Philosophy of data: Why?

Jonathan Furner

Graduate School of Education and Information Studies, University of California, Los Angeles, 300 Young Dr N, Mailbox 951520, Los Angeles, CA 90095-1520, USA
Tel.: +1 310 825 5210; Fax: +1 310 206 4460; E-mail: jfurner@ucla.edu

Philosophy of data should not be dismissed as a cluster of scholastic puzzles whose solutions are of limited practical value. On the contrary, philosophy of data should be recognized as constituting the core of a field of data studies that is informed by, but far from equivalent to, statistics, computer science, and library and information studies.

Keywords: Data, philosophy of data, philosophy of information

1. Introduction

Some people have started talking about “philosophy of data” [1,2]. Is there really such a thing? How could it be different from philosophy of information? And why should anyone working in library and information studies care? In this paper, I provide brief responses to these questions.

In the first section that follows, I examine two mid-twentieth-century uses of the phrase “philosophy of data”, downplay their relevance to current concerns, and instead propose a simple way of drawing rough boundaries around subject matter of present interest. Noting, in a second section, the centrality of ontological questions to the area so defined, I review concepts of data that have emerged in three fields for which such conceptualization is especially important, and take a structuralist-rationalist approach to explication of the relationship between the concept of data and that of information, ultimately coming to advocate for a conception of data as representational concreta. The fundamental distinction between data and information is thus revealed to be that between concreta and abstracta. I contend that, analogously, the kinds of philosophical questions that may be asked about concreta are fundamentally different from the ones asked about abstracta, and that it is therefore both possible and desirable to pursue a philosophy of data alongside (rather than as a mere component of) a philosophy of information. Finally, in a third section, I argue for the utility of the results of philosophy of data as potentially foundational for empirical sciences.

2. Is there really such a thing as philosophy of data?

Indeed there is. Initially, I imagined that one way of showing it could be to search
for occurrences of the term “philosophy of data” on Google, and do some sort of analysis of the results. But I decided against it once I had done the search. There are not a whole lot of occurrences of the term out there — and, of the ones that I managed to find, most of them are part of longer phrases like “philosophy of data analysis” and “philosophy of data processing”, the content of which turns out to be slightly different. If anything, a finding like this suggests an opposite, negative response to our opening question. What I would like to do instead is trace the outline of an area of inquiry that we might usefully identify as “philosophy of data”, regardless of the terminology that scholars use to label it. Having said that, as a preliminary I think it is worth taking a quick look at a couple of early uses of the term, if only as a means of distinguishing current, 21st-century developments from their predecessors.

One of the earliest uses of the term appears to be that made by E. W. Franks of the System Development Corporation (SDC) in his contribution to the AFIPS conference of 1966 [3]. Franks was laying out the design of SDC’s Time-Shared Data Management System (TDMS). “[T]o describe the user orientation features and file organization scheme”, he says (p. 80), “it is necessary to first say something about the underlying philosophy of data upon which TDMS is based”. He goes on to explain:

“A TDMS data file or data base, as it is often called, is a collection of information sets or entries. Each entry contains information about one object . . . but [typically] the objects are not all of the same kind. . . . TDMS allows . . . the accommodation of more than one logically consistent file in the same file structure. . . . The logical structure of the TDMS entry is a collection of predefined elements or descriptors. Each entry will have a subset of these elements appropriate to the object being described”.

Then, a little later on (p. 83):

“From the data user’s point of view, the data base appears to be a collection of values to which he may wish to refer. These values have two sets of associations.

In the first place each value is part of the total description of one of the objects in the data base. In the second place each value is both a value for a specified element and a member of the set of all values for that element. The TDMS organization of the data base reflects both types of value association, the element set and the object set”.

Here, Franks is using the label “philosophy of data” to refer to what we might nowadays call a “data model”. He certainly is not talking about a field of inquiry, like philosophy of science or philosophy of language; and he is not even talking about a proposed solution to a specified philosophical problem, such as an ontology of the fundamental categories of things in the world. His is what we might class as a design philosophy: a set of beliefs about what artifacts (such as TDMS) should be like if they are to meet their designers’ goals. When someone talks of “a philosophy
of X”, or “my philosophy of X”, it is normally the case that they are referring to such a set of personal beliefs rather than to a structured field of research.

A few years earlier, in 1962, the statistician John Tukey had related a story about a colleague who had “developed a clear and individual philosophy” about the way in which, in his judgment, outliers in samples should be handled [4]. Tukey would continue to develop his own ideas about a set of techniques that would come to be known as exploratory data analysis (EDA), and a body of justification for using such methods when certain conditions are satisfied [5,6]. Collectively, these ideas have since been characterized as Tukey’s “philosophy of data analysis” [7], and other such “philosophies” (i.e., sets of beliefs about the methods that are most useful in situations of different kinds) have been suggested over the years as alternatives. Few of them have much to do with philosophy of data per se; some, however, have drawn on Tukey’s ideas to make important contributions to philosophy of science [8].

For present purposes, I do not want to take “philosophy of data” to encompass anyone’s favorite way of structuring databases or doing data analysis. Instead, I suggest that it is more useful to view philosophy of data as comprising three somewhat overlapping branches:

1. **Ontology** (or metaphysics) of data: What, precisely, are data? Of what kind (or genus) of thing are they? Under what conditions can something count as a datum? What properties must something have if it is to be, or perhaps to play the role of, a datum? While it may seem arcane, ontology of data is potentially of great practical use, at least to the extent that knowledge of the properties of the phenomena with which we are dealing allows us to learn new ways of doing things to or with those phenomena, as well as to understand how and why things are done to the world by those phenomena.

2. **Epistemology** of data: What kinds of knowledge can we have about data (or about the concept data)? In what ways may we acquire or produce that knowledge? Epistemological questions and answers are useful to the extent that an understanding of the practices by which we try to find out about certain phenomena (in this case, the phenomenon of data) allows us to evaluate those practices and potentially develop new, better ones.

3. **Ethics** and **politics** of data: What kinds of value do we, could we, and should we place in data and data practices of various kinds? Awareness of the possibilities in these respects allows us to evaluate the phenomenon of data in ways that best support our policymaking goals. What are the social impacts of the collection, manipulation, and distribution, etc., of data? Are data intrinsically neutral means for the instrumental realization of different groups’ ends, or are they necessarily reflective of power relations or class interests?

---

1 When he found an ‘outlier’ in a sample he did not simply reject it. Rather he changed its value, replacing its original value by the nearest value of an observation not seriously suspect”. (p. 18).
One factor that can introduce some confusion into this tidy compartmentalization of philosophical approaches to the concept of data is the ordinary role of data in knowledge production. The fields of epistemology – philosophy of knowledge – and philosophy of science treat questions that involve the conceptual analysis of data insofar as data are considered to play a part in the acquisition of (scientific) knowledge in general. Strictly speaking, unless we were talking about methods of producing knowledge about data in particular, questions of these kinds would not normally be thought of as part of #2 (epistemology of data) per se. At the same time, answers to questions that fall under the heading of #1 (ontology of data) can greatly inform answers to questions in epistemology tout court.

3. How could philosophy of data be different from philosophy of information?

Three fields that have had especial reason to achieve consensus on the ontology of data are statistics, computer science, and library and information science.

In the first chapter of his Introduction to the Theory of Statistics (1st ed., 1911) [9], the British statistician Udny Yule summarizes the findings of a previous article [10] in which he traces some of the earliest uses in English of the terms “statistics” and “statistical” to the late eighteenth century. In the course of this account, Yule explains that early users of these terms “meant simply the exposition of the noteworthy characteristics of a state, the mode of exposition being…preponderantly verbal” (p. 2, emphasis added). Gradually, however, “statistics” came to stand not so much for verbal descriptions of a state’s characteristics, but more commonly for numerical descriptions: “After the commencement of the nineteenth century…the growth of official data was continuous, and numerical statements, accordingly, began more and more to displace the verbal descriptions of earlier days” (p. 2, emphasis added). By the time Yule was writing, the meaning of the word had expanded to encompass any “collection of numerical data, analogous to those which were originally formed for the study of the state, on almost any subject whatsoever” (p. 3, emphasis added).

Yule does not see the need to tell the story of “data” in the same way; instead, from the outset, he uses that term (as, for example, in the latter two quotations) to denote “descriptions” (i.e., “statements”) of “characteristics” of the phenomena (e.g., states) in question. He also talks (cf. p. 4) of “observations” of “facts”; later (p. 7), he distinguishes between data on “attributes”, consisting of counts of “objects or individuals” having or lacking a given characteristic, and data on “variables”, where the “actual magnitude” of some characteristic is “note[d] or measure[d]”. In cases of the first kind, “[t]he quantitative character… arises solely in the counting”; in cases of the second kind, “[t]he observations… are quantitative ab initio”.

Albert Kurtz and Harold Edgerton’s Statistical Dictionary of Terms and Sym-
bols [11],² first published in 1939 and reissued without revision in 1967, defines “data” as follows: “Measurements, observations, or estimates; numerical statements of facts in any department of inquiry (such as economics); a collection of statistical facts or figures”. Like Yule, here Kurtz and Edgerton distinguish between, on the one hand, “statements”, “measurements”, and “observations”, and, on the other, “facts”. But the water is muddied a little by the equating of “facts” with “figures” in the third definition. Do the figures not describe the facts? Or are facts themselves descriptions of phenomena?

By 2014, with the publication of the third edition of Graham Upton and Ian Cook’s Oxford Dictionary of Statistics [13], data are straightforwardly identified with structured information:

**data**

Information, usually numerical or categorical.

Somewhat unhelpfully, however, Upton and Cook do not provide a definition of “information”.³ In their 2016 dictionary of computer science (different field, same publisher), Andrew Butterfield and Gerard Ngondi [14] provide two definitions of “data” as a subcategory of “information”. The first distinguishes information of the kind that makes up programs (instructions), and information of the kind on which programs operate (data); the second distinguishes structured information on which programs operate (data) from unstructured information on which programs operate (e.g., text, graphics, speech, image):

**data**

1. Information, in any form, on which computer programs operate. The distinction between program (instructions) and data is a fundamental one in computing….

2. In a more limited sense, data is distinguished from other contrasting forms of information on which computers operate, such as text, graphics, speech, and image. The distinguishing characteristic is that it is organized in a structured, repetitive, and often compressed way. Typically the structure takes the form of sets of fields… The ‘meaning’ of such data is not apparent to anyone who

---

²Psychology professors Albert K. Kurtz (1904–92) and Harold A. Edgerton (1904–2003) were both active in the Psychometric Society, founded in 1935 by Kurtz and others for academic educationalists and psychologists interested in using quantitative methods. Kurtz and Edgerton completed the work begun by the members of the American Statistical Association’s Committee on a Dictionary of Statistical Terms and Symbols that had convened in 1931 [12]. Four of the seven original members of that Committee were among the thirty who constituted Kurtz and Edgerton’s “Advisory Council”, set up to review a preliminary, unpublished edition of the Dictionary in 1935. The authors claimed to have systematically combed 150 books, “as well as many issues of current periodicals” (p. v), in an effort to identify as many relevant terms and as authoritative a set of definitions as possible. Such detail is provided here since the published result is somewhat of a model of its genre.

³There is an entry for “information”, but it redirects the reader to an entry for “Fisher information”, defined as “The amount of information that a sample provides about the value of an unknown parameter”.

does not know what each field signifies. . . . That characteristic gives rise to the popular fallacy that ‘data is meaningless.’ . . .

Unlike Upton and Cook, Butterfield and Ngondi do also provide a full entry for “information”:

**information**
Generally, information is whatever is capable of causing a human mind to change its opinion about the current state of the real world. Formally, . . . information is whatever contributes to a reduction in the uncertainty of the state of a system . . . . Information must be distinguished from any medium that is capable of carrying it. A physical medium (such as a magnetic disk) may carry a logical medium (data, such as binary or text symbols). The information content of any physical objects, or logical data, cannot be measured or discussed until it is known what range of possibilities existed before and after they were received. . . .

Unfortunately, however, reading the two entries together serves only to confuse. The second entry clearly distinguishes medium as carrier and information as content, and lists “data” as an example of a medium. So, whereas the entry for “data” characterizes data as information of a special kind, the entry for “information” characterizes data and information as entirely separate categories.

Since its foundation in 1947, the International Organization for Standardization (ISO) has worked towards global standardization of the quantitative and qualitative properties of technologies, tools, and other artifacts, so that industrial designs and finished products are interoperable with one another, and sharing, trade, and reuse are not just possible but also increasingly efficiently organized. One strand of the ISO’s work has been to compile structured glossaries of words and phrases together with standardized, prescriptive definitions of the specialized meanings assigned to those terms in various communities of practice. The aims of the published standard ISO 5127:2001 [15], for example, are to identify terms used by information specialists and to specify (what ISO experts have determined to be) the consensus among members of that community as to the primary senses that are to be attributed to those terms regardless of local context. ISO 5127:2001 includes entries for “data”, “information”, “knowledge”, “meaning”, and many other of the terms that require careful analysis if we are to arrive at a convincing theory of data. Unfortunately, however, the definitions accompanying the terms, when considered together, display inconsistencies and circular reasoning, to the extent that, for example, it turns out that there is no simple way to evaluate a definition of “data”, given the absence of entries in the ISO vocabulary for terms like “representation” and “interpretation” (both used in the ISO’s definition of “data”), and the knots in which we would have to tie ourselves in order to maintain clear distinctions between “data” and related terms:

**data**
representation of information(1) in a formalized manner suitable for communication, interpretation and processing
information(1)
knowledge that is communicated

knowledge
cognizance which is based on reasoning and passes verification

information(2)
message used to represent information(1) within a communication process in order to increase knowledge

message
data prepared for communication purposes

From a surface reading, nonetheless, we may infer that some knowledge is information(1), that information(1) can be represented by data, that some (if not all) data are messages, that information(1) can be represented by messages, and that some (if not all) messages are information(2). What this means for an understanding of the nature of data is unclear. Perhaps a different approach to the conceptual analysis of data is needed.

Approaches of at least three kinds are recognizable in published efforts in this direction:

1. Those taking what might be called a cultural–historical approach ask, What have most of the people in a specified group at a specified point in space-time understood by the term “data”? An assumption made here is that it is more of a priority to understand what people actually mean when they talk about data than it is to understand what data actually are in objective reality (if anything). Typically, the goal may be described in pragmatic terms as one of understanding how conceptions of data, and the extents to which they have proved useful, have been shaped by the specific cultural and historical contexts in which they have been produced.

2. Those taking what might be called a naturalist approach ask, What are the observable properties of data? An assumption made here is that understanding what data are in objective reality is more important than any other goal. We might wish to evaluate the results in terms of the degree to which our identification of the objective properties of data correspond with “the facts” or “reality”. In practice, few have taken this approach since data is typically understood to be a social rather than a natural kind.

3. Those taking what might be called a rationalist–structuralist approach ask, How is the concept of data best located in an existing conceptual framework, or in a top-level structure of fundamental categories? Is it possible to situate data in a way that does not negatively impact the internal coherence of the framework, and that avoids unnecessary redundancy and/or circularity in its relationships to other concepts?
Jonathan Furner [16], drawing on Luciano Floridi’s work [17], identifies nine culturally-specific senses in which the term “data” has been used in English. Brian Ballsun-Stanton [18] describes a study which, while limited in scope, also exemplifies the first kind of approach. He reports a series of interactions with a small number of participants from which “three distinct philosophies of data” emerged: a “technological” understanding of “data as bits”, i.e., as media that contain digitally encoded information; a “scientific” understanding of “data as hard numbers”, i.e., as the product of objective, reproducible measurements; and an “engineering” understanding of “data as observations”, i.e., as the product of subjective perceptions (p. 123). He also reports a universal distinction made between “raw” data directly obtained from senses or sensors, and “derived” data indirectly produced from calculations performed on raw data (p. 120).

Ballsun-Stanton’s typology of positions on philosophy of data is reminiscent of others’ statements of the options available in philosophy of information, in which distinctions are similarly made between subjective and objective, containers and contents, digital and analog, raw and cooked. Clearly, the vast majority of commentators recognize the existence of some sort of relationship between the two concepts of data and information. For some, they are one and the same; for others, they are quite different. Yet, given the close connection identified even by those in the latter group, why should anyone bother to distinguish between philosophy of data and philosophy of information as two separate and equally worthy areas of inquiry?

Here, along the lines of the third of the types of approach identified above, a framework is provided for considering various candidates for the most useful characterization of the relationship between data and information. Clearly, we have two general options: either “data” and “information” are different names for the same concept; or the two terms pick out different concepts. The possibilities include the following:

1. Data are signifiers; information is the signified. Data stand for something; information is that thing. Data have meaning; information is meaning.
2. Data are concrete, and/or particular, and/or physical, and/or external; information is abstract, and/or universal, and/or mental, and/or internal.
3. Data are raw, and/or objective, and/or empirical, and/or unprocessed; information is cooked, and/or subjective, and/or interpretive, and/or processed.
4. The category of data includes natural signs; information is artifactual (human-made) data.
5. Information can be qualitative or quantitative; data is quantitative information.
6. Information can be structured (i.e., numerical or categorical) or unstructured; data is structured information.

Clearly, it is not the case that each choice from three possible positions on each of nine dimensions is independent of all others; but if it were, this framework would suggest 19,683 different candidate theories of data.
7. Information can be analog or digital; data is digital information.
8. Data are binary digits (bits), or whatever can be represented by them; information is informativeness (measured in bits).
9. Data can sometimes be false, and/or old, and/or irrelevant, and/or meaningless; information is data that is true, and/or new, and/or relevant, and/or meaningful.

Data = information

1. Both data and information are signifiers; or, neither are.
2. Both data and information are concrete (particular, etc.); or, neither are.
3. Both data and information are raw (objective, etc.); or, neither are.
4. Both categories include natural signs; or, neither do.
5. Both data and information can be qualitative; or, neither can.
6. Both data and information can be unstructured; or, neither can.
7. Both data and information can be analog; or, neither can.
8. Both data and information are binary digits (bits), or whatever can be represented by them; or, neither are.
9. Both data and information can be false (old, etc.); or, neither can.

Progress towards a theory of any class of things X requires that we answer fundamental questions. What properties, if any, do all and only instances of X have in common with one another? What are the individually necessary and jointly sufficient conditions, if any, under which something is an instance of X? Of what basic category among those constituting our favored upper ontology is X a subset? In what kinds of relationship does X stand with respect to other classes of things? If we are to present a convincing theory of data, for example, we should be prepared to justify any claims that we might make about the basis on which we count something as an instance of data, and about where, exactly, data fits within the ontology to which we are committed.

A justification of an entire upper ontology is outside the scope of this paper. In its place, I make a general assumption that it is useful to distinguish among instances of two general categories:

1. concreta: both
   (a) ordinary objects: datable and locatable objects (including states of affairs, events, etc.) that are medium-sized and readily distinguishable from their contexts; and
   (b) extraordinary objects: objects (states, events) that exist at microscopic or macroscopic levels and that, due to current limitations in human capabilities, are unobservable – including such things as events that last longer than a human lifetime, elementary particles, and thoughts; and
2. abstracta: such as concepts, kinds, properties, and propositions.

Of the concrete things, some (but not all) sometimes (but not always) play the role of symbols: i.e., under certain conditions, and in accordance with certain conventions, it can be useful for us to assign meaning to certain concreta – in other
words, to interpret those concreta as representing or standing for certain abstracta. This category of representational concreta, we may call data.\(^5\)

About any human interaction with any given symbol (or sequence of symbols), we may ask: (a) Does it result in an interpretation (or set of interpretations) that matches the one intended? and (b) Does it result in an interpretation (or set of interpretations) that is valuable? The first is never a question about the truth or falsity of any propositional component of the interpretation. It is more a contingent matter of the success with which that content is accurately and fully communicated. The relevance of the concept of truth to the second question, meanwhile, will depend on one’s position on scientific realism.

Scientific realism is a fairly common set of ideas about the nature of the relationship between theory and reality. Scientific anti-realism consists in denial of any or all of these beliefs. The realist will argue that any interpretation is valuable to the extent that it is true, i.e., in correspondence with the facts; the anti-realist will value usefulness ahead of “truth”. Claims that a realist typically makes include the following:

- That there are at least some things that exist in the real world, independently of anyone’s thoughts about them;
- That those things have properties, and at least some of those properties are natural (i.e., their having those properties does not depend on anyone’s thoughts about them);
- That at least some of those properties are observable, or measurable, or accessible by instrumentation, in ways that are objective, reliable, and not wholly dependent on the perspective taken by the observer;
- That it is not the case that theory is always underdetermined by our observations (i.e., it isn’t always impossible to choose among hypotheses for lack of relevant evidence);
- That it is at least sometimes possible to estimate the likelihood with which a given theory or proposition about the world is true;
- That, historically, presumptions of the truth of a good proportion of the propositions that have (as yet) evaded falsification have allowed us to make very substantial technological advances;
- That (inferring to the best explanation!) the historical success of science is due to the correspondence of its propositional content to the facts; and
- That, by using the scientific method, we are progressively getting closer to the truth about the way the world really is, i.e., closer to the facts.

It seems as if it could be useful, along similar lines, to distinguish between data realism and data anti-realism. The data realist would be someone who believes, for example, that it is not always the case that data stand for things other than what

\(^5\)It is worth noting that this is a category that includes not only ordinary objects, such as marks on a page, but also extraordinary objects, such as thoughts.
they purport to; and that at least some of the records purporting to serve as true and authentic representations of observations are in fact true and authentic, largely free of human influence and error.

One way of moving further beyond the difficulties faced by the compilers of field-specific vocabularies is to borrow, from philosophy of language and semiotics, a set of ideas about the relationships between thought, expression, and reality. In this framework, any concrete utterance or inscription may be characterized by its meaning, structure, form, and time and place of creation, among other properties (see Table 1). Each of a given pair of utterances may not instantiate any of the same characteristics as the other; may instantiate the same meaning but not the same structure; may instantiate the same meaning and structure but not the same form; may instantiate the same meaning, structure, and form, but not the same time and place of creation.

Applying these ideas to the concept of data allows for the following characterization:

1. Data are **phenomenal**. They are concrete entities that exist in the real world. However, not all phenomena are data.
2. Data are **signifying**. They are sign-vehicles or sequences of sign-vehicles, i.e., utterances and inscriptions (as well as, conceivably, “speech” acts accessible via smell, taste, and touch) intentionally created by humans for the purpose of conveying meaning.
3. Data are thus **representational**, in the same semiotic sense that any sign-vehicle “stands for” its meaning.
4. Data are **symbolic**. They perform their representational function, not through their similarity to their meaning (as icons are), nor through their being produced as an effect of their meaning (as indices are), but through a system of

<table>
<thead>
<tr>
<th>Property</th>
<th>Concept-level</th>
<th>Proposition-level</th>
<th>Work-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meaning</td>
<td>Concept: e.g., love</td>
<td>Proposition: e.g., that Lubetzky is wise</td>
<td>Work: e.g., Moby-Dick</td>
</tr>
<tr>
<td>Structure</td>
<td>Word-type: e.g., “love”</td>
<td>Sentence, lexeme-/morpheme-/phoneme-/grapheme-sequence: e.g., “Lubetzky is wise”</td>
<td>Expression, version, text, translation: e.g., a French translation of Moby-Dick</td>
</tr>
<tr>
<td>Form</td>
<td>Glyph-sequence: e.g., “Love”</td>
<td>Glyph-sequence: e.g., “LU-BETZKY IS WISE”</td>
<td>Manifestation, edition: e.g., a 1963 edition of a French translation of Moby-Dick</td>
</tr>
</tbody>
</table>

### Table 1

Properties of sign-vehicles
shared conventions by which certain data are understood to stand for certain meanings that are associated with them by contention.

5. Data are *propositional*. The meanings for which they stand are propositions that certain things are the case.

6. Data are *evidentiary*, in two senses. The propositional “content” of data – the propositions that people interpret as the meaning of symbols – can potentially be pressed into the service of the scientific method, as evidence from which hypotheses of explanations may be inferred by deduction, induction, or abduction. The data themselves, considered as physical records of abstracta, can also potentially be used as evidence of the circumstances in which they were created.

7. Data are *assertive*. By producing data, their creators commit themselves to certain things’ being the case; their creators perform speech acts through which belief in the truth of the propositions asserted is expressed.

8. Data are *descriptive*. The propositions for which they stand are about certain real-world phenomena. (Sometimes this descriptive quality is defined in terms of representation; it should be clear that this is representation of a different kind from that treated in #3.)

9. Data are *observational*. Typically, they are intentionally produced as a result of their creators’ acts of observing the world.

10. It is often the case, certainly, that data are *numerical* (i.e., quantitative) or categorical, but normally there is no particular reason to exclude qualitative or unstructured descriptions from the category.

In other words, the category of *data* should be equated not with the category of *phenomena*; nor with that of *acts of observing*; nor with that of *observations*; nor with that of *representations of observations*. Rather, data are *concrete instantiations of symbolic representations of descriptive propositions*, informed by empirical observation, about the quantitative and qualitative properties of real-world phenomena.

This is an opportune moment to ask again: What is there to stop us treating the conceptual analysis of data simply as a component of philosophy of information, as conceived by, for example, Floridi [19, pp. 20–21]? There are several potential directions in which to take a response.

1. Floridi’s conception is of *data* as a super-category of *information*: i.e., all information is data (of a certain kind); there is no information that is non-data. So, it would be more rational for philosophy of information to be considered part of philosophy of data, rather than vice versa.

2. A comparison with the relationship between philosophy of language and philosophy of literature might be illuminating. Clearly, language and literature are different things: for one thing, not all instances of language (in the sense of *parole*, or language-use) are instances of literature. As a consequence, different questions are asked in philosophy of language than in philosophy of literature, and these two field-names denote distinct areas of interest with different (if
partly overlapping) content. Suggestions that one should be encompassed by the other are rare.

3. Even though it may at first seem plausible that all instances of literature are instances of language-use, cases such as wordless novels are potential counterexamples. Analogously, we might wish to argue that, in fact, there exist instances of information that are not simultaneously instances of data, as well as instances of data that are not simultaneously instances of information. Perhaps this could be explained as a result of the informationhood of an entity lying in its capacity to inform, i.e., to cause a positive change in a person’s knowledge-state, while the datahood of an entity lies in its capacity to mean, i.e., to represent a proposition. Since the capacity to inform and the capacity to mean are quite different qualities, it does not seem to make sense to include study of one within study of the other.

4. We might be persuaded by the argument above, that the class of data-instances is a category of concreta and the class of information-instances is a category of abstracta. In that case, there can be no possibility of overlap between the two. Again, given the fundamental difference between concreta and abstracta, as little as it may make sense to study one without studying the other, it still makes no sense to subsume the study of one within study of the other.

4. Why should anyone care about philosophy of data?

Philosophy of data is an area of interest to scholars working in many distinct fields, ranging from those in which realist approaches to questions of explanation and truth are the norm (e.g., computer science, data science, information science) to those more hospitable to constructivist approaches (e.g., archival studies, digital humanities, science and technology studies). Certain elements and aspects of philosophy of data are also considered in various overlapping subfields of philosophy (philosophy of language, philosophy of mind, philosophy of technology). For reasons already mentioned, however, perhaps the most closely related of philosophy of data’s cognate fields is philosophy of science.

In discussions of the nature and role of scientific explanation – regardless of whether explanation is modeled as deductive inference from observed evidence to a hypothesis that cannot be denied without logical contradiction, as inductive inference from evidence of observed things of a given kind to a hypothesis about unobserved things of the same kind, or as abductive inference from observed evidence to “the best” (e.g., the simplest) hypothesis – “data” is a term that is frequently used interchangeably with “evidence”, “observations”, “facts”, and “phenomena”. Seldom are distinctions routinely drawn among, for example, the phenomena that are observed, the acts of observing those phenomena, the propositions in which the phenomena are described, and the physical records through which those propositions are reported.
and/or communicated, with the result that the intended reference of the term “data” is sometimes unclear.

When “data” is considered admissible for use in descriptions of “the scientific method” – an empirical procedure for constructing and evaluating explanations – the sequence of steps is typically presented along the following lines:5

1. Collect some data: i.e., make some observations or take some measurements of real-world phenomena;
2. Formulate a hypothesis: i.e., construct an argument that explains those data;
3. Deduce some new data: i.e., predict some different data (not previously observed) by deduction from the hypothesis;
4. Carry out a test: i.e., make more observations or take more measurements to determine whether the prediction is correct; then
5. If the prediction is correct, return to step 3; if the prediction is not correct (i.e., if it is disconfirmed), return to step 2.

One important assumption underlying presentations like this is an epistemological one – viz., the claim that it is possible to make accurate, non-contextual, objective representations of phenomena whose existence is independent of humans’ ability to perceive it. Whether the data collected should be treated merely as surrogates for real-world phenomena or as real-world phenomena themselves is left open in this presentation of the scientific method. James Bogen and James Woodward [20] argue for the importance of recognizing that the categories of data and phenomena are distinct;7 that “in most cases” (p. 306) phenomena are not observable; that data can “play the role” (p. 305) of evidence for phenomena, and that phenomena can similarly serve as evidence for general scientific theories, but that data “typically” (p. 305) cannot be predicted or explained by general theories. In other words, the only things that are observable are data, while the only things that can be explained by general theories are phenomena.

For example, suppose the phenomenon in which we are interested is the melting point of lead. Bogen and Woodward point out that “one does not determine the melting point of lead by observing the result of a single thermometer reading” (p. 308; emphasis in original): instead, “one must make a series of measurements”, which provide “a scatter of individual data-points” (p. 308). If we can assume that these observations are normally distributed, then we can take their mean to be “a good estimate of the true melting point” – even though this mean “does not represent a property of any particular data-point” (p. 308; emphasis added). “So”, Bogen and Woodward conclude, “while the true melting point is certainly inferred or estimated

5Presentations of this kind are said to be based on a hypothetico-deductive model of confirmation, in which to confirm (or corroborate) a hypothesis is to observe no evidence that is not logically entailed by it.

7Bogen and Woodward slip easily between talk of, on the one hand, data and phenomena in themselves, and on the other, facts (or claims) about data and facts (or claims) about phenomena.
from observed data . . . [no] report of an experimental distribution of the melting point of lead . . . literally describe[s] what is perceived or observed” (p. 309; emphasis in original). Moreover, “it is easy to see that a theory of molecular structure which explains why the melting point of lead is approximately 327 degrees could not possibly explain why the actual data-points occurred” (p. 309).

Bogen and Woodward allow that the calculation of the mean of a distribution of thermometer readings could be characterized as “a case of ‘theory-laden’ observation” (p. 310). “But”, they say, “it is not clear . . . why anyone should think it would be an illuminating thing to say”, since the problems that arise in the example are not problems to do with perception or with whether theory-unladen observation is possible, but rather are “problems of data-analysis and statistical inference” – e.g., “what conditions have to be met if the true melting point is to be estimated reliably from the observed data, what sort of properties different estimating procedures have, how to determine the likely error in one’s estimate, and so forth” (p. 310).

It is difficult to imagine a more fundamental assumption than one that specifies the kind of relationship that exists between phenomena and data. As we have seen, there are a number of ostensibly plausible options available to those wishing to commit to such an assumption, and conceptual analysis of the sort that is necessary to make a case in support of one or other of these options is a core component of philosophy of data. Any field of inquiry, then, whose goals include the construction of explanations for real-world events should find a place for philosophy of data among its foundations.

5. Conclusion

Current issues in data science tend to be expressed in terms of the impact of “big data”, the “data deluge”, etc., on social activities and on policymaking, without sustained reference to the kinds of method used or the kinds of conclusions drawn by philosophers. They are enormously important issues, to be sure. But philosophy of data should not be dismissed as a cluster of scholastic puzzles whose solutions are of limited practical value. On the contrary, philosophy of data should be recognized as constituting the core of a field of data studies that is informed by, but far from equivalent to, information studies, science and technology studies, and statistics. Exponential growth in the amount of work undertaken in the name of philosophy of data can be expected in the next few years.

References