Regulation & Governance

Regulation & Governance (2022) 16, 156–176 doi:10.1111/rego.12349

Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms?

Raphaël Gellert

Radboud Business Law Institute , Interdisciplinary Hub for Security, Privacy and Data Governance (iHub), Radboud University, Nijmegen, The Netherlands

Abstract

The notion of information is central to data protection law, and to algorithms/machine learning. This centrality gives the impressions that algorithms are just yet another data processing operation to be regulated. A more careful analysis reveals a number of issues. The notion of personal data is notoriously under-defined, and attempts at clarification from an information theory perspective are also equivocal. The paper therefore attempts a clarification of the meaning of data and information in the context of information theory, which it uses in order to clarify the notion of personal data. In doing so, it shows that data protection law is grounded in the logic of knowledge communication, which stands in stark contrast with machine learning, which is predicated upon the logic of knowledge production, and hence, upon different definitions of data and information. This is what ultimately explains the failure of data protection to adequately regulate machine learning algorithms.

Keywords: algorithmic regulation, artificial intelligence, data protection, information theory, machine learning.

> The fact that the concept of knowledge communication has been designated by the word information seems, prima facie, a linguistic happenstance.
> (*Capurro & Hjørland 2003*, p. 344)

1. Introduction: The problem with information, data protection, and algorithmic regulation

The notion of information has existed since the antiquity (see, e.g. Logan 2012, p. 70), yet it has encountered a phenomenal success since around the mid-20th century and the development of information technologies (see, e. g. Chaitin 1977). Even though its contemporary vernacular meaning is associated to the communication of knowledge, its exact meaning remains as vague as its usage is widespread. As Martin (1995) argues, "what is information? (…) Although the question may appear rhetorical, there is a sense in which the answer is that nobody really knows." In particular, the respective definitions and distinctions between information and data remain a matter of discussion (see, e.g. Zins 2007). Maybe, the very success of this notion resides in its lack of clear definition (Capurro & Hjørland 2003)?

It is argued that this very lack of definitional clarity – with a particular focus on the relation between information and data, which is itself also unclear – might be a key – yet overlooked factor as far as the regulation of machine learning algorithms through data protection law is concerned. This is of particular importance given the various shortcomings and criticisms at play.

The development of machine learning technology is an ongoing one, and the numerous societal challenges it raises have long been acknowledged (see, e.g. Pasquale 2015; Committe of Experts on Internet Intermediaries (MSI-NET) 2018). In this context, there is therefore a need for algorithmic regulation. By algorithmic regulation, this paper does not refer to Yeung's analysis of data-driven processes of regulation (Yeung 2018), which Bellanova and De Goede have coined as regulation *through* algorithm (Bellanova & De Goede 2022). Rather, and

Correspondence: Raphaël Gellert, Montessorilaan 10, 6525 HR Nijmegen, The Netherlands. Email: r.gellert@jur.ru.nl

Accepted for publication 28 June 2020.

to keep using Bellanova and De Goede's distinction, it refers to the regulation *of* algorithms. Therefore, as the deployment of machine learning into the fabric of our societies is increasingly pushed and promoted, such pushes are also accompanied with calls for adequate regulation of this technology (see European Commission 2020). In the quest for finding appropriate regulatory mechanisms (see, e.g. Matus & Veale 2022), data protection law has been widely seen as a key regulatory *locus* (see, e.g. European Commission 2020, p. 21). One likely reason for this turn toward data protection is because the notion of information is central both to data protection law, (which defines personal data as "any information relating to an identified or identifiable natural person"), and to machine learning algorithms, which are predicated upon the possibility of extracting information from data, pursuant to the canonical DIKW (data, information, knowledge, wisdom) pyramid, as first conceptualized by Ackoff (1989). It is this centrality that gives the impressions that algorithms and machine learning processes are just yet another data processing operation to be regulated.

However, current debates around the data protection law-oriented regulation of algorithms show that there are still broad interrogations concerning the protection afforded thereby. Beyond interrogations concerning the best way to put in place the specific algorithmic safeguards which the EU General Data Protection Regulation (GDPR) provides for (in particular Art. 22, see among many, Bayamlıoğlu, this special issue),[1] there are broader concerns that the GPDR provisions on algorithmic regulation are inadequate (Bygrave 2017). More radical critics argue that data protection as such is simply ill-suited for such regulatory purpose (see, e.g. de Vries 2016; Gürses & van Hoboken 2017).

This contribution echoes this last strand of criticism insofar as it argues that the under-explored meaning of information might account for such inadequacy. Albeit mobilized in both fields (data protection law and machine learning), it seems that information and its meaning are taken for granted, and in so doing justify the regulatory reach of data protection in the first place. Yet, as it transpires from data protection law discussions (see Bygrave 2014; Hallinan & De Hert 2016), the meanings of personal data, information, and their mutual relation, are far from being clear. It is therefore interesting to observe the similarity of situations between information theory and the personal data notion: both rely upon contested definitions of data and information. In this sense, this contribution can be seen as addressing an interlocking set of issues, since the attempted clarification of the terms from an information theory perspective is considered directly instrumental to the clarification of the notion of personal data under data protection law. These clarifications of information theory terms and of the notion of personal data are then contrasted to definitions of data and information at play in machine learning, which are different, and are premised upon a different relation between information and data. This is what ultimately explains the failures of data protection law to adequately regulate machine learning algorithms.

From a methodological viewpoint, this contribution thus contrasts conceptualisations of information and data stemming from various fields and literatures. It builds upon the widely acknowledged lack of univocity of the notion of information in the field of information theory (see beginning of this introduction) and the scarce echo this issue received in the data protection literature. By combining insights from data protection law scholarship with information theory and science, information philosophy, science and technology studies, and machine learning literature, this contribution can be qualified as pertaining to sociolegal studies. As Cownie and Bradney argue, sociolegal studies are hard to define because of the diverse range of scholarship carried under this label (Cownie & Bradney 2017, p. 41). For this reason, it is characterized by a lack of (formalized) methodology, and by a plurality of approaches and "ways of doing" (see Economic and Social Reaserch Council 1994). One of these "ways of doing" definitely consists in analyzing ideas and concepts from other disciplines in order to incorporate them into legal research and produce new ideas about law (Cownie & Bradney 2017, p. 48). This is clearly what this contribution has endeavored to do. In this sense, it echoes Maniglier's account of transdisciplinarity, which, following Deleuze and Bachelard, he equates to the way in which each discipline provides for a singular and generic way of structuring a problem (or issue) (Maniglier 2019).

The contribution will therefore start by trying to clarify the meaning and mutual relation of data and information in the light of information theory. This is considered crucial because the lack of definitional consensus surrounding the notion of personal data can be traced back to this original definitional quagmire. The contribution will therefore discuss a widely accepted account in information theory, which data protection authors have adopted. The latter neatly separates between data as the digital support of information, which is then mostly a human cognitive process of interpretation and attribution of meaning. Information theory however points to information as a multidimensional concept of information endowed with both cognitive and representational dimensions. This is confirmed by an historical analysis of the concept of data, the emergence of which as a representational device is a happenstance. One should note, however, that such concept of information only makes sense within a broader logic of communication (of information or knowledge), which is at the heart of information theory.

On this basis, it will be shown that a renewed understanding of the notion of personal data is consistent with the information theory-based conceptual exploration of the data-information relation. This then leads to a broader point – going beyond the notion of personal data, which argues that the whole edifice of data protection law is grounded in information theory and in the logic of communication of information which underpins it. This can be evidenced through the GDPR definitions of data flows and data processing operations that echo the communicative logic, and through its safeguards – such as data minimization, purpose limitation, which also echo this logic by aiming at minimizing the amount of data processed.

Turning to machine learning, the present conceptual exploration shows that it embodies dissimilar notions of data and information. Whereas data protection is based on the vernacular idea of information as the communication of information (or knowledge), machine learning is predicated on a different premise, namely the creation of knowledge. This leads to different concepts of data (as an ensemble of features that allow for the abstraction rather than the representation of real-world entities) and information (as the organization of these features in a way that enables the machine learning process to be successful).

On this basis, the contribution discusses one of the GDPR's main provisions for algorithmic accountability: Art. 22. As a confirmation of the conceptual exploration, it shows that some of its intrinsic shortcomings can be traced back to these different definitions of information and data.

This leads to final conclusions on the GDPR's potential for algorithmic regulation. Because the GDPR is a regulatory instrument based on definitions of information enshrined in the logic of knowledge communication, it might simply be inadequate in order to regulate a technology that is based upon a fundamentally different logic: the production of knowledge.

2. Definitional quagmires in information theory and in data protection law

2.1. Information theory and the birth of the modern concept of information

One can argue that the word information is not a novelty, as it can be traced back to Latin and ancient Greek. For Logan, the English word for information appears first in 1386, and is linked to Latin and French, by combining "inform," meaning to give a form to the mind, and "ation," denoting a noun of action. It is therefore about the training, or molding, of the mind (Logan 2012, p. 70). Machlup agrees by stating that in several modern languages, the meaning of the word information comes from the Latin word *informare*, meaning "to put to form" (1983, p. 642). This is in line with the definitions provided by Oxford English Dictionary (OED), which a number of authors take as a starting point for a definition of information (e.g. Machlup 1983, p. 642; Nunberg 1996, p. 108). According to the OED, the verb to inform means to "form (the mind, character, etc.) especially by imparting learning or instruction."[2]

The emergence of the modern concept of information, however, is linked to developments in information technology, information theory and science, and cybernetics following the work of Shannon and Hartley, which focused on the possibility to transmit information between devices (Hartley 1928; Shannon 1948). This mathematical theory of information is concerned with the possibility of storage, the communication to and between machines, and the quantitative measure of information as a mathematically defined entity (Logan 2012, p. 70). If Shannon can be considered "the father" of modern information theory (see Logan 2012, p. 71), his theory of information is limited to mathematical, statistical, and technical problems related to the transmission of signals. In this sense, it is more a theory of communication than a theory of information (see, e.g. Machlup 1983; Burgin 2010). The limitations of this theory soon spurred the need for a semantic theory of information, that is, a concept of information going beyond signals communication, that is, concerning the communication of ideas (Burgin 2010, p. 301). This was deemed particularly necessary in the context of information science, a field and practice descending from and intimately linked to information technologies and information theory (see, e.g. Bates 1999), which nonetheless called for a broader conception of information. Accordingly, information science is "a study of all aspects of information: information processes, properties, functions, relations, systems, etc." (Burgin 2010, p. 10).

The development of information theory as mainly concerned with the transmission of signals and messages has therefore led to the vernacular understanding of information as communicated knowledge (even though a similar definition can be traced back as far as 1450, see, Logan 2012, p. 70). The OED thus also defines information as the fact of "impart[ing] knowledge of some particular fact or occurrence to; to tell (one of something)."[3] For Machlup, information refers "to telling something or to the something that is being told." It is therefore about the transmission and reception of messages (Machlup 1983, p. 660). The modern concept of information is thus intrinsically linked to the notion of messages and messengers (Capurro 2000).

Yet, defining semantic information beyond its vernacular meaning remains a challenge. Whereas the statistical and mathematical definition of information of information is usually considered unproblematic (see, e.g. Logan 2012, p. 70), the same cannot be said about semantic concepts of information, or about a possible overarching concept of information (that would encompass both approaches).The literature has approached information in various ways (not to mention that the field of semantic theories of information is itself divided into various subfields, as shown in [e.g. Adriaans & Van Benthem 2008; Floridi 2008]). These include information as the following: a language structure; a process (i.e. that of becoming informed); a content (such as a specific informing item); a change in a knowledge system; some type of knowledge (e.g. personal beliefs or recorded knowledge); some type of data; an indication; intelligence; lore; wisdom; advice; signals; facts; acts; messages; things; meaning, and so on (see among others, Buckland 1991; Barwise & Seligman 1997; Stonier 1997; Wersig 1997; Flückiger 1999).

2.2. From information to data and their relation

In addition to information, a key term is data. Similar to the definition of information, the definition of information, data, and their mutual relationship is also a matter of controversy. The most eloquent example is probably the work of Zins who compiled a list of 42 such definitions, coming from a panel composed of information scholars (Zins 2007). Zins acknowledges that information and data form the bedrock of information theory and science, yet "the nature of the relations among them is debatable, as well as their meanings" (Zins 2007, p. 479). His broad overview does not succeed in bringing definitional clarity and univocity. In particular, and among the many definitions and relations that he investigates, he insists that data can refer both to sensory stimuli and to a set of signs that represent empirical stimuli. Similarly, information can refer both to the meaning of these stimuli, and to a set of signs which represent empirical knowledge (Zins 2007, p. 487).

2.3. Data protection law and the data-information dyad

In data protection law, the notion of personal data is defined as "*any information relating to an identified or identifiable natural person ('data subject')*" (Article 4(1) GDPR). It is interesting to observe that the notion of personal data is also underpinned by the data-information dyad, and that just as is the case for information theory, nor the definition of the terms or their relation is clear. As Bygrave argues, the "precise meaning of the terms 'data' and 'information' has been largely taken for granted" by lawmakers, and "the terms 'data' and 'information' appearing in such definitions have rarely been analysed systematically" (Bygrave 2010, p. 13). Hallinan and De Hert concur by reminding us that as a result of such lack of clarity, and for lack of better alternatives, some have used the terms as synonyms (Hallinan & De Hert 2016, p. 130). Some scholars such as Manson have lamented this equating of terms (Manson 2009, pp. 26–28).

In his work, Bygrave resorts to information theory in order to make sense of the interplays between information and data in the definition of personal data. On this basis, he argues that data protection law relies upon definitions of data and information that are imported from information theory, and more in particular, from semantic theories of information (Bygrave 2010, pp. 14, 22; see also Nys 2004). Hallinan and De Hert argue that "although this is not specifically clarified in any policy document, it seems a reasonable assertion" (Hallinan & De Hert 2016, pp. 132–133). Relying upon a semantic theory of information makes sense insofar as is clear from its definition, a personal data is an information *relating to* an individual. This is clearly a semantic element, related to meaning.

In his further exploration of the notion of information, Bygrave therefore emphasizes its semantic dimension, which he argues, is mainly a cognitive phenomenon (i.e. the attribution of meaning to information takes place inside the recipient's mind). In doing so, he relies on a popular account of the meanings of information and data, perhaps most famously represented by Floridi's general definition of information (GDI). Following the latter, "information equals data plus meaning" (see Floridi 2010), thereby entailing that information cannot be "dataless" (Floridi 2010, p. 17), or that "data is a necessary constituent of information" (Bygrave 2014, p. 26). Data are thus endowed with a sigmatic/syntactic function, that is, is representational (or is the carrier of information) (Bygrave 2010, p. 23), as opposed to information, which has no materiality, and can only exist through its material support, data. Information is therefore, simply a structure or a process, limited to cognitive activities of knowledge creation (through meaning giving and interpretation), which are solely situated in the receiver's mind, and depends upon the receiver's knowledge (Flückiger 1999).

Accordingly, "the notion of 'data' usually denotes signs, patterns, characters or symbols which potentially represent something (a process or object) from the 'real world' and, through this representation, may communicate information about that thing" (Bygrave 2010, p. 14, see also the references therein). Such understanding of data as being representational is indeed in line with accepted accounts of data, which is often equated to "facts, quantities, or conditions derived from systematic observation or experimentation" (Bygrave 2014, p. 23). In this sense, data has a *representational dimension* (see, e.g. Kitchin 2014), and is a conveyer of information. This is why Bygrave writes that data "typically denotes signs, patterns, characters, or symbols which potentially represent some object or process and, through this representation, can communicate information about that object or process" (Bygrave 2014, p. 23). Information in turn "denotes the semantic content of the data communicated to a person"; it has mainly a semantic and cognitive function (Bygrave 2010, p. 14), that is, to produce changes in knowledge and meaning. In other words, such view purports an understanding of data as "essentially a formalised representation of objects or processes," whereas information is seen as "a cognitive element involving comprehension of the representation" (Bygrave 2010, p. 14). It is in line with the ISO's definition of both information and data. It defines data as "a representation of facts, concepts or instructions in a formalised manner suitable for communication, interpretation or processing by human beings or by automatic means." Information on the other hand is defined as "the meaning assigned to data by means of conventions applied to that data" (see Bygrave 2014, p. 22).

However popular and well spread, these accounts of data and information might be (notably because they provide for easy and ready-to-use concepts), they have nonetheless been the target of criticism within the information theory and science literature. For instance, reducing every representational device (e.g. a tree, stone, sheet of paper, etc.) to data appears as quite a reductive move (see, e.g. Burgin 2010, pp. 120, 198). Equally, the meaning conveyed by data differs according to context so that in some cases, data might not convey any information at all (see, e.g. Hjørland 1998). From this perspective, the efforts of the literature to clarify the notion of personal data only succeed in further emphasizing the lack of clarity that surrounds the notion of information.

2.4. Beyond the purely cognitive dimension: Information as a multi-folded concept

The lack of clarity concerning the notion of personal data can be seen as an opportunity to address the information-data dyad, and simultaneously contribute to the discussion on the definition of personal data.

The account of information discussed in Section 2.3, which characterizes it as being solely a semantic process taking place at the level of the receiver's cognitive activity can be considered as problematic. It is argued that such a cognitive view focusing exclusively on interpretation, relevance, and meaning, obliterates a more material and syntactic view of information itself (Capurro & Hjørland 2003, p. 345).[4] According to Capurro and Hjørland's seminal paper on the concept of information, the main distinction is not between signs and their meaning (i.e. which is how the issue is framed in the data protection law context), but rather between an objective view of information concerned with quantifying it (e.g. Shannon 1948; Kolmogorov 1956), and a more subjective one concerned with meaning (Capurro & Hjørland 2003, p. 396). Yet, and crucially, they do not equate such subjective view of information with the one described in Section 2.3. On the contrary, they argue that this subjective view refers to "information as a sign; that is, as depending on the interpretation of a cognitive agent" (Capurro & Hjørland 2003, p. 396). In this sense they indicate that the semantic dimension of information is not only limited to cognitive (or pragmatic) processes, but also embodies the syntactic dimension of information, which the usual framing of the issue traditionally attributes to data. In this sense, information also includes signs such as written language, drawings, but also a piece of wood, a byte, or even the horse's head in the movie "The Godfather" (see Bygrave 2014, p. 26).

From this viewpoint, information is a twofold concept: it is both the signs that represent an objective reality, and the knowledge brought about by such signs. Such twofold dimension of information is also in line with general, vernacular definitions whereby information is what is contained in the message, and the communication of the message as such (Capurro & Hjørland 2003, p. 373). This is also confirmed by Benthall's information literature review, whereby he argues that as far as semantic information is concerned, "Library and Information Sciences (LIS) has fruitfully analysed the term and discovered that information can be both a process and a thing" [or in our case, more precisely, a sign] (Benthall 2018, p. 49).[5]

Capurro and Hjørland refer to Wersig's threefold definition of information, which encompasses the syntactic level (relation of signs to signs), the semantic level (relations of signs to meanings), and the pragmatic level (relation of signs – or meaning represented thereby – to humans) (see Wersig 1997). As they put it, information has syntactic, semantic, and pragmatic dimensions (Capurro & Hjørland 2003, p. 362). As a matter of fact, this syntactic view of information (i.e. information as a sign or a code) is also quite prevalent within semantic theories of information. This is the case with the objectivist philosophies of Dretske, Devlin, etc. (Dretske 1982; Devlin 1991) for whom there is so-called pure information, or information in itself, which exists independently of any receiver and the interpretative processes that go along. The same logic is at work behind Stonier's notion of "unit of information" (Stonier 1997), Teilhard de Chardin's "noosphere" (Teilhard de Chardin 1964), as well as behind the discovery following which information is constitutive of DNA material as such. This holds true only insofar as one relies upon a syntactic notion of information. Indeed, the nucleotides of a strand of DNA are nothing else but coded signs, yet they are referred to as information (see, e.g. Loewenstein 1999).

In other words, the semantic notion of information encompasses issues of information and meaning given to it, but does not exclude the syntactic dimension of information. What Capurro and Hjørland try to say is that the two dimensions of information are coextensive to one-another. Following this reasoning, even though sign do exist as such, their meaning and the knowledge it leads to is context dependent, and agent dependent. Signs do not make sense without an interpretative process.

2.5. On the relation between data and information

The other lesson that can be drawn from the overview of information herein above, is that the notion of data is notoriously absent from the discussions. Indeed, if both syntactic and semantic functions can be subsumed under the notion of information, where and how does the notion of "data" enter the equation?

In order to briefly answer this question, it might be useful to briefly study the notion of data. As indicated earlier, data is traditionally understood as being purely syntactic, as being representational. As a matter of fact, the etymology of data usually refers to data as being the plural of datum, which literally means "that which is given" (Davies 2015).[6] Beyond such considerations (which comfort the idea that data is syntactical), Rosenberg's historical account of the emergence of data might explain the possible overlap between two concepts (i.e. information and data) relating to signs and representation. In his history of the – etymology – of the word "data," Rosenberg acknowledges that the contemporary meaning of data is very much grounded in "the computer, information technology, and information theory" (Rosenberg 2013, p. 34), thanks to which it gained its current understanding of "data as information in numerical form" (Rosenberg 2013, p. 33). The latter is in line with the accepted definition of data as representational (and more specifically Rosenberg argues, as digital representation).

Following Rosenberg, this largely accepted meaning of data relies upon the meaning acquired by data during the 18th century, which refers to what was given as a result of an argument in the context of the emerging scientific rationality and epistemology, that is, facts determined by experiment, facts given as and resulting from an experiment (Rosenberg 2013, pp. 33–36). This is the reason why Rosenberg argues that when data rose to prominence during the 20th century, "it was already a well-established concept, but it remained largely without connotative baggage." The arrival of computer technology and information theory gave new relevance to the base concept of data as established in the 18th century (Rosenberg 2013, p. 34). In other words, data had already acquired its meaning as a factually, empirically observable and produceable given (and thus representable), but the emergence of information technology will anchor this meaning with a particular technological flavor. One can understand Rosenberg's brief history of data to mean that even though data has etymologically been associated to issues of representation, its current syntactic dimension, in particular in the context of computing technology (i.e. digital data) is more of a happenstance than anything else.[7] In other words, one can argue that the notion of data has highjacked the syntactical dimension of information; at least in the field of computers and other so-called "digital technologies."

However, it results from the previous that is it is not absolutely clear that "data" is the adequate and obvious vocable for designating the syntactic dimension of information. As a matter of fact, even the expression "digital data" is misleading. Accordingly, computer and other digital technologies are concerned with the processing of information, not data. Following Ceruzzi for instance, the computer is foremost concerned with the coding of information in binary form (Ceruzzi 2012, p. x).[8] From this point of view, the byte, which is the most essential unit of code is actually constitutive of information. This is confirmed by Dunn who argues that, "'Computers' store and process digital information, which can be thought of as a series of bits (1, 0)" (Dunn 2008, p. 592). Ceruzzi seems to agree with such a statement, by arguing that coded information (i.e. bits and bytes) is precisely constitutive of information stored in digital form (Ceruzzi 2012, p. 8). A look into the definition of a byte confirms this hypothesis. Accordingly, a byte can be defined as "a unit of digital information" (see Ceruzzi 2012, p. 170). In other words, even the computer, the digital data technology per excellence can actually be seen as an information processing technology, relying upon information coded into binary form, and thereby endowed with the representational features generally attributed to data.

2.6. Information theory, information flows, and the overarching logic of communication

To summarize this overview on the meaning of (semantic) information, one can argue that information is a twofold concept: it is both the signs that represent a reality, and the knowledge brought about by such signs. This is the reason why it was argued that information has syntactic, pragmatic, and semantic dimensions. However, one should keep in mind that such understanding of information is framed within a process of communication. As discussed in the Section 2.1, the modern concept of semantic information, rooted in information technology and theory, is intrinsically embedded into the logic of the communication of knowledge. In other words, one can argue that information refers to signs that are being communicated, and which lead to knowledge through an interpretation process involving the endowment of meaning to information (i.e. for the information to generate knowledge it must have some meaning for the recipient, which comes into play during the interpretation phase).

One of the key aspects therefore lies in the syntactic quality of information, that is, the sign at stake is duly representative of the knowledge it is meant to convey (subject to the interpretation process). This is what Weizsacker expresses in the context of language (but could also be expanded to other types of signs), when he argues that signs (as the syntactic dimension of information) should be univocal, even though he is well aware of what he refers to as the "unavoidable circle between language and information" (von Weizsäcker 1974). This points to the distinction between "word plurivocity" and conceptual univocity, as a characteristic of exact knowledge (von Weizsäcker 1974). In other words, even though signs should be univocal, they do not constitute information in and of themselves. The latter can only take place through a process of interpretation, which corresponds to the semantic and pragmatic dimension of information. Adriaans and Van Benthem also point out to the issue of representativeness of information by underlying its "aboutness" feature: "the information is always about some relevant described situation of the world" (Adriaans & Van Benthem 2008, p. 15). They go on to argue that "a situation carries information about another if there is a stable correlation between the two," as is the case with dots on a radar screen carrying information about the airplanes flying in the sky (Adriaans & Van Benthem 2008, p. 20). In other words, information is crucially *about something*, and is therefore at play in a relation between a receiving situation and a described, or sending information (Adriaans & Van Benthem 2008, p. 20, emphasis in the original). They go on to conclude that the quality of information depends essentially on the reliability of the correlation (Adriaans & Van Benthem 2008, p. 20).

In other words, a communicative approach to information implies among others, that the information is already existing, is "out there" through its syntactic dimension. Whether it leads to knowledge will depend upon the interpretation process (which entails giving meaning to the information received) (cf., semantic and pragmatic dimension). For information to be pre-existing, "out there," has a direct influence concerning the operations that can be performed on information, and hence on the definition of information flows. The American Society for Information Science Technology defines information science as "the generation, collection, organisation, interpretation, storage, retrieval, dissemination, transformation, and use of information" (see, e.g. Griffith 1980). As one can see, this definition of information flows fits indeed the idea of information as part of a communication process designed for such activities as storage and retrieval of information, which are characteristic of information science, and information technology. This is confirmed by Kelleher and Tierney, who argue that early information technology was predicated upon such notion of information and precisely for this reason was characterized by "high volumes of simple operations such as 'select', 'insert', 'update', and 'delete'" (Kelleher & Tierney 2018, p. 8). Mackenzie also argues along the same line, stating that pre-big data standard programming techniques of "sorting, indexing, counting, and searching have built the information retrieval systems that both support and trigger these data flows" (Mackenzie 2013, p. 395).

3. Beyond the definition of personal data: Information theory and data protection

Can any useful insight from this overview of information theory literature be drawn as far as the definition of personal data is concerned? One could argue that in conformity with the overview, the "any information" part of the personal data definition corresponds indeed to the syntactic dimension of information. This is actually confirmed by the Article 29 Working Party's (Art. 29 WP) Opinion 4/2007 on the concept of personal data, which takes the following as examples of information: digital data, but also other types of signs such as drawings, voice recordings, images, etc. (Art. 29 WP 2007, pp. 6–9) (Since data protection applies not only to computer but also to "filing systems," Art. 2(1) GDPR). This confirms the syntactic dimension of information. Further, one can argue that the definition of personal data also includes the interpretative process, which turns a syntactic information into a semantic one, through the "relatability" requirement. When an information relates to a data subject (identified or identifiable), a process of interpretation and meaning generation has obviously been at play so as to uncover such relation. If anything, then, the legal definition of personal data is not based upon a sharp distinction between information and data. On the contrary, it is based upon a subsumption of data and information. The constitutive elements of the notion of personal data are encompassed by the notion of information in both its syntactic and semantic dimensions. This is consistent with the present overview of information theory. It is also consistent with the "inner logic" of the definition of personal data. Contrary to Bygrave or Hallinan and De Hert who rely upon a distinction between data and information, at a very literal level, the definition itself equates both terms (i.e. "personal data **is** information").

If the notion of personal data is consistent with information theory at the definitional level, this entails that the whole data protection edifice would also be consistent therewith. This can be witnessed at various levels. First, the definitions of data flows are nearly identical in information theory/information science and in data protection law. More in particular, it is remarkable to observe the extent to which this definition of information flows within a communicative paradigm – illustrated in Section 2.6 – is similar to the definition of data processing operations in the GDPR. Namely the "recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction" (Art. 4(2) GDPR). The overlap is quasi-perfect, and this is probably not a coincidence. Data protection law emerged in the 1970s as a response to concerns created by the computer (see González Fuster 2014). The legal definition of data processing therefore mirrors closely the information flows that characterized this information technology at the time. This is in line with a quote from Hondius, one of the founding fathers of data protection law, who observed that "most data protection laws have laid down their provisions on the handling of personal information in the same sequence as that followed by computer operations" (Hondius 1975, p. 107). Such quote dates back from the 1970s, a time when computer operations and the coextensive information flows were indeed characterized by the "simple operations" of information science. This is also confirmed by Ceruzzi, who argues that the definition of "data processing" in early computers mainly referred to issues of storage and retrieval of information (Ceruzzi 2003, pp. 47–48). Second, this can also be witnessed at the level of data protection safeguards. Seminal data protection principles such as data minimization, purpose limitation, storage limitation, etc. are typically predicated upon the logic of minimizing as much as possible the quantity of data processing operations taking place. There is an assumption that the fewer data processed, the less harmful it will be. This again, matches the simple data processing operations in line with the logic of information communication.

4. Boundary moments: From information technology to knowledge technology

Following Sections 2.1 and 2.6, information technology (including the computer) is based upon a notion of information embedded into a communicative logic. This observation carries a number of consequences in terms of the data flows at stake: they do fit the idea of communicating and transmitting information (i.e. the information is already "out there"), and thus amount to linear and otherwise simple flows (Kelleher & Tierney 2018, p. 8).

There is a sense however, that if the first computers were indeed concerned with these types of processing activities (and even though such activities are in line with the definition of personal data in the GDPR), contemporary computer and digital devices are capable of much more diverse operations on information. This would suggest that even though information and data are at stake in contemporary processing operations, there is nonetheless a conceptual change at stake. Such conceptual change can be located within the underlying communicative logic of information. As discussed, these non-complex information flows can be explained in terms of communication: knowledge is achieved through the communication of information. This is what information technologies are about. However, Capurro and Hjørland point to a boundary moment, which mobilizes information around a shifting logic. It is this very shifting logic that changes the meaning of information (or so it is contended). This boundary moment is the shift from a so-called information society, to a so-called knowledge society. As they put it, "the terminological shift from information society to knowledge society signals that content, and not information technology, is the main challenge for the economy as well as for society" (Capurro & Hjørland 2003, p. 374). Contrary to the information society and information technology, the knowledge society frames information as "isolated pieces of meaningful data that, when integrated within a context, constitute knowledge" (Capurro & Hjørland 2003, p. 374, see the references contained therein). They are adamant that this other "semantic concept of information, located between data and knowledge, is not consistent with the view that equates information (management) with information technology." The reason being that, contrary to the information society and information technology that are predicated on the communication of information, this knowledge society (which, following Rifkin (2000), they situate in the 1990s) is based upon the production and creation of information (Capurro & Hjørland 2003, p. 374), something potentially radically different than the "mere" communication thereof.

These various, simultaneous, and competing definitions of information can be seen at play in the computer; which itself can be seen as a boundary technology, featuring both definitions. As indicated previously (see Section 3), the computer has traditionally been referred to as an information technology (notably following the work of Berkeley (1949)). However, Haigh points to the ambiguous status of the computer as an information technology. More in particular, he argues that information technology includes "everything from ancient clay tablets through scrolls and books to communication systems of all kinds, newspapers, street signs, libraries, encyclopaedias" etc. (Haigh 2011, p. 432). Yet, "treating information technology and computing as interchangeable concepts generally makes [him] a little uneasy" (Haigh 2011, p. 432). Such uneasiness can be traced back to the dual status of the computer, which was first conceived as a "super calculating machine," before it was realized that such calculative capabilities could also be turn the computer into an information processing device (especially for administrative purposes at first) (Agar 2003; Haigh 2011, pp. 433–440). This dual status of the computer is precisely what makes it an information technology but also a knowledge technology. As it is argued, the process of information creation that is characteristic of the knowledge society is precisely what machine learning (and other statistical technologies relying upon databases such as data analytics and mining) is about. This can also be observed in evolutions in database technologies, since "computer technology" is predicated upon databases (see, e.g. Ceruzzi 2003, 2012). Early computers were based upon so-called "hierarchical databases." One can argue that these databases store "simple and linear information," which is why information stored therein is referred to as "records" (Manovich 1999, p. 81). These are perfectly suited for the information retrieval tasks of information science and technology (Colonna 2013, p. 322), and espouse the idea of pre-existing information. Developments in database technology will lead the way to so-called “object-oriented databases.” Contrary to hierarchical databases, the information stored in these databases is complex and non-linear, which is why it is referred to as “objects” (Manovich 1999, p. 81). These new databases allow for the deployment of data mining, knowledge discovery in databases (KDD) and machine learning, and the knowledge production that characterizes such practices (see Frawley *et al.* 1992).

In other words, increasing computational and statistical power, and advances in database technology in the 1990s have led the way to machine learning, the goal of which is not to communicate data, but to learn therefrom and in doing so, to create new knowledge (as opposed to merely communicating pre-existing one) (see, e.g. Chisholm 1989; Kelleher & Tierney 2018). This shift of the computer from an information technology to a knowledge technology is also confirmed by Mackenzie, when he argues that the advent of big data has led to a change in programming practice: machine learning, which according to him embodies “different forms of programming thought and practice” (Mackenzie 2013, pp. 394–395).

The next section of this contribution will therefore explore the meanings of data and information in the context of knowledge production and technology, that is, machine learning.

5. From communication to learning: A different understanding of data and information in the context of machine learning

This section will explore the meaning of data and information in the context of machine learning. It should be clear that even though the focus of the present contribution is machine learning, many of the present findings can be adequately applied to a number of practices that partially overlap with machine learning, such as data mining, data analytics, KDD, visualization, pattern recognition, etc. (see Colonna 2013; Schutt & O’Neill 2013). In sum, to contemporary practices that are based upon probabilities, and statistical learning in databases (see Fisher 1925).[9]

5.1. Data: From representation to abstraction

Going into the machine learning literature, one learns that in its most basic form a data is “an abstraction of a real-world entity (person, object, event)” (Kelleher & Tierney 2018, p. 39). *Prima facie*, this is very similar to the definition of data analyzed in Section 2.5, whereby data represents something from the “real world.”

However, the overlap is not as full as it appears. In the context of machine learning, data are abstractions of real-world entities not because they are signs that represent such entity, but because they are an ensemble of features or attributes, which, put together, will allow for a representation of such entity. In other words, a single data represents only a single feature of the real-world entity at stake. And in order to correctly represent it, a number of such features/attributes/data are needed. *This is a crucial difference*. And this is exactly what Kelleher and Tierney mean when they say that “each entity is typically described by a number of attributes” (Kelleher & Tierney 2018, p. 39). They give the example of a book that can be represented through a number of features such as: author, title, topic, genre, publisher, price, date, word count, number of chapters, number of pages, edition, etc. In other words, the traditionally accepted understanding of data at play in data protection law and in information in a communicative context would argue that the word “book” is information about the specific book at hand, through the syntactic, semantic, and pragmatic dimensions of information. Crucial to this process is the quality of the correlation (between the sign and what it stands for) (cf., Adriaans & Van Benthem 2008, p. 20, *supra* section 2.6). As opposed to this understanding, machine learning puts forth a notion of data as attribute (or feature). Contrary to that, machine learning is less interested in the quality of the correlation between a sign and the entity it represents than in the possibility of adequately representing (i.e. defining) such entity through a number of signs, or features. This is what Mackenzie says when he argues that in the context of machine learning, “people are materialised as a bundle of features” (Mackenzie 2013, p. 398). The key then, is to represent individuals (or any entity subject to a machine learning process) with the most appropriate features/attributes (Kelleher & Tierney 2018, p. 47).

In other words, there is a shift from data as representation to data as abstraction. Again, Kelleher and Tierney confirm this by explicitly stating that “data are generated through a process of abstraction” (Kelleher &

Tierney 2018, p. 46). Rather than being a sign representing the real-world entity as such, a data is the representation of a real-world entity through the reduction (or abstraction) of such entity to a number of features that are deemed sufficient in order to adequately represent it. So, even though data refers to signs in both cases (meaning that each data also represents the feature it stands for), one can see that they are nonetheless endowed with different meaning (a sign that represents as such, versus a sign that represents only one feature of a bigger ensemble of features).[10]

5.2. From data to information, and hence, from communication to learning: Another meaning of meaning

Contrary to information theory and literature, machine learning specifically conceptualizes the relationship between data and information. Such conceptualization emerges from Ackoff's seminal article, which introduced the so-called data, information, knowledge, wisdom (DIKW) pyramid (Ackoff 1989), which underpins the machine learning process (see Fig. 1) (see, e.g. Kelleher & Tierney 2018, p. 56).

In this context, it has been traditionally accepted that information is "meaningful data." There are endless variations of such formula to be found in the literature. Laudon and Laudon for instance argue that "information is data that have been shaped into a form that is meaningful and useful to human beings" (Laudon & Laudon 2006). Similarly, (and among many others), Jashapara defines information as "data that is endowed with meaning, relevance and purpose" (Jashapara 2005). In other words, in machine learning the link between data and information seems to be the addition of meaning to data.

Yet, what exactly is meant by such "meaning"? In the context of information theory meaning is crucial to the interpretation process at play in the semantic and pragmatic dimension of information (cf., Section 2.4).This vernacular use of "meaning" mostly takes place in the recipient's mind: through a cognitive process of interpretation, the information might make sense (i.e. have some meaning) for the recipient, and thereby lead to the communication of knowledge (Rowley 2007, p. 175, see also the references contained therein).

However, it can be argued that the "meaning of meaning" at play in machine learning and in Ackoff's DIKW pyramid is radically different. For a start, Ackoff is the first to acknowledge that the difference between data and information is functional and not structural (Ackoff 1989). As a matter of fact, Ackoff argues that data and information are one and the same thing (Ackoff 1989, p. 1). As he puts it, "data are symbols that represent the properties of objects and events". Information equally "represents the properties of objects and events." The only difference between the two then lies in the fact that information represents properties of real-world entities in a more compact and useful way than data (Ackoff 1989, p. 1).

Rowley provides probably the most enlightening lines concerning the "meaning of meaning." She argues that meaning can be understood in two ways. In the vernacular understanding of the word, meaning is located in the recipient's cognitive schema, and information can as a result only reside in the recipient's mind (Rowley 2007, p. 175). However, meaning can also be understood in a different, functional, way, which is the way Ackoff intended it to be understood in the first place. From this perspective, meaning derives from the organization and structuration of data in a dataset/database. As Rowley puts it, the organization and structuration of the data, "lends the data relevance for a specific purpose or context, and thereby makes it meaningful, valuable, useful, and relevant"

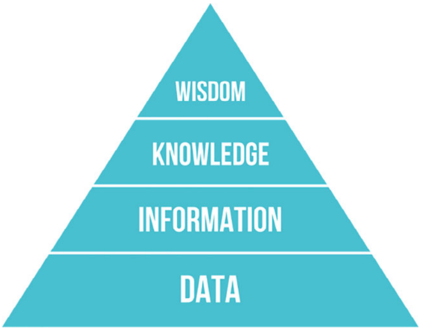


Figure 1 The data, information, knowledge, wisdom pyramid. https://ccsearch.creativecommons.org/photos/00c3863e-c1c1-4f40-bd1f-f055f97eb148

(Rowley 2007, p. 174). In other words, she argues that "structuring data according to a schema that has meaning and relevance for an individual, community or task, endows meaning, or perhaps the potential for meaning" (Rowley 2007, p. 174). It is this functional distinction which justifies the fact that information exists not only in the mind of the receiver, but also in the information systems that process it (Rowley 2007, p. 175). In other words, meaning just means that the symbols at stake (whether they be called data or information) have been structured and organized in a way that is meaningful, that makes sense for the purpose of the processing. This is a very different meaning of meaning.

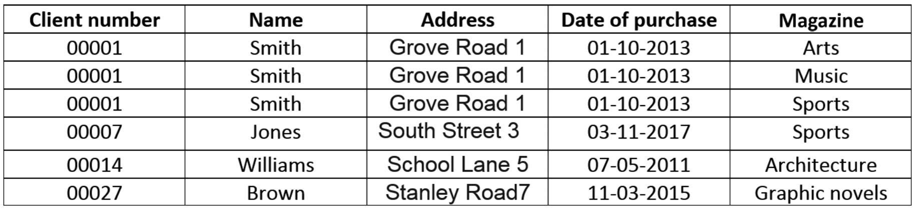
In order to make things more concrete, one can take an example from Adriaans and Zantinge's seminal book on data mining, which provides an interesting example of the process of the transformation of data into information (Adriaans & Zantinge 1996, p. 39).[11] As they argue, the whole point of machine learning is to find new useful patterns in a dataset. It is for this reason that the dataset must be arranged and cleaned, precisely in order to facilitate the pattern finding process (Adriaans & Zantinge 1996, p. 39).

Adriaans and Zantinge use the concrete example of a publisher who wants to set up a direct marketing exercise. For this reason, they need to find interesting clusters of clients, so that their advertising is as targeted and efficient as possible. They therefore need to find new patterns and correlations in order to constitute the clusters (and thus ask questions such as: "what is the typical profile of a reader of a car magazine?" or "is there any correlation between an interest in cars and an interest in comics?") (Adriaans & Zantinge 1996, p. 39).

They start with an original dataset, which is based on data they have managed to gather from the publisher's invoicing system. They therefore create a dataset with the following data/features/attributes: client number, name, address, date of subscription, type of magazine (see Fig. 2). In other words, this "raw" dataset contains a number of data, or attributes, that is, a way of representing each magazine buyer. The question however, is whether such data/features enact a representation that allows for the discovery and learning of useful patterns, that is, whether this data is meaningful for the purpose of the processing operation. If yes, then one can argue that the data is already constitutive of information, without any further need to "make it meaningful." However, Adriaans and Zantinge further explore the process of transformation of data into information.

Beyond the necessary step of cleaning the data of any inaccuracy (e.g. misspelling, false information provided, etc.), Adriaans and Zantinge argue for two essential steps: enrichment of the data, and coding of the data (the latter being strictly speaking, the process of transforming data into information) (Adriaans & Zantinge 1996, p. 42). Some additional data may be needed in order to uncover the patterns. In the case at hand, such data includes the date of birth, annual income, amount of credit, and whether they own a house or a car.[12]

Next, coding is the stage of creative transformation of the data (i.e. structuration and organization) in order to make it meaningful, that is, in order to learn therefrom. In the case at hand, they argue that the data at stake is way too detailed for any meaningful correlation to be drawn (Adriaans & Zantinge 1996, p. 44). They take the example of dates of birth, which are too detailed: an age class on a 10 years basis is amply sufficient (e.g. 10, 20, 30, etc.). Similarly for addresses: the zip (regional) code suffices (as opposed to the full address). As they put it, there could be millions of different addresses, which is much too detailed for the present purpose. They therefore decide to recode the addresses into 4 area codes. They also decide to divide the income by 100, thus creating a same order of magnitude as for the age class, making it easier to compare. The same goes for the credit. They also decide to convert car and house ownership from a "yes-no" to a "1–0." Finally, they convert the purchase data into months starting from 2011 (in our derived example). So, a purchase made on 1 January 2011 is equal to 1, and a purchase made in December 2012 is equal to number 24. This allows to perform analysis on time data, and in particular, on general time dependencies (the situation would be different for instance if one were interested in



Client number	Name	Address	Date of purchase	Magazine
00001	Smith	Grove Road 1	01-10-2013	Arts
00001	Smith	Grove Road 1	01-10-2013	Music
00001	Smith	Grove Road 1	01-10-2013	Sports
00007	Jones	South Street 3	03-11-2017	Sports
00014	Williams	School Lane 5	07-05-2011	Architecture
00027	Brown	Stanley Road7	11-03-2015	Graphic novels

Figure 2 Original dataset.

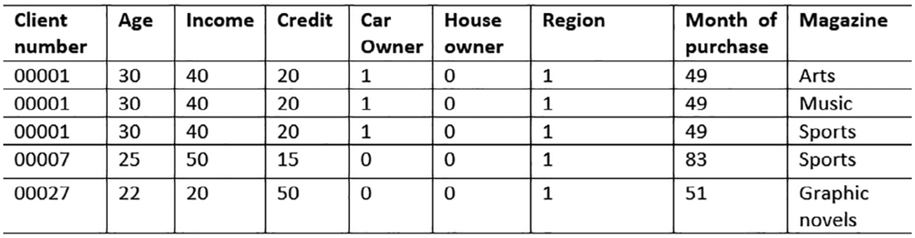
particular buying patterns at specific dates such as Christmas or Easter [Adriaans & Zantinge 1996, p. 45]). Again, the choice of structuration and organization of the data is also dependent on the purpose (the patterns one wants to uncover). In this case, the goal is to find relations between readers of different magazines (i.e. classes of products) rather than between a product and the moment of subscription (Adriaans & Zantinge 1996, pp. 44–45). This can be seen in Figure 3.

However, as it stands the data set is still not fully adequate for the machine learning operation at stake. At present, the dataset is still structured along individual transactions. But it would be more useful to have an overview of all the magazine subscribed by each reader, and thus, to create one entry per reader. The subscription to each magazine is represented in a binary way, by way of "1"or "0" (Adriaans & Zantinge 1996, p. 48). This gives way to the following and final dataset (Fig. 4).

The present author retains this example from Adriaans and Zantinge to be an extremely informative one insofar as it proves the point on the "meaning of meaning" (and more broadly on the difference between data and information) in the context of machine learning algorithms. In line with Ackoff and Rowley, the distinction between data and information is merely a functional one. Information indeed appears to "merely" be data that has been structured and organized in a way that is meaningful to achieving the purpose and goals of the machine learning task at hand. As one can see, the operation of giving meaning has little to do with the interpretative and cognitive processes that are at play in the context of information as knowledge communication. In both cases, data and information are one and the same thing: symbols that are representative of specific features of real-world entities insofar as they are part of a bigger ensemble of such features. They key difference then, is whether such ensemble of features allows for an abstraction of the real-world entity that lends itself to the learning goal of the machine learning process at hand. This also confirms Rowely's point that Ackoff's pyramid was intended in the first place from an "information systems perspective" (Rowley 2007, p. 166).

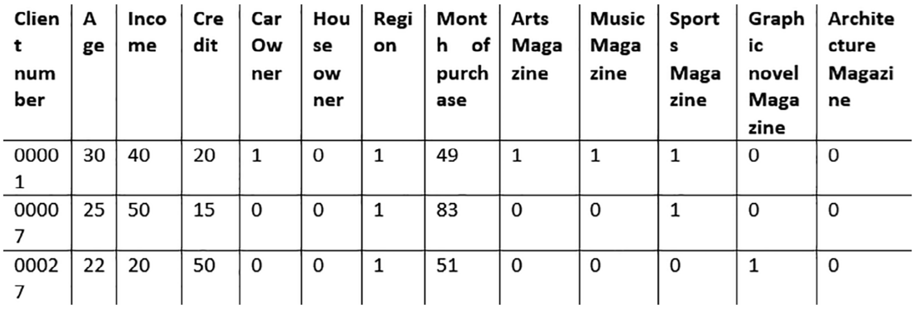
6. Data and information in data protection and machine learning: Different definitions coextensive to different practices

One can draw a number of points on the basis of these various definitions. First, one can see that the relation between data and information is dissimilar. In the data protection law/information technology context, data is a mere happenstance, the role of which is unclear. Further, the focus of information theory, information science,



Client number	Age	Income	Credit	Car Owner	House owner	Region	Month of purchase	Magazine
00001	30	40	20	1	0	1	49	Arts
00001	30	40	20	1	0	1	49	Music
00001	30	40	20	1	0	1	49	Sports
00007	25	50	15	0	0	1	83	Sports
00027	22	20	50	0	0	1	51	Graphic novels

Figure 3 Modified dataset.



Clien t num ber	A ge	Inco me	Cre dit	Car Ow ner	Hou se ow ner	Regi on	Mont h of purch ase	Arts Maga zine	Music Maga zine	Sport s Maga zine	Graph ic novel Maga zine	Archite cture Magazi ne
0000 1	30	40	20	1	0	1	49	1	1	1	0	0
0000 7	25	50	15	0	0	1	83	0	0	1	0	0
0002 7	22	20	50	0	0	1	51	0	0	0	1	0

Figure 4 Final dataset.

and information technology is precisely… information! There seems to be little room for the concept of data. Contrary to this, machine learning is theorized upon a clear and foundational distinction between data and information. Even more so, data and information are encased into a hierarchical relation (first data, then information), along the lines of the DIKW pyramid.

Second, even if they differ subtly, we see that data and information are endowed with different meanings in the two fields. In the first case, data exist as a sort of overlapping concept for the syntactic dimension of information and have a reprensentational function. Information, is both syntactic and semantic, the main purpose of which being the communication of knowledge, which is possible through a cognitive interpretation process taking place in the recipient's mind, giving meaning to data, and thus allowing for knowledge to be communicated. In stark opposition, data in an algorithmic/machine learning context are also signs, but the point of these signs is less to represent in a linear way (i.e. relation between what represents and what is being represented), than to represent real-world entities through what machine learning refers to as a concept, that is, as an ensemble of features, which taken together represent the "idea" (or concept) of the real world entity at stake (see, e.g. Alpaydın 2016, p. 75). Information in turn does not refer to any cognitive process, but rather to the way in which these features are structured. One can argue that data is transformed into information when the chosen ensemble of features of the concept are meaningful for the overall goal of the processing operation (which is to make predictions on the basis of available information).

In any case, this last insight only makes sense because of the overarching logic of which these definitions take part. As indicated in Section 2.6, information theory (and hence data protection) is embedded into a logic of knowledge communication (cf., information technology), whereas machine learning's definitions stem from a logic of knowledge creation (cf., knowledge technology). These differences in terms of definitions and overarching logics are not merely analytical. On the contrary, these diverse definitions are embedded into competing practices, thereby showing how such diverging discursivity is itself coextensive to distinct sociotechnical practices (see, e.g. Latour 2005). In other words, given definitions of data and information also partake of specific information (and data) flows and processing practices, which, it is argued, are not identical. As discussed, as a means for information communication, information technology is predicated upon a high volume of nonetheless simple processing operations. Algorithms and machine learning on the other hand, are predicated upon much more complex operations, which are precisely rendered thinkable because of the definition of data as abstractions, and because of the overarching logic of knowledge production. Such complex operations have to do with the production of new knowledge, which in this context takes place through learning – from data. That is, from inferring new information from existing one.

7. Information theory-based data protection: A sufficient basis for algorithmic regulation?

In the context of the search for algorithmic regulation principles, the differences observed herein above cast a serious doubt as to data protection's capacity to meaningfully regulate machine learning algorithms. Crucially, it is argued that data protection law does not give due account of the multiple data processing operations that are at stake in order to achieve such knowledge production, precisely because it is based on a notion of information that does not make room for learning from information, but only for the communication of such information. It is probably this observation that has inspired a number of radical critics of data protection's algorithmic regulation provisions.

In their paper on "privacy after the agile turn," Gürses and van Hoboken precisely argue that data protection is ill-suited to an agile mode of production of software, that is, to a predicament where every software, and more generally, every form of digital functionality is transformed into a data intensive machine learning product (Gürses & van Hoboken 2017, p. 597). They argue that data protection and its traditional safeguards (such as consent or purpose limitation) assume that digital services are based upon "the planned and up-front development" of software, rather than the real-time iterative optimisation that the shift to machine learning allows for (Gürses & van Hoboken 2017, pp. 591–592; see also, Overdof *et al.* 2018). In other words, data protection is better suited for the simple, "linear" flows of information highlighted above, as opposed to the real time optimisation of the digital that the process of learning allows for. This is the very reason why they argue that data protection regimes are concerned with "the governance of information collection and processing activities" (i.e. constraints upon information flows), but they are silent as to the specificities of machine learning (Gürses & van Hoboken 2017, pp. 579–581). On this basis, they question the reasonableness of treating information flows as the central concern for the protection of privacy and personal data in a machine learning context (Gürses & van Hoboken 2017, p. 597). For them, it is clear that a new regulatory framework, more attuned to the specificities or machine learning is needed, even though this constitutes a considerable challenge (Gürses & van Hoboken 2017, p. 598).

De Vries seems to be perfectly in line with this type of argument, and puts it in a much more explicit way. According to her, the process of regulating information flows from a data protection perspective is characterized by "a silence as deep as the one in the eye of a hurricane" as far as "the particular of knowledge production through automated inference algorithms" is concerned (de Vries 2016, p. 420). That is not to say that machine learning algorithms completely escape the regulatory gaze of data protection (de Vries 2016, p. 421). Simply, when such practice is translated in the data protection legal language (thus having to do with the lack of control over one's data, the security of the storage, the disproportionality of the means and the aims pursued, the lack of transparency of the processing, etc.), something is lost in the process (de Vries 2016, p. 421). These observations lead de Vries to the radical conclusion that given the way in which data protection misses out on several key aspects of machine learning algorithms (or as she puts it the specific mode of truth production of algorithms is currently ignored), there is simply to date, no "informational right that has ever encountered a profiling algorithm as a profiling algorithm" (de Vries 2016, pp. 418, 422). Wachter and Mittelstadt have voiced a similar criticism, arguing that data protection law focuses quasi exclusively upon "input" data, whereas the core of machine learning consists in transforming input data into "output" data (Wachter & Mittelstadt 2019).

One could argue that mobilizing a literature that points at certain data protection shortcomings concerning the regulation of machine learning, does not entail *ipso facto* validation of the present hypothesis. After all, data protection faces many issues for regulating algorithms, and not all of them are amenable to the underlying definitions of data and information. This is the case as far as the material and personal scope is concerned. The literature has long argued that machine learning-based practices are problematic because they often rely upon anonymous data set (see, e.g. Barocas & Nissenbaum 2014). Similarly, algorithms often have harmful consequences not for individuals as such but insofar as they are constitutive of a group. The latter, however, is not encompassed in the personal scope of data protection law (see, e.g. Mittelstadt 2017). While this is not denied, this contribution does not address scope-related issues.

Furthermore, the legitimacy of data protection's pretence to regulate machine learning can be questioned from various perspectives. On the one hand, some might argue that the scope of data protection law has been steadily expanding over the past decades (as a result of the increasing digitisation of society), and it has come to include a number of issues that go well beyond its original ambition to regulate the processing of personal data. These include journalistic and freedom of expression issues (see Van Hoboken 2014), competition issues (see Graef *et al.* 2018), and, possibly, machine learning as well. From this perspective the inadequate regulation of machine learning stems from the undue stretching of the flexible and open-ended data protection rules. This is reminiscent of Diver's seminal work on the optimal precision of rules (which should be transparent, congruent, and simple) (Diver 1998). On the other hand, one can argue the opposite stance. Namely, that the purpose of data protection has been to address profiling and social sorting since its very inception. This appears quite clearly from early data protection literature (see, e.g. Rule 1973), or from policy documents such as the influential so-called "Ware Report," which underpinned the adoption of a number of early data protection statutes (US Department of Health Education & Welfare 1973). This is maybe the reason why, during the reform process from the 1995 Directive to the GDPR, the European institutions stated that the "core" data protection principles (such as data minimization, purpose limitation) were still valid and future-proof and therefore did no need to be modified (European Commission 2010, p. 3). However, it would be a stretch to equate the profiling practices from the 1960s and 1970s, with the advanced profiling algorithms that exist nowadays. As seen, the DIKW pyramid which underpins machine learning has only been formalized in the late 1980s, and the technology only gained prominence around more or less the same time (Frawley *et al.* 1992; Fayyad *et al.* 1996). This is the reason why, for the relevant literature at the time, the main danger was the possibility to collect individuals' records, and to constitute dossier about them (see, e.g. Packard 1964; Miller 1971). The fact that the literature highlighted the constitution of records as one of the main dangers is perfectly in line with the database technology at the time, which only allowed for limited information retrieval operations on what was precisely known as records (cf., *supra* Section 5). These "records-based profiles" are something very different from what machine learning does today, but which is perfectly in line with safeguards that aim foremost at minimizing the data flows.

Be it as a result of its expanding scope, or as a way to try and uphold its original regulatory goals, modern data protection statutes such as the GDPR have included novel – mostly formal and procedural – provisions meant to address machine learning, such as Art. 35 on data protection impact assessments (see, e.g. Council of Europe 2017), or Art. 22 on machine learning-based automated decision-making. These provisions, however, are not without shortcomings. It is argued that these shortcomings are bound to exist precisely because they illustrate the main point of this contribution. Namely, that data protection is unable to adequately regulate machine learning insofar as both practices are underpinned by different concepts of data and information. This is particularly clear in the case of Art. 22 GDPR. Without entering into the technicalities of Art. 22, the present contribution solely focuses on the safeguards it contains, which are enshrined in Art. 22(3). Accordingly, if a machine learning process falling under the scope of Art. 22 is allowed to take place, it should at least feature the right to obtain human intervention on the part of the data controller; it should allow the data subject to express his or her point of view (which includes the right to receive an explanation of the decision); and it should allow to contest/challenge the decision.

Two observations can be made as far as these safeguards are concerned. First, such provision is extremely open-ended, and its level of prescription in terms of methods is minimal/non-existent, hence an intense debate in the literature on the most adequate transparency, explainability methods (see Kaminski's (2019) overview of the debate). There is however a more fundamental observation to be made, which is more in line with the argument of the present paper. Namely, and to re-quote de Vries (2016, p. 422), does Art. 22 "encounter a profiling algorithm as a profiling algorithm?" According to Lehr and Ohm the answer is a definitive no (Lehr & Ohm 2017). Their main argument is that current debates on algorithmic regulation have crucially treated machine learning algorithms as an abstraction, thus overlooking what an algorithm actually is and how it is made. The latter results in sub-optimal regulatory solutions (Lehr & Ohm 2017, p. 655). More specifically, they provide for a taxonomy of iterative data flows involved in the machine learning process, namely problem definition, data collection, data cleaning, summary statistics review, data partitioning, model selection, model training, and model deployment (Lehr & Ohm 2017, p. 655). The first seven steps are concerned with the "construction" of the algorithm as such (a step they refer to as the "playing with the data" stage). The eighth step is concerned with the deployment of the machine learning algorithm in "the real world," as a decision-making tool (a stage they refer to as the "running model" stage) (Lehr & Ohm 2017, p. 655). On this basis, they argue that the algorithmic issues and safeguards currently discussed and enshrined in the GDPR only address the later stage of algorithm deployment.

One can take the issue of discrimination which is crucial in the context of algorithmic decision-making (see, e.g. Council of Europe 2020). The Art. 29 WP Guidelines on Automated individual decision-making and Profiling make it clear that discrimination counts as one of the effects on the data subject that would trigger the application of Art. 22 (Art. 29 WP 2018, pp. 21–22). The Guidelines seem to implicitly argue that discrimination stems from "errors or bias in collected or shared data or an error or bias in the automated decision-making" (Art. 29 WP 2018, p. 27). This can be tackled through frequent assessments and audits of the data sets, which would allow to "check for any bias," as well as to prevent prejudicial elements such as over-reliance on correlations, or the inaccuracy and/or relevance of the profile created (Art. 29 WP 2018, p. 28). Lehr and Ohm argue that addressing algorithmic discrimination in such a way is exclusively "data related." For instance there is bias because the collected data fits into pre-existing biases, or because the data is non-representative, etc. (Lehr & Ohm 2017, pp. 703–704, and the references therein). As they argue, it is no wonder that the issue of bias is data-related. The is nothing in the technical details of an algorithm that will, as such, make it perform better or worse on certain type of groups. Thus, the critical factor is indeed the data fed to the algorithm (Lehr & Ohm 2017, p. 704). However, this widely spread argument does nonetheless miss a crucial point, which has nothing to do with the data as such, namely the issue of overfitting. In a nutshell, overfitting takes place at the training stage. If the algorithm model fits the training data too-well, there is a chance that it will be difficult to generalize, because it is too tailored to the training set (this is known as the free lunch theorem, namely that there is always a balance to be struck between the accuracy of the algorithm and its capability at reproducing its result on any data set, i.e. to generalize). There is therefore a chance that an algorithm that overfits its training data will produce discriminatory results, even though the initial data set is completely bias-free (provided there exists such a possibility) (Lehr & Ohm 2017, p. 704).

This is just but one examples that illustrates that current data protection-based algorithmic regulation provisions can be traced back to a "linear" notion of information and data flows that has little to do with algorithmic definitions of information and data, and in particular with how they learn.

8. Conclusions: Algorithmic regulation: From the regulation of information technology to the regulation of knowledge technology

The previous section highlighted some serious criticism concerning the ability of data protection law to adequately regulate machine learning algorithms. More in particular, these criticisms can be seen as a validation of one of the main points of the present contribution. Namely, that data protection as a regulatory framework for information and communication technologies (ICTs) – at the center of which, the computer – espouses the definitions of information that are associated thereto. In a way, these discussions allow to "square the circle" of the multiple, interlocking, and overlapping issues concerning data protection and information theory.

This contribution started by trying to make sense of the notion of personal data in the light of information theory. Such an endeavor is not self-evident since the field itself portrays notions of information and data as multiple and equivocal. That being said, the exploration of the meaning of data, information, and their articulation allowed to refine the understanding of personal data. Instead of neatly distinguishing between data and information as two separate concepts, information theory points to information as a multidimensional concept endowed with both cognitive and representational dimensions. The notion of data is shown to be a "mere" happenstance. One can therefore argue that in the legal definition of personal data, data **is** information. This is in line both with a literal reading of the GDPR definition (art. 4.1 GDPR), and with the present overview of information theory. Grounding the notion of personal data and data protection in information theory also frames it within a specific logic: that of the communication of knowledge. This is confirmed by the definition of data processing operations and data flows, and the types of safeguards at play. Ultimately, it is this grounding of data protection law in definitions of data and information stemming from information theory that can – at least partially – explain the inability of data protection law to meaningfully regulate machine learning algorithms. The latter are indeed predicated on different definitions of data and information (an ensemble of features that must be organized in a meaningful way in order to learn therefrom). In other words, from the perspective of the regulation of machine learning through data protection law, resorting to information theory not only provides for conceptual clarifications such as those pertaining to the notion of personal data. It also allows to better grasp the way data protection relates to and has attempted to regulate information and computer technology, and thus, to also better understand its shortcomings in relation to machine learning which the literature has highlighted (cf., de Vries 2016; Gürses & van Hoboken 2017; Wachter & Mittelstadt 2019). This in turn, could be helpful in devising ways to enhance the regulation of machine learning technology, which remains sub-optimal at present (see, e.g. Council of Europe 2020).

Therefore, the exploration of the different meanings of data and information at stake in data protection law and in machine learning and their different yet mixed-up meaning, point to the need for a set of new regulatory principles. These would go beyond data protection insofar as they would espouse the specificities of the information/data flows and processes at stake in machine learning. Lehr and Ohm's taxonomy of data flows is a useful example of this type of proposition and theoretical intervention. However, it just a first and incomplete step, as the authors themselves acknowledge it (Lehr & Ohm 2017, p. 676).

Just like Hondius argued that data protection law was fashioned alongside computers' data processing sequences, one useful way forward can be to model algorithmic regulation on the basis of the DIKW pyramid, which is a summary of the step involved in the machine learning process. This could be a way to address de Vries' lament concerning the absence of legal principles that "encounter a profiling algorithm as a profiling algorithm." What such regulatory regime would look like, and which exact legal framework it would build upon (e.g. data protection, anti-discrimination, consumer protection, or even legal frameworks regulating scientific research, which already include anonymous data in their scope) goes beyond the scope of the present contribution. However, as elements for further discussion one can hypothesize that it could include some of the following elements.

The possibility to anticipate the type of profiles that may be constructed and applied to us (Vedder 1999; Hildebrandt 2006; Custers 2018; Wachter & Mittelstadt 2019). One could also think about requirements pertaining to the structure of the dataset, or substantive requirements concerning the way the analytics should be performed on the data. Furthermore, one could even envisage a radical re-interpretation of the notion of personal data in the light of machine learning definitions of information and data. These are mere suggestions. In any case, they do point to the need for a shift from the regulation of information technology to the regulation of knowledge technology.

Acknowledgments

This contribution is based on the results of the INFO-LEG project (Understanding information for legal protection of people against information-induced harms), which has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 716971). The paper reflects only the author's view and the ERC is not responsible for any use that may be made of the information it contains. The funding source had no involvement in study design, in the collection, analysis, and interpretation of data, in the writing of the report, or in the decision to submit the article for publication.

Endnotes

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[2] See, https://www.oed.com/viewdictionaryentry/Entry/95568 (last consulted 30 October 2019).

[3] See, https://www.oed.com/viewdictionaryentry/Entry/95568 (last consulted 30 October 2019).

[4] Which as a matter of fact, in the history of the theory of information even preceded the more human-based cognitive paradigm. Following Capurro and Hjørland, "with an objectivist view from the world of information theory and cybernetics," before turning "to the phenomena of relevance and interpretation as basic aspects of the concept of information" (Capurro & Hjørland 2003, p. 345).

[5] See also (Capurro & Hjørland 2003, p. 368) "On one hand, information is a thing, on the other, a psychic construction."

[6] Even though Rosenberg warns that data in fact appeared before datum (around 1646), even though it already encompassed the idea of a given (Rosenberg 2013, p. 18).

[7] This is also confirmed by Rosenberg's point following which data can be conceived of as a rhetorical device, which lies at the intersection of concepts such as facts and evidence (Rosenberg 2013, p. 18).

[8] See also (Ceruzzi 2012, p. 5): the computer relies upon a number of significant functions, among which "the automatic storage and retrieval of information in coded form"; and (Ceruzzi 2012, p. 11): "it was at Bell labs where much of the theory of information coding, transmission, and storage was developed."

[9] As a matter of fact, machine learning itself is hard to define insofar as it builds upon various practices, hence the efforts of Jordan to provide scientific criteria of validity for machine learning practice (Jordan 2019). The present contribution builds mainly upon information philosophy, Science and Technology Studies (STS), and machine learning textbooks and literature.

[10] In this sense, one can argue that abstraction is a specific modality of representation. Thanks to Jake Goldenfein for pointing this out to me.

[11] The figures that follow are the work of the present author, but they are based on those of Adriaans and Zantige. If the following lines follow the authoritative example of Adriaans and Zantinge, this should not be taken to mean that the author is not aware of other relevant literature on the topic such as (Frawley *et al.* 1992; Fayyad *et al.* 1996).

[12] Such data can actually be quite easily gathered, for instance buying demographic data on average income for certain neighborhoods (Adriaans & Zantinge 1996, p. 42). This is all the more true in a big data context.

References

Ackoff R (1989) From Data to Wisdom. *Journal of Applied Systems Analysis* 16, 3–9.

Adriaans P, Van Benthem J (2008) Introduction: Information Is What Information Does. In: Adriaans P, van Benthem J (eds) *Philosophy of Information*, pp. 7–29. Elsevier, Amsterdam.

Adriaans P, Zantinge D (1996) *Data Mining*. Addison Wesley Longman Limited, Harlow.

Agar J (2003) *The Government Machine: A Revolutionary History of the Computer*. The MIT Press, Cambridge, MA; London.

Alpaydın E (2016) *Machine Learning: The New AI*. Cambridge, Massachusetts; London, England: The MIT Press.

Art. 29 WP (2007) *Opinion 4/2007 on the concept of personal data*.

Art. 29 WP (2018) *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation* 2016/679 *– Adopted on 3 October 2017 As last Revised and Adopted on 6 February 2018*.

Barocas S, Nissenbaum H (2014) Big Data's End Run around Anonymity and Consent. In: Lane J, Stodden V, Bender S, Nissenbaum H (eds) *Privacy, Big Data, and the Public Good*, pp. 44–75. Cambridge University Press, Cambridge.

Barwise J, Seligman J (1997) *Information Flow: The Logic of Distributed Systems*. Cambridge University Press, Cambridge.

Bates JM (1999) The Invisible Substrate of Information Science. *Journal of the American Society for Information Science* 50 (12), 1043–1050.

Bellanova R, & de Goede M (2022). The algorithmic regulation of security: An infrastructural perspective. *Regulation & Governance* 16(1), 102–118.

Benthall S (2018) *Context, Causality, and Information Flow: Implications for Privacy Engineering, Security, and Data Economics*.

Berkeley EC (1949) *Giant Brains or Machines that Think*. Wiley, New York.

Buckland M (1991) Information as Thing. *Journal of the American Society for Information Science* 42(1), 351–360.

Burgin M (2010) *Theory of Information: Fundamentality, Diversity and Unification*. World Scientific, Singapore.

Bayamlıoğlu, E. (2020, forthcomming). The right to contest automated decisions under the GDPR: Article 22 from a techno-regulatory perspective. Regulation & Governance.

Bygrave LA (2010) The Body as Data? Biobank Regulation via the "Back Door" of Data Protection Law. *Law, Innovation and Technology* 2(1), 1–25.

Bygrave LA (2014) Information Concepts in Law: Generic Dreams and Definitional Daylight. *Oxford Journal of Legal Studies* 35(1), 1–30.

Bygrave LA (2017) *EU data protection law falls short as desirable model for algorithmic regulation* (No. 85). *Algorithmic Regulation*.

Capurro R (2000) Ethical Challenges of the Information Society in the 21st Century. *International Information and Library Review* 32(3–4), 257–276.

Capurro R, Hjørland B (2003) The Concept of Information. *Annual Review of Information Science and Technology* 37(1), 343–411.

Ceruzzi PE (2003) *A Histrory of Modern Computing*. The MIT Press, Cambridge, MA; London.

Ceruzzi PE (2012) *Computing: A Concise History*. The MIT Press, Cambridge, MA; London.

Chaitin GJ (1977) Algorithmic Information Theory. *IBM Journal of Research and Development* 21(4), 350–359.

Chisholm R (1989) *Theory of Knowledge*. Prentice Hall, Englewood Cliffs, NJ.

Colonna L (2013) A Taxonomy and Classification of Data Mining. *SMU Science and Technology Law Review* 16(2), 309.

Committe of Experts on Internet Intermediaries (MSI-NET) (2018) Study on the Human Rights Dimensions of Automated Data Processing Techniques (in Particular Algorithms) and Possible Regulatory Implications. Strasbourg.

Council of Europe (2017) *Guidelines on the Protection of Individuals With Regard To the Processing of Personal Data in a World of Big Data*.

Council of Europe (2020) *Draft Recommendation of the Committee of Ministers to Member States on Human Rights Impacts of Algorithmic Systems*.

Cownie F, Bradney A (2017) Socio-Legal Studies: A Challenge to the Doctrinal Approach. In: Watkins D, Burton M (eds) *Research Methods in Law*, 2nd edn, pp. 40–65. Routledge, Oxon New York.

Custers B (2018) Profiling as Inferred Data. Amplifier Effects and Positive Feedback Loops. In: Bayamlioglu E, Baraliuc I, Janssens L, Hildebrandt M (eds) *BEING PROFILED:COGITAS ERGO SUM: 10 Years of Profiling the European Citizen*, pp. 112–115. Amsterdam University Press, Amsterdam.

Davies W (2015) *The Happiness Industry*. Verso, London; New York.

de Vries K (2016) *Machine Learning/Informational Fundamental Rights: Makings of Sameness and Difference*.

Devlin K (1991) *Logic and Information*. Cambridge University Press, Cambridge.

Diver CS (1998) The Optimal Precision of Administrative Rules. In: Baldwin R, Scott C, Hood C (eds) *A Reader on Regulation*, pp. 219–269. Oxford University Press, Oxford.

Dretske F (1982) *Knowledge and the Flow of Information*. The MIT Press, Cambridge, MA.

Dunn JM (2008) Information in Computer Science. In: Adriaans P, van Benthem J (eds) *Philosophy of Information*, pp. 589–616. Elsevier, Amsterdam.

Economic and Social Reaserch Council (1994) Review of Socio-Legal Studies: Final Report. Swindon.

European Commission (2010) *Communication from the Commission to the European Parliament, the Council, the Economic and social Committee and the Committee of the Regions: A Comprehensive Approach on Personal Data Protection in the European Union*.

European Commission (2020) *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*. Brussels.

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From Data Mining to Knowledge Discovery in Databases. *AI Magazine* 17(2), 37–54.

Fisher RA (1925) Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* 22(5), 700–725.

Floridi L (2008) Modern Trends in Philosophy of Information. In: Adriaans P, van Benthem J (eds) *Philosophy of Information*, pp. 117–136. Elsevier, Amsterdam.

Floridi L (2010) *Information: A Very Short Introduction*. Oxford University Press, Oxford.

Flückiger F (1999) Towards a Unified Concept of Information: Presentation of a New Approach. In: Hofkirchner W (ed) *The Quest for a Unified Theory of Information*, pp. 101–112. Routledge, New York.

Frawley WJ, Piatetsky-Shapiro G, Matheus CJ (1992) Knowledge Discovery in Databases: An Overview. *AI Magazine* 13(3), 57–70.

González Fuster G (2014) *The Emergence of Personal Data Protection as a Fundamental Right of the EU*. Springer, Dordrecht.

Graef I, Husovec M, Purtova N (2018) Data Portability and Data Control: Lessons for an Emerging Concept in EU Law. *German Law Journal* 19(6), 1359–1398.

Griffith BC (ed) (1980) *Key Papers in Information Science*. Knowledge Industry Publications, New York.

Gürses S, van Hoboken J (2017) Privacy after the Agile Turn. In: Selinger E, Polontesky J, Tene O (eds) *Cambridge Handbook of Consumer Privacy*, pp. 579–601. Cambridge University Press, Cambridge.

Haigh T (2011) The History of Information Technology. *Annual Review of Information Science and Technology* 45, 431–487. https://doi.org/10.1002/aris.2011.1440450116.

Hallinan D, De Hert P (2016) Many Have it Wrong – Samples Do Contain Personal Data: The Data Protection Regulation as a Superior Framework to Protect Donor Interests in Biobanking and Genomic Research. In: Mittelstadt B, Floridi L (eds) *The Ethics of Biomedical Big Data*, pp. 119–137. Cham, Switzerland: Springer.

Hartley RVL (1928) Transmission of Information. *Bell System Technical Journal* 7(3), 535–563.

Hildebrandt M (2006) Profiling: From Data to Knowledge. The Challenges of a Crucial Technology. *Datenschutz Und Datensicherheit* 30(9), 548–552.

Hjørland B (1998) Theory and Metatheory of Information Science: A New Interpretation. *Journal of Documentation* 54(5), 606–621. https://doi.org/10.1108/EUM0000000007183.

Hondius FW (1975) *Emerging Data Protection in Europe*. American Elsevier, Amsterdam; Oxford; New-York; North-Holland.

Jashapara A (2005) *Knowledge Management: An Integrated Approach*. FT Prentice Hall, Harlow.

Jordan MI (2019) Artificial Intelligence—The Revolution Hasn't Happened Yet. *Harvard Data Science Review*. 1(1), 1–9.

Kaminski ME (2019) The Right to Explanation, Explained. *Berkeley Technology Law Journal* 34(1), 10–17.

Kelleher JD, Tierney B (2018) *Data Science*. The MIT Press, Cambridge, MA; London.

Kitchin R (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences*. Sage, Los Angeles; London; New Delhi; Singapore; Washington, DC.

Kolmogorov AN (1956) *Foundations Theory of Probability*, p. 84. Chelsea Publishing Company, New York.

Latour B (2005) *Reassembling the Social an Introduction to Actro-Network-Theory*. Oxford University Press, Oxford.

Laudon KC, Laudon JP (2006) *Management Information Systems: Managing the Digital Firm*. Pearson Prentice Hall, Upper Saddle River, NJ.

Lehr D, Ohm P (2017) Playing with the Data: What Legal Scholars Should Learn about Machine Learning. *UC Davis Law Review* 51(2), 653–717.

Loewenstein WR (1999) *The Touchstone of Life: Molecular Information, Cell Communication, and the Foundations of Life*. Oxford University Press, Oxford.

Logan RK (2012) What Is Information?: Why Is it Relativistic and What Is its Relationship to Materiality, Meaning and Organization. *Information* 3(1), 68–91.

Machlup F (1983) Semantic Quirks in Studies of Information. In: Machlup F, Mansfield U (eds) *The Study of Information: Interdisciplinary Messages*, pp. 641–671. John Wiley & Sons, New York.

Mackenzie A (2013) Programming Subjects in the Regime of Anticipation: Software Studies and Subjectivity. *Subjectivity* 6(4), 391–405.

Maniglier P (2019) Problem and Structure: Bachelard, Deleuze and Transdisciplinarity. *Theory, Culture and Society*, 2.

Manovich L (1999) Database as Symoblic Form. *Convergence* 5(2), 80–99.

Manson N (2009) The Medium and the Message: Tissue Samples, Genetic Information and Data Protection Legislation. In: Heather W, Mullen C (eds) *The Governance of Genetic Information: Who Decides?* pp. 15–37. Cambridge University Press, Cambridge.

Martin WJ (1995) *The Global Information Society*. London and New York: Routledge.

Matus KJM, & Veale M (2022). Certification Systems for Machine Learning: Lessons from Sustainability. *Regulation & Governance* 16(1), 177–196.

Miller AR (1971) *The Assault on Privacy: Computers, Data Banks, and Dossiers*. The University of Michigan Press, Ann Arbor, MI.

Mittelstadt B (2017) From Individual to Group Privacy in Big Data Analytics. *Philosophy and Technology* 30(4), 475–494. https://doi.org/10.1007/s13347-017-0253-7.

Nunberg G (1996) Farwell to the Information Age. In: Nunberg G (ed) *The Future of the Book*, pp. 103–138. University of California Press, Berkeley, CA; Los Angeles, CA.

Nys H (2004) Report on the Implementation of Directive 95/46/EC in Belgian Law. *Implementation of the Data Protection Directive in Relation to Medical Research in Europe*, pp. 29–41. Ashgate, Aldershot.

Overdof R, Gürses S, Balsa E (2018) *POTs: The Revolution Will Not be Optmized?*

Packard V (1964) *The Naked Society*. Van Rees Press, New York.

Pasquale F (2015) *The Black Box Society: The Secret Algorithms that Control Money and Information*. Harvard University Press, Cambridge, MA.

Rifkin J (2000) *The Age of Access*. TarcherPerigee, New York.

Rosenberg D (2013) Data before the Fact. In: Gitelman L (ed) *Raw Data Is an Oxymoron*, pp. 15–40. The MIT Press, Cambridge, MA; London.

Rowley J (2007) The Wisdom Hierarchy: Representations of the DIKW Hierarchy. *Journal of Information Science* 33(2), 163–180.

Rule JB (1973) *Private Lives and Public Surveillance*. The Trinity Press, Worcester, MA; London.

Schutt R, O'Neill C (2013) *Doing Data Science. Straight Talk from the Frontline*. O'Reilly, Sebastopol, CA.

Shannon CE (1948) The Mathematical Theory of Communication. *Bell System Technical Journal* 27(1), 623–656.

Stonier T (1997) *Information and Meaning: An Evolutionary Perspective*. Springer, London.

Teilhard de Chardin P (1964) *The Future of Man*. Harper & Row, New York.

US Department of Health Education & Welfare (1973) *Records, Computers, and the Rights of Citizens*. Washington, D.C.: The MIT Press.

Van Hoboken J (2014) The European Approach to Privacy. *TPRC Conference*, pp. 3–32.

Vedder A (1999) KDD: The Challenge to Individualism. *Ethics and Information Technology* 1(4), 275–281.

von Weizsäcker CF (1974) *The Unity of Nature*. Deutscher Taschenbuch Verlag, Munich.

Wachter S, Mittelstadt B (2019) A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and Ai. *Columbia Business Law Review* 2(494), 494–620.

Wersig G (1997) Information Theory. *Encyclopaedic Dictionary of Library and Information Science*, pp. 220–227. Routledge, London.

Yeung K (2018) Algorithmic Regulation: A Critical Interrogation. *Regulation and Governance* 12(4), 505–523.

Zins C (2007) Conceptual Approaches for Defining Data, Information and Knowledge. *Journal of the American Society for Information Science and Technology* 58(4), 479–493.