Ten Years of DataMod: The Synergy of Data-Driven and Model-Based Approaches

Antonio Cerone(✉)

Department of Computer Science, School of Engineering and Digital Sciences, Nazarbayev University, Nur-Sultan, Kazakhstan
antonio.cerone@nu.edu.kz

Abstract. DataMod was founded in 2012 under the acronym of MoKMaSD (Modelling and Knowledge Management for Sustainable Development). The original aim of the Symposium, established by the United Nations University, was to focus on modelling and analysing complex systems while using knowledge management strategies, technology and systems to address problems of sustainable development in various domain areas. The focus was soon expanded to generally addressing the complementarity of model-based and data-driven approaches and the synergetic efforts that can lead to the successful combination of these two approaches. Hence the new name of the Symposium: "From Data to Models and Back" and the new acronym, which have been used since 2016.

In this paper we will start from the origins and history of DataMod and will look into the community formed around the Symposium in order to identify some research areas that have been addressed during its first 10 years of life. In this perspective, we will try to understand to which extent the two components of the DataMod community managed to integrate and which synergies between data-driven and model-based approaches have emerged. We will also have a look at what is happening outside DataMod and we will finally discuss which research and collaboration challenges should be addressed by future editions of DataMod.

Keywords: Data-driven approached · Model-based approaches · Formal modelling · Machine learning · Process mining

1 Introduction

Although DataMod (the *International Symposium "From Data to Models and Back"*) was officially founded in 2012 under the acronym of MoKMaSD (*Modelling and Knowledge Management for Sustainable Development*), the ideas underlying the symposium developed during the previous years and found their first

Work partly funded by Project SEDS2020004 "Analysis of cognitive properties of interactive systems using model checking", Nazarbayev University, Kazakhstan (Award number: 240919FD3916).

J. Bowles et al. (Eds.): DataMod 2021, LNCS 13268, pp. 7–24, 2022.
https://doi.org/10.1007/978-3-031-16011-0_2

implementation in a special track on "Modelling for Sustainable Development" at the 9th International Conference on Software Engineering and Formal Methods (SEFM 2011). The keynote talk by Matteo Pedercini, Millennium institute, and the four accepted papers of this special track, which were included in the SEFM 2011 proceedings [7], ranged over a number of application domains: agro-ecosystems, policy analysis, energy consumptions, and knowledge transfer process. Such application domains were explored by using agent-based systems, system dynamics and formal methods.

This first effort, initiated through a collaboration between the International Institute for Software Technologies of the United Nations University (UNU-IIST), which was later renamed the United Nations University Institute in Macau[1], and the Millennium Institute[2], aimed at the use of mathematical/formal modelling approaches and had a strong focus on sustainable development as an umbrella for the pool of considered application domains, in line with the common objectives of the two organising institutions.

However, during the discussion at the Conference, it was recognised that mathematical/formal modelling was not enough to fully characterise the application domains addressed by the special track contributions. The large quantity and heterogeneous quality of the information involved in the modelling process could not be entirely mathematically/formally represented and manipulated, but required a multidisciplinary approach integrating a variety of knowledge management techniques. Following these considerations, it was then decided to start a symposium series with a larger scope than the one of the special track.

The rest of the paper is structured as follows. Sects. 2 and 3 go through the history of the Symposium, with Sect. 2 devoted to the MoKMaSD events (2012–2015) and Sect. 3 devoted to the DataMod events (2016–2020). Section 4 compares and discusses the approaches, extends our overview with important, synergetic work that has been carried out outside DataMod, and and explores research challenges in the context of the scope of DataMod. Finally, Sect. 5 suggests possible directions and initiatives in the ambits of the DataMod community and DataMod future events.

2 Building up an Interdisciplinary Community and Elaborating on Its Scope: MoKMaSD 2012–2015

MoKMaSD 2012, the first edition of the Symposium, was held on 2 October 2012 in Thessaloniki, Greece [23]. The aim of the symposium, as highlighted in a short position paper [22],

> was to bring together practitioners and researchers from academia, industry, government and non-government organisations to present research results and exchange experience, ideas, and solutions for modelling and analysing complex systems and using knowledge management strategies,

[1] https://cs.unu.edu.
[2] https://www.millennium-institute.org.

technology and systems in various domain areas, including economy, governance, health, biology, ecology, climate and poverty reduction, that address problems of sustainable development.

As a result of the discussion at the SEFM 2011 special track, it was decided that formal modelling could not be the exclusive focus of the new symposium. Instead, there was an explicit attempts to put together two distinct communities, the modelling community and the knowledge management community, with the initial objective of exchanging experiences, presenting ideas and solutions to each other and trying to explore possible ways of collaboration and integration of the approaches. This first edition of MoKMaSD comprised a reasonable balance of modelling and knowledge management contributions.

On the modelling side the focus was on ecosystems. Corrado Priami's invited talk [39] introduced LIME, a modelling interface featuring the intuitive modelling of plant-pollinator systems and their automatic translation into stochastic programming languages equipped with analysis support aiming at providing an assessment of the functional dynamics of ecosystems. Barbuti *et al.* [6] presented a formalism, inspired by concepts of membrane computing and spatiality dynamics of cellular automata, for modelling population dynamics. Bernardo and D'Alessandro [8] analysed different policies for sustainability in the energy sector, through a dynamic simulation model aiming at evaluating the dynamics of different strategies in terms of their economic impact.

On the knowledge management side there were two contributions. Bolisani *et al.* [11] analysed the knowledge exchange that occurs in online social networks and revisited the literature available at that time in order to identify modelling and simulation approaches that may support the analysis process. Gong and Janssen [33] defined a framework for the semantic representation of legal knowledge, aiming at the automatic creation of business processes by selecting, composing and invoking semantic web services. In particular, these two works on knowledge management tried to look at ways for combining some forms of rigorous modelling with informal knowledge management techniques, thus fully addressing MoKMaSD scope and providing important ideas for the future directions of the Symposium.

However, since MoKMaSD originated from a modelling community, mostly interested in applications to biology and ecology, and was collocated with SEFM, a formal methods conference, during the next two editions [17,27] there were stable contributions from the original modelling community, with a large predominance of the applications of formal methods to the domains of biological systems and ecosystems.

MoKMaSD 2013 [27] featured four papers on ecosystem modelling and three papers on knowledge management. Two of the latter went beyond the integration of modelling in a knowledge management setting and considered a *linked data* perspective to combine data from different sources and facilitate data extension and management [16,52].

MoKMaSD 2014 [17] featured three papers on ecosystem modelling (including one position paper [25]), one paper on biological modelling in a

medical context [59], two papers on the application of data mining to the analysis of social networks [51] and to people's mobility flow [30], and one paper on the formal modelling of learning processes that are originated by collaboration activities within an open source software (OSS) community [47]. The application domains were expanded, with no longer an explicit focus on sustainable development. This was formalised by changing the symposium name to "Modelling and Knowledge Management applications: Systems and Domains" while preserving the acronym. The emphasis on data, which had already appeared in the 2013 symposium, was further strengthened in 2014. A *data mining* community joined the symposium contributing not only in terms of the new analysis techniques but also in terms of application domains: *social networks* [51] and *social contexts* [30,36,47]. These new injections gave the Symposium a very interdisciplinary flavour. And, actually, interdisciplinarity was explicitly recognised as a key research challenge by a contribution in the area of ecosystem modelling [25]. Moreover, a twofold role of the data in the modelling process started to appear. On the one hand, data can be used to calibrate the model, as seen in ecosystem modelling and biological modelling [59] and, on the other hand, it can contribute to the actual definition of the model itself [30,36,51].

MoKMaSD 2015 [9] represented an important milestone for the scope of the Symposium. The event programme featured two inspiring keynote talks: "Constraint Modelling and Solving for Data Mining" by Tias Guns and "Machine Learning Methods for Model Checking in Continuous Time Markov Chains" by Guido Sanguinetti. Tias Guns' talk reviewed motivations and advances in the use of constraint programming for data mining problems. In addition, in a contributed paper in the same area Grossi *et al.* presented expressive constraint programming models that support the extension of a given standard problem with further constraints, thus generating interesting variants of the problem [35]. Guido Sanguinetti's talk considered models with uncertain rates, which often occur in biology, and presented interesting ideas for using machine learning to synthesise those parameters that were to be quantified in order to run the model and use it to carry out reachability analysis and model checking.

During the discussion at the Symposium, it was explicitly recognised that modelling, even when considered in the large by incorporating formal or informal knowledge management techniques, was not enough to deal with the more and more increasing complexity of the system and size of the data. In fact, it was clear that, especially when dealing with big data, it is essential to consider real-data and compare them with the model prediction in order to validate the model and be able to use it in a practical context. As proposed in Sanguinetti's keynote talk, *machine learning* can provide a tool for making up for uncertainty and enriching incomplete models, thus, to some extent, allowing real-world data to drive the modelling process. A similar role can be played by *process mining* by extracting a descriptive process model from big data and use such a synthesised model in a number of analytical ways [19]. These synergetic aspects of modelling and data analysis paved the way for the new main scope of the Symposium.

3 Exploring Synergetic Approaches and Encouraging Interdisciplinary Collaboration: DataMod 2016–2020

DataMod 2016 [45] was the first edition using the new name of the Symposium: "From Data to Model and Back". Since then, the call for papers and the website have been clearly stating that the Symposium

> aims at bringing together practitioners and researchers from academia, industry and research institutions interested in the combined application of computational modelling methods with data-driven techniques from the areas of knowledge management, data mining and machine learning. Modelling methodologies of interest include automata, agents, Petri nets, process algebras and rewriting systems. Application domains include social systems, ecology, biology, medicine, smart cities, governance, education, software engineering, and any other field that deals with complex systems and large amounts of data. Papers can present research results in any of the themes of interest for the symposium as well as application experiences, tools and promising preliminary ideas. Papers dealing with synergistic approaches that integrate modelling and knowledge management/discovery or that exploit knowledge management/discovery to develop/synthesise system models are especially welcome.

Data mining and machine learning joined knowledge management as key areas on the data-driven side. And, most importantly, there was a new strong emphasis on synergistic approaches. Finally, as a follow-up of the Symposium, an open call for a special issue of the Journal of Intelligent Information Systems (JIIS) on "Computational modelling and data-driven techniques for systems analysis" was devoted to research results in any of the themes of interest of DataMod 2016 and the previous MoKMaSD editions [42].

The programme of DataMod 2016 started with a keynote talk titled "Mining Big (and Small) Mobile Data for Social Good", in which Mirco Musolesi explored challenges and opportunities in using big (and small) mobile data, which can be collected by means of applications running on smartphones or directly by mobile operators through their cellular infrastructure, within the modelling process and to solve systems-oriented issues. He also discussed the societal and commercial impact of this approach. Guidotti *et al.* showed that musical listening data from the large availability of rich and punctual online data sources can be exploited to create personal listening data models, which can provide higher levels of self-awareness as well as enable additional analysis and musical services both at personal and at collective level [37]. Inspired by process mining, Cerone proposed a technique, called *model mining*, to mine event logs to define a set of formal rules for generating the system behaviour. This can be achieved either by sifting an already existing *a priori* model by using the event logs to decrease the amount of non-determinism [20,21] or by inferring the rules directly from the events logs [21].

In a second keynote talk titled "The Topological Field Theory of Data: a program towards a novel strategy for data mining through data language",

Emanuela Merelli presented a topological field theory of data by making use of formal language theory in order to define the potential semantics needed to understand the emerging patterns that exist among data and that have been extracted within a mining context. Along the same lines Atienza *et al.* used *topological data analysis* to separate topological noise from topological features in high-dimensional data [3,4]. In a more practical perspective, Reijsbergen compared four approaches for generating the topology of stations in a bicycle-sharing systems and found out that a data-driven approach, in which a dataset of places of interest in the city is used to rate how attractive city areas are for station placement, outperformed the other three approaches [56].

DataMod 2017 [24] was the first two-day edition. It featured a tutorial titled "On DataMod approaches to systems analysis" by Paolo Milazzo and three keynote talks. The tutorial provided a taxonomy of approaches based on levels of knowledge of internal logic that is externalised by the system behavioral description mechanisms, ranging from purely data-driven approaches, which are essentially black boxes, to model-based, in which the internal logic is fully externalised. It then focused on approaches that combine different levels of knowledge, e.g. process mining, statistical model checking and applications of machine learning in formal verification.

Two keynote talks were on *smart cities*: "Understanding and rewiring cities using big data" by Bruno Lepri and "Exploring change planning in open, complex systems" by Siobhán Clarke". The third keynote talk was "Applications of weak behavioral metrics in probabilistic systems" by Simone Tini. Another special issue on "Computational modelling and data-driven techniques for systems analysis" was organised, this time in the Journal of Logical and Algebraic Methods in Programming (JLAMP) and restricted to revised and improved version of contributions accepted at DataMod 2017 [31].

Human cognition and *human behaviour* were interesting application areas that appeared in DataMod contributions for the first time in 2017. Broccia *et al.* proposed an algorithm for simulating the psychology of human selective attention, aiming at analysing how attention is divided during the simultaneous interaction with multiple devises, especially in the context of safety-critical systems [14]. Nasti and Milazzo took a neuroscience perspective in modelling human behaviour and proposed a hybrid automata model of the role that the dopamine system plays in addiction processes [49,50]. Finally, Carmichael and Morisset proposed the use of a methodical assessment of decision trees to predict the impact of human behaviour on the security of an organisation. In their approach, learning the behaviour from different sets of traces generated by a designed formal probabilistic model appears as an effective approach to building either a classifier from actual traces observed within the organisation or building a formal model, integrating known existing behavioural elements, which may both present high complexity, the former in selecting the best classifier and the latter in selecting the right parameters [18].

The 1st Informal Workshop on *DataMod Approaches to Systems Analysis* (**WDA 2018**), held on 5 February 2018 in Pisa, aimed at promoting the

development of a research community on DataMod approaches by bringing together researchers from the computational modeling and data science communities in order to let them discuss on potential collaborations in a friendly and informal context.

DataMod 2018 [43] included various contributions on *human-computer interaction* (HCI): the keynote talk titled "Data-driven analysis of user interface software in medical devices" by Paolo Masci, two contributed papers and two presentation-only contributions. The other two keynote talks were "Safe Composition of Software Services" by Gwen Salaün and "Computational oncology: from biomedical data to models and back" by Giulio Caravagna.

The two contributed papers on HCI showed a clear synergetic flavour. Cuculo *et al.* used automatic relevance determination to classify personality gaze patterns, but with the additional, important aim of showing how machine learning-based modelling could be used for gaining explanatory insights into relevant mechanism of the studied phenomenon [28]. This form of *explainable machine learning*, which is fundamental in the synergy of model-based and data-driven approaches, will be further discussed in Sect. 4.2. Cerone and Zhexenbayeva used the CSP process algebra, to define a compositional formal model of a language learning app, a user's profile and a formal representation of data from the real usage of the application. Such a formal model is then used in a model-checking framework that supports the validation of research hypotheses relating the learner profile to the user behaviour during interaction [26]. In this synergetic approach, a formal modelling and verification method is used in a data-driven way to carry out some form of validation.

DataMod 2019 [60] featured two keynote talks: "Verification of Data in Space and Time" by Mieke Massink and "Diagrammatic physical robot models in RoboSim" by Ana Cavalcanti . Massink presented spatial and spatio-temporal logics and their use within efficient model checking methods to investigate spatial aspects of data (which may be gathered in various domains ranging from smart public transportation to medical imaging). Cavalcanti presented RoboSim, a diagrammatic tool-independent domain-specific language to model robotic platforms and their controllers and automatically generate simulations and mathematical models for proof.

The Symposium also featured a number of synergetic contributed papers. Two HCI contributions build up on previous DataMod papers. Broccia *et al.* put their previous work on modelling selective attention [14] in a synergetic context by presenting its validation against data gathered from an experimental study performed with real users involved in a"main" task perceived as safety-critical, which was performed concurrently with a series of "distractor" tasks [13]. Based on Cerone and Zhexenbayeva work [26], Aibassova *et al.* presented a language learning application equipped with instrumentation code to gather data about user behavior and use such data for testing new forms of exercises and their combination on samples of users and, after converting raw data into a formal description, also for formal verification and validation purposes [1]. Garanina *et al.* proposed an approach to automatically extracts concurrent system requirements from the

technical documentation and formally verify the system design using an external or built-in verification tool. [32]. Bursic *et al.* proposed an automated approach to log anomaly detection. Their approach requires no hand-crafted features and no preprocessing of data and uses a two part unsupervised deep learning model, one applied to text for training without timestamp in order to learn a fixed dimensional embedding of the log messages, and the other applied to the text embeddings and the numerical timestamp of the message in order to detect anomalies [15].

DataMod 2020 represents an important milestone in the history of the Symposium. For the first time an LNCS volume was entirely devoted to the Symposium and titled after it [12]. This was an important step in officialising the already well-established identity of the Symposium. The keynote talk "Towards AI-driven Data Analysis and Fabrication" by Michael Vinov, IBM, presented two complementary methods for data management: a rule-based method that provides declarative language to model data logic and data rules in order to fabricate synthetic data using a constraint satisfaction programming solver, and machine-learning-based methods both for analysis of existing data and for creation of synthetic data. The latter methods themselves already match the DataMod philosophy to go "from data to model and back", but a combination with the former rule-based method could enrich machine-learning-based direction "from data to model" by adding knowledge of the behavioural logic in the spirit of explainable machine learning. In fact, a possible future combination of the two IBM approaches was discussed during the talk.

The Symposium featured 3 session: Machine Learning, Simulation-Based Approaches, and Data Mining and Processing Related Approaches. A number of approaches were presented in the area of healthcare, some of which obviously related with the COVID-19 pandemic.

Silvina *et al.* modelled and analysed the initial stage of the pathway of a cancer patient, from suspected diagnosis to confirmed diagnosis and start of a treatment (cancer waiting time). The approach is data-driven, in the sense that the structural information present in graphical data is exploited by aggregating information over connected nodes, thus allowing them to effectively capture relation information between data elements [61]. Bowles *et al.* presented an approach to evaluate the actual performance and quantify the capacity of an emergency department to support patient demand with limited resources in pre- and post-COVID-19 pandemic scenarios [?]. Milazzo considered the SIR model for spread of disease, which is based on the number of susceptible, infected and recovered individuals, and applied it to the COVID-19 pandemic to carry out stochastic analysis using the stochastic model checker PRISM. The SIR model is modified in order to include governmental restriction and prevention measures and make use of parameter estimation based on real epidemic data [44].

In addition to these simulation-based approaches, there were two more contributions in the area of healthcare. Rahman and Bowles proposed an approach to automatically infer the main components in clinical guideline sentences, using model checking to validate their correctness and combining them with the

information gained from real patient data and clinical practice in order to give more suitable personalised recommendations for treating patients [53]. Munbodh *et al.* developed a framework that, for a given standardised radiation treatment planning workflow, provides real-time monitoring and visualisation and supports informed, data-driven decisions regarding clinical workflow management and the impact of changes on the existing workflow, depending on optimisation of clinical efficiency and safety and new interventions incorporated into clinical practice. [57].

Finally, Barbon Junior *et al.* studied how the performance of process mining depends on the used encoding method [5], and Nasti *et al.*, building on work first presented at DataMod 2017 [49,50], showed that the current online social networks' notifications system triggers addictive behaviors [48].

4 Looking Around and Planning for the Future: DataMod Beyond 2021

In Sects. 2 and 3 we have gone through the history of DataMod and encountered essentially two kinds of model-driven approaches to system analysis, simulation and model checking, and a number of kinds of data-driven approaches, including machine learning, deep learning and data mining. All approaches have been used within several application domains: biology, ecology, medicine, healthcare, smart cities, IoT, social networks, human/social behaviour, human cognition and HCI.

Section 4.1 compares the three fundamental kinds of approaches aiming at analysing system dynamics: formal methods (model simulation and model checking), machine learning (including deep learning) and process mining. Section 4.2 summarises the synergetic work that has currently being carried out as well as some recent results and explore research challenges in the context of the scope of DataMod.

4.1 Comparing the Approaches

In our comparison we start considering model-based approaches and, in particular formal methods, which represent the most rigorous and mathematical approach to *define* a model of a real-world system. The key word in formal approaches is "*define*" and must be intended as *formal definition* or *formal specification*, aiming not just at describing the system but, more precisely, at providing a mechanistic way to generate the system behaviour. An important aspect of this system specification process is that the resultant system model also provides an explanation of the way the system work. Although this means that we need to start from our idea and vision of the real system, which represent a sort of *a priori* explanation driving our design process, the resultant model is not just a product of our creative effort but is largely affected by the real system we observe, its behaviour, the data it receives as an input and the way such data is transformed in the output. This role of the real-world data is always present in the design process, at least implicitly, and is sometime made explicit by specific

design approaches, such as iterative design, participatory design and, more in general, by all forms of agile design. However, the use of real-world data is normally an informal one and is carried out through an abstraction process. This also means that what we model is not a full representation of the real-system, but one of its possible abstract representations. Which abstraction we consider depends on what is the goal of our modelling effort, that is, on whether we aim at performing simulation and at which level of detail or we aim at proving properties and which kinds of properties.

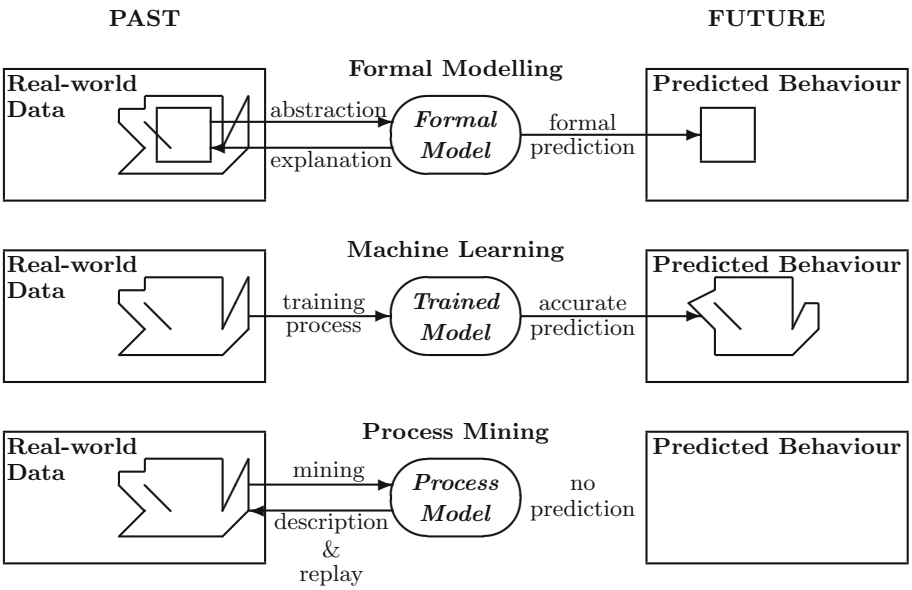

Fig. 1. Comparing Formal Modelling, Machine Learning and Process Mining

Therefore, as shown in the top part of Fig. 1, ***Formal Modelling*** starts from an informal use of real-world data (depicted at the top of Fig. 1 as the irregular shape on the left) to carry out an abstraction process that leads to the creation of a formal model, which provides an explanation of the relevant aspect of the past behaviour of the system (depicted at the top of Fig. 1 as the squared shape on the left) and supports the prediction of its future behaviour, still limited to the considered aspects (depicted at the top of Fig. 1 as the squared shape on the right).

As shown in the middle part of Fig. 1, ***Machine Learning*** processes real-world data trough a training process in order to learn the system behaviour and be able to carry out an accurate prediction of its future behaviour. The real-world data may comprise only input data (*unsupervised learning*), or also include the desired output (*supervised learning*), or may be explored in a dynamic environment trying to maximise the provided feedback, which is intended as a

sort of reward (*reinforcement learning*). The training process, which is essentially a generalisation process, aims at achieving an accurate prediction of the future behaviour (depicted in the middle of Fig. 1 as the irregular shape on the right, which is very similar to the one on the left). Obviously, since the training data is finite and the future is uncertain, there is no guarantee of the performance of the training process. Therefore, it is quite common to consider probabilistic bounds in order to quantify generalisation error.

Process Mining exploits real-world data by structuring datasets in a specific way in order define a *process model* that describes the system behaviour. Here the key point is to structure the data, which is expressed in terms of *event data*, also called *event logs*. An *event* is seen as consisting of at least three components: case, activity and timestamp. An *activity* names what has happened, a *case* correlates different events under the same identifier to describe an instance of the process, and a *timestamp* is the time when the event occurs, which also provides a partial ordering of the events. There are many ways for choosing what a case, an activity and a timestamp are, while structuring the event. Different choices lead to different perspectives in describing the system. But, in general, as shown in the bottom part of Fig. 1, with respect to the considered perspective, the model is described exactly as it is observed in the real-world (depicted at the bottom of Fig. 1 as the irregular shape on the left) and can be even replayed to reproduce the exact observed behaviour. Moreover, the description is formal, typically in some variant of Petri nets, and, although a full description tends to be very complex and appears as a spaghetti-like network, it is possible to lower the detail threshold (for example by removing the least frequent process instances) to make it understandable.

Each of these three fundamental approaches have pros and cons. The *pros* can be summarised as follows.

Formal Modelling Pros. Formal Modelling allows us to both explain the past and predict the future. Moreover, having a model that explains the behaviour, we can also explicitly modify the model to carry out some form of intervention or perform control on the future behaviour. And we actually understand what we are doing.

Machine Learning Pros. Trained models may reach a very high degree of accuracy and allow us to predict the future in very fine details. We can also aim at some form of control on the system, although this is normally somehow implicitly achieved by the training algorithm in some obscure way.

Process mining Pros. Process mining supports a formal description, which is virtually a perfect description of the past. Moreover, looking at the mainstream behaviour allows us to explain the past and identify deviations and bottlenecks.

The *cons* can be summarised as follows.

Formal Modelling Cons. A first problem is that by working on an abstract representation we have only a partial description of the real-world system that focuses on specific aspects but it is not accurate. A second problem

is the complexity of the analysis algorithms, which may cause state-space explosion or require too long a computational time.

Machine Learning Cons. Trained models are not in general explainable and this has two negative consequences. First, there are a number of contexts in which it is necessary to explain to customers and/or users how the prediction is attained. This is the case for safety critical systems, which require to prove that the used training algorithms are correct and the prediction accuracy covers the most critical cases, as well as for informing the customer on how the system carries out its tasks. Second, it is not possible to explicitly modify the model in order to carry out a deliberate intervention, since changes to the model can only implicitly occur through learning.

Process Mining Cons. Process models are descriptive, but in general not predictive. The full description is too large and complex to be understandable and usable and the mainstream behaviour is normally not sufficiently detailed to attain a correct prediction.

4.2 Current Research on Synergetic Approaches

In Sect. 3 we have seen a number of synergetic approaches presented and/or discussed at DataMod. Guido Sanguinetti's keynote talk at MokMaSD 2015 showed how to use machine learning to synthesise parameters that are needed to carry out reachability analysis and model checking, paving the way to the general use of machine learning for making up for uncertainty and enriching incomplete models. Then, starting with Mirco Musolesi's DataMod 2016 keynote talk, we have seen attempts to use real-world data to synthesise a formal model [18,21] and to enrich [53], improve [1,20] or validate [13] an existing system model. We have seen that data from documentation may be used to extract system requirements, which can then be fed to a formal verification tool [32]. And we have also seen that real-world data may be combined with a formal model in order to support the validation of research hypotheses [26].

As Milazzo discussed in the tutorial he presented at DataMod 2017, data-driven approaches are essentially black boxes whose internal logics is not visible. As we mentioned in Sect. 4.1, normally this is also the case for machine learning. However, among the several works on machine learning presented at DataMod events, Cuculo *et al.* showed how machine learning-based modelling could be used for gaining explanatory insights into relevant mechanisms of the studied phenomenon [28].

Explainable machine learning may be seen as an implementation of the *right to explanation* [29], whereby the user or the customer acquires knowledge about the internal logic of the system. Domain knowledge is a necessary prerequisite for effectively using machine learning to get scientific results but is not sufficient in order to understand how a specific model operates and the underlying reasons for the decisions made by the model. In fact, three further elements are necessary in order to gain scientific insights and discoveries from a machine learning algorithm: *transparency*, *interpretability* and *explainability* [58]. Transparency is achieved if the processes that extracts model parameters from training

data and generates labels from testing data can be described and motivated by the approach designer (what has been used and why). Interpretability refers to presentation of properties of the trained model making their meanings understandable to a human, that is, mapping the abstract concepts that define the meanings to the domain knowledge of the human (how we understand the training model). The concept of explainability, however, still belongs to a gray area, due to its intrinsic vagueness, especially in terms of completeness and degree of causality. Although a recent work [46] partitions explanatory questions into three classes, what-questions (e.g., What event happened?), how-questions, (e.g., How did that event happen?) and why-questions (e.g., Why did that event happen?), it is the actual machine-learning user's goal that must drive the choice of the appropriate questions. This means that even if transparency and interpretability are met, they can only lead to a satisfactory explanation if we ask the appropriate questions.

Therefore, the lack of a joint concept of explainability has resulted in the development of several alternative explainability techniques, each of them with a different emphasis and different advantages and disadvantages. Islam *et al.* carried out a survey of 137 recently published papers in the area of explainable artificial intelligence (another name for 'explainable machine learning') finding that most of the work is in the safety-critical domains worldwide, deep learning and ensemble models are the most exploited models, visual explanations are preferred by end-users and robust evaluation metrics are being developed to assess the quality of explanations [38]. However, there is a lack of work that aims at exploiting synergies between machine-learning-based and model-based approaches. One of the few exceptions is the work by Liao and Poggio, who suggested to convert a neural network to a symbolic description to gain interpretability and explainability [41]. They propose to use "objects/symbols" as a basic representational atom instead of the N-dimensional tensors traditionally used in "feature-oriented" deep learning. This supports the explicit representation of symbolic concepts thus achieving a form of "symbolic disentanglement" that makes properties interpretable. Although little explored so far, Liao and Poggio's proposal appears as a promising direction towards a higher explainability of machine learning models. In Sect. 3 we discussed another proposal along these lines: in his keynote talk at DataMod 2020, Michael Vinov presented IBM ideas for combining their complementary methods for data management, i.e., a rule-based method and a machine-learning-based method.

In general, we can say that the DataMod community represents the ideal ensemble to pursue the objective of exploiting synergies between machine-learning-based and model-based approaches. Working together towards this objective should be set as a priority goal for the future DataMod events and initiatives.

Finally, we would like to consider the large amount of open data made available in OSS repositories. These data repositories not only include code but also documentation, emails and other forms of communication, test cases, bug reports, feature proposals, etc. Traditional data mining techniques as well as

process mining have been used to analyse various aspects of OSS communities, project and the processes involving them, such as contribution patterns, learning and skill acquisition [19]. Machine learning and deep learning techniques have also been used for probabilistic learning of large code bases (called *big code*) from OSS repositories [2]. Here the objective is the exploitation of the information extracted from such existing code bases in order to provide statistically likely solutions to problems that are hard to solve with traditional formal analysis techniques (*statistical code modelling*). Examples of such problems are: program synthesis [40,55], code property prediction [54] and code deobfuscation [10]. The dual of statistical code modelling is *probabilistic programming*, which aims at deploying programming language concepts to facilitate the programming of new machine-learning algorithms [34]. Exploiting the big code available on OSS repositories through process mining and machine learning could be an other objective on which the DataMod community should join forces and focus in the future.

5 Conclusion and Future Initiatives

We have gone through the history of the DataMod symposium since its beginning in 2012 under the acronym MoKMaSD. We have highlighted some important works that have been contributed during these ten years of life of DataMod. In particular, we have considered those contributions that addressed synergies between data-driven and model-based approaches. We have also compared different approaches to modelling and put DataMod contributions in the context of the current research. Finally, we have identified two priority areas in which the heterogenous DataMod community has the potential to work successfully: explainable machine learning and the application of data-driven approaches to big code. We have proposed these areas as common objectives on which the DataMod community should join forces and focus in the future.

We would like to conclude that putting this proposals into practice requires the commitment of the DataMod community to invest time and resources into their concrete implementation. In particular, organisational and collaborative efforts are needed to create motivations and the appropriate context to enable us to productively work together. Time seems to be mature for a second informal workshop with a similar scope as WDA 2018, possibly with a more flexible, hybrid format (physical and virtual), which would encourage extensive participation. A number of further initiatives could be used to foster and focus collaboration, including the creation of working groups on specific research topics/challenges and/or addressing specific objectives, the preparation of one or more joint position paper(s) and the inclusion of talks from experts in the considered priority areas within the programme of the next WDA workshop.

Acknowledgments. We would like to thank the Program Co-chairs of DataMod and Paolo Milazzo for their comments during the presentation of this work at DataMod 2021. Paolo also provided important materials and pointers to further information, which were essential in the preparation of both the presentation and this paper.

References

1. Aibassova, A., Cerone, A., Tashkenbayev, M.: An instrumented mobile language learning application for the analysis of usability and learning. In: Sekerinski, E., et al. (eds.) FM 2019. LNCS, vol. 12232, pp. 170–185. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-54994-7_13
2. Allamanis, M., Barr, E.T., Devanbu, P., Sutton, C.: A survey of machine learning for big code and naturalness. ACM Comput. Surv. (CSUR) **51**(4), 81 (2018)
3. Atienza, N., Gonzalez-Diaz, R., Rucco, M.: Separating topological noise from features using persistent entropy. In: Milazzo, P., Varró, D., Wimmer, M. (eds.) STAF 2016. LNCS, vol. 9946, pp. 3–12. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50230-4_1
4. Atienza, N., Gonzalez-Diaz, R., Rucco, M.: Persistent entropy for separating topological features from noise in vietoris-rips complexes. J. Intell. Inf. Syst. **52**(3), 637–655 (2017). https://doi.org/10.1007/s10844-017-0473-4
5. Barbon Junior, S., Ceravolo, P., Damiani, E., Marques Tavares, G.: Evaluating trace encoding methods in process mining. In: Bowles, J., Broccia, G., Nanni, M. (eds.) DataMod 2020. LNCS, vol. 12611, pp. 174–189. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70650-0_11
6. Barbuti, R., Cerone, A., Maggiolo-Schettini, A., Milazzo, P., Setiawan, S.: Modelling population dynamics using grid systems. In: Cerone, A., et al. (eds.) SEFM 2012. LNCS, vol. 7991, pp. 172–189. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54338-8_14
7. Barthe, G., Pardo, A., Schneider, G. (eds.): SEFM 2011. LNCS, vol. 7041. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24690-6
8. Bernardo, G., D'Alessandro, S.: Transition to sustainability: Italian scenarios towards a low-carbon economy. In: Cerone, A., et al. (eds.) SEFM 2012. LNCS, vol. 7991, pp. 190–197. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54338-8_15
9. Bianculli, D., Calinescu, R., Rumpe, B. (eds.): SEFM 2015. LNCS, vol. 9509. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-49224-6
10. Bichsel, B., Raychev, V., Tsankov, P., Vechev, M.: Statistical deobfuscation of android applications. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM (2016)
11. Bernardo, G., D'Alessandro, S.: Transition to sustainability: Italian scenarios towards a low-carbon economy. In: Cerone, A., et al. (eds.) SEFM 2012. LNCS, vol. 7991, pp. 190–197. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54338-8_15
12. Bowles, J., Broccia, G., Nanni, M. (eds.): DataMod 2020. LNCS, vol. 12611. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70650-0
13. Broccia, G., Milazzo, P., Belviso, C., Montiel, C.B.: Validation of a simulation algorithm for safety-critical human multitasking. In: Sekerinski, E., et al. (eds.) FM 2019. LNCS, vol. 12232, pp. 99–113. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-54994-7_8
14. Broccia, G., Milazzo, P., Ölveczky, P.C.: An algorithm for simulating human selective attention. In: Cerone, A., Roveri, M. (eds.) SEFM 2017. LNCS, vol. 10729, pp. 48–55. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74781-1_4
15. Bursic, S., Cuculo, V., D'Amelio, A.: Anomaly detection from log files using unsupervised deep learning. In: Sekerinski, E., et al. (eds.) FM 2019. LNCS, vol. 12232, pp. 200–207. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-54994-7_15

16. Cáceres, P., Cuesta, C.E., Cavero, J.M., Vela, B., Sierra-Alonso, A.: Towards knowledge modeling for sustainable transport. In: Counsell, S., Núñez, M. (eds.) SEFM 2013. LNCS, vol. 8368, pp. 271–287. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05032-4_20
17. Cimatti, A., Sirjani, M. (eds.): SEFM 2017. LNCS, vol. 10469. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66197-1
18. Carmichael, P., Morisset, C.: Learning decision trees from synthetic data models for human security behaviour. In: Cerone, A., Roveri, M. (eds.) SEFM 2017. LNCS, vol. 10729, pp. 56–71. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74781-1_5
19. Cerone, A.: Process mining as a modelling tool: beyond the domain of business process management. In: Bianculli, D., Calinescu, R., Rumpe, B. (eds.) SEFM 2015. LNCS, vol. 9509, pp. 139–144. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-49224-6_12
20. Cerone, A.: Refinement mining: using data to sift plausible models. In: Milazzo, P., Varró, D., Wimmer, M. (eds.) STAF 2016. LNCS, vol. 9946, pp. 26–41. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50230-4_3
21. Cerone, A.: Model mining. J. Intell. Inf. Syst. **52**(3), 501–532 (2017). https://doi.org/10.1007/s10844-017-0474-3
22. Cerone, A., Garcia-Perez, A.: Modelling and knowledge management for sustainable development. In: Cerone, A., et al. (eds.) SEFM 2012. LNCS, vol. 7991, pp. 149–153. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54338-8_12
23. Cerone, A., et al. (eds.): SEFM 2012. LNCS, vol. 7991. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54338-8
24. Cerone, A., Roveri, M. (eds.): SEFM 2017. LNCS, vol. 10729. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74781-1
25. Cerone, A., Scotti, M.: Research challenges in modelling ecosystems. In: Canal, C., Idani, A. (eds.) SEFM 2014. LNCS, vol. 8938, pp. 276–293. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15201-1_18
26. Cerone, A., Zhexenbayeva, A.: Using formal methods to validate research hypotheses: the Duolingo case study. In: Mazzara, M., Ober, I., Salaün, G. (eds.) STAF 2018. LNCS, vol. 11176, pp. 163–170. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04771-9_13
27. Counsell, S., Núñez, M. (eds.): SEFM 2013. LNCS, vol. 8368. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05032-4
28. Cuculo, V., D'Amelio, A., Lanzarotti, R., Boccignone, G.: Personality gaze patterns unveiled via automatic relevance determination. In: Mazzara, M., Ober, I., Salaün, G. (eds.) STAF 2018. LNCS, vol. 11176, pp. 171–184. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04771-9_14
29. Edwards, L., Veale, M.: Slave to the algorithm? why a 'right to an explanation' is probably not the remedy you are looking for. Duke Law Technol. Rev. **16**(1), 18–84 (2017)
30. Gabrielli, L., Furletti, B., Giannotti, F., Nanni, M., Rinzivillo, S.: Use of mobile phone data to estimate visitors mobility flows. In: Canal, C., Idani, A. (eds.) SEFM 2014. LNCS, vol. 8938, pp. 214–226. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15201-1_14
31. Galpin, V., Milazzo, P., Monreale, A.: Guest editors' foreword. J. Logic. Algebraic Methods Program. **109**, 1–2 (2019)

32. Garanina, N., Anureev, I., Sidorova, E., Koznov, D., Zyubin, V., Gorlatch, S.: An ontology-based approach to support formal verification of concurrent systems. In: Sekerinski, E., et al. (eds.) FM 2019. LNCS, vol. 12232, pp. 114–130. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-54994-7_9
33. Gong, Y., Janssen, M.: A framework for translating legal knowledge into administrative processes: dynamic adaption of business processes. In: Cerone, A., et al. (eds.) SEFM 2012. LNCS, vol. 7991, pp. 204–211. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54338-8_17
34. Gordon, A.D., Henzinger, T.A., Nori, A.V., Rajamani, S.K.: Probabilistic programming. In: FOSE 2014: Future of Software Engineering Proceedings, pp. 167–181. ACM (2014)
35. Grossi, V., Monreale, A., Nanni, M., Pedreschi, D., Turini, F.: Clustering formulation using constraint optimization. In: Bianculli, D., Calinescu, R., Rumpe, B. (eds.) SEFM 2015. LNCS, vol. 9509, pp. 93–107. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-49224-6_9
36. Guidotti, R., Monreale, A., Rinzivillo, S., Pedreschi, D., Giannotti, F.: Retrieving points of interest from human systematic movements. In: Canal, C., Idani, A. (eds.) SEFM 2014. LNCS, vol. 8938, pp. 294–308. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15201-1_19
37. Guidotti, R., Rossetti, G., Pedreschi, D.: Audio ergo sum. In: Milazzo, P., Varró, D., Wimmer, M. (eds.) STAF 2016. LNCS, vol. 9946, pp. 51–66. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50230-4_5
38. Islam, M.R., Ahmed, M.U., Barua, S., Begum, S.: A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Appl. Sci. **12**(3), 1353–1391 (2022)
39. Kahramanoğulları, O., Lynch, J.F., Priami, C.: Algorithmic systems ecology: experiments on multiple interaction types and patches. In: Cerone, A., et al. (eds.) SEFM 2012. LNCS, vol. 7991, pp. 154–171. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54338-8_13
40. Liang, P., Jordan, M.I.: Learning programs: A hierarchical Bayesian approach. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pp. 639–646. ACM (2010)
41. Liao, Q., Poggio, T.: Object-oriented deep learning. CBMM Memo No. 70, Center for Brains, Minds, and Machines, McGovern Institute for Brain Research, MIT. (2017)
42. Matwin, S., Tesei, L., Trasarti, R.: Computational modelling and data-driven techniques for systems analysis. J. Intell. Inf. Syst. **52**(3), 473–475 (2019). https://doi.org/10.1007/s10844-019-00554-z
43. Mazzara, M., Ober, I., Salaün, G. (eds.): STAF 2018. LNCS, vol. 11176. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04771-9
44. Milazzo, P.: Analysis of COVID-19 Data with PRISM: parameter estimation and SIR modelling. In: Bowles, J., Broccia, G., Nanni, M. (eds.) DataMod 2020. LNCS, vol. 12611, pp. 123–133. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70650-0_8
45. Milazzo, P., Varró, D., Wimmer, M. (eds.): STAF 2016. LNCS, vol. 9946. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50230-4
46. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
47. Mukala, P., Cerone, A., Turini, F.: An abstract state machine (ASM) representation of learning process in FLOSS communities. In: Canal, C., Idani, A. (eds.) SEFM

2014. LNCS, vol. 8938, pp. 227–242. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15201-1_15
48. Nasti, L., Michienzi, A., Guidi, B.: Discovering the impact of notifications on social network addiction. In: Bowles, J., Broccia, G., Nanni, M. (eds.) DataMod 2020. LNCS, vol. 12611, pp. 72–86. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70650-0_5
49. Nasti, L., Milazzo, P.: A computational model of internet addiction phenomena in social networks. In: Cerone, A., Roveri, M. (eds.) SEFM 2017. LNCS, vol. 10729, pp. 86–100. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74781-1_7
50. Nasti, L., Milazzo, P.: A hybrid automata model of social networking addiction. J. Logic. Algebraic Methods Program. **100**, 215–229 (2018)
51. Nozza, D., Maccagnola, D., Guigue, V., Messina, E., Gallinari, P.: A latent representation model for sentiment analysis in heterogeneous social networks. In: Canal, C., Idani, A. (eds.) SEFM 2014. LNCS, vol. 8938, pp. 201–213. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15201-1_13
52. Perez, A., Larrinaga, F., Curry, E.: The role of linked data and semantic-technologies for sustainability idea management. In: Counsell, S., Núñez, M. (eds.) SEFM 2013. LNCS, vol. 8368, pp. 306–312. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05032-4_22
53. Rahman, F., Bowles, J.: Semantic annotations in clinical guidelines. In: Bowles, J., Broccia, G., Nanni, M. (eds.) DataMod 2020. LNCS, vol. 12611, pp. 190–205. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70650-0_12
54. Raychev, V., Vechev, M., Krause, A.: Predicting program properties from "big code". In: Proceedings of the 42nd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL 2015). ACM (2015)
55. Raychev, V., Vechev, M., Yahav, E.: Code completion with statistical language models. In: Proc. of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 419–428. ACM (2014)
56. Reijsbergen, D.: Probabilistic modelling of station locations in bicycle-sharing systems. In: Milazzo, P., Varró, D., Wimmer, M. (eds.) STAF 2016. LNCS, vol. 9946, pp. 83–97. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50230-4_7
57. Munbodh, R., Leonard, K.L., Klein, E.E.: Deriving performance measures of workflow in radiation therapy from real-time data. In: Bowles, J., Broccia, G., Nanni, M. (eds.) DataMod 2020. LNCS, vol. 12611, pp. 206–216. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70650-0_13
58. Roscher, R., Bohn, B., Duarte, M.F., Garcke, J.: Explainable machine learning for scientific insights and discoveries. IEEE Access **8**, 42200–42216 (2020)
59. Sameen, S., Barbuti, R., Milazzo, P., Cerone, A.: A mathematical model for assessing KRAS mutation effect on monoclonal antibody treatment of colorectal cancer. In: Canal, C., Idani, A. (eds.) SEFM 2014. LNCS, vol. 8938, pp. 243–258. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15201-1_16
60. Sekerinski, E., et al. (eds.): FM 2019. LNCS, vol. 12232. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-54994-7
61. Silvina, A., Redeker, G., Webber, T., Bowles, J.: A simulation-based approach for the behavioural analysis of cancer pathways. In: Bowles, J., Broccia, G., Nanni, M. (eds.) DataMod 2020. LNCS, vol. 12611, pp. 57–71. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70650-0_4