*, vol. 133, no. 7 (2005): 1865-1885. DOI: https://doi.org/10.1175/MWR2949.1.

Mildenberger, Carl D. "EVE Online". In: *Economics and Social Conflict*. London: Palgrave Macmillan, 2013. DOI: https://doi.org/10.1057/9781137281890_5.

Mohajerin, Nima, and Steven L. Waslander. "Multistep Prediction of Dynamic Systems with Recurrent Neural Networks". *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11 (2019): 3370-3383. DOI: https://doi.org/10.1109/TNNLS.2019.2891257.

mSpy. "mSpyTM Cell Phone Tracker: Your #1 Monitoring Tool". *Mspy.com* (n.d.). [Accessed: November 22, 2022]. https://www.mspy.com/.

NLTK. "Natural Language Toolkit". *Nltk.org* (n.d.). [Accessed: November 14, 2022]. https://www.nltk.org/.

OECD. "OECD Environmental Outlook to 2050: The Consequences of Inaction - Key Facts and Figures". *OECD* (n.d.). [Accessed: October 17, 2018]. https://www.oecd.org/env/indicators-modelling-outlooks/oecdenvironmentaloutlookto2050theconsequencesofinaction-keyfactsandfigures.htm.

Raza, Hasnain. "Fast-SRGAN: A Fast Deep Learning Model to Upsample Low Resolution Videos to High Resolution at 30fps". *Github* (n.d.). [Accessed: October 17, 2018]. https://github.com/HasnainRaz/Fast-SRGAN.

Reigeluth, Tyler Butler. "Why Data Is Not Enough: Digital Traces as Control of Self and Self-Control". *Surveillance & Society*, vol. 12, no. 2 (2014): 243-254. DOI: https://doi.org/10.24908/ss.v12i2.4741.

Schumm, Walter R. "Confirmation Bias and Methodology in Social Science: An Editorial". *Marriage and Family Review*, vol. 57, no. 4 (2021): 285-293. DOI: https://doi.org/10.1080/01494929.2021.1872859.

Shannon, Claude E. "Communication in the Presence of Noise". *Proceedings of the IRE*, vol. 37, no. 1 (1949): 10-21. DOI: https://doi.org/10.1109/JRPROC.1949.232969.

Skiena, Steven S. *The Data Science Design Manual*. Springer, 2017. DOI: https://doi.org/10.1007/978-3-319-55444-0.

Statista. "Cloud infrastructure services vendor market share worldwide from 4th quarter 2017 to 1st quarter 2022". *Statista* (n.d.). [Accessed: October 17, 2018]. https://www.statista.com/statistics/967365/worldwide-cloud-infrastructure-services-market-share-vendor/.

Synced. "LeCun vs Rahimi: Has Machine Learning Become Alchemy?". *Medium* (December 13, 2017). [Accessed: October 27, 2018]. https://synced.medium.com/lecun-vs-rahimi-has-machine-learning-become-alchemy-21cb1557920d.

Twint. "TWINT - Twitter Intelligence Tool". *PyPI* (n.d.) [Accessed: October 17, 2018]. https://pypi.org/project/twint/.

Wan, Renzhuo, Shuping Mei, Jun Wang, Min Liu, and Fan Yang. "Multivariate Temporal Convolutional Network: A Deep Neural Networks Approach for Multivariate Time Series Forecasting". *Electronics*, vol. 8, no. 8 (2019): 876. DOI: https://doi.org/10.3390/electronics8080876.

CV

Diego Díaz
Universitat Jaume I (UJI)
daz@uji.es

He teaches classes at the Jaume I University, Spain, for the Video Game Design and Development degree. From 2004 to 2006, he was a Research Associate at the Entertainment and Interaction Research Center at Nanyang Technological University, Singapore. Along with Clara Boj, his research has been presented internationally in prestigious renowned centres and indexed journals. Together, they have enjoyed creation and research residencies at institutions such as Hangar in Barcelona, Interface Culture Lab in Linz (Austria), Mixed Reality Lab (Singapore) and Symbiotic System Lab in Kyoto (Japan). Among other distinctions, they were awarded the Alfons Roig Research Scholarship from the Valencia Provincial Council in 2006 and the Production Incentive in the International Vida 13.2 Art and Artificial Life contest. They developed the research project entitled *Reset Mar Menor: Laboratory of imaginaries for a landscape in crisis*, financed by the Carasso Foundation within its Call for Citizen Art for Scholarships. More recently, Diaz received the 2019 Leonardo Scholarship for Cultural Researchers and Creators, BBVA Foundation. More information at www.lalalab.org.

Clara Boj
Universitat Politècnica de València (UPV)
claboto@esc.upv.es

Professor at the Polytechnic University of Valencia, Spain. From 2004 to 2006, she was a Research Fellow at the Interaction and Entertainment Research Center of Nanyang Technological University, Singapore. Together with Diego Díaz, her research has been presented internationally in prestigious renowned centres and indexed journals. Together, they have enjoyed creation and research residencies at institutions such as Hangar in Barcelona, Interface Culture Lab in Linz (Austria), Mixed Reality Lab (Singapore) and Symbiotic System Lab in Kyoto (Japan). Among other distinctions, they were awarded the Alfons Roig Research Scholarship from the Valencia Provincial Council in 2006 and the Production Incentive in the International Vida 13.2 Art and Artificial Life contest. With other researchers, she developed the research project entitled *Reset Mar Menor: Laboratory of imaginaries for a landscape in crisis*, financed by the Carasso Foundation within its Call for Citizen Art for Subsidies. In the field of artistic mediation, in which Clara Boj received her doctorate in 2003, she is currently co-director of the Permea Master's Degree, organized by the Consortium of Museums of the Generalitat Valenciana and the University of Valencia, coordinator of the Valencian node of the Red Planea, and director of the *transversalia.net* project organized by the Generalitat Valenciana and the Teacher Training Service (CEFIRE) of the Generalitat Valenciana. More information at www.lalalab.org.