# TRANSFORMING DATA INTO ADDED-VALUE INFORMATION: THE DESIGN OF SCIENTIFIC MEASUREMENT MODELS THROUGH THE LENS OF DESIGN THEORY

**Barbier, Raphaëlle;**
**Le Masson, Pascal;**
**Weil, Benoit**

MINES ParisTech

## ABSTRACT

Transforming data into added-value information is a recurrent issue in the context of "big data" phenomenon, as new sources of data become increasingly available. This paper proposes to offer a fresh look on how data and added-value information are linked through the design of specific models. This investigation is based on design theory, used as an analysis framework, and on a historical example in the Earth science field. It aims at unveiling the reasoning logic behind the design process of models combining data science and domain knowledge in specific ways, especially involving not only knowledge about the physical phenomena but also on the measuring instrument itself. More specifically, this paper shows how specific efforts on exploring the originality of the new instrument compared to existing ones can result in designing performant models to transform new sources of data into information. This also suggests several important competencies to be involved in the model-design process: (1) a detailed understanding of the limitations of existing models (2) the ability to explore both the originality of the new source of data compared to existing ones (3) the ability of leveraging independent data sources.

**Keywords**: Design theory, Big data, Design process, data science, information design

**Contact**:
Barbier, Raphaelle
MINES ParisTech
Centre for Management Sciences
France
raphaelle.barbier@mines-paristech.fr

**Cite this article:** Barbier, R., Le Masson, P., Weil, B. (2021) 'Transforming Data into Added-Value Information: The Design of Scientific Measurement Models through the Lens of Design Theory', in *Proceedings of the International Conference on Engineering Design (ICED21),* Gothenburg, Sweden, 16-20 August 2021. DOI:10.1017/pds.2021.585

# 1 INTRODUCTION

In the 1980s, within a European research project, three research teams were asked to participate to a kind of scientific competition, to make use of a new source of data to transform it into added-value information - an issue that seems very contemporary today with the rise of "big data" phenomenon and organisation of specific challenges to address it (Sitruk and Kazakçi, 2018). In the 80s project, the three teams had the same starting point and objective: *given a set of satellite data (new data source), propose the best possible model to predict solar radiation received on ground (already measured by in-situ sensors or other empirical methods), the performance criteria being clearly defined as minimizing the prediction error*. The difference between the teams thus lied in the way they designed their respective model. A synthetic document was published to shed light on each of the three models and compare them. It is particularly interesting as it opens the "black box" of the modelling activity: not only does it illustrate the classic opposition between so-called "data-driven" and "physics-driven" approaches, but it also illustrates how significantly better a model can be designed when building on the two previous approaches and going beyond each of them. Moreover, details of the approaches give hints on how to design a powerful model linking physical phenomena and data - and how to get a strong performance by creatively make use of all available knowledge, in particular knowledge of the physics of the instrument (the satellite) and not only knowledge of the physics of radiation throughout the Earth atmosphere. This example is fascinating insofar as (1) it addresses in the 1980s, almost 40 years ago, a really contemporary issue: making sense of data to get precise, unquestionably added-value information; (2) it addresses also a problem that is as old as science: the design of scientific measures (information) based on a new instrument (new source of data).

This paper proposes to investigate this historical example to give insights on a surprising blind spot in the successful development of data-based services, that is describing the design efforts that remain even when the type of information to be derived from data and its related value are already identified. Indeed, even when data are available (open data for instance) and when their added-value for the service is unquestionable (taking the form of valuable information), there can still be an issue in transforming available data into added-value information. This apparently small step in the chain linking data to value formally consists in using or building a model that relates data to information in a reliable way. As for model building in data science, two recent trends have emerged: on the one hand, it can be considered as a "technical" statistical black box that can be addressed by relying on the most advanced data-science algorithms (GAN, CNN..); on the other hand, more recently, scholars remind that the wealth of scientific knowledge should be leveraged in data science models (Karpatne et al., 2017; Reichstein et al., 2019). In these two trends, it is clear that models are designed, but the design process and its underlying reasoning logic that intertwines data science and domain knowledge remain allusive. Moreover, the specific role of domain knowledge related to the instrument itself is not explicitly described. Our paper thus aims at investigating the specific process of designing models to transform a new source of data into information, more specifically addressing the following question: *in which possible ways can new sources of data be leveraged in the process of designing models to transform data into added-value information?*

To investigate this question, in a first part we show how literature proposed many models relating data to added-value information but leaves a blind spot on the question of the reasoning logic behind the design process of these models, and its specific link with the new instrument. In a second part we propose to build a theoretical framework derived from design theory to analyse the design process associated to model design in the historical case mentioned above, that is then elucidated in a third part. A fourth part highlights contributions and limitations of this paper.

# 2 LITERATURE REVIEW AND RESEARCH QUESTIONS

## 2.1 Data-driven design: new opportunities stemming from the use of data

In recent years, the development of internet, new sensors, and computational means has dramatically increased the flow of data in almost every business, industry and research area. This phenomenon, commonly referred as "big data", has largely been discussed in the literature, shedding light on its definition, opportunities and challenges, especially the issue of how value can be created out of this new flow of data (Gandomi and Haider, 2015; Günther et al., 2017). Literature in design has also largely described the new opportunities arising from the use of data. (Parraguez and Maier, 2017)



https://doi.org/10.1017/pds.2021.585 Published online by Cambridge University Press

highlight the potential benefits of using open-data from various sources (e.g. patents, publications, business registries, company websites, social networks) for the engineering design research. The variety of usages that could be made from data is often emphasized, for example through the 20 contributions of the special issue of the Journal of Mechanical Design (Kim et al., 2017) covering topics as various as discovering future design and technological opportunities thanks to patent mining techniques, modelling complex parts of the body in new manners, giving new insights on the critical functions to be included in the design of new products and services. Literature in design also more specifically reports on the beneficial use of data in the concept development phase of the design process (Escandón-Quintanilla et al., 2018; Bertoni, 2020).

The transformation of data into usages is often described through the "Data Information Knowledge Wisdom" hierarchy (Rowley, 2007) or more recently the "data-information-knowledge" chain (Abbasi et al., 2016). Despite different definitions of these terms (Zins, 2007), they describe in a similar way the different types of design efforts to be made in order to effectively turn data into usages: (1) transforming data (that is contextualized, related to the measuring settings) into information (that should be more generally understandable, and meaningful in the context of reuse) ; (2) transformation of information into knowledge, generally referring to information used in a certain context. Thus knowledge refers to a certain "usage" or "value". These two terms will be preferred in the present paper, to avoid confusions with the term "knowledge" used in C-K design theory.

In design literature, an important stream of works has focused on the way information could be transformed into usages, through the design of appropriate content management or visualisation tools (Huron et al., 2014; Dammak and Gardoni, 2018). Regarding the transformation of data into information, scholars report on several issues. (Bertoni, 2020) notices the tendency to rather resort to relatively easy-to-use data (such as text mining of social networks) rather than building new data generation methods that would bring more valuable information (such as resorting to the use of sensors giving information on the product in use) because of the higher complexity of such approaches. The design effort to be made for extracting new types of data is also reported as an issue by (Montecchi and Becattini, 2020) in the context of using data to encourage sustainable behaviours. These issues are often related to the ability of implementing complex algorithms and specific data science techniques, that are said to be sometimes poorly understood (Parraguez and Maier, 2017). So this stream of works *suggests well the significant design effort to be made to transform data into information, however it does not fully describe the design process underlying this transformation.*

## 2.2 Designing models to transform data into information

Other branches of literature offer insights on how information is designed, highlighting the importance of designing appropriate models. First, literature about instrumentation design and metrology reminds us that information coming from a measuring system is always designed (even for basic direct-reading instruments). Indeed, designing specific models of the measurement process is required to adequately relate the "indication" given by the instrument and the "measurement outcome", that is information that can be attributable to the object under consideration and not to other factors related to the instrument or the environment (Mari et al., 2012; Giordani and Mari, 2012; Tal, 2017). This literature also highlights that several model-design approaches might exist, emphasizing two extreme archetypal cases: considering the system as a "black-box" where the model is derived from measures done for a number of known standard states, or considering the system as a "white box" where the model is determined by representing the physical process in details (Tal, 2017). This literature distinguishes two types of models and describes how to parameterize a (given) model in certain situations - still the question of the design of the base model remains unanswered.

Second, these considerations on how models are designed are also discussed in the literature related to the use of data for scientific activities. The question on how to build good models is common in science. To give a few examples, already in the nineties, reflexive works of several disciplines around Earth science on their modelling practices were carried out, e.g. in hydrology (Beven, 1989; Barnes, 1995), or for solar irradiance estimation for which different approaches - classified as either physical or statistical - were listed and compared (Noia et al., 1993a, 1993b). More recently, (Karpatne et al., 2017) made a similar distinction between "physics-only" (or "theory-only") models, that are built by modelling the different underlying physical processes, and "data-only" models, that are built without using scientific theories by leveraging the large amount of available data through various data science techniques. The same authors emphasize the limitations of both approaches, calling for a new "theory-



https://doi.org/10.1017/pds.2021.585 Published online by Cambridge University Press

guided data science" paradigm, that would consist in combining scientific knowledge and data science. This article also gives a broad overview of the different combination possibilities by categorizing them in five main types: (1) theory-guided design of data science model families, (2) theory-guided learning of data science models given a model family, (3) theory-guided refinement of data science models outputs, (4) constructing hybrid models, (5) improving theory-based models according to observational data. Hence the authors propose two main types of models and explain how to combine them. But it remains unclear, what is the design logics in "data-only" and "physics-only" and, more importantly, whether there might another type of model-design, different from the two previous ones and their combinations. This calls for more explicitly showing the underlying reasoning logic of the design process, especially the specific manner domain and data science knowledge bases are leveraged and might interact and evolve during the process. In this perspective, a quick analysis of "physics-only" and "data-only" models reveals for instance that the role played by the domain knowledge related to the new instrument (and not only on the physical phenomena) remains unclear. The present paper will therefore address the following research question: *how are new sources of data leveraged to design an appropriate model transforming these data into added-value information?* In particular, we will wonder what is the specific role played by the domain knowledge related to the instrument and how this instrument domain knowledge could help design specific model(s) transforming instrument data into added-value information.

## 3 METHOD

As underlined in the literature presented above, transforming data into information involves designing a specific model, leveraging both domain and data science knowledge. Our methodology is twofold: first, C-K design theory is used as a theoretical framework to represent this model-design process, second this framework is applied on a specific historical case study to further unveil interesting features of what makes a relevant model-design logic.

### 3.1 C-K theory to represent the reasoning logic of designing models

Our investigation relies on C-K theory as it sheds light on the reasoning logic underlying a design process (Hatchuel and Weil, 2003, 2009). Such a process is indeed described as the interaction and the expansion of two spaces: a space K of knowledge and a space C of concepts. The K-space gathers all the knowledge the designers activate and progressively acquire during the design process (technical knowledge, user preferences, standards and regulations, etc). The C-space is the space where new ideas, concepts are explored. The interactions between the two spaces are represented through four different operators: K→C ("disjunction" where C-space is expanded thanks to available knowledge in K-space), C→K ("conjunction" where available knowledge in K is expanded and triggered by the concept expansion in C), K→K (self-expansion of knowledge based on logic rules, e.g. proving new theorems), C→C (expansion of concepts through partitioning of concepts).

Thanks to this framework, we can represent the problem of designing a model (M) to better estimate a certain information (Y) based on new sources of data (X) as follows (see Figure 1):

**(a)** The starting point is the initial concept C0 "*designing a model M for a better estimation of Y through the use of X"*

**(b)** This calls for investigating in K-space (C→K) what are the available models to estimate Y (knowledge base on models) and how to use X in those models (knowledge base on the instrument, i.e. existing and new sources of data). Two main types of models can be considered:

- The *existing physics-driven models M_physics* based on parameters describing the atmosphere (cloud properties, aerosols, etc.), where previously existing sources of data are used to estimate the parameters of the model.
- The *existing data-driven models M_data* (e.g. multiple linear regression) whose parameters are statistically estimated based on known pairs of (X,Y).

**(c)** Based on these knowledge base on models, a subsequent operation K→C makes appear the two archetypal approaches called "physics-driven" or "data-driven", corresponding respectively to the concepts "*M built on M_physics using X to better estimate physical parameters of M_physics"* and "*M built on M_data using X to estimate the parameters of the statistical model M_data"*. Regarding the domain knowledge on the instrument, these approaches only rely on the capacity of the instrument to



https://doi.org/10.1017/pds.2021.585 Published online by Cambridge University Press

be integrated in existing models, i.e. the dimension of the instrument that can be expressed relatedly to the existing sources of data (referred as the knowledge base "correlation" in Figure 1).

**(d)** A third approach, coined "hybrid", can be generated by leveraging both knowledge of M_physics and M_data and their respective limitations, as mentioned in literature with the approach of "theory-guided data science". This concept could be formulated as "*M built by using X to overcome limitations of existing models"*. In this third approach, the role of the domain knowledge on the instrument remains unclearly described in literature. C-K design theory helps us to formulate some first insights. Indeed, it predicts that a good design process relies on the use of independent knowledge bases: thus the best approach should investigate to what extent the new source of data is "orthogonal" to the existing ones, i.e. what are the additional independent knowledge it could bring compared to the previously used data sources. The analysis of the historical case study aims at further elucidating these elements.

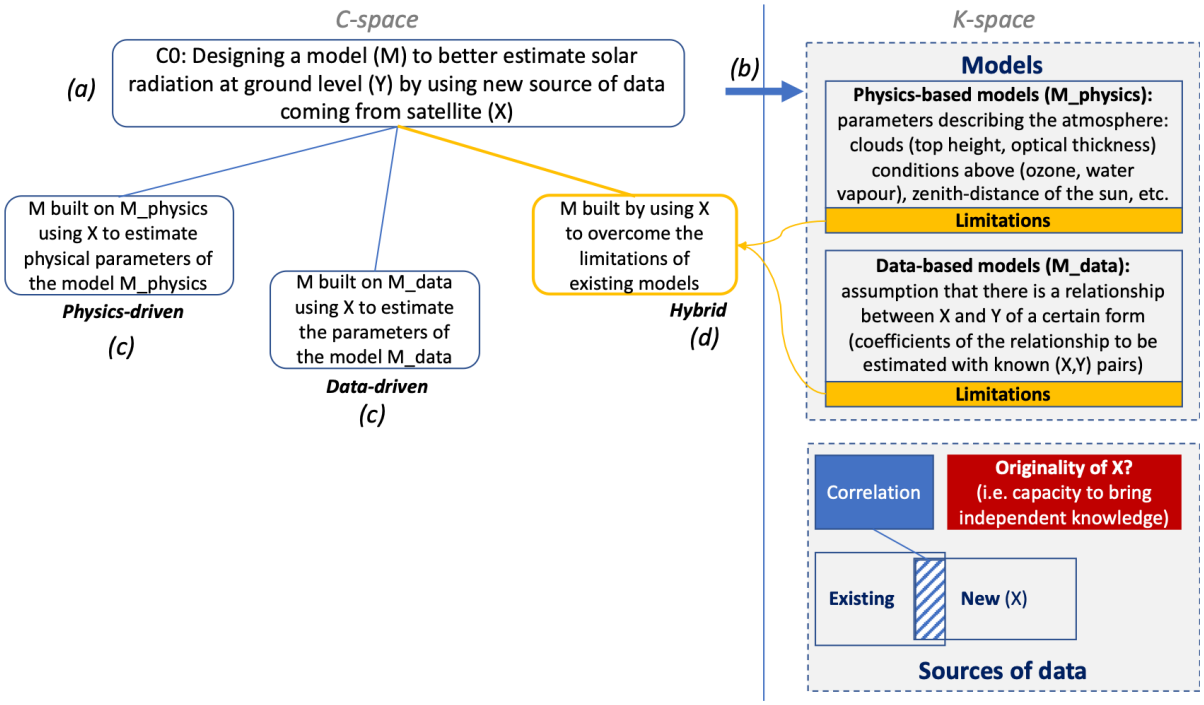


*Figure 1. C-K theory used to represent the model-design reasoning logic*

## 3.2 Case study: comparison of three model-design approaches

The case study used in this paper is particularly adapted to investigate the research question formulated above. Indeed, it corresponds to a situation where there is a new source of data with an already identified usage: it thus allows us to avoid other debates in literature related to the identification of usages for new data and to rather focus on the design efforts that remain to be made to transform data into information, already identified as useful. Our case study is based on the research work of a research organisation based in Sophia-Antipolis (France) on solar radiation estimation from satellite data, carried out in the 1980s. As mentioned in the introduction, this organisation was involved in a project supported by the European Commission's Solar Energy R&D Programme. The project aimed at assessing solar radiation more precisely and reliably, especially by integrating new data coming from satellites (whereas at the time solar radiation was mainly derived from networks of "in-situ" solar instruments, that were installed in a limited number of locations). Within this project, Sophia-Antipolis research institute, along with two other research teams, were in charge of developing a model to link solar radiation estimates and Earth observation data including new satellite data. Each research team developed a different model, based on its respective expertise. Their different model-design approaches were compared in the final report of the project for the European Commission (Grüter et al., 1986). We also had access to the PhD thesis detailing the specific modelling approach developed by the Sophia-Antipolis team, and conducted semi-structured interviews (6 hours in total) with the researcher of this team who had been working on the development of the solar radiation methods from this European project in the 80s up to 2018.



https://doi.org/10.1017/pds.2021.585 Published online by Cambridge University Press

# 4 ANALYSIS OF THE CASE STUDY: SPECIFIC ROLE OF THE DOMAIN KNOWLEDGE RELATED TO THE INSTRUMENT

## 4.1 Three teams corresponding to the three archetypal approaches found in literature

The three approaches developed by the different teams correspond well to the different archetypal approaches found in literature: a physics-driven approach (Cologne team), a data-driven approach (Stuttgart team) and a hybrid approach (Sophia-Antipolis team) - see also Figure 2:

- *Physics-driven approach (Cologne team):* this team designed a model based on a "radiative transfer model" that explicitly describes the physical processes (e.g. absorption, scattering) occurring in the atmosphere. In this approach, satellite data are used to estimate existing parameters of the physical model. The limitations of this approach lie in the need of additional sources of data coming from other sources to determine some of the parameters of the model, that usually involves averaging the results over larger areas thus degrading the resolution of the final product.
- *Data-driven approach (Stuttgart team):* this team resorted to a statistical approach: the model is based on statistical regressions between satellite data and solar radiation measurements at the Earth's surface, measured by "in-situ" stations within the considered area. Satellite data are used to estimate the coefficients of the statistical law (starting with 360 features describing each satellite image, 25 parameters were kept, being most correlated to the ground solar radiation). The limitations of the approach result from the large number of regression parameters to be estimated without considering much physical-consistency, as the parameters are mainly deduced from the texture analysis of the satellite images.
- *Hybrid approach (Sophia-Antipolis team):* this team resorted to an approach aiming at overcoming the limitations of the two previous approaches. This approach more specifically relied on the introduction of an intermediary variable coined "cloud index", describing the level of cloudiness. This hybrid approach had proved to be the most efficient one, in terms of quality of the estimation (see error histograms on Figure 1) but also easiness of processing (almost ten times quicker than the physical one). Our empirical materials help us to further elucidate the role played by the domain knowledge on the instrument in this approach.

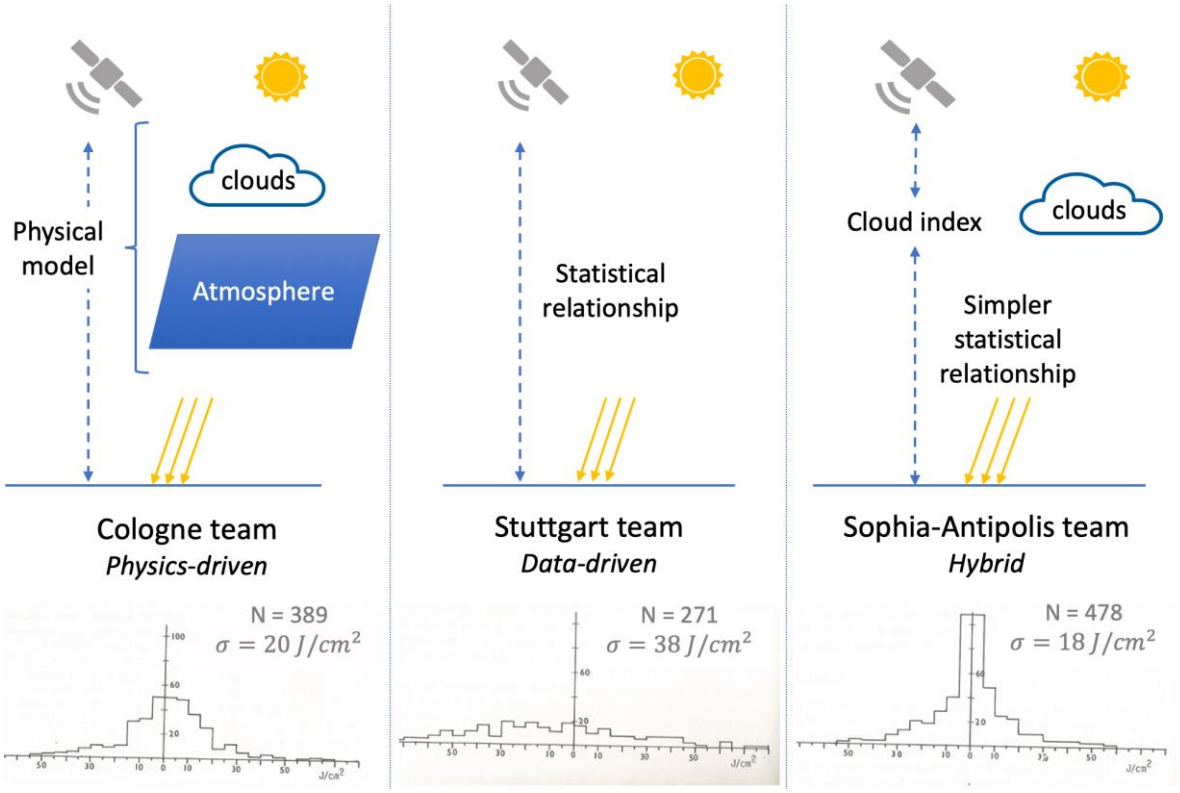


*Figure 2. The three competing approaches and their respective results (histograms of errors - difference between predicted estimation and measurement at ground level)*



https://doi.org/10.1017/pds.2021.585 Published online by Cambridge University Press

### 4.2 Closer look at the domain knowledge related to the instrument: intertwined exploration of the originality of the new instrument and the way of taking advantage of it to overcome identified limitations of existing models

Based on what predicts C-K theory, the performance of such an approach could be explained by the way domain knowledge on the instrument is leveraged, especially investigating the "orthogonal" dimension of the new instrument compared to existing sources of data. The introduction of the "cloud index" by Sophia-Antipolis team corresponds to such an approach of investigating the originality of satellite data compared to existing sources of data. Indeed, (Cano, 1982) explicitly mentions that this cloud index is specifically built in order to be only determined by satellite data, without resorting to other parameters that would require to be assessed through other data sources. To do so, the estimation of this variable takes advantage of a specific property of the satellite data, i.e. the provision of time series (images of the same location at different moments in time). It is thus clear that Sophia-Antipolis team resorts to domain knowledge on the new instrument in a very specific way: rather than relying on the dimension of new sources of data correlated to existing ones, the researchers e*xplore the originality of the instrument, here exploiting a specific property of satellite data.*
A second interesting element to be noted is that *this knowledge expansion about the possibilities of the instrument is made in close interaction with the exploration of how existing models' limitations can be addressed*. Indeed, Sophia-Antipolis researchers highlight in (Grüther et al., 1986) that the cloud index can then be statistically linked to solar radiation with a simple linear relationship, thus reducing the number of regression parameters compared to fully statistical approaches, and avoiding rough estimation of some parameters of the physical approach that could not be directly estimated by satellite data. These conclusions lead us to refine the "hybrid" approach by distinguishing between two forms of "hybrid" models (see Figure 3):

- "Combinatory" hybrid models that would combine parts of physics-driven and parts of data-driven models relying on a partial domain knowledge of the instrument, related to its dimensions that can be correlated to existing sources of data.
- "Expansive" hybrid models that would leverage the originality of the new instrument compared to existing ones to generate expansion on how the model is designed, as highlighted in this Sophia-Antipolis case.

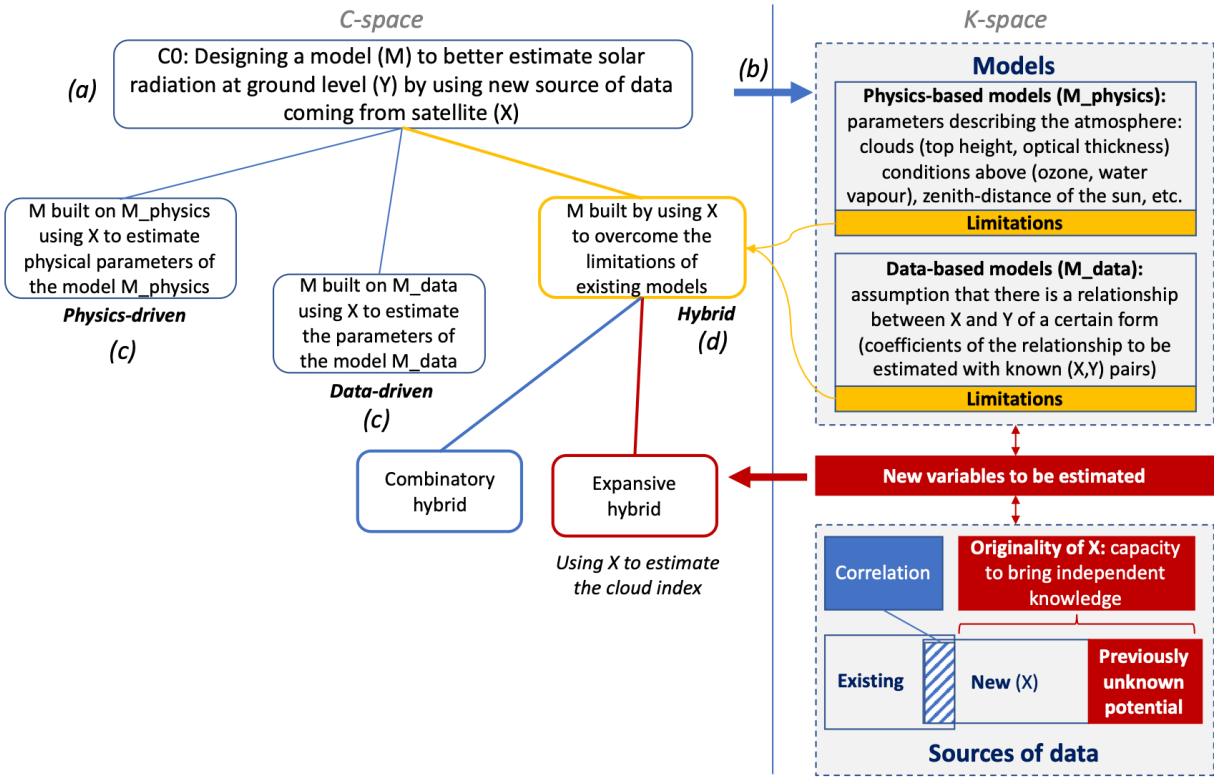


*Figure 3. Representation of the model-design reasoning logic using C-K theory, completed with the case study analysis*



https://doi.org/10.1017/pds.2021.585 Published online by Cambridge University Press

## 5 DISCUSSION AND CONCLUSION

This paper proposes several contributions. First, it gives particular insights on the current academic discussions related to data-driven economy, where a lot of attention is either directed to the potential usages that new flows of data could create, or the new types of data to be exploited or created in order to address defined use cases. The case of solar radiation estimation offers an interesting configuration where these two considerations are already addressed: indeed, there is already a clearly identified type of information (solar radiation) to be derived from new data sources (satellite data), for a known usage (solar radiation information being of direct interest for the European Commission building at that time a European Solar Radiation Atlas). This example allows us to concentrate on the design efforts to be made to transform data into information and to show interesting features of the reasoning logic behind the design of a "hybrid" model combining domain and data science knowledge. In particular, this paper highlights the specific role of the domain knowledge related to the instrument in designing models to transform new sources of data into information. More specifically, it shows that a performant approach can result from *making specific efforts on exploring and leveraging the originality of the new instrument compared to existing ones, suggesting a specific way of building hybrid models that goes beyond a simple combinatory logic*.

These considerations can be helpful to better understand some elements found in literature. First, the difficulties of resorting to new types of data mentioned by (Bertoni, 2020) and (Montecchi and Becattini, 2020) can be better understood: our results indeed suggest that the tendency to avoid resorting to new data might result from the intricate design effort to be made with specific competencies in order to transform new data into added-value information. The importance of *orienting the design process in a way that takes into account specific limitations of the usual models* can also be found in other contexts, e.g. in (Kazakçı, 2015) highlighting that in the context of "HiggsML challenge organised to gain insights into the study of Higgs boson in particle physics by means of machine learning algorithms", the participants who succeeded were the ones that did not simply apply the usual workflow of machine learning techniques but were able to take into account the specificity of the challenge involving an usual objective function. More interestingly, some examples of a "theory-guided data science" approach given in (Karpatne et al., 2017) can be better understood as either "combinatory" or "expansive" hybrid approaches, by considering the way domain knowledge related to the instrument is leveraged. For example, the problem of mapping surface water dynamics with satellite data starts with the *analysis of the limitations of "data-only" models*: "Remote sensing data from Earth observing satellites presents a promising opportunity for monitoring the dynamics of surface water body extent at regular intervals of time. It is possible to build predictive models that use multi-spectral data from satellite images as input features to classify pixels of the image as water or land. However, these models are challenged by the poor quality of labeled data, noise and missing values in remote sensing signals, and the inherent variability of water and land classes over space and time." From this analysis, *a way of addressing these challenges is investigated* thanks to additional domain knowledge, noticing that "locations at a lower elevation are filled up first before the water level reaches locations at higher elevations". Thus, to improve the model, *information on the elevation is identified as a new variable to be estimated* to assist classification models (it would be used as a constraint of the classifier minimizing training errors). However, such information obtained from other instruments (sonar instruments) is not available at the required granularity. Thus, *a new way of using satellite data is imagined to derive information on the elevation*, by "using the history of imperfect water/land labels produced by a data science model at every location over a long period of time", suggesting here an "expansive" hybrid approach.

This paper also contributes to practice as these results shed light on *several interesting competencies that a model designer should have to successfully develop a model combining domain and data science knowledge.* It is first highlighted that domain knowledge does not only involve understanding the physical processes but also understanding and exploring the potential of the instrument providing new sources of data. Second, the competencies of model designers should not be described as only picking in a model manual (either physics-based or data-based) given a certain situation, but should rather involve *(1) a detailed understanding of the limitations of existing models - either "physics-driven" or "data-driven" ; (2) the ability to explore both the originality of the new source of data compared to existing ones and on how it could help overcome the limitations of existing models*. This might lead to introduce new variables to be estimated from data, which might also consequently



https://doi.org/10.1017/pds.2021.585 Published online by Cambridge University Press

involve building new ways of considering the output data of the new instrument. Finally, these elements also suggest a third competency: *(3) the ability of leveraging together independent, potentially heterogeneous, data sources*. Indeed, the analysis of the model limitations and originality of the instrument can result in building new variables to be estimated from data, that would require looking for several new independent sources of data. This specific competency could be related to "data fusion", that had been interestingly investigated by the same Sophia-Antipolis team (Wald, 1998), as "a formal framework in which are expressed means and tools for the alliance of data originating from different sources, in order to obtain information of greater quality".

Several limitations of our research can be identified. First, the paper relies on a specific case study, highlighting the relevance of an expansive hybrid approach that leverages domain knowledge related to the new instrument by exploring its originality compared to existing sources of information. However, other interesting approaches could also result from other types of expansion that are not only related the instrument domain knowledge expansion, but that could come from the enrichment of knowledge bases related to the physics-driven or data-driven models (especially through new machine learning techniques that are currently developed). These types of expansions could be further described through other case studies corresponding to such contexts. Moreover, as highlighted in this paper, our case study corresponds to a situation where the new source of data and its usage (transformation into a certain type of added-value information) are already given. It would be interesting to investigate how the design process described in this paper could be leveraged in cases where the usefulness of information or the type of instrument to be used are still be to be explored. Finally, it is also worth noticing that our case study relies on a specific type of data (i.e. scientific or instrumental type). The relevance of our results for other types of data (such as data bases of patents or data on consumers' preferences) could be discussed. At first sight, we could consider our case study as an extreme case that helps rediscuss basic notions. In this perspective, even with non-scientific types of data, we could assume that information is also derived from data through the use of specific models (although maybe not as complex as for Earth science). In some contexts, these models might be implicitly used and little designed, and the specific model-design competencies highlighted in this paper could open up new possibilities, by designing models that makes most use of knowledge about limitations of data-science techniques and description of the considered phenomenon (not necessarily physical processes, but for example modelling of customer behaviours), and exploration of the specificities of the data collection process (that might be different from scientific instruments). Further investigations would be interesting to further examine this question.

## ACKNOWLEDGMENTS

This research work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 820852.

## REFERENCES

Abbasi, A., Sarker, S., Chiang, R.H.L., 2016. Big Data Research in Information Systems: Toward an Inclusive Research Agenda. J. AIS 17, 3. https://doi.org/10.17705/1jais.00423

Barnes, C.J., 1995. The art of catchment modeling: What is a good model? Environment International, Water Modelling 21, 747–751. https://doi.org/10.1016/0160-4120(95)00082-V

Bertoni, A., 2020. Data-Driven Design in Concept Development: Systematic Review and Missed Opportunities. Proceedings of the Design Society: DESIGN Conference 1, 101–110. https://doi.org/10.1017/dsd.2020.4

Beven, K., 1989. Changing ideas in hydrology — The case of physically-based models. Journal of Hydrology 105, 157–172. https://doi.org/10.1016/0022-1694(89)90101-7

Dammak, H., Gardoni, M., 2018. Improving the Innovation Process by Harnessing the Usage of Content Management Tools Coupled with Visualization Tools, in: Chiabert, P., Bouras, A., Noël, F., Ríos, J. (Eds.), Product Lifecycle Management to Support Industry 4.0, IFIP Advances in Information and Communication Technology.pp. 642–655. https://doi.org/10.1007/978-3-030-01614-2_59

Escandón-Quintanilla, M.-L., Gardoni, M., Cohendet, P., 2018. Improving concept development with data exploration in the context of an innovation and technological design course. International Journal on Interactive Design and Manufacturing (IJIDeM) 12, 161–172. https://doi.org/10.1007/s12008-017-0380-5

Gandomi, A., Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35, 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Giordani, A., Mari, L., 2012. Measurement, Models, and Uncertainty. IEEE Transactions on Instrumentation and Measurement 61, 2144–2152. https://doi.org/10.1109/TIM.2012.2193695

Grüter, W., Guillard, H., Möser, W., Monget, J.M., Palz, W., Raschke, E., Reinhardt, R.E., Schwarzmann, P., Wald, L., 1986. Solar Radiation Data from Satellite Images: Determination of Solar Radiation at Ground Level from Images of the Earth Transmitted by Meteorological Satellites - An Assessment Study, Solar Energy R&D in the Ec Series F: Springer Netherlands.
Günther, W.A., Rezazade Mehrizi, M.H., Huysman, M., Feldberg, F., 2017. Debating big data: A literature review on realizing value from big data. The Journal of Strategic Information Systems 26, 191–209. https://doi.org/10.1016/j.jsis.2017.07.003
Hatchuel, A., Weil, B., 2009. C-K design theory: an advanced formulation. Research in Engineering Design 19, 181–192. https://doi.org/10.1007/s00163-008-0043-4
Hatchuel, A., Weil, B., 2003. A new approach of innovative design: an introduction to C-K theory. Presented at the International Conference on Engineering Design, International Conference on Engineering Design, Stockholm.
Huron, S., Carpendale, S., Thudt, A., Tang, A., Mauerer, M., 2014. Constructive visualization, in: Proceedings of the 2014 Conference on Designing Interactive Systems, DIS '14. Association for Computing Machinery, New York, NY, USA, pp. 433–442. https://doi.org/10.1145/2598510.2598566
Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. IEEE Transactions on Knowledge and Data Engineering 29, 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168
Kazakçı, A.O., 2015. Data Science as a New Frontier for Design. Presented at the Proceedings of the 20th International Conference on Engineering Design (ICED15), p. 10.
Kim, H.H.M., Liu, Y., Wang, C.C.L., Wang, Y., 2017. Special Issue: Data-Driven Design (D3). J. Mech. Des 139. https://doi.org/10.1115/1.4037943
Mari, L., Carbone, P., Petri, D., 2012. Measurement Fundamentals: A Pragmatic View. IEEE Transactions on Instrumentation and Measurement 61, 2107–2115. https://doi.org/10.1109/TIM.2012.2193693
Montecchi, T., Becattini, N., 2020. Design for sustainable behavior: opportunities and challenges of a data-driven approach.Proceedings of the Design Society: DESIGN Conference 1, 2089–2098. https://doi.org/10.1017/dsd.2020.147
Noia, M., Ratto, C.F., Festa, R., 1993a. Solar irradiance estimation from geostationary satellite data: I. Statistical models. Solar Energy 51, 449–456. https://doi.org/10.1016/0038-092X(93)90130-G
Noia, M., Ratto, C.F., Festa, R., 1993b. Solar irradiance estimation from geostationary satellite data: II. Physical models. Solar Energy 51, 457–465. https://doi.org/10.1016/0038-092X(93)90131-7
Parraguez, P., Maier, A., 2017. Data-driven engineering design research: Opportunities using open data Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 7: Design Theory and Research Methodology, Vancouver, Canada, 21-25.08.2017.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1
Rowley, J., 2007. The wisdom hierarchy: representations of the DIKW hierarchy , The wisdom hierarchy: representations of the DIKW hierarchy. Journal of Information Science 33, 163–180. https://doi.org/10.1177/0165551506070706
Sitruk, Y., Kazakçi, A., 2018. Crowd-Based Data-Driven Hypothesis Generation from Data and the Organisation of Participative Scientific Process. Presented at the 15th International Design Conference, pp. 1673–1684. https://doi.org/10.21278/idc.2018.0510
Tal, E., 2017. Calibration: Modelling the measurement process. Studies in History and Philosophy of Science Part A, The Making of Measurement 65–66, 33–45. https://doi.org/10.1016/j.shpsa.2017.09.001
Wald, L., 1998. Data fusion: a conceptual approach for an efficient exploitation of remote sensing images. Presented at the 2nd International Conference "Fusion of Earth Data: merging point measurements, raster maps and remotely sensed images", Sophia-Antipolis (France), p. 8.
Zins, C., 2007. 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



https://doi.org/10.1017/pds.2021.585 Published online by Cambridge University Press