

Mass personalization: Predictive marketing algorithms and the reshaping of consumer knowledge

Big Data & Society
July–December: 1–14
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2053951720951581
journals.sagepub.com/home/bds

Baptiste Kotras

Abstract

This paper focuses on the conception and use of machine-learning algorithms for marketing. In the last years, specialized service providers as well as in-house data scientists have been increasingly using machine learning to predict consumer behavior for large companies. Predictive marketing thus revives the old dream of one-to-one, perfectly adjusted selling techniques, now at an unprecedented scale. How do predictive marketing devices change the way corporations know and model their customers? Drawing from STS and the sociology of quantification, I propose to study the original ambivalence that characterizes the promise of a *mass personalization*, i.e. algorithmic processes in which the precise adjustment of prediction to unique individuals involves the computation of massive datasets. By studying algorithms in practice, I show how the active embedding of local preexisting consumer knowledge and punctual de-personalization mechanisms are keys to the epistemic and organizational success of predictive marketing. This paper argues for the study of algorithms in their contexts and suggests new perspectives on algorithmic objectivity.

Keywords

Marketing, machine learning, Big Data analytics, personalization, algorithm, knowledge

Suddenly, [an internet user] is silently captured in a database and will soon receive information through the mail tailored to specific interests. What was learned cruising the internet has been vacuumed, and converted to a targeted selling proposition.

Donald Libby, “Cruising and Vacuuming the Internet”, *Internet Marketing News*, January 20, 1995.

Cited by Turow (2006), p.74.

As the above quote shows, since the 1990s, marketing professionals have seen the emerging Internet as a means of profoundly renewing their approaches and tools. As datafication makes individual behaviors measurable (Gitelman, 2013; Kitchin, 2014), marketers have been considering digital technologies not only as a new communication channel, but also and more importantly, as a new field of experimentation for calculating consumers and developing new marketing techniques and strategies (Beauvisage and Mellet, 2019). I will focus here on predictive algorithms, whose current momentum, 25 years after the first

email marketing campaigns, perpetuates the digitalization of marketing methods.

Indeed, during the last few years, a new kind of algorithms, often labeled as “artificial intelligence” but more precisely belonging to the field of machine learning, has spread at an unprecedented scale. Their main feature is their ability to make statistical predictions from very large sets of heterogeneous, possibly unstructured data.¹ Born in the 1960s at the margin of the statistics field, these algorithms are now enjoying a wide diffusion thanks to the increase of computing power and to the datafication of society (Jones, 2018; Plasek, 2016). Now designed not only by mathematicians, but also by a new population of “data scientists, programmers and hackers” (Cardon et al., 2018: 210),

Laboratoire Interdisciplinaire Sciences Innovation Société (LISIS), Institut National de Recherche pour l’Agriculture, l’Alimentation et l’Environnement (INRAE), Paris, France

Corresponding author:

Baptiste Kotras, Cité Descartes, 5 Boulevard Descartes, 77454 Marne-la-Vallée Cedex 2, France.
Email: bkotras@gmail.com



this new kind of algorithm has considerably expanded the realm of predictive computing (Mackenzie, 2015).

Within the marketing field in particular, specialized service providers (startups, software vendors and consulting firms), as well as in-house data scientists, are increasingly using machine learning to predict consumer behavior for large companies. A customer's interest in a given product, their risk of churn, or fraud, is predicted through algorithmic models that assign probabilistic scores to individual customers. These scores rely on data describing their past behaviors, whether collected from the web, or already held and stored in the companies' internal databases (Alemany Oliver and Vayre, 2015).

These methods, which I here refer to under the generic term of predictive marketing, are defined by their promise of individualized customer knowledge, much more granular than knowledge produced by traditional market research (questionnaires, focus groups, etc.). They are part of the longstanding project of personalizing market strategies, which traces back to the 1920s (Lauer, 2012), but has been given new prominence since the 1990s (Turow, 2006). Predictive marketing also borrows from well-established customer scoring practices, which originated in the banking sector to rationalize credit (Poon, 2007), and have been used since the 2000s in customer relationship management (Benedetto-Meyer, 2014). By systematically modelling a very large number of variables, these algorithms are said to enable the anticipation of individual behaviors, and thus produce a much better match between goods and services and customers.

How do these predictive algorithms for personalized marketing change the way corporations know and act upon their customers? As "personalization" constitutes the horizon of countless contemporary algorithmic devices (Lury and Day, 2019; Mackenzie, 2018), I study how this general promise unfolds in action, analyzing at a fine-grain level the discussions and material practices of the actors involved in the conception and use of predictive marketing algorithms. Drawing from STS and the sociology of quantification, I consider the epistemic and political consequences of these very practices and assemblages on the production of consumer knowledge (Bowker, 2005; Diaz-Bone and Didier, 2016; Espeland and Stevens, 2008). In particular, this qualitative study contributes to the reflections on the new forms of social ordering performed by Big Data analytics, and how these technologies shape and account for individuality (Bolin and Schwarz, 2015; Couldry et al., 2016). It focuses on the original ambivalence of personalized algorithmic marketing, which draws on both instrumental and humanistic arguments, as it simultaneously aims to optimize market strategies, and to take better account of *persons*, i.e. of each

unique customer, defined by her specificities, needs and life trajectory.

To this end, I rely on 24 semi-structured interviews with data scientists, data engineers, client advisors and marketing analysts, conducted in different settings: the "data labs" and marketing departments of a major telco company, a retailer, a banking group, and a startup.² These interviews lasted between 1 and 2 hours, and were mainly focused on the technical tasks and choices realized by these various professionals, based on accounts of actual cases. I also attended a series of three online training seminars organized by this startup, and accessed a number of professional and commercial documents (whitepapers, presentations, brochures, etc.). This in-depth qualitative material is essential to move beyond hype and commercial statements made by Big Data analytics companies; it allows to understand "algorithms in practice" (Christin, 2017), i.e. the material settings in which they are conceived, experimented and interpreted.

In this paper, I specifically focus on two distinct cases: the data lab of the banking group, which I will call "The Bank", and a predictive marketing startup, which I will call Predicto. While these two organizational contexts differ, data scientists from both the Bank's data lab and Predicto find themselves in a position where they strive to articulate the requests addressed to them, and the pre-existing data infrastructures of their (internal or external) clients: customer databases, purchase records, data management platforms, etc. The cases presented, set in the worlds of banking and insurance, make it possible to observe how predictive marketing takes place in universes that have historically been populated (or even saturated) with all kinds of calculations (Lazarus, 2012; Porter, 1995). Despite this specificity, the results developed in this paper have a more general reach, as the practices studied here are rather similar to what I observed in other domains where predictive marketing was used (namely retail and telecommunications).

I show that there is more to algorithmic predictive marketing than the issues of surveillance and control raised by many critical studies of Big Data. In contrast, this paper aims to understand the paradox posed by contemporary *mass personalization*, i.e. algorithmic processes in which the precise adjustment of prediction to unique individuals involves the computation of massive datasets, compiling the behaviors of very large populations. As noted by Lury and Day, "Personalization is not only personal: it is never about only one person, just me or just you, but always involves generalization" (2019: 2). Here, I show how attention to *persons*, unique individual subjects with specific histories and relations to corporate organizations, is embedded within the material

practices of machine learning-based marketing. I show that the actors constantly organize the porosity between computations and human representations in such a way that algorithms and the social worlds they act upon mutually adjust to each other throughout the prediction process.

In the first section, I discuss earlier research on personalization through Big Data analytics, highlighting the theoretical framework and contributions of this paper. The next three sections describe how predictive algorithms are successively negotiated, tuned and interpreted within collective practices and how it affects the conception of the individual customer. The conclusion discusses the implications of these results and suggests future lines of research.

Personalization as a disputed moral ground

Personalization is the cornerstone of contemporary algorithmic devices. What we buy, the news we read, the music we listen to and so many more components of our everyday lives increasingly depend on algorithmic suggestions, supposedly tailored to fit our personal interests. What Gerlitz and Helmond (2013) call the “like economy” has extended its roots deep in our daily behaviors, and allows algorithms to classify and treat people according to the digital expression of their tastes, thus intertwining profoundly our *liking* (what we like) and *likeness*, or who we are like (Lury and Day, 2019; Seaver, 2012). As noted by MacKenzie (2018), the central issue in recent debates about Big Data, privacy, commercial and political targeting, filter bubbles and so on, is how individual can or should be taken into account in the calculation, and with what consequences on society. Personalization simultaneously represents a promise of emancipation from the broad statistical categories of private and public bureaucracies (Cardon, 2015; Thévenot, 2019), and a potential threat to our privacy and freedom of choice, as it often implies surveillance, targeting and nudging (Lyon, 2014; Yeung, 2017).

Personalization for commercial uses has mostly been criticized in Big Data scholarship. Many authors see it as a mere cynical communication artifice, a “fairytale vision” used by marketers to dissimulate the “algorithmic manipulation of consumers” (Darmody and Zwick, 2020: 2). In this perspective, personalization would essentially mean a strengthened grip on individual behavior and the optimization of marketing strategies (Beckett, 2012). Thanks to the increasing computability of a datafied world, marketing would now consist of a constant surveillance of individuals’ actions, colonizing everyday more aspects of social life

(Pridmore and Zwick, 2011). This “new marketing paradigm” (Arvidsson, 2002) would thus allow unprecedented control of companies over their consumers, through sophisticated algorithms and ever-increasing data volumes. This configuration is sometimes even described as an emergent fully-fledged modality of capitalism (Srnicek, 2017; Zuboff, 2018).

Importantly, this line of research considers predictive algorithms as the advent of a new mode of government that would definitively dissolve the very reference to the central figure of our liberal democracies and economies: the individual subject. In a Deleuzian perspective, Rouvroy and Berns state that “the [*algorithmic*] measure of all things is ‘dividual’, both infra- and supra-personal” (Rouvroy and Berns, 2010: 94). Predictive marketing technologies would thus produce “self-grounded” representations of consumers, “out of clustered data points of disembodied interests, behaviors, opinions and demographics” (Cluley and Brown, 2015: 108). Algorithms would know only singular data points, abstracted from their social context of meaning, and in particular from the individual subjects who produced them in the first place.

These critical approaches raise significant questions, drawing attention to the dynamics of power and value extraction at stake in the algorithmic shaping of the social world. Nevertheless, they present two important limitations that this paper aims to address. First, many contributions on this subject mainly rely on various publicly available material, such as conferences, news stories and public communication of Big Data analytics companies. They lack first-hand empirical accounts of how algorithms are designed for consumer measurement. This leads to a sometimes general and relatively mundane denunciation of surveillance (Castagnino, 2018), that blends very different types of actors (from the overly famous GAFAM to data brokers, historical software vendors, startups, etc.), and in which the power of algorithms is often taken for granted, rather than empirically described (Beer, 2016).

Second, and somehow consequently, these contributions tell us little about what it actually means to build and use algorithms for personalized consumer knowledge, and what it does to the shaping of the consumer. In particular, they largely overlook the practices and motivations of data scientists and other professionals involved in these activities that cannot be reduced to a mere instrumentalism. As shown by Mackenzie (2018) and Thévenot (2019), personalization has become an essential moral underpinning of Big Data quantification devices, justifying practices, on the users’ side (Siles et al., 2020), but also on the part of developers and practitioners. Putting aside “common-sense distinctions [...] between ethical critics and unethical practitioners, positivist programmers and interpretive

ethnographers, and so on” (Moats and Seaver, 2019: 3), this investigation aims to account for the varied ways in which predictive marketing algorithms are collectively adjusted to fit with local forms of customer knowledge (Kiviat, 2019). It describes how data scientists and their interlocutors constantly strive to produce rich, detailed accounts of their clients’ identities, by accumulating thick data about their biographies, preferences and the events in their lives. As suggested by Matzner (2019), the algorithmic sorting of people does not necessarily erase the figure of liberal individuality, but rather reshapes it, in a way that we need to analyze in detail.

This paper thus studies how actors resolve in practice the tensions of algorithmic mass personalization, between the massiveness of calculation and the promise of personalization, and what it does to the way clients are represented and acted upon by corporations. To this end, in the continuity of an emerging research stream, I follow as closely as possible the actors involved in the design and circulation of algorithmic models. I thus study predictive marketing as an activity, inscribed in an ecology of practices, rooted within organized work collectives (Christin, 2017; Jaton, 2017). Such an approach goes beyond a strictly semiotic analysis of algorithms and their general spirit, or “fetichization of code” (Chun and Kyong, 2008). It allows to account for the numerous negotiations and frictions that occur in the design and circulation of predictive marketing devices (Bechmann and Bowker, 2019; Seaver, 2017).

Summoning the world and embedding it into algorithms

I will first show that predictive algorithms do not act *in abstracto*, from data alone. On the contrary, during my investigation, I was able to observe how the Bank’s data scientists, in their ordinary practice, are concerned with representing and integrating, through dedicated procedures, the different ways in which the customer exists within the organization. By so doing, preexisting customer knowledge is incorporated into the algorithmic models.

Data marketing as a science of “life” and a relational crutch

This attention to people stems from what might be called the humanistic ambition of personalized marketing, shared by the employees of the data lab, a structure created in 2015 within the Bank’s marketing department.³ Its Scientific Director, 42 years old, worked in several large industrial companies before joining the Bank in 2015. He particularly values the variety of

problems facing his profession, all of which, in his opinion, have a common denominator: “Data science really deals with every aspect of our lives. And for me, what really interested me was the human being, who is at the center of it all”.⁴ Among the many solicitations that receives a man of his experience in the young and trendy field of data science (Brandt, 2016), the banking sector was not the one that attracted him most, at first. He was nonetheless convinced by the variety of problems related to the Bank’s relationship with its clients. “Ok, [working for the Bank] is about banking, insurance, finance, etc. But there is also security, publishing, real estate. . . In short, everything that affects a person’s life”.⁵ The epistemic plasticity of data science and its ability to pervade a wide variety of social worlds (Dagiral and Parasie, 2017) underpin the trajectory of this interviewee, who sees his work as solving the many different problems of people’s lives.

As a result, many of the activities and projects deployed by the data lab are described as a way to improve the quality of commercial relationships, both for customers and customer advisers. Indeed, the Bank’s client advisers have to manage portfolios of several hundred clients each, which makes individual follow-up difficult. Data science is therefore considered here as the means to better achieve the objective of a quality service relationship, supposedly based on the knowledge of the client’s life trajectories and projects (all the more so in the banking sector, a world of long-lasting commercial relations). The scores thus constitute a form of crutch, or support, for interpersonal relation. As infrastructural devices, they capitalize on the clues needed to organize and prioritize the relational and commercial work of advisers. As a result, data lab employees see their activity as quite distinct from the “*very tactical*” scoring that have been produced for years by the “*IT guys*” of the regional branches to target commercial campaigns.

The [*traditional*] targeting models are very, very short term, it’s like: I want to do a campaign; these are the people [*to contact*]. At the data lab, we’ve pushed models that are more about long-term support, we’re predicting one-year trajectories, instead. To be able to say: for this client, here is the trajectory that we estimate for the next year. And behind that, here is what needs to be implemented to support him. For example, today he has no specific need, but in three months maybe he will need to buy a car, in six months, maybe he was a student, but he will be starting to work, and so forth. In a year’s time, perhaps, a real estate project. That’s what I’m talking about. It’s smarter, and it allows the adviser to really have. . . to build the trajectory with the client. Not to be just in a reactive approach, actually. The idea is really. . . It’s the

very meaning of the bank's relational approach, to anticipate more, [...] because we think that's what will increase client satisfaction, that's what will keep the client healthy, and that's what will strengthen our relationship with him. Scientific Director of the data lab (interview, July 2018)

Rather than maximizing sales for a given product, based on punctual business objectives, employees of the data lab aim to understand their customers' needs in the longer term, in order to support a so-called "relational" approach. In addition to better attending the customer, the data lab's activity symmetrically aims to take good care of the morale of the client advisers, which is affected when they are constrained by strictly commercial objectives and have to call imperfectly identified, reluctant customers. "An adviser who calls a hundred customers, and finally sees that it was useless, well, he'll lose motivation".⁶ Building long-term scores therefore also means taking care of colleagues in the sales front line, of their workload, and of how they can find satisfaction in a job well done. Predicting biographical ruptures (the "life moments") of clients thus helps to reduce the frustration of advisers faced with rejection or disinterest.

The client and her spokespeople

Taking people into account as much as possible then implies a particular organization which consists, as for Pasteur in the canonical article by Bruno Latour (1983), in transporting the world into the closed and controlled enclosure of the (data) laboratory. During what Slota et al. (2020) describe as the "prospection" phase, data scientists audit existing data infrastructures to gather and organize usable data for their projects. To this end, they organize meetings where multiple authorized spokespersons from the world of the customers are invited, in order to include their points of view in the design of measurement projects.

The idea is not to start from scratch; it is to contribute something else, something complementary, to enrich current knowledge. So we ask them: for this output, this variable that we want to predict, in your opinion, what could be the relevant variables? What are the indicators that you are used to build? Even beyond existing data, we often ask the question of the interest for a product, for example, why a customer might be interested? So we make lists. And then we're going to prospect: does this type of information exist in internal databases? Can we look for it in external databases? Scientific Director of the Data Lab (interview, July 2018)

This was the case in a project designed to identify at an early stage, among the Bank's "professional clients",⁷ those who are going to experience financial difficulties in the coming months, to contact them in advance and possibly take facilitating measures (debt spreading, etc.). Each "professional client" had therefore to be given an individual score reflecting the probability for him to experience, in the short or medium term, a significant period of trouble. The data lab then organized workshops over several weeks, to which different departments of the Bank were invited, in order to gather their expertise on the modelling project. Due to their heterogeneous backgrounds, these interlocutors can be analyzed as Latourian spokespersons of the various facets of the "professional client" (Latour, 1989):

- (i) What I call the *day-to-day client*, presented by representatives of the eight regional branches involved in the project, in particular client advisers, custodians of a form of client knowledge rooted in a proximity relationship;
- (ii) The *macroscopic client*, embodied by representatives of the Bank's headquarters, in charge of the strategy towards professionals, for their vision of what the professional client is (and should be) at an aggregated level;
- (iii) The *computerized client*, represented by regional IT specialists in charge of storing and maintaining client data, in order to identify a possible "data perimeter" for the project;
- (iv) The *legal subject client*, finally, represented by members of the Legal Affairs Department, who ensured that the data processed and the calculations applied respected the rights granted to the client by the Bank in its Commitment Charter, as well as by legislation, notably on commercial discrimination and privacy.

This polyphonic arrangement guaranteed the representation of different modes of relationships between customers and the Bank, and brought out the ordinary forms of client knowledge, by asking participants questions such as: which product might be of interest to which category of customer? What indicators already exist? etc. It also made it possible to gather the opinion of professionals on the content and properties of future algorithmic models, the variables deemed relevant, or the data they would like to use – and of course, the legal and technical constraints regarding the use of said data. We can thus observe the relative epistemic modesty of data scientists, who actively solicit the expertise of the usual custodians of the problem they seek to address through computation.

What concrete variables should then be taken into account to define the “financial difficulties” whose occurrence they were trying to predict? For data scientists, the obvious thing to do was to define these difficulties based on the variables contained in the litigation databases: late payment, overdraft, etc. According to them, such events, faithfully recorded in a dedicated database, were a rather plausible indicator of a customer’s difficulty. However, this first version of the predictive model, which was submitted to the assessment of their colleagues, was widely criticized by customer advisor representatives.

We started with that, because it’s something that is traced in our information systems. But what we want to do is to avoid getting to that point, so we’re discussing with the advisers, because this was the first deliverable they received after the various workshops. And now we think we’re getting in rather late. [...] What the advisers told us is: we want to avoid being in front of a client who is already in default, etc. This client, I can’t help anymore: it becomes a litigation, a debt recovery. Scientific Director of the data lab (interview, July 2018)

According to the customer advisers, the initial choice of the data scientists led the algorithm to detect only situations that were already too serious to remedy, quickly leading to bankruptcies and recoveries. The algorithm therefore needed to be trained to predict difficulties way upstream, so as to maintain the possibility of corrective actions. After further consultations, the calculations were then modified. The “financial difficulties” learned by the machine learning algorithm finally became a set of relatively classic variables, including each client’s turnover (and its possible decline), as well as the activity of its economic sector and geographical area, calculated by integrating the zip code of each company with economic open data published by INSEE.⁸

The way in which the early stages of designing algorithmic models were organized shows that an essential condition for the epistemic as well as organizational success of predictive marketing devices lies in the way data scientists make themselves permeable to the varied knowledge locally held by their colleagues. In this case, data scientists and client advisers negotiated the person that the algorithm had to model and predict. At first, the model predicted the wrong person, a client already in too much distress to be rescued. Training data had to be adjusted and redefined in order to predict a still-rescuable client. The action of algorithms on the world of people is therefore not only an effect of the scores and rankings they produce: it intervenes early on, by anticipation, in their very design.

When blindness is robustness

We will now study the precise moment when personalization becomes massive, i.e. when data centralization and unification make it possible to calculate “likenesses” between large populations of individuals (Seaver, 2012). Until now, making “good” predictions meant paying close attention to the many facets of the client. Once the predictions are imported into the data lab, however, building effective predictions require data scientists to deliberately “blind” themselves, through dedicated procedures, to specific traits of people described in the selected data. Only then is it possible to perform calculations whose heuristic value can be transposed outside the data lab. We will thus see that the modeling of clients does not mechanically result from “feeding” or “throwing” data to the algorithm (Amoore and Piotukh, 2015), as a recurring food metaphor would suggest, but from the work of enrichment and selection of variables deployed by data scientists.

The first challenge for them is to build the training environment in which machine learning algorithms will explore the most significant correlations, between a set of descriptive variables of customer behavior, and one (or several) “target variables” (for example: the customer’s future financial difficulties). To this end, they need to unify multiple and heterogeneous data sources, often stored in variously accessible databases: transaction records, purchase histories, calls to technical support, litigation, current contracts, etc. In this way, data scientists seek to constitute what they call a “master file”, or “ground truth” (Jaton, 2017), a table in which each customer is described by one single line, with its multiple descriptive variables in columns. This table materializes unique, calculable and therefore predictable individual behaviors. The challenge here is to transform the fragmented and scattered presence of clients into individualized profiles, which is similar – in its spirit, if not in its technical means – to the work of rearranging individual files described by Denis (2018).⁹

Once the master file has been compiled, it is then paradoxically important for data scientists to “blind” the algorithms to the persons represented in it, in several ways. First, two voluntary blinding measures consist of anonymizing the lines representing clients, so that data workers cannot access nominative personal information. The other one is to make the gender variable disappear: while machine learning algorithms do aim to *discriminate* customers, in the mathematical sense, they are constrained by the legal distinction between authorized discrimination (based on income, for example) and unauthorized discrimination (based on gender or race). Even though names and gender identities could provide useful predictive variables,

data scientists are therefore formally forced to remove them.¹⁰ It is only at the end of the process that the regional branches are allowed to remove the encryption applied to individual identities, to reestablish an interpersonal relationship – now algorithmically equipped.

The learning algorithms can then come into play. In what is called *supervised learning*, as for example in the case of the Bank's professional clients, the aim is to predict a target variable (or a group of target variables), based on records of thousands of individual cases and their past behaviors, each being marked with the "real result" (i.e. having or having not experienced financial difficulties). The algorithm then has to build a mathematical function which, by associating and weighing a number of variables, allows to predict *in as many cases as possible* the target variable, i.e. the occurrence of financial difficulties. After iteratively testing a very large number of correlations, the algorithm thus produces a mathematical model with a prediction rate, which has to be maximized: for example, a rate of 80% thus means that in 80% of individual "learned" cases, the mathematical function built by the algorithm produces the real result – which is considered to be a good rate.

In the controlled environment of the data lab, everything goes quite smoothly: provided with a large amount of data and variables, the machine learning algorithm almost always manages to produce models with a good prediction rate. But how can data scientists be sure that this mathematical rule will retain its predictive power when applied to new, real-world cases, to future customers? This raises the question of the robustness of the model, i.e. its ability to be transposed and used outside the data laboratory.

This is where data scientists deliberately organize the blindness of their algorithms. First, it means hide part of the data available for learning from the algorithm, in order to be able to "regularize" the predictions *a posteriori*. The master file is thus divided into two parts: the first one – the most important – is the training database for the algorithm; the second part will then be used to test the predictive potential of the model on new "unknown" data, which are then treated as if they were really new and unknown cases. The aim is then to obtain a comparable prediction rate for the training set and the regularization set. If the gap is too large, then a new work of selection and construction of new variables begins. Called feature engineering (by combining variables already present, or by enriching the data, for example), it is a crucial (and understudied) part of the know-how of data scientists (Domingos, 2012; Mackenzie, 2015).

But the essential blinding operation here is linked to the risk of *overfitting*, which weighs on any machine learning approach. It is the tendency for the algorithm

to develop sophisticated mathematical models, extremely efficient for describing training data – for which they obtain very high prediction rates – but little transposable to new cases, because they are too mathematically dependent on the individual cases used during training. Taking a fictitious example – predicting what a group of people order in a fast food restaurant – a data scientist from Predicto explains:

An algorithm is always going to be able to say: if he had a red hat and sneakers, he'd take this; if he had this, this and that, he'd buy that. It is able to learn everything by heart. That's overfitting. That's when it learns everything by heart. If there are too many specificities, he'll learn that John Doe buys the Big Mac, and the problem is that if you know that John Doe buys the Big Mac, but one day you're introduced to someone who's not John Doe, then you won't be able to generalize. In fact, you have to find all the common points between different profiles to say: when it's a similar profile, I know what he's buying. If you know all the specificities of the person, in fact, you're not learning a typical profile: you're learning John Doe. [...] You don't need to know what *he* is eating, you need to know what profile he looks like. Data scientist, Predicto (interview, May 2018)

As a consequence, preparing the statistical model to get out of the data lab, and ensuring its grasp on the real world may involve what data scientists call "dragging down"¹¹ the algorithm, by imposing random errors that prevent it from "learning by heart" the data. It thus enables it to build predictive models that can be transposed to future populations, not only valid for the training dataset.

There are mathematical techniques that allow data to be randomly and artificially dragged down. So the algorithm doesn't learn well. So that it doesn't give too much weight to certain values. So that he doesn't learn everything by heart. Randomly, I'm going to introduce some error. When it tries to minimize the error, it basically tries to minimize the weights so that the error is as small as possible. So it adjusts the weights less finely. It does something coarser, less efficient, but more robust. You take off its glasses, so its vision becomes blurry: it can't recognize John Doe any more. Data scientist, Predicto (interview, May 2018)

This "dragging down" may include deliberately eliminating variables (or groups of variables) that are most closely correlated with the variable to be predicted, to prevent the algorithm from learning strong but punctual correlations, unlikely to occur in the real world with future data. As illustrated schematically in

Figure 1, from the data points learned by the algorithm (in red), this trick prevents the algorithm from inferring a complex function linking all the points (the blue curve) but too specific to detect the general movement of the observed behaviors (the green one).

The construction of predictions, far from relying on the mechanical “ingestion” of large datasets, involves complex work in which the data are centralized, aggregated and tested by data scientists. Paradoxically, the promise of personalization of data marketing implies during this phase to make the algorithms blind to some specificities of the customers, and sometimes to the most significant data points of their behaviors. The algorithm must not know too precisely the unique individuals included in its training environment, otherwise an overly personalized calculation will subsequently prevent it from effectively predicting the behaviors of new individuals. Algorithmic learning thus puts in tension the attention to the people, which is its horizon, and the need for generalization, of making predictions that will apply to future data, and not only to the people included in the training data.

Interpreting scores, reconstructing explainable consumer figures

Finally, once the algorithmic model is out of the data lab, an important task for data scientists is to interpret its results, which in themselves are not self-evident. They carry out a specific articulation work (Strauss, 1988), seeking to reduce the frictions between the world predicted by the algorithm, and the world of

real customers. As in the case of digital advertising (Bolin and Schwarz, 2015), their objective is then to make predictive marketing intelligible and to prove its specific value, which implies linking its results to knowable figures of the consumer and interpretable macroscopic behaviors. This work is all the more important as the predictive model generates apparently incoherent results, as in the two examples analyzed here.

More targeting, less conversion? The paradoxes of “ultra-personalization”

I focus on the case of the startup company Predicto, created in 2015 by a computer scientist and a sales engineer.¹² One of its clients, a mutual insurance company, was seeking to optimize its commercial efforts to sell funeral insurance policy to their clients that did not have one already. The ambition was to better target potentially interested clients, to reduce the costs of the thousands of letters sent, but also not to “miss” potentially interested customers. This type of optimization, one of Predicto’s standard services, relied on what the startup calls its “ultra-personalization” algorithms, as shown below. It is based on a positivist epistemology in which interested customers preexist the commercial strategies deployed to “reach” them. This allows to consider direct marketing not as a technique to construct or trigger the client’s interest, but as an optimization problem.

Figure 2 illustrates the difference between contacting the same number of clients (4% of the whole population), based on traditional segmentation methods (first graph), and based on machine-learning “ultra-personalized” segmentation that takes into account both “strong” and “weak signals” (second graph). Both are compared to randomly contacting clients (the black line).

French titles respectively say:

*Model based on the main strong signal” and
“Combination of models based on a set of strong and weak signals*

While the insurance provider used to target its clients based on their age (the “strong signal”), Predicto built a scoring model based on eight sources of data, including customer record, history of reimbursements, of calls to customer service, products held, past marketing actions and sociodemographic open data. The algorithm then ranked all of the clients that did not have funeral insurance policy, by statistical similarity to the clients that did have one, in regard of these variables. The company then sent direct mail to the top-40,000 clients during the (very funereal) month of November,

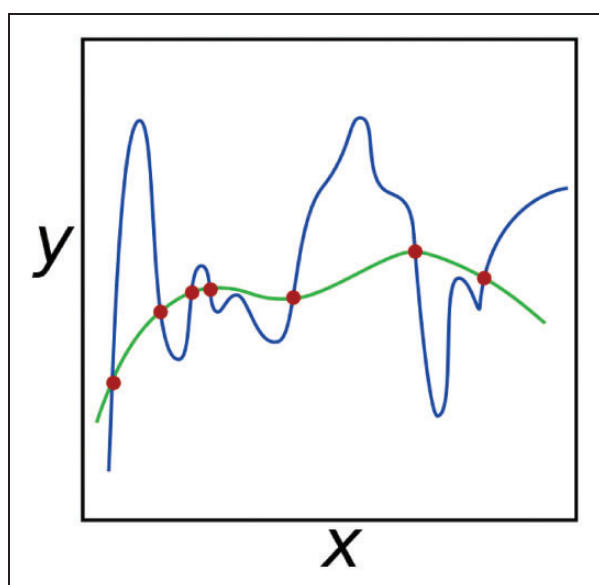


Figure 1. Graphical representation of the regularization of overfitting.¹³

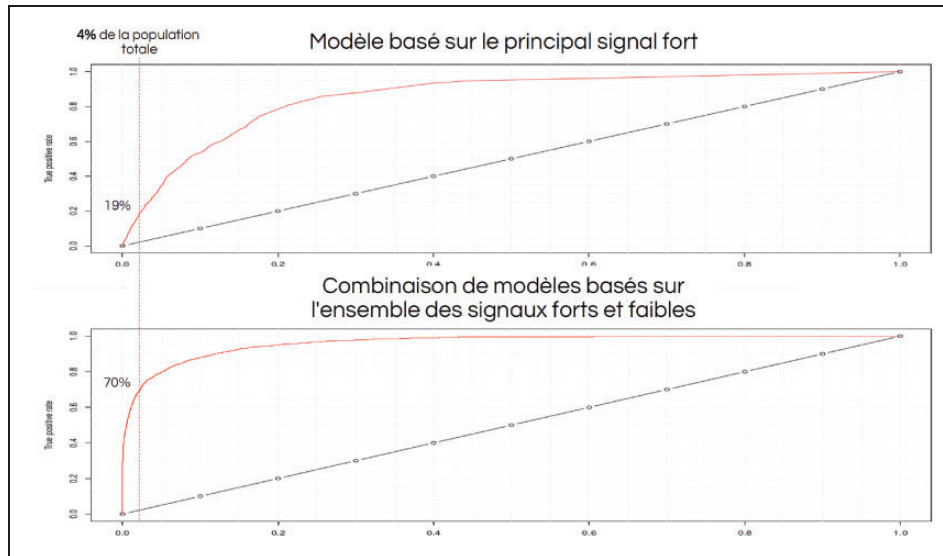


Figure 2. Training seminar “From Segmentation to Ultra-Personalization”, Oct-18.

inviting them to subscribe to a policy. Afterwards, a classical protocol of A/B testing was set to compare the results of the new machine learning-based method, and the old one.

This comparison generated some frictions. The first surprise was that, although the scoring carried out by Predicto did indeed allow selling more policy contracts in total, it was because it produced more requests for quotes (+30%) from the customers contacted. However, it produced a lower conversion rate than traditional methods; in other words, a smaller proportion of these requests for information resulted in an effective subscription.

Let’s say that with the traditional method, they make 50 [requests for quotes] and we make 80; out of the 80, we convert half of them, i.e. 40 contracts. And they convert 60%, so they’re going to get 30. They convert more, but in the end we sell more. Since we target better, we target people who may already have the project of buying what the client offers. Therefore, potentially, they are also considering competitors. Whereas there was less competition in their case because they were targeting less. Because when you have less competition, it’s easier to sell. Co-founder, Predicto (interview, April 2018)

According to the co-founder of the start-up, the traditional targeting methods converted more easily, because they reached a more random audience. The clients who asked for a quote thus tended to stay captive of the initial offer they received, and were therefore more inclined to sign directly with the insurance provider. On the contrary, more targeted methods reached

more interested people, which may be more likely to ask for a quote, but also to be already in a comparison process with competitors. The transformation rate was therefore lower when dealing with a public that was less influenced by the insurance provider’s commercial offer. At the end of the day, what improved sales was the targeting of already interested clients, which generated more requests for quotes. The interpretative work deployed by Predicto’s cofounder reflects a paradoxical justification of personalization: in this perspective, far from a method that would tend towards 100% efficiency (one mail, one sale), the economic utility of predictive algorithms ultimately derives from maximizing the affected audience.

The perfect client: When the algorithm predicts too well

The post-campaign evaluation generated another inconsistency, when Predicto realized that the customers who actually purchased a policy were not the top-ranked ones: they did have a good ranking, but the top-scored customers proved to be surprisingly unresponsive to the company’s solicitations. This observation called for another bit of interpretative work.

You can have a funeral policy with Roc Eclerc [*a funeral home*], you can have a funeral policy with a lot of service providers. Simply being a client [*of the health insurance company*] won’t make you take *their* policy. Our criteria was to address all those who did not have a funeral convention with our client. So, in particular, those who potentially had one with another provider. And that’s where we should have been more vigilant, it

was a youthful error, somehow. We targeted the perfect profile of the client who, indeed, already had a policy. Co-founder, Predicto (interview, July 2018)

The top-ranked clients were the ones without the funeral policy, that were most similar to those who already had it, for example the oldest clients, with high health expenditures. According to him, they were actually very likely to have one already with a competitor, for example a specialized provider or a funeral home. He then told me: what we should have done is make the algorithm learn the characteristics of the clients *at the time* they purchased the policy, not their present characteristics. Because, he said, “when you’re 80, you do have the policy, but you bought it ten years ago!”

Here is another tension regarding the person produced by algorithmic prediction, and the necessity for the actors to interpret and contextualize the results of algorithms, to establish their value. It is linked to a structural limitation of predictive scoring, which necessarily depends on the data available to train the algorithm and the choices made by data scientists in defining its desired outcomes (Grosman and Reigeluth, 2019). By measuring the likelihood to subscribe by comparison with actual policyholders, the algorithm may have assigned the highest scores to people who resembled them too strongly for them not to have a policy elsewhere.

The cases studied here allow us to depart from the ordinary representation of algorithmic targeting that would produce conversion rates tending towards 100% (Beer 2017; Cardon, 2015). In reality, quite often, algorithms do not do exactly what they are supposed to. Moreover, they call for human interpretation work, not only because machine learning produces black boxes, but to establish the utility and value of the calculation. What the algorithm *does*, the type of prediction it produces, constitute new grounds for what Kiviat (2019) calls processes of “causal theorizing” and “de-commensuration”, in which the actors argue and negotiate the logic and fairness of the algorithm. In our case, the algorithm did not predict the exact expected customer: it did not unearth latent individual preferences, but allowed to better identify customers who are already interested, active and informed about the state of the market.

Learning algorithms thus displace uncertainty in customer knowledge. In accordance with the historical ambition of market research (Berghoff et al., 2012; Cochoy, 1998), they quite considerably reduce the uncertainty surrounding the customer’s tastes and behavior. They do, however, induce a new type of uncertainty, linked to the explanation of scores (black box effect), but above all to their interpretation. Contrary to symmetric critics and praises regarding

machine learning algorithms, they do not produce a cold, seamless, and automated process of knowing and governing data points. Until the last moment, humans adjust the perimeter and nature of algorithmic action on people.

Conclusion

As we have seen from the cases presented here, mass personalization involves an iterative, back-and-forth process between the world and the data lab. Contrary to a widespread representation (Darmody and Zwick, 2020; Steiner, 2012; Zuboff, 2018), algorithms do not “manipulate” the social world from the outside, at the end of a linear process of learning, calculation and prediction. The world and the algorithms define each other, through the articulation work carried out not only by the “little hands” of the information society (Bowker and Star, 1999; Dagiral and Peerbaye, 2012), but also by data scientists. They centralize customer knowledge, harmonize the available data, interpret the results and reintegrate them into explanatory schemes specific to the predicted universes. As Grosman and Reigeluth (2019) point out, both learning objectives and training datasets are constructed, negotiated and adjusted by humans, themselves embedded in local knowledge regimes, organizations and cultures. In the case of predictive marketing, even the most abstract mathematical operations (such as the techniques to limit overfitting) are inseparably computational and social.

What happens to *persons* in this process? Far from being dissolved within algorithmic computation (Rouvroy, 2018), their taking into account constitutes both the horizon and the material support of a set of varied practices. The first section showed how organizational settings actively aim to embed the usual categories and perceptions of the consumer in the choice of analyzed data, and in the types of questions for the algorithm to solve. The last section showed how making use of algorithmic predictions demands a specific interpretation work, or “causal theorizing”, that reattaches them to “palatable” figures of the consumer and plausible courses of action (Kiviat, 2019). These moments, before and after the calculation, are key conditions of its possibility, success and performativity upon the social world. In other words, it is because data and algorithms are continuously readjusted to exogenous preexisting forms of knowledge that mass personalization can happen. Here, classical categorizations of the consumer are essential to the organizational as well as to the epistemic success (i.e. producing “useful” predictions) of predictive marketing.

Although Amoore and Piotukh (2015) rightly highlight the centrality of algorithmic calculation for making sense out of Big Data, this does not simply imply “throwing data at the algorithm” (Amoore and Piotukh, 2015: 343). “Ingestion” of data is actually a complex process involving non-mathematical, local epistemologies in a crucial manner. It does not abstracts data from their social context, or at least, not entirely. If anything, the reason why algorithms can predict consumer behavior is because their conception and interpretation are continuously nurtured by exogenous epistemologies of the consumer, produced at the interplay of working collectives, routines and inherited data infrastructures (Bowker and Star, 1999; Christin, 2017). In this respect, Big Data quantification is more of a “remediation”, than a disruption (McGuigan 2019: 3). As described by Bolin and Schwarz (2015) in the case of media planning, this remediation may involve a back-and-forth process between ancient and new modes of representation, between “Gaussian” and “Paretian” statistics (Bolin and Schwarz, 2015: 3). It may be due to “institutional inertia”, but it may also be – as it is here, or as shown in previous work (Kotras, 2015) – a crucial condition for epistemic success. This result argues for a deepened investigation into the hybridizations of the various forms of knowledge involved in the “manufacture of the public” (Bermejo, 2009).

This result also suggests that we need to take seriously the moral claims made by data science, which are keys to its growing pervasiveness. Here, the promise of supporting better market relations, more adjusted to individual authentic aspirations, is part of a longstanding criticism of traditional statistical categories, considered as insufficiently precise to do justice to the specificity of individuals (Boyd and Crawford, 2012; Desrosières, 2011). The description of data science as a science of “life” in general, as posited by the Bank’s data lab scientific director at the beginning of this paper, is a widespread moral horizon that must be studied as such. Of course, this does not mean taking the humanistic discourse of data marketers for granted. This article is rather an invitation to study its concrete consequences in terms of practices, attitudes and expectations. Documenting the way in which algorithms are constantly informed, parameterized, adjusted by exogenous knowledge (Bechmann and Bowker, 2019), also appears as a suitable way to overcome the divide between data scientists and social scientists, in which the latter are ultimately reduced to representing “ethics”, at the end of an opaque process (Moats and Seaver, 2019). As Lee et al. (2019) write, it is then less a matter of denouncing the “biases” of algorithms, or assessing *a posteriori* their propensity for justice or injustice, than of observing the way they “fold” the

world to link heterogeneous objects, knowledge, representations of the world. It is the ambition of this article to have contributed to documenting the construction of such assemblages.

Finally, these results call for further contributions about algorithmic objectivity. Following Angèle Christin (2016), we can test the classic contribution of Lorraine Daston and Peter Galison (2010), who showed how objectivity has historically been defined as the ability to produce knowledge disconnected from the humans who produce it. “To be objective is to aspire to a knowledge that bears no trace of the knower – knowledge unmarked by prejudice or skill, fantasy or judgment, wishing or striving” (Daston and Galison, 2010: 17). Algorithmic modelling seems to be a mechanistic knowledge instrument, in which the automated abstraction of calculation should guarantee the elimination of human preconceptions and biases, and would thus produce more “objective” results than those generated by traditional market research. Nevertheless, this investigation into the practice of data marketing has shown the constant and decisive attention that data scientists pay to local forms of knowledge, and to the people who are the bearers as well as the subjects of this knowledge. Further contributions are needed to document the interweaving of knowledge embedded in humans and machines in the production of algorithmic objectivity.

Acknowledgements

This paper owes much to the precise and rich feedback it received from Geoffrey Bowker, Éric Dagiral, Sylvain Parasie, Ashveen Peerbaye and Jean-Marc Weller. I am also very grateful to the reviewers and editors of *Big Data & Society* for their most helpful comments.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Recommendation automats, e.g., for online cultural products (YouTube, Netflix, Spotify, etc.), are a canonical example of this type of algorithm: by comparing the tastes (i.e. clicks) of one user with those of millions of others, they suggest the products most likely to appeal to her as well (Eriksson et al., 2019; Seaver, 2012).

2. Findings presented in this paper are coherent throughout these various settings, although it mostly relies on two of them. All fieldwork was conducted in France between 2018 and 2019, along my postdoctoral research in sociology on predictive marketing at IFRIS.
3. The Data Lab is part of a top management policy of promoting the use of data science throughout the company, particularly for modelling customer behavior. Counting with approximately 15 employees (data scientists, data engineers, etc.), it can be called upon by group entities, such as its regional branches. It can also take on subjects deemed innovative and important on its own.
4. Interview, Jul-18. All interviews are translated from French by the author.
5. Idem.
6. Idem.
7. At the Bank, “professional clients” are very small businesses, usually artisans.
8. The French National Institute for Statistics and Economic Studies.
9. A crucial issue in this harmonization work is the existence or absence of a unique identification key that can be used to circulate between multiple databases (customer service, litigation, sales, etc.).
10. It does not mean by any way that data scientists cannot, if needed, produce mathematical estimations of gender or race based on stored data. See Sweeney (2002).
11. In French : “plomber”.
12. Predicto was bought after the interviews by a business intelligence software vendor.
13. Source: [https://en.wikipedia.org/wiki/Regularization_\(mathematics\)](https://en.wikipedia.org/wiki/Regularization_(mathematics)) (consulted 03/04/2020).

ORCID iD

Baptiste Kotras  <https://orcid.org/0000-0003-3323-8392>

References

- Aleman Oliver M and Jean-Sébastien Vayre (2015) Big data and the future of knowledge production in marketing research. Ethics, digital traces, and abductive reasoning. *Journal of Marketing Analytics* 3(1): 5–13.
- Amoore L and Piotukh V (2015) Life beyond big data. Governing with little analytics. *Economy and Society* 44(3): 341–366.
- Arvidsson A (2002) On the ‘Pre-History of the panoptic sort’: Mobility in market research. *Surveillance & Society* 1(4): 456–474.
- Beauvisage T and Mellet K (2019) Mobile consumers and the retail industry: The resistible advent of a new marketing scene. *Journal of Cultural Economy* 13(1): 1–17.
- Bechmann A and Bowker G (2019) Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society* 6(1): 1–11.
- Beckett A (2012) Governing the consumer: Technologies of consumption. *Consumption Markets & Culture* 15(1): 1–18.
- Beer D (2016) The data analytics industry and the promises of realtime knowing: Perpetuating and deploying a rationality of speed. *Journal of Cultural Economy* 10(1): 21–33.
- Beer D (2017) The social power of algorithms. *Information Communication & Society* 20(1): 1–13.
- Benedetto-Meyer M (2014) Du Datamining Aux Outils de Gestion. Enjeux et Usages Des ‘Scores’ Dans La Relation Client. In: Alexandre M and Emmanuel K (eds) *La Fabrique de La Vente. Le Travail Commercial Dans Les Télécommunications*. Paris: Presse de l’Ecole des Mines, pp.147–169.
- Berghoff H, S Philip and S Uwe eds. (2012) *The Rise of Marketing and Market Research*. New York, NY: Palgrave Macmillan.
- Bermejo F (2009) Bermejo F (2009) audience manufacture in historical perspective: From broadcasting to Google. . *New Media & Society* 11(1–2): 133–154.
- Bolin G and Schwarz J (2015) Heuristics of the algorithm: Big data, user interpretation and institutional translation. *Big Data & Society* 2(2): 1–12.
- Bowker G (2005) *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.
- Bowker G and Star SL (1999) *Sorting Things out: Classification and Its Consequences. Inside Technology*. Cambridge, Mass: MIT Press.
- Boyd D and Crawford K (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662–679.
- Brandt P (2016) *The Emergence of the Data Science Profession*. Thesis of the Degree of Doctor of Philosophy, Graduate School of Arts and Sciences, Columbia University, New York, NY, USA.
- Cardon D (2015) *À quoi Rêvent Les Algorithmes. Nos Vies à L’heure Des Big Data*. Paris: Seuil.
- Cardon D, Cointet J-P and Mazières A (2018) La revanche des neurones. L’invention des machines inductives et la controverse de l’intelligence artificielle. *Réseaux* 5(211): 173–220.
- Castagnino F (2018) Critique des surveillance studies. Eléments pour une sociologie de La surveillance. *Déviance et Société* 42(1): 9–40.
- Christin A (2016) From daguerreotypes to algorithms. Machines, expertise, and three forms of objectivity. *ACM Sigcas Computers and Society* 46(1): 27–32.
- Christin A (2017) Algorithms in practice. Comparing web journalism and criminal justice. *Big Data & Society* 4(2): 1–14.
- Chun WH and Kyong (2008) On ‘sourcery,’ or code as fetish. *Configurations* 16(3): 299–324.
- Cluley R and Brown SD (2015) The dividualised consumer: Sketching the new mask of the consumer. *Journal of Marketing Management* 31(1–2): 107–122.
- Cochoy F (1998) Another discipline for the market economy: Marketing as a performative knowledge and know-how for capitalism. *The Sociological Review* 46(1): 194–221.
- Couldry N, Fotopoulou A and Dickens L (2016) Real social analytics. A contribution towards a phenomenology of a

- digital world. *The British Journal of Sociology* 67(1): 118–137.
- Dagiral É and Parasie S (2017) La « science Des Données » à La Conquête Des Mondes Sociaux. Ce Que Le « Big Data » Doit Aux Épistémologies Locales. In: Pierre-Michel M and Simon P (eds) *Big Data et Traçabilité Numérique. Les Sciences Sociales Face à La Quantification Massive Des Individus*. Paris: Collège de France, pp.85–104.
- Dagiral É and Peerbaye A (2012) Les mains dans les bases de données : connaître et faire reconnaître le travail invisible. *Revue D'anthropologie Des Connaissances* 6(1): 191–216.
- Darmody A and Zwick D (2020) Manipulate to empower: Hyper-relevance and the contradictions of marketing in the age of surveillance capitalism. *Big Data & Society* 7(1): 1–12.
- Daston L and Galison P (2010) *Objectivity*. Paperback ed. New York, NY: Zone Books.
- Denis J (2018) *Le Travail Invisible Des Données: Éléments Pour Une Sociologie Des Infrastructures Scripturales*. Paris: Presse de l'École des Mines.
- Desrosières A (2011) *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge, MA: Harvard University Press.
- Diaz-Bone R and Didier E (2016) The sociology of quantification. Perspectives on an emerging field in the social sciences. *Historical Social Research* 2(41): 7–26.
- Domingos P (2012) A few useful things to know about machine learning. *Communications of the ACM* 55(10): 78–87.
- Eriksson M, Fleischer R, Johansson A, et al. (2019) *Spotify Teardown. Inside the Black Box of Streaming Music*. Cambridge, MA: MIT Press.
- Espeland WN and Stevens ML (2008) A sociology of quantification. *European Journal of Sociology* 49(3): 401–436.
- Gerlitz C and Helmond A (2013) The like economy: Social buttons and the data-intensive web. *New Media & Society* 15(8): 1348–1365.
- Gitelman L, ed. (2013) *"Raw Data" Is an Oxymoron*. Cambridge, MA: The MIT Press.
- Grosman J and Reigeluth T (2019) Perspectives on algorithmic normativities: Engineers, objects, activities. *Big Data & Society* 6(2): 1–12.
- Jaton F (2017) We get the algorithms of our ground truths. Designing referential databases in digital image processing. *Social Studies of Science* 47(6): 811–840.
- Jones ML (2018) How We became instrumentalists (again): Data positivism since World War II. *Historical Studies in the Natural Sciences* 48(5): 673–684.
- Kitchin R (2014) Big data, new epistemologies and paradigm shifts. *Big Data & Society* 1(1): 1–12.
- Kiviat B (2019) The moral limits of predictive practices: The case of credit-based insurance scores. *American Sociological Review* 84(6): 1134–1158.
- Kotras B (2015) L'opinion autorisée. Requalification communautaire de l'espace social et techniques d'échantillonnage. *Revue d'anthropologie. Des Connaissances* 9(3): 311–329.
- Latour B (1983) Give me a laboratory and i will raise the world. In: Karin K-C and Michael JM (eds) *Science Observed: Perspectives on the Social Study of Science*. London: Sage, pp.141–170.
- Latour B (1983) (1989) *La Science en Action*. Paris: La Découverte.
- Lauer J (2012) Making the ledgers talk. Customer control and the origins of retail data mining, 1920–1940. In: Hartmut B, Philip S and Uwe S (eds) *The Rise of Marketing and Market Research*. New York, NY: Palgrave Macmillan, pp.153–169.
- Lazarus J (2012) Prévoir la défaillance de crédit : l'ambition du scoring. *Raisons Politiques* 48(4): 103–118.
- Lee F, Bier J, Christensen J, et al. (2019) Algorithms as folding: Reframing the analytical focus. *Big Data & Society* 6(2): 1–12.
- Lury C and Day S (2019) Algorithmic personalization as a mode of individuation. *Theory, Culture & Society* 36(2): 17–37.
- Lyon D (2014) Surveillance, snowden, and big data: Capacities, consequences. *Big Data & Society* 1(2): 1–13.
- Mackenzie A (2015) The production of prediction: What does machine learning want? *European Journal of Cultural Studies* 18(4–5): 429–445.
- Mackenzie A (2018) Personalization and probabilities. Impersonal propensities in online grocery shopping. *Big Data & Society* 5(1): 1–12.
- Matzner T (2019) The human is dead – Long live the algorithm! Human-Algorithmic ensembles and liberal subjectivity. *Theory, Culture & Society* 36(2): 123–144.
- McGuigan L (2019) Automating the audience commodity. The unacknowledged ancestry of programmatic advertising. *New Media & Society* 21(11–12): 2366–2385.
- Moats D and Seaver N (2019) 'you social scientists love mind games': Experimenting in the 'divide' between data science and critical algorithm studies. *Big Data & Society* 6(1): 1–11.
- Plasek A (2016) On the cruelty of really writing a history of machine learning. *IEEE Annals of the History of Computing* 38(4): 6–8.
- Poon M (2007) Scorecards as devices for consumer credit: The case of fair, Isaac & company incorporated. In: Michel C, Yuval M and Fabian M (eds) *Market Devices*. Malden, MA: Blackwell Publishers, pp.284–306.
- Porter T (1995) *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.
- Pridmore J and Zwick D (2011) Marketing and the rise of commercial consumer surveillance. *Surveillance & Society* 8(3): 269–277.
- Rouvroy A (2018) Governing without norms: Algorithmic governmentality. *Psychoanalytical Notebooks* 1(32): 99.
- Rouvroy A and Berns T (2010) Le nouveau pouvoir statistique: ou quand le contrôle s'exerce sur un réel normé, docile et sans événement car constitué de corps 'numériques'. *Multitudes* 40(1): 88–103.
- Seaver N (2012) Algorithmic Recommendations and Synaptic Functions. *Limn* 2. Available at: <https://limn.it/articles/algorithmic-recommendations-and-synaptic-functions/> (accessed 6 August 2020).

- Seaver N (2017) Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4(2): 1–12.
- Siles I, Segura-Castillo A, Solís R, et al. (2020) Folk theories of algorithmic recommendations on spotify: Enacting data assemblages in the global South. *Big Data & Society* 7(1): 1–15.
- Slota SC, Hoffman AS, Ribes D, et al. (2020) Prospecting (in) the data sciences. *Big Data & Society* 7(1): 1–12.
- Srnicek N (2017) *Platform Capitalism. Theory Redux*. Cambridge, UK; Malden, MA: Polity.
- Steiner C (2012) *Automate This: How Algorithms Came to Rule Our World*. New York: Portfolio/Penguin.
- Strauss A (1988) The articulation of project work: An organizational process. *The Sociological Quarterly* 29(2): 163–178.
- Sweeney L (2002) K-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(5): 557–570.
- Thévenot L (2019) Measure for measure: Politics of quantifying individuals to govern them. *Historical Social Research/Historische Sozialforschung* 44(2): 44–76.
- Turow J (2006) *Niche Envy: Marketing Discrimination in the Digital Age*. Cambridge, MA: MIT Press.
- Yeung K (2017) ‘hypernudge’: Big data as a mode of regulation by design. *Information, Communication & Society* 20(1): 118–136.
- Zuboff S (2018) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. 1st edition. New York: PublicAffairs.