

Data Mining Design and Systematic Modelling

© Co Yannic Kropp

Christian Albrechts University Kiel, Department of Computer Science, D-24098 Kiel, Germany
yk@is.informatik.uni-kiel.de

© Bernhard Thalheim

thalheim@is.informatik.uni-kiel.de

Abstract. Data mining is currently a well-established technique and supported by many algorithms. It is dependent on the data on hand, on properties of the algorithms, on the technology developed so far, and on the expectations and limits to be applied. It must be thus matured, predictable, optimisable, evolving, adaptable and well-founded similar to mathematics and SPICE/CMM-based software engineering. Data mining must therefore be systematic if the results have to be fit to its purpose. One basis of this systematic approach is model management and model reasoning. We claim that systematic data mining is nothing else than systematic modelling. The main notion is the notion of the model in a variety of forms, abstraction and associations among models.

Keywords: data mining, modelling, models, framework, deep model, normal model, modelling matrix

1 Introduction

Data mining and analysis is nowadays well-understood from the algorithms side. There are thousands of algorithms that have been proposed. The number of success stories is overwhelming and has caused the big data hype. At the same time, brute-force application of algorithms is still the standard. Nowadays data analysis and data mining algorithms are still taken for granted. They transform data sets and hypotheses into conclusions. For instance, cluster algorithms check on given data sets and for a clustering requirements portfolio whether this portfolio can be supported and provide as a set of clusters in the positive case as an output. The Hopkins index is one of the criteria that allow to judge whether clusters exist within a data set. A systematic approach to data mining has already been proposed in [3, 17]. It is based on mathematics and mathematical statistics and thus able to handle errors, biases and configuration of data mining as well. Our experience in large data mining projects in archaeology, ecology, climate research, medical research etc. has however shown that ad-hoc and brute-force mining is still the main approach. The results are taken for granted and believed despite the modelling, understanding, flow of work and data handling pitfalls. So, the results often become dubious.

Data are the main source for information in data mining and analysis. Their quality properties have been neglected for a long time. At the same time, modern data management allows to handle these problems. In [16] we compare the critical findings or pitfalls of [21] with resolution techniques that can be applied to overcome the crucial pitfalls of data mining in environmental sciences reported there. The algorithms themselves are another source of pitfalls that are

typically used for the solution of data mining and analysis tasks. It is neglected that an algorithm also has an application area, application restrictions, data requirements, results at certain granularity and precision. These problems must be systematically tackled if we want to rely on the results of mining and analysis. Otherwise analysis may become misleading, biased, or not possible. Therefore, we explicitly treat properties of mining and analysis. A similar observation can be made for data handling.

Data mining is often considered to be a separate sub-discipline of computer engineering and science. The statistics basis of data mining is well accepted. We typically start with a general (or better generic) model and use for refinement or improvement of the model the data that are on hand and that seem to be appropriate. This technique is known in sciences under several names such as inverse modelling, generic modelling, pattern-based reasoning, (inductive) learning, universal application, and systematic modelling.

Data mining is typically not only based on one model but rather on a model ensemble or model suite. The association among models in a model suite is explicitly specified. These associations provide an explicit form via model suites. Reasoning techniques combine methods from logics (deductive, inductive, abductive, counter-inductive, etc.), from artificial intelligence (hypothetic, qualitative, concept-based, adductive, etc.), computational methods (algorithmics [6], topology, geometry, reduction, etc.), and cognition (problem representation and solving, causal reasoning, etc.).

These choices and handling approaches need a systematic underpinning. Techniques from artificial intelligence, statistics, and engineering are bundled within the CRISP framework (e.g. [3]). They can be enhanced by techniques that have originally been developed for modelling, for design science, business informatics, learning theory, action theory etc.

We combine and generalize the CRISP, heuristics, modelling theory, design science, business informatics,

statistics, and learning approaches in this paper. First, we introduce our notion of the model. Next we show how data mining can be designed. We apply this investigation to systematic modelling and later to systematic data mining. It is our goal to develop a holistic and systematic framework for data mining and analysis. Many issues are left out of the scope of this paper such as a literature review, a formal introduction of the approach, and a detailed discussion of data mining application cases.

2 Models and Modelling

Models are principle instruments in mathematics, data analysis, modern computer engineering (CE), teaching any kind of computer technology, and also modern computer science (CS). They are built, applied, revised and manufactured in many CE&CS sub-disciplines in a large variety of application cases with different purposes and context for different communities of practice. It is now well understood that models are something different from theories. They are often intuitive, visualizable, and ideally capture the essence of an understanding within some community of practice and some context. At the same time, they are limited in scope, context and the applicability.

2.1 The Notion of the Model

There is however a general notion of a model and of a conception of the model:

A **model** is a well-formed, adequate, and dependable instrument that represents origins [9, 29, 30].

Its criteria of well-formedness, adequacy, and dependability must be commonly accepted by its community of practice within some context and correspond to the functions that a model fulfills in utilization scenarios.

A well-formed instrument is *adequate* for a collection of origins if it is *analogous* to the origins to be represented according to some analogy criterion, it is more *focused* (e.g. simpler, truncated, more abstract or reduced) than the origins being modelled, and it sufficiently satisfies its *purpose*.

Well-formedness enables an instrument to be *justified* by an *empirical corroboration* according to its objectives, by *rational coherence* and *conformity* explicitly stated through conformity formulas or statements, by *falsifiability* or *validation*, and by *stability* and *plasticity* within a collection of origins.

The instrument is *sufficient* by its *quality* characterization for internal quality, external quality and quality in use or through quality characteristics [28] such as correctness, generality, usefulness, comprehensibility, parsimony, robustness, novelty etc. Sufficiency is typically combined with some *assurance evaluation* (tolerance, modality, confidence, and restrictions).

2.2 Generic and Specific Models

The general notion of a model covers all aspects of

adequateness, dependability, well-formedness, scenario, functions and purposes, backgrounds (grounding and basis), and outer directives (context and community of practice). It covers all known so far notions in agriculture, archaeology, arts, biology, chemistry, computer science, economics, electro-technics, environmental sciences, farming, geosciences, historical sciences, languages, mathematics, medicine, ocean sciences, pedagogical science, philosophy, physics, political sciences, sociology, and sports. The models used in these disciplines are instruments used in certain scenarios.

Sciences distinguish between general, particular and specific things. Particular things are specific for general things and general for specific things. The same abstraction may be used for modelling. We may start with a general model. So far, nobody knows how to define general models for most utilization scenarios. Models *function* as *instruments* or tools. Typically, instruments come in a variety of forms and fulfill many different functions. Instruments are partially independent or autonomous of the thing they operate on. Models are however special instruments. They are used with a specific intention within a utilization scenario. The quality of a model becomes apparent in the context of this scenario.

It might thus be better to start with generic models. A **generic model** [4, 26, 31, 32] is a model which broadly satisfies the purpose and broadly functions in the given utilization scenario. It is later tailored to suit the particular purpose and function. It generally represents origins of interest, provides means to establish adequacy and dependability of the model, and establishes focus and scope of the model. Generic models should satisfy at least five properties: (i) they must be accurate; (ii) the quality of generic models allows that they are used consciously; (iii) they should be descriptive, not evaluative; (iv) they should be flexible so that they can be modified from time to time; (v) they can be used as a first “best guess”.

2.3 Model Suites

Most disciplines integrate a variety of models or a *society of models*, e.g. [7, 14] Models used in CE&CS are mainly at the same level of abstraction. It is already well-known for threescore years that they form a *model ensemble* (e.g. [10, 23]) or horizontal *model suite* (e.g. [8, 27]). Developed models vary in their scopes, aspects, and facets they represent and their abstraction.

A **model suite** consists of a set of models $\{M_1, \dots, M_n\}$, of an association or collaboration schema among the models, of controllers that maintain consistency or coherence of the model suite, of application schemata for explicit maintenance and evolution of the model suite, and of tracers for the establishment of the coherence.

Multi-modelling [11, 19, 24] became a culture in CE&CS. Maintenance of coherence, co-evolution, and consistency among models has become a bottleneck in development. Moreover, different languages with

different capabilities have become an obstacle similar to multi-language retrieval [20] and impedance mismatches. Models are often loosely coupled. Their dependence and relationship is often not explicitly expressed. This problem becomes more complex if models are used for different purposes such as construction of systems, verification, optimization, explanation, and documentation.

2.4 Stepwise Refinement of Models

Refinement of a model to a particular or special model provides mechanisms for model transformation along the adequacy, the justification and the sufficiency of a model. Refinement is based on *specialization* for better suitability of a model, on *removal* of unessential elements, on *combination* of models to provide a more holistic view, on *integration* that is based on binding of model components to other components and on *enhancement* that typically improves a model to become more adequate or dependable.

Control of correctness of refinement [33] for information systems takes into account (A) a focus on the refined structure and refined vocabulary, (B) a focus to information systems structures of interest, (C) abstract information systems computation segments, (D) a description of database segments of interest, and (E) an equivalence relation among those data of interest.

2.5 Deep Models and the Modelling Matrix

Model development is typically based on an explicit and rather quick description of the ‘surface’ or *normal model* and on the mostly unconditional acceptance of a *deep model*. The latter one directs the modelling process and the surface or normal model. Modelling itself is often understood as development and design of the normal model. The deep model is taken for granted and accepted for a number of normal models.

The deep model can be understood as the common basis for a number of models. It consists of the grounding for modelling (paradigms, postulates, restrictions, theories, culture, foundations, conventions, authorities), the outer directives (context and community of practice), and basis (assumptions, general concept space, practices, language as carrier, thought community and thought style, methodology, pattern, routines, commonsense) of modelling. It uses a collection of undisputable elements of the background as grounding and additionally a disputable and adjustable basis which is commonly accepted in the given context by the community of practice. Education on modelling starts, for instance, directly with the deep model. In this case, the deep model has to be accepted and is thus hidden and latent.

A (modelling) matrix is something within or from which something else originates, develops, or takes from. The matrix is assumed to be correct for normal models. It consists of the deep model and the modelling scenarios. The modelling *agenda* is derived from the modelling scenario and the utilization scenarios. The modelling scenario and the deep model serve as a part

of the *definitional frame* within a model development process. They define also the capacity and potential of a model whenever it is utilized.

Deep models and the modelling matrix also define some frame for adequacy and dependability. This frame is enhanced for specific normal models. It is then used for a statement in which cases a normal model represents the origins under consideration.

2.6 Deep Models and Matrices in Archaeology

Let us consider an application case. The CRC 1266¹ “*Scales of Transformation – Human Environmental Interaction in Prehistoric and Archaic Societies*”

investigates processes of transformation from 15,000 BCE to 1 BCE, including crisis and collapse, on different scales and dimensions, and as involving different types of groups, societies, and social formations. It is based on the matrix and a deep model as sketched in Figure 1. This matrix determines which normal models can still be considered and which not. The initial model for any normal model accepts this matrix.

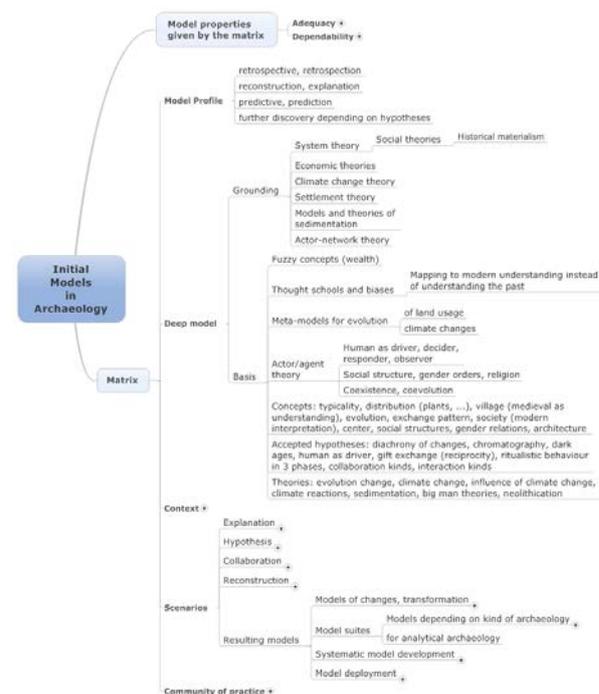


Figure 1 Modeling in archaeology with a matrix

We base our consideration on the matrix and the deep model on [19] and the discussions in the CRC. Whether the deep model or the model matrix is appropriate has already been discussed. The final version presented in this paper illustrates our understanding.

¹ <https://www.sfb1266.uni-kiel.de/en>

2.7 Stereotyping of a Data Mining Process

Typical modeling (and data mining) processes follow some kind of ritual or typical guideline, i.e. they are stereotyped. The *stereotype* of a modelling process is based on a general modelling situation. Most modelling methodologies are bound to one stereotype and one kind of model within one model utilization scenario. Stereotypes are governing, conditioning, steering and guiding the model development. They determine the model kind, the background and way of modelling activities. They persuade the activities of modelling. They provide a means for considering the economics of modelling. Often, stereotypes use a definitional frame that primes and orients the processes and that considers the community of practice or actors within the model development and utilization processes, the deep model or the matrix with its specific language and model basis, and the agenda for model development. It might be enhanced by initial models which are derived from generic models in accordance to the matrix.

The model utilization scenario determines the function that a model might have and therefore also the goals and purposes of a model.

2.8 The Agenda

The agenda is something like a guideline for modeling activities and for model associations within a model suite. It improves the quality of model outcomes by spending some effort to decide what and how much reasoning to do as opposed to what activities to do. It balances resources between the data-level actions and the reasoning actions. E.g. [17] uses an agent approach with preparation agents, exploration agents, descriptive agents, and predictive agents. The agenda for a model suite uses thus decisions points that require agenda control according to performance and resource considerations. This understanding supports introspective monitoring about performance for the data mining process, coordinated control of the entire mining process, and coordinated refinement of the models. Such kind of control is already necessary due to the problem space, the limitations of resources, and the amount of uncertainty in knowledge, concepts, data, and the environment.

3 Data Mining Design

3.1 Conceptualization of Data Mining and Analysis

The data mining and analysis task must be enhanced by an explicit treatment of the languages used for concepts and hypotheses, and by an explicit description of knowledge that can be used. The algorithmic solution of the task is based on knowledge on algorithms that are used and on data that are available and that are required for the application of the algorithms. Typically, analysis algorithms are iterative and can run forever. We are interested only in convergent ones and thus need termination criteria. Therefore, conceptualization of the data mining and analysis task consists of a detailed

description of *six main parameters* (e.g. for inductive learning [34]):

(a) The *data analysis algorithm*: Algorithm development is the main activity in data mining research. Each of these algorithms transfers data and some specific parameters of the algorithm to a result.

(b) The *concept space*: the concept space defines the concepts under consideration for analysis based on certain language and common understanding.

(c) The *data space*: The data space typically consists of a multi-layered data set of different granularity. Data sets may be enhanced by metadata that characterize the data sets and associate the data sets to other data sets.

(d) The *hypotheses space*: An algorithm is supposed to map evidence on the concepts to be supported or rejected into hypotheses about it.

(e) The *prior knowledge space*: Specifying the hypothesis space already provides some prior knowledge. In particular, the analysis task starts with the assumption that the target concept is representable in a certain way.

(f) The *acceptability and success criteria*: Criteria for successful analysis allow to derive termination criteria for the data analysis.

Each instantiation and refinement of the six parameters leads to specific data mining tasks.

The result of data mining and data analysis is described within the knowledge space. The data mining and analysis task may thus be considered to be a transformation of data sets, concept sets and hypothesis sets into chunks of knowledge through the application of algorithms.

Problem solving and modelling considers, however, typically six aspects [16]:

(1) *Application, problems, and users*: The domain consists of a model of the application, a specification of problems under consideration, of tasks that are issued, and of profiles of users.

(2) *Context*: The context of a problem is anything what could support the problem solution, e.g. the sciences' background, theories, knowledge, foundations, and concepts to be used for problem specification, problem background, and solutions.

(3) *Technology*: Technology is the enabler and defines the methodology. It provides [23] means for the flow of problem solving steps, the flow of activities, the distribution, the collaboration, and the exchange.

(4) *Techniques and methods*: Techniques and methods can be given as algorithms. Specific algorithms are data improvers and cleaners, data aggregators, data integrators, controllers, checkers, acceptance determiners, and termination algorithms.

(5) *Data*: Data have their own structuring, their quality and their life span. They are typically enhanced by metadata. Data management is a central element of most problem solving processes.

(6) *Solutions*: The solutions to problem solving can be formally given, illustrated by visual means, and presented by models. Models are typically only normal models. The deep model and the matrix is already provided by the context and accepted by the community

of practice in dependence of the needs of this community for the given application scenario. Therefore, models may be the final result of a data mining and analysis process beside other means.

Comparing these six spaces with the six parameters we discover that only four spaces are considered so far in data mining. We miss the user and application space as well as the representation space. Figure 2 shows the difference.

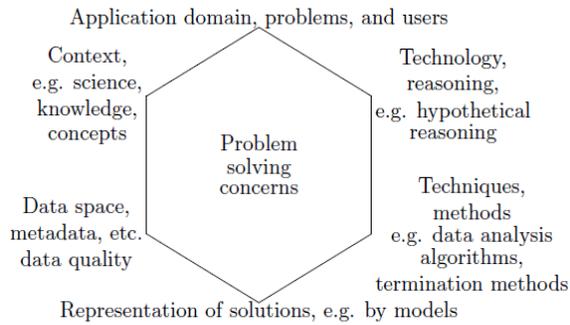


Figure 2 Parameters of Data Mining and the Problem Solving Aspects

3.2 Meta-models of Data Mining

An abstraction layer approach separates the application domain, the model domain and the data domain [17]. This separation is illustrated in Figure 3.

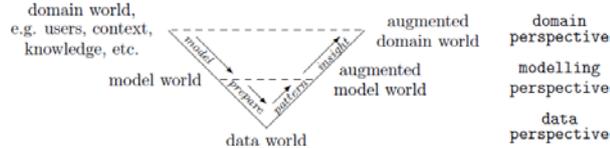


Figure 3 The V meta-model of Data Mining Design

The data mining design framework uses the inverse modeling approach. It starts with the consideration of the application domain and develops models as mediators between the data and the application domain worlds. In the sequel we are going to combine the three approaches of this section. The meta-model corresponds to other meta-models such as inductive modelling or hypothetical reasoning (hypotheses development, experimenting and testing, analysis of results, interim conclusions, reappraisal against real world).

4 Data Mining: A Systematic Model-Based Approach

Our approach presented so far allows to revise and to reformulate the model-oriented data mining process on the basis of well-defined engineering [15, 25] or alternatively on systematic mathematical problem solving [22]. Figure 4 displays this revision. We realize that the first two phases are typically implicitly assumed and not considered. We concentrate on the non-iterative form. Iterative processes can be handled in a similar form.

4.1 Setting the Deep Model and the Matrix

The problem to be tackled must be clearly stated in dependence on the utilization scenario, the tasks to be solved, the community of practice involved, and the given context. The result of this step is the deep model and its matrix. The first one is based on the background, the specific context parameter such as infrastructure and environment, and candidates for deep models.

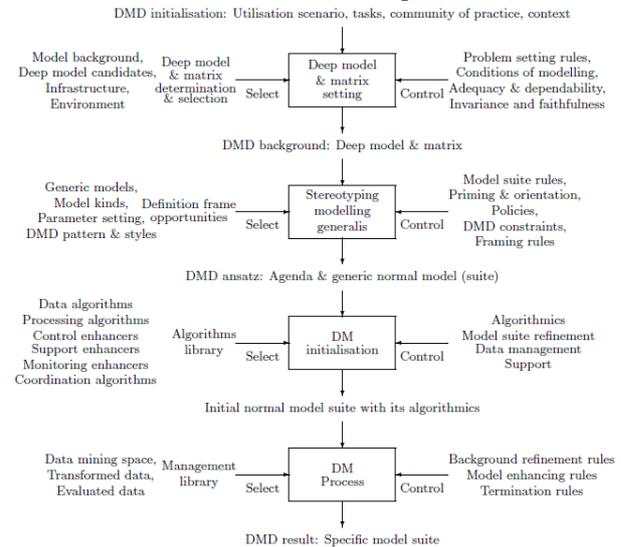


Figure 4 The Phases in Data Mining Design (Non-iterative form)

The data mining tasks can be now formulated based on the matrix and the deep model. We set up the context, the environment, the general goal of the problem and also criteria for *adequateness* and *dependability* of the solution, e.g. *invariance properties* for problem description and for the task setting and its mathematical formulation and *solution faithfulness properties* for later application of the solution in the given environment. What is exactly the problem, the expected benefit? What should a solution look like? What is known about the application?

Deep models already use a background consisting of an undisputable grounding and a selectable basis. The explicit statement of the background provides an understanding of the postulates, paradigms, assumptions, conceptions, practices, etc. Without the background, the results of the analysis cannot be properly understood. Models have their profile, i.e. goals, purposes and functions. These must be explicitly given. The parameters of a generic model can be either order or slave parameters [12], either primary or secondary or tertiary (also called genotypes or phenotypes or observables) [1, 5], and either ruling (or order) or driven parameters [12]. Data mining can be enhanced by knowledge management techniques.

Additionally, the concept space into which the data mining task is embedded must be specified. This concept space is enhanced during data analysis.

4.2 Stereotyping the Process

The general flow of data mining activities is typically implicitly assumed on the basis of stereotypes which form a set of tasks, e.g. tasks of prove in whatever system, transformation tasks, description tasks, and investigation tasks. Proofs can follow the classical deductive or inductive setting. Also, abductive, adductive, hypothetical and other reasoning techniques are applicable. Stereotypes typically use model suites as a collection of associated models, are already biased by priming and orientation, follow policies, data mining design constraints, and framing.

Data mining and analysis is rather stereotyped. For instance, mathematical culture has already developed a good number of stereotypes for problem formulation. It is based on a mathematical language for the formulation of analysis tasks, on selection and instantiation of the best fitting variable space and the space of opportunities provided by mathematics.

Data mining uses *generic models* which are the basis of normal models. Models are based on a separation of concern according the problem setting: dependence-indicating, dependence-describing, separation or partition spaces, pattern kinds, reasoning kinds, etc. This separation of concern governs the classical data mining algorithmic classes: association analysis, cluster analysis, data grouping with or without classification, classifiers and rules, dependences among parameters and data subsets, predictor analysis, synergetics, blind or informed or heuristic investigation of the search space, and pattern learning.

4.3 Initialization of the Normal Data Models

Data mining algorithms have their capacity and potential [2]. Potential and capacity can be based on SWOT (strengths, weaknesses, opportunities, and threats), SCOPE (situation, core competencies, obstacles, prospects, expectation), and SMART (how simple, meaningful, adequate, realistic, and trackable) analysis of methods and algorithms. Each of the algorithm classes has its strengths and weaknesses, its satisfaction of the tasks and the purpose, and its limits of applicability. Algorithm selection also includes an explicit specification of the order of application of these algorithms and of mapping parameters that are derived by means of one algorithm to those that are an input for the others, i.e. an explicit association within the model suite. Additionally, evaluation algorithms for the success criteria are selected. Algorithms have their own obstinacy, their hypotheses and assumptions that must be taken into consideration. Whether an algorithm can be considered depends on acceptance criteria derived in the previous two steps.

So, we ask: *What kind of model suite architecture suits the problem best? What are applicable development approaches for modelling? What is the best modelling technique to get the right model suite? What kind of reasoning is supported? What not? What are the limitations? Which pitfalls should be avoided?*

The result of the entire data mining process heavily depends on the appropriateness of the data sets, their properties and quality, and more generally the data schemata with essentially three components: application data schema with detailed description of data types, metadata schema [18], and generated and auxiliary data schemata. The first component is well-investigated in data mining and data management monographs. The second and third components inherit research results from database management, from data mart or warehouses, and layering of data. An essential element is the explicit specification of the quality of data. It allows to derive algorithms for data improvement and to derive limitations for applicability of algorithms. Auxiliary data support performance of the algorithms.

Therefore typical data-oriented questions are: *What data do we have available? Is the data relevant to the problem? Is it valid? Does it reflect our expectations? Is the data quality, quantity, recency sufficient? Which data we should concentrate on? How is the data transformed for modelling? How may we increase the quality of data?*

4.4 The Data Mining Process Itself

The data mining process can be understood as a coherent and stepwise refinement of the given model suite. The model refinement may use an explicit transformation or an extract-transform-load process among models within the model suite. Evaluation and termination algorithms are an essential element of any data mining algorithm. They can be based on quality criteria for the finalized models in the model suite, e.g. generality, error-proneness, stability, selection-proneness, validation, understandability, repeatability, usability, usefulness, and novelty.

Typical questions to answer within this process are: *How good is the model suite in terms of the task setting? What have we really learned about the application domain? What is the real adequacy and dependability of the models in the model suite? How these models can be deployed best? How do we know that the models in the model suite are still valid? Which data are supporting which model in the model suite? Which kind of errors of data is inherited by which part of which model?*

The final result of the data mining process is then a combination of the deep model and the normal model whereas the first one is a latent or hidden component in most cases. If we want, however, to reason on the results then the deep model must be understood as well. Otherwise, the results may become surprising and may not be convincing.

4.5 Controllers and Selectors

Algorithmics [6] treats algorithms as general solution pattern that have parameters for their instantiation, handling mechanisms for their specialization to a given environment, and enhancers for context injection. So, an algorithm can be derived based on explicit selectors and control rules [4] if we neglect context injection. We

can use this approach for data mining design (DMD). For instance, an algorithm pattern such as regression uses a generic model of parameter dependence, is based on blind search, has parameters for similarity and model quality, and has selection support for specific treatment of the given data set. In this case, the controller is based on enablers that specify applicability of the approach, on error rules, on data evaluation rules that detect dependencies among control parameters and derive data quality measures, and on quality rules for confidence statements.

4.6 Data Mining and Design Science

Let us finally associate our approach with design science research [13]. Design science considers systematic modelling as an embodiment of three closely related cycles of activities. The *relevance cycle* initiates design science research with an application context that not only provides the requirements for the research as inputs but also defines acceptance criteria for the ultimate evaluation of the research results. The central *design cycle* iterates between the core activities of building and evaluating the design artifacts and processes of the research. The orthogonal *rigor cycle* provides past knowledge to the research project to ensure its innovation. It is contingent on the researchers' thoroughly research and references the knowledge base in order to guarantee that the designs produced are research contributions and not routine designs based upon the application of well-known processes.

The relevance cycle is concerned with the problem specification and setting and the matrix and agenda derivation. The design cycle is related to all other phases of our framework. The rigor cycle is enhanced by our framework and provides thus a systematic modelling approach.

5 Conclusion

The literature on data mining is fairly rich. Mining tools have already gained the maturity for supporting any kind of data analysis if the data mining problem is well understood, the intentions for models are properly understood, and if the problem is professionally set up. Data mining aims at development of model suites that allows to derive and to draw dependable and thus justifiable conclusions on the given data set. Data mining is a process that can be based on a framework for systematic modelling that is driven by a deep model and a matrix. Textbooks on data mining typically explore in detail algorithms as blind search. Data mining is a specific form of modeling. Therefore, we can combine modeling with data mining in a more sophisticated form. Models have however an inner structure with parts which are given by the application, by the context, by the commonsense and by a community of practice. These fixed parts are then enhanced by normal models. A typical normal model is the result of a data mining process.

The current state of the art in data mining is mainly

technology and algorithm driven. The problem selection is made on intuition and experience. So, the matrix and the deep model are latent and hidden. The problem specification is not explicit. Therefore, this paper aims at the entire data mining process and highlights a way to leave the ad-hoc, blind and somehow chaotic data analysis. The approach we are developing integrates the theory of models, the theory of problem solving, design science, and knowledge and content management. We realized that data mining can be systematized. The framework for data mining design exemplarily presented is an example in Figure 4.

Acknowledgement. We thank for the support of this paper by the CRC 1266. We are very thankful for the fruitful discussions with the members of the CRC.

References

- [1] G. Bell. The mechanism of evolution. Chapman and Hall, New York (1997)
- [2] R. Berghammer and B. Thalheim., Methodenbasierte mathematische Modellierung mit Relationenalgebren. In: Wissenschaft und Kunst der Modellierung: Modelle, Modellieren, Modellierung, pp. 67–106. De Gruyter, Boston (2015)
- [3] M.R. Berthold, C. Borgelt, F. Höppner, and F. Klawonn. Guide to intelligent data analysis. Springer, London (2010).
- [4] A. Bienemann, K.-D. Schewe, and B. Thalheim. Towards a theory of genericity based on government and binding. In: Proc. ER'06, LNCS 4215, pp. 311–324. Springer (2006)
- [5] L.B. Booker, D.E. Goldberg, and J.H. Holland. Classifier systems and genetic algorithms. Artificial Intelligence, 40 (1–3): pp. 235–282 (1989)
- [6] G. Brassard and P. Bratley. Algorithmics - Theory and Practice. Prentice Hall, London (1988)
- [7] A. Coleman. Scientific models as works. Cataloging & Classification Quarterly, Special Issue: Works as Entities for Information Retrieval, 33, pp. 3-4 (2006)
- [8] A. Dahanayake and B. Thalheim. Co-evolution of (information) system models. In: EMMSAD 2010, LNBIB vol. 50, pp. 314–326. Springer (2010)
- [9] D. Embley and B. Thalheim (eds). The Handbook of Conceptual Modeling: Its Usage and Its Challenges. Springer (2011)
- [10] N.P. Gillett, F.W. Zwierns, A.J. Weaver, G.C. Hegerl, M.R. Allen, and P.A. Stott. Detecting anthropogenic influence with a multi-model ensemble. Geophys. Res. Lett., 29:31–34, 2002.

- [11] E. Guerra, J. de Lara, D.S. Kolovos, and R.F. Paige. Inter-modelling: From theory to practice. In *MoDELS 2010, LNCS 6394*, pp. 376–391, Springer, Berlin (2010)
- [12] H. Haken, A. Wunderlin, and S. Yigitbasi. An introduction to synergetics. *Open Systems and Information Dynamics*, 3(1): pp. 1–34 (1994)
- [13] A. Hevner, S. March, J. Park, and S. Ram. Design science in information systems research. *MIS Quarterly*, 28(1): pp. 75–105 (2004)
- [14] P.J. Hunter, W. W. Li, A. D. McCulloch, and D. Noble. Multiscale modeling: Physiome project standards, tools, and databases. *IEEE Computer*, 39(11), pp. 48–54 (2006)
- [15] ISO/IEC 25020 (Software and system engineering - software product quality requirements and evaluation (square) - measurement reference model and guide). *ISO/IEC JTC1/SC7 N3280* (2005)
- [16] H. Jaakkola, B. Thalheim, Y. Kidawara, K. Zettsu, Y. Chen, and A. Heimbürger. Information modelling and global risk management systems. In: *Information Modeling and Knowledge Bases XX*, pp. 429–446. IOS Press (2009)
- [17] K. Jannaschk. Infrastruktur für ein Data Mining Design Framework. PhD thesis, Christian-Albrechts University, Kiel (2017)
- [18] F. Kramer and B. Thalheim. A metadata system for quality management. In: *Information Modelling and Knowledge Bases*, pp. 224–242. IOS Press (2014)
- [19] O. Nakoinz and D. Knitter. *Modelling Human Behaviour in Landscapes*. Springer (2016)
- [20] J. Pardillo. A systematic review on the definition of UML profiles. In: *MoDELS 2010, LNCS 6394*, pp. 407–422, Springer, Berlin (2010)
- [21] D. Petrelli, S. Levin, M. Beaulieu, and M. Sanderson. Which user interaction for cross-language information retrieval? Design issues and reflections. *JASIST*, 57(5): pp. 709–722 (2006)
- [22] O.H. Pilkey and L. Pilkey-Jarvis. *Useless Arithmetic: Why Environmental Scientists Can't Predict the Future*. Columbia University Press, New York (2006)
- [23] A.S. Podkolsin. Computer-based modelling of solution processes for mathematical tasks (in Russian). *ZPI at Mech-Mat MGU, Moscow* (2001)
- [24] M. Pottmann, H. Unbehauen, and D.E. Seborg. Application of a general multi-model approach for identification of highly nonlinear processes – a case study. *Int. Journal of Control*, 57(1): pp. 97–120 (1993)
- [25] B. Rumpe. *Modellierung mit UML*. Springer, Heidelberg (2012)
- [26] A. Samuel and J. Weir. *Introduction to Engineering: Modelling, Synthesis and Problem Solving Strategies*. Elsevier, Amsterdam (2000)
- [27] G. Simsion and G.C. Witt. *Data modeling essentials*. Morgan Kaufmann, San Francisco (2005)
- [28] M. Skusa. *Semantische Kohärenz in der Softwareentwicklung*. PhD thesis, CAU Kiel, (2011)
- [29] B. Thalheim. Towards a theory of conceptual modelling. *Journal of Universal Computer Science*, 16(20): pp. 3102–3137, (2010)
- [30] B. Thalheim. The conceptual model \equiv an adequate and dependable artifact enhanced by concepts. In: *Information Modelling and Knowledge Bases XXV*, pp. 241–254. IOS Press (2014)
- [31] B. Thalheim. Conceptual modeling foundations: The notion of a model in conceptual modeling. In: *Encyclopedia of Database Systems*, Springer (2017)
- [32] B. Thalheim and M. Tropmann-Frick. Wherefore models are used and accepted? The model functions as a quality instrument in utilisation scenarios. In: I. Comyn-Wattiau, C. du Mouza, and N. Prat, editors, *Ingenierie Management des Systemes D'Information* (2016)
- [33] B. Thalheim, M. Tropmann-Frick, and T. Ziebermayr. Application of generic workflows for disaster management. In: *Information Modelling and Knowledge Bases*, pp. 64–81. IOS Press (2014)
- [34] B. Thalheim and Q. Wang. Towards a theory of refinement for data migration. In: *ER'2011, LNCS 6998*, pp. 318–331. Springer, (2011)
- [35] T. Zeugmann. Inductive inference of optimal programs: A survey and open problems. In: *Nonmonotonic and Inductive Logics*, pp. 208–222. Springer, Berlin (1991)