EasyChair Preprint

№ 9248

Implementing a Data Infrastructure to Enable Business Analytics in the Public Sector: a Case Study

Margunn Aanestad, Jens Christian Haatvedt and Annette Alstadsæter

November 5, 2022

Implementing a Data Infrastructure to Enable Business Analytics in the Public Sector: a case study[1]

Margunn Aanestad[1,2][0000-0003-3731-4241], Jens Christian Haatvedt[3,4] and Annette Alstadsæter[4,3][0000-0002-2554-3365]

[1] University of Agder, P.O.Box 422, 4604 Kristiansand, Norway
[2] University of Oslo, P.O.Box 1072 Blindern, 0316 Oslo, Norway
[3] Helfo Fredrikstad, Dokka 6, 1671 Kråkerøy, Norway
[4]NMBU, Universitetstunet 3, 1433 Ås, Norway

margunn.aanestad@uia.no, jens.Christian.Haatvedt@helfo.no, annette.alstadsater@nmbu.no

Abstract.
This paper describes the initial stages of the process of implementing the data infrastructure required to develop analytics capabilities in a public sector organization. Helfo (the Norwegian Health Economics Administration) is responsible for making payments to healthcare actors who submit reimbursement claims. An important task for Helfo is also to prevent and detect errors, and the organization is currently strengthening this capacity though employing data analytics and artificial intelligence. Implementing data analytics entails more than a "plug-and-play" process, and we analyze the initial stages of the implementation process as a sociotechnical change process. As a starting point we employ the CRISP-DM process model and enrich this with additional steps and tasks that was found to be central in the case. We describe in more detail the preparatory work relating to the technical setup and data infrastructures, and the implications for the information processing routines of the organization more broadly. The case study shows that also the early-phase improvements in data access and utilizing basic analytics capabilities yielded instant benefits to the organization, even before employing advanced analytics and artificial intelligence. The rich description of the early stages of the implementation process can be valuable for other public sector organizations that seek to build data analytic capabilities.

Keywords: Data analytics, Data infrastructure, Organizational capabilities, Sociotechnical change.

[1] The authors are affiliated with Helfo on the Innovation project financed by the Resarch Council of Norway, «*Lærende kontrollvirksomhet for å sikre riktig refusjon fra Helserefusjonsordningene*» #321044, Author 1 as project member, Author 2 as Helfo employee, and Author 3 as project leader in a project-based 10% position at Helfo. The views and observations expressed in this article does not express the views of Helfo, and any subjective opinions presented or perceived in the article aret those of the authors in their independend role as academics. Any errors or mispresentation of facts are our own.

1 Introduction

Detection and prevention of fraudulent behavior is crucial to ensure efficient use of public funds and to maintain trust in the government among the population. As part of this, government agencies run internal audits to ensure good governance. AI and the use of existing and new internal data and known red flags from machine learning is an efficient manner to reduce manual audits, increase audit frequency, and to concentrate scarce manual audit resources on the most serious cases. Also, insights from this can be used to tailor information efforts to agents to increase compliance and reduce involuntary errors. Well-functioning and state-of the art digital infrastructures within public agencies are essential to realize this, as is documentation of the implementation processes to ensure knowledge spill-over and realize efficiency gains also in other government agencies. In the current paper we describe as a case study the process of implementing data analytics capabilities in Helfo[2] (The Norwegian Health Economics Administration), who relies on shared data infrastructures administered by another public sector organization.

Implementing novel business intelligence and analytics technologies requires an organizational learning and adaptation process. This poses new challenges and demands to the data infrastructure, enterprise architecture as well as to organizational capabilities [1]. While Information Systems research has started to describe what such processes look like, there is still a dearth of empirical studies of implementation processes. Moreover, the existing debate is mainly oriented towards gains and challenges for firms operating within the competitive commercial sector. There is less insight into what public sector actors experience [2] and in particularly little about applications within an audit and control context. Public sector organizations operate in a different institutional context, within different political and regulatory governance arrangements. Some regulations are compulsory for all organizations, such as the General Data Protection Regulation[3] and Equality and Anti-Discrimination Act[4]. Public organizations in Norway must also comply with the Public Administration Act, Freedom of Information Act, and the Archives Act[5], which all have significant implications both for the regular case

[2] Original name: Helseøkonomiforvaltningen, established in 2004 under Rikstrygdeverket, from 2006 NAV Helsetjenesteforvaltning, since 2009 an agency under the Directorate of Health

[3] Lov om behandling av personopplysninger (personopplysningsloven), generell personvernforordning. https://lovdata.no/dokument/NL/lov/2018-06-15-38/gdpr#gdpr

[4] Act relating to equality and a prohibition against discrimination (Equality and Anti-Discrimination Act) https://lovdata.no/dokument/NL/lov/2017-06-16-51

[5] Act relating to procedure in cases concerning the public administration (Public Administration Act) https://lovdata.no/dokument/NLE/lov/1967-02-10 Act relating to the right of access to documents held by public authorities and public undertakings (Freedom of Information Act) https://lovdata.no/dokument/NLE/lov/2006-05-19-16 Archives Act/Lov om arkiv (arkivlova) https://lovdata.no/dokument/NL/lov/1992-12-04-126

handling processes and on the IT systems in use. Because public sector organizations are in a unique position to for instance require submission of data from citizens, they are also held to high standards with regards to the usage of these data. Moreover, innovation and efficiency are not rewarded through increased market share and profit as in commercial companies, rather the public organizations' budgets are subject to yearly revisions from decision makers within both the administrative and political systems.

To exemplify the importance of audits for good governance, let us consider healthcare expenditure that constitutes 11.3 % of Norway's GDP (2020). The "fee for service" system allows healthcare providers to be reimbursed for incurred costs after patient consultations. This system covers general practitioners (family doctors), private specialists, dentists, physical therapists and others. Helfo processes the reimbursement claims and distributes refunds in an automized and trust-based system. Helfo annually receives over 1 million reimbursement claims and pays out 42 billion NOK (2021). Helfo has implemented automatic controls of incoming reimbursement claims, and in 2020 these controls stopped erroneous claims worth 800 million NOK. Any claim that passes the automatic controls will be reimbursed automatically within a few days. However, also manual post-payment audits are performed. As resources are scarce, the number of such audits is low, with only 0.14% of practitioners being audited annually. As emphasized in a recent report by the Auditor General[6], these manual audits are focused on the perceived most serious cases of erroneous claims. This can for instance take the form of a health care provider (e.g., a medical doctor) reporting procedures which have not actually been provided, selecting a more profitable procedure code than the correct one, or charging separately for activities that are part of one procedure.

In an effort to make these audits more efficient and enable more post-payments audits, Helfo seeks to leverage recent advantages in business intelligence and data analytics. These technologies may have potential impact on several of the stages of the audit process: Data processing can be rationalized using automation technologies such as Robotic Process Automation (RPA). An RPA bot can for instance prepare data for audits, copy and paste data, annotate data, organize files, integrate data from multiple files, and run basic audit tests [3]. While traditional audits may rely on statistical sampling of evidence, employing machine learning allows larger data sets (e.g. whole data sets or document archives) to be scanned and analyzed for trends and anomalies, thus expanding the data base for audits. Machine learning also offers the promise of more pro-active audits, based on predictions, pattern analysis and use of non-traditional data sources [4]. Several strategies for detecting erroneous transactions exist. Individual transactions can be flagged if they deviate too much from the expected value(s). Also, risk scores can be computed for the transactions based on classification or regression models, although this depends on the auditors having sufficient information about the underlying incentives and opportunities for errors [5]. Data visualization can enhance the usage of information and its value for decision support. Together, these technologies thus offer promises of enabling a more comprehensive audit model where risk assessments are based on larger data sets instead of samples. It also enables moving

[6]https://www.riksrevisjonen.no/globalassets/rapporter/NO-2022-2023/helsedirektoratetsetterkontrollhelserefusjoner.pdf

towards near-continuous auditing, where assessments are updated on an ongoing basis instead of regular, scheduled audits as before, realizing long-standing promises of "continuous auditing" [6].

However, in order to realize this innovation potential, substantial changes and updates are required, both in information infrastructure and in the organizational structure and workflow within the organization and towards the system owner. We wish to highlight the changes this implementation process entailed for the agency, with a focus on the requirements for building the required information infrastructure. Our perspective is sociotechnical, addressing the interplay between social/organizational aspects and technical aspects. The study's primary aim is to provide a systematic description of the experiences of such an implementation process for the benefit of other similar organizations.

In the next section we present relevant research and also our own conceptual basis for the study – the CRISP-DM process model. In section three we describe the case and the data collection and analysis process. Then we describe the implementation process in section four, before we in section five discuss the findings and generate an enriched process model adapted to the specific nature of a public sector organization.

2 Related research and conceptual basis

Business Intelligence and Analytics (BI&A) is often used as an umbrella term that cover "the techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions" [7, p. 1166]. To implement business analytics requires far more than purchasing and implementing a well-defined technology. It implies a deep and long-term change process, involving both organizational and technical aspects. A socio-technical perspective that assumes that technology and organization is deeply intertwined, is therefore our point of departure. In the following sections, we first describe relevant research that offer insights into this socio-technical change process and then our conceptual framework.

Information Systems researchers have for some time investigated the impact on organizations of big data, improved analytics capabilities and artificial intelligence. Some works provided high-level and general accounts of potentials and risks of big data analytics [8,9], while other discuss practical issues relating to data acquisition, application of analytic methods, and visualization of data [10]. In this paper, it is of particular interest to consider research that focus on the requirements this pose both to individuals and organizations. Successful implementation of business intelligence and data analytics requires building data analytics capability in the organization. Mikalef at al. [11] defines this as "the ability of a firm to capture and analyze data towards the generation of insights by effectively orchestrating and deploying its data, technology, and talent". Through a systematic literature review and based on a resource-based view of the firm, they identify three main categories of resources; "tangible resources (e.g. infrastructure, IS, and data), intangible resources (e.g. data-driven culture, governance, social IT/business alignment), and human skills and knowledge (e.g. data analytics

knowledge, and managerial skills)" (p. 561). In a mixed method study, researchers investigated in more depth how value creation, both strategic and operational emerged from big data analytics [12]. They found differing patterns in different operational contexts. Some of the challenges identified in the effort to leverage big data analytics were related to the lack of top management involvement in defining a strategic direction, the organizational inertia in implementing data-driven decision making, and the implications of ethics and legislation. While these studies identified central factors that enable productive use of data analytics, we wished to build on research that provided a process model, i.e. an overview and description of the sequence activities and stages in the implementation process.

We searched for a suitable process model and have chosen to employ the CRISP-DM model (Cross Industry Standard Process for Data Mining) as a conceptual structure for our empirical account. This model was developed in the 1990s in order to provide a shared process model for data mining projects [13] but is generic enough to be used more generally also today when the data analytics capabilities are more developed. The model provides an overview of the life cycle, with six phases: Business Understanding, Data Understanding, Data Preparation, Data Modelling, Evaluation, and Deployment (see figure 1). The progression from phase to phase is not unidirectional and linear but can go back and forth.

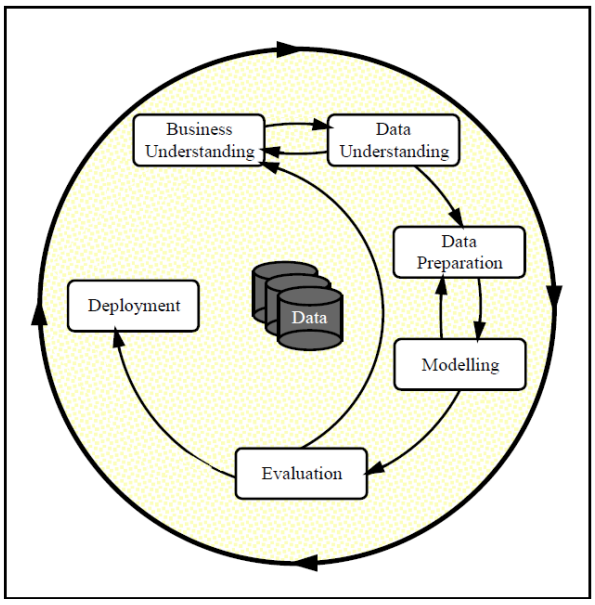

Figure 1. The phases of the DM implementation process (from [13])

The arrows indicate the most important and frequent dependencies between phases, but the sequence of the phases is not strict and the overall process is cyclical (as indicated by the outer arrow). For each phase, core generic and specialized tasks are indicated (see table 1)

Table 1. Core tasks in each phase (from [13])

Phase	Tasks
Business understanding	Determine Business Objectives, Assess Situation, Determine Data Mining Goals, Prepare Project Plan
Data Understanding	Collect Initial Data, Describe Data, Explore Data, Verify Data Quality
Data Preparation	Select Data, Clean Data, Construct Data, Integrate Data, Format Data
Modeling	Select Modeling Technique, Generate Test Design, Build Model, Assess Model
Evaluation	Evaluate Results, Review Process, Determine Next Steps
Deployment	Plan Deployment, Plan Monitoring and Maintenance, Produce Final Report, Review Project

While the model provides useful heuristic support at a general level, it also makes some assumptions that may not hold for our case. For instance, the underlying assumption is that a stand-alone firm will make its own decisions related to data it owns and controls. Several aspects are different in our case context: a public sector organization operates within a specific institutional and regulatory environment, and the data and digital infrastructures are not owned and controlled by the organization itself, but shared beyond the organization's boundaries. Our research question is thus: How does the implementation process look like in a public sector organization relying on shared data infrastructures? We aim to generate an enriched process model for public sector organizations that operate within societal domains and with shared data infrastructures.

3 Research method

We have conducted a qualitative case study of the early stages of a process of strengthening organizational data analytics capabilities. This was done in the context of an ongoing innovation project in Helfo[7]. Because there are few in-depth empirical studies of organizational implementation of data analytics in public sector and governmental agencies, we considered this to be a revelatory case study [14,15]. In the following we will provide background information about the case, the data collection and data analysis approach.

3.1 Case background

Helfo's responsibility includes controlling the reimbursements claims from health personnel. Submitted claims are screened for errors when they are submitted by an automated validation engine in which more than 2000 rules are embedded. If the claim passes these checks, payment is issued in an automized manner within days. In addition, Helfo conducts manual risk based post-payment audits. These audits are resource intensive as they require much manual work by highly specialized staff and may involve

[7] Project name: «Lærende kontrollvirksomhet for å sikre riktig refusjon fra helserefusjonsordningene». Funding from the Norwegian Research Council, NRC project number 321044

lengthy legal processes. While the post-payment audits are successful in that they uncover actual irregularities, it is also well known that these measures (submission checks and post-payment audits) are not sufficient to detect all errors.

A digitalization project called EDiT was run between 2018-2021[8], and ensured that the information infrastructure was upgraded across several domains for Helfo. It was followed by an innovation project that was initiated in 2021 with the aim to utilize digital technology and the already available data better, in order to improve the audit work by building on the foundations from EDiT. At the time when the project was designed, there were at least three ways in which this was expected to help: a) implementing data-driven decision support in selecting candidates for post-payment audits would imbue a more systematic approach to the current process that was based on manual information processing, b) the audit work itself could be supported by improved data availability, such as easy access to historical data or comparative data in a specific case, and c) task automation and better decision support might speed up the audit process, allowing more frequent audits and the possibility to issue reactions closer in time to when the error occurred.

The project plan and funding application was developed in 2020, and the project formally started on March 1, 2021, with funding from the Norwegian Research Council. The project owner was Helfo, and while the project was administratively located in the Audit department, it enjoyed strong support from top management.

3.2 Data collection process

The authors of the current study participate in the innovation project, author 1 as researcher, author 2 as a work package leader, and author 3 as academic project leader. As such, we have access to a wealth of information and insights about the implementation process, both documented and undocumented. Such a strong involvement will necessarily create insider bias. However, because the paper does not aim to provide any evaluation of the implementation process or project work, but rather to give a more factual account of the experiences, steps, challenges, and solutions during the implementation process, we consider the risk to be lessened. Still, the account in this paper is primarily based on 14 formal interviews with organizational members conducted by the researcher, who is not a member of the organization. In addition, project documents were consulted.

3.3 Data analysis

Our analysis strategy is mainly inductive. We first created a chronological timeline based on project documents. The interviews were recorded and transcribed, and in the first step of analysis the information provided in the interviews were related to the temporal evolution of activities in the project. In the second step of analysis, as a way to structure the empirical account, the information from interviews was related to the

[8] EDiT – Enklere Digitale Tjenester: https://www.helfo.no/om-helfo/digitale-tjenester-fra-helfo/enklere-digitale-tjenester-edit--raskt-enkelt-og-riktig

phases of the conceptual model presented in section 2. For each phase, information provided by the informants was combined, and the narrative presented in the next section constructed. Our account focuses on actual actions, decisions and events, rather than on the informants' individual perceptions, expectations and attitudes (in other words, we do not aim for an interpretative analysis), thus it does not include verbatim quotes from the interviews. In the next section we present the outcome of our analysis as a narrative that progresses through the different phases (ref. figure 1).

4 Findings from Case Study

In the following description, we mainly emphasize the work done within the ongoing innovation project, organized according to the phases of figure 1.

4.1 Business Understanding: do we know what we need?

The work of articulating the concrete organizational needs had mainly been done before the innovation project started in 2021. Thus, the overall aim of the project was already well defined – to strengthen the quality and capacity of audits within the existing resource limitations. However, effort was still required to make the project objectives shared across the organization as a whole and in general to ensure "organizational anchoring", as well as finetuning the concrete work packages as needs were further understood and articulated as the project moved forward. A central aspect of this related to clarifying the innovation project's role vis a vis other, already ongoing improvement projects in the organization. In particular, it could be seen as rather overlapping with an ongoing project to address and improve compliance. Also, it was very much in line with a long-term and broad initiative to implement quality management. The intentions to work based on data, increase traceability, and improve the process of selecting audit candidates were seen as contributing to these larger projects. The innovation project was connected to the compliance project through becoming an early 'use case' in which the compliance work could be operationalized. In addition, the necessary legal considerations regarding novel data use (more developed in the next section), necessitated alignment with ongoing efforts to determine implications of the GDPR regulations. The fact that the activities were organized as a formal project with deadlines and deliverables was essential for ensuring the required commitment and resource allocations.

4.2 Data Understanding and Data Preparation

Much of the required data were already available in the national data infrastructure. Two of the most central databases were owned by the Directorate of Health, and the data quality is generally good. During the EdiT project a data warehouse and an analytics sandbox (with RStudio and PowerBI functionality) was implemented in the National Health Network infrastructure. The decisions and technical actions required to achieve this necessitated collaboration with the external actors. This enabled easier access to data and software tools. The data warehouse contained a subset of the data,

which allowed swifter analysis. Previously, analysis was done based on importing data from precompiled reports into Excel, doing manual data merging across spreadsheets whenever information was scattered across various reports. This was time-consuming and could be error-prone. The new analysis facility was potentially more reliable and quicker. This enabled more flexible use and a potential increase in capacity to run ad hoc analyses.

Acquiring technical access to the necessary data was not sufficient. Also the legal basis for the planned work had to be secured. One work package in the project addressed the legal considerations and questions around applying additional data analytics and machine learning methods to the data. Questions that had to be resolved were e.g.: is this data use within the legal scope, i.e. is the purpose for which the data was originally collected covering also this usage? Will the combination of various data sources into one more comprehensive data model for decision support be violating privacy rights? Achieving the necessary conclusion on these and related issues was also time-consuming, because the legal personnel in the involved organizations (Helfo and the Directorate of Health) were in high demand. In 2021, a Data Protection Impact Assessment was finalized and the actual work with the data could start.

The two databases contained the reimbursement claims, which were the main data element for the analysis. In addition, relevant data existed in other of Helfo's own internal systems. For instance, historical data from previous audits was stored as reports in the archive system. However, these were differently formatted and stored in different structures. In order to become a valuable data source for a data-driven organization, the registration practices of the various case handlers would have to be standardized.

4.3 Modeling: Which questions to ask?

The pre-existing audit process relied on manual processes, but also employed some analytics. For instance, to support the selection of the most appropriate candidate for audit, a benchmarking strategy was used. The total amount of reimbursement over a year was calculated for each provider, and the providers with highest claims were flagged. This was combined with other information, e.g. tips coming in from the public or other health providers, into so-called "risk lists" (in an Excel sheet format). These risk lists were an important component in the decisions (made by management) on whom to select for a formal audit.

A large part of analytics resources in the project has been spent developing scripts, pre-calculating time series of various risk indicators in the data warehouse, doing quality assurance, and expanding these risk indicators to fields of health care providers which have received less audit attention in the last years. This enables speedy access to data analytics in order to support the decision-making process when deciding which health care providers should be audited. For areas that had not been given priority for audits in the last few years, e.g. reimbursements of drug claims, it was also necessary for the data analyst to spend time to "get a sense of the data", meaning what data elements the database contained, what they meant, and in general, assessing what kinds of analysis would be feasible and useful.

Developing data models to support risk assessment required an understanding of what the problem was and what to look for. With limited historical information on the actual amount and nature of errors, the experience-based insights from employees such as case handlers, personnel at the help desk, and investigators was considered relevant. In the early phase of the project, several workshops were organized, each focusing on a specific domain (medical doctors, dentists, drugs, laboratories/imaging/outpatient clinics, physiotherapy). In these workshops the project team elicited knowledge on typical error patterns and on what the employees considered indicators of these. This information, combined with the archival data of audit reports and previously known risk indicators, generated a number of non-compliance scenarios that informed the construction of risk indicators to be semi-automated and precalculated.

4.4 Evaluation and deployment

In general, there is a lack of information about the "ground truth" about non-compliance. This is the case for all the domains that Helfo cover, not just the fields with a shorter history of audits. Not only is there not sufficient historical data available, also the relatively frequent changes in the reimbursement and other regulatory rules implied that historical data might not be relevant to learn from. This meant that supervised machine learning methods that are trained on labelled data (i.e. need to be fed linked input-outcome data sets) to predict non-compliance could not be implemented immediately. The work therefore started out with semi-automating and expanding previously manual processes for transformation and calculation of risk indicators, which would primarily be used to detect outliers in some relevant dimension. Quality assurance of calculations and the underlying data used significant analytical resources. The selection of which dimensions to include required insight into the empirical domain. In order to know if results are significant, meaningful and usable, dialogue with case handlers and others was required.

4.5 Future developments

We have described some of the challenges of rolling out the required data infrastructure for data analytics. This is only the initial steps of an ongoing process of strengthening the capabilities for data analytics in Helfo. The project will run until 2024, but the perceived gain to the organization is so prominent that it has been decided that the process is to continue beyond the project period and become a part of ordinary operations in Helfo.

In order to harvest the full potential of data analytics, the models will need further development, the organizational culture and practices need to accommodate and utilize the potential for data-driven support to the existing processes, and the underlying data resources need to be extended for the full potential of machine learning methods to be realized. Thus, the deployment will not be characterized by a roll-out of a finalized product that is put into production. Continuous change is to be expected: health providers reporting practices will adapt to changes over time in legislation, in reimbursement coverage, in billing technologies and in medical procedures. Also, errors can be

expected to change along with increased information, opportunities for deliberately misreporting, and perceived detection risks. Ongoing learning and adaptive capacity are therefore also required. Today, the audit activities happen with a significant delay, and mainly focuses on past actions, and may thus have limited effect on preventing future errors. A future ideal scenario would have tighter coupling with the health providers' information systems, to be able to indicate errors or issue warnings as the information is being recorded. Such a continuous and near real-time feedback is the 'holy grail' of the 'continuous auditing visions [6].

5 Discussion: Implementing data analytics in the public sector

Our aim is to provide insights into the preparatory work required for developing data analytics capabilities in a public sector organization. Much research concerns companies in the private sector, and a public sector organization will operate in different conditions on many respects. Core differences are related to the data itself: while companies can collect and use their own data, public sector organizations often rely on using data that comes from public, shared infrastructures. The public infrastructures often have a well-defined mandate for collecting the data from citizens, therefore data are not legally available for any type of use. In our case, this necessitated collaboration with other entities both to ensure data access and to conduct the necessary structural changes in data flow. The organizational relations among actors in the public sector are more complex that the competitive and collaborative relations in a market system. Finally, while private companies usually can prioritize economic rationality and profit making, public sector organizations balance multiple goals, related to both economic, political, administrative, and regulatory contextual factors.

We therefore seek to extend the generic process model presented in Section 2 with additional tasks and challenges that we encountered in our case. In Table 2, we add a third column to the table of core activities, where we account for tasks that our case revealed as significant concerns during the initial stages of organizational implementation of data analytics capabilities.

While the original process model operates from assumptions that the process occurs within company boundaries and mainly involve data experts, our case shows that several external stakeholders are involved and that other internal stakeholders are also crucially important. Also, the work to ensure both technical and legal access to data is another major difference. This is a highly important step which is often under-communicated in much of the literature that emphasizes the benefits of data analytics, but has attracted focus and attention in research more attentive to actual practice on the ground [see e.g. 16,17].

We hope to have illustrated the significance of preparatory work to build a suitable data infrastructure, as well as given a sense of the need to continue with building organizational capacity. While we have emphasized the different operating conditions of public sector organizations in terms of legal and political governance arrangements, we do think that the experiences we account for here can be found also in private sector

organizations, especially if they do not own and control the data infrastructures themselves.

Table 2. Extending the process model

Phase	Tasks	Additional tasks
Business Understanding	Determine Business Objectives Assess Situation Determine Data Mining Goals Prepare Project Plan	Work with organizational anchoring: • communicate the added value and relevance of the project • locate the activities within the organizational structure • make productive links with ongoing projects
Data Understanding	Collect Initial Data Describe Data Explore Data Verify Data Quality	Ensure access to data: • Establish legal basis for accessing, merging, and using data • Negotiate technical access to data and data infrastructure • Build technical infrastructure for analysis
Data Preparation	Select Data Clean Data Construct Data Integrate Data Format Data	Improve and strengthen data resource • Ensure sufficient infrastructural data • Improve data registration practices • Work on how to integrate various formats and sources
Modeling	Select Modeling Technique Generate Test Design Build Model Assess Model	Find relevant questions: • Enroll domain experts to select analysis dimensions and model features • Analyse data to detect testing possibilities
Evaluation	Evaluate Results Review Process Determine Next Steps	Ensure that ambitions matched preconditions • Build domain insights on insufficient data • Include feedback from domain experts
Deployment	Plan Deployment Plan Monitoring and Maintenance Produce Final Report Review Project	Build long-term learning capacity: • Learn from operations • Enable adaptation of data, tools and work processes

6 Conclusion

We have reported from a qualitative, in-depth case study that covered the early stages of implementing data analytics capabilities in a public sector organization. We see that the efforts to establish a sufficient information infrastructure required much effort in the early stages, which are often glossed over in both research accounts and consultant reports.

While there is certainly a large initial cost to establishing the required information infrastructure long before any advanced artificial intelligence tools can be utilized, we also see that the initial steps have provided positive gains and increased the operational efficiency of the organization. Analyses can be done with less time and resource costs, and the audit work can be better supported. These are important stepping stones along the way to also build sufficient organizational commitment to continue and explore the potential of more advanced analytic technologies.

Building capacity to utilize new technological advancements in analytics will also imply capacity for continuous learning, for instance through well-planned field experiments. In Helfo's context, the next steps of the project is to design and implement experiments to produce new insights in how to achieve better compliance as well as information about the prevalence of errors and the total reimbursement gap. The plan is to implement such experiments in an adaptive fashion, such that effects may be evaluated along the way. This enables continuously updating of experimental methodology based on innovation and insights realized in the frames of the project. A fundamental underlying condition for all this is the need for continued organizational development, structural changes in the IT systems, and thorough registration of experiment data and quality control.

The limitations of this study are on the one hand connected to our insider role. While this is a strength in that it offers access to information and insight into the process, it may also be a limitation and an independent researcher might have seen other aspects that we are blind to. Also, our story only covers the early stage of the implementation efforts. More insights on how to build more advanced analytic capabilities will hopefully emerge with time. We hope that also the description and analysis of these early-phase experiences may be useful to other public sector organizations thar seek to strengthen their data analytic capabilities.

References

1. Tarafdar, M., Beath, C. M., & Ross, J. W. (2019). Using AI to enhance business operations. MIT Sloan Management Review, 60(4), 37-44.
2. Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. Government Information Quarterly, 36(2), 368-383.
3. Cohen, M., & Rozario, A. (2019). Exploring the use of robotic process automation (RPA) in substantive audit procedures. The CPA Journal, 89(7), 49-53.
4. Dickey, G., Blanke, S., & Seaton, L. (2019). Machine Learning in Auditing. The CPA Journal, 89(6), 16-21.
5. Ekin, T., Ieva, F., Ruggeri, F., & Soyer, R. (2018). Statistical medical fraud assessment: exposition to an emerging field. International Statistical Review, 86(3), 379-402.
6. Institute of Internal Auditors (IIA), 2012. https://www.theiia.org/en/content/guidance/recommended/supplemental/gtags/gtag-continuous-auditing/
7. Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 1165-1188.
8. Constantiou, I. D., & Kallinikos, J. (2015). New games, new rules: big data and the changing context of strategy. Journal of Information Technology, 30(1), 44-57.

9. Woerner, S. L., & Wixom, B. H. (2015). Big data: extending the business strategy toolbox. Journal of Information Technology, 30(1), 60-62.
10. George, G., Osinga, E.C., Lavie,D. and Scott, B.A. (2016): From the editors: Big data and data science methods for management research The Academy of Management Journal, Vol. 59, No. 5 (October 2016), pp. 1493-1507
11. Mikalef, O., I.O. Pappas, J. Krogstie, M. Giannakos: Big data analytics capabilities: A systematic literature review and research agenda. Information Systems and e-Business Management, 16 (2018), pp. 1-32).
12. Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big data analytics and firm performance: Findings from a mixed-method approach. Journal of Business Research, 98, 261-276.
13. Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1, pp. 29-39).
14. Yin, R. K. (2009). Case study research: Design and methods (Vol. 5). sage.
15. Sarker, S., Sarker, S., Sahaym, A., & Bjørn-Andersen, N. (2012). Exploring value cocreation in relationships between an ERP vendor and its partners: a revelatory case study. MIS quarterly, 317-338.
16. Passi, S., & Jackson, S. J. (2018). Trust in data science: Collaboration, translation, and accountability in corporate data science projects. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 1-28.
17. Parmiggiani, E., Østerlie, T., & Almklov, P. G. (2022). In the Backrooms of Data Science. Journal of the Association for Information Systems, 23(1), 139-164.