Theoria, 2025, 40(2), 172-186
https://doi.org/10.1387/theoria.27509

Ὁ Θεὸς ἀριθμετίζει
THEORIA
ISSN 0495-4548 – eISSN 2171-679X

Models of data and the representation of phenomena: a pattern-based inference

(Modelos de datos y la representación de fenómenos: una inferencia basada en patrones)

José V. Hernández-Conde*
University of Valladolid
https://orcid.org/0000-0002-8502-6570

María Caamaño-Alegre*
University of Valladolid
https://orcid.org/0000-0002-7005-9257

Keywords

Pattern
Phenomenon
Statistical inference
Nested modeling
Isomorphism

ABSTRACT: Since the 1960s, the distinction between data and phenomena has fueled debates in the philosophy of science, with scholars arguing that data must be modeled in order to serve as evidence for phenomena. We claim that the modeling of data to obtain evidence for phenomena involves four levels: data, sample structure, population structure and phenomena. Our analysis suggests that the notion of pattern is essential to fully grasp the inferential capacity of data models, where representation occurs through nested surrogative reasoning —typically in the form of an isomorphism that holds at different layers. We also explain how our taxonomy of pattern-based inferential steps could shed light on various aspects of nested data modeling, such as the risk of theoretical bias. To illustrate our proposal, we examine the Eddington experiment —which tested general relativity by observing the deflection of starlight near the Sun—, and show how patterns at different levels of the data modeling provide the basis for nested surrogative reasoning in this case. Transforming the data points identified on photographic plates into a representation of light deflection requires a multi-layered search for patterns, where each pattern takes us one step further in data modeling and one step closer to the target phenomenon.

Palabras clave

Patrón
Fenómeno
Inferencia estadística
Modelización anidada
Isomorfismo

RESUMEN: Desde la década de 1960, la distinción entre datos y fenómenos ha dado lugar a amplios debates en la filosofía de la ciencia, habiendo un claro acuerdo en que los datos deben modelarse a fin de que sirvan de evidencia para determinar la ocurrencia de fenómenos. Sostenemos que la modelización de datos, con el fin de obtener evidencia sobre fenómenos, involucra cuatro niveles: datos, estructura de la muestra, estructura de la población y fenómenos. Nuestro análisis sugiere que la noción de patrón es esencial para entender la capacidad inferencial de los modelos de datos, donde la representación se produce mediante razonamientos subrogatorios anidados —habitualmente en forma de isomorfismo entre diferentes capas. También explicamos cómo nuestra taxonomía de pasos inferenciales basados en patrones podría arrojar luz sobre diversos aspectos de la modelización de datos anidada, como aquellos relativos al riesgo de sesgo teórico. Para ilustrar nuestra propuesta, examinamos el experimento de Eddington —que puso a prueba la relatividad general observando la desviación de la luz de las estrellas cerca del Sol—, y mostramos cómo los patrones en diferentes niveles del modelado de datos proporcionan la base para el razonamiento subrogatorio anidado en este caso. Transformar los puntos/datos identificados en las placas fotográficas en una representación de la desviación de la luz requiere una búsqueda de patrones en varios niveles, en donde cada patrón constituye un paso adelante en el modelado de datos y nos acerca un paso más al fenómeno objetivo.

* ***Correspondence to:*** José V. Hernández-Conde. Department of Philosophy, Faculty of Philosophy and Arts, University of Valladolid, Plaza del Campus s/n, 47011, Valladolid, Spain – jhercon@uva.es – https://orcid.org/0000-0002-8502-6570

* ***Correspondence to:*** María Caamaño-Alegre. Department of Philosophy, Faculty of Philosophy and Arts, University of Valladolid, Plaza del Campus s/n, 47011, Valladolid, Spain – mariaconcepcion.caamano@uva.es – https://orcid.org/0000-0002-7005-9257

How to cite: Hernández-Conde, José V.; Caamaño-Alegre, María (2025). «Models of data and the representation of phenomena: a pattern-based inference»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 40(2), 172-186. (https://doi.org/10.1387/theoria.27509).

Received: 28 April, 2025; Final version: 24 July, 2025.
ISSN 0495-4548 - eISSN 2171-679X / © 2025 UPV/EHU Press

This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

1. Introduction

Scientific representation encompasses a wide and diverse range of practices. This paper narrows its focus to a specific domain within that practice: one in which models of data play the key representational role, by enabling inference from data to phenomena.

Since at least the 1960s, the distinction between data and phenomena has given rise to important debates in the philosophy of science. Both Suppes (1962) and, later, Leonelli (2009, 2015, 2019) argue that data can only constitute evidence for phenomena once they have been modeled, that is, assembled and arranged in a particular way. In a similar vein, Suárez (2024) has recently characterized the models of data, within certain domains of inquiry, as low-level mathematical models that represent those abstract features of the data that are relevant for the purposes at hand. His inferential conception of scientific representation emphasizes how mathematical models —such as statistical models of data— involve nested surrogative reasoning between their different layers (Suárez, 2024, p. 167).

Our purpose is to show the importance of the notion of pattern for understanding the inferential capacity of data models. We contend that, despite the fact that patterns play a crucial a role in the inference from data to phenomena, the notion of "pattern" has not been deeply investigated from an analytical point of view. This has often led to an ambiguous use of this notion, as noted by Bruce Glymour (2000), who proposes replacing the distinction between data and phenomenon with a classical distinction in statistics, namely, that between sample and population structure. We will suggest a revision of Glymour's proposal, advocating the need to distinguish between four distinct levels, i.e., data, sample structure, population structure and phenomena, and will explain the character and nature of each step necessary to move from one level to the next. We argue that the inferential, representational capacity of data models can be better understood as a three-step inferential process involving the above four levels. Our claim is that this multilayered account of data models can be useful to account for the kind of nested surrogative reasoning invoked by Suárez (2024) in his characterization of models of data.

From what has been said, it is clear that our concept of pattern is statistical-mathematical. The present analysis of the notion of pattern is limited to cases where the term "pattern" is applied to something that can be derived from a set of data. We stress the importance of patterns in identifying regularities in mathematical representations. These regularities can be expressed numerically. We distinguish three different senses in which the term 'pattern' is be used: (a) mathematical, as a distinction between signal and noise;[1] (b) statistical, which includes both descriptive and inferential uses of statistics; and (c) geometric or informational, as found in DNA chains, the growth patterns of shells and Brussels sprouts. The main sense in our proposal combines (a) and (b). Patterns that are neither mathematical nor statistical in nature —sense (c)—, such as those associated with the identification of the structure of DNA fall outside the scope of our article, since they determine an entity rather than a phenomenon. Consider X-ray crystallography, where evidence famously led Rosalind Franklin and her collaborators to attribute an antihelical structure to the 'dry' form of DNA. In this case, the properties of an entity are established based on the patterns seen in the photographs. However, no specific phenomenon is identified, unlike the melting point of lead, the deflection of light, or linear learning.[2]

We will illustrate the pattern-based inferential steps from data to phenomenon by using a well-known example (Laymon, 1982; Mayo, 1996; Kennefick, 2009, 2019), namely, the so-called "Eddington experiment", which was intended to empirically test the theory of general relativity by confirming the prediction about the deflection of light. In this case, the small sample size entails special difficulties that can only be overcome by introducing local

[1] This is the prevailing sense in McAllister's (1997) discussion of the inference from data to phenomena. While we share some of his concerns about the indeterminacy of this process, we also believe that our approach can also help to address this issue. However, we will leave this for another paper.

[2] See Schindler (2008; 2013, pp. 95-96).

checks and plausible assumptions about how the instruments function. This made it possible to determine a sufficiently representative sample pattern, i.e., the deviation of the points of light captured in the photographs from the center of the Sun as it appears in them. Based on this, an inference can be made about the phenomenon of light's path deviating when passing close to the Sun. This final step involves making a causal inference, since phenomena are conceived as the (often unobservable) cause of the observed data. Our example also shows how patterns, in the statistical-mathematical sense, play a crucial role in nested surrogative reasoning.

Our approach in this paper unfolds in three steps. First, we offer a concise overview of our perspective on the inferential process from data to phenomena. The next step is to examine how our framework can be useful in understanding data models that involve nested surrogative reasonings between their different layers. Finally, we provide an example to show how, in models of data such as those at play in Eddington experiment, the surrogative reasoning characteristic of representation manifests as isomorphisms holding between different layers of those data models.

2. *Bridging the gap between data and phenomena*

2.1. The data/phenomenon distinction

To frame the problem of patterns, we first examine the debate over the distinction between data and phenomena, where the concept of "pattern" proves highly relevant. As Woodward (2010, p. 792) points out, the introduction of this distinction was motivated by its relevance in addressing the epistemic problem of using data as evidence. The discussion stems from a question that Antonis Antoniou has recently articulated very clearly:

> How exactly are theoretical hypotheses eventually confronted and tested by experimental results given that the latter are often produced in the form of large datasets and in a language that is not accessible to the theory? (Antoniou, 2021, p. 2)

This question has been central in contemporary philosophy of science, where it is widely accepted that empirical phenomena emerge from a complex collection of data (Bogen & Woodward, 1988; Woodward, 1989, 2000, 2010; McAllister, 1997, 2011; Glymour, 2000; Massimi, 2007; Leonelli, 2015, 2019; Bokulich, 2018, 2020). Since the early characterization of the notion of data models by Patrick Suppes (1960, 1962), significant efforts over recent decades have explored how data models are applied in specific research contexts to infer phenomena from data (Leonelli, 2009; Antoniou, 2021; Bokulich & Parker, 2021).

All of the aforementioned authors agree that raw, unprocessed data do not serve as evidence. This debate already has a long history, which we do not aim to examine in detail. Rather, we assume the widely accepted view that scientific observations and experiments rarely allow for the direct observation or detection of phenomena. Instead, they detect certain effects that are captured in the form of data, which can, in principle, be attributed to the target phenomena. The epistemic problem lies in determining the occurrence of stable phenomena from a heterogeneous and unstable set of data, which almost always occurs in combination with noise and whose understanding is also limited (Woodward, 2010, p. 794).

Suppes addressed data models in the 1960s, conceptualizing them as statistical constructs that distill complex experimental processes, whose purpose is to facilitate comparisons between empirical results and theoretical predictions. Although Suppes did not explicitly focus on patterns, his emphasis on homogeneity, stationarity, and order tests suggests an implicit recognition of their importance in data modeling. Under his view, it is the patterns, not the raw data set or unmodeled experimental results, what can be interpreted as phenomena, whether consistent with predictions or not.

Later, Bogen and Woodward reinforced the distinction between data and phenomena. Arguing against van Fraassen's (1980, 1989) view, they held that theories are primarily concerned with unobservable phenomena, rather than observed data (Bogen & Woodward, 1988; Woodward, 1989). According to Bogen and Woodward, data are unstable, observable reports from experiments, unlike phenomena, which are stable, theory-explained features of the world inferred from reliable data patterns. In contrast to data, phenomena are unobservable but serve as evidence for theories because they are the underlying cause of observed data.

More recently, discussions have been focused on how, within different fields, scientists generate and process data to identify empirical phenomena, as exemplified by Leonelli's work in biology (2015, 2016, 2019) and Bokulich's (2018, 2020) in paleontology. Leonelli (2019) views data models, unlike unmodeled data, as actual, ready-to-use evidence, designed to systematically organize data for representing phenomena and supporting evidential claims. Similarly, Bokulich (2020) contends that data models surpass "raw data" in accuracy and reliability, and thus are more appropriate tools to science's epistemic goals. In this vein, Leonelli (2009, 2015, 2019), and Bokulich and Parker (2021) highlight the pragmatic nature of data models, as specifically designed tools to yield particular evidence to support a certain theory.

In McAllister's 1997 and 2011 works, the notion of pattern is central to his analysis of reasoning from data to phenomena, in an approach that is focused on the issue of the extreme underdetermination in this inference. As a result, he does not explicitly define patterns, beyond equating them with the notion of signal in a conception where data is split into signal and nose, and phenomena are identified with signals. In McAllister's view, the identification of patterns —or signals— depends on how data are decomposed into signal and noise, which allows for multiple interpretations of the same original data set. This leads to the concern that inference from data to patterns —and ultimately to phenomena— is fundamentally undermined by the ambiguity of the data, since any data set can yield an indefinite number of patterns.

Particularly interesting for our proposal is Glymour's (2000) critique of the vague use of the term in the data-to-phenomena inference debate, and his suggestion of replacing the philosophical distinction between data and phenomena with the statistical concepts of sample structure (that is, patterns recognized in the data sample) and population structure (that is, population model identified with the phenomenon). According to Glymour, the latter distinction offers greater precision, with sample structure consisting of statistical measures derived from the collected data, and population structure representing the generalization of these measures to the entire population. As a result, the inference from data to phenomena is seen as a two-step process: first, a pattern is identified within the sample, and second, the identified pattern is extrapolated to the entire population and used to define the phenomenon under study.

However, while Glymour's focus on statistical distinctions is valuable, his analysis conflates data with sample structure and population structure with phenomena, leaving the two senses of pattern present in the discussion unclear and mixed with other separate notions. In this regard, our analysis emphasizes the mediating role of patterns as essential for inference from data to phenomena, and the transformation of raw data into interpretable structures that, once generalized, provide a basis for characterizing the underlying phenomenon.

2.2. Patterns as mediators between data and phenomena

Patterns are essential to both statistics and contemporary scientific research. From a philosophical and statistical point of view, patterns are sets of properties in the form of regularities that characterize the members of a particular group. Recognizing patterns is the goal of several statistical methods, including cluster analysis, discriminant analysis, decision trees, and artificial neural networks. These approaches help distinguish between categories and uncover relationships between indirect measures and the phenomenon under investigation. Examples include discharge patterns in electronic particle detectors and reaction times in psychological studies (Bogen & Woodward, 1988).

We will now show how to distinguish data from patterns and patterns from phenomena by defining specific meanings of "pattern" and organizing the analysis into three sequential steps.

Step (a): From data to patterns in the data (sample statistics)

The initial step involves transforming raw empirical data into structured and interpretable sample statistics. This step involves processing raw empirical data collected in an experiment to identify statistical patterns, which includes computing descriptive statistics such as means, medians, standard deviations, correlations, and visual representations (e.g., histograms), together with data cleaning, variable selection, and outlier removal are performed to refine the data set.

This process is both statistical-methodological and partly epistemological. The raw data, often messy and unstructured, is subjected to cleaning (e.g., removing outliers, handling anomalies) and summarization (e.g., means, medians, correlations) to reveal trends or patterns. These patterns are not yet general claims about the world, nor claims about the target phenomenon; they are descriptive abstractions of the sample at hand. Techniques like histograms or preliminary clustering help distinguish signal from noise, making this a foundational act of pattern recognition. As a result, this step helps differentiate meaningful patterns from noise and provides a summarized representation of the sample behavior.

This first step has an interesting dual character. On the one hand, it is technical (relying on statistical-mathematical tools) but also it is interpretive, as decisions about which variables to prioritize or how to handle inconsistencies unavoidably shape the resulting patterns. This dual character is the norm in contemporary scientific practice, where data preprocessing is as critical as the analysis itself.

Step (b): From sample to population patterns (population statistics)

The second step generalizes the patterns found in the sample to the whole population. The transition from sample statistics to population statistics introduces a leap of generalization, which has a statistical-epistemological character. It requires assumptions about sample representativeness and usually involves techniques such as statistical inference, confidence intervals, and hypothesis testing. On this basis, this step hinges on inference —extrapolating patterns observed in a limited sample to a broader population. Statistical tools such as confidence intervals and hypothesis testing quantify the uncertainty of this jump, despite the fact that the inference is based on an assumption that cannot be conclusively proven, such as the representativeness of the sample.

As a result, the transition from sample statistics to population statistics introduces an epistemological gap, since it relies on probability-based estimations and, more importantly, on extrapolations. Unlike sample statistics, which are directly derived from data, population statistics remain theoretical and require careful validation. This point is where the epistemological weight becomes evident. Unlike step (a), which is grounded in the concrete data collected, step (b) moves into the realm of the unknown, estimating population parameters that cannot be directly observed. Despite this or, precisely, because of it, this step is crucial for science's aim to produce universal assertions and laws, and ultimately causal explanations that go beyond the limits of empirical knowledge.

Step (c): From population patterns to phenomena

The third step interprets population patterns as empirical representations of particular phenomena. Thus, this step moves from population statistics to the empirical characterization of the phenomenon itself, which entails a causal inference from data to phenomena (i.e., the phenomenon is conceived as the —often unobservable— cause of the observed data).

As a result, patterns attributed to the population are interpreted as manifestations of an underlying phenomenon. This is a distinctly epistemological move, as it involves hypothesizing a causal relationship between the observed patterns and the “hidden” reality they reflect. Researchers may use decision trees, cluster analysis, or dimensional reduction techniques —such as factor analysis or principal component analysis (PCA)— to identify key explanatory factors. However, the interpretive role of the researcher is critical —deciding which variables best capture the phenomenon.

Two-step or three-step approach?

Finally, depending on the analytical method adopted, the transition from data to phenomenon can follow either a three-step approach —progressing through sample statistics, population statistics, and empirical characterization of the phenomenon— or a two-step approach, where dimensional reduction could allow a direct identification of the phenomenon at the sample level.

The two-step approach —using techniques such as factor analysis or principal component analysis (PCA) directly on the cleaned data— constitutes a significant alternative. By identifying underlying factors or components early on, this approach collapses the three-step process into two: (1) from the data to the sample-level phenomenon, via dimensional reduction, and (2) from the sample-level phenomenon to the population-level phenomenon, via statistical validation. This approach bypasses the explicit construction of sample statistics as an intermediate step, and instead searches for the phenomenon in the underlying structure of the data.

3. *Inferential steps, patterns and nested modeling*

In the following sections, we will examine how our approach can reveal the nested nature of data models and the crucial role of patterns in the inferences that lead to the representation of phenomena. To this end, we will rely on Suárez’s view of models of data, which explicitly incorporates the concept of nested reasoning —a notion that we also deem to be important.

3.1. Nested modeling in Suarez’s view of data models

Although Suárez’s inferential account of representation excludes isomorphism between source and target structures as an essential condition for representation, he nevertheless acknowledges its significant role as a means of scientific representation (Suárez, 2024, pp. 167-168). He emphasizes that surrogate reasoning, when based on theoretical and mathematical models, is often facilitated by isomorphisms between source and target. Among the various domains of modeling where isomorphism is most likely to play this role, Suárez draws attention to the domain of data models. Thus, despite his rejection of isomorphism as a constitutive element of representation, he regards data models as a type of model in which isomorphism is most plausibly the means of representation.

Suárez (2023, pp. 46-54; 2024, p. 167) characterizes models of data as low-level mathematical constructs that represent abstract and relevant features of the data, often through isomorphisms. In his view, mathematical models of data are multilayered and involve nested surrogate reasoning between their different layers. Consequently, he refers to both isomorphism and nested modeling as features of data modeling, which together help to enable surrogative reasoning from source to target. In this context, he states the following:

> We saw that, while mathematical models —such as the Lotka-Volterra model or statistical models of data— are prima facie candidates for isomorphic relations, they turn out in more complex ways to involve nested surrogative reasonings between their different layers. These models are again no objection to [inf] since any means to promote surrogative reasoning from source to target is compatible with the source’s inferential capacities. (Suárez, 2024, p. 167)

In an earlier paper, Suárez (2023) examined an example of nested modeling in the field of asteroseismology —also discussed by Castellani & Schettino (2023). In this case, data models involve the identification of fluctuation patterns in the luminosity of stars which, supported by multiple modeling assumptions, allow for the inference of seismic oscillations within them. These seismic oscillations, in turn, provide highly valuable evidence for models of stellar structure. The data models that enable the identification of seismic oscillations are nested in a complex background of models, which Suárez analyzes in detail (2023, pp. 120-123). The expression 'nested models' intuitively refers to the idea that, in some contexts of inquiry, "(...) all models contain assumptions that involve or result from other models" (Suárez, 2023, p. 112). Suárez characterizes nested modeling not only as a presuppositional relationship between models, but also as one that involves the use of real (collected) data to enrich the empirical interpretation of theoretical models and to develop them further —thereby allowing the data to be inferred or estimated through these enriched, more sophisticated theoretical models. More specifically:

> It concerns the nested nature of modelling; the fact that most models operate against a background that incorporates further models, and where data is routinely tested backwards, as it were —by deducing the values of theoretical parameters from the data given some background assumptions, and where those parameters are then fed into new models that can appropriately generate the data. (Suárez, 2023, p. 112).

It is worth noting a peculiarity of the case of asteroseismology that is shared with other cases from geology and engineering, namely the use of forward or inverse modeling to explore how different modeling alternatives yield different estimates of the data. Stellar structure models serve as theoretical building blocks that provide preliminary and partial guidance for interpreting brightness patterns in terms of seismic oscillations. This guidance is subject to revision if the data do not agree with the predictions of the theoretical stellar structure models (*ibid.*, pp. 125-126). In the example analyzed by Suárez, the evidential —or "observational"— basis of stellar astrophysics is established by means of a complex combination of nested models required to obtain and interpret data on stellar seismic oscillations. These data are derived from the fundamental phenomenon of regular pulsations in the brightness of a star, which are caused by gravitational or acoustic oscillations generated by rotational or convective forces. Further modeling assumptions are required to go from the data on seismic oscillations to the oscillation patterns of nodes (radial and non-radial) for both pressure and gravity modes (*ibid.*, p. 121). The overall oscillatory profile of a star provides highly valuable evidence for evaluating not only stellar structure models, but also traditional modeling assumptions in the form of idealizations. Suárez emphasizes the need to re-examine the nested models once the data modeling is complete, especially if the results suggest that certain assumptions should be reconsidered. In the case of asteroseismology, assumptions such as blackbody radioactive equilibrium, spherical symmetry and uniform composition may be challenged by the empirical evidence gathered.

Suárez's description of nested models operating in the above case suggests a kind of circularity; however, he claims that this does not amount to an experimenter's regress in the sense described by Collins (1992). The latter would involve the assumption of theoretical models in data modeling, thereby (partly) determining the very evidence used to test those same theoretical models. As is well known, Collins' classic example is the detection of gravitational waves. In his view, theoretical assumptions —such as those about stellar structure— may be useful for establishing some tentative constraints on how data should be derived or explained, but this would be part of a wider procedure aimed at comparing actual data, obtained by independent means, with data inferred under certain theoretical assumptions. In Collin's example, theoretically predicted oscillation frequencies are compared with actual observed oscillations, and discrepancies lead to adjustments in the theoretical parameters or background models used to make the predictions. As Suárez argues, rather than introducing a theoretical bias, the nested structure of data models allows for a more thorough assessment of how evidence is constructed and evaluated.

3.2. Nested modeling as a pattern-based inference

We believe that our approach, which is in line with Suárez's view in several ways, can provide valuable insights into how data models are characterized. Although our approach is not committed to a particular theory of representation, it reinforces the idea that isomorphism plays a major role in how data models represent phenomena. In particular, we emphasize the importance of patterns in capturing regularities in mathematical representations, which makes them particularly well-suited for representing phenomena using isomorphisms. In our view, the notion of pattern is essential to understanding the inferential capacity of data models. Furthermore, we fully back the idea that, in typical cases of data modeling, surrogate reasoning —characteristic of scientific representation— relies on isomorphisms that hold across different layers of the model.

The Eddington experiment is also useful to see how, in principle, the nested modeling required for the development of data models is not necessarily subject to the problem of experimenter's regress. For the reasons set out by Kennefick (2019) and emphasized by Suárez (2025), we believe that the concerns raised by Earman and Glymour (1980) regarding Eddington's experiment are misguided. As convincingly argued by Kennefick in a chapter entitled "Not only because of theory" (2019, pp. 181-225), the background models supporting the light deflection inference were independent of the theoretical models from general relativity that led to Einstein's prediction. Although general relativity and Einstein's predictions were taken into account in drawing conclusions from the data, the data themselves were generated independently and stood on their own. Neither the generation of the data nor their selection and modeling depended on theoretical assumptions from Einstein's theory. Kennefick provides historical evidence to show that even if both Eddington and Dyson admired the theory as a theoretical achievement, neither of them assumed its adequacy in advance. It is therefore inaccurate to attribute to them a confirmation bias in data selection, as Earman and Glymour do, because neither of them assumed the adequacy of the theory. Other reasons given by Kennefick regarding the methodology used to select data further support the claim that the selection process was unbiased. An analysis of the nested models involved in this case sheds light on the quality of the data selected and the methodological soundness of their modeling. The discarded data —those more consistent with Newtonian predictions— came from photographic plates where interfering factors, known to affect the operation of the instruments, had clearly degraded the image quality. If assumptions about confounding factors were critical for selecting reliable data, assumptions about confounding factors were equally important for modeling the data to identify a pattern in the shift of starlight. As we will see in the next section, Einstein's prediction was only confirmed when the effect of the main confounding factor —the scale shift— was calculated and subtracted from the original observed pattern.

In the next section, we show what this inferential practice involving different layers would look like if our characterization of data models were supplemented with some insights from Suárez's approach. In applying our taxonomy of inferential steps, we show how the presuppositional relations between layers in nested modeling can be seen as corresponding to the presuppositional relations between layers in inference from data to phenomena via patterns. In both cases, models are built on a background of other models, and the quality of evidence can only be properly assessed by tracing the lineage of modeling assumptions involved in the nested structure.

4. *Application Example: Eddington Experiment*

4.1. Experiment presentation

An important example of inference from data to phenomena is the "Eddington experiment", designed to empirically test Einstein's general relativity by confirming the predicted deflection of starlight near the Sun during the 1919 solar eclipse starlight (Laymon, 1982, pp. 108-113; Mayo, 1996, pp. 132-139, chapter 8; Kennefick, 2009, 2019; Suárez, 2025). Woodward (2000, p. 164; 2010, pp. 792-797) frequently mentions this case to illustrate how data from photographic plates mix the target variable (light deflection) with confounding factors such as telescope peculiarities, thermal effects on instruments and chemical influences during plate development. As Woodward (2010, pp. 793, 797) argues, understanding these additional variables is essential to assess the reliability of inferences from data to phenomena. In our analysis, this knowledge aids in identifying patterns that serve as evidence for the phenomenon under study, and applying the concept of "pattern" enhances the precision of such inferences.

As is well known, British astronomers Frank Watson Dyson and Arthur Stanley Eddington organized two expeditions to gather evidence from the 1919 eclipse, in order to test Einstein's prediction of light bending near a massive body (Laymon, 1982; Mayo, 1996). The method relied on discrete data, namely, star positions on photographic plates taken during the eclipse and compared with night-time plates of the same field, captured months later when the Sun had shifted from the Hyades. Three data sets were produced by the two expeditions, the first was to Sobral in Brazil and the second was to the island Príncipe off the West African coast. Data sets 1 and 2 were generated using the 4-inch Sobral telescope and the Sobral astrograph, respectively. Data set 3 was produced with the Príncipe astrographic telescope. The data were cleaned and selected, excluding data set 2, as well as light point positions from photographic plates with images considered defective due to astronomical or atmospheric effects.[3]

In this experiment, the inference process involves a key pattern: the sample pattern, that is, the observed displacement of star positions —i.e., light points— in the eclipse photographs relative to the Sun's center. This pattern was used to infer the phenomenon of starlight deflection near the Sun. It took decades to expand the sample size to obtain a more robust, generalizable population pattern.[4] However, thanks to the efforts made in controlling secondary factors and analyzing sources of error, the sample pattern was found to be a good representation of the phenomenon. Although the similarity between the Príncipe and Sobral observational patterns did not in itself justify their selection, they were chosen (on the basis of other factors) for the subsequent extrapolation of their results to a broader population —namely, all instances where starlight passes near a massive body. In this case, the epistemological factor had much more weight than the statistical one, since the sample size (i.e., the number of locations from which the eclipse observation was made) was small, making the statistical dimension less relevant than it would have been had there been more locations. Although the photographic plates were considered representative, the astronomers recognized the need for further validation and confirmed the pattern's consistency through additional observations (Kennefick 2009: 38). This is why there is a two-step pattern-based inference in this example instead of a three-step one.

The first step involved identifying the sample pattern by superimposing eclipse and comparison plates, cleaning defective data (e.g., due to atmospheric effects), and isolating light deflection from confounding factors like scale shift —a radial displacement mimicking deflection. Unlike deflection, which peaks near the Sun's limb and diminishes with distance, scale shift affects stars farther from the image center most strongly. Hence, a key distinction exists between the two effects. Light deflection is most pronounced for stars near the Sun's limb and diminishes with distance, whereas scale shift has the greatest impact on stars far from the image center, with minimal effect near the center.

To accurately isolate the pattern tied to light deflection and rule out scale shift as the cause, calculating the scale shift effect is essential. The Cambridge team addressed this by taking "check plates" of a different star field at night, both in the UK and Príncipe. These plates served to detect significant scale changes between the eclipse and comparison photographs, enabling Eddington to account for variations arising from differences in timing, location, or

[3] As data sets 2 and 3 were both of poor quality, there has been a heated debate as to whether excluding only data set 2 —which was more in agreement with Newton's prediction— was a biased decision. While some argue that it was (Earman & Glymour, 1980), others oppose this idea (Kennefick, 2009, 2019; Schindler 2013), albeit for different reasons. For a brief overview of our position on this matter, see our comments in Section 3.2 above.

[4] For a survey of light deflection measurements from 1919 to 1960 see von Kluber (1969, pp. 58-63).

equipment setup. Using the positions on the check plates (white positions in the image), Eddington quantified the scale shift and subtracted its effect from the light point displacements observed when superimposing the eclipse and comparison plates. This subtraction isolated the light deflection effect, represented as the deviation α between the black and white positions of the same stars shown in Figure 1. Thus, in this experiment selecting the appropriate data and controlling the primary confounding factor (i.e., scale shift) are pivotal in identifying the sample pattern associated with light deflection.

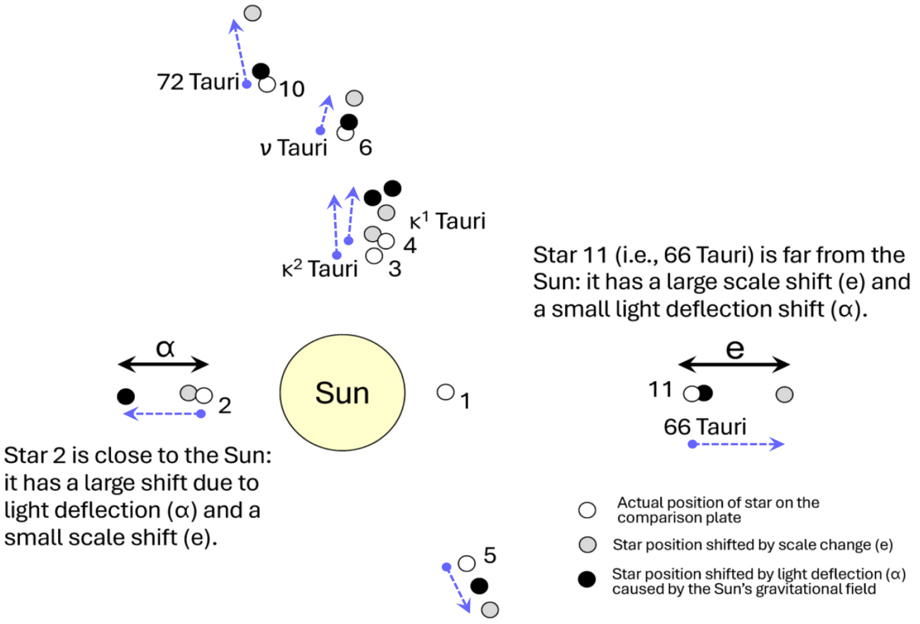
72 Tauri
10
ν Tauri
6
κ¹ Tauri
4
κ² Tauri
3
Star 11 (i.e., 66 Tauri) is far from the Sun: it has a large scale shift (e) and a small light deflection shift (α).
α
2
Sun
1
e
11
66 Tauri
Star 2 is close to the Sun: it has a large shift due to light deflection (α) and a small scale shift (e).
○ Actual position of star on the comparison plate
○ Star position shifted by scale change (e)
● Star position shifted by light deflection (α) caused by the Sun's gravitational field
5

Figure 1

Position of the stars closest to the Sun during the 1919 eclipse.
[Diagram adapted from Kennefick (2019, p. 191)]

The second step entails inferring the phenomenon from the established population pattern via causal inference. This step confirmed Einstein's prediction of light bending by interpreting the generalized pattern in step two as evidence of the phenomenon. Using standard astronomical statistical methods, the light point displacements were analyzed as deviations. The results revealed Sobral's deviation angle as 1.98 ± 0.12 arcsec and Príncipe's as 1.61 ± 0.30 arcsec. Given the Newtonian prediction of 0.87 arcsec for light grazing the Sun's surface, these findings aligned with Einstein's predicted value of 1.75 arcsec, confirming the phenomenon of light deflection.

4.2. Interpretation in terms of nested modeling and isomorphisms

Let us now examine how Eddington's experiment can be interpreted in terms of nested modeling and isomorphisms. Eddington's experiment exhibits the typical presuppositional relationships between models that characterize nested modeling. Rather than focusing on the theoretical model as the final result, we concentrate on the empirical model that underpins the observation of light deflection. This empirical phenomenon is established on the basis of other models, with the model representing the former being nested within those supporting the latter. In what follows, we will schematically outline the source-target relation at each level of modeling, along with the relation that enables the representation of the target by the source. To fully account for the role of isomorphism, we introduce an initial step that was ignored in our previous analysis.

The starting point of data handling in the experiment is the selection of photographic plates, from which the sample data are selected. In this case, the SOURCE$_{(0)}$ is the relative positions of light points in selected photographic plates, the TARGET$_{(0)}$ is the relative positions of stars in Hyades clusters, and the ISOMORPHISM$_{(0)}$ taking place between SOURCE$_{(0)}$ and TARGET$_{(0)}$ is the means to infer that points in the plates correspond to light coming from the stars (see Figure 2).

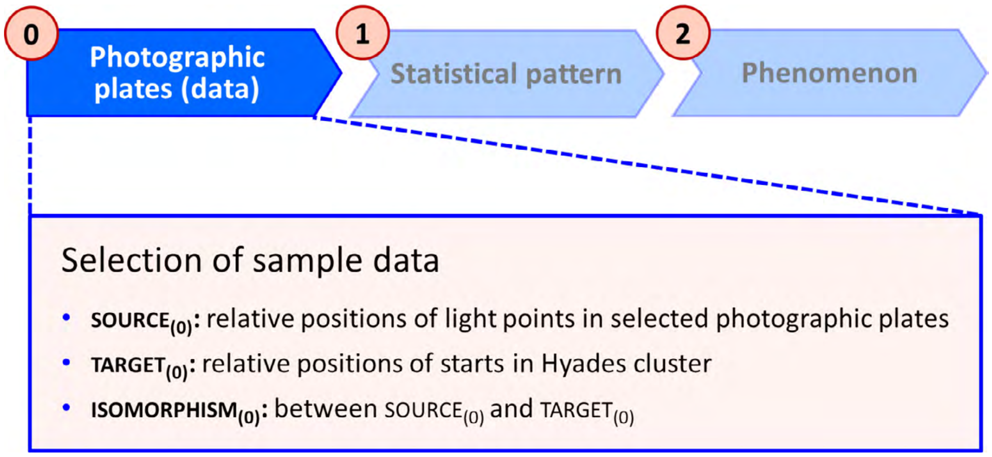
0
Photographic plates (data)
1
Statistical pattern
2
Phenomenon
Selection of sample data
• SOURCE$_{(0)}$: relative positions of light points in selected photographic plates
• TARGET$_{(0)}$: relative positions of starts in Hyades cluster
• ISOMORPHISM$_{(0)}$: between SOURCE$_{(0)}$ and TARGET$_{(0)}$

Figure 2

Isomorphism in the selection of sample data. [Initial step]

Then, researchers identified a statistical pattern in a sample of photographic plates. In this initial step, the SAMPLE$_{(1)}$ are the photographic plates (i.e., the relative positions of light points in selected plates, which was SOURCE$_{(0)}$), the SOURCE$_{(1)}$ is the statistical patterns of displacement of the light points in the photographic plates relative to the center of the Sun —as depicted in Figure 1—, and the TARGET$_{(1)}$ is the light's behaviour displacement in the sample depending on whether it passes or not close to the Sun. ISOMORPHISM$_{(1)}$ takes places between the SOURCE$_{(1)}$ (i.e., sample statistics from the places) and the TARGET$_{(1)}$ (i.e., trajectory of light in the observational scene), enabling the inference from source to target (see Figure 3). However, no isomorphism takes place between the SOURCE$_{(1)}$ (i.e., sample pattern) and the SAMPLE$_{(1)}$ (i.e., the photographic plates).

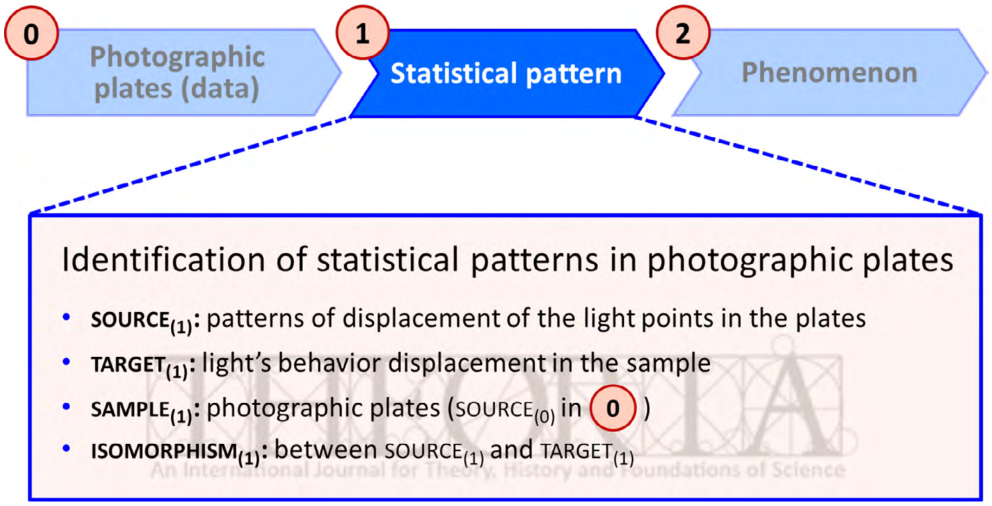
0
Photographic plates (data)
1
Statistical pattern
2
Phenomenon
Identification of statistical patterns in photographic plates
• SOURCE$_{(1)}$: patterns of displacement of the light points in the plates
• TARGET$_{(1)}$: light's behavior displacement in the sample
• SAMPLE$_{(1)}$: photographic plates (SOURCE$_{(0)}$ in 0)
• ISOMORPHISM$_{(1)}$: between SOURCE$_{(1)}$ and TARGET$_{(1)}$
THEORIA
An International Journal for Theory, History and Foundations of Science

Figure 3

Isomorphism in the identification of statistical patterns. [Step 1]

Steps b) and c) in our characterization of the inference from data to phenomenon (see Section 2.2) have no direct equivalents in the case of Eddington's experiment, since no additional instances of light behavior during eclipses were available at the time. In this historical example, population patterns do not come into play to infer the phenomenon —namely, the deflection of light by the Sun. However, the extrapolation of the result is indeed pursued. In this case, the full generalizability of the detected phenomenon requires steps (b) and (c) to be carried out subsequently, once new samples become available.[5] This would allow the inference of a population pattern generalizable across cases of the same type, by extending the statistical patterns of displacement observed in the photographic plates to patterns of displacement expected in the entire population. In this second step, an isomorphism holds between the population patterns and the general phenomenon of light deflection beyond Eddington's specific observational context. Moreover, a homomorphism[6] is present between the statistical patterns and the population patterns, and it is this homomorphic relation that licenses the inference from source to target.

Returning to the actual development of the inference in our example, the next step involves inferring the cause of the sample pattern —namely, the phenomenon of light deflection as it travels near the Sun. In this case, the SOURCE$_{(2)}$ is a particular value of starlight deviation inferred from the statistical pattern (see Figure 4), while the TARGET$_{(2)}$ remains the same as TARGET$_{(1)}$— namely, the deflection of starlight as it travels near the Sun.

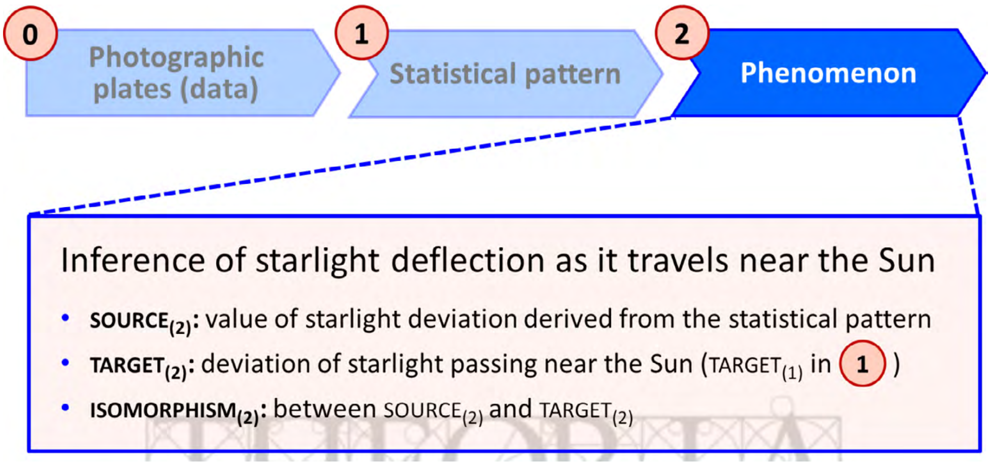
0
Photographic plates (data)
1
Statistical pattern
2
Phenomenon
Inference of starlight deflection as it travels near the Sun
SOURCE(2): value of starlight deviation derived from the statistical pattern
TARGET(2): deviation of starlight passing near the Sun (TARGET(1) in 1)
ISOMORPHISM(2): between SOURCE(2) and TARGET(2)

Figure 4

Isomorphism in the inference of starlight deflection when travelling near the Sun. [Step 2]

[5] The frequency of this type of case depends largely on the extent to which sample collection is constrained (i.e., on whether there are restrictions on extending the sample). When samples are readily available and affordable, the steps involving population patterns proceed as outlined in Section 2.2, that is, population patterns are inferred from sample patterns and serve as the basis for inferring the phenomenon. But even in the present case, the inferred particular phenomenon is considered as an instance of a general phenomenon that could be explained by general relativity.

[6] A homomorphism of systems is a weaker mapping relationship between two systems that preserves the structure of the source system in the target system. More formally, there is a homomorphism between two systems S and T, if f: $S \rightarrow T$ is a function that respects the relationships and operations defining the behavior and structure of system S within system T. On the other hand, an isomorphism is a homomorphism f that is injective (each component in S maps to a unique component in T), and surjective (each component in T has a preimage in S).

5. *Conclusions*

The preceding discussion emphasized the role of patterns as mediating in the inference from data to phenomenon. Our analysis reinforces the idea that data models are nested within background models that significantly influence data selection and processing —from assumptions about boundary conditions and control of hidden variables, to the statistical treatment required for pattern identification. We conclude that patterns play a crucial role in bridging data and the identification of phenomena, enabling surrogative inference through isomorphic relationships between source and target at each level of modeling.

Acknowledgments

Our work has greatly benefited from the helpful discussions we held in our reading group on Models of Data about 2024, especially those with Mauricio Suárez and Adán Sus. We would like to thank M. Suárez also for inviting us to contribute to the Symposium on Inference and Scientific Representation, at the *XI Congress of the Society for Logic, Methodology and Philosophy of Science in Spain*, where we got valuable feedback on our paper. Thanks also to Gui Sanches de Oliveira, as well as Aníbal M. Astobiza, for sharing interesting ideas on the subject of scientific representation.

This research was financially supported by the research project *The Philosophy of Statistical Modeling Science*, (PID2021-126416NB-I00, Spanish Ministry of Science and Innovation).

REFERENCES

Antoniou, A. (2021). What is a data model? *European Journal Philosophy of Science, 11*(4), 1-33. https://doi.org/10.1007/s13194-021-00412-2

Bogen, J. & Woodward, J. (1988). Saving the phenomena. *Philosophical Review, 97*(3), 303-352. https://doi.org/10.2307/2185445

Bokulich, A. (2018). Using models to correct data: Paleodiversity and the fossil record. *Synthese 198* (Supl. 24), 5919-5940. https://doi.org/10.1007/s11229-018-1820-x

Bokulich, A. (2020). Towards a taxonomy of the model-ladenness of data. *Philosophy of Science, 87*(5), 793-806. https://doi.org/10.1086/710516

Bokulich, A. & Parker, W. (2021). Data models, representation, and adequacy-for-purpose. *European Journal for Philosophy of Science, 11*(1), 31. https://doi.org/10.1007/s13194-020-00345-2

Castellani, E. & Schettino, G. (2023). Nested modalities in astrophysical modelling. *European Journal for Philosophy of Science, 13*, 11. https://doi.org/10.1007/s13194-023-00511-2

Collins, H. (1992). *Changing order: Replication and induction in scientific practice*. University of Chicago Press.

Earman, J. & Glymour, C. (1980). Relativity and eclipses: The British eclipse expeditions of 1919 and their predecessors. *Historical Studies in the Physical Sciences, 11*(1), 49-85. https://doi.org/10.2307/27757471

Glymour, B. (2000). Data and phenomena: A distinction reconsidered. *Midwest Studies in Philosophy, 31*(1), 184-201. https://doi.org/10.1023/A:1005499609332

Kennefick, D. (2009). Testing relativity from the 1919 eclipse —a question of bias. *Physics Today, 62*(3), 37-42. https://doi.org/10.1063/1.3099578

Kennefick, D. (2019). *No shadow of a doubt: The 1919 eclipse that confirmed Einstein's theory of relativity*. Princeton University Press. https://doi.org/10.1119/1.5123052

Laymon, R. (1982). Scientific realism and the hierarchical counterfactual path from data to theory. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1*, 107-121. https://doi.org/10.1086/psaprocbienmeetp.1982.1.192660

Leonelli, S. (2009). On the locality of data and claims about phenomena. *Philosophy of Science, 76*(5), 737-749. https://doi.org/10.1086/605804

Leonelli, S. (2015). What count as scientific data? A relational framework. *Philosophy of Science, 82*(5), 810-821. https://doi.org/10.1086/684083

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226416502.001.0001

Leonelli, S. (2019). What distinguishes data from models? *European Journal for Philosophy of Science, 9*, 22. https://doi.org/10.1007/s13194-018-0246-0

Massimi, M. (2007). Saving unobservable phenomena. *British Journal for the Philosophy of Science, 58*(2), 235-262. https://doi.org/10.1093/bjps/axm013

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

McAllister, J. W. (1997). Phenomena and patterns in data sets. *Erkenntnis, 47*(2), 217-228. https://doi.org/10.1023/A:1005387021520

McAllister, J. W. (2011). What do patterns in empirical data tell us about the structure of the world? *Synthese, 182*(1), 73-87. https://doi.org/10.1007/s11229-009-9613-x

Schindler, S. (2008). Model, theory, and evidence in the discovery of the DNA structure. *The British Journal for the Philosophy of Science, 59*(4), 619-658. https://doi.org/10.1093/bjps/axn030

Schindler, S. (2013). Theory-laden experimentation. *Studies in History and Philosophy of Science Part A, 44*(1), 89-101. https://doi.org/10.1016/j.shpsa.2012.07.010

Suárez, M. (2023). Stellar structure models revisited: Evidence and data in asteroseismology. In N. Boyd, S. de Baerdemaeker, K. Heng & V. Matarese (Eds.), *Philosophy of Astrophysics: Stars, Simulations and the Struggle to Determine What is Out There* (pp. 111-129). Springer. https://doi.org/10.1007/978-3-031-26618-8_7

Suárez, M. (2024). *Inference and representation: A study in modeling science*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226830032

Suárez, M. (2025). Daniel Kennefick's No Shadow of a Doubt, *BJPS Review of Books*, 2025. https://doi.org/10.59350/thebsps.13234

van Fraassen, B. C. (1980). *The scientific image*. Clarendon Press. https://doi.org/10.1093/0198244274.001.0001

van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford University Press. https://doi.org/10.1093/0198248601.001.0001

von Kluber, H. (1960). The determination of Einstein's light-deflection in the gravitational field of the sun. *Vistas in Astronomy, 3*, 47-77. https://doi.org/10.1016/0083-6656(60)90005-2

Woodward, J. (1989). Data and phenomena. *Synthese 79*(3), 393-472. https://doi.org/10.1007/BF00869282

Woodward, J. (2000). Data, phenomena, and reliability. *Philosophy of Science 67*, S163-S179. https://doi.org/10.1086/392817

Woodward, J. (2010). Data, phenomena, signal, and noise. *Philosophy of Science 77*(5), 792-803. https://doi.org/10.1086/656554

José V. Hernández-Conde is Associate Professor of Philosophy of Language at the University of Valladolid (Spain). In 2022, he was awarded a Leonardo Grant by the BBVA Foundation. His research focuses on the philosophy of language and the mind, as well as the philosophy of data science and artificial intelligence.

Address: Department of Philosophy, Faculty of Philosophy and Arts, University of Valladolid, Plaza del Campus s/n, 47011, Valladolid, Spain. E-mail: jhercon@uva.es – ORCID: 0000-0002-8502-6570

María Caamaño-Alegre is Associate Professor of Philosophy of Science at the University of Valladolid (Spain). She specializes in philosophy of science, with her research interests lying at the intersection of general philosophy of science, the methodology of science, the philosophy of language and epistemology.

Address: Department of Philosophy, Faculty of Philosophy and Arts, University of Valladolid, Plaza del Campus s/n, 47011, Valladolid, Spain. E-mail: mariaconcepcion.caamano@uva.es – ORCID: 0000-0002-7005-9257