A Philosophy of Data
Introduction

In 2017, The Economist headlined, "The world's most valuable resource is no longer oil, but data" (The Economist, 2017). Data are not only increasing in value but also affect our lives in the most profound ways — from guiding our love life (Rajan, 2019) to dictating the value of each citizen in China’s Social Credit System (Liang et al., 2018). At the same time, there is considerable debate on the social and ethical implications of data and data technologies. While this discourse is of critical importance, we argue that it is missing one fundamental point: If more and more efforts in business, government, science, and our daily lives are data-driven, we should pay more attention to what exactly we are driven by. Therefore, data ethics should include debate on what fundamental properties constitute data. It should be explicit about how the concept of data might differ between disciplines, as well as technologies, and how these differences interrelate with the concerns raised.

To this end, much work has been done to provide a philosophy of information (Floridi, 2011a). Nevertheless, it leaves open the substantive issue of whether and how data and information differ. Currently, most research on data ethics has avoided these fundamental questions, possibly because of the variability in the use of the term “data.” In fact, the task of providing a universal account of data appears little promising. In contrast, definitions of data in specific fields have already been provided, such as in biology (Leonelli, 2015). However, arguably the most prevalent and critical use of the term can be found in statistics, data science, and its applications such as artificial intelligence.

We argue that discourse on the ethical and social implications of statistics, data science, and data technologies would profit from a more active debate on the fundamental properties of statistical data. For this purpose, our paper proposes one possible definition of a statistical datum
based on the fundamental properties necessary for statistical computation. We outline how statistical data can only constitute a very specific representation of information. Based on this definition, we will differentiate two main types of data\(^1\) and exemplify how an explicit definition can help ground ethical and social debates. While data might constitute merely a specific representation of information, dynamics in the usefulness of data for statistical analysis also affect the semantics of data. We primarily focus on two ways of how these constraints, together with the proliferation of data, can alter our perspective and understanding of ourselves and the world around us.

**Ontology of Data**

**Localizing data**

Before defining statistical data, we shall at least outline some descriptions of statistics itself. Savage (1977), for instance, defines statistics as the study of uncertainty. Others define statistics as “the practice of gathering data and making inferences based on them” (Bandyopadhyay & Forster, 2011, p. 1). Agresti and Franklin (2009, p. 4) also identify statistics as “the art and science of learning from data.” As we lack an exact understanding of data themselves, such proposals leave open the substantive question of what exactly we are learning from. If we concede that those data-based definitions of statistics carry some truth, we can see how crucial an investigation into the nature of statistical data can be for understanding the epistemology of statistics, as well as its social and ethical implications.

\(^1\)For the remainder of this paper, we will refrain from repeating the qualifier “statistical,” and instead refer to data.
By looking at various definitions of statistics, it becomes evident that we can approach statistics from different angles, as well as understand its domain more or less broadly. Statistics might include stages such as observation, information collection, preprocessing, computation, visualization, and interpretation (Freund & Miller, 1965).

Within this process, different steps depend on different kinds and representations of information. The information gained at the point of observation can be vastly different from the processed information before computation in both semantics and representation. If we want to define data as the informational material of statistical analysis, we face the question at which stage we should seek to locate data.

To this end, there are three factors that we want to consider. First, we wish our definition of data to be sufficiently distinct from information in general. Only then will it be relevant for questions specific to statistics. Second, to reach meaningful insights about data, in general, we wish to define it at a stage where we have significant common properties of the kind and representation of information. Lastly, our definition of data should give us relevant insights on social and ethical issues. For this reason, it makes sense to locate data close to the step of greatest influence.

All these factors lead us to pinpoint data rather far along the statistical process. However, as data are the substance of statistical analysis, not only the product, we need to locate it before the primary inference-generating step. Therefore, we identify data as the processed information that can serve as the input for statistical calculation. Locating data at this place does not mean that only information immediately before computation is considered data, but that everything we consider to be data would be directly admissible for such calculation.
Consequently, we can only be sure that our definition of data is applicable if statistical work includes some sort of calculation. We believe this to be the case in the primary, or at least sufficiently relevant parts of statistics to warrant our definition’s generality. While the gathering and processing of data can make up significant parts of a statistician’s work, it is mostly preparation. A statistician does not collect or process data as an end in itself. He or she does so to enable inferences (Agresti & Franklin, 2009, p. 4; Bandyopadhyay & Forster, 2011, p. 1), even if these acts are temporally distant or carried out by different individuals. We also point out that our definition might still be admissible, even in cases where the primary purpose is visualization, not computation. Visualization might very well rest upon data as we define it, and not every visualization of information can lie within the practice of statistics. A more detailed analysis of these edge cases, however, would require us to analyze the various definitions of statistics in detail—a task beyond the scope of this paper.

Defining data

We identified data as information that can serve as the input for statistical calculation. To now arrive at our definition of a datum, we must look at the most basic properties that enable statistical calculation. Mathematical (or statistical) calculation is numerical. It follows that data must have numerical properties to be admissible for calculation. However, statistics is different from mathematics. Statistical computation is not self-sufficient but usually rests on real or imagined phenomena represented by data. Therefore, a statistical datum cannot be merely an isolated abstract number but must represent a substance behind it, with its respective properties.

We conclude that for data to be admissible for statistical computation, it needs to have both numerical, as well as substantive properties. Consequently, we arrive at our definition of a statistical datum:
**A datum is a numerical property representing a substantive property of a specific object.**

The question remains, how data is distinct from information. Our answer would be that data is a specific representation of information, not necessarily semantically specific information. Clearly, this depends on which conceptualization of information we entertain. In cases where information implicitly or explicitly entails both an object of information (*specific object*) and an interpretation (*substantive property*), we can represent such information as data even when that property is in itself not quantitative. We will illustrate this further in the next chapter, where we distinguish between types of data.

**Types of Data**

The numerical properties of a datum can have a wide range of domains. Recall that we identified data as the kind of information that one can immediately apply in statistical computation. There, it becomes the value of a variable. These variables can be categorical (also: qualitative) or quantitative. For this reason, we can also distinguish between categorical and quantitative data.

We could also distinguish data in other ways, such as based on their measurement scale as nominal, ordinal, interval, and ratio data (Stevens, 1946). We could use even finer groups, for instance, by differentiating between the ratio and absolute scale (Stevens, 1959, p. 34). Moreover, when defining a datum, we attributed it numerical properties, which can also include ranges or distributions of numbers that we could distinguish. Therefore our categorization of data is by no means authoritative. Our choice is purely teleological, in that we seek to illustrate ways how an explicit definition of data can ground different ethical and social concerns and relate them to different types of data. Other distinctions might prove more suitable for other goals.
Data That Mark: Categorical Data

In categorical data, some numerals are used only as labels. They serve the mere purpose of identifying the respective category. They are, as Espeland and Stevens (2008, p. 407) calls them “numbers that mark.” Therefore, they can be completely arbitrary. Moreover, such labels do not need to be numerical at all. A letter or other symbol can equally function as a label. So how can these be the numerical properties of data that enable statistical computation? They cannot.

Let us look at an example. We have categorical data on the sports a specific high school athlete is active in. Our categories are basketball, swimming, cheerleading, and soccer. While we could label the sports from 1 to 4 {basketball: 1; …; soccer: 4}, we could equally label them with their first letters {basketball: b; …; soccer: s}, or simply use their full name. These are abbreviations for the qualitative properties, not the numerical value of the individual datum.

To see what the numerical properties of categorical data are, let us now imagine that we have two students in our dataset. Joe swims and cheers, while Jessica plays soccer. The numerical property of the individual datum is the binary numerical answer to the question, “Does object$^2$ x have property y?“. In our case, we could ask, “Does Joe play Basketball?”.

Table 1

|       | Basketball | Swimming | Cheerleading | Soccer |
|-------|------------|----------|--------------|--------|
| Joe   | 0          | 1        | 1            | 0      |
| Jessica | 0         | 0        | 0            | 1      |

$^2$ Note that an object does not necessarily need to be an individual instance of something but a grouping can equally represent the object of a datum.
Here we can see the difference between a label for the substantive property and the numerical value of categorical data. By clearing this potential misconception, we can also understand how categorical data and its numerical property record the presence (1) or absence (0) of a property. It allows us to compare different objects for the properties that they have. However, it does not allow us to compare the property between objects. For the property can only be present, in which case it is identical in being present or absent, in which case there is nothing to compare.

Table 2
*Numerical Value and Substantive Property*

|                  | Plays Basketball (Substantive Property) |
|------------------|----------------------------------------|
| Joe (Object)     | 0 (Numerical Value)                    |

We wish to address one possible objection to our concept of categorical data: It is correct that we do not always represent categorical data in a numerical manner and might still habitually speak of “data.” For instance, if our table is represented in such a fashion:

Table 3
*Potential Categorical Data on High-School Athletes*

|          | Basketball | Swimming | Cheerleading | Soccer |
|----------|------------|----------|--------------|--------|
| Joe      | ✗          | ✓         | ✓            | ✗      |
| Jessica  | ✗          | ✗         | ✗            | ✓      |
Previously, we identified data asymptotically close to statistical calculation. The numerical representation is the form that is immediately admissible for such computation. The information can also be presented in other formats of identical semantical meaning but needs to be translated in a numerical format to be admissible for calculation. Therefore, our checkmark example constitutes a very close predecessor to data, that one might reasonably equate with data in many instances.

Categorical data only mark the presence or absence of a property in a specific object. While it may be of variable practical value, every information that includes an object and an interpretation can theoretically be represented as categorical data. This is contrary to quantitative data.

Data That Commensurate: Quantitative Data

While categorical data does only identify the presence or absence of a certain substantive property, other types of data allow for comparisons of a property by degree. Espeland and Stevens (2008, p. 408) call them “numbers that commensurate,” adding that “many of the most consequential uses of numbers entail commensuration – the valuation or measuring of different objects with a common metric.”

With quantitative data, we can quantitatively differentiate between the same property of different objects. For instance, can we not only say whether Jessica shot a goal or not (categorical), but we can see how many goals she shot and how she compares to Jennifer and Julia, who also shot goals.
However, to do so, we need to operationalize our property. That means we need to identify not only the property but also a way by which to measure it. This operationalization is of ethical and social significance.

**Ethics of Data**

We now wish to illustrate how our explicit definition of data can be of value for data ethics. Unfortunately, it is beyond the scope of the paper to provide a comprehensive overview. In the same manner that we appreciate different opinions on a possible ontology of data, we welcome diverging conclusions on how our definition supports or weakens specific ethical and social concerns. It is not our primary goal to provide an authoritative or comprehensive philosophy of data. Rather, we wish to incentivize a more lively discourse on the issue. To this end, we give two examples of dynamics that limit the kind of information available in data and touch on their ethical and social implications. First, we shed light on how statistic’s need for commensurability in the substantive properties of data may alter our perspective on the world. Second, we point out how the quantitative operationalization of data, provides rather constrained but ever more pervasive ontologies of properties.

---

**Table 4**  
*Quantitative Data on High-School Athletes*

|       | Goals shot | Time since last goal in days | Offensive value metric |
|-------|------------|------------------------------|------------------------|
| Jessica | 2          | 2 days                       | 3.2                    |
| Jennifer | 3          | 16 days                      | 3.4                    |
| Julia  | 1          | 2 days                       | 2.4                    |
Classification and our level of engagement with the world

As Blalock notes (1979, p. 12), “all other levels of measurement, no matter how precise, basically involve classification as a minimal operation.” With the proliferation of new statistical methods, such as machine learning, one remark is required here. It is correct that we increasingly make use of algorithms that make inferences based on numbers that are not categorized or identified at all (Lipton, 2016). Therefore, it does not constitute data as we defined it and, in fact, may not even be considered information (Floridi, 2011b). Nevertheless, exactly this difference, which has crucial implications, for instance, when it comes to responsibility and accountability, can be intuitively grounded through our definition of data.

In our framework, classification refers to the identification of a property. Essentially, it means that before we can have data, we need to settle what we have data on. While a datum might very well merely be a specific representation of information, teleological factors restrain the properties based on their usefulness in generating inferences through statistical computation.

There is much truth in the saying that one cannot compare apples to oranges. So, in order to compare properties across objects to the property of another, these properties need to be of the same kind. Useful classification, therefore, entails that we define properties that either allow for meaningful grouping (categorical data) or apply to a substantial number of objects under investigation (quantitative data).

Let us illustrate this point with an example. Students might be talented in very distinct ways. One student might be able to write exceptionally clearly and straight-forward, while the other can capture a reader with filigreed prose. Students may write in countless, very individual styles, but when we seek data for statistical computation, we would hardly classify each of these
styles as a distinct property. Instead, we would seek categories that allow for relevant groupings, such as “simple style,” or “ornamented style” (categorical) or that applies to all of them in varying degrees, such as “length of words” (quantitative).

We can see how the fundamental purpose of statistical data prescribes us a certain level of specificity. If more and more of our lives revolve around data, this might also affect our day to day conception of the world. Our increasing reliance on data, also called datafication (Cukier & Mayer-Schoenberger, 2019), might explicitly and implicitly influence at what level of specificity we engage with the world. It may shape us to see general properties, instead of individuality and to see distinct characteristics, instead of a complex whole.

Not only concerns about specific kinds of algorithms should entertain our interests. The ways in which a data-driven world changes our self-image and our place in the world, though more subtle, are by no means less decisive. For this reason, a philosophy of data must include efforts to understand these dynamics. If a philosophy of information leads to a “homo poieticus” (Floridi, 2013), a philosophy of data might need a “homo numericus” (Compiègne, 2010).

Quantification as ontological practice

Our second concern applies to quantitative data. To constitute data that commensurate, a defined property has to be measured in some form or the other. Recall our definition of a datum as a numerical property representing a substantive property of a specific object. A datum is the coming-together of numerical and substantive properties. This relationship between number and substance is where our definition can bring to light one of the most critical ethical and social implications of data: Quantitative operationalization of a property is ontological practice (Espeland & Stevens, 2008).
Let us start with an example. If we wish to collect meaningful data on biodiversity as a property of potential uses of land, we ultimately need to quantitatively operationalize it. In other words, we need to put a number on how biodiverse a forest or a shopping mall is. How we come up with this number then, in turn, becomes the operational definition of biodiversity at the basis of our calculations, algorithms, and the inferences made based on them. Moreover, does it implicitly shape our understanding of biodiversity as a society. In this manner, we can understand quantification as ontological practice. A fact that should make us even more conscious about the power that institutions or cultures may have in setting standards for data.

A historical example of how quantitative operationalization can coin our understanding of a concept is the Binot-Simon Intelligence Scale. As Carson (2007, p. 159) notes, “out of a variety of different ways of conceptualizing intelligence in play at the beginning of the century, one dominant theme had emerged. Intelligence was understood as a differential, quantifiable, unilinear entity that determined an individual’s or group’s overall mental power.” The success of the intelligence quotient and its proliferation shape our popular concept of intelligence to this day.

The importance of this dynamic can hardly be overstated as it effectively alters our understanding of ourselves, as well as the world around us and, thereby, our collective and individual actions. Data is rapidly becoming a pillar of modernity, and a significant amount of the use cases of data involve some act of statistical computation. At least in these cases, we outlined how specific in format and kind, the admissible information is. With the proliferation of data technologies, we are effectively ontologizing our reality in a manner that is constrained by the fundamental properties of data. It is for this reason that we need a philosophy of data.
Conclusion

In this paper, we provided one possible definition of data and showed how we can put it to work. We defined a datum as the coming-together of substantive and numerical properties and differentiated between categorical and quantitative data. Following, we stressed how two dynamics of datafication can implicitly but profoundly alter our conception of the world around us and of ourselves. First, data may influence with which degree of specificity we look at reality. Second, the quantitative operationalization of properties in data can be understood as influential ontological practice.

However, if there is a single thing that we hoped to achieve with our analysis, it is to demonstrate why we need an intensified debate about what exactly data is. If our collective and individual activities become increasingly “data-driven,” we should talk about what it is we are driven by.
Agresti, A., & Franklin, C. (2009). *Statistics: The Art and Science of Learning from Data*.

Bandyopadhyay, P. S., & Forster, M. R. (2011). Philosophy of Statistics: An Introduction. In *Philosophy of Statistics*. Elsevier Science & Technology.  
http://ebookcentral.proquest.com/lib/duke/detail.action?docID=713614

Carson, J. (2007). American Psychology and the Seductions of IQ. In *The Measure of Merit* (pp. 159–194). Princeton University Press; JSTOR. https://doi.org/10.2307/j.ctv346qgf.10

Compiègne, I. (2010). *La société numérique en question(s)*. Editions Sciences Humaines.  
https://doi.org/10.3917/sh.compie.2010.01

Cukier, K. N., & Mayer-Schoenberger, V. (2019, September 17). *The Rise of Big Data*.  
https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data

Espeland, W. N., & Stevens, M. L. (2008). A Sociology of Quantification*. *European Journal of Sociology / Archives Européennes de Sociologie*, 49(3), 401–436.  
https://doi.org/10.1017/S0003975609000150

Floridi, L. (2011a). The Philosophy of Information. In *The Philosophy of Information*. Oxford University Press.  
http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199232383.001.0001/acprof-9780199232383

Floridi, L. (2011b). *Semantic information and the veridicality thesis*. Oxford University Press.  
http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199232383.001.0001/acprof-9780199232383-chapter-4
Floridi, L. (2013). The constructionist values of homo poieticus. In *Ethics of information* (1st ed.). Oxford: Oxford University Press, 2013. https://find.library.duke.edu/catalog/UNCb7667329

Freund, J. E., & Miller, I. (1965). *Probability and statistics for engineers*. Englewood Cliffs, N.J., Prentice-Hall [1965]. https://find.library.duke.edu/catalog/DUKE001604840

Hubert M. Blalock, J. (1979). *Social statistics* (Rev. 2d ed.). New York: McGraw-Hill, c1979. https://find.library.duke.edu/catalog/DUKE000294170

Leonelli, S. (2015). What Counts as Scientific Data? A Relational Framework. *Philosophy of Science*, 82(5), 810–821.

Liang, F., Das, V., Kostyuk, N., & Hussain, M. M. (2018). Constructing a Data-Driven Society: China’s Social Credit System as a State Surveillance Infrastructure. *Policy & Internet*, 10(4), 415–453. https://doi.org/10.1002/poi3.183

Lipton, Z. C. (2016). The Mythos of Model Interpretability. *ArXiv:1606.03490 [Cs, Stat]*. http://arxiv.org/abs/1606.03490

Rajan, R. S. (2019, March 10). *This is How AI is Redefining Love*. Medium. https://medium.com/swlh/this-is-how-ai-is-redefining-love-53c78f0f1118

Savage, L. J. (1977). The shifting foundations of statistics. In *Logic, laws, & life: Some philosophical complications* (pp. 3–18). Pittsburgh: University of Pittsburgh Press. https://find.library.duke.edu/catalog/DUKE000223500

Stevens, S. S. (1946). On the theory of scales of measurement. *Science (New York, N.Y.)*, 103(2684), 677.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and Theories*. Wiley.
The Economist. (2017). The world’s most valuable resource is no longer oil, but data. *The Economist.* https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data