




# Data and Coherence Theories of Truth – Examples From a Data-Driven Geographical Information Science

Franz-Benjamin Mocnik 

University of Twente, Enschede, the Netherlands

Correspondence: Franz-Benjamin Mocnik ([franz-benjamin.mocnik@utwente.nl](mailto:franz-benjamin.mocnik@utwente.nl))

**Abstract.** The validity of information collections can be verified by their coherence, such as in the case of Volunteered Geographic Information. However, corresponding coherence theories of truth do not readily apply to collections of data if these consist of non-interpreted or virtually non-interpretable symbols, as is often the case with machine learning models and other black box systems. This paper argues why data-driven geography requires coherence theories, to then transfer the concept of coherence theories from the information to the data level. Finally, the relevant implications on the interpretation of data, especially in the context of black box systems and machine learning, are discussed.

**Keywords.** Theories of truth, coherence theories, data-driven geography, dichotomy between data and information, black box systems, machine learning

## 1 Introduction

The discipline of geography studies anthropogenic and physical structures and processes on the Earth's surface. For this purpose, information is collected in corresponding studies, which then finds its way into the social and natural science discourse. Corresponding data is generated and statistically analysed, for instance, in scientific measurement series and demographic surveys. In quantitative geography, data are traditionally used primarily in their function as an easily aggregated and summarized figure. This role has changed in recent decades. Increasingly, vast amounts of data are available and can be employed, in the sense of big data, for investigations that were previously impossible. In particular, improved data processing methods, including machine learning and artificial intelligence methods, allow a high degree of automation, which in turn enables analyses in previously unseen detail. These developments are not merely superficial but have far-reaching impact on the epistemological convictions of the discipline. The described development towards a data-driven

geography must therefore be seen as a new paradigm that complements existing ones (Miller and Goodchild, 2015).

The success of the data-driven paradigm relies on several factors. First, it seems important to establish an understanding of whether information inferred from data is reasonable in the context of the geographical environment and accurately describes it. Secondly, the data-driven paradigm is challenged by human conceptualization, which is difficult to capture formally. At the moment, however, data-driven geography retreats to formal and easily comprehensible points of view. Thirdly, the paradigm has yet to prove that it is capable of generating deeper understanding, insights, and knowledge. This list of challenges is by no means complete but outlines some of the important factors related to the success of the data-driven paradigm.

This paper deals with the first of the three factors mentioned: the question of whether data are reasonable in the context of the geographical environment. The question is of particular interest as geographical data are often difficult to check due to the large spatial extent of the geographies described; and they are very diverse in nature, especially because they are usually contributed by a large number of people (See et al., 2016; Mocnik et al., 2019), which means that there is no general one-fits-all strategy to verify the data. Initially, we fix a distinction between the data and information concepts that we use in this paper (Section 2), in order to then introduce coherence theories of truth (Section 3) and apply them to the context of data (Section 4). Finally, we discuss which consequences the processing of data by black box systems, potentially including machine learning techniques, have on the application of corresponding coherence theories (Section 5).

## 2 Concepts of Data and Information

There exist a host different yet related concepts of what data are (Zins, 2007). The conceptualization differs not only between the relevant fields of study, but also within

individual fields. In Geographical Information Science, for instance, the demarcation of data to information is fuzzy and unclear, as is the way in which data and information refer to each other, that is, the dichotomy between data and information. Floridi (2008) groups possible interpretations of the concepts of data and information into four classes, three of which we will distinguish and refer to in this paper: the epistemic, the computational, and the informational interpretation.

The first of these classes refers to *collections of propositions* about the geographical environment, based on an epistemic interpretation. These propositions make statements about the environment, and they may or may not be true. In some cases, it may be debatable or even undecidable whether these propositions are true. Such propositions are generally comprehensible to humans because they afford mental reasoning. They are cognitively accessible and part of our perception and thought process, which is why they do not necessarily have to be formally represented by symbols. Such concept needs to be distinguished from a computational interpretation, which refers mainly to the predominantly formal representation of propositions by means of symbols, such as those found in spatial datasets. In the context of this paper, we refer exclusively to such collections of symbols as *data*. That is, we do not consider collections of propositions themselves to be data but only the symbols that represent these. If, on the other hand, these symbols are interpreted, we speak of *information*. Following this definition, a collection of propositions needs to be considered to be information, too, if they are the result of interpreting data. Depending on how much such propositions and information in general are mentally processed and distilled, knowledge and wisdom can be generated (e.g., Ackoff, 1989). In this sense, there exist many types of information (including knowledge) other than collections of propositions about the (geographical) environment.

For the purposes of this paper, we will refer to collections of propositions, data, and information in the way these concepts have been characterized here.

### 3 Coherence Theories of Truth

Numerous research activities in Geographical Information Science focus on the quality of information (e.g., Frank, 2007; Chrisman, 1984; Shi et al., 2002; Senaratne et al., 2017; Ballatore and Zipf, 2015). At the core of such research is the question of which of the propositions apply to the geographical environment, and which do not. Regardless of what ‘apply’ means in the context of the natural and social sciences, such an assessment presumes access to the geographical environment (Mocnik et al., 2018). Classical *correspondence theories* thus attempt to determine the truth of propositions through appropriate explicit comparisons to the environment (Patterson, 2003). Extrinsic quality measures serve as typical examples of such correspon-

dence theories because they are usually based on comparisons with information obtained from reference data.

*Coherence theories of truth* have been proposed as an alternative to correspondence theories in philosophical discourse, since they take the intrinsic coherence of propositions among themselves as an indication of their applicability to the (geographical) environment and do not rely on the often hard to achieve epistemic and immediate access to the environment (e.g., Cohen, 1978; Rescher, 1973). Consistency, that is, the absence of contradictions in the propositions, is often assumed to be a prerequisite for coherence (e.g., Cohen, 1978; Stout, 1908). Coherence, however, goes beyond this in that it also assumes the various propositions to be mutually related and mutually supportive (e.g., Bartelborth, 1996; Blanshard, 1939; BonJour, 1985; Bradley, 1939; Petraschka, 2014). Corresponding concepts of coherence refer to the number of contradictions, the confirmation of already assumed propositions, the lineage of the propositions, explanatory anomalies, the degree of interrelatedness of the propositions, and many further aspects (e.g., Bartelborth, 1996; BonJour, 1985; Petraschka, 2014). Despite not having been formulated in this way in Geographical Information Science before, the use of intrinsic quality measures presupposes coherence theories of truth (cf., Mocnik et al., 2018). Such measures often refer to the lineage of the data as well as the logical and geometric consistency (e.g., Senaratne et al., 2017). Perspectives that explicitly refer to coherence theories have already been considered in relation to maps (Mocnik, 2023).

### 4 Defining Coherence Theories for Data

Coherence theories of truth have been widely discussed in the context of propositions about the environment. Yet, this is not the case in the context of data when being understood as a collection of symbols, in the sense of the computational interpretation. Only a (mental) interpretation of the data in the context of the geographical environment yields propositions and possibly further information. That is, data can only be *understood* in the context of the geographical environment if they are interpreted. Since coherence theories of truth are, however, based on the coherence of propositions and not on the coherence of symbols, they do not apply to data per se.

The difference outlined between the coherence of symbols and propositions is fundamental. This is aptly illustrated by the example of a text written in a language we are not familiar with. We cannot make sense of neither the symbols nor the composition of the symbols. If the same symbols are used repeatedly and if grammatical structures are recognizable – think here of ‘subject, predicate, object’ as an example – then this can be considered as an indication that the symbols are coherent to some extent. The text does not appear arbitrary. Whether the text describes the geographical environment of an existing place, originates

from the author's imagination, or even represents a collection of symbols without any meaningful interpretation is, however, independent of such structures within the text. Involving a person who knows the language can, nevertheless, provide information about whether the propositions about the geographical environment 'contained' in the text are mutually supportive and even coherent.

Since data largely determine their possible interpretations, it would be desirable to be able to apply coherence theories to data as well. In the example mentioned above, a person with corresponding linguistic proficiency will understand the text in a similar way as other people with the same linguistic proficiency. For the person who is able to interpret data, the data might in many cases and to an extent be synonymous with its interpretation. If one were to define the coherence of data as the coherence of the propositions arising from their interpretation, then this would constitute the basis for a coherence theory for data.

As the example of remote sensing data however shows, there exist multiple interpretations in many cases. Remote sensing images can be interpreted visually; machine learning approaches can automatically recognize objects in these images; and remote sensing images can be used to estimate live green vegetation by means of the normalized difference vegetation index (NDVI), among others. These interpretations can at times differ greatly. As an example, a faulty red band would lead to potentially incorrect information when using the NDVI, but at the same time still allow for a meaningful visual interpretation of the image.

To circumvent these issues, a definition<sup>1</sup> of the coherence of data can be approached in the context of a single interpretation  $\xi$ :

**The  $\xi$ -coherence of data  $A$  is defined as the coherence of the collection of all propositions that emerge from the interpretation  $\xi$ .**

Such definition can be applied independently of how coherence is conceptualized. In particular, it is meaningful even if coherence is not conceptualized as a number but, as in most cases, a complex quality.

Instead of focussing on single interpretations, this definition naturally extends to a collection  $\Xi$  of interpretations:

**The  $\Xi$ -coherence of data  $A$  is defined as the coherence of the collection of those propositions that emerge from at least one of the interpretations  $\xi \in \Xi$ .**

In case it is obvious from the context of the data  $A$  which interpretations are 'meaningful'<sup>2</sup>, the definition can even

<sup>1</sup>The definitions made here are to be understood as provisional and should be examined in more detail for their usefulness in further scientific discourse.

<sup>2</sup>The concept of meaningfulness is referred to here in contrast to arbitrariness. Only those interpretations are considered meaningful that stand out as being more useful than other ones, for instance, because they turn out to be of practical use. These interpretations need to be distinguished from arbitrarily chosen

be made independent of a choice of interpretations. For this purpose, the collection  $\Xi(A)$  of all possible 'meaningful' interpretations will be considered, despite this collection sometimes being difficult to grasp and to potentially only meaningfully exist in some cases. We define:

**The coherence of data  $A$  is defined as the  $\Xi(A)$ -coherence of  $A$ .**

The last two definitions are interesting in that they presuppose the concept of coherence to be applicable to collections of propositions even if these propositions arise from different interpretations. Although this is the case for virtually all concepts of coherence as long as the considered interpretations are commensurable, it often does not apply for corresponding operationalizations due to their dependence on practical context. For example, many intrinsic quality measures are specific to the case of OpenStreetMap and cannot be easily transferred to other use cases (Senaratne et al., 2017; Mocnik et al., 2018). In case that only one meaningful interpretation  $\xi$  exists, that is,  $\Xi(A) = \{\xi\}$ , coherence and  $\xi$ -coherence are equal according to their definitions.

The three definitions of coherence of data as provided here induce coherence theories. These state, analogously to the case of propositions, that data 'apply' to the geographical context if they are coherent. It should be noted, however, that 'apply' has a slightly different meaning here, because data cannot not be assigned truth values in the same way as is the case for propositions.

## 5 The Coherence of Data Used in Black Box Systems

The interpretation of data and thus the way from symbols to propositions about the geographical environment is potentially long and may involve machine processing of the data. In some cases, this is accompanied by the fact that the process of interpreting the data and the resulting propositions can no longer be fully comprehended by humans. Instead, the data is in many cases first processed and converted into new data before it is finally interpreted by humans. Only this final step might be fully comprehensible to humans while previous steps elude human insight. Machine learning-based methods are an example of such processing that is not well comprehensible to humans, because their models 'trained' by sample data consist of symbols that we cannot immediately comprehend. The 'predictions' generated from the model, however, consist of symbols that we ultimately can comprehend, such as in the case of text generated through machine learning, objects recognized in areal imagery, and categorizations of land cover created by means of machine learning.

In such examples, the definitions made earlier in Section 4 come into play. Consider the example of building footprints, which are represented by coordinates in the data.

ones, for which there are just as many reasons as for almost all other interpretations.

It can be argued that there is only one meaningful interpretation of these coordinates. This interpretation is easily comprehensible and might even suggest a direct translation mechanism between the coordinates and corresponding propositions, such as what these coordinates refer to in the geographical environment. Therefore, the difference between data and information is in this case of a more semantic nature. A distinction between the coherence of the two seems thus practically not necessary. However, when the interpretation of data involves black box systems, such translation mechanisms no longer exist. Consider the example of a model trained using machine learning. Such a model is formed by a collection of symbols, which can be used to generate data that can themselves be interpreted in the context of the environment. While it is easily possible to determine which part of the generated data relate to which proposition, this is no longer possible for the model itself in many cases. It seems largely impossible to determine which parts of the model (understood as data) a proposition refers to, or even to understand which individual symbols of the model (and their combination) a proposition generated through the application of the model and subsequent interpretation refers to. The definition of the coherence of data must therefore confine to the data in their entirety in such cases.

The fact that only the coherence of the data as a whole can be considered in many interpretations, such as those involving black box systems, limits the possibilities of a coherence theory. On the level of propositions, coherence theories are often used to improve a set of propositions, such as by checking how well a particular proposition fits into this collection by contributing to the coherence of the collection. In case this one proposition does not contribute much, it might be removed in order to strengthen the coherence of the collection. Due to the lack of correspondence between data and propositions, such removal of data is, however, often not possible if the interpretation involves black box systems. This poses problems, especially in the context of ethical considerations, such as related to the removal of unwanted bias (cf., Turilli and Floridi, 2009; Richardson, 2022; Zhou et al., 2022; Shrestha and Das, 2022).

## 6 Conclusion

The discourse of this paper has achieved three objectives. First, we have argued that data-driven geography requires coherence theories. Secondly, the concept of coherence theories, which is usually only defined for collections of propositions about the geographical environment, has been translated to data. And thirdly, we touched upon the problems that such a concept poses in the context of black box systems and machine learning.

The perspectives discussed demonstrate the need for further research. For instance and most importantly, it seems necessary to better understand how robustly the definitions made here depend on the collection of meaningful inter-

pretations. That is, how much a slight variation of the collection of meaningful interpretations – which interpretation is considered ‘meaningful’ can be judged differently – influences the assessment of the coherence of the data. Likewise, the role of the dichotomy between data and information should be better explored and conceptualized, since only this dichotomy renders the definition of coherence theories for data as presented in this paper necessary.

The epistemological consequences of the data-driven paradigm in geography are still poorly understood. This is partly due to the fact that the validity of propositions derived from data is insufficiently understood. By exploring the potential of coherence theories in relation to data in more depth, it can be hoped to better elucidate the epistemological consequences. For example, examining the influence of the relationship between data and information, which is defined by interpretation, on coherence theories can shed light on the epistemological hurdles that exist in the comprehensibility of the interpretation of data.

The systematic exploration of concrete examples can help to better understand which particular characteristics coherence theories have in case of geographical data. This would enable a better assessment of data quality in geography-related applications, as well as potentially lead to a better understanding of the characteristics of geographical data themselves. Also, it can be hoped that in this way, the implications of the impossibility of assessing the coherence of smaller parts of the data can be disclosed in more detail. In particular, the ethical problems arising from only assessing the coherence of the data in their entirety could be further explored and possibly even mitigated. The prospects of further research related to coherence theories seem manifold.

*Acknowledgements.* I would like to gratefully thank Laura Kühl for the discussions about belief systems and coherence theories in the context of maps and data.

## References

- Ackoff, R. L.: From data to wisdom, *Journal of Applied Systems Analysis*, 16, 3–9, 1989.
- Ballatore, A. and Zipf, A.: A conceptual quality framework for volunteered geographic information, *Proceedings of the 12th Conference on Spatial Information Theory (COSIT)*, p. 89–107, [https://doi.org/10.1007/978-3-319-23374-1\\_5](https://doi.org/10.1007/978-3-319-23374-1_5), 2015.
- Bartelborth, T.: *Begründungsstrategien. Ein Weg durch die analytische Erkenntnistheorie*, Akademie Verlag, Berlin, Germany, <https://doi.org/10.1515/9783050073514>, 1996.
- Blanshard, B.: *The nature of thought*, George Allen & Unwin, London, UK, 1939.
- BonJour, L.: *The structure of empirical knowledge*, Harvard University Press, Cambridge, MA, 1985.
- Bradley, F. H.: *Essays on truth and reality*, Clarendon, London, UK, 1939.

- Chrisman, N. R.: The role of quality information in the long-term functioning of a geographic information system, *Cartographica*, 21, 79–87, 1984.
- Cohen, L. J.: The coherence theory of truth, *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 34, 351–360, 1978.
- Floridi, L.: Data, in: *International encyclopedia of the social sciences*, Vol. 2, edited by Darity, Jr., W. A., p. 234–237, Macmillan, Detroit, MI, 2nd edn., 2008.
- Frank, A. U.: Data quality ontology: an ontology for imperfect knowledge, *Proceedings of the 8th Conference on Spatial Information Theory (COSIT)*, p. 406–420, [https://doi.org/10.1007/978-3-540-74788-8\\_25](https://doi.org/10.1007/978-3-540-74788-8_25), 2007.
- Miller, H. J. and Goodchild, M. F.: Data-driven geography, *GeoJournal*, 80, 449–461, <https://doi.org/10.1007/s10708-014-9602-6>, 2015.
- Mocnik, F.-B.: Why we can read maps, *Cartography and Geographic Information Science*, 50, 1–19, <https://doi.org/10.1080/15230406.2022.2127911>, 2023.
- Mocnik, F.-B., Mobasheri, A., Griesbaum, L., Eckle, M., Jacobs, C., and Klöner, C.: A grounding-based ontology of data quality measures, *Journal of Spatial Information Science*, 16, 1–25, <https://doi.org/10.5311/JOSIS.2018.16.360>, 2018.
- Mocnik, F.-B., Ludwig, C., Grinberger, A. Y., Jacobs, C., Klöner, C., and Raifer, M.: Shared data sources in the geographical domain—a classification schema and corresponding visualization techniques, *ISPRS International Journal of Geo-Information*, 8, 242, <https://doi.org/10.3390/ijgi8050242>, 2019.
- Patterson, D.: What is a correspondence theory of truth?, *Synthese*, 137, 421–444, 2003.
- Petrashka, T.: *Interpretation und Rationalität. Billigkeitsprinzipien in der philologischen Hermeneutik*, de Gruyter, Berlin, Germany, <https://doi.org/10.1515/9783110351163>, 2014.
- Rescher, N.: *The coherence theory of truth*, Oxford University Press, Oxford, UK, 1973.
- Richardson, S.: Exposing the many biases in machine learning, *Business Information Review*, 39, 82–89, <https://doi.org/10.1177/02663821221121024>, 2022.
- See, L., Mooney, P., Foody, G. M., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Miłčinski, G., Nikšič, M., Painho, M., Pödör, A., Olteanu-Raimond, A.-M., and Rutzinger, M.: Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information, *ISPRS International Journal of Geo-Information*, 5, 55, <https://doi.org/10.3390/ijgi5050055>, 2016.
- Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., and Haklay, M.: A review of volunteered geographic information quality assessment methods, *International Journal of Geographical Information Science*, 31, 139–167, <https://doi.org/10.1080/13658816.2016.1189556>, 2017.
- Shi, W., Fisher, P. F., and Goodchild, M. F., eds.: *Spatial data quality*, Taylor and Francis, London, UK, 2002.
- Shrestha, S. and Das, S.: Exploring gender biases in ML and AI academic research through systematic literature review, *Frontiers in Artificial Intelligence*, 5, 976838, <https://doi.org/10.3389/frai.2022.976838>, 2022.
- Stout, G. F.: Immediacy, mediacy and coherence, *Mind*, 17, 20–47, 1908.
- Turilli, M. and Floridi, L.: The ethics of information transparency, *Ethics and Information Technology*, 11, 105–112, <https://doi.org/10.1007/s10676-009-9187-9>, 2009.
- Zhou, N., Zhang, Z., Nair, V. N., Singhal, H., and Chen, J.: Bias, fairness and accountability with artificial intelligence and machine learning algorithms, *International Statistical Review*, 90, 468–480, <https://doi.org/10.1111/insr.12492>, 2022.
- Zins, C.: Conceptual approaches for defining data, information, and knowledge, *Journal of the American Society for Information Science and Technology*, 58, 479–493, <https://doi.org/10.1002/asi.20508>, 2007.