DATA-PUSHED PROJECTS: THE ROLE OF ANOMALIES TO BUILD DESIGN PROCESSES FOR SUBSEQUENT EXPLORATION

Bordas, Antoine;
Le Masson, Pascal;
Weil, Benoit

Mines Paris, PSL University, Centre for management science (CGS), i3 UMR CNRS, 75006 Paris, France

ABSTRACT

Data-pushed projects are common in companies and consist in the design of a model in order to deliver a desirable output. The design of data science models appears at the intersection of optimisation and creativity logic, with in both cases the presence of anomalies to a various extent but no clear design process.

This paper therefore proposes to study the possible design processes in data-pushed projects, highlighting distinct knowledge exploration logics and the role of anomalies in each. This research introduces a theoretical framework to study data-pushed projects and is based on design theory. Three case studies complete this theoretical work to examine each of the processes and test our hypothesis.

As a result, this paper derives three design processes adapted to data-pushed projects and put forward for each of them: 1) the various knowledge leveraged and generated and 2) the specific role of anomalies.

Keywords: Design theory, Design process, Big data, Data-pushed projects, Anomaly management

Contact:
Bordas, Antoine
Mines Paris
France
antoine.bordas@minesparis.psl.eu

Cite this article: Bordas, A., Le Masson, P., Weil, B. (2023) 'Data-Pushed Projects: The Role of Anomalies to Build Design Processes for Subsequent Exploration', in *Proceedings of the International Conference on Engineering Design (ICED23),* Bordeaux, France, 24-28 July 2023. DOI:10.1017/pds.2023.114

https://doi.org/10.1017/pds.2023.114 Published online by Cambridge University Press

1 INTRODUCTION

Many surveys acknowledge the massive resort to data science and its continuous development in companies, encouraging ever more companies to launch their data transformation. The literature in data science is therefore very fertile, especially when it comes to designing models, whether explicative or predictive. This is one direction of data science, that we call data-pushed projects in the following, where data is the primary component of these projects, and the aim is to design a model in order to deliver a "marketable" output. This goes backward from data-pulled projects where the data is what we want to design and obtain at the end, as described by (Trabucchi and Buganza, 2018).

Such data-pushed projects require transforming a huge and available database into something valuable through designing a mathematical model. The literature in data science and mathematics is very prolific when it comes to designing models, that often are designed thanks to the combination of various categories of statistical models, even though sometimes incorporating non data related sources of knowledge is considered. However, the underlying design process remains unclear and the role played by a recurring parameter, anomalies, raises questions. Hence, we wish to elucidate the various strategies data scientists have to design a model and the role of anomalies in each.

Consequently, we wonder *what are the various roles of anomalies in each of the design processes for designing models in data-pushed projects?* As such, this paper is an attempt to make engineering design and data science, through data-pushed projects, discuss with the opportunity to enrich one another.

To investigate this question, we will see that the literature positions data science at the intersection of optimization and creativity, two domains that can be coherently joined with engineering design. The second part will build a theoretical model, based on C-K theory, to derive the various design processes in data science, that we will afterwards illustrate on different case study. The deeper analysis of these case studies will reveal the role of anomalies in the subsequent exploration process, as described in the fifth section, before concluding the paper.

2 LITERATURE REVIEW AND RESEARCH QUESTION: MODEL DESIGN AS A CREATIVE TASK THAT CAN LEVERAGE ENGINEERING DESIGN

2.1 Data science models design comes under optimization strategies

Data pushed projects went along with the development of a recent discipline, data science which objective is "to design automated methods to analyse massive and complex data" (Kazakçı, 2015), what is done following an established workflow according to the same author. This workflow, very similar to a dominant design, starts with the selection of a category of model, supposedly the speciality of the data scientist who "knows what method to use when" and can in fine be read as an optimization problem. This optimization logic can be well illustrated by the gradient descent algorithm, widely used in data science for minimization of a loss function, hence determination of the coefficients of a models, as done for neural networks (Ruder, 2017). We can also find this strategy in (Corral et al., 2015) when the authors explain the modelling stage as following six steps from problem definition to evaluation and refinements of the candidate models, by way of selecting the right category of models. It is important to note that this methodology is largely based on a problem-solving approach, as the authors acknowledged it, and that it implicitly assumes we have a model provider that can feed the designer, who will then choose, in other term optimize. This dominant design in data science can be found and is conveyed by many data science books (Friedman et al., 2001; James et al., 2013).

In parallel to this historical trend, scholars developed new paradigms, such as theory-guided data science (Karpatne et al., 2017) or informed-machine learning (von Rueden et al., 2021), where the idea is to introduce non-data knowledge in the design of models. This was done mainly to answer the lack of efficiency of data-only-models and many calls for explainability, stemming from the legislator (European Parliament, 2016) or scholars (Adadi and Berrada, 2018). Yet, both these strategies, theory-guided data science and informed machine learning, fall under the same optimization logic that tend to incorporate additional knowledge in order to get rid of inconsistencies with respect to known physical principles. That can be illustrated by (Daw et al., 2021) where the principles of theory-guided data science are applied and where the authors enter an optimization logic once the architecture of their model defined and fixed. Yet, the inconsistencies mentioned are what scientists call anomalies (Barnett and

https://doi.org/10.1017/pds.2023.114 Published online by Cambridge University Press

Lewis, 1984; Hodge and Austin, 2004), and whatever might be the category of model, it will have a limited domain of validity, that can be extended thanks to improvement, calling sometimes on creativity.

2.2 What scientific modelling, hence data science, have to do with creativity

The role of creativity for the design of models in data-pushed projects could be seen as integrative, by designing new models in which anomalies would be integrated and understood. This brings us very close to the question of designing and constructing science and models, taking us to Einstein's famous letters to his friend Maurice Solovine, where he explains his way of constructing models. To do so, the physicist uses an illustrative cyclical diagram linking experiences, what is given to us, and axioms, from which we draw conclusions. The path between these two parameters is represented by a simple arrow, although, according to Einstein there is "no logical path from the E [experiences] to the A [axioms], but only an intuitive (psychological) connection, which is always subject to revocation" (Einstein, 1987). This reveals the creativity work in the job of a scientist and is a plea for creativity in science. (Runco, 1994) agrees in a chapter where it is explained that the activity of problem-solving (as well as problem-formulation) are creative ones. In the more specific context of data-pushed projects, scholars studied the relation between models (hence algorithms) and creativity (Cascini et al., 2022; Davis, 2013), what remains an active field of research.

Yet, the precise role of creativity in the design process is unclear, Einstein in his letters (Einstein, 1987) does not characterize more the process leading to axioms from experiences more than with an "arrow-tipped arch like a pulse of light" (Holton, 1981). Hints to answer this have been given by philosophers of science, among which (Kuhn, 2021) when he explains that the construction of a model comes from the response we give to anomalies, restraining the relevance of Einstein's set of experiences E. (Bloor, 1978), interpreting Lakatos's work, explains the different way to respond to anomalies and as such starts to explain the reasoning logic that allows to go from an anomaly to a theorem and a model.

2.3 Engineering design to help data science articulate between optimization and creativity

These two literatures point into two different directions to guide model design with data, positioning data science in-between creativity and knowledge-based systems, asking the question of the design process in such a situation. On one side, creativity have long been studied by engineering design researchers, that showed its importance and how to manage it in design processes (Howard et al., 2007; Redelinghuys and Bahill, 2006). For instance, creativity as a stimulus to foster innovation has been studied (Howard et al., 2010). On the other end, the design of knowledge-based systems has also been studied by engineering design. For instance, (Braha and Reich, 2003) formalized the coupled design process theory that boils design of systems down to a combination of knowledge and models. Through their stable closure spaces, they formulate a theory of navigation and assembly within models and knowledge (Hatchuel and Weil, 2008).

Both these dimensions of creativity and knowledge navigation in design process have been studied but remains the question of their articulation. Among researchers in engineering design, the systematic design, formalized by (Pahl et al., 2007) can be read as a good coordination of creative and knowledge layers: conceptual design making great use of creativity for working structures' development, while embodiment design requires circulation between models. Some years later, (Hatchuel et al., 2013) developed C-K theory, an innovative design process managing both creativity and knowledge, and guiding the coupling between both.

3 RESEARCH QUESTION

This literature review shows that many processes have been developed by data science model designers, in other terms data scientists, to respond to value creation from available data. The processes highlighted above go from the traditional optimization logic to more explorative ones, all with a special relation to anomalies. In such a context, the elucidation of design processes, characterized by their relation to anomalies can be raised and comes within the competency of engineering design. Hence, w*hat are the various roles of anomalies in each of the design processes for designing models in data-pushed projects?*

https://doi.org/10.1017/pds.2023.114 Published online by Cambridge University Press

4 METHOD: INTRODUCING A THEORETICAL FRAMEWORK TO DERIVE AND ILLUSTRATE THE VARIOUS DESIGN PROCESSES

To answer this question the methodology adopted is threefold: 1) we will introduce an adapted language, in order to have theoretical lenses adapted to our problem, namely the design of models in an environment rich in data 2) in order to derive the various processes to design scientific models in data-pushed project, we will resort to C-K theory and 3) we will illustrate and study the processes deduced from C-K theory with various case studies, with our theoretical lenses.

4.1 Common theoretical lenses to study design with data science

Data science manuals, among which (Friedman et al., 2001; James et al., 2013; Skiena, 2017) show a similar structure and framework to look at data-pushed problems. It all starts with an input dataset and a desired output data, that are often commonly called X and Y, the objective being to define a model M that will efficiently give Y from X. In even more formal terms, a data-pushed project can be characterized by a triplet (X, M, Y) ensuring Y=M(X). Yet, this formulation forgets an important specificity of such problems: we look for the best model fitting Y=M(X), what can be written in formal terms as $M^* = \underset{M \in \mathcal{M}}{\operatorname{argmin}} Y - MX$ where $\mathcal{M}$ designates a category of models. A category of models is to be understood as a family of models with the same logic, for instance within regression one can find several categories such as linear, logistic, ridge… (Fahrmeir et al., 2021). Hence, data-pushed problems can be written as an optimization problem, that formally boils down to a convergence one.

For illustrative purposes, let us take a quick example given in the "what is data science" chapter of (Skiena, 2017): IMDb database. This database gathers data on millions of movies, from running times, to genre by way of financial data and users' ratings. One of the data-pushed projects that can be carried out from such a database is determining the "probability that someone will like a given movie". Writing all that with our (X, M, Y) framework we have "X=IMDb data base (title, duration, cast, ratings…)", "Y=probability that each user will like a movie" and what is to be design is the model M minimizing the error in the prediction.

This formulation is particularly relevant for several reasons:

- it allows to clearly represent and catch the three main components of a data-pushed project, namely the input data, the output data and the model fitting between both,
- it allows to depict their evolution over time, for instance the addition of new characteristics in the input data (age of the user if we go on with the above given IMDb illustration) or a tightening of the output,
- it allows to represent the various logics of action that one has on each of these components and the impact it has on the other two.

4.2 C-K theory to derive a variety of design processes in data science

Having an interpretative framework and a formal formulation of what is to be designed in a data-pushed project, we will resort to C-K theory (Hatchuel and Weil, 2008, 2003). Briefly, this theory clarifies the reasoning logic of a design process, seeing it as the interplay between two spaces: a space of knowledge, K, and a space of concept C. The K space is composed of knowledge, formally propositions with a logical status, that can guide the designer in his design process. Whereas the C-space is composed of concepts, namely propositions without logical status, that the designer can explore. One of the major points of this theory are the interplays between these two spaces, allowing them to evolve and grow together, thanks to four operators:

- C→K, knowledge creation from concept exploration,
- K→C, concept extension from accumulated knowledge,
- C→C, concept extension from concept partition,
- K→K, knowledge creation from previously accumulated knowledge.

This theory has already been used and shown its value in many fields for the design of both product and process and it has recently been fruitfully applied to data science, especially to understand the way data scientist reason to design a model (Barbier et al., 2021; Kazakçı, 2015).

This framework, combined with the formal writing we derived above, allows us to represent the design logic between model design in data-pushed projects, and the mechanisms of C-K theory will lead us to various design processes, as illustrated on Figure 1.

https://doi.org/10.1017/pds.2023.114 Published online by Cambridge University Press

i) Let us start with the initial concept of a data-pushed project, our C_0 that we formalize in 4.1: given X and Y, design M^*such that $M^* = \underset{M \in \mathcal{M}}{\operatorname{argmin}} Y - MX$. We verify that it well is a concept since our knowledge base consists only of statistical knowledge, data knowledge (obtained from data exploration) and environmental knowledge.

ii) Looking at the knowledge we have, namely data and statistical one, the very first and common strategy to respond to this optimization problem is by selecting one category of models from which we will optimize. Let us remark that this is the logic adopted in many data science manuals, conveying the dominant design highlighted by (Kazakçı, 2015). This strategy follows the successive steps: selecting one category of models and optimizing the parameters defining the models within it, what boils down to a problem-solving approach. The main issue here is de facto in the design of the right category of model within which we will have to look for the optimal model and not so much in the effective design of the model itself. Let us note that this design strategy creates new knowledge, nothing other than an operator of optimization within the fixed $\mathcal{M}^*$category. The question of what happens when we are confronted to one (or several) points (X_0, Y_0) such that MX_0 is far from Y_0, in other terms we are facing an anomaly, can be raised. And in the context of this strategy, we can suppose that the designer will remove such points and consider them outside the scope of the model.

iii) Now, from this same base of knowledge and from the first strategy derived comes another one, consisting in refining the category of model we are in throughout the process of optimizing leading to the optimal model. We can think of this strategy as a sort of symbiosis between the "under-construction model" and the category since the latter will grow thanks to the learnings induced by the mere optimization process. Here it is important to note that the category of models one starts with is not fixed in advance and will be fine tunned during the process of designing the model. Hence, it induces more sophisticated knowledge, namely the circulation and layout between the successively refined categories of models. One can wonder how the circulation between the refined categories is managed. Here we can hypothesize that this is done thanks to integration of anomalous points with a logic of fine tuning. To illustrate, let us consider that we are optimizing within the category $\mathcal{M} = \mathcal{M}_1 \cap \mathcal{M}_2$, leading us to the model $\mathrm{M}_{1,2}$, that fails to map X_0 on Y_0 (what is an anomalous situation), then the fine-tuning logic will indicate to refine $\mathcal{M}$ into $\mathcal{M}_1 \cap \mathcal{M}_2 \cap \mathcal{M}_3$ to integrate this anomaly, and so on and so forth.

iv) The two previous design processes only resort on statistical knowledge and mathematical categories of models, forgetting another pocket of knowledge, the one of environmental knowledge. This is not necessarily mathematical and statistical knowledge but rather knowledge regarding for instance data creation, collection or needs the expected output Y should address. Resorting to this hidden pocket of knowledge opens a whole new design process, that completes the second one. This process consists in multiplying the categories in which to optimize and to resort to non-statistical one for some dimensions of the design problem when required. This is in fact a logic of exploration in the category of all categories of models, whether statistical or not. Consequently, this logic is creating new knowledge regarding an operator of coherent exploration of the unknown. Almost like the second process, we can suppose that this exploration of the categories of models is driven by anomalous points, not only they will indicate how to refine the categories, but they will also indicate when and where to look in $\mathcal{M}_\jmath$.

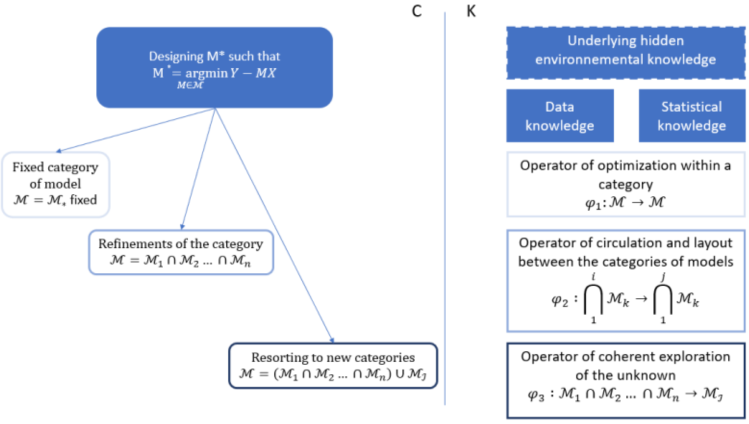

C
K
Designing M* such that $M^* = \underset{M \in \mathcal{M}}{\operatorname{argmin}} Y - MX$
Fixed category of model $\mathcal{M} = \mathcal{M}_*$ fixed
Refinements of the category $\mathcal{M} = \mathcal{M}_1 \cap \mathcal{M}_2 \dots \cap \mathcal{M}_n$
Resorting to new categories $\mathcal{M} = (\mathcal{M}_1 \cap \mathcal{M}_2 \dots \cap \mathcal{M}_n) \cup \mathcal{M}_\jmath$
Underlying hidden environnemental knowledge
Data knowledge
Statistical knowledge
Operator of optimization within a category $\varphi_1 : \mathcal{M} \to \mathcal{M}$
Operator of circulation and layout between the categories of models $\varphi_2 : \bigcap_1^i \mathcal{M}_k \to \bigcap_1^j \mathcal{M}_k$
Operator of coherent exploration of the unknown $\varphi_3 : \mathcal{M}_1 \cap \mathcal{M}_2 \dots \cap \mathcal{M}_n \to \mathcal{M}_\jmath$

Figure 1 - C-K theory used to represent the various processes to design in data-pushed projects. In the K-space, the blue background pockets refer to common initial knowledge, whereas the three pockets under refer to knowledge generated by each of the strategies

https://doi.org/10.1017/pds.2023.114 Published online by Cambridge University Press

4.3 Three case studies to illustrate the three processes theoretically obtained

The previous theoretical developments gave us three design processes and allowed us to formulate hypotheses regarding the role of anomalies and the varying extent of knowledge exploration in each. The aim of the coming case studies is to examine these hypotheses, relying upon the adapted language introduced in 4.1, that, we recall, is composed of three dimensions (X, M, Y) and allows to understand the interaction between these three overtimes. The three case studies, summarized in Figure 2, all take place in a data science context, with data as common "raw material".

	Case study 1	**Case study 2**	**Case study 3**
X	House characteristics	Outdoor warehouse stock images	Activity data from a mobile app
Y	Sales prices	Number of pallets in the warehouse	Cluster of users according to their use
Strategy adopted to design M	Optimisation of a model within a predefined category	Successive refinements of the category of models	Exploration and use of unknown categories of models

Figure 2 - Summary of the characteristics of each case study

4.3.1 Case study 1: dominant design in data science, illustration by the books

To illustrate and deepen our understanding of the first process, that is the dominant one in data science, we resort to a famous competition platform, Kaggle and selected the "house prices" one, more specifically focusing on the strategy adopted by (Lu et al., 2017). Being a competition for beginners it is expected to well represent the dominant strategy and being published in IEEM means peer reviewing hence validation of the relevance of the process in this paper. To characterize more this case, the authors were given a dataset of house characteristics in Ames Iowa, precisely 79 explanatory variables (X), such as address, condition, building year, superficies… The objective was to predict the sales price of the house (Y), for what the process adopted follows these steps, as formalized in Figure 3.

1. Exploration and understanding of X,
2. Selection of the relevant category of models for predicting this Y (more precisely, here within the regression, the authors chose to optimize in several categories in parallel, namely Ridge, Lasso and Gradient Boosting),
3. Optimization of the parameters to converge to the optimal model from the fixed category (what is done in step 1 while step 2 consists in the creation of a new category of models, that is combination of Lasso and Gradient Boosting, in which they will launch another optimization to find the optimal coefficients of the linear combination).

	X	M	Y
Step 0	House characteristics	None	Sales prices
Step 1	House characteristics	**Ridge regression, Lasso regression, Gradient Boosting**	Sales prices
Step 2	House characteristics	**Linear combination of Lasso regression and Gradient Boosting**	Sales prices

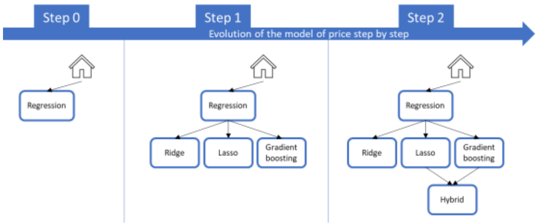


Figure 3 - Explication of the various steps followed by the competitors, with respect to the evolution of the triplet (X, M, Y)

Looking at this first process, it all starts with statistical and mathematical knowledge in order to fix a relevant category of model, what can be explicitly read in (Lu et al., 2017) when they choose three categories of models within which they will optimize to obtain the optimal model. It is to be noted that this initial pocket of knowledge is driven by the output Y. The second major source of knowledge for this strategy is the input data itself, after what these authors call "creative feature engineering". To summarize, the knowledge generated to start this process is independent from any environmental knowledge (in our case that we are looking at houses and sales price).

Regarding anomalies, the authors made it very clear when they say that they "remove 4 abnormal points", making us understand that in this optimization logic, the process is guided by the reduction of the number of anomalies. Yet, these anomalies do guide the process but do not shed light on the pursuit of the process.

https://doi.org/10.1017/pds.2023.114 Published online by Cambridge University Press

4.3.2 Case study 2: successive refinements of the category of model, illustration in a factory

Our second case study is based on an experimentation carried out by the supervision of a group of three engineering students, during a 3-months project, working for a French company. The students had access to the whole resources of the company (in terms of data, people and knowledge) and were given an input dataset made of outdoor warehouse stock images (X), from a Dutch factory, with whom they worked closely. The company asked them to improve their stock management in this factory from these images, what sums up to predicting the number of pallets in the warehouse (Y). We were able to follow the strategy adopted by the students thanks to weekly meetings with them and participation in steering committees with the company, what is summarized in Figure 4. Hence, their process can be described as follow:

1. Selection of a relevant category of models for predicting this Y (here K-means algorithms),
2. First round of optimization leading to new learnings to refine the category of models ($\mathcal{M} = \mathcal{M}_1 \cap \mathcal{M}_2$),
3. Optimization within this new $\mathcal{M}$ and iteration of this process until convergence and is exemplary of the second one described in 4.2.

	X	M	Y
Step 1	Stock images directly from camera	None	**Number of pallets**
Step 2	Stock images directly from camera	**K-means algorithm**	Number of pallets
Step 3	**Stock images modified by image analysis**	**K-means algorithm + image analysis (contrast, fuzzy...)**	Number of pallets
Step 4	Stock images modified by image analysis	K-means algorithm + image analysis (contrast, fuzzy...)	**Number of pallets for same orientation and time images**

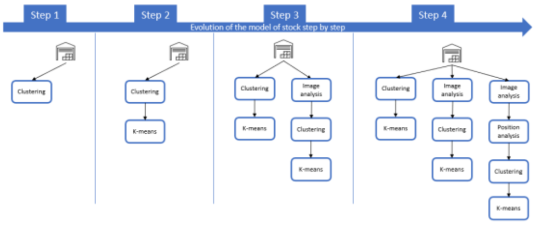


Figure 4 - Explication of the various steps followed by the students, with respect to the evolution of the triplet (X, M, Y)

This second process starts with the same pocket of knowledge, namely data and statistical one. As an illustration, let us say that when choosing K-means clustering algorithm (Alsabti et al., 1997), students did not take into account the fact they were dealing with stock of pallets. However, the main difference comes when they allow themselves to learn from the iteration during the optimization process, hence to learn from the models and use this knowledge to re-open the statistical pocket of knowledge. In other terms, they let themselves another lever of learnings to explore the world of statistical models and refine their initial category.

Regarding anomalies, their role tends to be more sophisticated since they indicate the moment when the statistical pocket of knowledge needs to be reopened, to sophisticate the category of model. For instance, in step 2 in Figure 4, anomalies indicate students the importance to add an image analysis model (hence resorting to a whole new category of model) to improve the efficiency of their model. This is the same logic that supports step 4 and the resort to repositioning models.

4.3.3 Case study 3: unexpected knowledge strategy, illustration from a healthcare company

Finally, our third case study is based on another experimentation, carried out with a French healthcare company, on a data-pushed project with the objective to understand user behaviour of a mobile app. Starting from usage data (X) gathered from a mobile application this company launched, the aim was to understand user behaviour and derive a model of users (M), hence the expectation was to create clusters of users (Y). The researcher had access to all the details of the project (interviews with experts and participation in steering committees), what allows us to describe the strategy adopted as follow:

1. Selection of an exploratory category of models to extent the knowledge base,
2. Resort to an orthogonal (yet statistical) category of model to create first clusters,
3. Qualitative actions to complete the model and refine the understanding of the clusters.

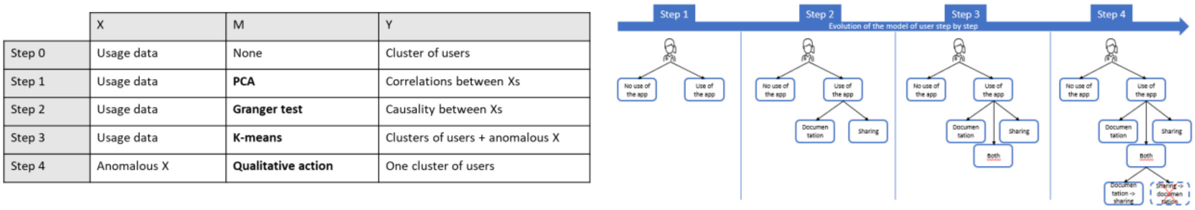
The construction of the clusters of usage during this process are depicted under on Figure 5.

This last process voluntarily starts with few explorations of the whole space of categories of models and limits itself to exploratory models. It is then only in a second phase that it explores more widely the statistical pocket of knowledge, guided by the conclusions drawn from the first exploratory analysis. Here, it is of major importance to note that the model designer resort to a whole new pocket

https://doi.org/10.1017/pds.2023.114 Published online by Cambridge University Press

of knowledge and goes out of the world of statistics with qualitative models of actions. This process then appears segmented with 3 distinct phases, each exploiting different pocket of knowledge.

Regarding anomalies, almost like in 4.3.2, anomalies indicate when and where to launch new explorations, especially for the qualitative actions. Not only do they indicate to reopen the statistical pocket of knowledge, but they also point to hidden, thus non considered knowledge. This has several implications: 1) reconsidering the initial category of models, 2) completing and developing the knowledge base, 3) adding new descriptors in the input data that were not initially available.



	X	M	Y
Step 0	Usage data	None	Cluster of users
Step 1	Usage data	**PCA**	Correlations between Xs
Step 2	Usage data	**Granger test**	Causality between Xs
Step 3	Usage data	**K-means**	Clusters of users + anomalous X
Step 4	Anomalous X	**Qualitative action**	One cluster of users

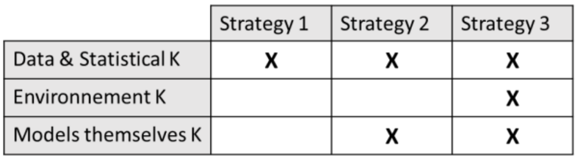
Step 1
Step 2
Step 3
Step 4
Evolution of the model of user step by step
No use of the app
Use of the app
Documen tation
Sharing
Both
Documen tation -> sharing
Sharing -> documen tation

Figure 5- Step-by-step evolution of the model of the user with this strategy, with respect to the evolution of the triplet (X, M, Y)

5 RESULTS: VARIOUS DESIGN PROCESSES GUIDED BY ANOMALIES

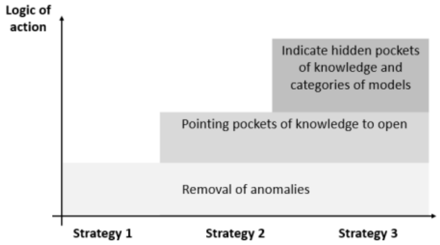
The theoretical framework introduced in 4.1, combined with C-K theory led us to highlight three main design processes in data-pushed projects in 4.2. We hypothesized that these three contrasted design processes were characterized by the various extent of knowledge exploration and the role of anomalies. To test the hypotheses made, three case studies were chosen to well illustrate each one of the design processes: fixed category of models in 4.3.1, closed sequence of categories in 4.3.2 and open sequence of categories in 4.3.3.

We present hereunder results from this study, that have been validated with practitioners, in data science and engineering design, from the two companies we worked with (for case studies 2 and 3).



	Strategy 1	Strategy 2	Strategy 3
Data & Statistical K	X	X	X
Environnement K			X
Models themselves K		X	X

Figure 6 - The various knowledge leveraged for each process study



Logic of action
Indicate hidden pockets of knowledge and categories of models
Pointing pockets of knowledge to open
Removal of anomalies
Strategy 1
Strategy 2
Strategy 3

Figure 7 - Non exclusivity of the logics of action with anomalies

The analysis of the case studies done in 4.3 confirmed the importance of an activity of exploration within the world of models, revealing an increasing intensity of this activity. Not only the intensity fluctuates but also the variety of used knowledge is different, with a multiplication of the sources going from strategy 1 to strategy 3, as summarized in Figure 6.

Moreover, as hypothesized it appeared that these explorations are guided by anomalies, in varying and differentiating ways. In the first process, the case study confirmed that they are to be removed and considered outside the scope of the models. But another dimension was revealed: they can be indicators of when to redesign, thus completely changing the fixed category of models within which the designer optimizes. In the closed sequence of categories process, as supposed, anomalies have a guiding role, in the sense that they indicate when and how to refine the category. What is in fact generalized in the third process: not only do they indicate how to refine the category, but they also open new pockets of knowledge and unknown categories of models.

As shown throughout the case studies, these three logics of actions supported by anomalies are not exclusive, but rather complementary and the elementary one can be combined with the more sophisticated ones, as illustrated in Figure 7. For instance, in the second case study, not only did anomalies allowed students to open new pockets of knowledge but they also guide the optimization thanks to removal. Considered as a whole, the three case studies showed that not only do anomalies guide the process within one process, but they also appear as the bounce point that indicates when to go from one process to another.

https://doi.org/10.1017/pds.2023.114 Published online by Cambridge University Press

This role of anomalies as a guide both in the exploration and the convergence towards on optimal model makes data-pushed projects dividable in smaller data-pushed ones. What allows us to argue that this design exercise is in fact very similar to the design and implementation of industrial processes and not the design of a product or services as many authors argue (Cao, 2018). Indeed, the design of an industrial process works with a succession of restriction and extension until stabilization of the operating domain, this convergence to the optimal operating domain being controlled by anomalies, as can be understood from statistical process control (Oakland, 2007). Explained like that, the parallel is striking: it is precisely the role of anomalies we observed throughout the case studies. Designing a model in data-pushed projects can now be seen as a convergence process towards an optimal operating regime in which the model M optimally ensure Y=M(X). What the case studies showed is the design of this operating regime, controlled by anomalies whose role is to guide exploration and subsequent design efforts.

6 FURTHER RESEARCH AND CONCLUSION

In this paper, we have decided to partition on models in our C-K model, what allowed us to derive conclusions on the role of anomalies in each of the design processes obtained. Yet, the reciprocal question could be interestingly raised: starting from strategies to manage anomalies in data-pushed projects, would we open new design processes in such projects? This question could also extend our work in the direction of a systematic design of model design in data-pushed projects. Even though our approach starts to give some hints, it does not answer to this question, nor does it say what would be the optimal model to manage anomalies.

Such further questions pave the way to more collaborations between engineering design and data-pushed projects. Even though data-pushed projects seem far from the consideration of engineering design, this work showed that the study of such projects with the techniques of engineering design is fruitful and could help in designing models. Conversely, data-pushed projects open new theoretically challenging questions that could enrich the corpus of engineering design.

BIBLIOGRAPHY

Adadi, A., Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alsabti, K., Ranka, S., Singh, V., 1997. An efficient k-means clustering algorithm.

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. Proc. Des. Soc. 1, 3239–3248. https://doi.org/10.1017/pds.2021.585

Barnett, V., Lewis, T., 1984. Outliers in statistical data. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics.

Bloor, D., 1978. Polyhedra and the Abominations of Leviticus. The British Journal for the History of Science 11, 245–272. https://doi.org/10.1017/S000708740004379X

Braha, D., Reich, Y., 2003. Topological structures for modeling engineering design processes. Res Eng Design 14, 185–199. https://doi.org/10.1007/s00163-003-0035-3

Cao, L., 2018. Data Science: A Comprehensive Overview. ACM Comput. Surv. 50, 1–42.

Cascini, G., Nagai, Y., Georgiev, G.V., Zelaya, J., Becattini, N., Boujut, J.F., Casakin, H., Crilly, N., Dekoninck, E., Gero, J., Goel, A., Goldschmidt, G., Gonçalves, M., Grace, K., Hay, L., Le Masson, P., Maher, M.L., Marjanović, D., Motte, D., Papalambros, P., Sosa, R., V, S., Štorga, M., Tversky, B., Yannou, B., Wodehouse, A., 2022. Perspectives on design creativity and innovation research: 10 years later. International Journal of Design Creativity and Innovation 10, 1–30. https://doi.org/10.1080/21650349.2022.2021480

Corral, K., Schuff, D., Schymik, G., Louis, R.S., 2015. Enabling Self-Service BI Through a Dimensional Model Management Warehouse. 2015 Americas Conference on Information Systems, AMCIS 2015.

Davis, N.M., 2013. Human-Computer Co-Creativity: Blending Human and Computational Creativity, in: Ninth Artificial Intelligence and Interactive Digital Entertainment Conference. Presented at the Ninth Artificial Intelligence and Interactive Digital Entertainment Conference.

Daw, A., Karpatne, A., Watkins, W., Read, J., Kumar, V., 2021. Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling. https://doi.org/10.48550/arXiv.1710.11431

Einstein, A., 1987. Letters to Solovine. Philosophical Library.

European Parliament, 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016.

https://doi.org/10.1017/pds.2023.114 Published online by Cambridge University Press

Fahrmeir, L., Kneib, T., Lang, S., Marx, B.D., 2021. Regression Models, in: Fahrmeir, L., Kneib, T., Lang, S., Marx, B.D. (Eds.), Regression: Models, Methods and Applications. Springer, Berlin, Heidelberg, pp. 23–84. https://doi.org/10.1007/978-3-662-63882-8_2

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. Springer series in statistics New York.

Hatchuel, A., Reich, Y., Le Masson, P., Weil, B., Kazakçi, A., 2013. Beyond Models and Decisions: Situating Design through generative functions, in: International Conference on Engineering Design. Séoul, South Korea.

Hatchuel, A., Weil, B., 2008. C-K design theory: an advanced formulation. Res Eng Design 19, 181.

Hatchuel, A., Weil, B., 2003. A new approach of innovative design: an introduction to C-K theory. DS 31: Proceedings of ICED 03, the 14th International Conference on Engineering Design, Stockholm 109-110 (exec.summ.), full paper no. DS31_1794FPC.

Hodge, V., Austin, J., 2004. A Survey of Outlier Detection Methodologies. Artificial Intelligence Review 22, 85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

Holton, 1981. L’imagination scientifique. Gallimard.

Howard, T., Culley, S.J., Dekoninck, E., 2007. Creativity in the Engineering Design Process. DS 42: Proceedings of ICED 2007, the 16th International Conference on Engineering Design, Paris, France, 28.-31.07.2007 329-330 (exec. Summ.), full paper no. DS42_P_493.

Howard, T.J., Dekoninck, E.A., Culley, S.J., 2010. The use of creative stimuli at early stages of industrial product innovation. Res Eng Design 21. https://doi.org/10.1007/s00163-010-0091-4

James, G., Witten, D., Hastie, T., Tibshinari R, 2013. An Introduction to Statistical Learning, New York: Springer. ed.

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 International Conference on Engineering Design.

Kuhn, T., 2021. The Structure of Scientific Revolutions, in: Philosophy after Darwin: Classic and Contemporary Readings. Princeton University Press, pp. 176–177. https://doi.org/10.1515/9781400831296-024

Lu, S., Li, Z., Qin, Z., Yang, X., Goh, R.S.M., 2017. A hybrid regression technique for house prices prediction, in: 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). Presented at the 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), pp. 319–323. https://doi.org/10.1109/IEEM.2017.8289904

Oakland, J., 2007. Statistical Process Control, 6th ed. Routledge, London. https://doi.org/10.4324/9780080551739

Pahl, G., Beitz, W., Feldhusen, J., Grote, K.-H., 2007. Engineering Design. Springer, London. https://doi.org/10.1007/978-1-84628-319-2

Redelinghuys, C., Bahill, A.T., 2006. A framework for the assessment of the creativity of product design teams. Journal of Engineering Design 17, 121–141. https://doi.org/10.1080/09544820500273136

Ruder, S., 2017. An overview of gradient descent optimization algorithms. https://doi.org/10.48550/arXiv.1609.04747

Runco, M.A., 1994. Problem Finding, Problem Solving, and Creativity. Greenwood Publishing Group.

Skiena, S.S., 2017. The Data Science Design Manual, Texts in Computer Science. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-55444-0

Trabucchi, D., Buganza, T., 2018. Data-driven innovation: switching the perspective on Big Data. European Journal of Innovation Management 22, 23–40. https://doi.org/10.1108/EJIM-01-2018-0017

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., Walczak, M., Garcke, J., Bauckhage, C., Schuecker, J., 2021. Informed Machine Learning -- A Taxonomy and Survey of Integrating Knowledge into Learning Systems. IEEE Trans. Knowl. Data Eng. 1–1. https://doi.org/10.1109/TKDE.2021.3079836

https://doi.org/10.1017/pds.2023.114 Published online by Cambridge University Press