Journal of Big Data

RESEARCH Open Access

DAPS diagrams for defining Data Science projects

Jeroen de Mast[1]* and Joran Lokkerbol[2]

*Correspondence:
Jeroen de Mast
j.demast@uva.nl
[1] Amsterdam Business School, University of Amsterdam, Plantage Muidergracht 12, Amsterdam 1018 TV, Netherlands
[2] Centre of Machine Learning, Netherlands Institute of Mental Health and Addiction, Utrecht, Netherlands

Abstract

Background Models for structuring big-data and data-analytics projects typically start with a definition of the project's goals and the business value they are expected to create. The literature identifies proper project definition as crucial for a project's success, and also recognizes that the translation of business objectives into data-analytic problems is a difficult task. Unfortunately, common project structures, such as CRISP-DM, provide little guidance for this crucial stage when compared to subsequent project stages such as data preparation and modeling.

Contribution This paper contributes structure to the project-definition stage of data-analytic projects by proposing the Data-Analytic Problem Structure (DAPS). The diagrammatic technique facilitates the collaborative development of a consistent and precise definition of a data-analytic problem, and the articulation of how it contributes to the organization's goals. In addition, the technique helps to identify important assumptions, and to break down large ambitions in manageable subprojects.

Methods The semi-formal specification technique took other models for problem structuring — common in fields such as operations research and business analytics — as a point of departure. The proposed technique was applied in 47 real data-analytic projects and refined based on the results, following a design-science approach.

Keywords Data mining, Machine learning, Big data analytics, Problem definition, Project management, CRISP-DM

Introduction

Data and analytics projects come in a wide variety [1, 2], including projects focused on data management or IT infrastructure and projects that explore the potential value in data [3]. An important class of projects are those that aim to develop a predictive, prescriptive or causal (diagnostic) analytics model [4–6]. Such projects deploy machine learning and AI to realize certain business goals, through, for example, improved decision support, forecasting for capacity planning or an early-warning system for detecting irregularities. The traditional model for structuring such data-analytics projects is CRISP-DM [7], and Martínez-Plumed et al. [3] observe that twenty years after its introduction, this is still the de-facto standard for such projects. Other project models include

Microsoft's Team Data Science Process (TDSP [8]) and IBM's Foundational Methodology for Data Science (FMDS [9]).

Key factors in the success of data-analytics projects are a proper definition and scoping of the data-analytic problem that the project will tackle (see, e.g., [8, 10–12]). This task is often described as difficult and challenging (e.g., [13, 14]). Specifying the business objectives of a project, and translating them into a data-analytic question, is done in CRISP-DM's Business Understanding stage (or, alternatively, in the stages of the same name in either TDSP or FMDS). The task requires both technical and domain knowledge: technical understanding of what machine learning and data analytics can and cannot do, and domain understanding of the goals and strategies of the business and the processes and customers possibly affected by the project. It is often suggested that a more senior data scientist should be responsible, in collaboration with a business expert. The translation of business goals to data-analytic questions is also the crux of the emerging role of analytics translators [15].

Data-analytic projects try to create value by translating business objectives into questions about Y (dependent) and X (independent) variables and modelling how they are related [4, 16]. CRISP-DM calls this the 'data-mining goals,' and we call it the data-analytic problem. In predictive analytics, such as classification or supervised learning, the crux is a model of the form $Y = f(X)$ that predicts the dependent from the independent variables. Prescriptive analytics use $Y = f(X)$ models to solve optimization problems. In diagnostic analytics, $Y = f(X)$ equations model cause-and-effect relations, allowing one to analyze what would happen if one intervened in the X variables. In descriptive and exploratory projects, emphasis is often on the discovery of potentially relevant Y and X variables, or describing their current distribution [4]. The central end term of CRISP-DM's Business Understanding stage is the definition of one or a few data-analytic questions about Y and X variables and their relations, and an explanation of how these questions are expected to contribute to the business goals. This is the link to the ensuing stages, where these Y and X variables are connected to data sources, and the relationship between them can be modelled.

Compared to the other stages in CRISP-DM, the task of translating a business question into a data-analytic question is relatively unstructured, lacking a clear, concrete approach for practitioners to follow. Our contribution in this paper is a technique that we call the Data-Analytic Problem Structure, or DAPS. It is a graphical technique that offers structure to the Business Understanding stage in CRISP-DM and to similar stages in alternative project models such as TDSP and FMDS. We propose DAPS as a semi-formal specification technique that offers guidance to data scientists and analytics translators for scoping and defining the goals of data-analytic projects. The next section presents the function of DAPS and describes its development and validation. Next, we provide a detailed description of the components of DAPS and how they should be used. Further, we describe how DAPS can be used to facilitate project scoping. We describe two real examples showcasing the use of DAPS. Finally, the Conclusions Section summarizes the key arguments for introducing and adopting DAPS.

Function of DAPS and its development

This section motivates why a novel technique is needed and what its function is, and then explains how we developed and validated DAPS.

The need for and purpose of DAPS

Currently available support for defining data-analytics projects consists largely of process models for an entire project. Besides the aforementioned CRISP-DM, TDSP and FMDS, other models that got substantial coverage in the literature include SEMMA, D3M and ASUM-DM (see Kurgan and Musilek [2] for an overview). Some of these, such as SEMMA [17], focus on the technical aspects of data-analytics projects, and offer little support for translating business goals to data-analytic questions. Most models offer a breakdown of the problem definition task into specific deliverables. CRISP-DM, for example, lists the following deliverables for its Business Understanding stage: Determine business objectives, Assess situation, Determine data mining goals, and Produce project plan [7], where each of these deliverables is further broken down into more detailed subtasks. The Snail Shell model of Li et al. [18] breaks down its Problem Formulation stage into a number of tasks, including: determine business objectives and success measures, determine boundaries, factor complex problems into sub-problems, and determine the data-analytical problem (type and goal).

From these process models, we can learn the deliverables of the Business Understanding stage of data-analytics projects (*what* to do?), and we derive the functions of our DAPS technique as summarized in Table 1. Note that DAPS focuses on problem definition. Typical other deliverables listed under Business Understanding, in particular project-management tasks such as resource planning and time planning, are beyond the function of DAPS.

Where these process models fall short, is in providing operational support in structuring the tasks listed in Table 1 (*how* to do it?). Given that the translation of business goals into well-structured and well-scoped data-analytics problems is recognized as a difficult task, support of a more operational nature is needed. Some process models suggest generic problem-definition techniques such as SMART [19], which is abstract and unspecific in the support it offers, and not geared to the particular structure of data-analytic problems. The Machine Learning Canvas [20, 21] offers a structure for defining and motivating a machine-learning project, including how an algorithm could be evaluated and maintained. The definition of the data-analytic problem, however, is not its primary focus, and it is too cursory to guide the ensuing Data Understanding and Modelling stages. The canvas visualization only loosely shows how elements of the data-analytic problem are related, but not as clearly as graph diagrams used in the problem-structuring literature, which we discuss below. The technique has only limited underpinning in the academic literature. Also the Analytics Canvas [6] does not reveal the structure of a data-analytic problem, and instead, is a format for describing use cases and the necessary data infrastructure.

Table 1 Functions of DAPS as derived from literature

Function	Supporting literature
Support the translation of business goals into a data-analytic problem (that is, into questions about Y and X variables)	[7–9, 18, 22, 23]
Help to define the anticipated business value, by articulating how the model is going to be used, and what value that is expected to bring	[7, 18, 22, 24]
Help to define the type of model that is needed	[2, 3, 6, 9, 18]
Support the breaking down of a long-term ambition into manageable chunks, and thus scope the project	[18, 25]

What is missing in the data-analytics literature, is a technique for defining the structure of data-analytics problems and their relationship to business value, using a visualization effective in showing problem structure, and geared to the structure of data-analytic problems. The approach should be precise enough to guide the ensuing stages of Data Understanding and Modelling, and its usefulness in its application context should be properly evaluated.

Developing DAPS as a problem-structuring device

Taking the criteria in Table 1 to represent the functions of DAPS, we describe below how we developed the technique and then how we evaluated and improved it. Instead of the literature discussed above, we found a more useful point of departure in the literature on problem structuring in the fields of operations research, business analytics and related fields.

Goals of data science projects are often not presented to an analytics team in a well-defined, clearly structured form [18]. Often, teams are initially given what's called a *mess*: a complex tangle of interrelated and interdependent issues, possibly involving subjective perceptions and incongruent goals [26, 27]. The field of operations research uses the phrase *problem structuring* for the task of structuring a mess into one or a few problems with clear objectives and clear constraints [28, 29].

The point of departure for developing DAPS were a number of diagrammatic problem-structuring techniques known in other fields, including:

- Cognitive maps in operations research and management science [29].
- The Current Reality Tree in the Theory of Constraints [30].
- The CTQ Flowdown in Six Sigma [31].
- Block diagrams in System Dynamics [32].
- The Goal Question Metric approach in software engineering [33].

Such diagrams frame the structure of a problem by means of directed graphs, where arrows represent relations between issues and factors. As a starting point to the development of DAPS, we adjusted such diagrams to the particular structure of data-analytic problems as described in De Mast et al. [4].

At the heart of data-analytic approaches is the modelling of relationships among X and Y variables based on data [34]. These models could be causal, correlational or deductive models [4], and for new observations they predict Y values (alternatively called outcomes, dependent variables, responses, or labels) based on the given X values (called features, independent variables, predictors, or explanatory variables). In DAPS, the definition of the data-analytic problem as a number of questions about relations between X's and Y's is captured in the lower part of the diagram (see Fig. 1).

The upper part of the diagram shows the business value of the project, by explicating to what performance indicators the application of a data-analytic model is expected to contribute and how that is envisioned to further the organization's goals. The data-analytic problem in the lower part and the intended business value in the upper part are connected by the decision framework in the middle, indicating how the model is going to be used. The decision framework specifies what decisions or actions are driven by the Y values that the model predicts. Better support for these decisions or actions should

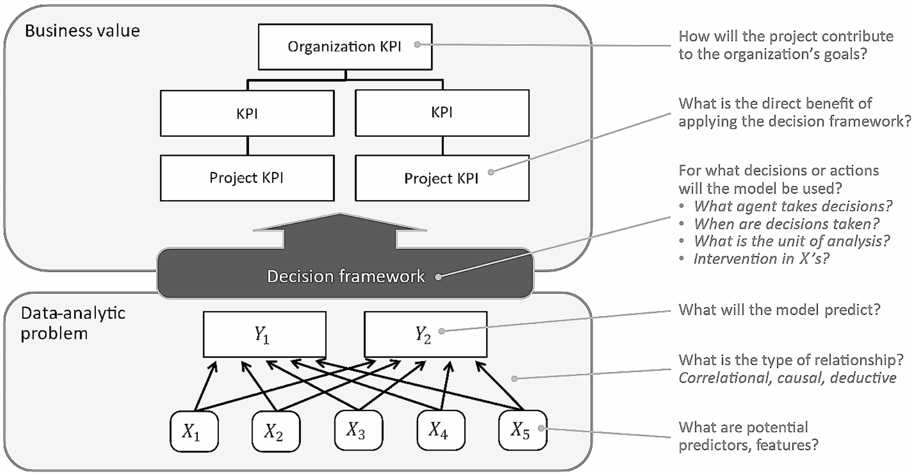
Business value
Organization KPI
KPI
KPI
Project KPI
Project KPI
Decision framework
Data-analytic problem
Y_1
Y_2
X_1
X_2
X_3
X_4
X_5
How will the project contribute to the organization's goals?
What is the direct benefit of applying the decision framework?
For what decisions or actions will the model be used?
• *What agent takes decisions?*
• *When are decisions taken?*
• *What is the unit of analysis?*
• *Intervention in X's?*
What will the model predict?
What is the type of relationship? *Correlational, causal, deductive*
What are potential predictors, features?

Fig. 1 Data-Analytic Problem Structure (DAPS)

bring the intended improvement in the performance indicators specified in the upper part.

Structuring a project's goals and translating them into a problem that is precise enough to be addressed by analytical techniques have the character of a search process, often involving multiple stakeholders representing multiple perspectives. Diagrammatic representations such as DAPS serve the purpose of helping individuals and groups to articulate their views, identify discrepancies and missing information, and collectively arrive at a shared understanding of the problem to be solved [29, 35]. This is usually not a linear process, where a group starts to build the diagram from the top and elaborates it towards the bottom. Rather, the process is iterative, with much going to and fro. Even after completion of the Business Understanding stage, in the later stages of Data Understanding, Modelling and elsewhere, it is common that advancing insights make it necessary to iterate back to the problem definition [18]. DAPS facilitates this iterative process, documenting insights as they evolve and converge, and at any moment presenting the current views as an invitation to new experts to comment on them and improve them from their points of view, until there is an internal consistency to the different components of DAPS, making it clear how pursuing the data-analytic problem helps to improve the organization's objectives.

Research strategy

We identify the development and refinement of DAPS into a problem-structuring device as research in the paradigm of design science [36, 37]. The goal of design science research is the design of an artefact, such as a model or a method, that meets functional performance requirements as determined by the application context or environment [38, 39]. Following Hevner [40], design science research follows three processes: the Relevance Cycle, Rigor Cycle and Design Cycle. The Relevance Cycle establishes the practical value of the artefact, by identifying the application domain and an important problem in need of a solution. Above, we have described the need for a technique such as DAPS in defining data-analytics projects, with Table 1 specifying what such technique should do. The Rigor Cycle grounds the artefact in established knowledge, methods, experience and theory in the application domain [39, 40]. Above, we described how theory on problem structuring and existing diagrammatic techniques were taken as a point of departure for developing DAPS.

The Design Cycle follows an iterative process of building a version of the artefact and then evaluating its performance in the application context [39, 40]. Hevner et al. [39] discuss evaluation methods in design science research. Due to the nontechnical nature of the technique that we develop and in view of the qualitative rather than quantitative nature of the requirements specified in Table 1, evaluation methods such as functional or structural testing, controlled experiments or architecture analysis (from Table 2 in [39]) are unsuited. Instead, we used an *observational* evaluation approach [39]. Subsequent versions of DAPS were applied in real data-analytics projects, which were done by students of professional and executive programs in data science offered at two distinct universities. A first version of DAPS was developed by taking the abovementioned existing problem-structuring techniques in other domains as a start, and the function specification in Table 1 as a goal, and established theory about the structure of data-analytic problems as a guide [4, 16, 34]. This version of the model was taught to a first group of students, who then applied it in the actual projects that they executed as part of their program (10 projects total). The projects were reviewed by the authors and discussed with the students, evaluating to what extent they gave a satisfactory fulfillment of the criteria in Table 1, and discussing with the students in how far DAPS had or had not offered effective support in achieving them. Each project was reviewed multiple times as it unfolded (typically four to eight times). Students applied DAPS in the Business Understanding phase of their projects, but the authors followed the projects also in the ensuing CRISP-DM phases to identify issues in the project definition that emerged later. The Appendix describes the criteria based on which the evaluations were done, as well as the practical setup of the evaluations.

On the basis of the findings in the first ten projects, the DAPS technique and its instructions were adjusted. The following shortcomings were observed and addressed in an improved version of DAPS:

- Students did not specify a specific action or decision that the data-analytic model's predictions were intended to support, and as a consequence, they failed to explicate how the data-analytic model was envisioned to deliver the anticipated business value.
- The definition of the Y characteristic that the model should predict was insufficiently precise (for example, in terms of prediction horizon or unit of analysis). As a consequence, the predictions might be ineffective in supporting the specified decisions or actions.
- Students did not specify whether their application required a causal $Y = f(X)$ model or a purely predictive (that is, correlational) model. This depends on whether the decisions or actions that the model is to support involve interventions in the X's (as explained later).

The improved version was again applied in a number of real projects (37 total), by next groups of students (Table 2 gives an overview). These projects were also reviewed by the authors, following the criteria and setup presented in the Appendix. This second round of applications did not reveal points for further improvement of the technique. Two of these projects, using the improved and final form of DAPS, are discussed in the sections below as illustrations of the technique.

Table 2 Overview of projects in which DAPS was tried and evaluated

Sector	Projects in 1st round	Later rounds
Industry	8	7
Energy and infrastructure		3
Finance		12
Healthcare		9
Government		2
Agrofood	1	
Retail & logistics		4
Sports and entertainment	1	
Total	10	37

The components of DAPS and how they should be used

This section explains the various parts of DAPS. The section then illustrates the technique by a real example.

Lower part of DAPS: the data-analytic problem

Following [4, 16, 34], data-analytic problems revolve around models that predict Y values from X values. The Y variable is the characteristic that the model predicts or explains or that we want to optimize. The model could be a correlational model, as in supervised learning. Such models can predict Y values of new observations provided that the new observations and the training data are representative samples from the same data-generating process [4]. If the new observations and the training data are not sampled from the same data-generating process, then the predictions may be biased. In particular, correlational models cannot predict the effects of interventions in the X's. Such predictions require a causal model, and the relations between X and Y are then one of cause-and-effect instead of merely correlational. Note that developing causal rather than correlational models requires more involved study designs, possibly involving randomized controlled experiments or structural causal modelling [41–43].

Even though DAPS is applied early in a data-analytic project, it is important to be specific in defining the Y variables that the model is to predict. This is different for the X's, however. Early in the project, it suffices to indicate the general sort of features that the team envisages can be used to predict Y, but identifying and sharpening the definition of the features is better done in later stages of CRISP-DM. Even if in the early stages the identification of the X's in the DAPS diagram is provisional and sketchy, we observed in the 47 projects that it is useful for a team to articulate the sort of X's they hope will be good predictors.

Some forms of analytics do not predict Y values from X values, but instead, merely characterize or visualize the current distribution of the Y values. Davenport and Harris [5] and Steenstrup et al. [44] refer to such applications as *descriptive analytics*. In such projects, the lower level of DAPS only lists Y variables, and no X variables.

Middle part of DAPS: decision framework

The middle part of DAPS specifies how the model is going to be used. Following [6, 44], analytics models are either used to support humans in taking decisions, or to directly steer actions. The decision framework in DAPS captures the decisions and actions to be supported by the model. The intended use of the model determines the type of model that is needed, the specific definition of the Y variable, and also, which X variables can be used. It is, therefore, important to be precise in defining the model's intended use. In the 47 projects, we observed that it is useful to be specific about the following elements of the decision framework.

1. Identify the agent that will use the model for taking decisions and actions.
2. Identify at what point in time (or at what frequency) decisions or actions are taken.
3. Specify the unit of analysis for predictions, decisions and actions.
4. Specify whether decisions or actions involve an intervention in the X's.

In the example at the end of this section ("Forecast-driven logistics"), the predicted characteristic is the volume of parcels to be handled (the Y variable). This is predicted for each of a number of distribution centers (DCs), so the unit of analysis (3) is a DC. The predicted volumes are used by a planning department of a logistics company for deciding on the number of bins to send to each DC, who are therefore the agent taking decisions (1). They apply a daily planning cycle, so the model should predict the daily parcel volumes with a forecast horizon of one day (2). The decisions taken by the planners do not involve an attempt to influence the parcel volumes by adjusting some of the X's (4), and therefore, a correlational model is suited (as opposed to a causal model). Being precise upfront in defining for what decisions or actions the model will be used, helps in sharpening the definition of the Y variable to be predicted, and in turn, it helps in identifying what features can and cannot be used in the model. In the example, the specifics of the decisions for which the model will be used determine that the Y variable is the daily volume of parcels in a DC, and the model can only use features whose values are available at the moment that the planners decide on the number of bins to be sent to the DCs.

Upper part of DAPS: business value

The upper part of DAPS specifies what goals the data science project pursues and what value that could bring to the business. Reading from bottom to top, the team has specified the data-analytic model they aim to develop, and they have explained how and for what decisions the model will be deployed. Then, in the upper part of DAPS, the project KPIs specify the performance indicators that are aimed to improve by deploying the data-analytic model. These are the indicators that are directly impacted by the project, and therefore represent the scope and direct goals of the individual project. Upwards, the DAPS diagram shows how these project KPIs are believed to contribute to the organization's goals or strategy by linking project KPIs to organization KPIs.

DAPS's upper part thus models the business value of a project by linking the impact on the level of the individual project to the impact that the project has on the level of the organization. This reflects the view that projects are more effective if they are aligned with strategic planning [45]. In strategic planning, organizations make choices as to where they should focus their efforts in order to remain valuable and competitive, and next, they translate these strategic focal points to programs and projects [46, 47]. Even in cases where the organization has no articulate strategy, explaining the business value of a project does require a project team to link the immediate benefits of their project to the general aims or mission of the organization as a whole.

Example 1: Forecast-driven logistics

One of the 37 projects using the final form of DAPS took place in a logistics company that collects parcels from the DCs of retailers, and delivers them to the end customers. Every day, they decide on the transportation capacity (number of bins) to be available on the next day at each DC. Overestimation of the required capacity results in losses due to overallocation of bins and truck runs to a DC, which can then not be deployed elsewhere and are instead underutilized. Likewise, underestimation of the required capacity results in relatively expensive 'settlement runs', where additional bins are transported to a DC at the last moment.

In the current way of working — in which capacity planning is based on expert opinion — underestimation and (severe) overestimation of the required capacity occur frequently, and the feeling is that much could be gained from a data-driven approach to estimating the required capacity at each DC. To come to a clearly defined and logically consistent data science project, the DAPS diagram in Fig. 2 was developed, indicating how the outcomes of the data-analytic approach contribute to the goals of the organization. The upper part of this DAPS took the current problems regarding under- or overestimation of required capacity as a starting point and indicated how less underestimation and less overestimation (project KPIs) would help company KPIs such as truck utilization and parcel costs. Achieving such reductions in under- and overestimation requires a model able to predict next day's volume of parcels in a retailer's DC. The bottom part of the diagram lists the preliminary ideas about some potentially relevant predictors, namely, week and seasonality patterns, trends and campaign information. Finally, the prediction generated by the model can be used to determine the required number of bins to be transported to a DC (middle part of the DAPS), taking into account the uncertainty in the predictions and strategic considerations regarding the desired balance between under- and overestimation.

The diagram in Fig. 2 presents the end result that emerged after about four sessions, where the project leader (himself a data scientist) discussed the project with domain experts and executives. The clarity and conciseness are a long way from the messy assignment that the team started out from. There was an intuitive understanding that better prediction

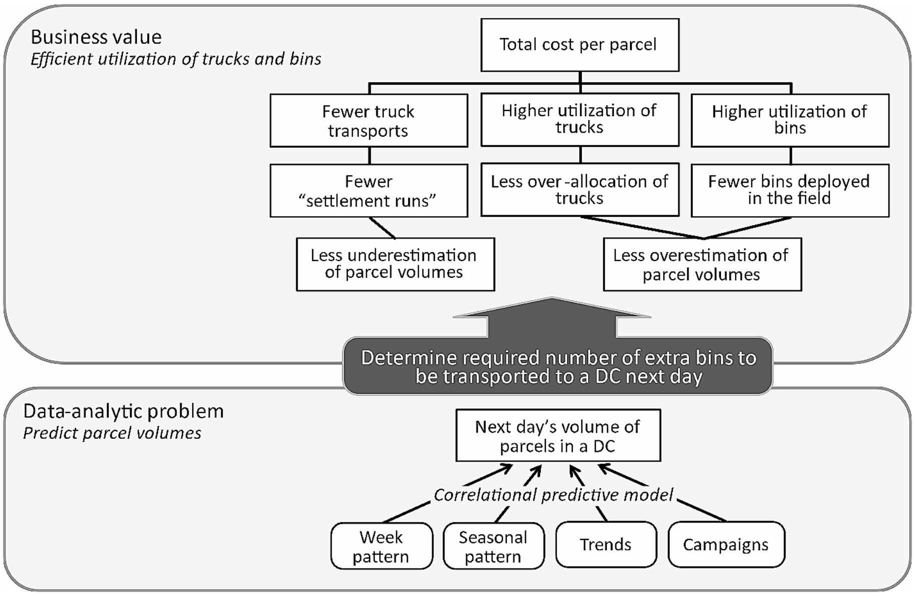
Business value
Efficient utilization of trucks and bins
Total cost per parcel
Fewer truck transports
Higher utilization of trucks
Higher utilization of bins
Fewer "settlement runs"
Less over-allocation of trucks
Fewer bins deployed in the field
Less underestimation of parcel volumes
Less overestimation of parcel volumes
Determine required number of extra bins to be transported to a DC next day
Data-analytic problem
Predict parcel volumes
Next day's volume of parcels in a DC
Correlational predictive model
Week pattern
Seasonal pattern
Trends
Campaigns

Fig. 2 Data-Analytic Problem Structure for Example 1: Forecast-driven logistics

of volumes would be a good thing, but it took a lot of thinking before the economic benefits were clearly understood. Also, it took a while before the team realized that they should address the problem on the level of a DC. This was very obvious to the logistics managers (who therefore didn't articulate it), but became apparent to the data scientist only after lengthy discussions. This illustrates DAPS as a communication technique that supports a team of different backgrounds in developing a shared understanding.

How DAPS is used to facilitate project scoping

Above, we discussed how a logistics company used DAPS to clarify the goals of a data-analytic project. The DAPS diagram (Fig. 2), however, sketched a long-term ambition, where the model is to predict parcel volumes for all retailers in the country, ranging from large international corporations with huge DCs to small niche companies with sporadic and irregular volumes. The prospect of developing a model able to handle this wide range of DCs, and accommodating all sorts of idiosyncrasies and special situations, was overwhelming. Adding to the team's hesitance to start were doubts about the feasibility of building a powerful predictive model in the first place.

To find a manageable scope to work on, the team divided the end goal up in chunks: a number of intermediate goals that could be attained in compact subprojects, each scoped to take 3 to 6 months' time. The first subproject focused on building proof-of-concept predictive models for one small, one medium and one large DC. If successful, a follow-up project could then focus on generalizing these models to be applicable in all DCs in the country. Breaking down a large ambition into manageable chunks is an acknowledged principle in project management, and DAPS was designed to facilitate this tactic (see the fourth function of DAPS in Table 1). In this section, we discuss the principle and we illustrate from a few examples how the DAPS idiom and structure support chunking.

Making projects manageable by chunking

The goals as laid down in a DAPS diagram may imply a large amount of work, too much to be done in a compact amount of time. The ambitions may also critically depend on unknowns: Are the data of sufficient quality? Can we predict Y with sufficient accuracy? Large-scale ambitions, possibly with substantial unknowns, are difficult to manage and difficult to translate to concrete actions, and this may lead to paralysis. Literature on project management acknowledges that long-term ambitions should be broken down in manageable, 'bite-sized' chunks. Thoms and Pinto [48] stated that it is a critical skill for a successful project leader to be able to break a project down into small manageable parts, which they defined as time chunking, "in order to maintain a focus on the final project result regardless of its current state of development". Clarke [49] and Lewis [50] identified breaking down a project into bite-sized chunks as critical to project success. Raz and Globerson [51] describe the Work Breakdown Structure as a principal tool for planning and controlling the work contents of a project, which through the hierarchical decomposition of a project into parts, can contribute to the probability of successful completion of a project.

By dividing the work laid out in DAPS into piecemeal, chunk-by-chunk subprojects, teams can spring into action and start on a first, concrete, manageable chunk. Along the way, as unknowns or assumptions become more clarified, the next chunk will become more concrete. Typical chunking patterns that the authors encountered in the 47 projects include:

- First building a proof of concept (subproject 1), then building a deployable model (subproject 2).
- First modeling the relation between the X's and Y for a subsample of the data (e.g., a limited number of sites, such as a small, a medium, and a large site) (subproject 1), then generalizing the results to the full population of interest (subproject 2).
- First creating an algorithm or infrastructure to measure outcomes or predictors, thus obtaining data on X or Y variables (subproject 1), then using these data to build a model (subproject 2).
- First building a model (subproject 1), then optimizing the decision framework given the reduced but remaining decision uncertainty (subproject 2).

Other examples of chunking in the 47 projects occurred when a project's goals implied more than one data-analytic problem. The project presented below ("Early-warning system in a power grid") involves three data-analytic problems:

- The core problem is to build a predictive model, which signals a problem called *asymmetry* in a power grid so that technicians can intervene and take measures to avoid nuisance for customers.
- This application is based on the assumption that nuisance experienced by customers is actually caused by asymmetry, which is however not undisputed. A secondary data-analytic problem is to test this assumption, and establish to what extent nuisance is caused by asymmetry.
- Asymmetry is not measured directly, and instead, is to be derived from raw sensor readings in the grid. The derivation of relevant characteristics from raw data is called feature engineering in data science, and the design of this algorithm that computes asymmetry values from available data was a premise for fitting the other two models.

Developing these three data-analytic models were the first three subprojects that the company undertook to realize the end goal of an effective early-warning system. We discuss the project below and demonstrate how DAPS was used to structure and visualize the project's strategy.

Example 2: early warning system in a power grid

Power is distributed and transported over longer distances in a three-phase system, where there are three 'hot' wires that are 120° out of phase with each other. For the grid to work optimally, power load should be distributed equally over the three phases, in which case there is virtually no current running through the neutral wire. When loads are unequal, this is called asymmetry, which results in energy losses and is also believed to be the cause of a fair number of problems experienced by end users.

A utility company operating a nationwide power grid had recently invested in installing sensors in the 35,000 transformation stations in the grid. The underlying vision was that the company wanted to evolve towards a more data-driven grid and asset management. The end goal of the initiative under study here, was to develop a model that predicts asymmetry even before it occurs, based on time-series data on voltages and loads, data on network interruptions and data on changes in the number of connections to the grid (primary data-analytic problem). This predictive model was to be used as an early warning system allowing technicians to intervene and adjust power delivery to prevent asymmetry from occurring (decision framework), thus reducing energy losses. Preventing asymmetry was also believed to mitigate complaints and issues experienced by customers (business value). The left-hand side of Fig. 3 shows this primary data-analytic problem and its presumed business value.

Attaining this end goal implied more work than merely fitting the predictive model. In particular, there were no data available on asymmetry and it was not clear to what extent customer complaints were actually related to asymmetry. The team made the following provisional roadmap:

- Subproject 1: Determine asymmetry from raw sensor data. The newly installed sensors in the transformation stations allowed the computation of asymmetry in each station every 15 min. The first subproject developed this algorithm and implemented the infrastructure needed for collecting and storing the resulting asymmetry data.
- Subproject 2: Establish whether complaints and issues are caused by asymmetry. The asymmetry data made available through the algorithm of Subproject 1 would then be collated with customer complaints to analyze what fraction of issues experienced by end users could be related to asymmetry.
- Subproject 3: Predict asymmetry. Provided the second subproject established that a substantial part of customer complaints is due to asymmetry, the third subproject would then develop a model that predicts asymmetry before it occurs. For training the model, the subproject used the asymmetry data created in the first subproject. The model predicts asymmetry in a transformation station (unit of analysis) every 15 min. If asymmetry is detected, this is signaled to the grid's operational management, who will then decide how to intervene.

Conclusions

The literature identifies proper project definition as crucial for the success of data-analytic projects, and also recognizes that the translation of business objectives into data-analytic problems is a difficult task. Common project models, such as CRISP-DM, TDSP and FMDS, specify *what* needs to be done. This results in the deliverables specified in Table 1. Unfortunately, such project models do not offer operational guidance for *how* to arrive at a problem definition, and support is at a high level of abstraction. Techniques

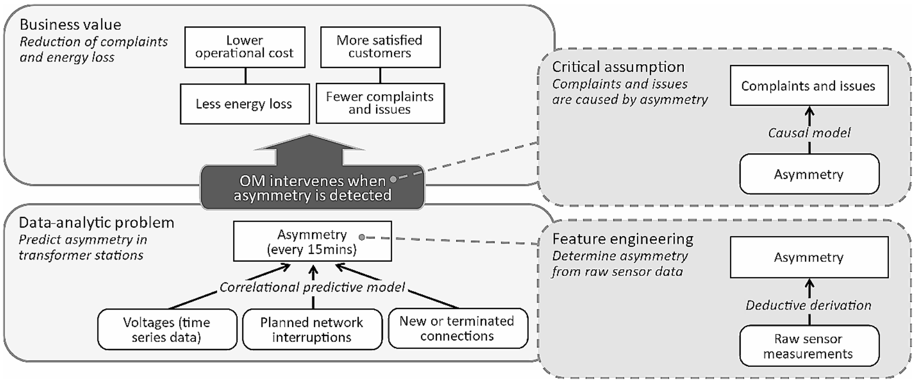
Business value
Reduction of complaints and energy loss
Lower operational cost
More satisfied customers
Less energy loss
Fewer complaints and issues
OM intervenes when asymmetry is detected
Critical assumption
Complaints and issues are caused by asymmetry
Complaints and issues
Causal model
Asymmetry
Data-analytic problem
Predict asymmetry in transformer stations
Asymmetry (every 15mins)
Correlational predictive model
Voltages (time series data)
Planned network interruptions
New or terminated connections
Feature engineering
Determine asymmetry from raw sensor data
Asymmetry
Deductive derivation
Raw sensor measurements

Fig. 3 Data-Analytic Problem Structure for Example 2: Early-warning system in a power grid

for problem structuring in other fields, including operations research and business analytics, are a stronger point of departure, but such techniques need to be adjusted to the particular structure of data-analytic problems. Departing from theory on the structure of data-analytic problems, and following an iterative design-science approach, we developed and refined the Data-Analytic Problem Structure diagram, DAPS, which the paper presents.

DAPS guides practitioners in defining a clear, logically consistent project where data drive a predictive model, which in turn is used to contribute to the goals of an organization. DAPS helps the project owner to be aware of the conditions and critical assumptions underlying the way in which the project can contribute to organizational KPIs, thereby helping the project owner to chunk the overall ambition into subprojects and go/no-go moments. This increases the likelihood of successful project execution, as unclear project goals or overly ambitious goals in data-analytic projects are known to be important reasons for project stagnation or failure.

DAPS acts as a communication tool by having relevant stakeholders from the perspectives of the business, the data and the data-analytic methods contribute to one clear, logically consistent flow in which data are used to contribute to an organization's business goals. This iterative process helps to prevent projects from departing from merely one of these perspectives, and missing logical flaws that are evident when considered from the other perspectives. Typical examples are projects that build predictive models using data that are available in an organization but that fail to contribute to the organization's goals, as they are not linked to a decision process. Or, the other way around, projects starting from the business perspective, hinging on unrealistic assumptions about what data are actually available or what value data-analytic methods can provide. DAPS aims to facilitate the smooth translation of ambitions and ideas from diverse stakeholders into realistic and manageable tasks, while assuring that the project's rationale is clear and logically consistent.

Applying a first version of DAPS in 10 real projects revealed several opportunities for improvement, specifically in the operationalization of the data-analytic problem (bottom-part) and decision-framework (middle-part). These opportunities were addressed by adding instructions to the DAPS-diagram (right side of Fig. 1). A second iteration of its use in 37 real projects revealed no further opportunities for improvement, and demonstrated that DAPS is a strong addition to the CRISP-DM framework and similar models.

Appendix: evaluation protocol

The research follows the framework of design science research, at the heart of which is the cycle of designing a version of the technique, and then evaluating it in its application context. As explained in the paper, the DAPS model was applied in 47 real data science projects, which were done by students of professional and executive programs in data science offered at two distinct universities. The projects were reviewed by the authors and discussed with the students, evaluating to what extent they gave a satisfactory fulfillment of the criteria in Table 1, and discussing with the students in how far DAPS had or had not offered effective support in achieving them.

In this Appendix, we present additional details on how the evaluations were done and what criteria were applied.

How were the 47 projects evaluated?

- Each project was reviewed multiple times as it unfolded (typically four to eight times in total). Students applied DAPS in the Business Understanding phase of their projects, but the authors followed the projects also in the ensuing CRISP-DM phases to identify issues in the project definition that emerged later.

Who did the evaluations?

- One or both authors, who are both experienced data scientists, as well as experts in theory on data science.
- The evaluations were done in discussion with the students applying the proposed DAPS technique.

What was evaluated?

- The DAPS model and the instructions used to explain the technique.

What were the criteria for evaluating an application of DAPS?

- Applications of DAPS in the projects were evaluated on how well the functions listed in Table 1 were fulfilled, how helpful the DAPS technique had been for fulfilling them, and how good the resulting project definition was. Below, we make these three main criteria more specific.

1. The evaluation assesses whether the four functions listed in Table 1 were fulfilled satisfactorily.

- Function 1: *Translation of business goal into data-analytic problem*. Based on theory in machine learning and data-analytics such as [4, 5, 16], the validity and preciseness of the data-analytic problem definition were assessed. For example, is the problem a valid predictive, prescriptive or diagnostic problem? Are the Y variables well-defined?
- Function 2: *Definition of business value*. The rationale and precision of the decision framework was assessed, as well as its relationship with project and organization KPIs. For example, how precise is the definition of the decisions or actions for which the algorithm is to be used? How convincing is the rationale for linking the decision framework to the anticipated improvement in the mentioned KPIs? Does the organization itself recognize the goals described as organization KPIs?
- Function 3: *Specification of the type of model*. Based on theory in statistical learning and machine learning [4, 41–43] it was assessed whether the type of model (causal, correlational or deductive) is suitable for the model's purpose. For example, when the model should be able to make predictions involving interventions in the X variables, was the model specified as *causal* instead of *correlational*?
- Scope the project into manageable chunks: it was assessed whether the technique facilitated the process of breaking down a large ambition into manageable subprojects.

2. The evaluation assesses in how far DAPS had or had not offered effective support in fulfilling the functions in Table 1.

- Students were asked whether the DAPS model and its instructions were helpful in making the project definition.

- Students were asked to identify elements of DAPS that they found unclear or difficult to apply, as well as suggestions for improvement.

3. Later in the projects, it was monitored whether students encountered issues in the original project definitions.

- All 47 projects followed the CRISP-DM stage model, where students applied DAPS in the Business Understanding stage. The authors followed the projects also in the ensuing stages (Data Understanding, Data Preparation, Modelling, Evaluation and Deployment), identifying issues in the original project definition.
- In case of issues, the authors discussed with the students whether these could have been prevented by improving the DAPS model or its instructions, or instead, that the issues were due to unknowns or complications in the project's context itself.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40537-024-00916-7.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
The first author conceived the study. Both authors conducted the research and wrote the manuscript text. Both authors reviewed and approved the final manuscript.

Funding
No funding received.

Data availability
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 23 February 2023 / Accepted: 7 April 2024

Published online: 12 April 2024

References

1. Sabharwal R, Miah SJ. A new theoretical understanding of big data analytics capabilities in organizations: a thematic analysis. J Big Data. 2021;8:159.
2. Kurgan L, Musilek P. A survey of knowledge discovery and data mining process models. Knowl Eng Rev. 2006;21(1):1–24.
3. Martínez-Plumed F, Contreras-Ochando L, Ferri C, Hernandez-Orallo J, Kull M, et al. CRISP-DM twenty years later: from data mining processes to data science trajectories. IEEE Trans Knowl Data Eng. 2021;33(8):3048–61.
4. De Mast J, Steiner SH, Nuijten WPM, Kapitan D. Analytical problem solving based on causal, correlational and deductive models. Am Stat. 2023;77(1):51–61.
5. Davenport T, Harris J. Competing on analytics: the New Science of winning (updated edition). Boston, MA: Harvard Business School Press; 2017.
6. Kühn A, Joppen R, Reinhart F, Röltgen D, Von Enzberg S, Dumitrescu R. Analytics Canvas — a framework for the design and specification of data analytics projects. Procedia CIRP. 2018;70:162–7.
7. Chapman P, et al. CRISP-DM 1.0: step-by-step data mining guide. Tech. Rep., The CRISP-DM Consortium; 2000.
8. Martinez I, Viles E, Olaizola IG. Data science methodologies: current challenges and future approaches. Big Data Res. 2021. https://doi.org/10.1016/j.bdr.2020.100183.
9. Rollins J. Foundational methodology for data science. 2015, https://www.ibm.com/downloads/cas/WKK9DX51. Accessed 18 Feb 2023.

10. Becker DK. Predicting outcomes for big data projects: big data project dynamics (bdpd): research in progress. IEEE Int Conf Big Data. 2017; 2320–30.
11. Hoerl R, Kuonen D, Redman TC. Framing data science problems the right way from the start. MIT Sloan Manag Rev Apr. 2022.
12. Elragal A, Klischewski R. Theory-driven or process-driven prediction? Epistemological challenges of big data analytics. J Big Data. 2017;4:19.
13. Das M, Cui R, Campbell DR, Agrawal G, Ramnath R. Towards methods for systematic research on big data. IEEE Int Conf Big Data. 2015: 2072–81.
14. Saltz J, Shamshurin I, Connors C. Predicting data science sociotechnical execution challenges by categorizing data science projects. J Assoc Inf Sci Technol. 2017;68(12):2720–8.
15. Henke N, Levine J, McInerney P. You don't have to be a data scientist to fill this must-have analytics role. Harv Bus Rev Feb. 2018.
16. Provost F, Fawcett T. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly: Sebastopol (CA); 2013.
17. SAS. From data to business advantage: Data Mining, SEMMA Methodology and the SAS System (White Paper). SAS Institute; 1997.
18. Li Y, Thomas M, Osei-Bryson K. A snail shell process model for knowledge discovery via data analytics. Decis Support Syst. 2016;91:1–12.
19. Doran G. There's a smart way to write management's goals and objectives. Manage Rev. 1981;70:35–6.
20. Dorard L. The machine learning canvas. Gumroad, 2019.
21. Takeuchi H, Ito Y, Yamamoto S. Method for constructing machine learning project canvas based on enterprise architecture modeling. Procedia Comput Sci. 2022;207:425–34.
22. Cabena P, Hadjinian P, Stadler R, Verhees J, Zanasi A. Discovering Data Mining: from concepts to implementation. Prentice Hall; 1998.
23. Angée S, Lozano S, Montoya-Munera E, Ospina Arango J, Tabares M. Towards an improved ASUM-DM process methodology for cross-disciplinary multi-organization big data & analytics projects. 13th International Conference, Knowledge Management in Organizations, Aug 2018, Proceedings 13: 613–624.
24. Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Comm ACM. 1996;39(11):27–34.
25. Anand S, Buchner A. Decision support using Data Mining. FT Management; 1998.
26. Ackoff RL, Vergara E. Creativity in problem solving and planning: a review. Eur J Oper Res. 1981;7:1–13.
27. Ho JKK, Sculli D. The scientific approach to problem solving and decision support systems. Int J Prod Econ. 1997;48:249–57.
28. Mingers J, Rosenhead J. Problem structuring methods in action. Eur J Oper Res. 2004;152:530–54.
29. Eden C. Analyzing cognitive maps to help structure issues or problems. Eur J Oper Res. 2004;159:673–86.
30. Rahman S. Theory of constraints: a review of the philosophy and its applications. Int J Oper Prod Manag. 1998;18(4):336–55.
31. De Koning H, De Mast J. The CTQ flowdown as a conceptual model of project objectives. Qual Manag J. 2007;14(2):19–28.
32. Karnopp DC, Margolis DL, Rosenberg RC. System dynamics: modeling, Simulation, and Control of Mechatronic Systems. 5th ed. New York: Wiley; 2012.
33. Bazili V, Caldiera G, Rombach D. Goal question Metric (GQM) approach. In: Encyclopedia of Software Engineering. Wiley; 2002.
34. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York, NY: Springer; 2013.
35. Shaw D, Westcombe M, Hodgkin J, Montibeller G. Problem structuring methods for large group interventions. J Oper Res Soc. 2004;55:453–63.
36. Akoka J, Comyn-Wattiau I, Prat N, Storey VC. Knowledge contributions in design science research: paths of knowledge types. Decis Support Syst. 2023;166:113898.
37. Van Aken JE, Chandrasekaran A, Halman J. Conducting and publishing design science research: inaugural essay of the design science department of the Journal of Operations Management. J Oper Manag. 2016;47:1–8.
38. Denyer D, Tranfield D, Van Aken JE. Developing design propositions through research synthesis. Organ Stud. 2008;29(3):393–413.
39. Hevner AR, March ST, Park J, Ram S. Design Science in Information Systems Research. MIS Q. 2004;28(1):75.
40. Hevner AR. A three cycle view of design science research. Scand J Inf Syst. 2007;19(2):87–92.
41. Pearl J. Causal inference in statistics: an overview. Stat Surv. 2009;3:96–146.
42. Pearl J, MacKenzie D. The Book of why — the New Science of cause and Effect. New York, NY: Basic Books; 2018.
43. Rubin D. Estimating Causal effects of treatments in Randomized and Nonrandomized studies. J Educ Psychol. 1974;66(5):688–701.
44. Steenstrup K, Sallam RL, Eriksen L, Jacobson SF. Industrial analytics revolutionizes Big Data in the digital business. In: Gartner Research. 2014. https://www.gartner.com/en/documents/2826118. Accessed 18 Feb 2023.
45. Grundy T. Strategy implementation and Project Management. Int J Proj Manag. 1988;16(1):43–50.
46. Mintzberg H. The rise and fall of Strategic Planning. Prentice Hall; 1994.
47. McElroy W. Strategic Change through Project Management. APM; 1995.
48. Thoms P, Pinto JK. Project leadership: a question of timing. Proj Manag. 1999;30(1):19–26.
49. Clarke A. A practical use of key success factors to improve the effectiveness of project management. Int J Proj Manag. 1999;17(3):139–45.
50. Lewis R. Take the `big' out of big projects: break them into manageable chunks. InfoWorld. 1996;18(20):24.
51. Raz T, Globerson S. Effective sizing and content definition of work packages. Proj Manag. 1998;29(4):17–23.

Publisher's Note