Ana Gallego Cuiñas

Literature Seen Through Big Data and Artificial Intelligence: Key Concepts and Critical Challenges

> In the clinic of the art of reading,
> the one with the best vision is not
> always the one who reads best
> Ricardo Piglia

These lines from *El último lector* (*The Last Reader*) (2005) by Ricardo Piglia clearly reflect the fact that the literary is a matter of perspective or scale, a radically historical ideological and aesthetic positioning that constructs a truth. For Piglia, the ideal reader is one who literally cannot read well – and in this statement there are indubitable echoes of Harold Bloom's *The Anxiety of Influence* (1973) – because their vision (or point of view) (Kittler 2010) compels them to read *up close*, like a short-sighted person who needs a magnifying glass to make out anything tiny and particular – the hidden structure of text that becomes a system of secret correspondences that have to be uncovered in every era. This metaphor perfectly illustrates the critical approach of *close reading* (Empson 1966; Richard 2004), based on hermeneutics and/or narratology, which has predominated in literary studies since the beginning of the last century. The critical obverse of this would be *distant reading* (Moretti 2016), sociological and/or quantitative in nature, which was developed in the second half of the twentieth century and would fit the metaphor of the *far-sighted* reader, who with the perspective of distance can access the general context of the texts to establish formal and material relations of a discursive, social, cultural and economic nature.

In this paper, I take this dual perspective as a basis for the following proposal: literary criticism of the twenty-first century needs to overcome this counterposition in approaches, misleadingly understood as opposites, to practise a combined mode of reading in which textual interpretation, materialism and dataism complement each other, for the sake of a more thorough and organic intellection of the literary fact and of its social function. The hypothesis I begin with is that, on the one hand, the distant reading that the social sciences and computational techniques adopt is ever more necessary to analyse the aesthetic and material function of literature in society, both in diachrony and synchrony. On the other hand, all data analysis requires careful – *close* – attention to the structures, qualitative and quantitative, that appear as a result of research. Thus the epistemic crossover between literature, sociology and big data is not only possible but

https://doi.org/10.1515/9783110753523-003

desirable, since through the intertwining of these scales – or methodological strategies (English and Underwood 2016) – we can achieve better findings with greater breadth and concision, which benefits both theory and literary criticism, and computational science. We could call this approach *cross-reading*,[1] which in turn suggests another ocular metaphor, *cross-eyed reading*, displaced and interposed – a form championed by Ricardo Piglia himself (2005) as a highly productive way of reading for the Argentine cultural field.

This essay therefore presents a *crisscrossed* and *situated* reflection on Literature and Big Data centred on two main themes that overlap and supplement one another: first, the use of Big Data and Artificial Intelligence (AI) in literary culture, based both on the mechanisms of production, circulation and consumption of literature in the market, and on the impact and utility of computational methods in the field of criticism; second, the use of certain literary and philosophical categories that could be advantageous for the – "situated" (Haraway 1995) – epistemic and political intellection of the functions of Big Data and AI. This proposal is undoubtedly only a starting point, which has the ultimate aim of contributing to the much-needed design of an agenda for the *literary criticism of data*[2] that contemplates the multiple possibilities of collaboration and dialogue between the Humanities, Sociology, Data Science and AI.

1 Use of Big Data in Culture and Literary Criticism

The first thing that needs to be stated in this introduction is that to think about the integration of big-data techniques or, what amounts to the same, about quantitative methods of measuring data, in the sphere of literature is to base our thought, first of all, in the social question (Halavais 2015) – that is to say, in the sociology of literature. This discipline examines the literary object as social fact or product, common and collective, which crystallizes the eternal conflict between technology (numbers) and culture (letters). The criticism of concentrating and standardizing cultural objects has its roots in Adorno and Horkheimer's *Dialectic of Enlightenment* (1944), since when this tension has been constantly re-

[1] Bootz and Laitano (2014) put forward the same name to designate a data visualization model based on Spinoza's ontology. My idea, however, points to the crossover and simultaneity of two methods (close- and distant-) that, in the historiography, have been seen as opposites.

[2] One could also write *literary dataism* or *computational literary criticism*, but these terms are more restrictive than the one I propose, which is more labile and versatile. This modality would fall within the epistemology of 'Critical Data Studies', based on the study of data and their different critiques, systematized by Dalton and Thatcher (2014).

peated one way or another in “literary culture” (Gallego Cuiñas 2022). Adorno, remember, takes up Benjamin’s thesis on the alienation of a work of art through its commercial reproducibility – circulation – which would have to be in opposition to its *aura*, the Kantian authenticity, of a cultural object. The attack on mass culture and mercantile utilitarianism is evident, as is Benjamin’s romantic and idealized view, by Adorno himself and later by Guy Debord in *The Society of the Spectacle* (1967). Today, the consumption of Art reveals itself in its pure contingency more than ever (Gallego Cuiñas 2021), in that being-in-the-moment and in the ephemeral that the School of Frankfurt and its followers despised. Therefore, if we wish to read from the present and out of contingency, as a privileged mode of literary production, in the studies of literature we would have to include a sociological focus and big data, the volume and speed of which have increased exponentially over the last decade with an impact in the culture sphere that is both material and symbolic, and which we cannot avoid:

> In the arts and humanities, the notion of big data is still in its embryonic stage, and only in the last few years, arts and cultural organizations/institutions, artists, and humanists are starting to investigate, explore, and experiment the deployment and exploitation of big data as well as understand the possible forms of collaborations. (Schiuma and Carlucci 2021: xxiv)

In the last five years, studies on the humanities and quantitative methods have been appearing with much more frequency, particularly in English-speaking academia, followed by the French, who historically have had more porous borders between the Humanities and the Social Sciences than the Hispanic world, which facilitates the transdisciplinary crossover. In the specific case of literary studies, the most distinguished researchers in data or computational criticism are North American: Paul Delany and George Landow, Matthew Jockers, Andrew Piper and Ted Underwood. In the Iberian-American world, some names of note are: Belén Gache, Claudia Kozak, Carolina Ferrer, Carolina Gainza, Germán Ledesma, Diana Roig-Sanz, Germán Sierra and Alex Saum-Pascual.[3]

This leads me to the second aspect I would like to make clear from the start: the utilization by sociology and data science in literary studies transcends the concept of *Digital Humanities*. I agree with Underwood that this label is more of a reaction to a tactic – which has strengthened the use of digital technology and open science, essentially through the idea of the archive, digitalizing and cataloguing historical texts that are difficult to access – than referring to an area of

[3] It needs to be made clear that Iberian-American literary criticism of data has left the sociology of literature to one side, which I believe is fundamental for the material understanding of the literary, not only the aesthetic understanding.

knowledge in itself. What has undoubtedly occurred is a *digital turn* (Dobson 2019), which in the second decade of this century has become a *computational turn* whose epistemic value is founded on the ecosystem of quantitative methods that the sociology of culture has traditionally used. Today these methods for measuring have also become methods for mediation – we are all now digital researchers – at the mercy of the aforementioned phenomenon of the intensification of datafication and of the advances in AI for creation and cultural consumption.

1.1 Literature, Market and Artificial Intelligence: Some Hypotheses

There are three main ideas that currently underpin the production, circulation and consumption of literature: the figure of the writer, the literary work and the reader. In criticism, these categories have become "zombies", insufficient to express the "new" state of what is literary (Gallego Cuiñas 2019) in which orality and print, the human and the digital, literature and literary culture live together. The space and the computational techniques have broadened the creative experience, to the point at which the digital environment has itself become a medium for the production and distribution of literature, with ever more enthusiasts. At the same time, we rely more and more on the predictive potentiality of big data (Sádaba Rodríguez 2020), both in reception studies and in the creative industry, to evaluate trends in the literary and artistic market, the quality of products and the degree of user satisfaction (Piper and Portelance 2016; Schiuma and Carlucci 2021). How therefore does Big Data and AI affect the ontology and epistemology of the literary?

Starting with this question, I propose some core ideas for reflection on the use of data analysis in the sociological and materialist approach of literature, which (re-)opens several lines of research for today's academia:

From author to '(artificial) writer'. Big Data is a highly advantageous instrument for the development – in literary studies – of what I call *writer criticism* (Gallego Cuiñas 2022), which is based on the sociological, materialist and aesthetic analysis of the figure of the writer, utilizing new methods and elements that have not received sufficient attention: the use of digitalized biographical archives; the production and reception of writer bots; the authorial image on social media; *bookporn*, interaction with other mediators of literary culture; performativity in the public sphere; the extent of education and literary professionalization, and so on.

However, in the creative sphere, *artificial creation* or the literary production of texts by an AI has imperilled the pristine category of 'author'. This takes us back to the same conundrum that the Frankfurt School detected regarding the loss of the work of art's aura in the first half of the twentieth century, now applied to the anthropocentric notion of the author figure as the intellectual property holder of a text, and questions the author's hegemony and validity (Badía Fumaz 2012; Berti 2015; Herrmann et al. 2022). In the same way, the romantic idea of the author as genius creator, based on the symbolic value of human and individual literary creation, which is difficult to put a value on, has become unsustainable. In contrast, a mode of artificial creation, collective in origin, is growing, with a symbolic value that is more easily quantifiable in material and economic terms. In both cases, in the literary field the trend is to talk of the (artificial) 'writer' rather than the author.

To this we can add the commercialization of the aforementioned creation of *artificial works* – to give them a name – made by an algorithm, most in open access, and aimed at mass consumption, which until now have been 'overseen' by human writers. As in other areas of creative industry (i.e. music, art, photography, graphic design, et cetera), this new mode of (digital) literary production is being proclaimed as an attractive mode of exploitation and extremely beneficial for the cultural industry and institutions, which, with the control and use of an AI in the creation of an artistic work, can also become the co-authors – not merely co-producers – of texts.

Artificial creation, therefore, generates at least three theoretical and political problem areas:

(i) First, the entry into crisis that I have already mentioned – a new death? – and resignification of the notion of authorship tied to the concepts of 'authenticity' and 'intellectual property', which would shift from being individual to collective (the final text is the result of an algorithm that works with the big data obtained through millions of works), from being human to technological. The literary algorithms, avatars or bots (Olaizola 2018; Sierra 2022: 13) that automatically create literary content are also an example of the way in which digital production contributes to the performativity of the category of author, which changes to that of 'writer' (Gallego Cuiñas 2022), given that the capitalist notion of authorship is being displaced, and it is becoming very difficult to distinguish intellectual property.[4]

(ii) Second, the place of the writer in the creative process transforms and shifts from the romantic value of the genius who produces an original and unique work, to the pre-capitalist value of co-creation, appropriation and communi-

[4] Literary (ro-)bots are algorithms that produce content, above all on social networks: "What distinguishes bots from other types of software is that they interact with and or produce content for human users, often taking on a human personality" (Olaizola 2018: 239).

tarian transmission of the work produced by an AI. Therefore, the writer would have to become a craftsman or a mixer, mediator or *gatekeeper* (Gallego Cuiñas 2022) of the resulting artificial work.[5]

(iii) Third, this mode of digital (re-)production results in the loss of bibliodiversity (cultural, of genres, authors, et cetera), and in the dangerous increase in colonial and gender biases (the majority of the works collected in databases are written by men and edited in cultural systems of the north – that is, of hegemonic cultural systems), unless, in the artificial creation phase we find ourselves in, there is a gatekeeper or guarantor of these egalitarian, decolonial and inclusive values.

From the work to the '(artificial) work under construction'. We first need to distinguish between the works that are born and are (re-)produced in the digital medium,[6] those that are hybrid (print publishing and digital technology), and those that are digitalized. However, they all share five essential traits: "digital or numerical representation, modular composition, variability, automatization, and transcoding" (Berti 2018: 139). Second, the alphabetic and binary codes are interpretable aesthetically and computationally based on common notions such as: mutability, contingency, collectivity, anonymity, fragmentation, brevity, materiality, and gamification. Based on these premises, I have come up with three hypotheses:

(i) From the creative point of view,[7] it is clear that the production of digital literature[8] is a modality that has been growing in Ibero-America[9] in recent years, mainly through the practice of poetry (cf. Gache 2006; Kozak 2010 and 2017; Berti 2015; Gainza 2019; Ledesma 2022; and Saum-Pascual 2022). This literature experiments with the signifier, with multimedia elements and with the archive[10] through algorithms (cf. Bolter 1991; Hayles 2008, or Côrtes Maduro 2017), which is why it is often associated with the notion of

[5] It is clear that this mode of artificial production of literature recovers – literally – the notion of tradition as the great producer of texts, something that Borges stood for.

[6] The large majority of texts today naturally appear digitally before in print form.

[7] Remember that it was precisely with the publication of Mary Shelley's *Frankenstein* when creativity became a great (anthropocentric) value associated with divinity.

[8] To mention a few notable names in Spanish-language digital literature: Belén Gache, Iván Marino, Luis Espinosa, Marina Zerbarini, Mariano Sardón, Gustavo Romano and Alex Saum. Argentines lead the list in numbers, followed by Spanish writers.

[9] It has been developed and studied more in the English-speaking world. See the following databases: Electronic Literature Knowledge Base https://elmcip.net/; Electronic Literature Collection https://collection.eliterature.org/ or NETescopio https://proyectoidis.org/netescopio/.

[10] On the one hand, digitality is a mode of production and a materiality, and on the other, digital literature works with the existing tradition, with what is repeatable becoming prime material.

avant-gardism[11] or with experimentalism[12] – not only aesthetic (of a formalist stamp) but also technical.[13]

The *live* creation of artificial work – cyberwriting (i.e. twitterature, instapoetry, literary memes and avatars, transmedia narrative or playwriting, Wonderbook, and literature made on WhatsApp or Wattpad) – has also been gaining greater visibility. This generates literary value that one could call *relational*, based on the participation of the reader – hence playing with Didi-Huberman's phrase "work under construction" (2015:16) – and on the 'live' or 'serialized' consumption of *readable* fiction, simple and direct.

Lastly, there is a boom in artificial works that are recycled or reworkings (remix or sampling), *works of works*, that incorporate GPS, audiovisual content or QR codes, where the extraordinary transmediality and performativity of the literary in the 21st century is at the forefront.

(ii) From the critical point of view, computational technology helps to probe into the nature of the written work, both from past eras and in the present: types of language, styles, biases, techniques or genres that have been adopted the most over the years and across cultures, which are easily studied through the mass digitalization of literary texts (i.e., Books Ngram Viewer, Blatt 2017 or CATMA). The data analysis of digitalized works lends itself to a predictive criticism that can calculate the success of a text, construct series (of texts) and evaluate their level of innovation.[14]

(iii) From the material point of view, the artificial literary work transcends the book-object as the receptacle of the text (Striphas 2011). The machines for making and selling literature cease to be the printing press and the distributors, for now it is the digital medium and its new formats that create and distribute it on platforms and social networks. The tools of production, publishing, reading and conservation of digital literature have changed radically in the second decade of this century, to the point at which some computer skill is required for it, although programs are being designed that are ever easier and more democratic to use.[15]

[11] I am thinking about visual, concrete and sound poetry and their performative performance.

[12] See the anthology of experimental literature compiled by Tomás Vera Barros (2014).

[13] Rafael Pérez y Pérez is one of the most outstanding researchers in computer creativity and has produced several books with AI (see http://www.rafaelperezyperez.com/). Questions abound: Do algorithms have an aesthetic? What form of appropriation is it by the author regarding the product generated by AI? If the authorship is held by the publishers, are we returning to a hegemonic authorship? To a post-human authorship?

[14] In this regard, in a few years Wattpad will be able to write its own stories based on the big data it has obtained from the success of certain stories on the platform.

[15] María Goicoechea de Jorge explains: "These types of programs have enabled a greater number of authors to access this genre who are not necessarily connected to the academic world or

From reader to (digital) ‘prosumer’. The reader of digital literature is always a co-producer or a prosumer (Villanueva 2022: 5), because the interaction with the work is a consubstantial part of the reading process. In fact, the artificial work acts as a kind of toy that is both literary and computer (with readings coded according to the text and to the use of technology), which goes along with a contingent and non-standardized use. The temporary nature of digital reading is manifold (it goes backward and forward, it ends, it breaks into parts), transmedial and simultaneous, non-linear, and successive like printed reading. But reading is also conceived in series and intermittency – like the nineteenth-century serialized novel or mass-culture subscription-based instalments – from the same digital setup, as occurs with Serial Box or in Spanish with the platform Black & Noir, which operate like distributors on mobiles and tablets of serialized literature.

Furthermore, Big Data has been heavily used in the study of audiences, with methods based mainly on Singular Value Decomposition (SVD).[16] The data analysis of reception measures the way in which rating patterns change and how literary prestige is formed and circulated. The ‘stock exchange’ that underlies aesthetic judgement has barely altered over the last century (Underwood 2019), since it has always been in the hands of the same authorities of the market and academia: institutions, universities, publishers, prizes, critics, et cetera. However, the democratization of taste that has gone hand-in-hand with digital and technological progress has undeniably impacted the appraisal of literary value, in such a way that not only are we witnessing an unprecedented proliferation of producers of literature and literary products, but also of readers/consumers, ‘digital prosumers’ who rate literary value on platforms such as the aforementioned Wattpad, or Goodreads, a social network of readers and writers who act as critics of other books and who influence the prescription of taste (Bourdieu 2002).[17] In the wake of this shift from academia and the cultural press as agents of literary value, we find *booktubers*,

to research. The following are three of the currently most popular programs with their most notable characteristics and differences: Twine, created by Klimas in 2009 with a free software licence; Inklewriter, a tool created by the games company Inkle, co-founded by the British mathematician and writer Jon Ingold; and Undum8, created by Millington in 2010 with an MIT licence (Figure 4). The importance of these types of programs is that they have democratized the use of this genre of digital literature, since advanced programming knowledge is no longer needed to write a narrative or interactive game” (2019: 175).

[16] Singular Value Decomposition (SVD) is a technique used in 2009 to predict user ratings for films on Netflix.

[17] Up until now, there are hardly any comments on self-published books on Amazon and similar platforms. Instead, the majority are from the usual publishers. We should also consider the content – book – recommendations that users make based on their consumer experiences or the data platforms based on algorithms (Vanoli 2019: 27–34).

bookstagrammers and Amazon algorithms – "symbolic expropriation" is what Jorge Carrión has come to call the *modus operandi* of the online retail site – which condition the book choices of (digital) mass culture upon the basis of patterns of consumption carried out using Big Data. Thus, Amazon acts like a virtual and democratic bookshop (Lefort-Favreau 2021: 79) that prescribes taste and dictates the norm – being a new instance of value appraisal that is consumerist and populist in style – by virtue of the quantitative concentration of information, which always entails a certain standardization: the tyranny of the masses, which replaces the former tyranny of the elite, of the mesocratic bourgeoisie that has dominated the construction of literary value in the modern world (Gallego Cuiñas 2019).

To finish this section, we cannot forget that the opposite phenomenon also exists: the appraisal of literary value with digital parameters that perpetuate and defend an elitist community, a ghetto, of prosumers of literature: "The electronic art and visual poetry market have adopted the NFT (non-fungible token) as the preferred format for diffusion and sale. But it no longer only applies to the visual arts, but also to texts, mixed artworks and even novels" (Sierra 2022: 13).

1.2 Literary Criticism and Big Data: A New Challenge for the Sociology of Literature

The sociological study of literature, which was prevalent in the 1960s and 1970s in Latin America, today only makes up between one and two percent of academic publications on literature in the Spanish language (Gallego Cuiñas 2022). The majority today take Pierre Bourdieu's perspective,[18] yet few dare to use quantitative and computational methods for literary analysis (i.e., Roig-Sanz 2019 and Gallego Cuiñas 2022). This shows that the question about the nature of the sociology of literature and about which methodological instruments they should use continues to be relevant and highly debated today. In this context, the use of a *critical dataism* seems like a very fertile space of expansion and experimentation both for the sociology of culture and for the literary studies of the future.

Let us remember that the articulation of the sociology of literature as framework of thought dates back to Marxist structuralism, but it did not begin to be developed as a discipline until the sixties, with the Birmingham School. Subsequently, in the seventies and eighties, a generation of cultural and literary sociologists emerged that carried it on into its brightest period. The Marxist approach was, from the 1990s, then displaced by the advance of New Materialism and the

[18] See Moraña (2014) and Maltz (2020) on the colonization of Bourdieuan thought.

application of an anti-hermeneutic and anti-aesthetic methodology that Moretti called “distant reading”. The weaknesses of this clearly positivist method have already been pointed out, although this does not, in my opinion, invalidate the idea that the sociology of literature and dataist technique, with statistical and computational methods, could prove politically advantageous for twenty-first century literary criticism, given that literature is historical and ideological merchandise, tied to the real economy, and it depends, in its dispositions, sociabilities and affects (Brouillette 2017: 280), on the numerical logic of the economy, on the possibilities of the digital and on the big data of the market. Value can undoubtedly be extracted from these – they give us a pattern, new forms of production, association and a forecast – because they represent and transform, materially and symbolically, literary taste. In other words, this new form of producing knowledge can lead us to the configuration of a new epistemic field.

How can we therefore give legitimacy to a sociology of literature based on dataism today? There is no avoiding the fact that one of the most important problems in literary criticism is precisely the legitimacy of the method or theoretical approach of the researcher. In reality, this is a question of strategy in the academic struggle for intellectual capital, safe from self-absorbed and centripetal methodological trends, where what is really at stake is the professional standpoint of the critic, not the conceptual make-up of a field (the knowledge or cognition), but recognition (Morgan 2013) in a sparsely populated ivory tower of specialists. Hence quantification – being associated with the social sciences and positivism – becomes a twofold enemy for the critic and theorist of literature, since it presupposes both the mix with sociology and an attack or questioning of the qualitative methods inherent in the humanistic field. However, one thing does not exclude the other, and in the third decade of the 21st century we cannot keep turning our backs on the myriad of resources that digital culture makes available for the theoretical, sociological and historical study of literature. Its use offers us a tool, not a substitute for but a complement to criticism and creation, already commonplace in the sphere of linguistics and historiography (Lemercier and Zalc 2019) – which are highly familiarized with working with corpus and archive – and becoming more so in other arts, although up until now they have barely used the databases, sources and samples of data on a large scale in the humanities.

To speak plainly, many humanists question the validity of quantitative methods, which they brand as neoliberal,[19] without understanding them or having tried out their uses politically, which in some cases represent real challenges for left-

[19] We cannot deny that only universities in the northern world can use these highly expensive methods, which enable access to data. In the end, information is power.

wing, materialist literary criticism. Neither numbers nor quantifications are intrinsically objective or bad: they are merely signs, and as such, depend on (ideological) interpretation. This is why “numbers are becoming more useful in literary study for reasons that are theoretical rather than technical” (Underwood 2019: xi). Why? What can data science provide literature with? The techniques of quantification expand the scope of our study toward new forms of representation – such as data visualization (Karsdorp et al. 2021) – and toward new ends, at the same time as providing the aforementioned *relational value* – as Saussure and structuralism understood it – through the configuration of different statistical, digital and computational ‘models’ or ‘structures’, which strengthen theoretical and critical analysis, focused on themes, problems, genre(s), characters, periods, authors, et cetera (Piper 2017). Likewise, the access and handling of big data (re-)opens sociological lines of research – not widely explored in Hispanism – that can be developed through this approach, without giving in to data fetishism. The three that I believe have the most political repercussion are:

i. **Study of invisibilized works.** This is the area of interest of Moretti (2016), focused on the possibility of accessing the big data provided by texts that have been marginalized – historiographical blind spots and gaps – by the hegemonic mechanisms of recognition, which have generated the canon of literature and its modes of representation (Bode 2017; Roig-Sanz 2019).

ii. **Study of taste**. There are two options here: one aimed more at academic criticism, which Carolina Ferrer calls “criticometría” (“criticometry”) and which entails the bibliometrical or citation analysis of certain critical theories and trends in academic publications – in different times and spaces – on the database of semantic associations, of repetition and generalization, in the geopolitical context of their utterance. This helps to trace the global map of academic geopower relations, and of their capital, in every era (Goldstone and Underwood 2014; Ferrer 2015; Espino 2020). The second option is focused on the cultural field, through the analysis of newspapers, notes, journals, digital content, prizes and other discourses that can contribute to learning the way in which literary value has been appraised and how social prestige is constructed outside the academy (Underwood 2019, 69),[20] also taking into account its variations in time and space (Martínez-Gamboa 2016; Posada 2019). One example is the book by Archer and Jocker, *The Bestseller Code: Anatomy of the Blockbuster Novel* (2016).[21]

[20] For example, in his study Underwood shows with quantitative methods that the way we judge a literary work generally changes every thirty years.

[21] The problem is that up until now, this type of study has omitted the material analysis of gatekeepers – publishers, translators, agents, etc. – which is essential, from my point of view, for con-

iii. **Study of figurations and networks of sociability**. In the area of research pioneered by de Nooy (1991), analyses have been carried out – first in psychology and then in sociology (Lemercier and Zalc 2019, 101) – of networks in this new era of Big Data (i.e., Jean So and Long 2013; Gallego Cuiñas et al. 2020) on the connections that are produced in order to build *value networks* on digital platforms, using content and profiles on social media (Twitter, Facebook, Instagram, Linkedin). This seems to be a highly productive opportunity for understanding the way in which the figures and figurations of the writer are currently constructed, as I stated earlier, but also for examining the role that algorithms, avatars, bots and intermediaries (i.e. publishers, agents, other writers, Granta, festivals, et cetera) take in the promotion of a work, a genre or an author, and their symbolic and financial resources.

Despite the new research areas that this sociology of literature based on the analysis of big data and on AI opens up, we cannot ignore the fact that, currently, the traditional *close reading* is still predominant in academic publications, and therefore provides much more professional assurance than this new agenda that, at the moment, does not enjoy the same prestige in our field. The price a humanist has to pay for expanding their discipline's horizons is high, since not only are they faced with another discipline that they have to learn but also with institutional and material problems deriving from the lack of technical training and infrastructure,[22] as well as the academic loss of worth as judged by the agents who control the field (Underwood 2019: xviii): journals, publishers, assessment agencies, departments, institutes, and so on. In short, the impact of *new* methods and study aims is always slow, and at first incurs rejection in the disciplines of origin, thus making the decision to opt for this type of research evidently riskier and more unrewarded.[23]

structing literary value in the contemporary world. The combination of both perspectives would give a more thorough and complete interpretation of the modes of production and circulation of value in the literary field.

[22] Thus the wealthier northern academia, with better material conditions, are always the pioneers in taking up innovative methodologies.

[23] This is why much of computational criticism carried out on humanist objects is being done by computer scientists, specialists in information and communication sciences, economists, engineers, and so on (Schiuma & Carlucci 2021).

2 Use of Literary Categories in Data Science

There is no doubt that the fact of dataism needs an interpretation, needs a *situated* narrative meaning. This assertion opens the door to the possibility that literary criticism has something to contribute to data science and not only the other way around.[24] What am I referring to? That we find theoretical categories and critiques of analysis – principally from Russian formalism and from (post-)structuralism – that help to explain the functioning of the algorithm and to articulate a kind of *Big Data hermeneutics* that will illuminate the critical and political thinking of computational methods and results. Namely:

i. **Close Reading** (Empson and Richard). This eminently literary strategy is fundamental for supervising the algorithms and for data interpretation that guarantees the certainty and efficacy of the results (Koskimaa 2005). The creation of the algorithm is also a *reading machine*, to use the Deleuzean metaphor; in other words, it is a model for reading. Hence every data reader is a co-producer of a *significant* structure, which comes from micro-thinking, not only macro-thinking: “thinking small in order to think big” (Piper 2018: 9). As well as knowing how to read between the lines, this involves acting as a kind of mediator or guardian of knowledge – that is, a gatekeeper (Gallego Cuiñas 2019) – that vouches for the value of the knowledge generated. This entails the differentiation and discerning of information, the removal of bias, and ensuring the quality of data: *Smart Data*. Thus, in computational criticism, the humanist (the ethnographer, the philosopher, and the philologist) becomes a gatekeeper because the authority, “the law” – in the Kafkian sense from the parable “Before the Law” – is still essential not only for giving *meaning* but also to *situate* and make visible gender, geopolitical and colonial inequalities that the algorithms do not *see*:

> The proficient and valuable use of big data needs the personal and organizational capacity of asking the right questions and in the right way. Big data is powerful only if it is generated, combined, or supported by the creation of strong narratives, organizationally and contextually framed. This means that the big data has to be “thick”, i.e., not only quantitative but most importantly qualitatively relevant (Schiuma and Carlucci 2021: xxv).

[24] “The hypothesis of the mutation of art due to digital transformation has been widely accepted, but it is also worth reversing these suppositions. As Kenneth Goldsmith states, ‘if one thinks about it, the engine that drives the internet is literature […]. It gives the possibility of cutting, copying and pasting, imitating the movements of language. Language has never been moved in the way that we are moving it today (14 February 2014)’.” (Helgueta Manso 2022: 43).

ii. **Series and construction** (Tynyanov). These concepts belong to Russian formalism. The former refers to the property that the literary text has of breaking down into different units of meaning, the same procedure that algorithms use today. The latter refers to the "constructive function" and "relational function" of literary works, texts and units in similar "series" or systems of correspondence, as occurs in computing. Obviously the selection of texts, topics or units of meaning has a subjective or immanent component in literary criticism, as the algorithmic training of data processing also has, which presupposes a "value in itself" of the elements (i.e., *Topic Modelling*). This is why human – and humanist – readers are needed, to supervise the constructed computational models, since the aforementioned *relational value* is responsible for the recontextualization of texts in series – and one must remember here that a context is a point of view – as well as for its decontextualization.

iii. **Intertext** (Bakhtin and Kristeva). The theory of intertextuality, structuralist in origin, is based on the assumption that all text refers to other texts (in ideas and statements, in diachrony and in synchrony), in a more or less evident way (Pozuelo Yvancos 1994). This *value* of repetition or of the quotation also works as the constituent principle of algorithms that work with big data to come up with the correlations of a "series" or accumulation of meanings. Moreover, it is interesting to bring in here other literary notions such as "parody" and "irony", which require human involvement for their interpretation: machines operate with quotations or literal reproductions and this distorts the meaning.

iv. **Rhizome and Diagram** (Deleuze and Guattari). The epistemological definition of rhizome is well known and appeals to concepts that explain, many years in advance, the functioning of data science: multiplicity, modification, lines of flight, the calque and replication, connections and associations, as well as the absence of a centre and of a hierarchical model. The same occurs with the Deleuzean notion of diagram, which is chaos and seed, a "possibility of fact" and a "modulation". Both ideas appear to me to be fundamental for thinking theoretically about the form and procedure of *Big Data.*

To conclude: methods of analysis "tend to be concealed, are legitimized as neutral in themselves, as supposedly independent" (Rodríguez 2011: 95), but they are not. The problem lies in that one must know how to grasp those interdependent and transdisciplinary relations, which are often "invisible" (cf. Merleau-Ponty 1979). It is humanists who can do this, because they are the ones who have the competence of *crossed* and *situated* reading, although the task represents an epistemic and academic challenge. I am convinced that these days there is no sense in separating literary criticism – its ideological construction – and data analysis – quantitative and computational methods – although the former deals with the object in a simultaneous order – synchronic and micro – to unravel its principles and its limits, while the latter does so in a chronological order – diachronic and macro – to situate specific literary productions in a historical process that answers to a given social matrix, not exempt from colonial and gender biases. Computational criticism supplies the appropriate set of tools for the theoretical, historical, material and aesthetic knowledge of the literary work, but in turn this science is modified and is augmented with humanistic, philosophical, feminist and decolonial tools. Ultimately, the relationship between literature, big data and artificial intelligence does not only point to *other* forms of knowledge and representation, but to new *crossovers* between the theory and the praxis that create value: social, cultural and academic.

Bibliography

Altamirano, Carlos y Beatriz Sarlo (1991): *Literatura y sociedad*. Buenos Aires: Centro Editor de América Latina.

Archer, Jodie y Jocker, Matthew L. (2016): *The Bestseller Code: Anatomuy of the Block-buster Novel*. New York: St. Martin’s.

Bajtín, Mijail (2012): *Estética de la creación verbal*. México: Siglo XXI.

Berti, Agustín (2015): *From Digital to Analog. Agrippa and Other Hybrids in the Beginnings of Digital Culture*. Nueva York: Peter Lang Publishers.

Badía Fumaz, Rocío (2012): “Muerte del autor y literatura digital”, in *Eikasia: revista de Filosofía*, 44, pp. 113–127.

Blatt, Ben (2017): *Nabokov’s Favorite Word Is Mauve: What the Numbers Reveal About the Classics, Bestsellers, and Our Own Writing*. New York: Simon & Schuster.

Bloom, Harold (2009): *La ansiedad de la influencia. Una teoría de la poesía*. Madrid: Trotta.

Bode, Katherine (2017): “The Equivalence of ‘Close’ and ‘Distant’ Reading; or, Toward a New Object for Data-Rich Literary History”, in *MLQ*, 78, pp. 77–106.

Bolter, Jay David (1991): *Writing space: The computer, hypertext, and the history of writing*. Hillsdale, NJ: Lawrence Erlbaum.

Bootz Philippe and Maria Ines Laitano (2014): “Cross-Reading: un outil de visualisation de close readings”, in *Revue Formules*, 18, pp. 271–288.

Boudieu, Pierre (2002): *Las reglas del arte*. Barcelona: Anagrama.

Brouillette, Sarah (2017): “Neoliberalism and the Demise of the Literary”, in Michum Huehls and Rachel Greenwald (eds.). *Neoliberalism and Contemporary Literary Culture*. Baltimore: John Hopkins University, pp. 277–290.

Côrtes Maduro, Daniela (ed.) (2017): *Digital Media and Textuality. From Creation to Archiving*. Bielefeld: transcript.

Dalton, Craig and Jim Thatcher (2014): “What does a Critical Data Studies look like and why do we care? *Society + Space*”. https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care

Debord, Gay (2005): *La sociedad del espectáculo*. Valencia: Pre-textos.

Delany, Paul, Landow George P. (eds.) (1991): *Hypermdia and Literary Studies.* Cambridge/London: MIT Press.

Deleuze, Gilles (2008): *Pintura. El concepto de diagrama*. Buenos Aires: Cactus.

De Nooy, Wouter (1991): "Social networks and classification in literature", in *Poetics*, 20, pp. 507–537.

Didi-Huberman, Georges (2015): *En la cuerda floja*. Santander: Shangrila.

Dobson, James E. (2019): *Critical Digital Humanities*. Urbana: University of Illinois Press.

Empson, William (1966): *Seven Types of Ambiguity*. New York: New Directions.

English, James and Underwood, Ted. (2016): "Turbulent Flow: A Computational Model of World Literature", in *MLQ*, 77, pp. 345–367.

Espino, Francisco (2020): "Big data, criticometría y el estudio de las literaturas nacionales en la bibliografía crítica: el caso excepcional de la literatura cubana", in *Revista de Humanidades Digitales*, 5, pp. 66–85.

Ferrer, Carolina (2015): "Digital Humanities, Big Data, and Literary Studies: Mapping European Literatures in the 21st Century", in *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 1, pp. 1–11.

Gache, Belén (2006): *Escrituras nómades*. Gijón: Trea.

Gainza, Carolina (2019): *Narrativas y poéticas digitales en América Latina: producción literaria en el capitalismo informacional*. México: Secretaría de cultura / Centro de Cultura Digital.

Gallego Cuiñas, Ana (2019): *Las novelas argentinas del siglo 21. Nuevos modos de producción, circulación y recepción*. New York: Peter Lang.

Gallego Cuiñas, Ana, Esteban Romero-Frías y Wenceslao Arroyo (2020): "Independent publishers and social networks in the 21st century: the balance of power in the transatlantic Spanish-language book market", in *Online Information Review*, 44, pp. 1387–1402.

Gallego Cuiñas, Ana (2021): "La cuestión de la literatura latinoamericana y española en el siglo XXI", in *Novísimas. Las narrativas latinoamericanas y españolas del siglo XXI*. Madrid: Iberoamericana, pp. 11–41.

Gallego Cuiñas, Ana (2022): *Cultura literaria y políticas de mercado. Editoriales, ferias y festivales*. Berlin: De Gruyter.

Goicoechea de Jorge, María (2019): "La literatura digital y los nuevos formatos de la edición literaria", in *Revista de Humanidades Digitales*, 4, pp. 162–186.

Goldstone, Andrew and Underwood, Ted (2014): "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us", in *New Literary History*, 45, pp. 359–384.

Halavais, Alexander (2015): "Bigger sociological imaginations: framing big social data theory and methods", in *Information, Communication & Society*, 5, pp. 583–594.

Haraway, Donna (1995): *Ciencia, cyborgs y mujeres. La reinvención de la naturaleza*. Madrid: Cátedra

Hayles, Katherin (2008): *Electronic Literature: New Horizons for the Literary*. Indiana: University of Notre Dame Press.

Helgueta Manso, Javier (2022): "Antiguos y modernos en la comunicación multimedial de la literatura. Citas, videoreseñas, tuiteros y booktubers", in *Insula*, 907–908, pp. 43–46.

Herrmann, Sebastian et al. (eds.) (2022): *Beyond Narrative. Exploring Narrative Liminality and Its Cultural Work*. Bielefeld: transcript.

Horkheimer, Max y Theodor W. Adorno (19989: *Dialéctica de la ilustración*. Madrid: Trotta.

Jean So, Richard and Long, Hoyt (2013): "Network Analysis and the Sociology of Modernism", in *Boundary*, 40, pp. 147–182.

Jockers, Matthew L. (2013): *Macroanalysis: Digital Methods and Literary History*. Chicago: University of Illinois Press.

Karsdorp, Folgert et al. (eds.) (2021): *Humanities Data Analysis. Case Studies with Python*. Princeton: Princeton University Press.

Kittler, Friedric (2010): *Optical Media*. Cambridge: Polity.

Kristeva, Julia (2002): *Estética I*. Madrid: Fundamentos.

Koskimaa, Raine (2005): “Close reading: hipertextos de ficción”, in Laura Borrás (ed.). *Textualidades electrónicas. Nuevos escenarios para la literatura*. Barcelona: Editorial UOC, pp. 177–191.

Kozak, Claudia (2012): *Tecnopoéticas argentinas. Archivo blando de arte y tecnología*. Buenos Aires: Caja Negra.

Kozak, Claudia (2017): “Esos raros poemas nuevos. Teoría y crítica de la poesía digital latinoamericana”, in *El jardín de los poetas. Revista de teoría y crítica de poesíaLatinoamericana*, 4, pp. 1–20.

Ledesma, Germán (2022): *El susurro de los mercados. Capitalismo financiero y literatura digital*. Rosario: UNR Editora.

Lemercier, Claire and Zalc, Claire (2019): *Quantitative Methods in the Humanities. An Introduction*. Virginia: University of Virginia Press.

Maltz, Hernán (2020): “Discusión sobre sociología de la literatura”, in *Políticas de la Memoria*, 20, pp. 261–271.

Martínez-Gamboa, Ricardo (2016): “Big Data en humanidades digitales: de la escritura digital a la “lectura distante””, in *Revista Chilena de Literatura*, 94, pp. 39–58.

Merleau-Ponty, Maurice. Merleau-Ponty (1968): *The Visible and the Invisible*. Evanston: Northwestern University Press.

Moraña, Mabel (2014): *Bourdieu en la periferia. Capital simbólico y campo cultural en América Latina*. Santiago de Chile: Cuarto Propio.

Moretti, Franco (2016): *Lectura distante*. Buenos Aires: FCE.

Morgan, Nick (2013): “¿Olvidar el latinoamericanismo?: John Beverley y la política de los estudios culturales latinoamericanos”, in *Cuadernos de Literatura*, 34, pp. 18–45.

Olaizola, Andrés (2018): “Bots sociales literarios y autoría. Un aporte desde la retórica digital”, en *Virtuales*, 9, pp. 237–259.

Piglia, Ricardo (2005): *El ultimo lector*. Barcelona: Anagrama.

Piper, Andrew (2017): “Think Small: On Literary Modeling”, in *PMLA*, 132, pp. 651–658.

Piper, Andrew (2018): *Ennumerations. Data and Literary Study*. Chicago: University of Chicago Press.

Piper, Andrew and Eva Portelance (2016): “How Cultural Capital Workds: Pizewinning Novels, Bestsellers, and the Time of Reading”, in *Post45*, https://post45.org/2016/05/how-cultural-capital-works-prizewinning-novels-bestsellers-and-the-time-of-reading/.

Posada, Adolfo R. (2019): *Big Data*: el crítico frente al algoritmo (Lectura distante de la narrativa española editada en el año 2019), in *Colindancias*, 10, pp. 17–40.

Pozuelo Yvancos, José María (1994): *Teoría del lenguaje literario*. Madrid. Cátedra.

Rodríguez, Juan Carlos (2011): *Tras la muerte del aura (En contra y a favor de la Ilustración)*. Granada: Editorial de la Universidad de Granada.

Roig-Sanz, Diana and Reyne Meylaers (2019): *Literary Translation and Cultural Mediators in ‘Peripheral’ Cultures: Customs Officers or Smugglers? (New Comparisons in World Literature)*. New York: Palgrave MacMillan.

Rosseti, Miguel (2014): “A contraluz: World Literature y su lado salvaje”, in *CHUY*, 1, pp. 60–93.

Richards, Ivor A. (2004): *Practical Criticism: A Study of Literary Judgment: A Study of Literary Judgement*. New York: Routledge.

Sádaba Rodríguez, Igor (2020): “El no tan nuevo espíritu del predictivismo: de la estadística moderna al *Big Data*”, in *Revista Crítica Pneal y Poder*, 19, pp. 57–77.

Saum-Pascual, Alex (2022): “Materialidad digital, algoritmos y otras abstracciones modernas”, in Daniel Escandell (ed.). *Escrituras hispánicas desde el exocanon*. Madrid: Iberoamericana, pp. 23–36.

Schiuma, Giovanni and Carlucci Daniela (eds.) (2021): *Big Data in the Arts and Humanities*. London/ New York: CRC Press.

Sierra, Germán (2022): “Caos frio”, in *Insula*, 907–908, pp. 11–14.

Striphas, Ted (2011): *The Late Age of Print: Everyday book culture from consumerism to control*. New York: Columbia University Press.

Tinianov, Iuri (1972): *El problema de la lengua poética*. México: Siglo XXI.

Underwood, Ted (2019): *Distant Horizons. Digital Evidence and Literary Change*. Chicago: University of Chicago Press.

Vanoli Hernán (2019): *El amor por la literatura en tiempos de algoritmos. 11 hipótesis para discutir con escritores, editores, gestores y demás militantes*. Buenos Aires: Siglo XXI.

Vera Barros, Tomás (comp.) (2014): *Escrituras objeto. Antología de literatura experimental*. Buenos Aires: Interzona.

Villanueva, Darío (2022): “Muerte de la literatura, posliteratura, literatura digital”, in *Insula*, 907–908, pp. 2–7.