Data-push projects and their unique feature: managing with anomalies

Antoine Bordas, Pascal Le Masson, Benoit Weil

Data-push projects and their unique feature: managing with anomalies

Antoine Bordas
Centre de Gestion Scientifique
Mines Paris – PSL University
Paris, France
antoine.bordas@minesparis.psl.eu

Pascal Le Masson
Centre de Gestion Scientifique
Mines Paris – PSL University
Paris, France
pascal.le_masson@minesparis.psl.eu

Benoit Weil
Centre de Gestion Scientifique
Mines Paris – PSL University
Paris, France
benoit.weil@minesparis.psl.eu

***Abstract*—Data-push projects are defined as projects whose objective is to derive valuable insights from an initial database, through model design. They have become increasingly common with the ever-rising availability of data and encompass several other terms found in the literature. Even though there is a growing body of research on how to manage such projects, it is noted by both scholars and practitioners that this remains a main challenge. This paper therefore proposes to address this issue with brand new lenses, after gathering insights from the philosophy of mathematics. This literature is particularly relevant because mathematics is at the heart of data-push projects. Thanks to a longitudinal case study we have been able to demonstrate the role of anomalies, especially putting forward three dimensions in which they act as a resource for the management of data-push projects. As such this paper complements the literature in management of data-push projects, in the brand-new direction of anomalies, therefore opening the path for more investigations in this promising direction.**

Keywords— Data projects, Anomalies, Big Data, Data Science, Project Management

I. INTRODUCTION: THE CHALLENGES OF DATA PROJECTS

Companies have launched their data transformation and are willing to use vast amount of data, to draw value for their customers and for their business. In this march towards data, two approaches have been identified. Either the aim is to design a set of data, to respond to a predefined objective (for instance, given a question, what data can one design to answer it), that we call a *data-pull* mode [1]. The other way consists in starting from a given database and trying to derive valuable insights from it, that we call a *data-push* mode, a mode especially within the competency of data science. This *data-push* mode is frequent in companies, and falls under a great variety of names, such as "data science" or "data mining" among others. Both scholars and practitioners acknowledge difficulties in managing them, even though there is a growing body of research on this matter. This research therefore aims to propose a new approach to manage data-push projects. Hence, remembering the high mathematical and statistical foundations of data science, at the heart of this *data-push* mode of projects, gets us in a theoretical journey in the philosophy of mathematics. In this literature, and more generally in the philosophy of science, scholars highlight the guiding role of anomalies towards a theorem or model. Besides their mention in the philosophy of mathematics, this concept of anomalies has been leveraged in many scientific disciplines, as well as in management science, hence leading us to assume that they have a role for these specific *data-push* projects. Consequently, we wonder *how can anomalies be a resource for enhanced data-push project management?*

The next section will develop the literature review, clarifying the notion of *data-push projects* and assessing the related managing difficulties. Recalling the mathematical specificities of such projects justifies a detour in the philosophy of mathematics, that leads us to hypothesize the role of anomalies for managing data-push projects and formulate our research question. To investigate this question, we will see in the fourth section that we resorted to a longitudinal case study and will briefly introduce theoretical lenses, allowing us to make visible the role of anomalies. In the fifth section, we show that this work confirms our assumption and refines it by explicating three dimensions in which anomalies can help for managing data-push projects. The last section is dedicated to a discussion with the literature and outlines the research perspectives envisaged.

II. LITTERATURE REVIEW: DATA-PUSH PROJECTS AND THE POTENTIAL ROLE OF ANOMALIES

A. Clarifying the concept of data-push projects

The literature is very fertile when it comes to discussing so called "data projects" with several different terms each with their specificities. Four main terms can be found in the literature in management and information systems: "data mining project", "data science project", "big data project" and "analytics project". An extensive research on the Scopus database allowed us to identify several papers each proposing a definition of these projects, that we have been able to conceptualized, drawing inspiration from first and second order concepts [2]. This led us to derive the main characteristics behind these four terms, that we summarize here below:

- "Data mining project": focus on uncovering hidden knowledge and patterns in the data that should be actionable by the organization.
- "Data science project": focus on deriving insights from diverse datasets by resorting on machine learning and statistical techniques.
- "Big data project": focus on deriving insights from data, with a data management dimension, with the 4Vs specificities of big data [3].
- "Analytics project": focus on deriving insights enhancing decision-making and business competitiveness resorting to data analysis techniques.

Hence with the quick analysis done and the definitions derived above, it appears that these four terms each have a precise perimeter but still intersect one another. This can be seen also when looking at crossed references in some papers, such as, for instance [4] dealing with "data science projects", while citing [5] themselves dealing with "analytics projects".

These crossed references can be explained by the thin frontiers in-between these terms. For instance, "big data projects" are considered very similar to traditional IT projects [6]. Consequently, they often encompass issues regarding data management and storage issues, that are not really shared by the three other terms. Still they share some similarities with "data science projects", like the use of statistical techniques [7], [8]. Moreover, the evolution over the years of these terms is also a first indication of the relationships they have with one another: "big data project" appearing as the overarching one over the past ten years, whereas "data mining project" tends to disappear in favour of "data science project". This can be explained by the rather circumscribed definition of "data mining projects" [9], [10] that is now covered by "data science projects". Still, all these terms have for common point that the objective of the project is to derive value only from the data, often forgetting other projects within the company. Even though the objective is to derive value, diving into the definitions of these terms, it appears that the sought value can differ: "analytics projects" will concentrate on enhancements of an existing business, whereas "data science projects" are future-oriented, with anticipation of demand or new business creation, as evidenced by the rise of data products [11], [12].

As hinted in the previous paragraph, there is a need to introduce a new term to well define the category of projects we are looking at, that is to say projects which start with a set of data and have for objective to derive value from it. Such projects can be seen as the opposite of the ones described by [1], where data is the expected product. Relying on an analogy with the famous "technology-push" versus "market-pull" dichotomic framework in technology management, we propose to call the projects we are looking at "*data-push* projects". Like "technology-push" situations start with an emerging technology or combination of existing technology [13], [14], *data-push* ones start with an emerging database or combination of various databases. [15] considers that technology-push start with "shelved ideas and products with no apparent use", what could be transposed in the realm of data: a *data-push project* starts with shelved data, waiting to create value for a customers or businesses. Defined as such, *data-push* projects share several characteristics with technology-push ones, such as an unknown time to market, a high level of unknown and technology uncertainty, as well as a high level of required R&D efforts [16]. However, the difficulties of data-push might not be exactly the same as the ones encountered in technology-push ones. [15] put forward difficulties around users' comprehension and perceived usefulness of the product, what is expected in the context of *data-push projects* would rather be a long-required time of research and design of a model. Data-push projects start with shelved data but data alone is no product and mathematical transformation is required in order to transform it into information [17]–[19]. These authors show that data is not a ready-to-use product, but rather that it requires several technological and mathematical developments, showing that *data-push projects* do not start with "shelved ideas" but only shelved data still to be transformed.

This term *data-push project* is therefore interesting because it embraces all the ones given above, being independent of the objective value and the various techniques of data transformations employed, what were precisely the two segmenting dimensions of the four terms studied above.

B. Managing data-push projects

The aim of such data-push projects lies in the definition of an efficient model, that will link an input data into a valuable output [20]. Hence these projects have a high mathematical nature, with a strong focus on data science, a scientific discipline that began its journey in 1970 and accelerated it in the 2010s [21]. As a consequence, data-push projects face a double injunction and need to ensure an internal consistency, in the sense that the defined model must be statistically coherent [22], [23]. While, on the other side, they are to live within an environment and therefore must ensure an external validity and be coherent with the purpose they are designed for [24], [25]. This situation is also described by [26] when he explains the two criteria for a good theory, in other terms a good model: "inner perfection" and "external validation". This situation draws two extremes that can be found in companies: either the model is over-optimized, ensuring a high mathematical consistency, but lacks external actionability [27]; whereas the other extreme would be a constant discussion with the stakeholders, leading to difficulties in modelling [28].

In order to overcome this difficulty and after [29] highlighted a need for better management processes, scholars have developed a number of techniques in order to carry out such projects.

One of the first processes derived by the literature, known as Knowledge Discovery in Databases (KDD), is now almost 30 years old [30] and consists of five steps to transform raw data into knowledge. Another widely used methodology to manage such projects is Cross Industry Standard Process for Data Mining (CRISP-DM), initially defined for data mining projects and mentions six phases, from business understanding to deployment [31]. A number of other methodologies have been derived based on these two, as explained by [32], yet all these methodologies appear as basically similar and none of them is widely accepted, as evidenced by the massive use of individual

methodologies [33]. A more detailed review and assessment of existing methodologies, as well as their brief history, can be found in [34]. These authors show also that most strategies tend to focus only on one aspect, be it project management, team management or data management, whereas all three are essential to reduce the failure rate of such projects.

This is in line with the numerous failures admitted in data-push projects, even with a growing body of research [35], that have difficulties in finding the right balance between internal and external validity. This situation is symptomatic of the lack of specificities of these methodologies, that forget the specificities of data-push projects [34], [36]. Let us remark that this points into two directions. First, it brings us back to the logic of industrial research for which [37] explained the strategies developed by companies to ensure the right balance between consistency of the designed scientific models and actionability for the company. Second, it calls for more focus on the specificities of such projects, namely their high mathematical nature [20].

C. Anomalies as a potential resource?

We will follow the second path, remembering the high mathematical nature of *data-push projects*, whose objective is, we recall, the design a mathematical model to derive valuable insights, we propose to dive into the philosophy and epistemology of mathematics to gather insights. [38] in a famous interpretation of [39] reveals the major role of anomalies as guiding the modelling process. The author recognizes the role of anomalies in knowledge constitution and details some possible responses (among which are exclusion or integration) to deal with such data, on the example of Euler's polyhedron formula. At that point, it is to be noted that the term "anomaly" is vaguely defined, as "things which do not fit into the boxes and boundaries of accepted ways of thinking". In the same direction, [40] discussed the role of anomalies in statistics, estimating they are "a signal […] from which we may learn a lesson", hence guiding exploratory projects. Yet the author does not give a precise definition either and speaks of "apparently wild or outlying observations". Among the references in [40], [41] makes a first attempt to rigorously define an anomaly: it will be an observation generated by a variation of the normal population, distributed according to the normal distribution. This modelling of an anomaly makes us understand the term "contaminated data", used by [41] to describe such anomalies, and it allows him to give procedures to exclude them from the analysis. Interestingly, the author explains it could be possible to draw insights on the "normal distribution" of data by focusing on the anomalous ones, yet under the strong hypotheses of his modelling. Several decades later, [42] confirms the opportunity that anomalies are in the context of data science and proposes to shift from the classical detection framework [43], dominant in data science, to a detection and analysis ones, in other terms to manage anomalies.

The previous paragraph focuses on a literature centred on mathematics, whether it be its epistemology or statistics, but it is to be noted that anomalies is a concept widely leveraged in many disciplines. A quick look at the Scopus database shows that anomalies have been first mentioned by [44] and more generally in a medical context to describe patients' state such as malformations. Over the decades, anomalies have been considered in many disciplines such as physics or engineering, to qualify states of matter for instance [45] and very recently to find "patterns in the data that do not conform to expected behaviour" [43]. All this literature gives us insights to derive a definition and a modelling of what an anomaly is. Yet it does not give us insights on how to deal with such points and to what extent can they be useful in guiding an exploratory process [46], [47].

Once again looking at the philosophy of science can give insights. For instance, [48] give a list of possible responses scientists can give to anomalies, illustrated by several historical examples, all in the context of physical science. Among these responses, we find several levels of ignoring anomalies, as well as several levels of model change, what has been completed some years later [49]; adding an eighth dimension, namely uncertainty regarding the anomalous data. The use of anomalies in the making of science and models have also been largely discussed by [50].

Finally, it is of major importance to note that anomalies and alike notions have been also leveraged by management scholars. For instance, [51] put forward four roles of what they called paradoxes, "contrary or contradictory assumptions", in building organization theories, what has been continued afterwards [52], [53]. These papers in management use anomalies within a theoretical setting aiming to improve the understanding of theory creation from an academic point of view. In a close, yet different stream of research, [54] have also leveraged anomalies, what they call "paradox", to study Sydney Opera House project. Doing so, they demonstrate the relevance of anomalies and their generative role in project management, when dealt in a dialogical way.

The literature brought above from the philosophy of mathematics showed that anomalies are a key parameter for managing model design in this discipline [38]–[40]. An assumption that has been reinforced by philosophers of other disciplines, such as physics [49], and also management scholars when they use anomalies to design theories or study management principles [51], [54]. Coming back to data-push projects, anomalies hence appear as a relevant concept for their management, given:

- Their high mathematical grounding [20], that pointed into the direction of the philosophy of mathematics from which anomalies emerged,
- Their high exploratory level and the consequent numerous unknowns [34], [55], [56], what is a commonality with the papers seen in A.

Yet, remains to conceptualize the notion of anomaly, since it suffers a proper definition, as evidenced by the various terms used and the various definitions used: "data points that deviate markedly from others" [52], "occurrences in a dataset that are in some way unusual and do not fit the general pattern" [42] or "data that appears as inconsistent with theory A" [48]. Then, roughly speaking, an anomaly is an observation that is far from

an initial a priori model, what, in more formal terms, means an anomaly is a triplet made of: an observation O, an a priori model $M$and a metric d, such that $d(O,M) > 0$. To illustrate this example, let us have a quick look at Mercury perihelion [57]: when Le Verrier observed Mercury's precession rate, O in our framework, the accepted model M was Newton's two body system and the metric d was the alignment with the predicted precession, hence we see this historic anomaly. These three dimensions of the definition we propose of an anomaly already indicate its fruitfulness for managing data-push projects, since it is in line with some responses to anomalies given in various stream of research [48], [58].

III. Assumption and Research Question

The literature review showed the existence of at least four terms to speak about "data projects", each one having a precise definition but thin distinctions with the others. Moreover, none of them truly embraced what we want to discuss, namely projects starting from the data and with the objective to derive value from it. Management scholars especially acknowledge high failure rates in data-push projects, even though they studied and derived several methodologies to manage them, calling for a new perspective on these issues.

Recalling the mathematical specificities of these projects, we proposed to look at the epistemology of mathematics in order to gather insights. This leads to the notion of anomaly, that appeared as a guiding parameter in model design, what has been confirmed by other streams of literature, ranging from physical to management sciences.

This unique discussion between mathematics, epistemology and management allows us to formulate the main assumption of this research, namely the role of anomalies for managing data-push projects. Hence the research question of this paper:

how can anomalies be a resource for enhanced data-push project management?

IV. Methodology: a Case Study in the Context of Healthcare Digitalization

A. Empirical setting

To answer this question, we use a qualitative research method, based on one longitudinal in-depth case study [59] from a French healthcare organization. The data-push project we selected takes place in a digitalizing context, where the company entered the market of mobile applications for healthcare professional and consists in segmenting users. Organizationally, this project was initially in the R&D department before the recent creation of a digital department, to which it is now attached.

This project is particularly relevant to study our question for several reasons. First, it is representative of the class of data-push projects, on the particular and classic question of users' segmentation. Second, it does not suffer the traditional flaws of such projects, namely accessibility of the data and lack of resources, allowing us to control classical problems of such projects, hence focusing on the role played by anomalies. One of the authors collaborated closely with the company's data scientist to design the models, what gave us access to a unique empirical material: interviews with experts, participation in project meetings and steering committees. Two model design strategies have been carried out during this data-push project, that thanks to this investigation have been characterized with their relations towards anomalies. One "anomaly agnostic", in tune with the standards given in data science manuals and the second, "anomaly pragmatic", both will be detailed and described in more details below.

B. Data analysis

The material collected regarded each of these two strategies had to be analysed to highlight anomalies and their role, what we describe hereunder.

a) An adapted framework to detect and decipher anomalies

The framework we developed is to a large extent inspired from data science manuals [20], [22], [23], where data-push projects are read with three components:

- X, the input data, in our case usage data of the mobile application,
- Y, the output data, in our case clusters of users,
- M, the mathematical or statistical model,

the aim being to transform X into Y with the to be defined model M, in more formal terms to ensure $Y = M(X)$. In this framework and in line with the literature, an anomaly is defined as a data point that does not conform to a defined notion of normality. The literature in data science often considers anomalies in X, referring to data quality issues [60], whereas anomalies in Y can reveal a lack of quality in the modelling process.

Hence, we will follow the various steps of each strategy with these parameters, emphasizing on X and Y, as explained in b).

b) Encoding accordingly the material for each strategy

Consequently, we will follow X and Y, step by step, throughout the whole project, a step being defined as any change in either X or Y. This will allow to reveal anomalies and the role they played. It is of major importance to note that the framework introduced above is relevant to study organizational and managerial dimensions, since the three components, X, Y and M, all have an organizational twin: X refers to data providers, Y to clients and M to experts.

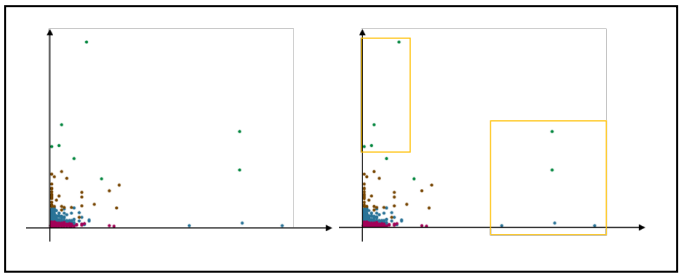

Figure 1- Results of the clustering algorithm and subsequent anomalies

After having participated in the project, especially steering committees and internal documents, we have been able to complete the following tables (see Table 1 and Table 2).
The logic described in Table 2 is to be understood as iterative and systematic. As illustrated on the right part of Figure 1, the dots in yellow squares are considered as anomalous ones and each one is a candidate for further investigations. During the project, this strategy of further investigations was carried out on several of these anomalous points, each time revealing new descriptors that had to be gathered with non-mathematical methods. To take an example, investigations carried out on the dots in the top-left corner yellow square revealed a new descriptor, that has to do with "users business models", what was clearly out of the input database and would hardly be identified without such anomalies.

Table 1- Course of the "anomaly agnostic" strategy in the introduced framework

Anomaly agnostic	***X***	***M***	***Y***	***Comments***	***Anomalies***
Step 0	App usage data		Clusters of usage	Defining expected input and output	
Step 1	App usage data	Data transformation	Usage frequency	Pre-processing of the data	Indistinguishable
Step 2	Usage frequency	Clustering algorithm (K-mean)	4 clusters of users	Defining a clustering model	Indistinguishable

Table 2 Course of the "anomaly pragmatic" strategy in the introduced framework

Anomaly pragmatic	***X***	***M***	***Y***	***Comments***	***Anomalies***
Step 0	App usage data		Clusters of users	Defining expected input and output	
Step 1	App usage data	Principal Component Analysis	Action classes	First exploratory phase to gather knowledge	
Step 2	App usage data + Action classes	Clustering algorithm (K-mean)	4 clusters of user	Defining a first clustering model	Anomaly 1: pointing out new descriptor and further investigations
Step 3	App usage data + Action classes + user business model	Clustering algorithm (K-mean) + acquired knowledge	4 clusters of users	Refinement of the previous models	Anomaly 2: design implications

V. FINDINGS: ANOMALIES, A CONFIRMED RESOURCE FOR DATA-PUSH PROJECTS

As developed in IV.B, we have been able to compare, thanks to our case study, two different strategies carried out during a data-push project. The first strategy, called "anomaly agnostic", that applies the good practices given in data science manual starts with a data processing and transformation step. What we revealed is the perverse results of this step since it makes potential anomalies invisible and even indistinguishable. What in fine reduces the learning potential of the data scientist in charge of the project. Whereas, in the second strategy, characterized as "anomaly pragmatic", it revealed the various roles of anomalies. Let us note that overlooking anomalies is always a managing strategy in this context, even though our research suggests going further and using them as means to help in management of such projects, as described hereunder and summarized on Figure 2.

a) Managing anomalies for database refinements

The research carried out here with the systematic focus made on different anomalous points shows two dimensions in which anomalies are relevant for enhancing the input database:

- Indicating new descriptors to add in the initial database, that were initially and otherwise inaccessible to the data scientist and allowed comprehension of the anomalous phenomenon,
- In the meantime, this knowledge creation allows to refine the analysis and indicate directions in which to further investigations to achieve beforehand defined business goals.

It is important to note that the new descriptors mentioned in the first dot cannot be deduced from the initial ones in the database. In that sense they are orthogonal to the already available ones and allow new dimensions in the analysis, thus they open a new phase of the project where it is important to relaunch the analysis with the completed database. Gathering these new descriptors is precisely guided by anomalies, that by essence indicate ways to obtain them.

b) Managing anomalies for organizational repositioning

It is important to note that the second point has for consequence to blur the lines of the data-push project with the rest of the organization, in the sense that gathering a new descriptor might require new competencies within the project. For instance, still on the first anomaly case, working on "users business model" implies to work closely with initially unexpected departments of the company (such as marketing and sales for instance). Not only blurring the lines, anomalies also act as indicators of when the project needs to open to the exterior to ensure the right balance between internal and external validity. Once investigations carried out on the anomaly naturally indicate the new actors required, sometimes for a specific sub-project but sometimes for the whole rest of the project. Moreover, as seen on the example described above, it also gives insights on how to ensure this balance and what should be the interactions to

design, whether within the company's organization and with external stakeholders.

c) Managing anomalies for risk analysis

Another use of anomalies that can be seen through this case study is their quantifying role in risk analysis. In addition to indicating further investigations, as described in the previous paragraphs, they implicitly indicate the risk taken by the project manager when he decides not to carry in the indicated direction. Once again, we can read it on the first anomaly case: it indicated to focus efforts on one kind of users, hence indicating both the operating mode of the model and the subsequent risk taken by not modifying it. More formally, as can be observed on Figure 1, a set of homogenous anomalies in one square represent a percentage of the whole population. What allows to estimate probabilities, hence risk taken, as suggested in the framework of decision analysis [61].

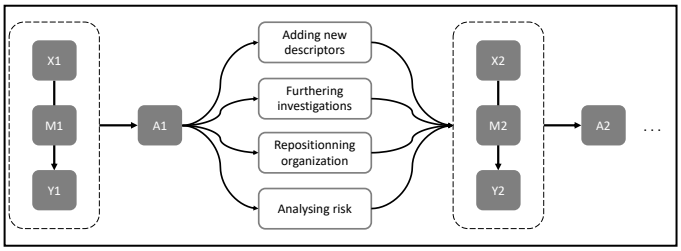

Figure 2 - Illustrative summary of the various logics of action with anomalies in a data-push project

VI. Implications and discussion

A. Anomalies: a bridge between management and mathematics around data

As explained, this research highlighted that one of the learning dimensions of an "anomaly pragmatic" strategy indicates new descriptors to complete the initial database. Yet, we must be careful not to assimilate this strategy to the ones typically found in data science and statistical learning [62]. In statistical learning, the dominant strategy is actually to grow the initial database or improve its quality, often with the same type of data [60], [63], whereas our research aims rather at a frugal completion of the database, focusing on new descriptors. This frugal completion of the database appears in accordance with some of the last results in statistics, for instance:

- [64] showed that more data asymptotically hurt the confidence interval for the least square estimator,
- [65] showed that increasing data size can reduce model performance.

This logic of frugal database augmentation as allowed by anomalies encourages to nuance the various Vs defining Big Data. To recall, the three main admitted characteristics defining Big Data are volume, variety, velocity [3], [66] and each is discussed in the light of anomalies. Looking at volume for instance, the case studied in this paper pleas for a controlled flow of data generation, on the contrary to variety where this same case study pleas for adding new but precisely pertinent data.

B. Data brings back the traditional logic of R&D

Second, behind completion of the initial database, anomalies tend to indicate directions in which to further investigations to generate learnings, as described in V. As such, they appear as an extension of the probe-and-learn strategy [67], emphasizing the need and major role of knowledge production, modelling and learning from them in such data-push projects. This logic is well-known and typically found in R&D projects, dating us back to the beginnings of industrial research, when knowledge production was systematized and the relation with the rest of the company institutionalized [37], [68]. Yet, as we have seen, this project was considered a failure and could not survive in the context of an R&D department, thus was transferred to a business one. In other terms, the project we studied, that we think is representative of data-push projects, and more generally emblematic of digitalization, deploys techniques of R&D management outside an R&D department. Consequently, this research pleads for applying R&D methods [69] behind the only scope of the R&D department.

All in all, anomalies appear close to what the literature calls "weak signals", as was first developed by [70], defined as "advanced indicators of change phenomena". This parallel between anomalies and weak signals have already been done, especially by the latter mentioned paper. Yet the author, as many of the related literature, follows the direction of "weak signals that indicate threats", often seeing them as indicators of what to prevent, an anticipation tool. Instead, our research suggest that this consideration comes down to managing anomalies for risk analysis, hence staying in one managing mode, whereas we show the existence of at least two other managing modes.

VII. Conclusion and next steps

Overall, this paper brings a new perspective regarding the data transformation of companies. We introduced the concept of *data-pushed projects*, with the aim to unify the various terms currently used in the literature. Focusing on the mathematical specificities of such projects, we drew insights from the philosophy of mathematics that put forward anomalies as a guiding concept for such projects. A longitudinal case study allowed us to confirm the assumption regarding the role of anomalies for managing data-push projects, showing they are a mean for management in three different ways: to learn, to ensure the external/internal balance and to analyse risk. Hence it responds to the literature in technology management, especially by extending the probe-and-learn strategy and precising the agile stage-gate methodologies, with this anomaly-driven management. It also complements the literature in data projects management by introducing a brand-new point of view, one enabled by anomalies, and paving the way to further research. Moreover, for practitioners, it positions anomalies as key milestones during data-push projects, with actionable managing strategies associated: sometimes indicating ways to carry on, sometimes new ways to follow. It could therefore suggest new ways to evaluate data-push projects as well as the required competencies within a team of data scientists, what after this work remain open questions.

This research hence proves the relevance of anomalies, that act as weak signals, yet is limited to the context of one case study in a healthcare company. It could therefore be of interest to complement it with other case studies from other industrial contexts and in companies with a different maturity regarding data issues.

REFERENCES

[1] D. Trabucchi and T. Buganza, 'Data-driven innovation: switching the perspective on Big Data', *European Journal of Innovation Management*, vol. 22, no. 1, pp. 23–40, Jan. 2018, doi: 10.1108/EJIM-01-2018-0017.

[2] D. A. Gioia, K. G. Corley, and A. L. Hamilton, 'Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology', *Organizational Research Methods*, vol. 16, no. 1, pp. 15–31, Jan. 2013, doi: 10.1177/1094428112452151.

[3] A. Gandomi and M. Haider, 'Beyond the hype: Big data concepts, methods, and analytics', *International journal of information management*, vol. 35, no. 2, Art. no. 2, 2015.

[4] J. Saltz, I. Shamshurin, and C. Connors, 'Predicting data science sociotechnical execution challenges by categorizing data science projects', *Journal of the Association for Information Science and Technology*, vol. 68, no. 12, pp. 2720–2728, 2017, doi: 10.1002/asi.23873.

[5] J. Gao, A. Koronios, and S. Selle, 'Towards A Process View on Critical Success Factors in Big Data Analytics Projects', presented at the Americas Conference on Information Systems, 2015. Accessed: Aug. 02, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Towards-A-Process-View-on-Critical-Success-Factors-Gao-Koronios/247bfe6fa3365d74bd98c2c460785d62c3d7561d

[6] M. Volk, N. Jamous, and K. Turowski, *Ask the Right Questions: Requirements Engineering for the Execution of Big Data Projects Full Paper*. 2017.

[7] T. Aho, O. Sievi-Korte, T. Kilamo, S. Yaman, and T. Mikkonen, 'Demystifying Data Science Projects: A Look on the People and Process of Data Science Today', in *Product-Focused Software Process Improvement*, M. Morisio, M. Torchiano, and A. Jedlitschka, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 153–167. doi: 10.1007/978-3-030-64148-1_10.

[8] J. Saltz, I. Shamshurin, and C. Connors, 'A Framework for Describing Big Data Projects', in *Business Information Systems Workshops*, W. Abramowicz, R. Alt, and B. Franczyk, Eds., in Lecture Notes in Business Information Processing. Cham: Springer International Publishing, 2017, pp. 183–195. doi: 10.1007/978-3-319-52464-1_17.

[9] K. Julisch, 'Data Mining for Intrusion Detection', in *Applications of Data Mining in Computer Security*, D. Barbará and S. Jajodia, Eds., in Advances in Information Security. , Boston, MA: Springer US, 2002, pp. 33–62. doi: 10.1007/978-1-4615-0953-0_2.

[10] Ó. Marbán, G. Mariscal, E. Menasalvas, and J. Segovia, 'An Engineering Approach to Data Mining Projects', in *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007, pp. 578–588. doi: 10.1007/978-3-540-77226-2_59.

[11] T. H. Davenport and S. Kudyba, 'Designing and Developing Analytics-Based Data Products', *MIT SMR*, Sep. 2016, Accessed: Aug. 11, 2023. [Online]. Available: https://sloanreview.mit.edu/article/designing-and-developing-analytics-based-data-products/

[12] J. Meierhofer, T. Stadelmann, and M. Cieliebak, 'Data Products', in *Applied Data Science: Lessons Learned for the Data-Driven Business*, M. Braschler, T. Stadelmann, and K. Stockinger, Eds., Cham: Springer International Publishing, 2019, pp. 47–61. doi: 10.1007/978-3-030-11821-1_4.

[13] S. R. Chidamber and H. B. Kon, 'A research retrospective of innovation inception and success: the technology–push, demand–pull question', *International Journal of Technology Management*, vol. 9, no. 1, pp. 94–112, Jan. 1994, doi: 10.1504/IJTM.1994.025565.

[14] C. Herstatt and C. Lettl, 'Management of "Technology Push" Development Projects', *International Journal of Technology Management*, vol. 27, Jan. 2000, doi: 10.1504/IJTM.2004.003950.

[15] W. E. Souder, 'Improving Productivity Through Technology Push', *Research-Technology Management*, Jan. 2016, Accessed: Jul. 31, 2023. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/08956308.1989.11670582

[16] T. J. Gerpott, *Strategisches Technologie- und Innovationsmanagement*. Schäffer-Poeschel, 2005.

[17] S. Fosso Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, 'How "big data" can make big impact: Findings from a systematic review and a longitudinal case study', *International Journal of Production Economics*, vol. 165, pp. 234–246, Jul. 2015, doi: 10.1016/j.ijpe.2014.12.031.

[18] W. A. Günther, M. H. Rezazade Mehrizi, M. Huysman, and F. Feldberg, 'Debating big data: A literature review on realizing value from big data', *The Journal of Strategic Information Systems*, vol. 26, no. 3, pp. 191–209, Sep. 2017, doi: 10.1016/j.jsis.2017.07.003.

[19] J.-L. Monino, 'Data Value, Big Data Analytics, and Decision-Making', *J Knowl Econ*, vol. 12, no. 1, pp. 256–267, Mar. 2021, doi: 10.1007/s13132-016-0396-2.

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

[21] L. Cao, 'Data Science: Challenges and Directions', *Commun. ACM*, vol. 60, no. 8, pp. 59–68, Jul. 2017, doi: 10.1145/3015456.

[22] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1, no. 10. Springer series in statistics New York, 2001.

[23] G. James, D. Witten, T. Hastie, and Tibshinari R, *An Introduction to Statistical Learning*, New York: Springer. 2013. Accessed: Apr. 27, 2022. [Online]. Available: https://link.springer.com/book/10.1007/978-1-0716-1418-1

[24] L. Breiman, 'Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)', *Statistical Science*, vol. 16, no. 3, pp. 199–231, Aug. 2001, doi: 10.1214/ss/1009213726.

[25] D. Edwards, *Guide to Mathematical Modelling*, Second edition. New York, NY: Industrial Press, Inc., 2007.

[26] G. Holton, 'Constructing a Theory: Einstein's Model', *The American Scholar*, vol. 48, no. 3, pp. 309–340, 1979.

[27] O. A. Montesinos López, A. Montesinos López, and J. Crossa, 'Overfitting, Model Tuning, and Evaluation of Prediction Performance', in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, O. A. Montesinos López, A. Montesinos López, and J. Crossa, Eds., Cham: Springer International Publishing, 2022, pp. 109–139. doi: 10.1007/978-3-030-89010-0_4.

[28] J. Saltz, I. Shamshurin, and K. Crowston, 'Comparing Data Science Project Management Methodologies via a Controlled Experiment', *Hawaii International Conference on System Sciences 2017 (HICSS-50)*, Jan. 2017, Accessed: Aug. 08, 2022. [Online]. Available: http://hdl.handle.net/10125/41273

[29] J. S. Saltz, 'The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness', in *2015 IEEE International Conference on Big Data (Big Data)*, Oct. 2015, pp. 2066–2071. doi: 10.1109/BigData.2015.7363988.

[30] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, 'From Data Mining to Knowledge Discovery in Databases', *AI Mag.*, 1996, Accessed: Feb. 03, 2023. [Online]. Available: https://www.semanticscholar.org/paper/From-Data-Mining-to-Knowledge-Discovery-in-Fayyad-Piatetsky-Shapiro/a1874aafa8730bdd4b28f29d025141c13ee28b58

[31] P. Chapman *et al.*, 'CRISP-DM 1.0: Step-by-step data mining guide', 2000. Accessed: Feb. 03, 2023. [Online]. Available: https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72

[32] G. Mariscal, Ó. Marbán, and C. Fernández, 'A survey of data mining and knowledge discovery process models and methodologies', *The Knowledge Engineering Review*, vol. 25, no. 2, pp. 137–166, Jun. 2010, doi: 10.1017/S0269888910000032.

[33] G. Piatetsky, 'CRISP-DM, still the top methodology for analytics, data mining, or data science projects', KDnuggets. Accessed: Aug. 03, 2023. [Online]. Available: https://www.kdnuggets.com/crisp-dm-still-the-top-methodology-for-analytics-data-mining-or-data-science-projects.html

[34] I. Martinez, E. Viles, and I. G. Olaizola, 'Data Science Methodologies: Current Challenges and Future Approaches', *Big Data Research*, vol. 24, p. 100183, May 2021, doi: 10.1016/j.bdr.2020.100183.

[35] G. Reggio and E. Astesiano, 'Big-Data/Analytics Projects Failure: A Literature Review', in *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Aug. 2020, pp. 246–255. doi: 10.1109/SEAA51224.2020.00050.

[36] H. E. Lange, P. Drews, and M. Höft, '"Ideation is Fine, but Execution is Key": How Incumbent Companies Realize Data-Driven Business Models', in *2021 IEEE 23rd Conference on Business Informatics (CBI)*, Sep. 2021, pp. 191–200. doi: 10.1109/CBI52690.2021.00030.

[37] S. Lenfle, '"Out of the dusty labs". Really? Bell Labs, the transistor and the myth of isolated research', in *R&D Management Conference*, Ecole Polytechnique, Palaiseau, France, Jun. 2019.

[38] D. Bloor, 'Polyhedra and the Abominations of Leviticus', *The British Journal for the History of Science*, vol. 11, no. 3, Art. no. 3, Nov. 1978, doi: 10.1017/S000708740004379X.

[39] I. Lakatos, *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press, 2015.

[40] W. H. Kruskal, 'Some Remarks on Wild Observations', *Technometrics*, vol. 2, no. 1, pp. 1–3, 1960, doi: 10.2307/1266526.

[41] W. J. Dixon, 'Processing Data for Outliers', *Biometrics*, vol. 9, no. 1, pp. 74–89, 1953, doi: 10.2307/3001634.

[42] R. Foorthuis, 'On the nature and types of anomalies: a review of deviations in data', *Int J Data Sci Anal*, vol. 12, no. 4, pp. 297–331, Oct. 2021, doi: 10.1007/s41060-021-00265-1.

[43] V. Chandola, A. Banerjee, and V. Kumar, 'Anomaly Detection: A Survey', *ACM Comput. Surv.*, vol. 41, Jul. 2009, doi: 10.1145/1541880.1541882.

[44] S. W. Fearn, 'Monstrosity. ISngular anomaly of the urinary organs', *The Lancet*, vol. 24, no. 610, pp. 178–179, May 1835, doi: 10.1016/S0140-6736(02)83228-7.

[45] D. G. Westlake, 'Anomalies in the physical properties of vanadium the role of hydrogen', *Philosophical Magazine*, vol. 16, no. 143, pp. 905–908, 1967, doi: 10.1080/14786436708229683.

[46] S. Lenfle, 'Exploration and project management', *International Journal of Project Management*, vol. 26, no. 5, pp. 469–478, Jul. 2008, doi: 10.1016/j.ijproman.2008.05.017.

[47] T. Shibata, Y. Baba, and J. Suzuki, 'Managing exploration persistency in ambidextrous organizations', *R&D Management*, vol. 52, no. 1, pp. 22–37, 2022, doi: 10.1111/radm.12468.

[48] W. F. Brewer and C. A. Chinn, 'Scientists' Responses to Anomalous Data: Evidence from Psychology, History, and Philosophy of Science', *PSA: Proceedings of the*

Biennial Meeting of the Philosophy of Science Association, vol. 1994, pp. 304–313, 1994.

[49] C. A. Chinn and W. F. Brewer, 'An empirical test of a taxonomy of responses to anomalous data in science', *Journal of Research in Science Teaching*, vol. 35, no. 6, pp. 623–654, 1998, doi: 10.1002/(SICI)1098-2736(199808)35:6<623::AID-TEA3>3.0.CO;2-O.

[50] T. Kuhn, 'The Structure of Scientific Revolutions', in *Philosophy after Darwin: Classic and Contemporary Readings*, Princeton University Press, 2021, pp. 176–177. doi: 10.1515/9781400831296-024.

[51] M. S. Poole and A. H. van de Ven, 'Using Paradox to Build Management and Organization Theories', *The Academy of Management Review*, vol. 14, no. 4, pp. 562–578, 1989, doi: 10.2307/258559.

[52] H. Aguinis, R. K. Gottfredson, and H. Joo, 'Best-Practice Recommendations for Defining, Identifying, and Handling Outliers', *Organizational Research Methods*, vol. 16, no. 2, pp. 270–301, Apr. 2013, doi: 10.1177/1094428112470848.

[53] B. De Keyser, A. Guiette, and K. Vandenbempt, 'On the Use of Paradox for Generating Theoretical Contributions in Management and Organization Research', *International Journal of Management Reviews*, vol. 21, no. 2, pp. 143–161, 2019, doi: 10.1111/ijmr.12201.

[54] M. Gaim, S. Clegg, and M. Cunha, 'In Praise of Paradox Persistence: Evidence from the Sydney Opera House Project', *Project Management Journal*, Apr. 2022, doi: 10.1177/87569728221094834.

[55] D. K. Becker, 'Predicting outcomes for big data projects: Big Data Project Dynamics (BDPD): Research in progress', in *2017 IEEE International Conference on Big Data (Big Data)*, Dec. 2017, pp. 2320–2330. doi: 10.1109/BigData.2017.8258186.

[56] A. Feelders, H. Daniels, and M. Holsheimer, 'Methodological and practical aspects of data mining', *Information & Management*, vol. 37, no. 5, pp. 271–281, Aug. 2000, doi: 10.1016/S0378-7206(99)00051-8.

[57] R. Baum and W. Sheehan, *In Search Of Planet Vulcan: The Ghost In Newton's Clockwork Universe*. Basic Books, 2003.

[58] S. W. Evans and M. J. Palmer, 'Anomaly handling and the politics of gene drives', *Journal of Responsible Innovation*, vol. 5, no. sup1, Art. no. sup1, Jan. 2018, doi: 10.1080/23299460.2017.1407911.

[59] R. K. Yin, *Case Study Research: Design and Methods*. SAGE, 2009.

[60] Y. Zhu, 'The Challenges of Data Quality and Data Quality Assessment in the Big Data Era', vol. 14, no. 0, Art. no. 0, May 2015, doi: 10.5334/dsj-2015-002.

[61] H. Raiffa, *Decision analysis: introductory lectures on choices under uncertainty*. in Decision analysis: introductory lectures on choices under uncertainty. Oxford, England: Addison-Wesley, 1968.

[62] T. Kavzoglu, 'Increasing the accuracy of neural network classification using refined training data', *Environmental Modelling & Software*, vol. 24, no. 7, pp. 850–858, Jul. 2009, doi: 10.1016/j.envsoft.2008.11.012.

[63] M. Aleedy, E. Atwell, and S. Meshoul, 'Using AI Chatbots in Education: Recent Advances Challenges and Use Case', in *Artificial Intelligence and Sustainable Computing*, M. Pandit, M. K. Gaur, P. S. Rana, and A. Tiwari, Eds., in Algorithms for Intelligent Systems. Singapore: Springer Nature, 2022, pp. 661–675. doi: 10.1007/978-981-19-1653-3_50.

[64] Z. Li, C. Xie, and Q. Wang, 'Provable More Data Hurt in High Dimensional Least Squares Estimator'. arXiv, Aug. 14, 2020. doi: 10.48550/arXiv.2008.06296.

[65] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, 'Deep double descent: where bigger models and more data hurt*', *J. Stat. Mech.*, vol. 2021, no. 12, p. 124003, Dec. 2021, doi: 10.1088/1742-5468/ac3a74.

[66] P. Russom, 'Big data analytics', *TDWI best practices report, fourth quarter*, vol. 19, no. 4, Art. no. 4, 2011.

[67] L. Magazzini, F. Pammolli, and M. Riccaboni, 'Learning from Failures or Failing to Learn? Lessons from Pharmaceutical R&D', *European Management Review*, vol. 9, no. 1, pp. 45–58, 2012, doi: 10.1111/j.1740-4762.2012.01027.x.

[68] L. S. Reich, *The making of American Industrial research: science and business at GE, 1876-1926*, Cambridge University Press. 2002.

[69] H. J. Thamhain, 'Managing innovative R&D teams', *R&D Management*, vol. 33, no. 3, pp. 297–311, 2003, doi: 10.1111/1467-9310.00299.

[70] H. I. Ansoff, 'Strategic issue management', *Strategic Management Journal*, vol. 1, no. 2, pp. 131–148, 1980, doi: 10.1002/smj.4250010204.