Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N. J. A. H., Ramírez-Quintana, M. J., & Flach, P. A. (2019). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2019.2962680

CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories

Fernando Martínez-Plumed, Lidia Contreras-Ochando, Cèsar Ferri, José Hernández-Orallo, Meelis Kull, Nicolas Lachiche, María José Ramírez-Quintana and Peter Flach

Abstract—CRISP-DM (CRoss-Industry Standard Process for Data Mining) has its origins in the second half of the nineties and is thus about two decades old. According to many surveys and user polls it is still the *de facto* standard for developing data mining and knowledge discovery projects. However, undoubtedly the field has moved on considerably in twenty years, with *data science* now the leading term being favoured over *data mining*. In this paper we investigate whether, and in what contexts, CRISP-DM is still fit for purpose for data science projects. We argue that if the project is goal-directed and process-driven the process model view still largely holds. On the other hand, when data science projects become more exploratory the paths that the project can take become more varied, and a more flexible model is called for. We suggest what the outlines of such a trajectory-based model might look like and how it can be used to categorise data science projects (goal-directed, exploratory or data management). We examine seven real-life exemplars where exploratory activities play an important role and compare them against 51 use cases extracted from the NIST Big Data Public Working Group. We anticipate this categorisation can help project planning in terms of time and cost characteristics.

Index Terms—Data Science Trajectories, Data Mining, Knowledge Discovery Process, Data-driven Methodologies.

1 Introduction

Towards the end of the previous century, when the systematic application of data mining techniques to extract knowledge from data was becoming more and more common in industry, some companies and institutions saw the need of joining forces to identifying good practices as well as common mistakes in their past experiences. With funding from the European Union, a team of experienced data mining engineers developed a generally applicable data mining methodology which over time would become widely accepted. In 1999 the first version of the *CRoss-Industry Standard Process for Data Mining*, better known as CRISP-DM, was introduced [1]. This straightforward methodology was conceived to catalogue and guide the most common steps in data mining projects. It soon became "*de facto* standard for developing data mining and knowledge discovery projects" [2], and it is still today the most widely-used analytic methodology according to many opinion polls.

In the last two decades the ubiquity of electronic devices and sensors, the use of social networks and the capacity of storing and exchanging these data all have dramatically increased the opportunities for extracting knowledge through data mining projects. The diversity of the data has increased – in origin, format and modalities – and so has the variety of techniques coming from machine learning, data management, visualisation, causal inference and other areas. But, more importantly, compared to twenty years ago there are many more ways in which data can be monetised, through new kinds of applications, interfaces and business models. While the area of deriving value from data has grown exponentially in size and complexity, it has also become much more exploratory under the umbrella of *data science*. In the latter, data-driven and knowledge-driven stages interact, in contrast to the traditional data mining process, starting from precise business goals that translate into a clear data mining task, which ultimately converts "data to knowledge". In other words, not only has the nature of the data changed but also the processes for extracting value from it.

Clearly these changes did not happen overnight, and new methodologies have been proposed in the meantime to accommodate some of the changes. For instance, IBM introduced ASUM-DM [3], and SAS introduced SEMMA [4], and many others, as we will review in more detail in the following section. However, the original CRISP-DM model can still be recognised in these more recent proposals, which remain focused on the traditional paradigm of a sequential list of stages from data to knowledge. We would argue that they are still, in essence, data mining methodologies that do not fully embrace the diversity of data science projects.

In this paper we investigate the extent to which, after twenty years, the original CRISP-DM and the underlying data mining paradigm remain applicable for the much wider range of data science projects we see today. We identify new activities in data science, from data simulation to narrative exploration. We propose a general diagram containing the possible activities that can be included in a data science project. Based on examples, we distinguish

F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo and MJ. Ramírez-Quintana are with Universitat Politècnica de València, Spain e-mail: {fmartinez,liconoc,cferri,jorallo,mramirez}@dsic.upv.es.

M. Kull is with University of Tartu, Estonia e-mail: meelis.kull@ut.ee.

N. Lachiche is with Université de Strasbourg, France e-mail: nicolas.lachiche@unistra.fr.

P. Flach is with the University of Bristol and the Alan Turing Institute, U.K. e-mail: peter.flach@bristol.ac.uk.

particular *trajectories* through this space that distinguish different kinds of data science projects. We propose that these trajectories can be used as templates for data scientists when planning their data science projects and, in this way, explore new activities that could be added to or removed from their workflows. Together, they represent a new Data Science Trajectories model (DST).

On one hand, this DST model represents an important overhaul of the original CRISP-DM initiative. However, we have been careful not to discard CRISP-DM completely, as it still represents one of the most common trajectories in data science, those that go from data to knowledge when there is a clear business goal that translates into a data mining goal. One could say that DST is "backwards compatible" with CRISP-DM, while allowing the considerable additional flexibility that twenty-first century data science demands. In this paper we identify some other trajectories that capture the common routes of data science projects, but the flexibility of the DST map makes it possible to incorporate current and new methodologies in the development and deployment of data science projects.

The contributions of the paper are the following:

- Recognition of the limitations of the original CRISP-DM and other related methodologies considering the diversity of data science projects today.
- Identification of more exploratory activities that are common in data science but not covered by CRISP-DM, leading to a more flexible and comprehensive DST map.
- Recognition of popular trajectories in this space describing well-known practices in data science, which could be used as templates, making the DST model exemplary rather than prescriptive.
- Some general suggestions on how the DST model can be coupled with actual project management methodologies in order to be customised to different organisations and contexts.

The rest of the paper is organised as follows. Section 2 revisits CRISP-DM and other related variations that have been introduced in the last two decades. The identification of new activities and the formulation of the DST map is included in Section 3. Section 4 illustrates these trajectories on real cases of data science projects, using a precise notation on trajectory charts. In Section 5 we discuss data science project management by considering the three kinds of activities, looking at these seven real cases plus 51 use cases from the NIST Big Data Public Working Group. Section 6 compares the model with software methodologies and the scientific method, suggesting how organisation can couple this with existing and new methodologies, as well as particular ethical issues and the challenge of data science automation. The appendix includes more detail about the experimental analysis over the 7 + 51 use cases covered in the paper.

2 CRISP-DM and Related Process Models

In this section we give a succinct description of the most used and cited data mining and knowledge discovery methodologies, providing for each an overview of its evolution, basis and primary characteristics. For a more comprehensive description of these methodologies we refer the reader to [5], [6]. Fayyad, Piatetsky-Shapiro and Smyth define *Knowledge Discovery in Databases* (KDD) as "the overall process of knowledge discovery from data, including how the data is stored and accessed, how algorithms can be scaled to massive datasets and still run efficiently, how results can be interpreted and visualised, and how the overall human-machine interaction can be modeled and supported" and data mining as a single step in this process, turning suitably pre-processed data into patterns that can subsequently be turned into valuable and actionable knowledge [7]. However, data mining is often used as a synonym for KDD, and we will not distinguish between the two meanings in this paper.

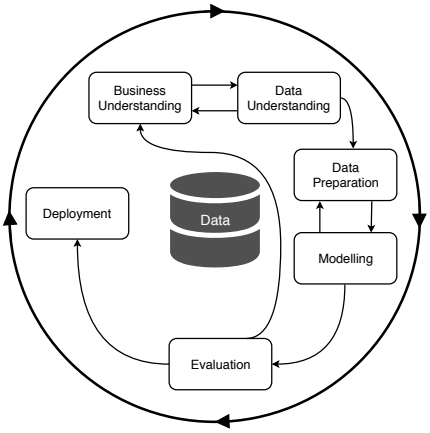

Fig. 1. The CRISP-DM process model of data mining.

As already mentioned in the introduction, CRISP-DM [1] can be viewed as the canonical approach from which most of the subsequent proposals have evolved (both for data mining and data science process models). It elaborates and extends the steps in the original KDD proposal into six steps: Business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment. Figure 1 depicts the six steps of CRISP-DM and the way they are sequenced in a typical data mining application.

Several process models and methodologies were developed around the turn of the century using CRISP-DM as a basis, but with varying objectives. Some examples include:

- *Human-Centered Approach to Data Mining* [8], [9], which involves a holistic understanding of the entire Knowledge Discovery Process, considering people's involvement and interpretation in each phase and putting emphasis on that the target user is the data engineer.
- *SEMMA* [4], which stands for Sample, Explore, Modify, Model and Assess, is the proprietary methodology developed by SAS[1] to develop Data Mining products and is mainly focused on the technical aspects.
- Cabena's [10] model, used in the marketing and sales domain, this being one of the first process models which took into account the business objectives;
- Buchner's [11] model, adapted to the development of web mining projects and focused on an online customer (incorporating the available operational and materialised data as well as marketing knowledge).

1. www.sas.com

- *Two Crows* [12], which takes advantage of some insights from (first versions of) CRISP-DM (before release), and proposes a non-linear list of steps (very close to those from KDD), so it is possible to go back and forth.
- *D^3M* [13], a domain-driven data mining approach proposed to promote the paradigm shift from "data-centered knowledge discovery" to "domain-driven, actionable knowledge delivery".

There are also some other relevant approaches not directly related to the KDD task. The *5 A's Process* [14], originally developed by SPSS[2], already included an "Automate" step which helps non-expert users to automate the whole process of DM applying already defined methods to new data, but it does not contain steps to understand the business objectives and to test data quality. Another approach that tries to assist the users in the DM process is [15]. All these were influential for CRISP-DM. In 1996 Motorola developed the 6σ approach [16], which emphasises measurement and statistical control techniques for quality and excellence in management. Another approach is the *KDD Roadmap* [17], an iterative data mining methodology that as a main contribution introduces the "resourcing" task, consisting in the integration of databases from multiple sources to form the operational database.

The evolution of these data mining process models and methodologies is graphically depicted in Figure 2. The arrows in the figure indicate that CRISP-DM incorporates principles and ideas from most of the aforementioned methodologies, while also forming the basis for many later proposals. CRISP-DM is still considered the most complete data mining methodology in terms of meeting the needs of industrial projects, and has become the most widely used process for DM projects according to the KDnuggets polls (https://www.kdnuggets.com/) held in 2002, 2004, 2007 and 2014. In short, CRISP-DM is considered the *de facto* standard for analytics, data mining, and data science projects.

To corroborate this view from data science experts, we also checked that CRISP-DM is still a very common methodology for *data mining* applications. For instance, just focussing on the *past four years*, we can find a large number of conventional studies applying or slightly adapting the CRISP-DM methodology to many different domains: healthcare [18], [19], [20], [21], signal processing [22], engineering [23], [24], education [25], [26], [27], [28], [29], logistics [30] production [31], [32], sensors and wearable applications [33], tourism [34], warfare [35], sports [36] and law [37].

However, things have evolved in the business application of data mining since CRISP-DM was published. Several new methodologies have appeared as extensions of CRISP-DM, showing how it can be modernised without changing it fundamentally. For instance, the *CRISP-DM 2.0* Special Interest Group (SIG) was established with the aim of meeting the changing needs of DM with an improved version of the CRISP-DM process. This version was scheduled to appear in the late 2000s, but the group was discontinued before the new version could be delivered. Other examples include:

2. http://www.spss.com/

- Cios et al.'s *Six-step discovery process* [38], [39], which adapts the CRISP-DM model to the needs of the academic research community (research-oriented descriptions, explicit feedback mechanisms, extension of discovered knowledge to other domains, etc.).
- *RAMSYS* (RApid collaborative data Mining SYStem) [40], a methodology for developing collaborative DM and KD projects with geographically diverse groups.
- *ASUM-DM* (Analytics Solutions Unified Method for Data Mining/Predictive Analytics) [3], a methodology which refines and extends CRISP-DM, adding infrastructure, operations, deployment and project management sections as well as templates and guidelines, personalised for IBM's practices.
- *CASP-DM* [41], which addresses specific challenges of machine learning and data mining for context change and model reuse handling.
- *HACE* [42], a Big Data processing framework based on a three tier structure: a "Big Data mining platform" (Tier I), challenges on information sharing and privacy, and Big Data application domains (Tier II), and Big Data mining algorithms (Tier III).

The aforementioned methodologies have in common that they are designed to spend a great deal of time in the business understanding phase aiming at gathering as much information as possible before starting a data mining project. However, the current data deluge as well as the experimental and exploratory nature of data science projects require less rigid and more lightweight and flexible methodologies. In response, big IT companies have introduced similar lifecycles and methodologies for data science projects. For example, in 2015 IBM released the *Foundational Methodology for Data Science* (FMDS) [43], a 10-stage data science methodology that – although bearing some similarities to CRISP-DM– emphasises a number of the new practices such as the use of very large data volumes, the incorporation of text analytics into predictive modelling and the automation of some of the processes. In 2017 Microsoft released the *Team Data Science Process* (TDSP) [44], an "agile, iterative, data science methodology to deliver predictive analytics solutions and intelligent applications efficiently" and to improve team collaboration and learning.

At a high level, both FMDS and TDSP have much in common with CRISP-DM. This demonstrates the latter's flexibility, which allows to include new specific steps (such as analytic and feedback phases/tasks) that are missing in the original proposal. On the other hand, methodologies such as FMDS and TDSP are in essence still data mining methodologies that assume a clearly identifiable goal from the outset. In the next section we argue that data science calls for a much more exploratory mindset.

3 From Goal-Directed Data Mining Processes to Exploratory Data Science Trajectories

As is evident from the previous section, the perspective of CRISP-DM and related methodologies is that data mining is a process starting from a relatively clear business goal and data that have already been collected and are available for further computational processing. This kind of process is

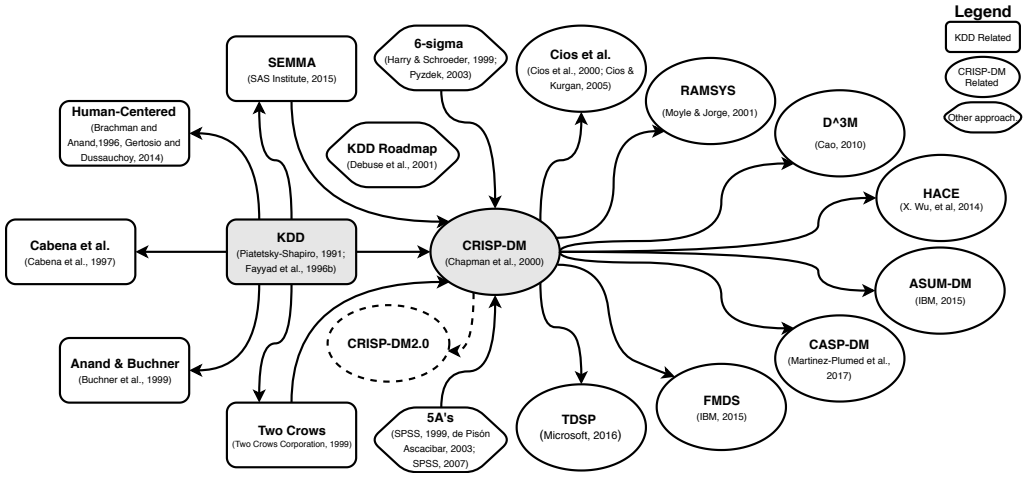

Fig. 2. Evolution of most relevant Data Mining and Data Science models and methodologies (in white and light blue, respectively). KDD and CRISP-DM are the 'canonical' methodologies, depicted in grey. Adapted from [6]. The years are those of the most representative papers, not the years in which the model was introduced.

akin to mining for valuable minerals or metals at a given geographic location where the existence of the minerals or metals has been established: data are the ore, in which valuable knowledge can be found. Whenever this kind of metaphor is applicable, we suggest that CRISP-DM is a good methodology to follow and still holds its own after twenty years.

However, *data science* is now a much more commonly used term than *data mining* in the context of knowledge discovery. A quick query on Google Trends shows that the former became a more frequent search term than the latter in early 2016 and now is more than twice as common. So what is data science? There seem to be two broad senses in which the term is used: (a) the science OF data; and (b) applying scientific methods TO data. From the first perspective, data science is seen as an academic subject that studies data in all its manifestations, together with methods and algorithms to manipulate, analyse, visualise and enrich data. It is methodologically close to computer science and statistics, combining theoretical, algorithmic and empirical work. From the second perspective, data science spans both academia and industry, extracting value from data using scientific methods, such as statistical hypothesis testing or machine learning. Here the emphasis is on solving the domain-specific problems in a data-driven way. Data are used to build models, design artefacts, and generally increase understanding of the subject. If we wanted to distinguish these two senses then we could call the first *theoretical data science*; and the second, *applied data science*. In this paper, we are really concerned with the latter and henceforth we use the term 'data science' in this applied sense.

The key difference we perceive between data mining twenty years ago and data science today is that the former is goal-oriented and concentrates on the process, while the latter is data-oriented and exploratory. Developed from the goal-oriented perspective, CRISP-DM is all about processes and different tasks and roles within those processes. It views the data as an ingredient towards achieving the goal – an important ingredient, but not more. In other words, from the data mining perspective, the process takes centre stage. In contrast, in contemporary data science the *data* take centre stage: we know or suspect there is value in these data, how do we unlock it? What are the possible operations we can apply to the data to unlock and utilise their value? While moving away from the process, the methodology becomes less prescriptive and more inquisitive: things you *can* do to data rather than things you *should* do to data.

To continue with the 'mining' metaphor: if data mining is like mining for precious metals, data science is like *prospecting*: searching for deposits of precious metals where profitable mines can be located. Such a prospecting process is fundamentally exploratory and can include some of the following activities:

Goal exploration: finding business goals which can be achieved in a data-driven way;
Data source exploration: discovering new and valuable sources of data;
Data value exploration: finding out what value might be extracted from the data;
Result exploration: relating data science results to the business goals;
Narrative exploration: extracting valuable stories (e.g., visual or textual) from the data;
Product exploration: finding ways to turn the value extracted from the data into a service or app that delivers something new and valuable to users and customers.

While it is possible to see (weak) links between these exploratory activities and CRISP-DM phases (e.g. goal exploration relates to business understanding and result exploration relates to modelling and evaluation), the former are typically more open-ended than the CRISP-DM phases. In data science, the order of activities depends on the domain as well as on the decisions and discoveries of the data scientist. For example, after getting unsatisfactory results in data value exploration performed on given data it might be necessary to do further data source exploration. Alternatively, if no data are given then data source exploration would come before data value exploration. Sometimes neither of these

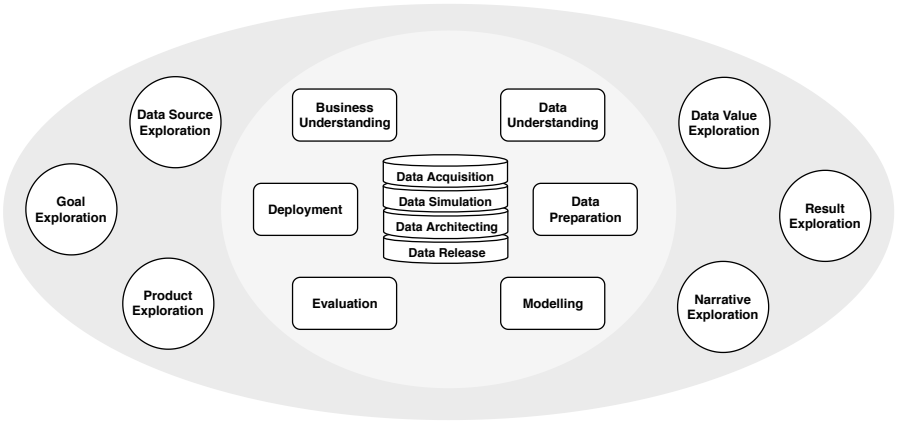

Fig. 3. The DST map, containing the outer circle of exploratory activities, inner circle of CRISP-DM(or goal-directed) activities, and at the core the data management activities.

activities is required, and sometimes these activities would be run several times.

Data science projects are certainly not only about exploration, and contain more goal-driven parts as well. The standard six phases of the CRISP-DM model from business understanding to deployment are all still valid and relevant. However, in data science projects it is common to see only partial traces through CRISP-DM. For example, sometimes there is no need for activities beyond data preparation, as the prepared data are the final product of the project. Data that is scraped from different sources, integrated and cleansed can be published or sold for various purposes, or can be loaded into a data warehouse for OLAP querying. The CRISP-DM phases are also often interrupted by further exploratory activities, whenever the data scientist decides to seek more information and new ideas.

We hence see a successful data science project as following a *trajectory* through a space like the one depicted in Figure 3. In contrast to the CRISP-DM model there are no arrows here, because the activities are not to be taken in any pre-determined order. It is the responsibility of the project's leader(s) to decide which step to take next, based on the available information including the results of previous activities. Even though the space contains all the CRISP-DM phases, these are not necessarily run in the standard order, as the goal-driven activities are interleaved with exploratory activities, and these can sometimes set new goals or provide new data.

Data take centre-stage in data science, and the terms 'data preparation' and 'modelling' do not fully capture anymore the variety of practical work that might be carried out on the data. Two decades ago, many applications, especially those falling under the term business intelligence, were based on analysing their own data (e.g., customer behaviour) and extracting patterns from it that would meet the business goals. But today, many more options are considered.

For instance, causal inference [45] has recently been pointed out as a new evolution of data analysis aimed to understand the cause-effect connections in data. Causal inference from data focuses on answering questions of the type "what if" and relies on methods that incorporate causal knowledge (such as the Structural Causal Models [46], the Potential Outcomes Framework [47] or the Linear non-Gaussian acyclic models [48]). Hernan et al. [49] discuss how data science can tackle causal inference from data by considering it as a new kind of data science task known as *counterfactual prediction*. Basically, counterfactual prediction requires to incorporate domain expert knowledge not only to formulate the goals or questions to be answered and to identify or generate the data sources, but also to formally describe the causal structure of the system. This task and others performing causal inference go well within CRISP-DM (under the modelling step) but expert knowledge becomes crucial (and, as a result, the inner stages of the CRISP-DM process are harder to automate). For its part, the business understanding phase reinforces its first-stage position in these circumstances as this must be the place where the expert understanding of the domain has to be converted into models and queries which are needed for the subsequent steps (data understanding, preparation, modelling and evaluation).

However, under the causal inference framework, data science must play a more active role with the data. Data is not just an input of the system: "a causal understanding of the data is essential to be able to predict the consequences of interventions, such as setting a given variable to some specified value" [48]. This suggests a more iterative process where we could need to generate new data, for instance through randomised experiments or performing simulations on the observed or generated data, using the expert's causal knowledge in the form of graphical models together with other kinds of domain knowledge or extracted patterns. All these operations are difficult to integrate in the CRISP-DM model and may require new generative activities for data acquisition and simulation.

Another relevant area where CRISP-DM seems to fall short is when thinking about "data-driven products", such as a mobile app that takes information from the location of their users and recommends routes to other users, according to their patterns. The *product* is the data and the knowledge extracted from it. This perspective was unusual two decades ago, but it is now widespread. Also, nowadays the data might have multiple uses, even far away from the context or

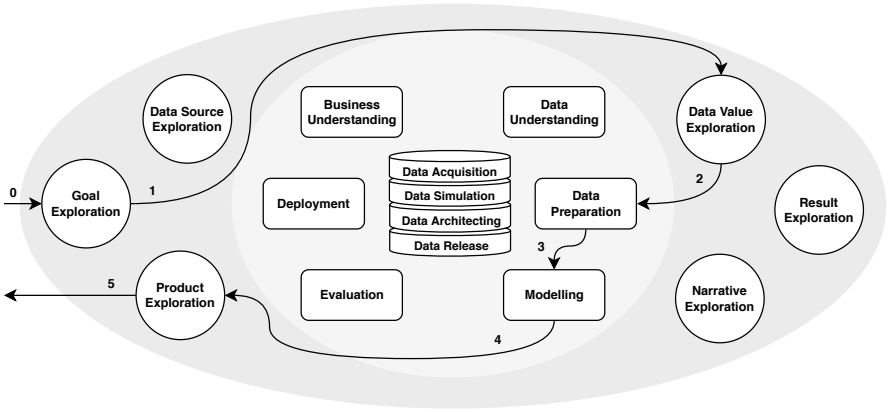

Fig. 4. Example trajectory through a data science project.

domain where they were collected (e.g., the data collected by an electronic payment system can be bought and used by a multinational company to know where a new store will be best located, or can be used by an environmental agency to obtain petrol consumption patterns). The huge size and complexity of the data in some applications nowadays also suggest that handling the data requires important technical work on curation and infrastructure. In other words, the CRISP-DM model included the 'data' as a static disk cylinder in the middle of the process (see Figure 1), but we want to highlight the activities around this disk, going beyond data preparation and integration[3]. Given the variety of scenarios for using the data from others or from yourself, for your own or others' benefit, we consider the following data management activities.

Data acquisition: obtaining or creating relevant data, for example by installing sensors or apps;

Data simulation: simulating complex systems in order to produce useful data, ask causal (e.g., what-if) questions;

Data architecting: designing the logical and physical layout of the data and integrating different data sources;

Data release: making the data available through databases, interfaces and visualisations.

Once the set of activities has been introduced, a trajectory is simply an acyclic directed graph over activities, usually representing a sequence, but occasionally forking to represent when things are done in parallel (by different individuals or groups in a data science team). An example of a trajectory through the DST map is given in Figure 4, where the goal is established as a first step in a data-driven way (*goal exploration*), and relevant data is then explored to extract valuable knowledge (*data value exploration*). Classical CRISP-DM activities are performed to clean and transform the data (*data transformation*) which will be used to train a particular machine learning model (*modelling*). Finally, the most appropriate end-user product and/or presentation is explored (*product exploration*) in order to turn the value extracted from the data into a valuable product for users and customers. This example will be visited in full detail in section 4.1.

3. Despite disk cylinders not being cognitively associated with activities as a representation, we have decided to use them to emphasise the correspondence with the original CRISP-DM model

As we will do in the next section, we can represent trajectories more compactly, by removing those activities that are not used. Still, if an activity happens more than once in a trajectory, we only show the same activity once. For these DST charts, we use numbered arrows to show the process (possibly visiting the same activity more than once)[4]. More precisely, a trajectory chart is defined as follows:

- A DST chart is a directed graph that only includes activities (once) and connections (transitions) between them (as directed solid arrows).
- All arrows are numbered from 0 to N, showing the sequence of transitions between activities. Consequently, we cannot have unlimited loops.
- We use three different types of boxes for activities (circles for exploration activities, rounded squares for CRISP-DM activities, and cylinders for data management activities).
- If two or more arrows have the same number, it means that they take place in parallel (or their sequential order is unattested or unimportant).
- A trajectory can go through the same activity more than once. If the trajectory moves from A to B more than once, we will annotate this as a single arrow with a single label, showing as many transition numbers as needed, separated by commas.
- Every trajectory has an entrance transition (with number 0 and not starting from any activity) and an exit transition (with number N and not ending in any activity).

By following the transitions from 0 to N, we derive one single trajectory from the chart (remember that repeated numbers are not alternatives, but things going in parallel). Once introduced the graphical notation for the charts that completes our DST model, in the following section we present some real-life scenarios and discuss the order of exploratory, goal-directed and data management activities in these scenarios.

4. Note that a trajectory chart represents one single trajectory, and it is *not* a pattern for a set of trajectories. CRISP-DM is actually a pattern and not a single trajectory chart, as CRISP-DM admits several trajectories, especially through the use of the backwards arrows.

4 Examples of Data Science Trajectories

The set of cases we include in this section is not meant to be exhaustive, but aims to show a diverse range of common data science trajectories that illustrate alignments and especially misalignments with parts (and most of them the whole) of the CRISP-DM model, by showing exploratory and data management activities. The exemplar trajectories are also useful to illustrate the graphical notation that we use for the trajectory charts. For each case, we explain the domain and context in a separate subsection while the sequence of activities is explained in the captions of the corresponding figures.

4.1 Tourism recommender

With the increasing popularity of location-based services, there is a large amount of this sort of data being accumulated. For instance, real-time data is being collected from drivers who use the *Waze*[5] navigation app as well as from pedestrians who use the public-transportation app *Moovit*[6], or the popular social network for athletes *Strava*[7], which monitors how cyclists and runners are moving around the city, giving it an unprecedented view on thousands of moving points across the cities. All this information can be collected from thousands of smartphones being walked or driven around a city, and can be used by many different companies that could be interested in this information with very different purposes. For instance, a tour operator would be interested in answering questions related to location recommendation (if we want to do something, where shall we go?) or activity recommendation (if we visit some place, what can we do there?). By exploiting the information retrieved from the aforementioned networks, the company then decided to create a collaborative smart tourism recommendation system to provide personalised plan trips as well as suitable and adequate offers and activities (accommodation, restaurants, museums, transports, shopping and other attractions) appropriate to the users' profile. We find real-word examples such as *Google Travel*[8], a service developed to plan for upcoming trips with summarising info about the users destination in several categories such as day plans, reservations, best routes, etc. In this example, a possible trajectory is shown in Figure 5.

Fig. 5. Tourism recommender: A possible trajectory for the development of a location and activity recommendation system (Section 4.1) may imply that, once the goal is established as a first step (*goal exploration*), the company would decide to use the users' location and activity histories as relevant data (*data value exploration*) from the data which has been retrieved from third party location based services and networks. Then, the *data preparation* activity starts to create a user-location-activity rating tensor which could be used to implement and train a recommendation system (*modelling* stage). Once the best model is selected and evaluated (note that the evaluation against business goals in CRISP-DM is not necessary here), the company may explore the most appropriate end-user product and presentation (*product exploration*), either through simple visualisations or through the development of mobile/web apps.

4.2 Environmental simulator

Simulation processes are an effective resource that may be used to create a whole system in order to generate data that is usually difficult (or expensive) to collect. Moreover, the simulation of complex systems also provides additional advantages such as the possibility of analysing different scenarios and, in this way, estimating the costs and consequences of the alternatives. For instance, agencies and researchers can integrate traffic simulation models with real data about meteorological conditions (e.g., obtained from weather stations located around the city) for building models about pollution spread for different pollutants which are generally linked to fuel combustion as by-products of these processes. The generated system can be used not only to predict the level of the pollutants, but also for simulating the effect on pollution of, for instance, restricting the circulation of cars in certain parts of the city since temporal and spatial resolution of emissions is essential to predict the concentration of pollutants near roadways. A trajectory of working with a simulated system for predicting traffic and pollution is shown in Figure 6.

Transportation agencies and researchers in the past have estimated emissions using one average speed and volume on a long stretch of roadway. With MOVES, there is an opportunity for higher precision and accuracy. Integrating a microscopic traffic simulation model (such as VISSIM) with MOVES allows one to obtain precise and accurate emissions estimates. The proposed emission rate estimation process also can be extended to gridded emissions for ozone modeling, or to localised air quality dispersion modeling, where temporal and spatial resolution of emissions is essential to predict the concentration of pollutants near roadways.

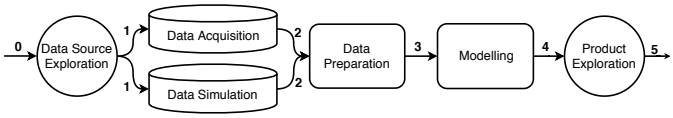

Fig. 6. Environmental simulator: Possible trajectory of an application for predicting pollution in cities (Section 4.2). The first activity must select the data sources for traffic parameters and topology of a city, as well as real meteorological data, all done by means of a *data source exploration*. The real data about weather conditions can then be collected by the sensors distributed along the city (*data acquisition*) and simulated data about traffic can be generated(*data simulation*). In order to make predictions, all the collected data have to be converted (*data preparation*) to a format or structure suitable for being processed by the machine learning techniques (*modelling*). The generated models are then evaluated according to a certain quality criterion (again not against any business goal), and the best model is further used to make the predictions. Finally, the municipalities can explore the most appropriate end-user presentation (*product exploration*), e.g, web or mobile app, and the most effective way to communicate the alerts (e.g, text messages, email alerts or Pop-Up Mobile Ads).

4.3 Insurance refining

Insurance companies can use driving history records, locations and real-time data based on ubiquitous Internet of Things (IoT) sensors to offer context-based insurance plans

5. https://www.waze.com/
6. https://moovit.com/
7. https://www.strava.com/
8. https://www.google.com/travel/

an behavioural policy pricing to their clients. This data can be used to create much more complete user profiles including, for instance, how much time the vehicle is in use, frequent destinations, whether drivers change lane excessively, their driving speeds, to what extent they respect traffic rules, or if they use their smartphone while driving, among many other things. All this information may be used to allow safer drivers to pay less for auto insurance. This may be considered as a special data science project where the insurance company has already deployed a data mining-based product (customer profiling) which could be potentially enriched by means of different new data explorations. This would make a shift from the insurer companies being reactive claim payers to a proactive risk managers. Some major auto insurance companies are already using this sort of data[9]. Fig. 7 shows the trajectory followed which, apart from the classical CRISP-DM cycle used to develop their current customer profiling product, involves new activities.

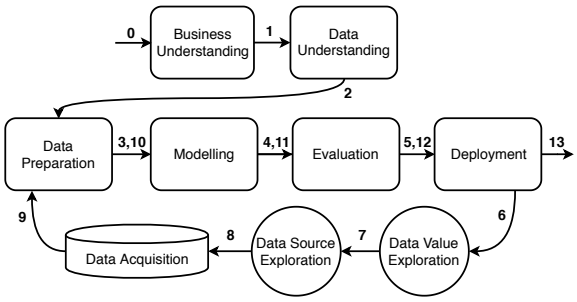

Fig. 7. Insurance refining: For insurance companies (Section 4.3) aiming at improving already deployed products (e.g., customer profiling using data mining), a possible trajectory may imply, after a complete CRISP-DM trajectory, the *exploration of the value of the data*, where the insurance company realises that combining analytical applications (e.g., behavioural models based on customer profile data) with streams of real-time data (e.g., driver's behaviour, vehicle sensors, satellite data, weather reports, etc.) could be an important source for refining the products and services offered; *exploring new data sources*, where the company decides what should be acquired and/or sensorised to create detailed and personalised assessments of risks; and, finally, *data acquisition*, where it is to be decided which kind of sensor technology, smart or wearable devices should be used and where/how they should be installed/used to obtain relevant data.

4.4 Sales OLAP

In a supermarket, managers regularly analyse information regarding the results of merchandise sales since, as a critical resource, it influences directly the operational efficiency of commercial enterprises. For this purpose, managers usually look at the results of various predefined queries, reports and indicators, and can also refine their queries to get a better understanding of the sales. Such managers either write their own queries or use reporting tools. But they need an appropriate representation, a star (or multidimensional) schema, and data organised into datamarts. These datamarts usually come with supporting software (OLAP tools) to make human analysis easier both by lowering the cognitive load of the user to understand/manipulate the data and by speeding up the database system itself. The data analyst, who is here a manager using user-oriented

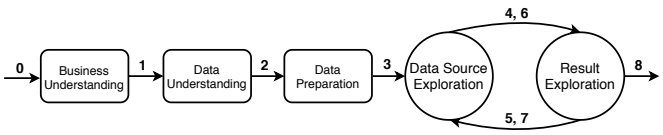

Fig. 8. Sales OLAP: Trajectory for the analysis of sales in retailing (Section 4.4). The first developments imply the preparation of the data mart, led by a data scientist that goes through the first activities of a data mining project: *business understanding*, choosing a process of interest; *data understanding*, identifying the needed data: what are the facts, the dimensions, their hierarchies?; and *data preparation* thus building the datamart. These activities are usually performed using the so-called ETL tools (Extract, Transform and Load) in data warehousing, helping in the progress of migrating and integrating data from the original data sources to the data warehouse. The second part of the trajectory involves possibly several analysts/managers extracting value from the datamart by getting the right data (*data source exploration*), and analysing the results (*result exploration*), and iterating loops until they come to decisions.

tools, can iteratively explore data and results, through typical drill-down and roll-up operations along the hierarchies in order to visualise key business issues. The trajectory consists of two main developments, the creation of a data warehouse, which can be assimilated to the first stages of CRISP-DM and a more explorative period at the end, as illustrated in Figure 8.

4.5 Repository publishing

Data publishing means curating data and making it available in a form that makes it easy for others to extract value. So data is both the starting point and the product. Some amount of data value exploration has happened as part of this process, but there is not a very concrete business goal (yet) for which the data is being made available. Some data mining has happened to support the data value exploration and data understanding process, but data publishing takes the place of deployment. In this way a data repository can be created serving as a data library for storing data sets that can be used for data analysis, sharing and reporting. Many examples of data repositories can be found through platforms such as *re3data*[10], which allows users to search among a vast number of different data repositories along the world by a simple or advance search using different characteristics. Another similar example is *paperswithcode.com*, a free resource for researchers and practitioners to find and follow the latest state-of-the-art machine learning-related papers and code. The company behind this (Atlas ML) has explored the way to present data regarding trending machine learning research, state-of-the-art leaderboards and the code to implement it. This way users could have access in a unified and genuinely comprehensive manner to papers (fetched from several venues, repositories and open source and free license related projects) and to its code on different repositories, which can help with reviewing content from different perspectives to discover and compare research. A possible trajectory for both examples is shown in Figure 9.

4.6 Parking App

Smart cities are an emergent concept that refers to an urban area that uses types of electronic data collection sensors

9. http://fortune.com/2016/01/11/car-insurance-companies-track/

10. https://www.re3data.org/

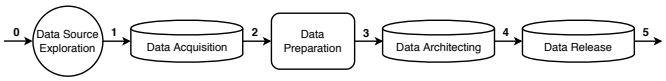

Fig. 9. Repository publishing: A possible trajectory for generating a data repository (Section 4.5) that might have been taken includes the activites of *data source exploration*, when data comes from external sources, and *data acquisition*, where the required data is downloaded, scraped and explored; *data preparation* where data is parsed, curated and structured; *data architecting*, where data is annotated, stored and managed in order to provide an easy access to the users; and *data release*, where both the data and the automatic data extraction pipelines are shared under different licenses for public use.

to supply information which is used to manage assets and resources efficiently. Smart cities technology allows to monitor what is happening in the city and to make decisions to improve the city evolution. Local governments and city councils usually realise that these real-time raw data collected (e.g., from citizens, sensors, devices, etc.) could be an important source for enhancing the quality of their living environment by improving the performance of urban services such as energy, transportation and utilities in order to reduce resource consumption, wastage and overall costs. For example, the open CityOS platform[11] is an Open source software that supports the visualisation of real time data and mobile applications of smart cities. This platform has been adopted by several smart cities projects. One of the developed applications is a smart parking app for the city of Dubrovnik (*Smart Parking Dubrovnik*) that allows drivers to find vacant parking spots, visualising them in an interactive map. In this example, a possible trajectory for this application is shown in Figure 10.

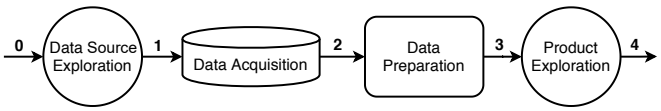

Fig. 10. Parking App: A possible trajectory for the development of the Smart Parking Dubrovnik app (Section 4.6). The first step is to determine what data should be acquired (*data source exploration*) and how to collect them (*data acquisition*), which may imply the development of specific sensors. Then, the following actions are performed in real time: the data gathered by the sensors are transformed to a format (*data preparation*) that allows to determine which parking spots are free and which ones are occupied. Finally, an app is developed for visualising the vacant parking spots in a map on the screen of users' mobiles (*product exploration*).

4.7 Payment geovisualisation

Credit card transactions are a rich source of data that banks and other payment platforms can exploit in many ways. BBVA, one major Spanish bank, through their Data & Analytics division, has been exploring several ways of making this data valuable. They realised that the historical information of what is bought by different people (nationalities) at different times and dates, and different locations could be an important source for monetisation, as many other companies (retailers, restaurants, etc.) could be interested in this information. They decided to create an interactive representation, so that users could learn about

Fig. 11. Payment geovisualisation: A possible trajectory for the tourism spending example (Section 4.7). This includes the steps of *data value exploration* where the bank systematically looked through the data it held; *goal exploration* where the bank considered the potential goals and chose to do an interactive website; *data preparation* where the data were integrated and prepared to be queried for visualisation; *result exploration* where the visualisations were analysed to decide which companies to offer particularised applications for; and *narrative exploration* where example stories were compiled in order to attract the audience to the visualisation tool.

the spending behaviour of tourists in Spain by having access across several variables to this information, with a general free demo application and particularised (or more detailed) applications for companies. The application, which reveals the data simply and clearly, was made attractive with stories such as: "Ever wondered when the French buy their food?", "Which places the Germans flock to on their holidays?", or "Sit back and discover the dynamics of spending in Spain"[12]. In this example, a possible trajectory that might have taken is shown in Figure 11.

5 Activity Types for Project Management

In the previous section we have seen a rich variety of data science trajectories. Some include the data mining process (in part or entirely) as a key component of the trajectory, but others mostly exclude it. We have even seen some cases where the conversion of data into knowledge by modelling or learning is not part of the process, but they are still considered genuine data science trajectories, as data is used to generate value. This ranges from projects only featuring the non-inferential part of "business intelligence" (e.g., building a data warehouse and obtaining aggregated tables, cubes and other graphical representations from the data) [50], but also those that follow more exploratory or interactive scenarios, such as those common in visual analytics [51].

Such variability across data science projects poses challenges for project managers, who need to hire suitable people and make time and cost estimates. Exploratory activities require expert data scientists and increase time and cost uncertainty, whereas data management activities require more data engineers and are more easily contained within a fixed time interval and budget. The DST model (see Figure 3) can help project planning by clearly separating exploratory, CRISP-DM (goal-directed) and data management activities, which each have different time and cost characteristics.

In order to better understand the nature of our seven illustrative examples from the previous section, Figure 12 shows a Venn diagram of the three kinds of activities. We can see which of the seven use cases in section 4 fall in each of the possible regions according to how relevant (in number or importance) the three kinds of activities are (details of the methodology to estimate this are given in the Appendix A). For example, for the Tourism recommender case (location-based services, with DST in Figure 5) both the exploratory and the CRISP-DM activities play an important role, and

11. http://cityos.io

12. The result can be found here: http://bbvatourism.vizzuality.com

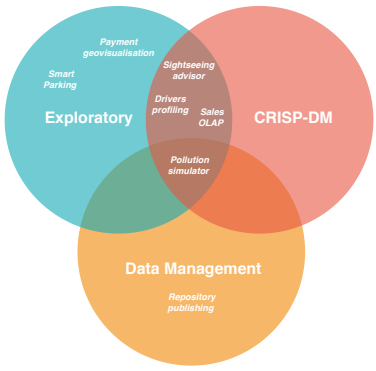

Fig. 12. Venn diagram of the three kinds of activities (*exploratory*, *CRISP-DM* and *data management*) and the seven use cases introduced in section 4.

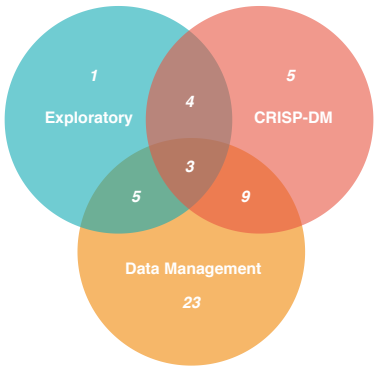

Fig. 13. Venn diagram of the three kinds of activities (*exploratory*, *CRISP-DM* and *data management*) and the number of use cases from NIST Big Data Public Working Group Use Cases [52] which fall into each region.

this is shown by their location in the Venn diagram. Overall, we see that most of the use cases are located in regions where exploration is important, as expected.

However, this picture should not be mistaken as representative of the whole range of data science applications, many of which may follow a more traditional CRISP-DM workflow or may give more relevance to data management. In section 2 we not only referred to polls that recognised CRISP-DM as the methodology that is still prevalent for data scientists (despite its limitations) but included a bibliographic survey covering the last past four years, with an important number of domains where CRISP-DM is still used extensively. All the applications reviewed there fit CRISP-DM well, with no or very little adaptation over the original formulation and including mostly CRISP-DM activities. This shows that CRISP-DM is still fit for purpose for one of the areas in Figure 12.

Apart from projects that fit in the CRISP-DM category, and those that are more explorative, it may be worth looking at some other projects that can have a stronger component in the data management part. In order to do this, we have examined the NIST Big Data Public Working Group Use Cases [52], as per their version 3.0. This is a very comprehensive set of 51 real use cases and their requirements gathered by the NBD-PWG Use Cases and Requirements Subgroup at the US National Institute for Standards and Technology (NIST). Following the approach used for our seven illustrative cases, we went through the 51 NIST cases. The first significant insight is that we did not find any activity that is not represented in Figure 3. This shows that our model is comprehensive, and captures a wide range of activities associated with any kind of data science project, including those that are more data-heavy. Also, when we look at the distribution of activities we also see clear patterns, which confirm what we already knew about the types of applications that are included in this NIST collection. In particular, Figure 13 shows a Venn diagram of the NIST cases and how many of them fall into each of the possible regions that emanate from the three kinds of activities.

Some further insights can be extracted from this diagram. Unsurprisingly, since this is a collection of Big Data projects, we find nearly half of them located in the Data Management (only) region. But there are also some other cases that are combined with the exploratory and/or CRISP-DM activities. Interestingly, even if this collection is about Big Data projects, we have at least one exemplar in each region.

This focus on the three kinds of activities and possible regions of overlap provides a useful characterisation of data science projects. Data science teams and their organisations can do a similar analysis of their projects and compare a new project specification against them. We recommend the following procedure: (1) Even at very early stages of a project, it is already possible to identify the activities that will be required. By analysing how many and how significant they are for each kind (exploratory, CRISP-DM or data management) it is possible to identify to which region of the Venn diagram they belong. If the project has one or more strong exploration components, it will be more open-ended. Consequently, more expert data scientists will be needed, with good knowledge about the domain and its casual models. Furthermore, planning will be more involved. If the project has a strong data management component, more data engineers will be needed, as well as more hardware and software resources. (2) By comparing to other projects of that region, one can estimate the project costs more accurately than by comparing against the whole collection of projects, and use some of the trajectories in that region as patterns or prototypes for the appropriate DST for the project. As a result, the types of activities in Figure 3 (exploratory, CRISP-DM and data management) are a practical, yet powerful, way of describing a data science project, prior to going into the more detailed flow of its trajectory, which can be useful for predictive and explanatory questions about the project.

6 Discussion

Standardised processes are not the same as methodologies [53], and many methodologies do not necessarily include guided processes, where one can follow a series of steps linearly. Two cases that are close to data science are quite illustrative. The first case is software engineering, which has many methodologies [54], and none of them seems to be the best methodology for all situations, depending on many internal and external factors. Software development, like many other engineering problems, has a structure that

resembles CRISP-DM in many ways (starting with business needs and ending up in deployment and maintenance of the outcome of the process), but it would be likewise inappropriate to use the same linear flow for all problems and circumstances. The similarities have suggested the application or adaptation of software development methodologies for data science (or big data) projects [55], but it is perhaps the general project management methodologies that may be more appropriate, or some specific ideas such as design patterns [56]. Also, we can learn from some novel lightweight methodologies, such as Extreme Programming (XP) [57], which attempted to add flexibility to the process, allowing teams to develop software, from requirements to deployment, in a a more efficient way.

The second case is methodology in science. The whole process of scientific discovery is usually question-driven, rather than data-driven or goal-driven, but is generally much more flexible in the initial trajectories (surprising observations, serendipity, etc.) – while more strict when it comes to hypothesis testing, replicable experimental design, etc. Despite the analogies between some trajectories in data science and the methodologies in science, there is an ongoing controversy whether the traditional scientific method is obsolete under the irruption of data science [58], [59], or whether data science methodologies should learn more from the general scientific method [60], [61].

In the absence of more rigid schemes, this diversity of methodologies and trajectories may create uncertainty for project management. This is mitigated by three important aspects of our DST model. First, we define trajectories over a well-defined collection of activities, which can be encapsulated and documented, similar to the original substages in CRISP-DM. DST thus allow data scientists to design their data science projects as well as explore new activities that could be added to or removed from their workflows. This is especially useful for teams, as they can agree and locate themselves (and subteams) in some of the subactivities of the trajectory. Secondly, existing trajectories can be used as templates so that new projects can use them as references. A new project may find the best match in the catalogue of trajectories rather than forcing it to fit a process model such as CRISP-DM that may not suit the project well and may cause planning difficulties and a bad estimation of effort (e.g., resources, costs, project expertise, completition plans, etc.). Actually, if the estimations of resources and costs using DST are more accurate than using CRISP-DM, this would be evidence for validity and usefulness in an organisation. Thirdly, trajectories can be mapped with project plans directly, assigning deadlines to transitions, and assigning personnel and budget to activities. Iterations on activities are explicit in the trajectories, which also allows for spiral models where subparts of the trajectory are iterated from small to big or until a given criterion is met (or a resource is exhausted).

All this paves the way to the introduction of proper data science project management methodologies, and the reuse of statistics and experiences from activities used in previous projects. Techniques from the area of workflow inference and management could also be applied to analyse trajectories [62], estimate costs and success rates, and extract patterns that fit a domain or organisation.

While the trajectory perspective may allow for a more systematic (and even automated) analysis at the process level, it is no surprise that the more flexible, less systematic, character of the new activities (exploration and data management) highlights the challenges for the automation of data science. For instance, while the automation of the modelling stage of CRISP-DM has been achieved to a large extent under the AutoML paradigm [63], [64], many other parts of CRISP-DM are still escaping automation, such as data wrangling or model deployment. Beyond data mining, many new competences have been identified as necessary for a data scientist, including both technical and non-technical skills, such as communicating results, leading a team, being creative, etc. [65], [66], [67], [68], and they are usually associated with the exploration activities. Data scientists are expected to cover a wide range of soft skills, such as being proactive, curious and inquisitive, being able to tell a story about the data and visualise the insights appropriately, and focus on traceability and trust. Most of the new explorative steps beyond CRISP-DM identified in this paper imply these soft skills and the use of business knowledge and vision that is far from the capabilities that AI provides today, and will be harder to automate in the years to come.

The trajectory model does not yet explicitly address all the ethical and legal issues around data science [69], an area that is becoming more relevant in data science over the previous data mining paradigm, even if problems such as fairness and privacy already existed for data mining. The increased relevance comes especially from the incentives behind many data science projects, which focus on the monetisation of the data, through the exploration of new data products. This usually implies the use of data for purposes that are different from those that created the data in the first place, such as social networks, digital assistants or wearable devices. The most relevant ethical issues will appear in the new activities: goal exploration, data source exploration, data value exploration, result exploration, product exploration, and data acquisition. These are also the parts of the trajectories where more senior data scientists will be involved, assuming higher awareness and training on ethical issues [70] than other more technical, less senior data scientists or team members.

The DST is also motivated by the causal approach to data science. In this case, it is not that much that new exploratory activites are needed, but new *data management* activities, required to *generate* data for the discovery of the causal structure: data acquisition and simulation. These are a series of activities that are becoming more and more relevant, as we have also seen in the large Big Data NIST repository and the associated trajectories that we explored in section 5.

In conclusion, CRISP-DM still plays an important role as a common framework for setting up and managing data mining projects. However, the world today is a very different place from the world in which CRISP-DM was conceived over two decades ago. In this paper we have argued that the shift from data mining to data science is not just terminological, but signifies an evolution towards a much wider range of approaches, in which the main value-adding component may be undetermined at the outset and needs to be discovered as part of the project. For

such exploratory projects the CRISP-DM framework will be too restrictive. We have proposed a new Data Science Trajectories (DST) framework which expands CRISP-DM by including exploratory activities such as *goal exploration*, *data source exploration* and *data value exploration*. Entry points into, trajectories through and exit points out of this richer set of data science steps can vary greatly among data science projects. We have illustrated this by means of a broad range of exemplar projects and the trajectories they embody.

Data science is still a young subject, with many open questions regarding its nature and methodology. While other authors approach these questions from a top-down perspective [71], what we have attempted here is more bottom-up, starting from something that is generally accepted to be productive in the data mining context, and investigating how it can be generalised to account for the much richer data science context. We hence see this as part of a larger, ongoing conversation and hope that the perspective offered here will be received as a positive contribution.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their comments, which motivated the analysis in Section 5. This material is based upon work supported by the EU (FEDER), and the Spanish MINECO under grant RTI2018-094403-B-C3, the Generalitat Valenciana PROMETEO/2019/098. F. Martínez-Plumed was also supported by INCIBE (Ayudas para la excelencia de los equipos de investigación avanzada en ciberseguridad), the European Commission (JRC) HUMAINT project (CT-EX2018D335821-101), and UPV (PAID-06-18). J. H-Orallo is also funded by an FLI grant RFP2-152.

REFERENCES

[1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 step-by-step data mining guide," 2000.

[2] O. Marbán, J. Segovia, E. Menasalvas, and C. Fernández-Baizán, "Toward data mining engineering: A software engineering approach," *Information systems*, vol. 34, no. 1, pp. 87–107, 2009.

[3] IBM, "Analytics solutions unified method," ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf, 2005.

[4] SAS, "Semma data mining methodology," http://www.sas.com/technologies/analytics/datamining/miner/semma.html, 2005.

[5] L. A. Kurgan and P. Musilek, "A survey of knowledge discovery and data mining process models," *The Knowledge Engineering Review*, vol. 21, no. 1, pp. 1–24, 2006.

[6] G. Mariscal, O. Marban, and C. Fernandez, "A survey of data mining and knowledge discovery process models and methodologies," *The Knowledge Engineering Review*, vol. 25, no. 02, pp. 137–166, 2010.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996.

[8] R. J. Brachman and T. Anand, "Advances in knowledge discovery and data mining," U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996, ch. The Process of Knowledge Discovery in Databases, pp. 37–57.

[9] C. Gertosio and A. Dussauchoy, "Knowledge discovery from industrial databases," *Journal of Intelligent Manufacturing*, vol. 15, no. 1, pp. 29–37, 2004.

[10] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi, *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc., 1998.

[11] A. G. Buchner, M. D. Mulvenna, S. S. Anand, and J. G. Hughes, "An internet-enabled knowledge discovery process," in *Proc. of the 9th Int. Database Conf., Hong Kong*, vol. 1999, 1999, pp. 13–27.

[12] H. A. Edelstein, *Introduction to data mining and knowledge discovery*. Two Crows, 1998.

[13] L. Cao, "Domain-driven data mining: Challenges and prospects," *IEEE Trans. on Knowledge and Data Engineering*, vol. 22, no. 6, pp. 755–769, 2010.

[14] C. Brunk, J. Kelly, and R. Kohavi, "Mineset: An integrated system for data mining." in *KDD*, 1997, pp. 135–138.

[15] A. Bernstein, F. Provost, and S. Hill, "Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification," *IEEE Trans. on knowledge and data engineering*, vol. 17, no. 4, pp. 503–518, 2005.

[16] M. J. Harry, "Six sigma: a breakthrough strategy for profitability," *Quality progress*, vol. 31, no. 5, p. 60, 1998.

[17] J. Debuse, B. de la Iglesia, C. Howard, and V. Rayward-Smith, "Building the kdd roadmap," in *Industrial Knowledge Management*. Springer, 2001, pp. 179–196.

[18] O. Niaksu, "CRISP data mining methodology extension for medical domain," *Baltic J. of Modern Computing*, vol. 3, no. 2, p. 92, 2015.

[19] D. Asamoah and R. Sharda, "Adapting CRISP-DM process for social network analytics: Application to healthcare," *21th Americas Conf. on Information Systems, Puerto Rico, 2015*, 2015.

[20] N. Njiru and E. Opiyo, "Clustering and visualizing the status of child health in kenya: A data mining approach." *International Journal of Social Science and Technology I*, 2018.

[21] N. Azadeh-Fard, F. M. Megahed, and F. Pakdil, "Variations of length of stay: a case study using control charts in the CRISP-DM framework," *International Journal of Six Sigma and Competitive Advantage*, vol. 11, no. 2-3, pp. 204–225, 2019.

[22] A. Dåderman and S. Rosander, "Evaluating frameworks for implementing machine learning in signal processing: A comparative study of CRISP-DM, semma and kdd," 2018.

[23] M. Rogalewicz and R. Sika, "Methodologies of knowledge discovery from data and data mining methods in mechanical engineering," *Management and Production Engineering Review*, vol. 7, no. 4, pp. 97–108, 2016.

[24] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model," *Procedia CIRP*, vol. 79, pp. 403–408, 2019.

[25] C. Barclay, A. Dennis, and J. Shepherd, "Application of the CRISP-DM model in predicting high school students' examination (csec/cxc) performance," *Knowledge Discovery Process and Methods to Enhance Organizational Performance*, p. 279, 2015.

[26] D. B. Fernández and S. Luján-Mora, "Uso de la metodología CRISP-DM para guiar el proceso de minería de datos en lms," in *Tecnología, innovación e investigación en los procesos de enseñanza-aprendizaje*. Octaedro, 2016, pp. 2385–2393.

[27] L. Almahadeen, M. Akkaya, and A. Sari, "Mining student data using CRISP-DM model," *International Journal of Computer Science and Information Security*, vol. 15, no. 2, p. 305, 2017.

[28] D. Oreski, I. Pihir, and M. Konecki, "CRISP-DM process model in educational setting," *Economic and Social Development: Book of Proceedings*, pp. 19–28, 2017.

[29] E. Espitia, A. F. Montilla *et al.*, "Applying CRISP-DM in a kdd process for the analysis of student attrition," in *Colombian Conference on Computing*. Springer, 2018, pp. 386–401.

[30] V. Tumelaire, E. Topan, and A. Wilbik, "Development of a repair cost calculation model for daf trucks nv using the CRISP-DM framework," Ph.D. dissertation, Master's thesis, Eindhoven University of Technology, 2015.

[31] F. Schäfer, C. Zeiselmair, J. Becker, and H. Otten, "Synthesizing CRISP-DM and quality management: A data mining approach for production processes," in *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*. IEEE, 2018, pp. 190–195.

[32] E. G. Nabati and K.-D. Thoben, "On applicability of big data analytics in the closed-loop product lifecycle: Integration of CRISP-DM standard," in *IFIP International Conference on Product Lifecycle Management*. Springer, 2016, pp. 457–467.

[33] H. Nagashima and Y. Kato, "Aprep-dm: a framework for automating the pre-processing of a sensor data analysis based on CRISP-DM," in *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2019, pp. 555–560.

[34] S. B. Gómez, M. C. Gómez, and J. B. Quintero, "Inteligencia de negocios aplicada al ecoturismo en colombia: Un caso de estudio

aplicando la metodología CRISP-DM," in *14th Iberian Conference on Information Systems and Technologies, CISTI 2019*. IEEE Computer Society, 2019, p. 8760802.

[35] R. Ganger, J. Coles, J. Ekstrum, T. Hanratty, E. Heilman, J. Boslaugh, and Z. Kendrick, "Application of data science within the army intelligence warfighting function: problem summary and key findings," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006. International Society for Optics and Photonics, 2019, p. 110060N.

[36] R. P. Bunker and F. Thabtah, "A machine learning framework for sport result prediction," *Applied computing and informatics*, 2017.

[37] R. Barros, A. Peres, F. Lorenzi, L. K. Wives, and E. H. da Silva Jaccottet, "Case law analysis with machine learning in brazilian court," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2018, pp. 857–868.

[38] K. J. Cios, A. Teresinska, S. Konieczna, J. Potocka, and S. Sharma, "A knowledge discovery approach to diagnosing myocardial perfusion," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 19, no. 4, pp. 17–25, 2000.

[39] K. J. Cios and L. A. Kurgan, "Trends in data mining and knowledge discovery," in *Advanced techniques in knowledge discovery and data mining*. Springer, 2005, pp. 1–26.

[40] S. Moyle and A. Jorge, "Ramsys-a methodology for supporting rapid remote collaborative data mining projects," in *ECML/PKDD 2001 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning: Internal SolEuNet Session*, 2001, pp. 20–31.

[41] F. Martínez-Plumed, L. C. Ochando, C. Ferri, P. A. Flach, J. Hernández-Orallo, M. Kull, N. Lachiche, and M. J. Ramírez-Quintana, "CASP-DM: context aware standard process for data mining," *CoRR*, vol. abs/1709.09003, 2017. [Online]. Available: http://arxiv.org/abs/1709.09003

[42] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[43] J. Rollins, "Why we need a methodology for data science," 2015. [Online]. Available: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IMW14824USEN

[44] R. B. Severtson, "What is the team data science process?" 2017. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

[45] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic Books, 2018.

[46] J. Pearl, "The seven tools of causal inference, with reflections on machine learning." *Commun. ACM*, vol. 62, no. 3, pp. 54–60, 2019.

[47] G. W. Imbens and D. B. Rubin, "Rubin causal model," *The new palgrave dictionary of economics*, pp. 1–10, 2017.

[48] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, pp. 2003–2030, Dec. 2006. [Online]. Available: http://dl.acm.org/citation.cfm?id=1248547.1248619

[49] M. A. Hernán, J. Hsu, and B. Healy, "A second chance to get causal inference right: A classification of data science tasks," *CHANCE*, vol. 32, no. 1, p. 42–49, Jan 2019. [Online]. Available: http://dx.doi.org/10.1080/09332480.2019.1579578

[50] S. Chaudhuri, U. Dayal, and V. Narasayya, "An overview of business intelligence technology," *Communications of the ACM*, vol. 54, no. 8, pp. 88–98, 2011.

[51] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual analytics: Definition, process, and challenges," in *Information visualization*. Springer, 2008, pp. 154–175.

[52] D. N. B. D. I. Framework, "NIST big data interoperability framework: Volume 3, use cases and general requirements," *NIST Special Publication*, vol. 1500, p. 344, 2019.

[53] J. Saltz, K. Crowston *et al.*, "Comparing data science project management methodologies via a controlled experiment," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[54] L. R. Vijayasarathy and C. W. Butler, "Choice of software development methodologies: Do organizational, project, and team characteristics matter?" *IEEE software*, vol. 33, no. 5, pp. 86–94, 2016.

[55] V. D. Kumar and P. Alencar, "Software engineering for big data projects: Domains, methodologies and gaps," in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 2886–2895.

[56] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design patterns: elements of reusable object-oriented software*. Pearson Education, 1995.

[57] K. Auer and R. Miller, *Extreme programming applied: playing to win*. Addison-Wesley Longman Publishing Co., Inc., 2001.

[58] C. Anderson, "The end of theory: The data deluge makes the scientific method obsolete," *Wired magazine*, vol. 16, no. 7, pp. 16–07, 2008.

[59] R. Kitchin, "Big data, new epistemologies and paradigm shifts," *Big Data & Society*, vol. 1, no. 1, p. 2053951714528481, 2014. [Online]. Available: https://doi.org/10.1177/2053951714528481

[60] S. Carrol and D. Goodstein, "Defining the scientific method," *Nat Methods*, vol. 6, p. 237, 2009.

[61] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2318–2331, 2017.

[62] W. Van Der Aalst, K. M. Van Hee, and K. van Hee, *Workflow management: models, methods, and systems*. MIT press, 2004.

[63] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-weka: Combined selection and hyperparameter optimization of classification algorithms," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 847–855.

[64] I. Guyon, L. Sun-Hosoya, M. Boullé, H. Escalante, S. Escalera, Z. Liu, D. Jajetic, B. Ray, M. Saeed, M. Sebag *et al.*, "Analysis of the automl challenge series 2015-2018," 2017.

[65] "8 Skills You Need to Be a Data Scientist," https://blog.udacity.com/2014/11/data-science-job-skills.html, Nov. 2014.

[66] V. Dhar, "Data science and prediction," *Communications of the ACM*, vol. 56, no. 12, pp. 64–73, 2013.

[67] M. Loukides, *What Is Data Science?* "O'Reilly Media, Inc.", Apr. 2011.

[68] E. Commission, "European e-Competence Framework," 2016. [Online]. Available: http://www.ecompetences.eu/

[69] M. Taddeo and L. Floridi, "Theme issue 'the ethical impact of data science'," 2016.

[70] S. Russell, S. Hauert, R. Altman, and M. Veloso, "Ethics of artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 415–416, 2015.

[71] L. Cao, "Data science: a comprehensive overview," *ACM Computing Surveys (CSUR)*, vol. 50, no. 3, p. 43, 2017.

[72] M. Ponsen, K. Tuyls, M. Kaisers, and J. Ramon, "An evolutionary game-theoretic analysis of poker strategies," *Entertainment Computing*, vol. 1, no. 1, pp. 39–45, 2009.

APPENDIX

In section 5 we portray summarised information about 51 use cases extracted from the NIST Big Data Public Working Group [52]. In this appendix we give more information about this source of cases and the methodology we used to process them. The National Institute of Standards and Technology (NIST) sought to establish relations among industry professionals to further the secure and effective adoption of Big Data and develop consensus on definitions, taxonomies, secure reference architectures, security and privacy, and, from these, a standards roadmap. With this aim, the NIST Big Data Public Working Group (NBD-PWG) was launched with extensive participation by industry, academia, and government. The results from this group are reported in the NIST Big Data Interoperability Framework series of volumes which, among definitions, taxonomies, requeriments, etc., contains a set of 51 original use cases gathered by the NBD-PWG Use Cases and Requirements Subgroup[13]. The report includes examples in the following broad areas: government operations (4 cases), commercial (8), defense (3), healthcare and life sciences (10), deep learning and social media (6), research (4), astronomy and physics (5), earth,

13. https://bigdatawg.nist.gov/show_InputDoc.php

environmental and polar sciences (10) and energy (1). For each use case, the report presents their requirements and challenges.

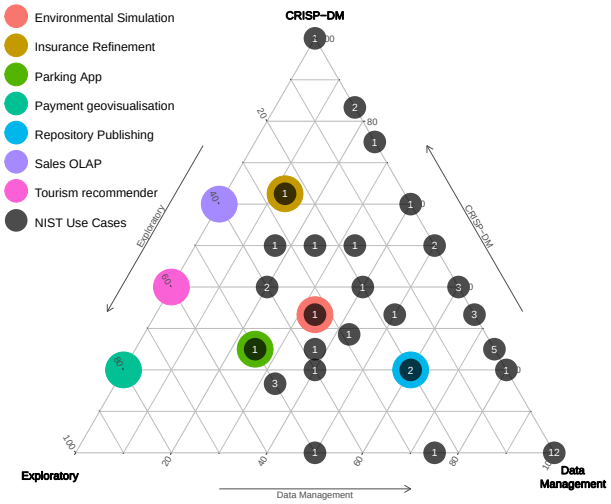

Fig. 14. Ternary plot depicting the proportions of the three activity types (*exploratory*, *CRISP-DM* and *data management*) for the seven use cases in section 4 and the 51 use cases from NIST Big Data Public Working Group Use Cases [52] (numbers show how many NIST use cases fall in the same point).

Aiming at better understanding the nature of these 51 uses cases, we classify them according to how relevant the three kinds of activities (exploratory, CRISP-DM and data management) are. In this regard, each use case is modelled as a DST following their definition from [52]. We then determine whether a case has a significant number of activities for each of the three groups of activities. We have three possible variables (i.e., type of activity) and 2^3 potential combinations ("application types") depending on how many activities of each type an use case involves. In this regard, we set a threshold to determine whether there is a significant use or not of a specific type of activity in terms of the number of activities used. Particularly, for the present study we set this threshold on minimum 2 activities. The results are those shown in Figures 12 and 13 in section 5

On the other hand, and in order to support the analysis performed in section 5, we have also analysed the percentage of the three types of activities as positions in a ternary plot (or *simplex* plot in game theory [72]) for all the illustrative examples from section 4 as well as the NIST use cases. This way, Figure 14 visualises the relative importance of the three activity types for each point (use case), where their positions in the plot represent their different compositions. Using percentages or ratios (instead of absolute numbers) here makes sense as there are no big differences in the number of activities involving each use case (i.e., they range from 3 to 7 activities, with 4.2 ± 1.3 activities on average).

The previous classifications show two things: (1) there is no case which has an activity that is not captured by our set of activities; (2) while our selection of illustrative examples in section 4 was made to emphasise the exploratory activities, which are more distinctive in the new conception of data science, the use cases in the NIST dataset are more related to data management.