Big data challenge for social sciences and market research: from society and opinion to replications

By Dominique Boullier

EPFL Social Media Lab

(2017), "Big data challenge for social sciences and market research: from society and opinion to replications", in Cochoy, F., Hagberg, J., Petersson McIntyre, M. and Sörum, N. (eds.), *Digitalizing Consumption, Tracing How Devices Shape Consumer Culture,* London and New York, Routledge.

Translated by Jim O'Hagan

Big data challenge for social sciences: from society and opinion to vibrations

Abstract

Big Data dealing with the social is taken into account by agencies specializing in the processing of this data to produce potentially predictive correlations primarily for the benefit of brands. The risk is that the Social Sciences would be lastingly disqualified from producing reflexivity that has hitherto been their raison d'être. Beyond "society" and "opinion" for which the text lays out a genealogy, appear the "traces" that must be theorized as "vibrations" by the Social Sciences in order to reap the benefits of the uncertain status of entities' widespread traceability. The third generation of Social Sciences currently emerging must assume the specific nature of the world of data created by digital networks, without falling back on the Sciences of "Society" or "opinion", that were built in the XXe century and produced largely shared conventions.

A new generation of Social Sciences is knocking at the door. To put it simply, marketing and "computer sciences" take ownership and generate monitoring tools for social life, in the form of monitoring of brands, reputations, communities, social networks, etc. that can do without the interpretations and models of the Social Sciences because they compensate with computing power and the unprecedented traceability of Big Data. Their main concern remains the action and reaction, not the analysis or understanding as the traditions of Sociology and other Social Sciences had defined. Traces rather than data, reactivity rather than reflexivity, the digital world finds itself shaped by principles that leave less and less room for Social Sciences. First, we shall discuss the peculiar status of these traces to understand the justifications and limitations of such a hype. Then, we shall compare our times of tremendous technical change to the ones of the 30's where sampling was invented and opinion made visible in the same socio-technical and marketing move. Finally, another time will serve in the comparison, when "Society" began to be thought as an entity of its own and computed with novel devices at the same time. Finally, we shall discuss how social sciences and market research can repurpose these digital sources for their own goals while brands are using them for their reactivity requirements. The three Ages of Social Sciences are not a mere question for quantification sociologists, they frame our collective reflexivity and the table of their features will give a striking evidence of the correspondences between these three ages.

1.1 The Digital Age

Neither people nor identities, traces are the raw material

For many years, but in a way extended with social networks, "Computer Sciences" calculate and model the social as if the traces collected allowed access to the "truth" about individuals in a more effective way than all 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 declared Sir Tim Berners-Lee, founder of the web in 1991 with René Caillau, wishing to emphasize the transition to a dimension of networks which is neither technological (Internet) nor documentary (WWW), but social (GGG for Global Giant Graph).
- Facebook for its part has managed the tour de force of "normalising" in terms of the actors themselves, the declaration of their true identity, that is to say those provided by the civil state, the name and surname, in opposition to the tradition of anonymity on the web. The platform thus claims to become the world of reference or even a civil-status-alternative, competing with Google in this regard.

  However, 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 that could possibly take on the form of civil status. This uncertainty should lead to a careful examination of results in order to differenciate registered accounts and active or engaging accounts and prevent anyone from allegations about "society" or "people" when using social networks analytics. For the scores that classify sites on a search engine such as Google, the resulting topology of sites and blogs never discusses their contents as such, but the inbound and outbound links that produce a rank of authority or hub, as defined in the network topology (Kleinberg et al, 1998) and not a civil status. It should be noted here at the outset

what we mean by traces in order to distinguish them from data. Traces can range from signals ("raw" 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, they fall outside relational databases. 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 fact 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. 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 factors which are rarely mobilized in correlations sought 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.

Traces are produced by platforms

Amazon or Apple do not focus on the same features as do Facebook and Google (since the web is no longer distributed but monopolized by these four GAFA platforms that centralize the majority of traffic, with Twitter extending this traces industry). It is not people who are put into

[1] D. Cardon (2013) has proposed a typology consisting of links, clicks, likes and traces.

relation with each other but, above all, tastes (books or music originally), expressed by traces of purchases, 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 resold. The ensuing marketing methods are largely based on the addressing of mass advertising or mail to IP addresses or mails that have clicked on an article (retargeting) but much more rarely via sophisticated links with other attributes of the supposed people attached to these addresses or those clicks (profiling).

Traces of digital behaviour are thus a particularly profitable “raw material”, without the need to appeal to the Social Sciences. How should the Social Sciences deal with this situation? Two options are open to them: either they are confined to their world of administrative data, surveys and polls, relativizing the interest of such traces and focusing on data; or they decide to take the bull by the horns and take these traces as raw material provided for “repurposing” it as proposed by Richard Rogers (2013). So they must accept being dependant on the platforms that produce these traces, without being able to hold any weight over their formatting or even being totally dependant on the conditions of production for such data, which may vary over time and across platforms. The powerful viral phenomenon specific to the Facebook platform and its ‘likes’ mechanism nevertheless, does not leave researchers indifferent, since they are so spectacular, as in pages generating tons of likes in a few days. This arouses all types of analyses from the most constructivist and critical (‘all the likes are bought’ i.e. artefacts) to the most realistic (‘it is solid proof that opinion, and even “the people” think that way’). 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 we find in boyd and Crawford in relation to 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). Other studies (Driscoll and Walker, 2014) tested the data produced from various access methods offered by e.g. Twitter, and showed that the Search API, the Streaming API and Gnip Power Track (paid service) provide very different results, the latter method collecting a much larger number of tweets, but not uniformly according to the requests! This means that the traces collected are entirely dependent on the mechanisms of collection, which is not surprising but which we do tend to forget for other, older methods that have become conventional.

The brands’ grip on traces

Where does this fascination with traces despite their limitations come from, compared with data from registries and surveys? The traces are actually a key resource for brands to monitor the effects of their actions on the public. Reputation and notorietyno longer translate audience measurement, that would be a simplistic import of measures lengthy built for the mass media. On networks, one must measure both a form of audience (the reach), 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 all activities, whether commercial, cultural, political, institutional or even interpersonal when each must measure his excellence with rankings, as researchers are requested to do (Bruno and Didier, 2013). Thus, it is the brands' methods which 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, 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 closer the relationship with the devices that monitor the activity and the offers made to the customer such as in CRM systems, the more efficient the reactivity. Algorithms that decide the prices of ads placement do not depend anymore on negotiations or decisions made by experts in pricing and market segmentation but only on the previous amount of traces collected by the systems (clicks for instance) and assembled in correlations that generate automated decisions. These methods and the calculation devices that have been built for market purposes were imported from the financial market where reactivity is the key factor, to the point of High Frequency Trading, where expectations and moves on the markets can be manipulated automatically at a millisecond pace. 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*(Boullier, 2013)

We 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 not new. Two other key moments in the existence of the Social Sciences, especially Sociology, Market Studies and Political Science, must be paralleled on the same basis to understand the scope of the changes that are underway. These two main periods of quantification for societies (Desrosières, 1998) must teach us how new methods and principles can be arranged in such a way that they transform themselves "into socio-technical conventions". The emergence of Big Data can be as challenging for Social Sciences as were sampling methods in the 30's for instance. Its ability to become a shared method to quantify social phenomena may produce the same lasting effects.

1.2 The construction of 'opinion'[2]

The contemporary situation is undoubtedly not that far from a key moment in the history of the Social Sciences, especially when dealing with consumption behavior issues, that would help us to understand what is happening: this is why we shall remind some features of the period to guess what are the equivalent in our times. If we gave the current era of digital traces the label "3G" for third generation, it would then have to give the emergence of public opinion in the late 30s the label "2G". Indeed, in 1936, George Gallup was able to predict the election of Roosevelt over Landon with a study of 50,000 people. So, founded on this dramatic gesture was the reliability of the survey and of investigative sampling methods[3], which certainly sacrificed the exhaustiveness of inquiries on entire populations but managed to produce correct results provided that the terms of *representativeness* were respected. The issues of smapling

[2] The works of Loïc Blondiaux (1998) and Jean Converse (1987) develop this story extensively.

[3] Even though Kaier tested them in 1891 and Bwoley established the principles of the probable error in 1912.

were addressed previously by the social survey of Rowntree (1901) studying urban poverty in York and using the Booth's "poverlty line as a statistical marker". However, as Converse (1987) told the story, it is certainly in the context of the mass media that their importance was recognized and their method systematized. 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. (Converse, XXX) Considerable media transformation, and the mass media (radio at the time), has established the conditions for the emergence and validation of a survey technique, which thus opens up a whole new era, for Marketing and Political Science at the same time. Moreover, it is "public opinion" itself which takes on a measurable existence with these sampling methods.

Communication agencies such as polling organizations indeed cannot live solely from their campaign activities even if they do bring them high visibility and notoriety. From the outset, their target is constituted by the mass media, as we said, for one essential reason: audience measurement becomes key to the distribution of advertising space, since the dawn of radio and then later with television (in 1941 the first ads are aired on American television for Bulova watches, during a baseball game). But these measures also allow us to monitor the effects of these campaigns on the minds of consumers, giving an unprecedented boost to marketing that drives increasingly sophisticated communication strategies (Cochoy, 1999). Agencies that provide the main reliable feedback on audiences are used to design the programs and the ads as well, targeted at the same populations and generating revenues from companies. Brands are thus present from the beginning in methods of inquiry into opinion via sampling from the moment when such investigations were aimed primarily at mass-media audiences.

Public opinion exists, I measured it!

The work done by Gallup for the operational side and Lazarsfeld for the scientific side is therefore not a simple marketing operation or a face lift for the social sciences: it provides

whole societies with methods with which to auto-analyse, to represent themselves- as opinions. Tarde (1901) has certainly highlighted the importance of these views; 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 assembly to hold on. 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 with the aim 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 comprehensiveness of the studied population. Such successful convention work focuses on the same assemblages of mediations already mentioned for traces:

- 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)

1.3 The fabrication of 'society'

This historical reference to opinion might seem too close to the digital networked world because of the involvement of the media and brands. Therefore, the world of traces produced on the web may ultimately be restricted to a permanent extension of the domain of brands and

other metrics. Yet, to us it seems that another historic moment for the Social Sciences would allow us to complicate the panorama and perceive it in the long term. Let us effectively pretend here that Durkheim succeeded in an operation identical to that of Gallup and Lazarsfeld when they invented "public opinion" because he managed to make "Society" exist. If the conventional nature of the concept of opinion may still be admitted, its stands that evidence about society does not bear discussion, not only for academia but for the layman and for the experts of markets as well. Especially since the term did not begin with Durkheim, although its history is not so long. 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. With "The Suicide" (1897), the method was set up to extend the discussion of the types that would reveal anomia to be a problematic situation. But reliance on data records produced by states, from their various components (ministries, prefectures, governments) becomes key to the demonstration. It is these aggregates that are explained or explanatory, using a method of comparison between countries, regions, counties or districts where possible and necessary. The method depends entirely on the available data and cannot afford to criticize or to question the procedures for the production of this data, despite the countless limitations identified upon publication. By organizing all his systems of proof around these national administrative statistics, Durkheim finds a quantitative analogue for his conceptual choice that puts "Society" in a separate status from all manifestations and individual behaviour. Durkheim's *whole* becomes an entity of the second degree, "Society", (Latour, 2005), while the censuses and other state-data-registers simply conduct the task of recovering individual, administrative events (marital status, judicial procedures, etc.), formatted in identical categories and aggregated to reveal the behaviour of populations. All Durkheim's force of conviction would have been to make these statistics exist as equivalent to his "society".

The statistical apparatus makes society visible in the same way that the survey makes opinion visible, and, regardless of the statistical validity, the framing that operates here gains power. It is indeed necessary to notice that a form of "objective alliance" was formed between data producers from the state administrations and the emerging social sciences. Together they would produce the entity “society” as the object to be tracked by the state for reasons of government and to be explained for scientific reasons. The result is the widely shared and obvious fact, ‘society’ exists and the methods that allow it to do so have no grounds to be questioned because they demonstrate both their scientific and operational value, they are “tools of proof” and “tools of government” as Desrosières (2014) put it. These processes and alliances look absolutely identical to those we encounter between the media and the polling organisations who get on well with one another in order to fabricate opinion and make it seem natural, *taken for granted*, after a long work of implementing conventions. Technical devices should be considered as parts of the assemblage for these conventions. In 1890, Hollerith used his machine to conduct the American census because the Census Bureau had failed to finish processing the previous census dating back to 1880 when it had had to start the next. Hollerith’s company would later be transformed into IBM by Watson in 1926 and would spread among all countries administrations. This specific feature of new devices that are enrolled in this new quantification era is exactly what is at stake in the digital revolution, as well as it was at the times of phone lines and mass media for opinion. This socio-technical setting is well known in Sciences and Technology Studies but become very useful in times of on-going innovations that generate so much disorientation. The ability of socio-technical designs to assemble a specific set of features, to make them last long after the innovation breakthrough and to become a “taken-for-granted” part of the “social environment” is what make technologies powerful ta maintain “a sense of the social structure” as Schutz (1962) used to say.

Durkheim’s performance would have been to hold an assemblage of very powerful mediations:

- Censuses
- Assembled and formatted by Public Administrations
- Under guarantee of exhaustiveness
- For States
- With a Government in mind
- To produce "Society" (based on population)
- Using tabulating calculation machines

1.4 What the Social Sciences can do with the digital and what the digital does to the Social Sciences

Replacing digital transformations in the long history of the Social Sciences allows us to better understand contemporary movements in the use of traces. Three positions are possible:

- One that aims to take up the course of the Social Sciences of previous generations and to apply their methods and concepts of “society” and “opinion”" to trace the web;
- One which accepts this new world of traces, immersing itself in its demands and principles by abandoning traditions and scientific requirements,
- The last one that confronts the radical novelty of this socio-technical configuration and which attempts to understand what the Social Sciences’ place might be in the production of new conventions to exploit these traces.

1.4.1 The digitalisation of opinion and Society traditions

The first direction consists of taking up the well-known methods and concepts and applying them to traces collected on the web, or, even more sensibly, exploiting the potential of digital networks to implement exactly the same methods. Thus, surveys via questionnaire and polls are not only computer-aided on platforms equipped for this, but can be conducted entirely on the Internet and sometimes permit the recovery of enough of the respondents' socio-demographic attributes to ensure the sample's representativeness. These surveys or online surveys are ideal for making opinion more responsive and tracked on a more frequent basis. so the digital **amplifies** (Eisenstein, 1991) "the reality of public opinion" and sampling is used in online surveys for market research that make the rationale of consumers tastes, opinions and judgements appear through their formatted individual expressions. Likewise, input modes for censuses could be equipped with computer terminals to speed up and standardize the collection of data, which makes the now visible 'Society' even more reliable. But Web Studies, a by-product of the Social Sciences, implement the same framework on these new media formats: Economic Studies preferably from Google queries, studies of sociability, longitudinal monitoring of "communities" around themes or specific sites, uses of "opinion mining" and "sentiment analysis" (Boullier and Lohard, 2012) methods are implemented to increase the monitoring of public opinion or the identification of trends in consumer behaviors. In this approach, digital traces on social networks or blogs are just one more way to get access to these opinions that are not questioned and not dependant on the platforms where they appear.

1.4.2 The end of social theory?

Another orientation is available, a rather radical one when phrased by Chris Anderson. The editor of the *Wired* magazine raised concerns with his short paper entitled "The End of Theory" in 2008. Extending his provocative statement to social networks analysis for instance,

one can say that the Facebook likes do not need theory: the platform picks up traces of the actions and clicks of Internet-users or machines, in a standardized format, it aggregates them and produces a score that is displayed and can be used by the platform itself to show trends that guide the placements of advertisers who also seek to achieve certain effects and to optimize their investment or communication choices. In a simplified format, this is the string of events that was produced. Social theory has virtually no use in such an operational system where the performative mechanism works almost the same way ashe audience measurement. Some then try to develop a critique showing that the likes aggregate very different sorts of behaviour including even purchased likes. But this hardly concerns operators, platforms or advertisers. Their action / reaction works in the performative mode, where the likes unearth 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, produced a shared agreement and evidence of this has come to forcefully impose itself, every morning in the direction of programs in the mass-media. The platforms of social networks and advertisers have not yet reached a stable compromise, which explains the proliferation of services that claim to be the standard, like Klout for example, and want to become the Nielsen of these measures. But for all these institutes no theory is necessary, if there is high, statistical quality (which is seldom the case!), as any theoretical advance in audiences and their processes would require a renegotiation of the agreements, which remains the only validity criterion of any theory. It is easy to see the difference between these principles and the traditions of the Social Sciences as G. Bowker does 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 with themselves over time. (…)

The unexamined term the “individual” is what structures the database and significantly excludes temporality.” (Bowker, art. Cit.)

Bowker has cause for concern from the point of view of “Society”, but the third generation of the Social Sciences are not so much interested in “society” as they are in other 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 “Society” matrix to make them say what they are not made for saying. 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. The vital statistics which are the reference base for the 3rd generation of the Social Sciences, are no longer the censuses, but totally agnostic traces about the entities that “are behind” because all act almost equivalently and cause the others to act.

1.4.3 The properties of the third generation of Social Sciences

What the third generation of the Social Sciences could be remains to be observed or even imagined if they:

(a) Assume the radically new character of these heterogeneous traces without falling back on the status of traces or symptoms of something “truly” social (“society” or “opinion”, “market” being a hybrid of both),

(b) Do not get caught up in the self-referential production system/ monitoring of traces that dispense with theory because of other aims.

We adopt a radical empiricism approach (James, 1890) in following the digital traces at their face value, looking for what they are (traces produced by platforms) and for how they transform and are transformed by the very milieu they live within, refusing to reduce them to

an equivalent of any other phenomena in society or in opinion. This constraint strongly limits the power of "explanation" of these traces but it complies with other approaches that consider "platforms bias" as a component of the analysis. These high frequency propagation phenomena cannot separate traces (elements) and platforms (milieu) and must account for the distribution of agency which is never settled as an a priori. Following them allows us to compute these imitation processes (that include opposition and invention as Tarde said)[4]. This approach does not contest the legitimacy of other analysis of the social as long term social structures, or as opinion movements of mid-range frequency (fashion and elections have almost the same wavelength) and should help more traditional social sciences to seriously consider the extension of the social to new entities, these vibrations that could not be captured before the digital era. We would like to avoid two traps at the same time: using digital traces for the documentation of the "social-as-usual" (society or opinion) or reducing them to a set of clever trick of methods. Our responsibility for the social sciences to get able to play their full role in the digital world dominated by platforms and brands is to build the conventions for a new layer of social sciences maintaining the requirements of scientific reflexivity.

The affinity of Big Data's quality criteria with the requirements of the Social Sciences is quite striking. They are often summarized with 3Vs: Volume, Variety and Velocity.

Volume and exhaustiveness

The volume in some way, mimics the need for exhaustiveness famous in social sciences, however resulting in a somewhat limited mode, because nobody and nothing can define the boundaries of the universe of data collected. We clearly need to mourn the death of

[4] Tarde Gabriel, *Les lois de l'imitation*, Paris, Alcan, 1890.

exhaustiveness when using web traces but that does not mean dispensing with the laying down of all conventional-frameworks for Social Science's approaches when dealing with digital traces.

Variety and representativeness

The second criterion, variety, is also a form of transcription for the representativeness-requirements that allowed all the Social Sciences to proceed with inquiries and surveys based on sampling. Again, the test is a loose version of representativeness, which assumes that we accept a *sufficient* level of variety. The establishment of a set of sources (sourcing) in studies of the web should then adhere to some criteria, specific to digital methods and to the field of study. Our work on opinion mining has led us to consider that no description of social-society, social-opinion or social-traces can be produced "in general" on digital networks. The Social Sciences must agree to deal solely with "issues" (Marres and Weltewrede, 2013), or on the focal points of attention, or on "oriented and situated engagements" (Hannerz, 1983), for which traces can be kept digitally, traces that are specific to each outcome or each engagement.

Velocity and traceability

The last criterion, velocity, hardly finds a parallel in the first and second generation of the Social Sciences. Indeed, these dynamic processes were neither their *forte* nor their concern. It was essential to seek primarily to represent the positions at a given moment *t*, to show the strength of "society" on the diversity of individual behaviour or to show how public opinion and consumers appraisal is structured beyond singular expressions obtained in surveys. Certainly, through a longitudinal follow-up of the same populations or with reusing the same questionnaire, it is possible to deliver the equivalent of dynamism, but without ever being able to track the mediations that would produce these changes. Velocity seems outside the scope of

conventional approaches. However, a branch of Web Science has also seized upon the issue of velocity in its own way by exploiting the *meme* traces that spread on the web. It is very significant that Kleinberg, the very man who had exported scientometric methods to the study of web topology, methods that were taken-up by Google, has for several years (2002), been interested in the technical development of a "meme tracker" with Leskovec (Leskovec and al., 2009). Their most famous study looked at the propagation of citations throughout the Obama campaign, which allowed them to achieve a spectacular visualisation of the focus of attention in rapid mounting and descending curves (*streams and cascades*) around certain incidents during the campaign. Their method aggregates all types of traces that can leave these citations, treated as strings of characters for which traces can be found throughout the entire web, and with it produces a metric anchored in time, day to day, even minute to minute now with Twitter (the unit of measurement has become the Tweet per Second). Taking memes into consideration seems promising to us, provided that we also track the transformations-translations of these memes (derived from memetics, Dawkins, 1976, Blackmore, 1999) in different environments. Therefore, it becomes possible to find an equivalent for the velocity of Big Data: *traceability*. It becomes the essential quality criterion for entities that can be studied.

The third generation of Social Sciences will hardly be able to do anything but associate with digital platforms and brands to produce a science of traces which would then be treated as "vibrations", as we propose below. The traces produced are platform-dependent; we can hardly expect to modify them at the source. Nevertheless, it is possible to exploit the traces produced by the platforms by diverting them from the purpose for which they were designed. The rule here is that we do not take any explanation at another level or another world into account, but that we might compare propagation speeds, rhythms, and possible transformations (e.g. contamination of other areas, etc.). The difference should be the ability to see the processes

that had not yet been identified, either because of the limits of pre-digital technology or the targets adopted by previous generations of Social Sciences. As R. Rogers has proposed in his pioneering work, this « repurposing »[5] of traces will suppose a fine tuning of the « query design », on Google or on any API. For him, this should rely on well defined hypothesis and not only follow the opportunity of inference offered by Big Data and machine learning technologies. N. Marres seems more concerned by the critics about the overwhelming dependency of scholars to the platforms that deliver the data. The way she handles socio-political controversies as "issues" has proved very inspiring by designing limits of validity to empirical researches (for instance, no general tracing of tweets- or weibos- or any other traces without delineating the arenas made by « issues »).

1.5 From traces to replications[6]

As we have highlighted, the production of traces is directly dependent on platforms that generate their own analyses. While cooperation with these platforms is required in order to reach conventional methods, it is still critical to produce the theoretical framework that would account for this phenomenon of traces. Let us then talk about "replications". It may help social sciences to accept the shift achieved vis-à-vis the notions of actors, strategies and representations. All these notions have their legitimacy in the context of other Social Sciences but do not allow for these circulating entities' **power to act** (that are the replications' agency), to be taken into account. We cannot say a priori what the size or status of these entities are because it is only the mass corpus investigations that allow us to identify them, when their vibration emerges from the sensors we exploit, certainly from platforms but according to *our*

5 Richard Rogers, *Digital Methods*, Cambridge, MA, MIT Press, 2013.

6 The original term used was 'répliques' which could also be translated as 'aftershocks' or 'tremors'

objectives. The principle of a Sociology of Replications relies on the need to follow the elements in order to detect waves, without knowing how they will join together to make a "whole" of variable geometry. The replication approach allows us to build an infinite combinatory, following extensions, propagations and repetitions, provided we remain focused on "issues" that carry replications. It is then necessary to focus on the moments of emergence and not on the peaks that function as aggregates. The object of this science of replications is surely the agency of replications that spread and end up enveloping us. Because people are actually traversed by ideas and ideas make us act not the inverse as Tarde clearly indicated (1893). "The imitation rays first and then the beings, whose existence we infer from the transformation they undergo with the flow of imitation" (Latour, 2011). It is then possible to study the properties of these replications, to potentially compare their chances of survival or contamination. This is made possible by differences in their properties, which are always directly related to the "issues" that they carry with them. We started this work in 1987 with the monitoring of TV conversations and their transposition onto workplaces to make "local public opinion". In two different research projects, we monitored the attributes of a photo from the Flickr database in the same way and to track the propagationof cultural signs (songs, flags, landscapes, ..) in a corpus of web sites linked to a region. In the first case, the attributes of the photo (e.g. crossed arms),became tag attractors and thus connected accounts or photos that would have no chance to get connected according to the traditional criteria of social explanatory variables. Tags or icons are replications that can be followed, even if they have neither the explicit character of verbatim or expressions as in the meme tracker nor their massiveness.

Potentially, all the traces that we have identified (such as likes, tweets, recommendations, etc.) may be the object of monitoring: however they require specific, tracking-tools that exist largely for Twitter only. But a detailed review of these tools should bedone to ensure that they meet

the specifications of a traceability of replications (not just traces for the sake of it or for the reactivity of brands).

Some approaches from these digitized corpora (not native to the digital) can give an idea of the potential of such methods. Work done on the n-grams studied from Google Books (Michel et al., 2011) showed the evolution of the English language (the preterit of irregular verbs). Lev Manovich (2012) created a base of over a million examples of manga to compare their most basic attributes such as the contrast and produce a unique insight into influences between trends. He used similar tools to conduct cultural comparisons between countries from millions of photos on Instagram or from the Maidan square in Kiev. The story of "JeSuisCharlie" as hashtag and a logo all over the social networks and media should be considered as the demonstration of the agency of an entity that is not related to a strategy, an intention, and not accounted for by simple society causations. By focusing on those entities that propagate, we change the distribution of agency and we pretend to account for the specific role played by the very features of these messages in the waving of the network. Some propagation patterns can be detected along with the social features of the nodes (e.g. Twitter accounts that are famous do not retweet but are retwitted) but the program to detect the agency of semiotic features of the messages (e.g. the tweet or the hashtag) has still to be built. This would give the opportunity of extending the agency apart from structures (the "society") and from individuals ("opinion" extracted from the mind of individuals) to actants such as messages, that frame some issues. This is why a theory of replications (or vibrations) follows some principles of ANT (Callon, Latour, 1981), by exploring this distribution of agency, thanks to the traceability of these elementary parts that produce the network and perform it and not only follow social paths already well known by social scientists. The role played by the founding fathers of ANT (Callon, Latour and Law) can be considered as two-folds: the first one is quasi technical through scientometrics, because by describing how scientific facts are produced in this web of

citations, they paved the way for the topological principles of the web as we told previously and also for the whole set of methods of traceability. The second one is of a more philosophical kind because by "following the actants" they give the opportunity to repopulate the description of how social worlds are produced and maintained (including non-humans and messages, as well as clicks, likes and so on) and to escape from the strategist view of the network. When computer scientists and market research enter the field of social networks, they immediately look for "influential" (Rogers, 1963) and treat these nodes as strategist actors. This is a extended version of decision making theories (including rational choices theories and game theory) but it misses the point when trying to account for virality and propagation patterns. The network is always "enacted" and reconfigured along with each issue, and with opportunities. This is why ANT always remain on the side of emergence theories, in order to account for what is not a mere effect of structure nor of rational choices. Tracing the network is only possible by following the actants, be they artefacts and messages, provided that the observer adopt the right techniques to account for their specific agency.The consequences on the way Market Reseach handles these traces can be listed as follows:

- The process under scrutiny, the central issues are not the same: from segmentation of the market (MR1G, marketing research 1st generation) or from trends (MR2G), we move to reputation that is the way brands use digital traces or to replications (MR3G) from the Market Research point of view.
- The entities that are populating this world and being computed are not socio-demographical categories nor market segments (MR1G) nor sociostyles, word of mouth, opinion leaders, influentials, consumer journey anymore as in MR2G approach. MR3G focuses on brands and communities when framed by companies and brands or on viral content features when framed by social sciences. This is an important move, because the agency is not distributed to the same kind of entities. It means that

segments or influentials do not make sense anymore from this point of view. It means that another set of entities emerge as candidates that had fallen into oblivion before the digital networks shed the light on them. And this is something content designers are eager to buy, even though they will get back to the omnipotent view of this viral approach, something Jenkins and al. (2013) criticized in favor of their « spreadability », with some reasons.

- The data collection devices are transformed from the CRM where all targets can be assembled and traced down (MR1G) and from polls and focus groups (MR2G) to social listening platforms and community management processes for the brands and to some kind of meme tracker for the social sciences (MR3G).
- The methods are intrisically different from the « targeting » borrowing on balistics (MR1G) and from the influence processes (MR2G) to propagation patterns, for brands as well as for social sciences (MR3G)
- The wavelengths that are investigated are different and do not compete between each other : the long waves of structural social features (MR1G), the mid length cycles of market trends (MR2G) and the high frequency waves of memes propagation (MR3G), typical of an emergence approach.

  Any market research strategy should be aware of the stratey it adopts among these 3 generations and should not believe that it accounts for all the richness of social life not that is should be easily combined. We still need to build the conventions to make replications analysis as reliable as censuses and opinion polls.

A table can summarize these features more clearly.

Table 1: Market research generations

	1st generation	2nd generation	3rd generation (from brands)	3rd generation (from social sciences)
Issues	**Segmentation**	**Trends**	**Reputation**	**Replications**
Entities	Socio-demographic categories + market segments (e.g. DINK)	Sociostyles + word of mouth+ opinion leaders + influentials+ consumer journey	Brands + communities	Memes + viral content features
Collection devices	CRM	Polls + focus groups	Social listening + community management	Meme tracker
Methods	**Targeting (balistics)**	**Influence**	**Propagation**	**Propagation**
Wave lengths	Long waves (structure)	Cycles (market)	High frequency waves (emergence)	High frequency waves (emergence)

Conclusion

A table summarizing the three ages of the Social Sciences allows the consistency of this approach to be made visible, yet at the same time, demands simplification and the elimination of the specificities of each age. Remember however that we are not dealing here with the so-called “qualitative” aspects of the methods of Social Sciences.

Table 2: The three generations of the Social Sciences

	1st generation	**2nd generation**	**3rd generation**
Concept of the social	**Society/(ies)**	**Opinion(s)**	**Replications(s)**
Collection devices	Censuses	Surveys/polls	Platforms/ Big Data
Validation principle	Exhaustiveness	Representativeness	Traceability
Co-construction institutions/ research	Registers/ inquiries	Audience/ Polls	Traces/ Repurposed digital methods
Major players of reference (and financiers)	States	Mass media	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 (population metrics)	Propaganda and media-influence (audience metrics)	Science and technology (scientometrics)
Technical conditions	Hollerith's machine (tabulating calculation)	Radio and telephone	Internet, the web and Big Data
Semiotic formats	Crosstabs and topographic maps	Curve and bar charts / pie charts	Graphs, timelines and dashboards
Metrics	Statistics (classic)	Sampling	Machine Learning
Technical criteria for data quality	Relevance, accuracy, timeliness, accessibility, comparability, coherence	Confidence intervals Probabilities	Volume, Variety and Velocity (Big Data)
The Social Science's Dominant modalities	Explanations	Descriptive and predictive correlations	Predictive correlations

Brands may benefit fromlearning to react using these metrics based on traces. Social Sciences of "society" and "opinion" may also benefit in further developing their approaches by using these sources. In this sense, we plead to make these approaches coexist, to learn to change points of view and to admit the conditions of possibility for each generation, relying on the States, media and brands. Each specific study of an issue arising from every-day-experience or raised by prescribers such as brands must lead to a combination of the three generations. Provided that research has a specific framework for these traces that invade our world. There is a new "raw" material that deserves a review of its own, and produces a third layer to the social, measurable according to other principles, and not reducible to "society" or to "opinion". "Society" ended up existing, "opinion" ended up existing, and "replitions" must eventually end up existing in the same way. Time may have come for the "buzz" to be translated in scientific terms and to gain some recognition of its own agency.

References

Blondiaux L. (1998). La fabrique de l'opinion. Une histoire sociale des sondages. Paris: Le Seuil.

Boullier, D. « Plates-formes de réseaux sociaux et répertoires d'action collective » in Najar, S. (ed.), Les réseaux sociaux sur internet à l'heure des transitions démocratiques, Paris, Editions Karthala, 2013, 492 p.

Boullier D. and A. Lohard (2012) Opinion mining et sentiment analysis. Méthodes et outils, Paris, Open Editions Press.

Boullier, D. (2016), Sociologie du numérique, Paris, Armand Colin.

Bowker G. (2014). The Theory/Data Thing. Commentary, International Journal of Communication 8, 1795–1799.

boyd D., & Crawford K. (2011). Six Provocations for Big Data, Paper Presented at the Oxford Internet Institute's A Decade In Internet Time: Symposium On The Dynamics Of The Internet And Society, Oxford, UK, University of Oxford.

Bruno I. and Didier E. (2013). Benchmarking. L'état sous pression statistique. Paris: La Découverte, Coll. Zones.

CALLON, Michel et Bruno LATOUR.- "Unscrewing the Big Leviathan : How Actors Macrostructure Reality and How Sociologists Help Them to Do So."in KNORR, Karin and Aaron V. CICOUREL (eds) : Advances in Social Theory and Methodology : toward an Integration of Micro and Macro Sociologies, London : Routledge and Kegan Paul, 1981.

Cardon D. (2013) Du Lien Au Like Sur Internet. Deux mesures de la reputation. Communications, 93, La réputation, (pp. 173-186).

Cochoy F.. (1999). Une histoire du marketing. Discipliner l'économie de marché. Paris: La Découverte.

Callon M., Law J., Rip A. (1986). Qualitative Scientometrics. In Callon M., Law J., Rip A., (Ed.). Mapping the Dynamics of Science and Technology. (pp.103-123), London: Macmillan,

Converse J. (1987). Survey Research in the United States. Roots and Emergence 1890-1960, University of California Press

Dawkins R.(1976). The Selfish Gene. Oxford: Oxford University Press.

Desrosieres A. (1998). The Politics of Large Numbers: A History of Statistical Reasoning. Cambridge Massachusetts: Harvard University Press.

Desrosieres A. (2014). Prouver et gouverner : une analyse politique des statistiques publiques, (Ed. Emmanuel Didier). Paris: La Découverte

Driscoll K. and S. Walker (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. (1897). Le suicide. Paris: Alcan.

Eisenstein, E. (1983). The Printing Revolution in Early Modern Europe. Cambridge: Cambridge University Press.

Hannerz U. (1983). Exploring the City. NY: Columbia University Press.

James W. (1890). The Principles of Psychology, 2 Vols. Dover Publications

Jenkins H., S. Ford and J. Green (2013) Spreadable Media. Creating Value and Meaning in a Networked Culture, New-York and London, New-York University Press.

Kleinberg, J. , D. Gibson, P. Raghavan, "Inferring Web Communities From Link Topology", In Proc. of the 9th ACM Conference on Hypertext and Hypermedia (HYPER-98), pages 225-234, New York, June 20-24 1998.

Kleinberg, J. (2002). Bursty and Hierarchical Structure in Streams, Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining.
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, Volume 63, Issue 4, (pp. 590–615).

Latour B. (2005). Reassembling the Social - An Introduction to Actor-Network-Theory, Oxford: Oxford University Press.

Latour B. (2011). Gabriel Tarde. La société comme possession. La preuve par l'orchestre. In Debaise D., Philosophie des possessions. Paris: Les Presses Du Réel.

Leskovec J., L. Backstrom, J. Kleinberg (2009). Meme-Tracking and the dynamics of the news cycle. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).

Marres N. and Weltevrede E. (2013). Scraping the Social? Issues in Live Social Research. Journal of Cultural Economy, 6(3), (pp. 313-335)

Rogers, E. M. , Diffusion of Innovations, Free Press, New-York, 1983 (1ère édition : 1963).

Rogers, R. (2013). Digital Methods, Cambridge Ma: Mit Press.

SCHUTZ, Alfred.- The Phenomenology of the Social World, Evanston : Northwestern University Press, 1962.

Tarde G. (1989). L'opinion et la foule, Paris: Puf, (1st Edition 1901).

Tarde, G. (2001). Les lois de l'imitation, Paris: Les Empêcheurs de Penser en Rond (1st Edition : 1895).