Mediascapes journal 17/2021

Finding a socio-semiotic role for Data Science: A review of applications and case studies and some critical reflections*

Daniela Ghidoli**
Università di Torino

Federico Montanari***
Università degli Studi di Modena e Reggio Emilia

The purpose of this paper is to provide some indications and directions for a possible critical approach to big data, albeit in broad terms, taking into account both authors such as Bruno Latour, socio-semiotic research in the wider sense, and scholars engaging Foucault and Deleuze's gaze on the question of the "society of control". At the same time, we review some examples and case studies involving the use of software and applications that analyze large amounts of data to produce representation of cultural trends (for fashion design and marketing). The intention is to start asking (from the point of view of a possible critical socio-semiotics) what this kind of qualitative and quantitative automated research can produce and how so. We try to show what values, rhetorics and narratives may be entailed in the production, use, and manipulation of data, including from a cultural and social perspective.

Keywords: Socio-Semiotics; Big Data; Critical Data Studies; Cultural Trends Analysis; Actor Network Theory

Between Socio-semiotics and Data Science: what dialogue?

The aim of this paper is to provide a tentative overview of some possible ties between social science, socio-semiotics and the field of quantitative and data-oriented research. We understand that this issue and horizon is vast, fluid and in continuous renewal and transformation, full of open questions and problems from theoretical ones to practical applications. Our purpose here is to present a critical gaze on some tendencies, research practices and uses – including some very concrete ones, such as relationships between social sciences on one side and applications in marketing and commercial communication analysis on the other – with specific attention to several case studies that make use of analytical tools. The general intent is also to propose a critical approach, albeit a provisional one, by observing these examples and cases through a lens inspired by a "Latourian" and Actor Network Theory approach as well as from a critical socio-semiotic

* Articolo proposto il 20/06/2021. Articolo accettato il 20/08/2021 This essay has been conceived, written and elaborated together. But the first three paragraphs have been concretely redacted by Federico Montanari, and the last four by Daniela Ghidoli.
** daniela.ghidoli@gmail.com
*** federico.montanari@unimore.it

and discursive point of view. What does it mean to adopt a critical socio-semiotic approach? It bears emphasizing that in this article we use "socio-semiotic" (or, perhaps, better to say, social-semiotic) approach to refer not only to what is normally meant by a specific school of Greimasian semiotics (in the work of authors such as Landowski, i.e., 1989, 1997; cf. also, Marrone, 2001); rather, we intend a wider perspective as proposed by several authors who have dealt with data from the point of view of semiotics applied to the social. And what does it mean in relation to Latour and Actor Network Theory (Latour et al. 2012; Latour 2010, 1992)? Firstly, this view consists in trying to understand what possible actors (with their connections, profiles, links and values) are involved in these tools and participating in big data settings, and what part they play in these settings. These questions will be our point of departure.

We use "critical approach" to mean, firstly, the rejection of an old-style "apocalyptic" concern about data but, at the same time, also rejecting an attitude that considers big data to be "objective entities" deriving from the realm of computing, information and computer sciences. In our view there is also a third risky of problematic stance: a sociological "everything goes" approach in which data, and big data, are taken as a given. Our idea is to think about data as social, collective objects, produced and, at the same time, to be viewed as actors: the producers of further links and transformations. By "critical approach" we do not mean that data are to be a priori judged negatively or in a sort of moralistic way: rather, we should address big data in terms of their social status, treating them as collective operators and actors and taking them into account with a focus on their further possible transformations. Furthermore, another important issue for a critical-socio-semiotic approach is the evaluation of "basic socio-cultural metaphors" and discursive-narrative social chaining in which big data take part. According to Bassett (2015: 549, emphasis added), "Big data promises an end to scarcity of a particular kind, declaring informational abundance. (...) Big Data sets, data analytic methods and data visualization techniques are powerful. (...) However, the potential of engaging with Big Data has largely not been explored by community and cultural organizations operating at local levels." According to Bassett, therefore, big Data should be conceptualized not only as "centralized" (dealing with forms of centred control, etc.) but also as a way of imagining what communities and local subjects could do with those data.

Secondly, again drawing on Bassett, the main question from a more general critical point of view is:

> to understand a re-distribution of expertise, organized through the advent of new form of automation, and to read this in relation to knowledge transformation and the material operations of social power. This engages with what, in Foucauldian terms, could be called Big Data 'procedures', understanding these as forms of governance – where the latter at once refers to the operations through which an order (an ordering) is made (for instance, capturing, finding, aggregating, organizing, interpreting, large data sets) and also to forms of (human) self-governance and self-limitation, what not to know, what not to practice, what not to desire and perhaps also what to forget. (Bassett 2015, ibid.).

As Bassett argues, this vision is closely connected to Foucault's thought and, particularly, Judith Butler and other researchers' (such as Michael Dieter) interpretation of Foucault's thought and its application in the sphere of digital technologies:

> In an article arguing for the virtues of a critical practice as a critique of governance, Foucault (1997) stated that critique should ask '[h]ow not to be governed like that, by that, in the name of those principles, with such and such an objective in mind and by means of such procedures, not like that, not for that, not by them'. Judith Butler (2002) engages with this aspect of Foucault's thinking in What Is Critique? An Essay on Foucault's Virtue, and Michael Dieter has used Butler's account to explore (Morozovian) solutionism, and the prospects for a critical technical practice (see Dieter, 2014; Morozov, 2013), partly in relation to accelerationism and its demands for rapacious, 'large-scale' engagement with the destructive force of technological innovation in the interests of building an 'alternative modernity'. (Bassett, ibid.)

In this passage Bassett likewise builds on Foucault to stress the importance of this approach in considering what in big data "is to be critiqued": the "epistemological claims" of big data and, more generally, of data infrastructures (and their general inadequacy and "wishful obfuscation", particularly in times of pandemic),[1] their "promises" and possible "solutions beyond disputes, because beyond human intervention" (ibid.). This is, in our view, an important point that points to a "non victimistic" and "non a-priori apocalyptic" critique of big data. It also addresses what other authors (see, for instance, Pasquinelli 2015; Manovich 1999, 2012, 2020) have written about big data and data analytics, or data as "a medium" (Manovich, 2020). Pasquinelli provides a similar definition of big data in close connection to the idea of Metadata society and the transformation from a society of control to a new pluri-composed and multi-faced dimension: "Metadata represent the shift to a different and higher dimensional scale in relation to information: they disclose the collective and political nature that is intrinsic to all information" (Pasquinelli 2015: 1). Metadata also describe the abstract coordinates of a new posthuman condition that is the matrix of the "dividuals" who compose the "superject" in the society of control (Deleuze 1992). Why use "dividuals" and "superjects" instead of merely "subjects" or "individuals" in Deleuze's prophetic view? Because today's forms of knowledge and sociability are undergoing a process of re-composition and recombination into unexpected kinds of matter.

Data take part in this activity of constructing, producing links, spaces, shared spatio-temporal features and situations between individuals, objects and different locations. As forecasted by Deleuze and underlined in Pasquinelli (ibid.): "In his famous 'Postscript on the Societies of Control' Deleuze envisions a form of power that is no longer based on the production of individuals but on the modulation of dividuals. Individuals are deconstructed into numeric footprints, or dividuals, that are administrated through 'data banks'. We no longer find ourselves dealing with the mass/individual pair. Individuals have become "dividuals", and masses, samples, data, markets, or 'banks' (Deleuze 1992)." Today, thanks to social and geo-locative media, all these numeric footprints are brought into connection with different spatial situations but also, as will be increasingly common in the

coming years, via the "internet of things". The effect here has to do with technological and social networks improving the capacity to build up complex patterns that can be processed and handled by either researchers (data analytics) or marketing profilers, as well as statual or non-statual agencies (from traffic to health-care control, as in the case of epidemic containment, or information about terrorist attacks, natural disasters, or extremist groups).[2]

It is important here to employ Latour's definition of social actors. Particularly in our web society and inside social media and the internet, according to Latour (2012: 3):

> The more you wish to pinpoint an actor, the more you have to deploy its actor-network. Let's take a simple example. We all have had the experience of preparing a meeting by searching on the web the name of the person we are soon to meet. If for instance we look on the web for the curriculum vitae of a scholar we have never heard of before, we will stumble on a list of items that are at first vague. Let's say that we have been just told that 'Hervé C.' is now 'professor of economics at Paris School of Management'. At the start of the search, it is nothing more than a proper name. Then, we learn that he has a 'PhD from Penn University', 'has written on voting patterns among corporate stake holders', 'has demonstrated a theorem on the irrationality of aggregation', etc. If we go on through the list of attributes, the definition will expand until paradoxically it will narrow down to a more and more particular instance. Very quickly, just as in the kid game of Q and A, we will zero in on one name and one name only, for the unique solution: 'Hervé C.'. Who is this actor? Answer: this network. What was at first a meaningless string of words with no content a mere dot, now possesses a content, an interior, that is, a network summarized by one now fully specified proper name.

At this point there is a further significant question to underline: the possible link and passage between a critical analysis provided by semiotics, and the move to envisage a political and economic role currently performed by big data as a part of more general forms of immaterial labour. Let us take seriously the role of "concrete metaphors" about data (metaphors working as basic concrete and cognitive as well as pragmatic schemas, cf. classic works such as Lakoff and Johnson, i.e., 1999). Today, data as "harvesting and profiling", data "as oil" or "coal" represent one of the key drivers of the contemporary economic system (see, i.e., Manovich 2012; Mayer-Schönberger, Cukier 2013). From the Data Economy to the so-called "sharing", "gig" or "platform" economy, many observers have noted that these economic systems are based on data manipulation: profiling, reputation, databases, and algorithms. Such "visual metaphors" are not simply labels: they represent concrete practices and discursive forms acting in social systems. And this brings us to a further significant point: how should we conceive of the passage between theoretical and critical analysis, about the capacity to think about the way in which quantitative and qualitative methods are currently intertwined? Finally, what possible effects and repercussions – such as whether we are really living in a sort of "big data fallout"[3] – can be detected using this socio-semiotic critical approach?

It is difficult to think of an area of science, industry, commerce, or government that is not involved in the data revolution in some way. Indeed, the growth of data science, including

big data and data analytics, has obliged a wide range of scholarly disciplines and commercial endeavours to work together. The interdisciplinary nature of this revolution means that a substantial collaborative effort is needed to understand its full potential in terms of transformative trends. The focus of data science is not simply the mathematical or statistical analysis of data. In fact, it is possible to conduct data science using semantics, linguistics and semiotics not necessarily related to math. For this reason, some semioticians (see for instance Rastier or Stockinger, 2015) stress a crucial point: the question is the way we deal with digital data, that is, their Volume, Velocity, Variety and the way we make contact with them. For instance, we are witnessing a heavy use of big data in political election campaigns. This trend can be seen particularly from the second Obama campaign to Clinton, and then Trump, albeit with some differences[4] and, obviously, results; such use involves various sources, from commercials, personal contacts, and data profiles of consumers' lifestyles based on extracted information, voting inclinations, and expressed undecided positions and opinions as captured via interviews, social networks, data mining, etc. (and, perhaps, hacking and possible cyber-attacks). For instance, Hillary Clinton took up Obama's campaign techniques but with a huge multiplication of statistical models, making her campaign "the most data-driven campaign in American history" with data extraction that transformed political campaigning from megaphones and slogans to "profiling", "capture" and "extraction".

Semioticians are trying to formulate their approaches and qualitative methodologies alongside the quantitative nature of Machine learning, Database management and Statistics. However, whether the area is commercial data analysis and marketing or political campaigns, analysts may have had "too much faith in what the data told them." At any rate, this possible exchange between semiotic and qualitative researchers and data scientists in the work of organizing extracted raw data may work not only with mathematical logics (such as coordinates, sets, matrices) but also through "semiotic pattern recognition". The first general question in relation to this point is, how is the socio-semiotic statute of data changing in a context in which a sort of "distributed artificial intelligence" prevails, in order to understand how data are coordinated, sifted, and harvested? Secondly, it is important that we interrogate the "new nature" of these quali-quantitative objects which are made-by-data and, at the same time, making the data. Once again, we need to devise a new "symbolic" approach concerning the way in which knowledge is not only represented but "extracted". In other words, the first possible step might be to put socio-semiotic models and theory into dialogue with some other computable representations that enable algorithms to recognize data from another point of view. This probably represent a new way of training machines to acquire knowledge, applying the socio-semiotic method and using their own knowledge "as an expert in the field". The only way to move in this direction is to collaborate with mathematicians, statisticians, data scientists and all the other experts who are able to introduce new data representations, but, at the same time, holding firm to the idea that "data, along with its science and infrastructure, is informed by specific histories, ideologies, and philosophies that tend to remain hidden, although there have been recent calls to investigate these areas" (cf., Iliadis, Russo, 2016, p. 2).

In this context, statistical methods allow semioticians to work on and to deal with forms of accumulated knowledge that need to be "re-layered" or stratified. Framing questions statistically allows them to make use of data resources to extract knowledge and obtain better answers. Statistical inference, having to do with the component of randomness in data, enables semioticians to formulate questions in terms of underlying processes and to quantify uncertainty in their answers. A statistical framework makes it possible to distinguish between causation and correlation and thus identify what interventions will bring about changes in outcomes. It also allows researchers to establish methods for prediction and estimation, to quantify the degree of certainty, and to do all of this by using algorithms that exhibit predictable and reproducible behaviour.

Experts have coined the term "integrated strategy machine" to refer to a system capable of formulating and organizing raw data while also integrating human intelligence in this process:

> An integrated strategy machine is the collection of resources, both technological and human, that act in concert to develop and execute business strategies. It comprises a range of conceptual and analytical operations, including problem definition, signal processing, pattern recognition, abstraction and conceptualization, analysis, and prediction. One of its critical functions is reframing, which is repeatedly redefining the problem to enable deeper insights. Within this machine, people and technology must each play their particular roles in an integrated fashion.[5]

Setting off from an example of "Integrated Strategy Machine" based on semiotics, and some general considerations

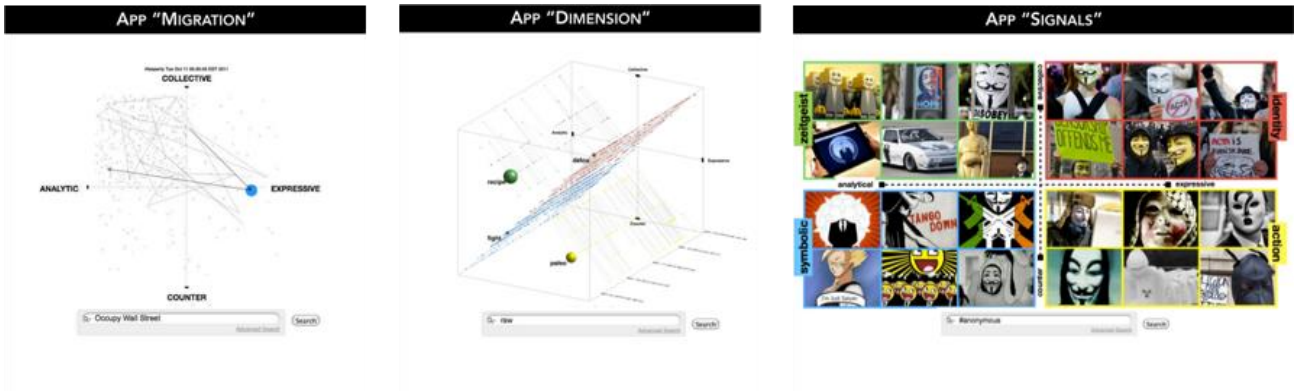
In this second paragraph we would like to present a case study of a qualitative analysis toolkit used to study fashion trends that integrates some semiotic tools and socio-semiotic categories. We seek to illustrate, albeit briefly, both the "internal mechanism" of this type of tool and the more general potential connections with the points outlined above regarding the more general role of data as social actors. The ScenarioDNA toolkit is a software application working to analyse and map cultural changes and trends using a bidirectional approach that combines semantic ontologies with social media data mining. The two engines converge in a specific tool, called "Cultural Mapping", in which it is possible to visualize the data and orient them in order to develop insights. The proprietary method is a semiotic framework that reveals the patterns of cultural signifiers:

> Our intent is to bring structure and efficiency to the elusive area of language. To that end, we are melding human insight with analytics. It's impactful to have machine and human working in consort.[6]

In this case semiotic data refers to the signs and symbols put out into the world, knowingly or not, by human beings. However, we should also remember that the principal object of semiotic research is not insulated symbols or signs, but rather textually based information

and relationships, narratives, and discursive devices that work under the textual surface. For socio-semioticians, the big challenge that has been emerging for the last few decades is how to collect general models of these textual units (cf., O'Halloran, 2013); how to undertake the analysis of textual layerings made up of narrative, discursive (such as spatio-temporal features), thematic and figurative structures. The humanness of visualising data lies in the randomness of the data and the patterns it creates or avoids creating. The purpose of semiotically mapping that data is to see the systems that are evolving and to help reveal tensions that might be otherwise fade into the background. This approach, for instance, is based on these assumptions:

> 1) Networks possess a language and grammar that provides structure to our expressions and how we process what things mean; 2) How culture & ideas migrate depends on the structure of the systems available to us, i.e. social networks, the institutions we belong to, etc.; 3) As our social graph fuels the transformation of culture and how we absorb and process it, the challenge for big data becomes making sense of the flow of the rather unstructured expressions we share across those networks — from status updates, to Tweets to 'Likes.' There is a significant challenge to interpret this data correctly; 4) There is a risk when we only measure the loudest signals within the cultural conversation. We risk missing how meaning and value were formed within that culture to begin with; 5) We need to train our brains a bit differently, to frame data, analyse signals and reveal patterns we can actually act on; 6) We start by evaluating the structure of a culture network; in that structure are the mechanics of what drives meaning; 7) We propose a matrix for cultural mapping to plot smaller networks and their relationships[7].



Figg.1 & 2 – ScenarioDNA tools

Starting from the idea that 'there are things we don't know that we don't know", in this example researchers have created a mapped visualization capable of flushing out and conjuring up all the things there are to know in machine-learning power. The intent is to bring researchers, strategists, engineers, designers and analysts "past the point where we knew all that we could know". Their "Cultural Mapping" tool provides multidimensional views of cultural data extracted by social media or based on independent corpus. Let us recall that mapping is currently one of the powerful resources in humanities and social

sciences. In our case study, the first view is provided by the app “Migration”, a 2D matrix that tracks the migration of culture over time. The second app, “Dimensions”, is a 3D graph that explores the varied meaning of words based on cultural context. The third app, “Signals”, is a map in which visual signals and shared imagery are organized according to modes of expression. Each app reflects a facet of what real thought processes actually look like in the development of meaningful ideas and connections.

This software is constantly mapping patterns and their evolution across genres, following the idea that exploring culture is an inductive process and there is no point in waiting for "briefings" to start looking for evidence. Although traditionally Semiotics has been considered to be based on hypothetical-deductive approach, in this case the inductive process is what connects socio-semiotics to design thinking with its more deductive approach. Whereas design thinking lands on a new concept, semiotic thinking allows researchers to see the cultural system over time and confidently invest in multiple future-oriented scenarios. Semiotic mapping is a powerful way to lend context, and semioticians know that this process entails more than looking for the cultural drivers of the moment because mapping can identify as-yet-unseen cultures, untold stories and information voids. Mapping can make sense of gestural and non-verbal communication; it reveals cultural nuances and the essence of culture, that is, a narrative-oriented essence. For all of these reasons, semiotic mapping allows researchers to forecast the future.

Semiotic mapping in an integrated strategy machine helps prove that objects are not incidental. The objects people use to express themselves are critical to identity. One item out of sync can alter the messaging people want to convey. Semiotic mapping seems to be able to organize knowledge and recognize the difference between cultural objects. After such recognition has been carried out, the algorithms are able to filter the materials, for example through geographic location tracking, to better understand the regional positioning of a brand and localize its messages according to each country and its cultural dimensions. With the machine approach, the corpus of materials becomes statistically significant, and it is possible to compare them synchronously and diachronically. Semiotically oriented qualitative research conducted by a human team would not be able to achieve the same results (without a weighty dose of quantitative integration).

This approach aids in analysing the cultural codes of the various generations over the years on the basis of social media content from all around the world. It could be a strong foundation for comparing different fields, topics and brands, something that is impossible using only human analysis but becomes a real possibility thanks to the speed of algorithms and databases. To return to the points outlined above with Stockinger (ib.), however, it is important to understand that mappings and scenarios such as those produced by ScenarioDNA are once again textual multi-units, corpora produced in collaboration between human observations and interpretation, machine and data-base retrieval, and an activity of montage and assemblage.

The merge with socio-semiotics helps to push analysts beyond our present-day knowledge and render visible the pattern of evolution as it is happening. It helps to offer a point of view that allows researchers to differentiate among categories, research problems

and to give them breathing space. For example, in 2009 ScenarioDNA conducted a project for Dorel/Cannondale bikes. Project researchers started by looking at Cannondale and understanding where the competition was located as well as where its sister brands were positioned:

> Today, we hardly think twice about bike culture, bike lanes and bike commuting. But, just five years ago, most of the talk teetered from Lance Armstrong to recreational biking. We were handed a foot-high stack of survey data from bike retailers to glean insight from. We crossed reference that information with social media postings. Bikes were evolving among people, but not at the retail level. Bikes were being hacked and augmented and used in ways no one was discussing — bike jousting, bike moves, anyone? Our initial cast out for information eventually lead us to urban planning. There, within community studies, we found a pent-up demand for bike commuting. The semiotic symbols were all there, but the language of the category and the codes of bicycles had not yet been set in the minds of marketers. We mapped the archetypes of bike culture that we uncovered as seen in this semiotic map (Stock, Tupot 2015).

The intended strength of ScenarioDNA consists of quantifying semiotic hypotheses with a solid base of quantitative data. It is able to supplement human intuition with data information. With this combination, insights can become active strategies. Moreover, the entire range of typical semiotics questions can be validated with this approach, for example recognizing emerging stories:

> About two years ago, we worked with Wieden + Kennedy to uncover deep insight for NIKE into the shifting perception of sport among youth in Japan. We collected data by tracking social media and conducting infield ethnography with teens in Tokyo, Osaka and Kamakura. Mapping the words, signs and imagery that we saw infield, it quickly became apparent that there were a multitude of sports stories emerging. A critical observation, seen in the map here, demonstrates a breaking away from traditional bukatsu sports into more visceral expressive scenarios. The resulting study helped NIKE visualise where kids were moving and understand what it means to be young and Japanese (Stock, Tupot 2015).

Or identifying information voids:

> Most recently, we had been involved in an education grant looking at higher education curriculum. The exercise prompted us to deploy our Culture Mapping method to visualise university syllabus. We recognized that there had been a lot of talk about education reform, but no one was looking at what already exists. Courses live as siloed as ever, never teaching students how to move out into the real world. We tested our theory by analysing course descriptions from a sample of 1000 free online classes found on Open Culture. When a syllabus is mapped according to our culture mapping, classes consistently fall across four quadrants: theory; technique; outreach; and reporting. These quadrants become education archetypes, reflecting classroom learning and community engagement. Today's students need educational experiences from each quadrant for a viable education. Mapping the Open Culture-listed courses, all from respected institutions, demonstrated that the courses available online are sorely lacking in outreach. No surprise. However, the imbalance also represents the state of most curriculum. When a student's education lacks balance, it is either because they do not know how to structure their choices, or the breadth is not available. If we can map a university's

> syllabus according to the cultural language of its course descriptions, we can see curriculum evolve as a system and begin to reveal its patterns of strength and weakness. An analytics toolkit was the result of this exercise. It visualizes higher education syllabus according to culture (Stock, Tupot 2015).

Using an information system capable to set Semiotics tools and procedures highlights the value of its results by expanding the data comfort zone and rendering the inductive discovery process more in-depth. The focus is less on the ready-made data and more on the inductive approach of the semiotics. The resulting visualizations are able to reveal the drivers and the logics behind them. For this reason, ScenarioDNA is collaborating on visualization with Brooklyn-based creative technologists Cousins & Sears:

> Culture mapping makes the most of human creativity in a world where machines tasked with exploration increasingly self-navigate. It's exciting to think about data being visualized in a subjective framework where its objective representation remains completely intact. Once clients can see the system that exists, it clears and opens their vision. It gives them an irrefutable lens to work with. And we can begin to think beyond their initial limits[8].

Why Semiotics and Data Science can improve each other, and some problems

Semiotics and socio-semiotics seems to work and collaborate with machine learning and data mining approaches for a number of different reasons. First, because semiotics starts by analysing symbols, signs or, more precisely, the structures behind the manifestation of these symbols and signs. A machine could be better suited to catch this information and extract from its correlations and then structures, because a person's thinking inevitably passes through her own assumptions. However, this poses the first crucial semio-critical and political problem. Machines deal with structured data. Not only are these symbols and signs circulated and put out into the world by people but, nowadays, they are also continuously "reworked": they are handled and sifted as part of data harvesting. Indeed, "afterthought" analysis is a laborious effort and job. This point represents the critical problem of the semiotics-data connection. We might say that "people" (i.e., consumers, "prosumers", data users) cannot elaborate this connection by themselves because "people" are always involved in a matrix of meaning and completely ignoring one's own interpretive background is difficult. "People are people" sang Depeche Mode band in the '80s, but in a transformed today's post-public opinion era, in which collective actors' roles are played by different segmented audiences (see, about this point, i.e., Bentivegna, Boccia Artieri, 2019) in their niches and networks of possible prosumers, a machine seems to be able to do it. The "lucidity" and transparency of an algorithm seems capable of bypassing our opinion to move in the direction of a sort of "objectiveness" (and, in some ways, reinventing the positivist myth of the absolute objectivity in science). This could be useful because it gives other scientific communities (and professional data users, such as marketing experts, as well as marketers and social scientists) an idea of what they are up

against, for instance which of these 'prejudices' are most amenable to change. And yet the problem seems to be not only "what we do with" data but also "what data do with us" in the sense of the transparency effect, filtering, emerging capacities, and so on. Which kinds of social objects are produced by data? According to Latour, when, we, humans, have to do with technologies in connection with social and cultural relationships and interactions, we always deal with forms of "delegation". A form of delegation can be defined in these terms, according to Latour (1992: 154): "I will define this transformation of a major effort into a minor one by the words displacement or translation or delegation or shifting." And here, at the heart of this question, we find the problem of the "black box" effect in semiotics that deals with "data-catching" in its quali-quantitative dimensions. We can trust that some data or analytic material goes in at one end and some conclusions come out at the other end, but what goes on in the middle is a black box, a mystery. People feel more confident in research when they understand how it was carried out. How can anyone be sure that semiotics findings are true and not simply subjective opinion? A machine that parses and organizes materials and results in an objective and replicable way contributes to a form of the objectivity that semioticians need when proposing their insights.

This point also dismantles another barrier semiotics often faces: because semioticians frequently focus on material that was already ready-made before the research began, sometimes the common feeling is that the results of socio-semiotic research do not really amount to news. An expert in that field or corpus may feel that he knows the material like the back of his own hand and wonders what more semiotics can tell him about it. Sometimes this problem is articulated in terms of subjective opinion. Other researchers would like to know what makes semiotics findings reliably true and, therefore, different from a subjective opinion that they or anyone could have formulated. However, semiotics – in particular semiotics and socio-semiotics, which started from the hjelmslevian and then greimasian idea of a "discipline with a scientific vocation" – is always a formal activity with a distinct set of tools and research procedure. These tools have been more or less codified, as far as structural semiotics is concerned, starting in particular in the work of authors such as Floch (1990), Greimas himself (see, also, Fabbri, 2005; Marrone, 2018; Fontanille, 2004, 2015). Some of these tools have been (even if partially) inserted, and used, in the research device and application we are discussing here in this paragraph. Outlined in brief, some of the objects in the semiotic toolkit, include[9]:

- Drawing of course on Saussure, linguistic signs are understood as not merely isolated words or phrases, but as stressed just above and at least after the work of Benveniste, signs are surface representations of what people communicate with each other, representing themselves in discursive activities and operating on the basis of provisional agreements. When a word has had some agreed-on meaning attached to it, it becomes a sign. Semioticians study words and phrases as signs and correlate those with a more complex sign system made up of images, words, sounds, gestures and so on. A word or phrase is more than an expressive form, and semioticians can observe the specific act of selection enacted by the author or

speaker behind the manifestation of such signs. Tone of voice and correlated elements can reveal, for example, the situation of enunciation, specific enunciative and narrative rules, enunciative intentions and agentivity. Some examples of expressions that semioticians are accustomed to analysing are anaphoric, deictic, personal index, exposition index, referentiality, modality, verbal classification, verbal tense and so forth. They correlate all this deeper information with other levels of analysis to understand the mining system (semantic but not only) in all its complexity.

- Visual signs: a complex and significant set of plastic formants, a variable composition of colours, graphic signs, cultural symbols, gestures and contexts. Semioticians study how the different elements are arranged relative to each other and their relative size. They examine what is in the foreground and what is more peripheral, and how space is used in the frame.
- Aural signs: Semioticians investigate regional accents, ambient sound, and music.
- The implied communication situation: in any piece of communication speech, writing, or visual art there is an implied speaker, the person doing the communicating, and an implied recipient, the person being addressed.
- Textual structures: the various technical features of how the text is constructed. These may include elements such as narrative structure, how it is designed so as to tell a story, or a set of rhetorical devices that have been used to give the impression of a truthful report.
- Information structure: As well as looking at the form or structure of a piece of text, it is also possible to look at its content. Researchers can observe what information is presented as new versus taken-for- granted and what is treated as the important point of the information versus what is treated as incidental.
- Genre: a class or category of human communication that has a distinctive style or set of conventions. For example, advertising draws on lots of different genres such as documentary-style realism, drama and fantasy.
- Binary and semantic oppositions and contrast pairs: a noticeable feature of human communication is how much people describe and understand things through reference to how the things in question are different from something else. Think of low-fat, low-tar, and new and improved or tabloid versus broadsheet, strategic versus tactical, and quantitative versus qualitative.
- Communication codes: the sets of unspoken rules and conventions that structure sign systems and link signs to meanings. When yesterday's people interpreted bright gold biscuit packaging as meaning luxury and today, they interpret brown paper wrappings in the same way, they were able to do so by using different sets of codes to make sense of what they saw. Because codes change over time, they are often characterised as being lapsed or residual (out of date), dominant (the main code or codes being used now) or emergent (new, on the way in).

- Enunciation level: the process needed to create a discourse, such as a dialogue or novel or TV commercial, has to organize a set of other indexes to express actantial roles, spatial and time information, and specific enunciative figures.
- Narrative level: at this level Semioticians explore themes, isotopies, narrative actantial roles, narrative scheme, narrative paths, values and narrative structure.

On the basis of these parameters, semioticians can observe different linguistic, enunciative, and narrative styles and highlight behaviours, habits, perceptions, opinions that are trackable by isotopies and monitorable over time.

Studying culture is an integrated approach that starts from codes

Semiotics should be able to examine what happens for consumers in a given cultural context. Although there could also be psychological factors, when lots of people provide the same interpretation, it is reasonable to argue that part of the cause is the fact of their sharing the same cultural resources. Probably the relevant feature of semiotics is that its focus falls on the culture instead of the consumer. When semiotics is able to understand unspoken cultural rules or codes about contemporary communications and define a sense people are giving to what everyone else in talking about, this kind of insight is useful for understanding the market much better.

As signs change over time, it is necessary to continue with semiotic research; signs must be updated based on evolution and differentiated by categories and sectors. Communication has to be based on actual or emerging language, thus keeping it culturally authentic and fresh because signs often change their meaning depending on associations. The communication strategy has to be developed accordingly to up-to-date codes and rules; semiotics helps to define an accurate market strategy and design.

Most qualitative researchers are strategists and advisors trying to integrate findings and techniques from different intellectual disciplines into their existing qualitative practice. The main task is to fill the gap between research and the practical framework, moving from data through to insight[10]. The idea that researchers are in the business of telling stories based on their data has not always found favour with researchers[11]. A coherent narrative is more likely to be acknowledged by the clients and such clients could adopt it to develop products, services and communication. Although narrative may be an old-fashioned concept (or, in recent years, on the contrary, an overly fashionable and inflated idea like that of "storytelling"), the narratives produced by research around data are really important for businesses to be successful and each narrative has to be driven by a method as to how data are interpreted. Researchers with more experience in a market are likely to be able to assign more accurate interpretations to specific market data, but these arguments fail to engage with the wider realities with which the research industry must contend. Research as a strategic tool means integrating qualitative and quantitative approaches in the most powerful and effective possible combination.

Research has adopted different approaches as a model based on a series of different intellectual disciplines and traditions. The idea is to build interdisciplinary insights into general principles that can be applied as a set of guiding principles or rules across a wide range of cases. Some researchers believe that market research should not draw on other disciplines in developing its own techniques; that it should be sufficient unto itself. Most other business service professions would be surprised by this position. For example, if we consider the quantitative data planners use in advertising or the very wide range of techniques and resources claimed by management consultants in the commercial field, it seems that, when one's own body of knowledge is not extensive enough, it is normal to borrow someone else's.

The bridge with the other sciences: further examples of socio-semiotic tools

Semiotics should help us to curb our tendencies to land on singular ideas, solitary icebergs. It helps us be comfortable in exploring riskier territories and investing incrementally, and to propagate ideas that we do not yet understand or see, and to be confidently provocative and impactful. Leveraging culture mapping to frame semiotics provides the critical context that is too often overlooked.

Semiotics seems to be everywhere around us. It is mentioned in the standard textbooks on researching consumer behaviour. It is rarely the only analysis or the main one that clients are interested in, however. More often, semiotics has been treated as a supplement to other form of research such as qualitative groups. The market is used to supplement semiotics with another methodology, such as groups, in order to add value to ordinary qualitative research. For this reason, semiotics is productively used before focus groups to test predictions with consumers and refine the objectives of qualitative research. These hypotheses can then be used to guide and structure the group discussion. In these cases, semiotics can offer a guide and an aid in identifying what to expect.

According to this point we would like to come back to other commercial examples of interdisciplinary machine structure with the combination of different approaches. One is the TNS Semiometrie methodology and its suite of tools (Semiogramm, SemioSelect, SemioDialog). Semiometrie is originally based on J. F. Steiner's method (1985)[12], where 210 terms and words validated in several European Countries (France, Germany, Great Britain and Spain) are associated with their emotional meanings. Semiometrie links a statistical survey with the Semiotic Square scheme, to identify the attitudes, basic settings and ideals of target groups. During the survey, respondents evaluate the 2010 words, such as hero, suffer, present, fire and so on, according to a seven-stage range from "very agreeable" to "very disagreeable" so as to quantify the person's values. The Semiotic Square organises the words/value map in quadrants, along four dimensions: sociality, vitality, individuality and responsibility. The data are then statistically evaluated for the specific product, service or brand. Semiogramm is the company's online tool for surveying

and visual mapping. SemioSelect is the panel tool used to ascertain the correct people to engage for a deeper qualitative analysis based on their answers to the survey (and their direct contacts). SemioDialog is their sub-brand to conduct qualitative analysis based on Semiometrie.

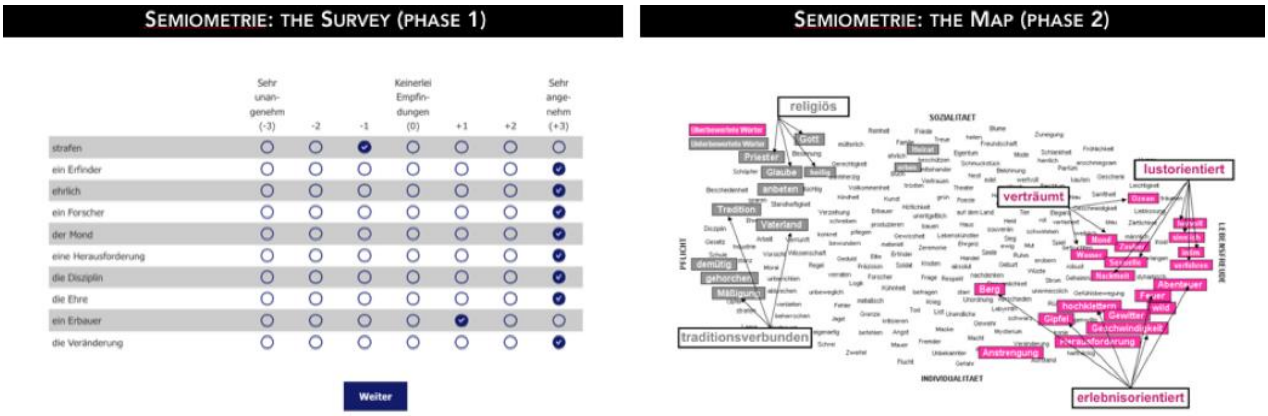


Fig. 3 – TNS Semiometrie

The academic community of socio-semioticians may not find Semiometrie very convincing because of its hybrid appropriation of semiotic terminology without a strong connection to their underlying principles. Indeed, it is more of a semantic map organized according to psychographic correlations. From 1990, however, it has stood as a demonstration that there are no technical limits to introducing a mapping structure using algorithms.

Likely encouraged by this fact, Concept Lab at Bologna University have adopted a methodology based on the concept of Semiometrie to develop a new method of ethnographically analysing and presenting data. This method employs a Digital Interactive Matrix thanks to which users can 'navigate' through the entire data collection extracted from the field.[13] This project's text mining target was to create statistical models of prediction using textual information (exogenous variables) for classifying variables of a qualitative nature (target variable) related to the diaries' textual contents. The idea is to extend the cognitive interpretative model used by qualitative experts to a large quantity of textual material (the rules); this model would be so large that it could be generated on the basis of a representative sample according to the traditional terms of sample statistic theory:

> The research employs a cocktail of in-depth methodologies that combine psychological methods with qualitative fieldwork. The research focused on young people (14–22) and mature adults (55–70) who have been asked to fill in a photo diary for a period of seven days and taking photos of the 'objects of their Happiness' (people, places, products, etc.) which they recognise as meaningful to the building of their day-to-day well-being. Diaries have been followed by in-depth interviews with the respondents on the basis of ad-hoc designed

> discussion guides that took into consideration people's cultural backgrounds, life-stage, and the content of the diary. The result is a collection of 1.200 stories of Happiness reported in people's own words and through visuals, symbols and drawings. This collection of stimuli has been classified within the interactive matrix that combines social forces such as Subjectivity, Sociability, Experimentation, Connectivity, Ethics, combined with Ethno-behavioural variables that we defined as: Domesticity, The Body Affair, Daily Responsibility, Leisure and Consumption, Commuting and Territory, Landscape and Nature, and Extra Occasions.
>
> The matrix of Happiness has been designed as a two-way insight methodology that allows access to information in two ways. On the one hand it is possible to 'enter' the world of the respondents directly, accessing their quotes, photos, history and cultural background. Details on the respondents' profiles have been collected and edited in the project so that at any time it is possible to refer to the protagonist of the experience by clicking on the diary. Each diary has been not only analysed but also reported without altering its original content. The opportunity to 'navigate' the matrix and then narrow down a specific focus of interest is certainly the most innovative factor that characterises this research tool, with respect to traditional research methods. Following a specific research interest, on the other hand, it is also possible to search by age, culture (country of provenience) or by selecting the Happiness Trends that most interest the reader (e.g., the area related to Technology or to the relationship with the home).[14]

In 2005, this experiment formulated a common ethnographic approach through a collaboration involving computer scientists, statisticians, engineers and ethnographers: multidisciplinar methodologies are important in developing this kind of strategic tools.

The activities of Guibourgé and Bonny in their company *Inevidence*[15] are another example of developments moving in this direction. Their methodology is similar to the most well-recognized semiotics approaches in that they combine semiotics and text-mining on the basis of Floch's work, with the use of consumers' semiotic square (1990) up to Fontanille, Zilberberg and tensive models[16]. In 2009, moreover, they focused on the opposition between the Critical and Ludic quadrants, and they were one of the first early adopters of a text-mining method based on SN open discourses (posts, blogs, comments...):

> Although there is a set of potential possibilities and characteristics of positions and trajectories, the value that we are interested in here is the bottom right (recreational or emotional value). We will give some rules of this diagram whose first is that you can't valorise your product both on practical and emotional values. This last is the negation of the practical value. (…) The second rule is that Meaning is generated by differences and all relations (arrows) on this diagram are combinations of these differences. The third rule is that Values are by what we promote our acts but are rarely articulated or directly accessible. (...) Text-mining methods based on Internet open discourses, as proposed here, might be more appropriated for innovation early phases. As such analysis can be launched (and re-launched) rapidly in order to refine the analysis and the related potential innovation leads.[17]

They have implemented the semiotic procedure using a top-down approach:

- Corpus Set-up: semiotics is used to structure the research plan, with filters created to categorize the raw data. After the preparatory desk analysis, the lexical and visual elements which are most impactful for the research objectives are selected and passed on to the text mining system.
- Data Processing: a semiotics-based categorization scheme organizes the corpus. The semiotics recognition patterns then orient the materials and extract the main data points in order to qualify the materials through a set of variables. There are three pattern recognitions: one for sensations, one for emotions, and one for verbs and modalities.
- Tensive-diagram: building statistical occurrence and co-occurrence and generating a tensive-diagram such as Intensity (to measure the emotional density linked to the component) and – Extensity (to measure the number of contents). Based on these elements, a 2D diagram can be drawn.
- Analysis: a human analyst works on the matrix and summarizes the insights.

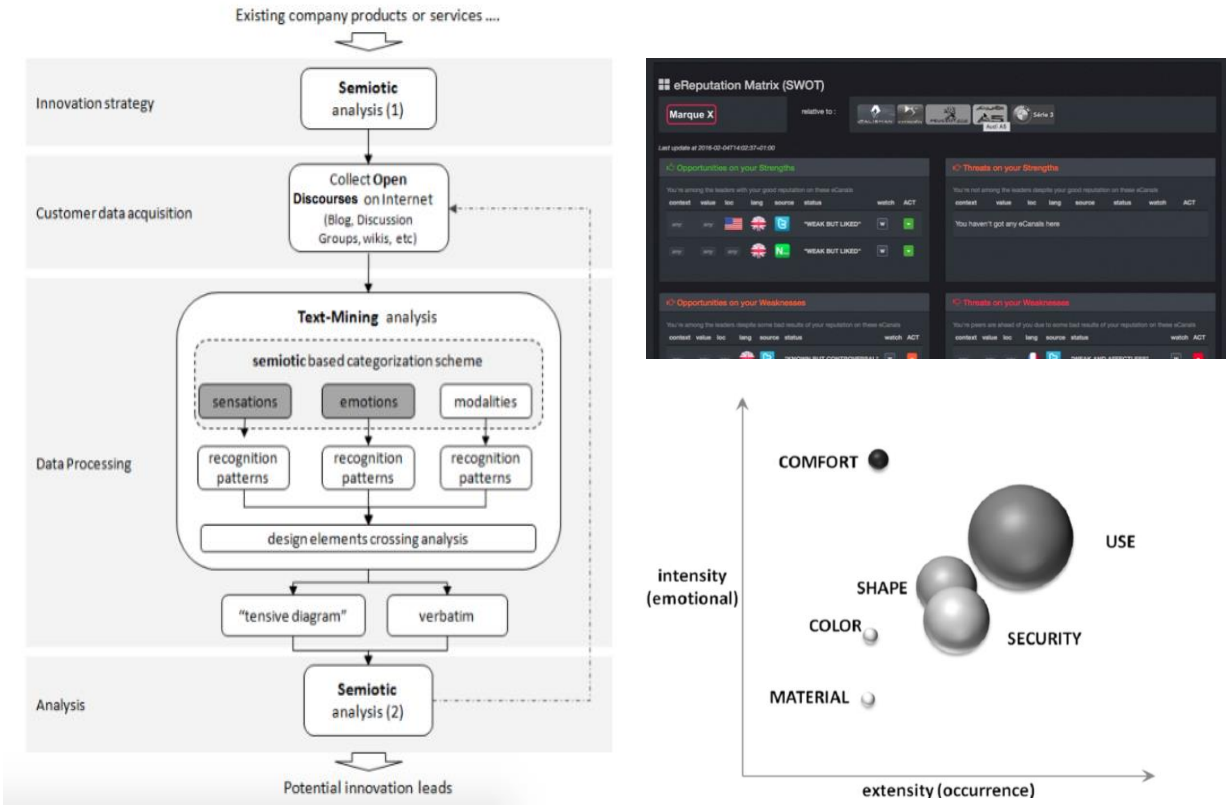


Fig. 4 – Inevidence Semiotic Text Mining approach and tools (tensive diagram and SWOT Matrix)

How Data Science and Semiotics work together

Semiotics is different from traditional qualitative research that normally adopts an inside-out perspective. Whereas interviews and groups are aimed at collecting psychological sensations such as perceptions, attitudes and beliefs from people, socio-semiotics has an outside-in approach. Semiotics may be used in combination with qualitative research, and such uses also demonstrate how semiotics serves as the bridge connecting ethnography to anthropology. Semiology allows researchers to make sense of documents collected via observation and develops the acceptable generalisations required by anthropology. It helps to provide a guide about what to expect. After fieldwork, semiotics may be used alongside the psychological types of qualitative analysis to interpret interview and focus group talk, thereby gleaning the maximum value from qualitative data.

Data science is a field with a very high degree of interaction between researchers: this collaboration with other scientific communities could be important for future decisions concerning political issues as well. All the disciplines have to learn to work together and learn from each other's research approaches, but this engagement must occur at all levels, from individuals to groups of researchers, academic departments, and professionals as a whole. In order to start managing raw data and end up with user-friendly, efficient implementations of principled scientific methods and the communication of substantive results, new strategies are needed, as well as an informed view concerning the risks associated with the use of these data.

Semiotic analysis is a formal activity with a set of tools and a definite procedure, often more clearly specifiable than the qualitative analysis that is conducted following ordinary interviews or focus groups. The socio-semiotic toolkit helps researchers to think in an organised way about what other researchers and data scientists are looking at, and to notice similarities and differences in the data within a category or sector.[18] Semiotics is a particular discipline that provides tools needed to build a solid meta-frame (or mapping) in which the raw data have to interact each other, sharing their information to help the analyst understand the system. In semiotic approach, data must enhance their information as a result of their interaction. Their oppositions offer analysts the interpretation drivers they need to organize the ranking, distribution, and positioning of the data in relationship with the system that they have auto-generated during the analysis phase.

For decades such analysis was carried out using only the human reasoning. With the "arrival" of big data, however, the flood of information subject to human intelligence has definitely increased. With a solid structure and consistent pattern of analysis, the input becomes manageable and thereby see-able. Furthermore, if analysts are overloaded with information and cannot manage it, a structure enables them to tighten the variables and guide how a subject is being examined. With algorithms capable of supporting human effort in these activities, there are more opportunities to manage greater quantities of data

without losing information or possible unexpected results. With a structure based on algorithms, moreover, researchers can better convey their thinking to others. In other words, they can more effectively wring out their brains and share their results in a more objective way.

One machine learning approach that is actually quite similar to socio-semiotic approach and which can contribute to growing this integration with Data Science is the Self-Organizing Map (or SOM). It was introduced by Teuvo Kohonen in 1980:

> The self-organizing map is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map. Self-organizing maps are different from other artificial neural networks as they apply competitive learning as opposed to error-correction learning, and they use a neighbourhood function to preserve the topological properties of the input space. (…) This makes SOMs useful for visualizing low-dimensional views of high-dimensional data, akin to multidimensional scaling. (…) and it operates in two modes: training and mapping. "Training" builds the map using input examples, while "mapping" automatically classifies a new input vector. A self-organizing map consists of components called nodes or neurons. Associated with each node is a weight vector of the same dimension as the input data vectors, and a position in the map space. The usual arrangement of nodes is a two-dimensional regular spacing in a rectangular grid. The procedure for placing a vector from data space onto the map is to find the node with the closest weight vector to the data space vector.[19]

During the 8th International Conference on Advances in Self-organising Maps (2011)[20] participants presented very extensive applications of SOM. For example, Mayer and Rauber introduced a multi-faceted application of SOM in the field of content analysis in which they applied SOM to analysing documents from the internet platform Wikileaks. They choose this document collection because of several aspects:

> For one, documents contained cover a rather large time-span, thus there should also be a quite divergence in the topics discussed. Further, the documents stem from a magnitude of different sources, thus different styles should be expected. Moreover, the documents have very interesting, previously unpublished content. Finally, while the WikiLeaks website provides a way to browse all documents published by certain meta-data categories such as creation year and origin of the cable, there is no way to access the documents by their content. Thus, the SOM offers a valuable alternative mean to provide access to the content of the collection by their content.[21]

They use the Java SOMToolbox framework, a toolbox available online[22]. This framework provides the user with a wealth of analysis and interaction methods, such as different visualisations options, zooming and panning, and automatic labelling of different levels of granularity, to help the user to rapidly acquire an overview of and navigate through the collection.

Zhao X., Li P., Kohonen T. (2010)[23] have introduced software that applies a corpus-based algorithm to derive semantic representations of words with specific examples from two languages (English and Chinese). Their software is open-source and can be used for specific cases. Fig. 4 shows the results of their analysis: a structured semantic

representation in the form of a self-organizing map model after 200 epochs of training on the numerical representations of the 300 most-common English words derived from a small-scale corpus (an edition of Grimm's Fairy Tales). The map size is 50×60, and the various lexical clusters are marked with different colours and shades.

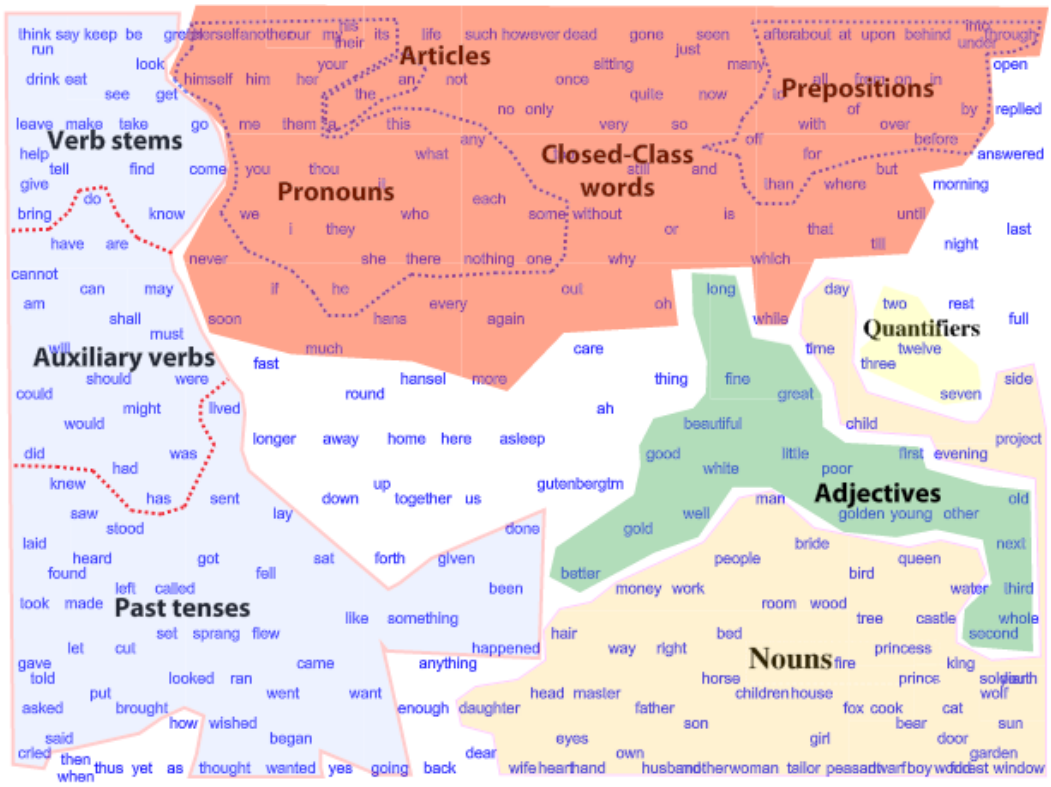


Fig. 5 – The semantic SOM by Zhao X., Li P., Kohonen T.

A similar semantic approach was used by Sjorb M. and Laaksonen J. (2011)[24] on visual data. They investigated two data sets, one with human-provided concept labels and another of unlabelled data for which they utilised automatically detected concept membership scores by using models trained on a labelled data set. By arranging the concept memberships of visual objects as components of a vector, these memberships can be used as the feature space for training a SOM. A visual and qualitative analysis of the SOM distributions of different concepts is augmented by a quantitative analysis. They set off from a video retrieval system based on measuring similarity combine low and high-level (or semantic concepts) visual features automatically extracted from the objects. Their main idea is to create semantic representations by extracting intermediate semantic levels (events, objects, locations, people, etc.) using low-level machine learning techniques, and then integrating these with semantic concepts. For example, they train detectors in semantic concepts such as "image containing a cat" or "video depicting an explosion" which can then be used as building blocks for higher-level querying of the database.

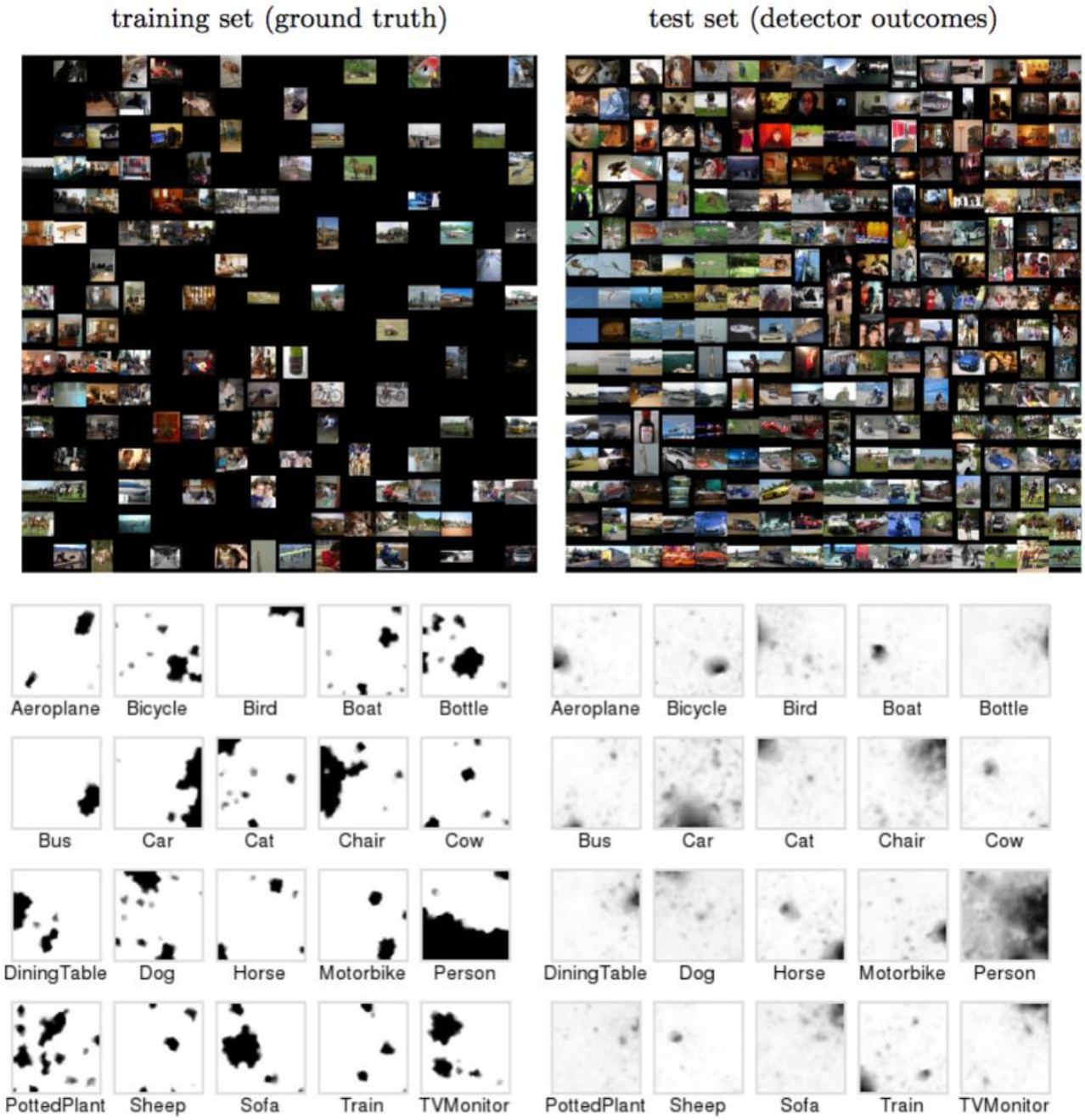


Fig. 6 – The visual semantics SOM of Sjorb M. and Laaksonen J.

In addition to visually and qualitatively inspecting the distribution of concepts on the SOM surface, they also propose a more quantitative analysis, in particular one focused on the “closeness" of concepts. The various distribution of components in the SOM model vectors represent different concepts and the closeness of concepts is thus estimated by

calculating the distance between these distributions arranged as vectors, with short distances indicating semantically close concepts. They apply the Earth Mover's Distance (EMD)[25] instead of Euclidean distance to produce better results.

O'Halloran and Kay (2015)[26] adopt their Multimedia Digital Semiotics approach[27] to develop the multimodal digital humanities research program[28], where they have structured an interesting analysis of Japanese Street Fashion[29] using SOM algorithms. This study demonstrates that SOM and topology learning can be used to reveal sociocultural patterns and trends. Using a SNN corpus-based, the algorithms:

> arrange the photographs of young Japanese people with a similar fashion look into a network, where clusters are represented by a single photograph (i.e. the node) which best display the fashion look of each cluster. The photographs in the nodes were linked to each other through lines (i.e. edges) to reveal the topological structure of the data. The thickness of the frame around the photograph in the node revealed the number of fashion designs in the cluster, and the thickness and numerical label of the edge represent the number of fashion designs which belonged to the path from one model to another.[30]

The representative photographs in each node were analysed using a social semiotic framework for clothing developed by Owyong (2009)[31].

Fig. 7 – The semiotic SOM of O 'Halloran, Kay L. for Japanese street fashion.

Fig. 8 – Some examples of Japanese street fashion.

The data set was composed of 2,249 images[32] automatically mined from the Japanese Fashion Association portal. Each image is linked to various metadata, such as the brand name, hair style, accessories and pattern tags as well as personal information about the person's age, occupation, music preferences and so on. To build the pattern recognition, the authors observed that there one important visual data for fashion that is capable of organizing the contents without any of the subjectivity inherent in the terms fashion analysts use to describe a fashion trend: colour. For more effective qualification, however, they managed the images using three strategies:

- "the source image was cropped to contain only fashion-related pixels"[33] and to eliminate background sounds;
- "the fashion area was separated into top and bottom segments"[34], because fashion designers use vertical symmetry in which the top and bottom elements are different;
- "the colour values for the top and bottom areas are averaged to form a six-dimensional features vectors, to represent every image by six-dimensional vectors".

They selected SOM to organize and visualize the data because it can explain the general-to-particular relationship between clusters. This means that the predominant styles are more represented, and the other combinations are able to express their connections.

The field of forecasting

A similar application of the integration of deconstructive, interpretative and reconstructive approaches with the potential to enhance results through quantitative modelling is the area of semiotics and trends. Ordinary qualitative research can be very efficient in gathering information about what is normal and mainstream today but, apart from some specialized groups such as early adopters and cool hunters, people are notoriously lazy at imagining a world more advanced than the one they already know.

Traditional socio-semiotic analysis can predict a range of possible consumer insights as well as identifying those that are likely. Semiotics is a methodology that can look to the future as well as understand the past. It helps to understand consumer culture and how such culture is changing. It can be used to spot emergent codes in given sectors, backed up by knowledge about the developments that are taking place elsewhere. Academics have developed a significant body of theory about consumer culture that helpfully explains how consumers express themselves through personal taste, how consumption interacts with lifestyle, and so on. This is very useful when the task is to anticipate how consumer behaviour might change in response to some new product, service or technological development.

While the study of semiotics has always fallen in the qualitative side of the qualitative/quantitative divide, there is no reason why we cannot support robustness in semiotic studies by quantifying the prevalence of signs or signifiers, estimating the prevalence of codes and assessing the rate at which these codes are becoming more or less dominant in a certain market or cultural context. Semiotics can help to look forward so as to 'confirm' trends, as the game of trend-spotting has burgeoned remarkably in recent years.

The application of similar techniques is particularly useful in the area of trends research, aligning qualitative foresight around social, cultural or sub-cultural change with powerful quantitative sizing and forecasting techniques that extend well beyond the count and display capabilities of online digital analytics. Nextatlas[35] is only one example of a business application for monitoring qualitative trends on the basis of a SNN corpus. The raw data, continuously acquired from individuals identified as trend-spotters through diverse social networks, are aggregated into a form suitable for identifying and studying trend birth, development and decline. NextAtlas is a web application that integrates human and machine generated data and insights related to the early identification of emerging trends.

More generally, to return to the question of a semiotic-critical analysis of data, according to semioticians such as Stockinger (2015: 1), we are currently facing at least five kinds of "big data": a) data from scientific calculus and research; b) data concerning profiles and digital identities (about behaviours, consumers, interactions etc.); c) statistical data concerning markets, brands, brand portfolios; d) data relating to citizens, government, public administrations, etc. and, finally, e) all the data produced by the big platforms governing today's internet (such as Youtube, Instagram, Facebook, Pinterest etc.): these compose the "backbone" of and, we might say, a sort of "translation and mediation channels" for other data. According to Stockinger, however, what is and will be increasingly important for a semiotics of big data is its capacity to detect and assess the form and value of data, their appropriation strategies, and cycles of production, management, and distribution strategies and, finally, their status as open data: as semio-cultural "objets de valeur".

Note biografiche

Daniela Ghidoli (daniela.ghidoli@gmail.com) (Phd in Science and Communication), has been working as a senior researcher and strategy manager for several research institutes on sentiment analysis, marketing intelligence, social and fashion trends and communication (such as, Accenture Interactive, Nextatlas, Armando Testa Group, Baba Consulting Milan). She is contractual professor at University of Turin, Brera Academy, IED.

Federico Montanari (federico.montanari@unimore.it), is Associate professor at the University of Modena-Reggio Emilia, in Sociology of Cultural and Communication Processes, and Semiotics of Media, after having taught in several other universities, such as Politecnico di Milano, University of Bologna, ISIA, IED, Iulm, and as visiting scholar at the University of California, San Diego. He works on sociosemiotic analysis applied to war and conflict situations, to the study of urban spaces and technologies, also in relation to cultural, visual and media studies. He also works on the philosophy of poststructuralism.

Bibliography

Aa.vv. (2011). Advances in self-organizing maps. Heidelberg: Springer-Verlag Berlin.

Andrejevic, M., Hearn, A., Kennedy, H. (2015). Cultural studies of data mining: Introduction. European Journal of Cultural Studies, 2015, Vol. 18(4-5) 379–394.

Bassett, C. (2015). Plenty as a response to austerity? Big Data expertise, cultures and communities. European Journal of Cultural Studies, 2015, Vol. 18(4-5) 379–394.

Boccia-Artieri, G., Bentivegna, S. (2019). Le teorie delle comunicazioni di massa e la sfida digitale. Bari-Roma: Laterza.

boyd D., Crawford K. (2012). Critical questions for big data. Information, Communication & Society 15(5): 662–679.

Burns, R. (2013). Digital community as data sharing: Science lab management apps. Paper to Digital culture: Promises and discomforts workshop. ECREA digital culture section, Berlin, 2–4 October.

Butler, J. (2002). What is critique? An essay on Foucault's virtue. Ingram D (ed.) The Political: Readings in Continental Philosophy. London: Basil Blackwell.

Deleuze, G. (1968). Différence et répétition. Paris: PUF.

Deleuze, G. (1992). Postscript on the Society of Control. October 59.

Fabbri, P. (2005). La svolta semiotica. Roma-Bari: Laterza.

Fletcher J., Smith D. (2002). A New Model for Converting Market Research Data into Actionable Insights. Citigate DVL Smith. Market Research Agencies.

Floch, J., M. (1990). Sémiotique, marketing et communication. Paris: PUF. Trad. it., Semiotica, marketing, comunicazione, Milano: Franco Angeli, 1992.

Fontanille, J. (2004). Textes, objects, situations et formes de vie. Les niveaux de pertinence de la sémiotique des cultures. E/C: www.ec-aiss.it.

Fontanille, J. (2015). Formes de vie. Liège: Presses universitaires de Liège.

Fontanille, J., Zilberberg, C. (1998). Tension et signification. Paris: Mardaga.

Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics, 43 (1), pp. 59–69.

Guilborgè, J., B., P. (2009). Internet emotional discourse analysis and innovation – A mix semiotic and text-mining method. Paris: Inevidence.

Iliadis, A., Russo, F. (2016). Critical data studies: An introduction. Big Data & Society, July–December, pp. 1–7.

Jaradat, S., Dokoohaki, N., et al. (2019). TALS: A Framework For Text Analysis, Fine-Grained Annotation, Localisation and Semantic Segmentation. IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), pp. 201-206.

Lakoff, G., Johnson, M. (1999). Philosophy in the Flesh, New York: Basic Books.

Latour, B. (1992). Where are the Missing Masses?. Shaping Technology/Building Society, Cambridge (Mass.): MIT Press. Trad. it., Dove sono le masse mancanti? Mattozzi, A., ed., Il senso degli oggetti tecnici, Roma: Meltemi 2006.

Latour, B. (1993). We Have never been modern, London, Harvester Wheatsheaf.

Latour, B. (1994). Une sociologie sans objet? Note théorique sur l'interobjectivité. Sociologie du travail, vol. 36.4, pp. 587-607.

Latour, B. Jensen, P., Venturini, T., Grauwin, S., Boullier, D. (2012). The whole is always smaller than its parts' – a digital test of Gabriel Tardes' monads. The British Journal of Sociology, 2012, Volume 63, Issue 4.

Lawes, R. (2002). Demystifying semiotics: some key questions answered, MRS Conference. International Journal of Market Research, vol. 44, n° 3 (WARC).

Lebart L., Piron M., Steiner J. F. (2003). La Sémiométrie, Essai de statistique structurale. Paris: Dunod.

Manchia, V. (2020). Il discorso dei dati. Note semiotiche sulla visualizzazione delle informazioni. Milano: Franco Angeli.

Manovich, L. (1999). Database as symbolic form. Convergence: The International Journal of Research into New Media Technologies, 5.2, 1999.

Manovich, L. (2012). Trending: The Promises and the Challenges of Big Social Data. In Matthew K. Gold (ed.) Debates in the Digital Humanities. Minneapolis: University of Minnesota Press.

Manovich, L. (2020). Cultural Analytics. Cambridge, Massachusetts, London, England: MIT Press.

Marrone, G. (2001). Corpi sociali. Torino: Einaudi.

Marrone, G. (2014). The invention of Text. Milano: Mimesis.

Marrone, G. (2018). Prima lezione di semiotica. Roma-Bari: Laterza.

Mayer R., Rauber A. (2001). On wires and cables: content analysis of wikileaks using self-organising maps. Advances in self-organizing maps, Heidelberg: Springer-Verlag Berlin, pp. 238-246.

Morace F., Traldi T., Camillo F. (2005). From Marketing To 'Societing' - Reading ethnographic material through the use of digital matrix and Semiometrie, ESOMAR, Innovate! Conference, Paris.

Morozov, E. (2013). To Save Everything, Click Here: The Folly of Technological Solutionism, Philadelphia: PA, Public Affairs.

Nagarajan, M., Gomadam, K., Sheth, A. P., Ranabahu, A., Mutharaju, R., Jadhav, A. (2009). Spatio-Temporal-Thematic Analysis of Citizen, Sensor Data: Challenges and Experiences. Web Information Systems Engineering - WISE 2009, Berlin, Springer, pp. 539-553.

O'Halloran, K. L. et al. (2013). Multimodal Digital Semiotics: the interaction of languages with other resources. Text and Talk: Special Edition for Michael Halliday (edited by Thompson G.) 33 (4-5) pp. 665-690.

O'Halloran, K. L. (2015). Multimodal Digital Humanity. International Handbook of Semiotics, Berlin: Springer, pp. 389-415.

Owyong, Y. S. M. (2009). Clothing semiotics and the social construction of power relations. Social Semiotics, 19(2), pp. 191-211.

Pasquinelli, M. (2015). Metadata Society. Braidotti, R., Hlavajova M. (eds.) Posthuman Glossary, London: Bloomsbury.

Podlasov A., O'Halloran, K. L. (2014). Japanese street fashion for young people: a multimodal digital human approach for identifying socio-cultural patterns and trends. In Djonov E., Zhao S., Critical multimodal studies of popular culture. New York: Routledge, pp. 71-90.

Reeves M., Ueda D. (2016). Designing the Machines That Will Design Strategy. Harvard Business Review, 2016. Source: https://hbr.org/2016/04/welcoming-the-chief-strategy-robot.

Rubner, Y., Tomasi, C., Guibas, L. J. (1998). The Earth Mover's Distance as a metric for image retrieval.Tech. Rep. CS-TN-98-86. Stanford: Stanford University.

Shelton, T. (2020). A post-truth pandemic? Big Data & Society, July–December, pp. 1–6.

Singh, V., K. (2012). Personalized Situation Recognition. PhD Dissertation, University of Irvine, Ca.

Sjorb M. and Laaksonen J. (2011). Analysing the Structure of Semantic Concepts in Visual Databases, in: Advances in self-organizing maps. Heidelberg: Springer-Verlag Berlin, pp. 338-347.

Stock, T., Tupot, M., L. (2009). Common denominators:what unites global youth?. World Advertising Research Center.

Stock T., Tupot M. L. (2015). Visualising Culture using Semiotics, Esomar publications.

Stockinger, P. (2015). Les apports de la sémiotique dans la valorisation des données numériques. L'exemple des archives audiovisuelles. Conference paper. Colloque Sémiotique et sciences humaines et sociales: la sémiotique face aux défis sociétaux du XXIème siècle, Centre de Recherches Sémiotiques de l'Université de Limoges, Limoges, 25-27 novembre 2015.

Vazquez P. (2012). How Do We Dissect Culture Networks? Psfk website, interview of Tim Stock, ScenarioDNA. Source: http://www.psfk.com/2012/03/dissecting-culture-sxsw.html

Zhao X., Li P., Kohonen T. (2010). Contextual self-organizing map: software for constructing semantic representations. Psychonomic Society, Inc. Source: https://sites.google.com/site/xiaoweizhao/tools

Zilberberg, C. (2006). Eléments de grammaire tensive, Limoges: Pulim.

Note

[1] Disclaimer: This article was conceived prior to the impact of the current Covid pandemic and its concrete examples are more concerned with the use and critique of data and big data in the context of marketing and social, cultural, and fashion trends. However, the authors believe that some key points, such as the critique of a "neutral" view of data, remains valid in relation to current events, particularly in terms of the use of data during the pandemic-caused crisis. For some insights on this topic, see, i.e., Manchia (2020); Shelton (2020).

[2] See, for instance, Nagarajan, *et al.* (2009), regarding the thematic and spatio-temporal analysis of events such as terrorist attacks from one side; or efforts to detect, via twitter and social media, natural risks and events such as the spread of wildfires or hurricanes and epidemics (see also, Singh 2012).

[3] For an interesting and up-to-date general survey of "Big data society" and "Data harvesting economy", see the website and online research (especially on Facebook): "Immaterial Labour and Data Harvesting Facebook Algorithmic Factory (1)" (August 2016): https://labs.rs/en/facebook-algorithmic-factory-immaterial-labour-and-data-harvesting/. And for a recent in-deep and interesting journalistic investigation (8 October, 2016): "Machine logic: our lives are ruled by big tech's 'decisions by data'", in: *the Guardian online*: https://www.theguardian.com/technology/ 2016/oct/08/algorithms-big-tech-data-decisions.

[4] For an interesting comparison of differences between Obama, Hillary Clinton and Trump's use of statistics and Big Data in their campaigns, cfr. *The Economist*'s dossier: "The role of technologies in the presidential election", *The Economist*, Nov., 20th, 2016: http://www.economist.com/news/united-states/21710614-fake-news-big-data-post-mortem-under-way-role-technology?zid=311&ah=308cac674cccf554ce65cf926868bbc2.

[5] Reeves M., Ueda D. (2016), "Designing the Machines That Will Design Strategy", Harvard Business Review, 2016. Source: https://hbr.org/2016/04/welcoming-the-chief-strategy-robot.

[6] Tim Stock interview (2015), "ScenarioDNA announces approval of culture mapping patent to visualize big data", Prnewswire.com, source: www.prnewswire.com/news-releases/scenariodna-announces-approval-of-culture-mapping-patent-to-visualize-big-data-300086263.html

[7] Vazquez (2012), How Do We Dissect Culture Networks? Psfk website, interview of Tim Stock, ScenarioDNA. Source: http://www.psfk.com/2012/03/dissecting-culture-sxsw.html

[8] Jonathan Cousins interview (2015), ScenarioDNA announces approval of culture mapping patent to visualize big data, Prnewswire. Source: www.prnewswire.com/news-releases/scenariodna-announces-approval-of-culture-mapping-patent-to-visualize-big-data-300086263.html

[9] See Fontanille (2015), for a partially similar summary scheme of levels of semiotic analysis, or "planes of immanence" (although, in Fontanille's case, the scheme is much more thorough and articulate and theoretically motivated). This list is also not far from other proposal analytical schemas coming from sociolinguistics and cultural linguistic and linguistic socio-semiotics, such as "SPEAKING" model, from Dell Hymes.

[10] "An insight, in the context of market research, is a new finding about customers in a market which, when acted on by an organisation, gives it a competitive advantage), and then through to effective marketing decisions", Fletcher J., Smith (2002), *A New Model for Converting Market Research Data into Actionable Insights*, Citigate DVL Smith.

[11] See for example Alt & Brighton (1981).

[12] Cf., Lebart, Piron, Steiner (2003), about "Sémiométrie".

[13] The method was designed for a European ethnographic study called "The Material Culture of Happiness", a permanent research program created by FCL to provide insight into day-to-day happiness experienced in eight European countries, namely Spain, France, United Kingdom, Italy, the Netherlands, Germany, Finland and Russia. For details about the methodology, see the work of Morace, Traldi, Camillo (2005), From Marketing To 'Societing' - Reading ethnographic material through the use of digital matrix and Semiometrie, ESOMAR, Innovate! Conference, Paris.

[14] For details about this methodology, see Morace, Traldi, and Camillo (2005).

[15] For a presentation of this methodology, see, Guilborgè, Bonny (2009): Internet emotional discourse analysis and innovation – A mix semiotic and text-mining method.

[16] Floch J. M. (1990); Fontanille, Zilberberg (1998); Zilberberg (2006).

[17] Guilborgè, Bonny (2009).

[18] See, Lawes (2002),
[19] See, Kohonen T. (1982, pp. 59-69), about the "Self-Organized Formation of Topologically Correct Feature Maps".
[20] See, Aa.vv. (2011).
[21] See, Mayer R., Rauber A. (2001), pp. 238-246.
[22] The Java SOMToolbox is developed at the Institute of Software Technology and Interactive System at the Vienna University of Technology and licensed under the Apache License, Version 2.0. Source: http://www.ifs.tuwien.ac.at/dm/somtoolbox/download-releases.php
[23] See, Zhao, Kohonen (2010).
[24] See, Sjorb, Laaksonen (2011), pp. 338-347.
[25] EMD measures the minimum cost of turning one distribution into the other, where in this case the cost is the value that needs to be moved times the Euclidean distance over the 2D map surface. In this analysis, the authors used the C implementation for EMD provided by Rubner, Tomasi, Guibas (1998), The Earth Mover's Distance as a metric for image retrieval. Tech. Rep. CS-TN-98-86, Stanford University.
[26] Cf., O 'Halloran (2015), pp. 389-415.
[27] O' Halloran et al. (2013), pp. 665-690.
[28] In the Multimodal Analysis Lab in the Interactive & Digital Media Institute at the National University of Singapore from 2008 to 2013.
[29] Podlasov, O'Halloran (2014), pp. 71-90.
[30] O 'Halloran (2015), pp. 389-415.
[31] Owyong (2009), pp. 191-211.
[32] See also, for similar methodologies, the research of Lev Manovich (2012; 2020).
[33] Podlasov, O'Halloran (2014), p. 76.
[34] Podlasov, O'Halloran (2014), ib.
[35] Nextatlas is a technology powered by iCoolhunt,