ISA eSymposium for Sociology

Big Data Challenge for Social Sciences:

From Society and Opinion to Replications

Dominique Boullier

Big Data dealing with the social produce predictive correlations for the benefit of brands and web platforms. Beyond 'society' and 'opinion' for which the text lays out a genealogy, appear the 'traces' that must be theorised as 'replications' by the social sciences in order to reap the benefits of the uncertain status of entities' widespread traceability. High frequency replications as a collective phenomenon did exist before the emergence of digital networks but now they leave traces that can be computed. The third generation of Social Sciences currently emerging must assume the specific nature of the world of data created by digital networks, without reducing them to the categories of the sciences of 'society' or 'opinion'.

Keywords: Big Data, social sciences, opinion, propagation, quantification

When in 2007 Savage and Burrows pointed out 'the coming crisis of empirical methods', they were not expecting to be so right. Their paper however became a landmark, signifying the social sciences' reaction to the tremendous shock triggered by digital methods. As they frankly acknowledge in a more recent paper, they did not even imagine the extent to which their prediction might become true, in an age of Big Data, where sources and models have to be revised in the light of extended computing power and radically innovative mathematical approaches. They signalled not just a debate about academic methods but also a momentum for 'commercial sociology' in which platforms acquire the capacity to add 'another major nail in the coffin of academic sociology claims to jurisdiction over knowledge of the social', because 'research methods (are) an intrinsic feature of contemporary capitalist organisations' (Burrows and Savage, 2014, p. 2). This need for a serious account of research methods is well tuned with the claims of Social Studies of Science that should be applied to the social sciences as well.

I would like to build on these insights and principles of Burrows and Savage to propose an historical and systematic account of quantification during the last century, following in the footsteps of Alain Desrosières, and in which we see Big Data and Machine Learning as a major shift in the way social science can be performed. And since, according to Burrows and Savage (2014, p. 5), 'the use of new data sources involves a contestation over the social itself', I will take the risk here of identifying and defining the entities that are supposed to encapsulate the social for each kind of method: beyond the reign of 'society' and 'opinion', I will point at the emergence of the 'replications' that are fabricated by digital platforms but are radically different from previous entities. This is a challenge to invent not only new methods but also a new process of reflexivity for societies, made available by new stakeholders (namely, the digital platforms) which transform reflexivity into reactivity (as operational quantifiers always tend to). This great

transformation is built by and for the sake of brands which are framing everyone's perceptions, from truly commercial brands to scholars for their H-index, cities for their attractiveness, youngsters on YouTube or politicians on Twitter. This is why 'the end of theory' (Anderson, 2008) concerns everybody. It also means the end of some kind of reflexivity, the end of the opportunity to connect long-term trends and opinion movements to personal experience by focusing only on replications, the high frequency dimension of the social, although the social is made of all three of these 'wavelengths' (Society, Opinion, Replications). My proposal is not to disqualify data and processes related to high-frequency replications, but to maintain a pluralistic range of analysis while agreeing to take up the challenge of Big Data and Machine Learning. Finally, there is no reason why the social sciences' 'authority' should escape the challenge to all kinds of authority, intermediation and power generated by the digital revolution. This seems more compatible with a pragmatist approach, which requires us to understand the social from within while not being trapped in the hype of the promotion of this new 'apparatus' that is yet to be established as a convention (Eymard-Duvernay *et al.*, 2004).

Therefore, the main challenges for the social sciences are twofold: fast data, and inductive methods whenever they are used for tracing social behaviours (and not in genomics or in astrophysics for instance). This framework can help compare 'epochs' of quantification, in order to understand the conditions required for these new conventions to be built.

1. Trouble over entities in the Digital Age

Neither people nor identities nor communities but traces are the 'raw' material

For many years, but in an extended way with social networks, computer science has calculated and modelled the social as if the traces collected allowed access to the 'truth' about individuals in a more effective way than do polls, surveys and censuses. Consider two examples, one academic and the other commercial:

• *'The Web does not just connect machines, it connects people.'* (Knight Foundation, 14 September 2008). There you are, this is what Sir Tim Berners-Lee, co-founder of the Web in 1991, with René Caillau, stated to emphasise the transition to a dimension of networks which is neither technological (III for International Information Infrastructure) nor documentary (WWW), but social (GGG for Global Giant Graph).

• Facebook has managed the tour de force of 'normalising' members' declaration of their true identity, that is to say, the features of identity provided by the civil registry, the name and surname, as opposed to the tradition of anonymity on the Web. The platform thus claims to have become the authority of reference or even a civil-registry-alternative, competing with Google in this regard. This is becoming usual whenever one uses a Facebook account for creating and certifying access to other apps, for instance.

Yet there is no guarantee whatsoever of any connection between the identities on Facebook, or Berners-Lee's 'people', and persons identified by their civil registry. What are connected are merely the retrieved accounts and data, and these are only the traces of activity from an entity, which could possibly take on the form of a legal civil status. Based on the scores that classify sites on a search engine, the resulting topology of sites and blogs never discuss their contents as such, but rather the inbound and outbound links that produce a rank of authority or hub, as defined in the network topology (Kleinberg *et al.*, 1998) and are not a civil status. It should be noted here from the outset what I mean by traces, in order to distinguish them from data. Traces can range from signals (so-called 'raw' ones, generated by objects) to unstructured verbatim; they can be traces exploited in databases (links, clicks, likes, cookies)[1] by operators or platforms, but also captured independently of this through the API and, as such, fall outside permanent relational databases (Bowker and Star, 1999). Traces are not necessarily pre-formatted for a specific calculation nor are they dependent upon aggregation that can then be applied. It is easy to argue that, despite everything, 'behind' these sites or 'behind' these clicks, there are most

certainly people, but that does not alter the fact that the algorithms themselves do not take this onto consideration and that, furthermore, no guarantees can be given in this regard. Traces understood in the restricted sense are produced by platforms and digital-technological-systems, but are not the 'signs' or evidence of anything other than themselves, as long as relationships with other attributes are not created and validated. Of course, they are transactional data because there is no trace without a distributed setting of relationships including humans and non-humans. This differs radically from the data that can be recovered en masse from client files or from administrative acts. Certainly, the Big Data methods for calculating can be applied here in both cases, but the traces are a priori independent of other attributes, in particular socio-demographic features, which are rarely used in correlations between traces. Relationships with more conventional parameters in data sciences are limited to time (a timestamp) and location (geo-location tags), which allow for the production of timelines and maps that become simplified modes of representation for traces. Can the social sciences accept this shift in the 'raw' material they usually process?

Traces are produced by platforms

In order to get into a thicker description of this social life of data, consider how it works for Amazon or Apple. The Web is no longer distributed but monopolised by these four platforms – GAFA that centralise the majority of traffic, with Twitter extending this traces industry, while Microsoft is trying to get back in the battle by purchasing the social network Linkedin. What I described above about traces and their detachment from the legal ID's features was more central for platforms such as Google and Facebook. The pretention of these platforms to perform their version of a 'society' should be noticed according to the three wavelengths of social sciences above mentioned. On such consumer platforms as Amazon and Apple, it is not people who are linked to one another but above all tastes (books or music originally), reflected in traces of purchases, and thus of choices, which can be treated en masse to produce patterns and profiles, independently of personal information. It should certainly not be forgotten that all these platforms without exception are also very fond of civil status-type data, phone numbers and other highly attractive resources to advertisers, to whom they are sold. Amazon and Apple are designing the perfect market environment in which preferences can be traced, they built their own sociological device focused on 'opinion', the second wavelength I mentioned above. The marketing methods thus developed are largely based on mass advertising or on emails addressed to IP addresses, or emails that have clicked on an article (retargeting), and much more rarely on sophisticated links with other attributes of the supposed people attached to these addresses or clicks (profiling). Traces of digital behaviour are thus a particularly profitable 'raw material', without the need to appeal to the social sciences, although platforms, social listening agencies and marketing experts readily make use of academically coined terms such as 'communities', 'networks', 'engagement', 'tastes' and so on. A distinction could therefore be made between social network platforms and consumer platforms, despite the trend towards some combination in many of them. On the one hand, the social registry that Facebook and Google tend to build is akin to what states and social sciences did when designing censuses in an attempt to assemble socio-demographic features of individuals and traces left while using apps and interacting with 'friends'. On the other hand, consumer platforms such as Apple or Amazon are more dedicated to a sort of mapping of affiliations, types, styles, and preferences, all of which are more ephemeral or more precisely cyclic than the traditional social features. This is quite close to the traditional social study of moods, fashion, and opinions in political life or in any marketing area, which consist of individual expressions, aggregated and mapped to make the opinion live, or of trends in cultural tastes. However, it is true that the social sciences and marketing research methods did use sampling and were indeed concerned about maintaining the connection with the socio-demographic features of the individuals they selected in their samples, which is not true for Amazon or Apple and any other consumer platforms. These two sets of digital data (socio-demographic features on the one hand,

with Google + and Facebook, and preferences and tastes on the other, with Amazon and Apple) are not so new, indeed. Due to the specific way of producing them in a natively digital format and sometimes on the fly, they cannot be considered as substitutes for other protocols such as censuses or surveys. This is nevertheless how they come to be used by many data scientists and by some social scientists seeking shortcuts to the social. Twitter plays its own part, which paves the way for social sciences of a third kind, accounting for the high frequency propagation processes.

2. The making of 'society'

As mentioned above, computer scientists do not hesitate to make use of social science categories such as communities, networks, and so on, even though they very often lack adequate knowledge of these concepts. This is a part of what I call 'algorithmic positivism'. Yet one cannot blame them for the use of a concept like 'society' since there is nothing more 'taken-for-granted' among members of society. This is why it may be risky to adopt a constructivist viewpoint and try to convince data scientists that 'society' was designed and promoted in such a successful move that no one dares to question it. If we want to understand the historical times that we are living in as regards quantification methods (and the social reflexivity that is attached to it), we need to look back to the times of the construction of this entity, 'society'. Let us pretend here that Durkheim succeeded in making 'Society' exist. The term was not coined by Durkheim, obviously, although its history is not a long one. The archaeology of the concept of society (Latour, 2005) could be further enriched by calling upon the work of Quetelet, who produced the 'average man' which long remained the key to all statistics. At the end of the nineteenth century, however, and largely thanks to Durkheim's genius, 'society' took a strong stance regarding 'community', which was still prevalent (see Tönnies, 1887). Durkheim's early work on the 'division of labour in society' (1893) was not based on statistical methods, but instead laid the foundation for a model of social types, aggregated in mechanical and organic solidarity. Detailed examination of legal systems served as demonstrations and therefore relied on the groups formed or being formed that are legal systems in their traditional or more modern aspects. With 'The Suicide' (1897), the method was set up to extend the discussion of the types that would reveal anomie to be a problematic situation. But reliance on data records produced by states, from their various components (ministries, prefectures, governments) became key to the demonstration. It was these aggregates that are explained or explanatory, using a method of comparison between countries, regions, counties or districts, where possible and necessary. The method depended entirely on the available data and could not afford to criticise or to question the procedures for the production of this data, despite the countless limitations identified upon publication (Douglas, 1967). By organising all his systems of proof around these national administrative statistics, Durkheim found a quantitative analogue for his conceptual choice that put 'Society' in a separate status from all manifestations and individual behaviours. Durkheim's *whole* became an entity of the second degree, 'Society', (Latour, 2005), while the censuses and other state-data-registers simply perform the task of recovering individual, administrative events (marital status, judicial procedures, etc.), formatted in identical categories and aggregated to reveal the behaviour of populations. Durkheim's strength of conviction would have been to make these statistics exist as equivalent to his 'society', where the quantification is able to account for a 'whole' through the quality of exhaustiveness, while the concept accounts for the agency of the social structure as such.

It is necessary to note that a form of 'convention' was formed between data producers from the state administrations and the emerging social sciences. Together they produced the entity 'society' as the object to be tracked by the state for the purpose of governing and to be explained for scientific reasons. The result is the widely shared and obvious fact that 'society' exists, and the methods that allow it to do so have no grounds to be questioned because they demonstrate both their scientific and their operational value: they are 'tools of proof' and 'tools of government' as Desrosières (2014) put it.

The age of calculations and calculators

Other historical proximities are noteworthy, that do not mean causality but that do allow for an understanding of the power-gains this approach affords in making society exist. In 1890, Hollerith used his machine (that he had invented a few years earlier and for which he filed a patent application in 1886) to conduct the U.S. census. The Census Bureau had not yet finished processing the previous census dating back to 1880 when it had had to start the next one. A change in technique was both necessary and available. Hollerith's tabulating calculation-machine did the work and was sold for doing censuses in several countries. His company would later be transformed into IBM by Watson, in 1926. We can see how the power gained in the counting and description of populations reinforces the status of the State and offers it supposedly useful sources of information for its governance. The pretence of the calculation's exhaustiveness seemed to fulfil the promise of the concept of society: a technical device capable of inputting all that existed, that as Hollerith's census-procedure-equipping machine. It should be noted to what degree the *investments of form* (Thévenot, 1986) that are censuses, end up being events in a given population and become indisputable, appearing almost ritually, to portray all that is social onto the members of this society.

Alain Desrosières had amply demonstrated this process by showing how Durkheim's concept of society took the genesis of nation states into account. Nation states rely as much on these figures as they do on infrastructures. From this point on, territory became a key mediation, of which we find traces in the emergence of social welfare policies or in the developing national commodity markets, owing to the railroads, and then in national electoral campaigns through the media (the press also circulated via the railroad, after which came radio, that would become a vector for the emergence of opinion).

In modern Western countries, the State gained its legitimacy through electoral processes that rely on nations, those 'imagined communities' (Anderson, 1991) that work as content while the State is the container (Boullier, 2011), or the 'materiality' and the 'statement' that constitute the apparatus in Foucaldian terms (Foucault, 1982). This is why 'society' is always enacted in various 'nations' although social theory tries to extract it from the limitations of national boundaries. The first generation of social sciences was indeed doomed to methodological nationalism (Beck, Sassen) and still has problems inventing the methods to account for a globalised world made of flows (finance, media, commodities, migrants, and so on). However, we shall see that when examining this digital world, it is quite difficult for us to escape 'methodological platformism', due to the total dependency on the traces data platforms deliver!

Durkheim's achievement has been to form an assemblage of very powerful mediations:

- Censuses
- Assembled and formatted by public administrations
- Under guarantee of exhaustiveness
- For States
- For government purposes
- To produce 'society'
- Using tabulating calculation machines

3. The construction of 'opinion'[2]

The contemporary situation is undoubtedly not that far from another key moment in the history of the social sciences that would help us to understand both what is happening and the conditions of felicity of new conventions. This is why we will consider some features of that period when opinion polls were invented, as an indication of the equivalents in our times. After the first generation of quantification, used by Durkheim to build the concept of 'society', we could give the label '2G' to the emergence of public opinion in the late 1930s. In 1936 George Gallup was able to predict the election of Roosevelt over Landon, based on a survey of 50,000 people. Roper and Crossley had done likewise at the same time. Gallup not only impressed the media and policy makers, he radically disqualified older methods (straw polls), including that of the Literary Digest, based on responses from 2 million people, whilst even predicting their own erroneous results (Osborne and Rose, 1999). This impressive demonstration lays the foundations of the survey's reliability and of investigative sampling

methods. The exhaustiveness of inquiries on entire populations was indeed sacrificed in the process, but the new approach managed to produce accurate results, provided that the terms of *representativeness* were respected. It nevertheless failed to predict the victory of Truman in 1948, whose voters changed their minds in the last ten days. Methods thus applied to political life and to life-size tests as important as a presidential election had previously been tested on readership studies for which Gallup had operationalised stratified sampling. In fact, these methods had already been applied by the Norwegian Kiaer in 1894. Similar statistical methods in the field of agriculture and later in unemployment in the early 1930s, in the USA, underwent profound changes, from the correspondents' method to random sampling based on probabilistic approaches (Desrosières, 1993) – as Emmanuel Didier has also shown (Didier, 2009). Quota methods based on 'sensible choices', where the selected sample is matched with certain properties of the population identified by the census, were however different from those methods of stratified random selection, and were even despised somewhat by statisticians[3] (cf. Stephan quoted by E. Didier). The data collected were also very different, since statisticians from agricultural or employment administrations wanted to obtain 'facts', but were nevertheless obliged to rely on statements, not measuring machines, even if they attempted the latter with the 'crop meters'. Yet the sampling-legitimisation-operation generally succeeded, primarily thanks to Gallup's performances (1939), which were dedicated entirely to other social worlds; those of 'public opinion' and not 'society'. The latter remained a reference of statisticians of the federal state and its offices. It was unquestionably in the context of the mass media that the importance of sampling was recognised. With Ogilvy, Gallup studied film audiences, and then with Crossley, at Young's and Rubicam's, he studied radio audiences, using telephone interviews before even making a proposal to conduct the election polls. From this point of view, Gallup's name must be associated with that of Lazarsfeld, who in the same period, in 1936, launched a 'Radio Research Program', based on audience-research-work begun in 1930.[4] Together with Merton they launched the *focus groups* method as early as 1941, and their study of Decatur in 1945 provided the data for the analysis of 'Personal Influence' published in 1955 (Katz and Lazarsfeld, 1955). The latter study established the framework for analysis of the 'two-step flow' in which mass media play a role, but through the mediation of various kinds of influential relationships.

The links between the mass media and politics are thus elements of new statistical methods. As Alain Desrosières noted (1998), a national election's predictability actually depended upon the formation of a common public media-space across the United States, and only the radio could do this in such a way as to make the voter's knowledge about electoral candidates comparable. Considerable media transformation and the mass media (radio at the time) established the conditions for the emergence and validation of a survey technique, which thus opens up a whole new era, most notably for political science and market research. Moreover, it is 'public opinion' itself which takes on a measurable existence with these sampling methods whose performative power will by far exceed their experimental phase. This move does not disqualify the previous assemblage made by censuses and mecanographic calculators that will soon be transformed in computers. But the role of radio (and telephone for the report of results and their aggregation at a faster speed) contributes to a new apparatus.

Markets and national publics: the scale of the media

The missing link in my description remains the vehicle of financial incentives for such investments, needed to understand a public. Communication agencies such as polling organisations cannot live solely from their campaign activities, even if they do bring them high visibility and renown. From the outset, their target was the mass media, as noted above, for one essential reason: audience measurement has been the key to the distribution of advertising space, since the dawn of radio and then later with television (in 1941 the first advertisements were aired on American television for Bulova watches, during a baseball game). But these measures also serve to monitor the

impacts of these campaigns on the minds of consumers, giving an unprecedented boost to marketing, which in turn drives increasingly sophisticated communication strategies (Cochoy, 1999). Brands have thus been present, from the outset, in methods of inquiry into opinion using sampling; that is, from the moment such investigations were aimed primarily at mass-media audiences. Market research on consumer goods developed at the same time, from the 1930s, and in the same movement of national standardisation of products, as Desrosières pointed out. The production of a unified national territory, through the media, that included transportation and mail, established a new condition of felicity for these survey methods. This allows me to draw a direct parallel with the recent creation of a global market, this time beyond the national ones, through the domination of digital platforms. Google, Apple, Facebook and Amazon have produced the same effect on a global territorial scale as radio and the railway had on the territory of national markets. This is in line with the work of McLuhan (1964), for whom the change of scale in itself constitutes another world far beyond the property or goods exchanged.

Public opinion exists, I measured it!

The work done by Gallup on the operational side and Lazarsfeld on the scientific side is therefore not simply a marketing operation or a face lift for the social sciences: it provides whole societies with methods with which to analyse themselves and to represent themselves – as opinions. Tarde (1901) has certainly highlighted the importance of these views, yet it is only when the metrics are established and produced in a conventional way that opinion finally exists. Only the media's control and their ability to produce a unified public in a national territory enabled this methodological assemblage to hold. The 'whole' referred to by the polls is in fact originally the *public* formed by the media, which allow the *audience* to emerge as *public opinion* and to make it permanently visible and measurable. This connection between audience measurement and monitoring methods for public opinion, a connection that is both technical and historical, must be regarded as the key to the device: the media want, above all, to measure audiences, as did Gallup for reading, but the techniques in place turned into predictive voting tools, which justified this betting on public opinion. The whole 'audience' or even 'public' has mutated into 'public opinion' and managed to detach itself from its own reference within the media (which measure themselves), for the purpose of being exploitable for brands to measure the influence of their campaigns. The parts (Latour *et al.*, 2012) that are individual expressions are preformatted to be recordable and calculable, but the link between the parts and everything else is made only by the pollsters' black boxes. The rigorous, scientific precautions are upheld through 'confidence intervals' (defined by Neyman in 1934), which keep a reference on the exhaustiveness of the studied population. Bowley (1906) proposed these principles in 1906, when speaking of the 'probable error', allowing for the clear linking of the polls and the emergence of statisticians' probabilities. But soon these precautions will disappear from the findings, as seen in the contemporary media. At this point, everyone knows that 'opinion exists', whatever the report about the artefacts needed to make it exist, and despite what Bourdieu said about it (1984).[5] It has been naturalised, 'taken for granted', and the sampling methods lie buried beneath the powerful performative effect of these immediate, aggregated indices. The approximation remains acceptable, especially with the repetition of the same questionnaire over time (by panel, independent rotating sample) under identical conditions, 'all things being equal'. It allows for the smoothing-out of biases which then become acceptable by convention. Such successful convention work focuses on the same assemblages of mediations already mentioned for society:

- the 'surveys' and 'polls' (from individual expressions framed by questions and thus made calculable)
- assembled and formatted by pollsters
- guaranteeing the representativeness of samples (sampling)
- for the media
- for the purpose of monitoring
- to generate public opinion (and audiences).

As Alain Desrosières (2008) put it, the essential thing is not whether these data are a reflection or mirror of society or something else, but rather whether they 'make something that stands by itself'. Note that there is a new element at work in this chain: that of the methodological limitation, expressed in terms of the representativeness of the samples, because this element is still missing in digital traces, which explains much of the uncertainty and suspicion on all results compared with the polls, for which the 'biases' are well known but have been controlled by convention since the 1940s. The 'consolidation' that Emmanuel Didier (2002, 2007) describes for statistics and surveys remains to be done. This rather long account of the successful fabrication of opinion was necessary not only to understand the similarities between that 'epoch' and our current one, but also to measure the distance and the work required to produce conventions of equal quality (Eymard-Duvernay *et al.*, 2004) that would make 'traces' exist as entities recognised by the social sciences. We certainly need to consider opinion to be a social reality that lives its life and is no longer called into question, thanks to the quality of technical and institutional arrangements that stabilised its mode of appearance, notwithstanding the many critiques that they still face. And they will face many critics more often due to the complete remodelling of the media conventions by the social networks activity. To be sure, the worlds of social science and marketing differ, yet for years they have used the same methods and even the same samples while being able to distinguish themselves from one another. Within the shared new world of traces emerging on the web, how can we invent the social sciences in a way that suits these traces whilst admitting the conditions of their production and utilisation?

4. Three generations, three points of view on the social

Building on these two historical landmarks of 'society' and 'opinion' and on the mediations which managed to make them exist as taken-for-granted entities, I propose a comparative table in which the next generation of social sciences is designed along the same criteria in order to devise some roadmap for the consolidation of the conventions. Digital devices generate new opportunities for quantification but not only in terms of volume. My main statement is to emphasise the emergence of new entities, apart from society and opinion; entities that have an agency of their own and that I call 'replications', propagated along networks in the material aspect of traces. The following table will make the comparison more visible and will introduce my understanding of the stakes of third-generation social sciences.

The excessive coherence of any table must not make us forget that what is at stake is the construction of a proposal for the third-generation of social sciences, which is in no way guaranteed. Actor Network Theory has indeed laid the foundations, building on the methods of scientometrics (and previously of bibliometrics), and Tarde (1890) did announce the principles as a general theory of imitation (including opposition and invention, not to be forgotten). But, for now, the trend is more at the 'end of theory' and the occupation of the field by the Web platforms (GAFA). They produce, calculate and publish themselves on these traces, and all for commercial purposes primarily because major brands are demanding these approaches. I shall describe the choices available to the social sciences in such a context. As the table suggests, not only does Big Data challenge the status of social sciences in terms of empirical capacity and of modelling without theory, but Big Data really needs Big Theory for the social sciences to keep their role alive. What begins as a historical comparison, where the succession of generations can let us believe that the next one makes the previous one obsolete, is turning into a more diplomatic statement, where each approach is able to grasp one specific aspect of the social that others cannot account for. This derives from the social studies of science led by Bruno Latour (1987), who has deeply transformed our understanding of the process of scientific knowledge. The insistence on the agency of devices, of scientific devices, can be used for social science as well, because one could say 'we've got the sciences of our devices'. The census and the polls have built entities that highlight some specific dimensions of the social which came to be framed as 'society'

Table 1: The three generations of social sciences

	1st generation	2nd generation	3rd generation
Concept of the social	**Society/(ies)**	**Opinion(s)**	**Replication(s)**
Collection devices	Censuses	Surveys/Polls	Platforms
Validation principle	Exhaustiveness	Representativeness	Traceability
Co-construction institutions/ research	Registers/ inquiries	Audience/ Polls	Traces/ Repurposed digital methods
Major players of reference (and funding)	States/ Nations	Mass media/ Audiences	Platforms/Brands
Operational actors	National Institutes	Polling organisations	Web platforms (GAFA)
Founding authors	Durkheim	Gallup, Lazarsfeld	Callon, Latour, Law
Key problems of scientific approaches	Division of labour and the welfare state	Propaganda and media-influence (audience measurement)	Science and technology (scientometrics)
Technical conditions	Hollerith's machine (tabulating calculation)	Radio and telephone	Internet, Web and Big Data
Semiotic formats	Crosstabs and topographic maps	Curves, bar charts / pie charts and topology of influences	Graphs, timelines, dashboards
Metrics	Statistics	Sampling	TPS (tweets per second) (scores) and similarity matrix
Technical criteria for data quality	Relevance, accuracy, timeliness, accessibility, comparability, coherence	Confidence intervals Probabilities	Volume, Variety and Velocity
The social science's dominant modalities	Explanations	Descriptive and predictive correlations	Predictive correlations, memetics

and 'opinion', but digital devices make something new appear.

Generations, waves or viewpoints on the social?

From this historical account and from this diplomatic move, a general pattern emerges. The social sciences adopt a rather limited number of perspectives related to these devices and to the entities that have been constructed. In short, 'society' is generated by a 'structure' approach; 'opinion' is produced by a 'market' approach; and 'replications' (those of the digital world) are discovered through an 'emergence' approach. When trying to account for the various controversies in the social sciences, one will eventually revert to this classification. And, more importantly, these old disputes can be considered as 'viewpoints' on the world, at the epistemic level, equipped with different devices and targeting different entities. We may talk of these differences in terms of 'wave lengths' as we did previously: 'structure' analysis focuses on long wavelengths, 'market' on mid-term and cyclical ones, and

'emergence' on high-frequency ones. This should remind historians, with a slight translation, of the Fernand Braudel's famous distinction: long-term history, cycles, events (Braudel, 1996). He was right not to condemn any of these viewpoints for the sake of another, and only to advocate sufficient diversity (against the trend, at the time, of focusing on events). All these three viewpoints could easily fall under a general theory of attraction, as accounting for the social in general: long-term attractions of social traditions, habits, and repetitions (the ones sociologists love to love); cyclic attraction of fashions, political opinions, tastes and preferences in general (the ones economics, marketing, psychology and political science are fond of); and high-frequency attraction of replications (that make the buzz as well as the financial speculation). Imitation is not a matter of long-term structure,thecause it emerges at the very moment it occurs, nor a matter of strategic decision, because the time lapse is so short that reactions are quasi 'unknown' to the subject, for these subjects are only the targets of replicators, entities that connect every mind in a millisecond. This existed before the digital continent emerged, such as in 'olas' or Mexican waves, crowd moves, rumours, and so on. But no social science has ever been equipped for documenting these contagious processes (Sperber, 1996) in which speed is so critical. Today the platforms amplify these processes and produce the leverage for measuring and tracking them.

This extension from a historical approach to different conceptual frameworks of the social has many implications and help understand why it is so difficult to speak of solid ground for social sciences: there are three points of view, none of which is wrong but only brings a different perspective to make different entities exist.

Generations	1st Society	2nd Opinion	3rd Replications
Points of view	Structure	Market	Emergence
Wavelengths	Long term	Mid-term, cycles	High frequency
Features of the networks	Structure	Nodes	Circulating entities
Main concern	Positions	Decision	Propagation
Process	Inheritance	Arbitrage	Neighbourhood
Status of human actors	Inheritors, determined subjects	Strategist, decision-maker, rational agent	Vehicle for memes' propagation

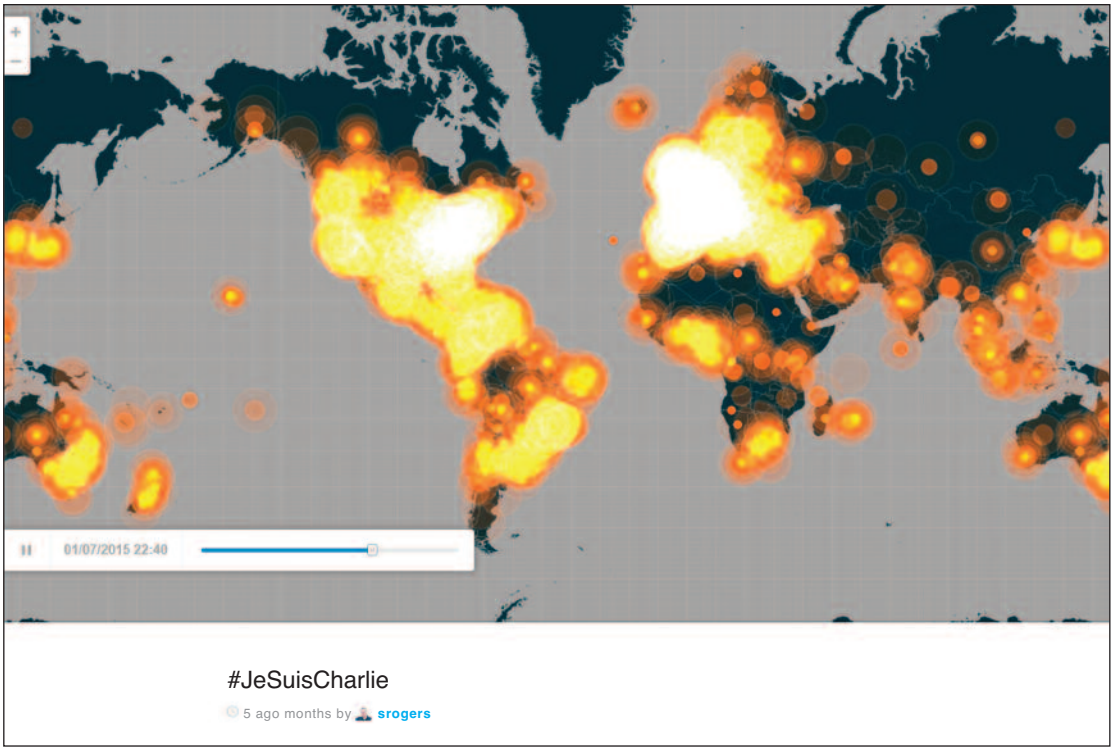
This table can lead to a kind of 'quantic sociology', because one cannot adopt all points of view at the same time and cannot sum up the views of 'the world out there' in a 'whole' by assembling the outputs of each method. When changing methods, tools and observers' points of view, the focus shifts from waves to particles, from positions to trajectories. Even the network analysis is not a shift as such since there are many ways of selecting the features of the network which are agencies of their own: its structure (homophily for instance), the nodes (influentials for instance, Watts, 2007), the entities that circulate and make the connection (goods, messages, memes for instance). This pluralist and constructivist view of social sciences approach offers the opportunity to distribute agencies among all entities and features while accepting the un-completeness of any account of the social.

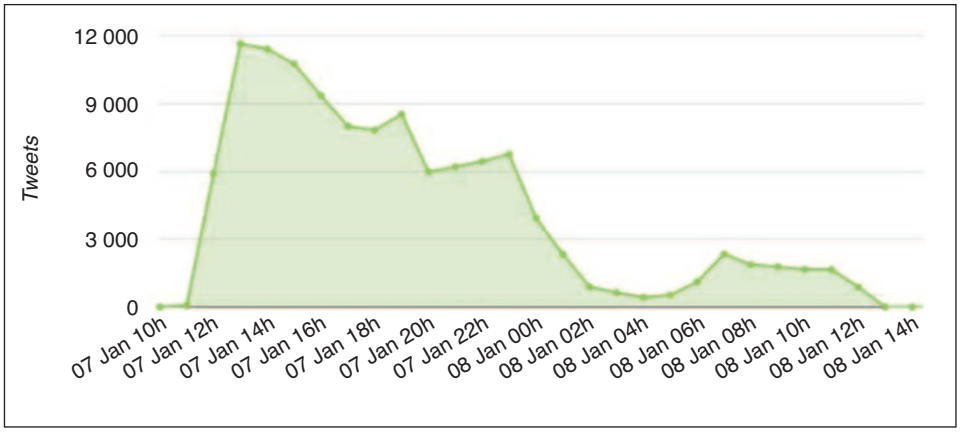
The wavelengths of JeSuis Charlie

Let's illustrate this with a discussion about 'Je suis Charlie', the global movement that followed the Charlie Hebdo and Hypercasher attacks in January

2015 in Paris. A controversy emerged in France around the book of historian and anthropologist (expert in demographics) Emmanuel Todd, a rather famous and maverick kind of scholar. He rushed to publish his book (*Qui est Charlie?*, in May 2015) based on the correlation between the number of protesters in the streets in France and the age-old history of Catholicism in each city. The book used well-known maps of the conflictual period of the French Revolution that were based on priests' acceptance or rejection of the new regime. I shall not get into the discussion of the main hypothesis, that demonstrations were stronger in the regions where Catholicism used to be strong and are a significant anti-Muslim signal of this 'zombie Catholicism' in France, where traces remain although the active practice of the religion has fallen sharply. However, the method is somewhat unbalanced in favour of long-term waves: the only indicator used for the analysis of the demonstrations is the number of participants estimated both by the police forces and by the leftist newspaper *Libération*, in a sort of reciprocal compensation. The figures are 'botched' as the author said, but will do in this state of emergency (in order to deflate the mood of unanimous republican celebration). What is more interesting, is that he did not consider it to be of any interest to ask the demonstrators any questions since, as he said, 'very often, they did not know how to explain their participation', as all of them were 'carried away by the mimetic intoxication of a saturated media space' (p. 21). The three generations in the same sentence, it seems: surveys do not make sense with 'cultural idiots' such as demonstrators and their 'opinion' does not exist since the media are reaching a level of excitement that is contagious (mimetic intoxication), and this does not need any investigation, except a critical stance from the point of view of the true and only social science, the one based on long-term indicators that crush all the mediations of events and of opinions in a single move. Of course, French experts in political science (Mayer and Tiberj, 2016) clearly documented the capacity of sampling surveys to account for a part of their motivations and their background (surveys were conducted on a large scale right after the demonstration). And Twitter specialists published dynamic maps of the contamination of the hashtag "**#JeSuisCharlie**" which managed to propagate all over the world in the following six hours. This global dimension of the process was not relevant for a social scientist who relied on its 'methodological nationalism', due to the idiosyncrasy of a general cause such as the French Revolution. More significant is the fact that even at the level of the demonstration, the specificity of the event was missed by the author: when two demonstrations took place, as in the case of Marseille (one for the left, one for the right), the score was just an addition of the two, as if a common demonstration does not make a difference in such circumstances. The limits of the model when applied to Paris or to Strasbourg were acknowledged, but never could they question the model itself. The demonstrations did not have a life of their own; they were unexpected emergences of a 'collective' that would be quite difficult to label ('the people'?), and for sure doomed to disappear the next day, but were still a striking experience for the participants. According to Todd, these demonstrators were just a number that can be correlated to the only deep, real and permanent causes of French political behavior, two centuries later (with this delay explaining the 'zombie' rhetorical trick, since the mediations – religious behaviour – disappeared from the long-term radar!). For opinion experts as well, the use of polls was the only way to account for this sudden change in the mind of so many people, and from their individual expressions they were able to build an image of what 'public opinion' was saying – a much more complex view than the simple causalities of long-term social sciences. However, they did not make any use of the tremendous amount of data generated on internet, via Twitter, Facebook, Instagram and other social networks, and on blogs, media websites' comments and so on. The public's emotional experience was expressed not only during demonstrations but also earlier, immediately after the attacks, through a contagious extension of some hashtags and especially through the propagation of the meme "**#JeSuisCharlie**". Thanks to the traceability provided by social network platforms, the global phenomenon of contagion becomes visible in real time, as shown on the following map and timeline.

(source:
http://www.reputatiolab.com/2015/01/analyse-de-je-suischarlie-sur-les-reseaux-sociaux/





These traces data seem to provide very little information except for the volumes that they reflect: nothing about the Twitter accounts, profiles or networks; just a very simple display along timelines, showing how strong yet ephemeral these kinds of propagations are, and another very basic spatial representation, powerful for demonstrating the extension of the contagion worldwide but not precise enough to allow correlations with any other spatially referenced dataset. This is enough however to trigger more investigation without discarding such a global phenomenon, and to challenge the social sciences to account for these high-frequency and short-term waves without using their traditional tools and concepts. All three points of view can bring some insights, none of which can be claimed to be the ultimate cause – in Tarde's critical terms, the *'cause-finaliers'* –, or to assemble all of them in an overarching vision of a 'whole' which is just an effect of the devices we adopt to delimit the social life continuum (Latour *et al.*, 2012). Some may look for the translations from high-frequency traces (expressions or hashtags) to the cycles of opinion or to the long-term memory of an event that will become a part of the globalised collective experience. Why not? But there is so much to do first to understand the specificity of these waves of traces and to be sure of what can be extracted from them, that a precautionary principle should apply to the reconnection of all these wavelengths. In order to do so, we need to build the conventions that will guarantee 'sufficient scientificity' to the social sciences of the Third Generation.

5. Brands' grip on traces

This may look like a gamble since all these traces are produced and controlled by digital platforms and, ultimately, by the brands that are funding these social networks for their own purposes: is there a way to escape the 'methodological platformism' (or 'platform bias'; Marres, 2017) that can be observed for instance through computer and social scientists' focus on Twitter? Where does this fascination with traces – as opposed to data from registries and surveys – come from, despite their limitations? Does it come from the impression that any observer can grasp the whole of the phenomenon on such open platform as Twitter or to the fact that Twitter's code encapsulates a quasi 'meme machine', i.e. the Retweet function? The business model seems to be a better rationale for this attractiveness. The traces are actually a key resource for brands to monitor the impact of their actions on the public. Reputation and renown no longer translate uniquely into audience measurements; this would be a simplistic import of measures built for the mass media. On networks, one must measure not only a form of audience (the reach) and the most basic activities of its uncertain public (likes, stars), but also more sophisticated activities such as comments, which constitute what is called 'the engagement rate'. Brands are fond of these traces and it is they who fuel the turnover of all these platforms, and thereby of the entire Web. The opinion mining and sentiment analysis tools (Boullier and Lohard, 2012) are thus the answer to 'the marketer's anxiety after the product launch'. However, the extension of this brand domain reaches into all activities, whether commercial, cultural, political, institutional or even interpersonal when everyone must measure her excellence with rankings, as researchers are requested to do (Bruno and Didier, 2013). It is the brands' methods that take precedence everywhere and impose their law and their pace, even on public services. But what concerns these brands primarily is not structured and constructed data to test e.g. causality, but many traces that function as indicators and alerts, even approximate ones, not at the individual level but at the level of trends. Similarly, it is not reflexivity that is sought but primarily reactivity, the ability to determine which lever to act upon in relation to the dimensions (features) of the brand that are affected. The political world itself is now caught up in the spiral of reactivity and its addiction to tweets led us to consider that we have entered the era of *High-Frequency Politics* in the image of the *High-Frequency Trading* of speculative finance (Boullier, 2016).

Platforms pick up traces of the actions and clicks of Internet-users or machines, in a standardised format that aggregates them and produces a score. This score is displayed and can be used by the platform itself to show trends to guide the placements of advertisers who also seek to achieve certain effects and to

optimise their investment or communication choices. In a simplified format, this is the string of events that was produced. The performative mechanism works almost identically to the audience measurement (Boullier, 2004). Some then try to develop a critique showing that the 'likes' aggregate very different sorts of behaviour, including even purchased likes. The limited quality of the traces is observable on all platforms, but these limits may be intrinsic, when they do not meet the criteria for traceability that we consider crucial in order to exploit them, or extrinsic when we criticise their lack of reliable relation to the 'real' world. It is the latter stance that we find in boyd and Crawford as regards Twitter: 'Some users have multiple accounts. Some accounts are used by multiple people. Some people never establish an account, and simply access Twitter via the web. Some accounts are "bots" that produce automated content without involving a person. Furthermore, the notion of an "active" account is problematic. While some users post content frequently through Twitter, others participate as "listeners". Twitter Inc. has revealed that 40 percent of active users sign in just to listen.' (boyd and Crawford, 2011, p. 6). Other studies (Driscoll and Walker, 2014) tested the data produced from various access methods offered by Twitter, for example, and showed that the Search API, the Streaming API and Gnip Power Track (paid service) provide very different results. The latter method for instance collected a much larger number of tweets, but not uniformly according to the requests! This means that the traces collected are entirely dependent on the collection mechanisms, which is not surprising but which we do tend to forget since other, older methods have become conventional.

But these limitations hardly concern operators, platforms or advertisers. Their action/reaction works in the performative mode, where the likes reveal/produce a reality that will initiate strategies to influence the likes, in a self-referential cycle to which one could also assign audience ratings. However, in the case of audience ratings, all advertisers and programmers have agreed on stable criteria and produced a shared agreement, and evidence of this has come to forcefully impose itself every morning in the management of programmes in the mass-media. Social network platforms and advertisers have not yet reached a stable compromise, which explains the proliferation of services that claim to be the standard, as I have shown in the case of Klout (Boullier and Lohard, 2015), and that want to become the Nielsen of these measures. It is easy to see the difference between these principles and the traditions of the social sciences, as G. Bowker does (2014), and to show their extreme reductionism: 'If I am defined by my clicks and purchases and so forth, I get represented largely as a person with no qualities other than "consumer with tastes". However, creating a system that locks me into my tastes reduces me significantly. Individuals are not stable categories – things and people are not identical to themselves over time. (This is argued in formal logic in the discipline of mereology and in psychiatry by, say, ethnopsychiatry.) The unexamined term the "individual" is what structures the database and significantly excludes temporality' (Bowker, 2014, p. 1797).

Bowker has cause for concern from the point of view of 'society', but the third generation of social sciences is not so much interested in 'society' as in other social processes created by other devices, but which, nonetheless, cause us to act. Brands, reputations and recommendations as they are exploited by Amazon can certainly be forcefully re-injected into a matrix 'society' to make them say what they are not made to say. But they also say something of themselves, from another world, that of the power of recommendations and contagions that the social sciences are reluctant to understand. It is as if the sociology of 'society' were reliving an analogous experience to the one that anthropology provoked, that of the necessary shift with the modern world and its categories. To be sure, Durkheim did first use it to analyse religions (Durkheim, 1912) and to employ traditional societies and totemism for his demonstration of the power of society on individuals. But Mauss (1950) made a great side step in recognising the power of things and the spirit that persists within them, the 'mana', for which Levi-Strauss criticised him.

It is useless to complain about the imperfection of the data and its approximation because we are now dealing with traces through a process of pervasive traceability. This traceability connects entities that did

not exist beforehand, but are now endowed with an IP address (thanks to the availability of IPv6, or 3.4 $\times 10^{38}$ addresses) and so can interact just as humans do via their machines. The vital statistics which are the reference base for the 3rd generation of the social sciences are no longer the censuses but an index of IP addresses, totally agnostic about the entities that 'are behind' because all act almost equivalently and cause the others to act. This shift may seem radical, but it helps to hold together the approaches of previous generations while watching the ever-present world with the tools and relevant categories at hand.

I have drawn up a table that merits systemisation. Digital networking generates:

- traces
- assembled and formatted by platforms
- for brands
- with a view to reactivity
- in order to produce rankings or patterns.

This situation is akin to the two other key moments in the existence of the social sciences, especially sociology and political science, discussed above. However, these new methods and principles have still to be arranged in such a way that they transform themselves into 'socio-technical conventions'.

Conclusion

Science and Technology Studies have always rely on historical and narrative approaches that help understand how the black boxes of any 'fact', 'result' or 'device' were designed along time, through careful assemblages of entities, properties and allies. And in the wake of this historical look to the emergence of new phenomena, very often the old and familiar assemblages appear in a new light and are reconsidered. This is what innovators and scientists do every time they make a move: their vision bears a reinterpretation of the previous taken-for-granted black boxes. The digital revolution brings so many disruption that the understanding of many 'old' technologies, methods and practices can be reassessed, as it is for media, music, printing, and so on: under the scrutiny of this hard pressure from digital innovation, social sciences appear to rely on a set of procedures, of devices and of selective attention to some features of the social. The omnipresent technology in this digital move obliges us to think again how was designed our previous assemblages for quantification. Of course, scholars like Desrosières paved the road for that, long before Big Data meteor hits the planet social sciences. But the need to understand what is occurring and why these changes are so troubling can be profitable for all social sciences, as Burrows and Savage had envisioned many years ago. I hope this tentative framework can help face the challenges of the new conventions to be built in a more informed way and to give a chance for social sciences to reorganise themselves in order to maintain their social role of reflexivity without sticking to disciplines that may be heading for their fossilisation and marginalisation when faced to the new entrants in the field of social analysis, i.e. digital platforms.

Notes

[1] Dominique Cardon has proposed a typology consisting of links, clicks, likes and traces (Cardon, 2013).
[2] The works of Osborne and Rose (1999) and Loïc Blondiaux (1998) develop this story extensively.
[4] The coupling of operational/academic consisted rather of Gallup-Cantrill on one side and Roper-Lazarsfeld on the other, but history has remembered mainly Gallup and Lazarfeld. See Blondiaux (1998) on this topic.
[5] Pp. 222-235. Stating his thoughts at the end of the article, he writes: 'Public opinion in this sense, implicitly admitted by those who carry out opinion polls or those who use the results, I'm just saying that *this* opinion does not exist'. Being the champion of social structure and its reproduction in long-term trends, Bourdieu could not allow room for other entities such as public opinion, which is more cyclical, to exist independently.

References

Anderson, B. (1991). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*.London and New York: Verso.
Anderson, C. (2008, June). The End of Theory: The Data Deluge Makes the Scientific Method

Obsolete. *Wired Magazine*.
Blondiaux, L. (1998). *La fabrique de l'opinion. Une histoire sociale des sondages*. Paris: Le Seuil.
Boullier, D. & Lohard A. (2012). *Opinion mining et sentiment analysis. Méthodes et outils*. Paris : Open Editions Press.
Boullier, D. (2004). *La télévision telle qu'on la parle. Trois études ethnométhodologiques*. Paris : L'Harmattan.
Boullier, D. (2016). *Sociologie du numérique*. Paris : Armand Colin.
Bourdieu P. (1984). Public Opinion Does Not Exist. In Mattelart A. and Siegelaub, S. (Eds.) *Communication and Class Struggle*. New York: International General.
Bowker, G. & Leigh Star S. (1999). *Sorting Things Out: Classification and its Consequences*. Cambridge: MIT Press.
Bowker, G. (2014). The Theory/Data Thing. Commentary. *International Journal of Communication* 8, 1795–1799.
Bowley, A.L. (1906). Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science. *Journal of the Royal Statistical Society*, *69* (3), 540-558.
Boyd, D., & Crawford K. (2011). Six Provocations for Big Data. Paper presented at *A Decade In Internet Time: Symposium On The Dynamics Of The Internet And Society*, Oxford Institute, UK, University of Oxford.
Braudel, F. (1996). *The Mediterranean and the Mediterranean World in the Age of Philip II*, Berkeley: University of California Press.
Bruno, I. & Didier, E. (2013). *Benchmarking. L'état sous pression statistique*. Paris: La Découverte.
Burrows, R. & Savage, M. After the crisis? Big Data and the methodological challenges of empirical sociology, *Big Data & Society*. Jun 2014, 1 (1), 1-6.
Cardon, D. (2013) Du lien au like sur Internet. Deux mesures de la réputation. *Communications*, 93, 173-186.
Cochoy, F. (1999). *Une histoire du marketing. Discipliner l'économie de marché*. Paris: La Découverte.
Desrosières, A. (1998). *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge Massachusetts: Harvard University Press.
Desrosières, A. (2008). *Gouverner par les nombres : L'argument statistique I*. Paris: Presses de l'école des Mines.
Desrosières, A. (2014). *Prouver et gouverner : une analyse politique des statistiques publiques*. Paris: La Découverte.
Didier, E. (2002). "Sampling and Democracy: Representativeness in the First United States Surveys", *Science in Context*, 15 (3), 427-445.
Didier, E. (2007). Do Statistics 'Perform' the Economy ? in *Do Economists Make Markets? On the Performativity of Economics*, MacKenzie, D. et al. (Eds.), Princeton University Press, 276-310.
Didier, E. (2009). *En quoi consiste l'Amérique? Les statistiques, Le New Deal et la démocratie*. Paris: La Découverte.
Douglas, J.D. (1967). *The Social Meanings of Suicide*. Princeton, NJ: Princeton University Press.
Driscoll, K. & Walker; S. (2014). Working within a Black Box: Transparency in the Collection and Production of Big Twitter Data. *International Journal of Communication 8*, 1745–1764.
Durkheim, E. (1893). *The division of labour in society*, Paris, Alcan.
Durkheim, E. (1897). *Le suicide*. Paris, Alcan.
Durkheim, E. (1912). *Les formes élémentaires de la vie religieuse*. Paris: Alcan.
Eymard-Duvernay F., Favereau O., Orlean A., Salais R. & Thévenot L (2004). L'économie des conventions ou le temps de la réunification dans les sciences sociales. *Problèmes Économiques*, 2838.
Foucault, M. (1982). *The Archaeology of Knowledge and the Discourse on Language*. New York: Vintage Books.
Gallup, G. (1939). *Public Opinion in a Democracy*. Stafford: Herbert L. Baker Foundation Little Lectures.
Katz, E. & Lazarsfeld, P. (1955). *Personal Influence: The Part Played by the People in the Flow of Mass Communication*. Glencoe: Free Press.
Kleinberg J., D. Gibson & P. Raghavan (1998). Inferring Web Communities from Link Topology. In *Proc. Of The 9th ACM Conference on Hypertext and Hypermedia (Hyper-98)*, (225-234), New York, June 20-24.
Latour, B (1987). *Science in Action, How to Follow Scientists and Engineers through Society.* Cambridge (Mass.): Harvard University Press.
Latour, B. (2005). *Reassembling the Social - An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.
Latour, B., Jensen B., Venturini T., Grauwin S. & Boullier D. (2012). The Whole Is Always Smaller Than Its Parts. A Digital Test Of Gabriel Tarde's Monads. *British Journal Of Sociology 63* (4), 590–615.
Marres, N. (2017). *Digital Sociology*. Cambridge: Polity Press.

Mauss, M. (1950). *Essai sur le don. Sociologie et Anthropologie*, Paris: Puf.
Mayer, N. & Tiberj V. (2016). Who were the « Charlie » in the Streets? A Socio-Political Approach of the January 11 Rallies. *International Review of Social Psychology*, 29 (1), 59-68.
Mc Luhan, M. (1964). *Understanding Media: The Extension of Man*, London: Routledge.
Osborne, T. & Rose, N. (1999). Do the social sciences create phenomena? The example of public opinion research. *British Journal of Sociology 50*(3), 367-396.
Savage, M. & Burrows R. (2007). The coming crisis of empirical sociology. *Sociology 41*(5): 885–899.
Sperber, D. (1996). *La Contagion des Idées*, Paris : Odile Jacob.
Tarde, G. (1890). *Les lois de l'imitation*. Paris: Alcan.
Tarde, G. (1901). *L'opinion et la foule.* Paris : Alcan.
Thévenot, L. (1986). Les investissements de forme. In Thevenot L. (Ed), Conventions Économiques 21-71, Paris: Cee-PUF.
Todd, E., (2015). *Qui est Charlie ? Sociologie d'une crise religieuse*. Paris : Le Seuil.
Tönnies, F. (1887). Gemeinschaft und Gesellschaft, Leipzig: Fues's Verlag.
Watts, D.J.& P.S. Dodds (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, vol. 34, no. 4, 441–458.

Dominique Boullier was trained as a sociologist and a linguist in Paris, EHESS, in 1987, and since 1981, his research has focused on information technologies. He created and managed a company, two testbeds (user labs) and 4 academic labs, including the médialab at Sciences Po Paris with Bruno Latour (2008-2015). In 2015, he joined the Digital Humanities Institute at the Ecole Polytechnique Fédérale de Lausanne (Switzerland).
dominique.boullier@epfl.ch