Qualitative and Quantitative Models

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

Abstract. Data are considered to be the oil of the 21th century. They are also a rich source for many sciences, especially those that use observational data for development of an understanding behind the data. They are used to gain an insight into the discipline based on (phenotypical) observations. This insight may result in a quantitative theory offer. The main target is however a theory that explains the data. We develop a model-backed approach to theory development based on quantitative theory offers. Models are becoming the mediator between quantitative and qualitative theories. Models can be systematically developed based on a layering approach.

Keywords: models, models as mediators, qualitative theories, middle-range theories, theory offer, quantitative theories, genotype models, phenotype models, phenotypical data, data science.

1 Introduction

1.1 From Empiric Sciences to Data Science

Data science is considered to be a new stage of scientific research. Data science is based on analysis of data resources. The analysis asks the right questions with efficient processing algorithms, machine learning and cognitive computing techniques, refined statistical models, and innovative visions of how to more effectively extract the relevant data assets and scrutinise them fast with more sophisticated results.

Sciences use all four dimensions in different proportions. We consider empirical science (describing natural phenomena and experimenting in social economics, archaeology, etc.) with a larger empirical investigation, theory-driven sciences (mainly mathematical and natural sciences) with orientation on theory-backed investigation, computation science (simulating complex phenomena, e.g. bioengineering, meteorology) with orientation on computational investigation, and data digging and exploration sciences (currently mainly data analysis and mining, partially AI) with data-oriented investigation. The last two kinds may also be considered as technology-enabled sciences.

The role of models also changes as displayed in Fig. 1:

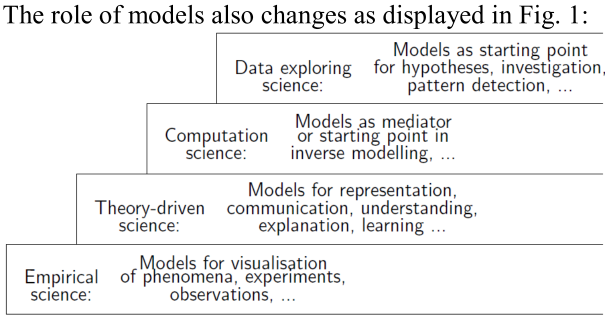

Fig 1. Models and their functions in sciences

Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018

Data science goes beyond empirical sciences, theory-driven science, and computation science [12]. Data science discovers pattern and generates insights in data sources or data proxies. It is based on raw data and builds these insights based on knowledge from the scientific discipline and application domain. It provides models, recommendations, and potential theories on how to interpret the data. It is based on a process of organising data for analysis including data proliferation, data collection organisation, cleaning, application of tools, and analysis. It may consider huge data collections as well as small data sets. The proxy data are compiled and may become 'smart' quantitative data for quantitative research. Data science is essentially the 'science' that is turning data proxies into narrative and into quantitative data. It thus develops an understanding of the data itself. In our case, we investigate rather thin data sets. The picture is however similar to the one with very large data sets.

1.2 From Proxy-Based Investigation to Quantitative and Qualitative Theories

Explorative and investigative theory development (e.g. [1, 18, 20]) starts with an investigation of data sources and develop some proxy-based observation concepts and a theory offer. A *theory offer* is a scientific, explicit and systematic discussion of foundations and methods, with critical reflection, and a system of assured conceptions providing a holistic understanding. A theory offer is understood as the underpinning of technology and science similar to architecture theory [23] and the approaches by Vitruvius [32}, and L.B. Alberti [2]. Theory offers do not constitute a theory on their own, rather are some kind of collection consisting of pieces from different and partially incompatible theories, e.g. sociology theories such as the reference group theory, network theories, economic theories such as the agent, Darwinian evolution theories, subjective rationality theories, and ideology theories. *Middle-range theories* lie between the hypotheses space and all-inclusive efforts to develop a unified holistic theory. They guide empirical enquiry. They are close enough to observed data. A

(scientific) *theory* is a "systematic ideational structure of broad scope, conceived by the human imagination that encompasses a family of empirical (experiential) laws regarding regularities existing in objects and events, both observed and posited. A scientific theory is a structure suggested by these laws and is devised to explain them in a scientifically rational manner. In attempting to explain things and events, the scientist employs (1) careful observation or experiments, (2) reports of regularities, and (3) systematic explanatory schemes (theories)." [6]
Typically, we start with some data, e.g. proxy data. Next we (g-)derive *proxy concepts* (or concepts) and form some proposals for (h-)formation of a proposal of a potential explaning theory, i.e. *a theory request* (or request for a theory offer).
Proxy sources can be aggregated and (f-)condensed and thus become quantitative sources which are the basis for (g-)formation of *quantitative concepts*. These quantitative concepts are (h-)embedded into *theory offers* (or, resp., theory requests for proxy requests) and are the basis for a theory offer that serves as an explanation for the theory request. Quantitative concepts can be (F-)mapped back to proxy concepts. Proxy-based research and quantitative research is well-integrated if the diagram is commuting.

$$\begin{tikzcd}
\text{PROXY SOURCES} \arrow[r, "f_s^s"] \arrow[d, "g_c^s"] & \text{QUANTITATIVE SOURCES} \arrow[r, leftrightarrow, "???", "??"'] \arrow[d, "g_c^s"] & \text{SUPPORTING SOURCES} \\
\text{proxy concepts} \arrow[d, hook, "h_t^c"] & \text{quantitative concepts} \arrow[l, "F_c^c"'] \arrow[d, hook, "h_t^c"] & \text{theoretical concepts} \arrow[u, "G_s^c"] \arrow[d, hook, "h_t^c"] \\
\text{theory request} & \text{theory offer} \arrow[l, "expla-"', "nation"] & \text{theory}
\end{tikzcd}$$

Fig. 2 The current state-of-art in the data science

Theories can be built on the basis of theoretical concepts which are supported by sources. Quantitative concepts should be associated with qualitative concepts. The association can only be developed in the case when the association among the data has been clarified. So far, the explanations that can be generated are mainly developed for explaining the observations made on the basis of proxies.
We arrive therefore with the following **research challenge**: *How we can close the gap between quantitative theory offers and qualitative theories within the setting of data science?*

1.3 A Typical Data Science Application

Investigative modelling at CRC 1266 [1, 16] aims at exploring and explaining transformations in societies as "processes leading to a substantial and enduring re-organisation" [1] of any or all aspects of the human social, cultural, economic, and environmental relations. Proxies are observations for main concepts[1] in Figure 3. These main concepts need however a quantitative underpinning and a number of theoretical concepts.

1.4 The Storyline of the Paper

We develop an approach to data science based on models. Models are instruments that function in utilisation scenarios. One of these scenarios might be the development of a theory for a theory offer. We will show in the sequel how this approach can be systematically applied to development of mediating models that close the gap in Figure 5. We start with a notion of model in Section 2. Six research questions are developed which are answered in Sections 3 and 4. Next**,** we develop a model construction approach in Section 3**.** Finally, we apply this approach to data science and use models as mediators in Section 4**.**

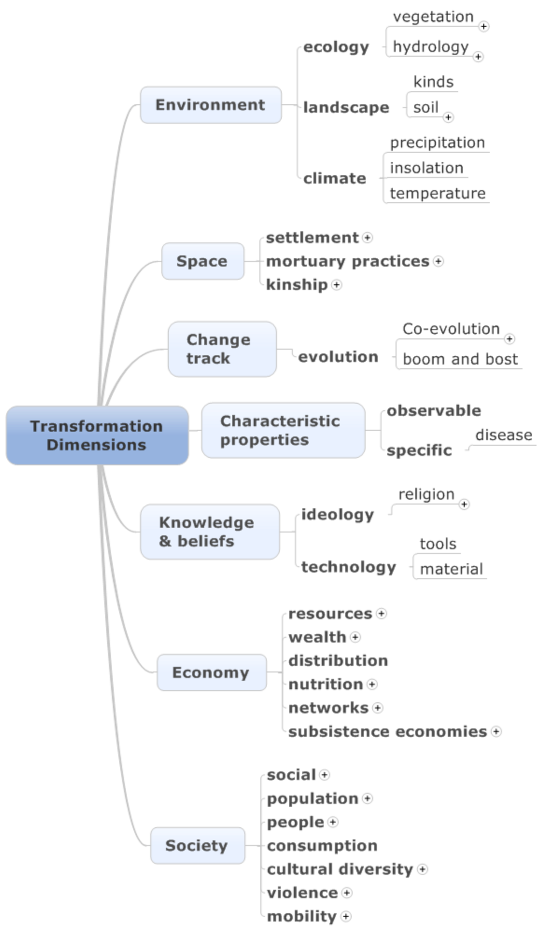

Fig. 3 Theoretical concepts to be investigated in the CRC 1266

2 Models

Models are widely used in life, technology and sciences. Their development is still a masterpiece of an artisan and

[1] We restrict the mindmap in Fig. 3 to main concepts and do not display the full concept network. For details see the website of the project.

not yet systematically guided and managed. The main advantage of model-based reasoning is based on two properties of models: they are focused on the issue under consideration and are thus far simpler than the application world and they are reliable instruments since both the problem and the solution to the problem can be expressed by means of the model due to its dependability. Models must be sufficiently comprehensive for the representation of the domain under consideration, efficient for the solution computation of problems, accurate at least within the scope, and must function within an application scenario.

Research question 1: *Can models be used for resolving the gap between theory offers in quantitative research and theories in qualitative research?*

Consider for instance the CRC 1266 application: Transformation is considered in this context as a phenomenon that requires detailed description of features and hence quantitative data are necessary for descriptions by empirical models and simulations.

Models mediate between quantitative theories and qualitative theories. Models are applied in hypothetical and investigative scenarios, should support causal reasoning as well as network-oriented reasoning, and are developed in an empiric framework.

2.1 The Notion of Model

Let us first briefly repeat our approach to the notion of model: A ***model*** is a *well-formed*, *adequate*, and *dependable instrument* that represents *origins* and that *functions* in *utilization scenarios* [10, 27, 28].

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

The model should be well-formed according to some well-formedness criterion. As an instrument or more specifically an artifact a model comes with its *background*, e.g. paradigms, assumptions, postulates, language, thought community, etc. The background is often given only in an implicit form. The background is often implicit and hidden.

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* characterisation for internal quality, external quality and quality in use or through quality characteristics [26] such as correctness, generality, usefulness, comprehensibility, parsimony, robustness, novelty etc. Sufficiency is typically combined with some assurance evaluation (tolerance, modality, confidence, and restrictions).

A well-formed instrument is called *dependable* if it is sufficient and is justified for some of the justification properties and some of the sufficiency characteristics.

2.2 Functions of Models

Models are used as instruments in certain utilisation scenarios such as communication, reflection, understanding, negotiation, explanation, exploration, learning, introspection, theory development, documentation, illustration, analysis, construction, description, and prescription. They have to fulfil a number of specific functions in these scenarios.

Typical functions of models as instruments in scenarios are (a) cognition, (b) explanation and demonstration, (c) indication, (d) variation and optimisation, (e) projection and construction, (f) control, (g) substitution, and (h) experimentation [31].

Our notion of a model and its functions also covers the notion of a middle-range theory. Middle-range theories have been developed for hypothetical investigation in empiric research. Models are more general and support also other functions. They provide a methodological power and allow inclusion of knowledge as long as the model (or the model suite) is still coherent. In the sequel we shall discover that models can be used as a mediator. Models may reflect qualitative as well as quantitative theories. Reflection is based on the function that is anticipated when selecting a utilisation scenario.

2.3 Model := Normal Model ₪ Deep Model

A model consists of a *normal model* that is combined with some *deep model* similar to the visible (or exterior) and invisible parts of an iceberg [16, 29, 30]. The deep model reflects (α) the intentions of the problem world, (β) the accepted understanding within the community of practice, (γ) the context of the application domain, (δ) the background that is commonly accepted in the problem and application domain, and (ε) the general restrictions to the origins that might be considered. The deep model allows partial derivation of the justification and adequacy of a model.

The normal model reflects the collection of origins that are currently under consideration. Both the deep and the normal model are dependent on the functions that a model should play in application scenarios. Development of models is often restricted to development of a normal model under the assumption that the deep model is given by the modelling method, the context, the community of practice, and the function that the model has to play in a given scenario. The modelling methods also determined the methods that are used for model development. It might also include the utilization methods.

Research question 2: *Can we separate the deep model from the normal model in such a way that the model can be composed of the deep model and of the normal model?*

If the answer to this question is positive then we might try to consider the model as an enhancement of the deep model. In this case, the development of a model can be layered.

The deep model also reflects a shared foundational set of beliefs and priorities that govern the way how a discipline handles and interprets its data. Deep models are not dominating. They are hidden and somehow form the model's obstinacy.

Research question 3: *How can be development of a model layered into the development of a deep model followed by the development of the normal model?*

We may now ask us whether this approach is universal. The answer will be negative if the notion of model also includes models with intractable deep models, e.g. for metaphors, parables, or physical representations. We might however concentrate on models in sciences and technology.

2.4 Model Suites

Models may be given as a holistic instrument that combines all aspects into one model. The approach is often too challenging. A simpler approach is the consideration of a model as a model suite (or model ensemble) [8,25] that consists of a coherent collection of models which are representing different points of view and attention. It is extended by an explicit association or collaboration schema among the models, controllers that maintain consistency or coherence of the model suite, application schemata for explicit maintenance and evolution of the model suite, and tracers for the establishment of the coherence.

Research question 4: *Exists there a systematic approach to model development that is based on a co-development of normal models and deep models? Which additional models should be integrated into the model suite?*

2.5 Generic and Specific Models

Model development does not start from scratch. We often start with generic models. A *generic model* [16] 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. Generic models can be calibrated to specific models through a process of data or situation calibration, refinement, concretization, context enhancement, or instantiation.

Research question 5*: Can we develop normal models starting with a generic model and are they still capable of being integrated with the deep model?*

If the answer is positive then generic normal models can be calibrated to specific normal models through a process of data or situation calibration, refinement, concretization, context enhancement, or instantiation.

2.6 Data Mining as a Success Story

In [16], we developed the V-model to data mining based on a separation of the data mining process into the domain perspective with its domain world of users from a community of practice, the modelling perspective with a model world, and the data perspective in a data world. Users are interested in solution of certain problems an application world, share the context and also the scientific and technological background. The classical data model mining process uses these perspectives for a stepwise development of a model that allows to solve their problems, e.g. (1) by modelling the problem and the issues under consideration, (2) by preparing the data world for development and enhancement of models, (3) by applying data mining algorithmics for pattern detection and model development, and (4) by using the model for development of some solution for the problems and thus augmenting the application domain world. The model development process itself can be understood a multi-iterative guided procedure that has its flow of activities.

This approach extends the classical CRISP framework [4] and other approaches to systematic data mining, e.g. [15]. Each of these approaches has its capacity and potential as well as its threats and limitations. The question is now:

Research question 6: *Can we generalize a data mining setting to model development for data science in such a way that models mediate between theory offers and theories?*

Data analysis and model suite development currently inherit success stories in a similar application. These success stories follow some kind of a meta-pattern and result in a specific data mining process as an example of a modelling method or modelling mould. Data mining starts with exploring and understanding the data mining project, its data, and a general setting of principles of modelling. After the project and the nature of the data is understood, data are preprocessed and prepared for the application of algorithms. Next pattern within the data are investigated. This pattern analysis results in clusters, maps, association rules, and some deviation analysis, i.e. we develop a model on the data space. This model is then used for development of explanations, e.g. via decision trees, (Bayesian) classifiers, regression, and (rule) learning approaches. We develop a second model on top of the first model. Next, the data space is considered in a general form by prediction analysis, e.g. nearest neighbor predictors, (artificial) neural networks, support vector machines, or other ensemble methods. The result is another model. Finally, the models are evaluated and potentially deployed. If the evaluation shows that the models satisfy quality criteria, we revise the models.

2.6 The Programme to Answer the Research Questions and Solving the Research Challenge

The research questions will now be handled by specific methodological mould and hypotheses that might support a positive answer to our research challenge:

1. Models may have a number of functions. One of the functions can be the mediator function. After knowing how to develop a model, we can investigate the mediator function of models for development of a shared view on both quantitative theory offers and qualitative theories.
2. We already discovered that the deep model is

somehow hidden inside the model and must be explicated. Since deep models are the intrinsic part of a model or model suite, we might try to develop deep models first before we develop normal models.
3. The layered approach has its potential and capacity. We should also understand the limitations and the risks of this approach.
4. A model suite consists of a number of associated models. Some of them are dependent on others. If we use a layered approach then we could also use generators for models within a model suite. A typical generator is applied for generation of informative and representation models.
5. Generic and reference models combine, abstract, and generalise the experience one has gained while developing a number of models in the past. It seems thus to be a good idea to collect generic and reference model as stereotypes or pattern and then to use these models for derivation of normal models.
6. Data mining nicely demonstrates how a model can be calibrated to its real purpose. Another good example is inverse modelling where a generic model is used for refinement, calibration customi-zation, settlement, fitting, pruning, parameter instantiation, and operational adaption to origins and the world under consideration.

3 A Layered Model Composition Approach

Let us now investigate whether our hypotheses can be supported and handled in a proper way. Models are masterpieces and would thus become work of an engineer if the model development process can be properly handled.
Engineering and software engineering (e.g. [13, 22]) distinguish between the five primary development dimensions:

- *Activities* ('how') describe the way how the work is performed and the practises for the work.
- *Work products* ('what') are the result of the specification and are used during specification.
- *Roles* ('who') describe obligations and permissions, the involvement of actors in the specification process.
- Aspects ('where') are used for separation of concern during the specification process.
- *Resources* ('on which basis') are the basis for the specification.

Since a model (suite) is also a work product we may refine this approach to model engineering. The modelling method we will use is similar to the modelling method in mathematics [3].

3.1 Systematic Layering for Model Suites

The layered approach has already often and successfully been used in Computer Engineering. Most program language realisations follow this approach since COBOL and ALGOL 60 development (e.g. infrastructure definition; variable space; program space; interpreted or compiled code) and application development (e.g. application case; infrastructure; design; specialisation & tuning; deliver). Layering has also been the guiding paradigm behind text processing.
Model suite development and deployment will be based on separation of concern into intrinsic and extrinsic parts of models. Models typically consist from one side of a normal model that displays all obviously relevant and important aspects of a model and from the other side of a deep model that intrinsically reflects commonly accepted intentions, the accepted understanding, the context, the background that is commonly accepted, and restrictions for the model. The model suite will be layered into models for *initialisation*, for *strategic setup*, for *tactic definition*, for *operational adaptation*, and for *model delivery.*

(I) The Initialisation Layer
The W*H specification pattern [9] can be applied to *model initialisation* as well an includes then the following set of statements: (i) a plan, function, and purpose dimension (model as a conception: 'wherefore', 'why', 'to what place or end', 'for when', 'for which reason') within a scenario in which the model is going to be used as an instrument; (ii) a user or CoP dimension ('who', 'by whom', 'to whom', 'whichever') that describes the task portfolio in the CoP and profile of users including beliefs, desires and intentions; (iii) an application and a problem dimension (`in what particular or respect' , 'from which', 'for what', 'where', `whence'); the added value dimension (evaluation). The initialisation layer may also be enhanced by a contrast space for user-related separation of a model and a relevance space that is dependent on the user [11]. The contrast and relevance spaces as a form of mind-setting also define what is not of interest.

(II) The Enabling Setup Layer
The *enabling intrinsic setup layer* defines the opportunity space and the infrastructure for the model. The results will be from one side a deep model and from the other side a modelling framework or modelling mould that guides and govern next activities. We define the *context* and the most of the *background* (the grounding (paradigms, postulates, restrictions, theories, culture, foundations) and the basis (assumptions, concept world, practices, language as carrier, thought community and thought style, methodology, pattern, routines, common sense)) of the model. The context, extrinsic, and strategic dimension answers question like 'at or towards which', 'where about', 'to what place or situation', and 'when'. Additionally, we decide which methodology and environment seem to be the most effective and purposeful. The development and deployment dimension ('how', 'whence', 'what in', 'what out', 'where') defines the modelling methodology, i.e. the *modelling mould*.

(III) The Extrinsic Source Reflection Layer
We separate the deep model elements from elements of the *normal model* at the *extrinsic source reflection layer.* According to the model function, the normal model

represents *extrinsic* elements of potential origins based on their content and thus answers questions such as ‘what’, ‘with which’, and ‘by means of which’. It reflects the extrinsic theory essentials that are necessarily to be represented, e.g. conceptions or pre-conceptions from the theory that is underpinning the application. The normal model can be built from scratch (‘greenfield’ modelling). It is more usual based on the experience gained so far. The latter case thus starts with a generic or reference model that might incorporate parameters. The extrinsic source reflection layer can be understood as a tactical layer.

(IV) The Operational Customisation Layer
Generic or general normal models are adjusted to those that a best fitted to those origins that are considered for the application in the *operational customisation layer*. This layer is sometimes holistically handled with extrinsic reflection. Inverse modelling uses this layer for adaptation of the model to the observational data (e.g. data adaption in astrophysics or parameter instantiation in most data mining processes). In some cases, this layer seems to be trivial. It is not trivial in the general case however. It instantiates parameters, adapts the normal model to those origins (or data sources) that are really under consideration, prepares the model for the special use and to the special - most appropriate - solution, and integrates the deep model with the normal model. The normal model is typically pruned in order to become simpler based on Solomonoff and Occam principled deviation [19, 24] and error-prone. The (normal) model might be enhanced by concepts and thus become a *conceptual model*.

(V) The Delivery and Product Layer
The *final result* of the modelling process is a model suite that is adequate for origins, properly justified, and sufficient at the *delivery and product layer*. We cannot expect that one singleton model is the best instrument for all members of the community of practice. A sophisticated model that integrates deep and specific normal models is delivered to some members. An informative model that is derived from this model can be better for other CoP members. Models delivered in the finalisation space are often enhanced by additional annotations, e.g. relating the model to the demands for members of the CoP by answering the ‘with’, ‘by which’, ‘by whom’, ‘to whom’, ‘whichever’, ‘what in’, and ‘what out’ questions. At the delivery and product layer we thus generate a number of associated models.

4 Models as Mediating Instruments

Model-backed reasoning is thus some kind of revisable reasoning depending on the stages of knowledge. Modelling becomes now a process that starts with deep models and continues with suites of generic models and revisable refinement according to data on hand. It should support handling of uncertainties and incompleteness of any kind and must thus make use of an integrated data management. Therefore, model-backed reasoning is properly based on layered model development.

4.1 Towards Models as Mediators

Models can be used to render the theory offer. At the same time models may also render a theory. We claim that these two views can be integrated. The model functions thus as mediator [17]. The rendering procedures are however different. We envision that this integration can be based on the mappings in Figure 4.

$$
\begin{tikzcd}
\text{quantitative research} & \text{models as mediators} & \text{qualitative theory-oriented research} \\
\text{QUANTITATIVE SOURCES} \arrow[r, "f^{(s,q)}_{(s,m)}"] \arrow[rr, bend left, "f^{(s,q)}_{(s,t)}"] \arrow[d, "g^{(s,q)}_{(c,q)}"'] & \text{SUPPORTING SOURCES} & \text{SUPPORTING SOURCES} \arrow[l, "f^{(s,t)}_{(s,m)}"'] \\
\text{quantitative concepts} \arrow[r, "f^{(c,q)}_{(c,m)}"] \arrow[d, hook, "h_q^{(c,q)}"] & \text{model concepts} \arrow[u, "G^{(c,m)}_{(s,m)}"] \arrow[d, hook, "h_m^{(c,m)}"] & \text{theoretical concepts} \arrow[l, "f^{(c,t)}_{(c,m)}"'] \arrow[u, "G^{(c,t)}_{(s,t)}"'] \arrow[d, hook, "h_t^{(c,t)}"] \\
\text{theory offer} \arrow[r, "\text{offer-}"', "\text{driven view}"] & \text{model} & \text{theory} \arrow[l, "\text{theory-driven}"', "\text{view}"]
\end{tikzcd}
$$

Fig. 4 Models as integrating and mediating instruments

Models can be understood as being composed of model concepts that are supported by data sources. We can now distinguish f-mappings at the same level, g-mappings between sources and concept, G--mappings from concepts to supporting sources, and h-embedding mappings from concepts to theory offers, models, or theories. Fig. 4 shows how the gap displayed in Fig. 2can be closed. A model consists of concepts which are derived from quantitative or theoretical concepts. The association between these two kinds is either handled similar to associations in a model suite or is directly integrated similar to integration of structural components from different database schemata.
Quantitative concepts are indicators or general quantitative properties. Model concepts are already abstractions from those quantitative concepts. Theoretical concepts in Fig. 3 are elements of a theory that is currently under development. The research task to be accomplished is the harmonisation of these mappings. This harmonisation can be based on the mappings for supporting resources if some commuting diagram properties are valid for model concepts and the model. For instance, quantitative concepts used in the CRC 1266 for support of reasoning as displayed in Fig. 3 are typicality of a property for the proxy concept, distribution (plants, ...) of properties within the proxies, the concept of a village (with an understanding from medieval times), evolution pattern that can be observed for proxies, exchange pattern, society and their members (within a modern interpretation), the variety of concepts of a center, concepts characterizing social structures and gender relations, and concepts characterizing architecture. Accepted hypotheses are, for example, are the existence of diachrony pattern for changes, chromatography of time series in the proxies, the existence of periods of dark ages, the human as a driver

of development, the existence of gift exchange between societies (reciprocity), pattern of ritualistic behavior in 3 phases, pattern of kinds and styles of collaboration, and pattern of interaction among centers.

Theories of qualitative research use, for instance, concepts of evolution change, of climate change, of influence of climate change, of climate reactions, of sedimentation, concepts expressed as big man theories, and a number of concepts that might characterize the neolithication.

We thus use the model as some kind of twofold medium which has 'Janus' head behavior: it is both (I) a view of the theory-offer and (II) a view of the theory. It is a model (i) *for* the theory offer as a reflective "epistemic thing" [21] of discovery and a presupposition and (ii) *of* a theory as a specific viewpoint representation. It thus comprehends what has been developed for theory offers and supports explanations of the theory.

The development of a model has, however, also a feedback turn both on the theory offer and the theory. The model is then at the same time an instrument, a mediator, a companion, a middle, and a medium in the sense of [5]. The model itself thus becomes an investigation instrument.

4.2 Evidence-Based Reasoning in Data Science

Let us finally discuss an obstacle of quantitative research that results in some obstinacy of models. Theory and model development are in both cases evidence-based due to the way how they are derived from proxies. The O(bservation)-C(laims/Hypotheses)-E(vidence)-R(easoning) pattern [7] starts with some observations and detection of hypotheses about these observations. Hypotheses are transformed into claims and research questions that form a research agenda. Evidences are then either systematically elicited from data, from previous investigations, or from the belief and knowledge space. Reasoning should then connect evidences to the claims. The results are some kind of Bayesian formulas representing the claim with the evidence. Evidence-based reasoning combines therefore inductive and abductive reasoning. It is enhanced by Occam's razor approaches [19] that allow to finalize the model development. It can be combined with Solomonoff induction [24] that enhances a result (1) by conduction of experiments that will test the claims and (2) by provisionally accepting the claim if the experiment confirms the claims. It can be combined with Epicurus' principle of keeping multiple explanations that allowing consideration of several models and theories as long as they are consistent with the observations. The reasoning schema follows the pattern: (1) *induction/abduction; (2) retrospection; (3) observation_concepts; (4) theory_offer.*

OCER pattern are the basis for evidence-based proxy reasoning (e.g. in the CRC 1266 [1]) that follow positive evidences. Evidence-based reasoning is based on the following principles:

Models represent only acceptable possibilities (each model captures a distinct set of possibilities to which the current description refers) which are consistent with the premises and the knowledge gained so far what makes them intrinsically uncertain because they mirror only some properties they represent.

Models are proxy-driven (the structure of the model corresponds to the proxies it represents. They might also include abstractions such as negation. Models represent only what has been observed and not what is false in contrast to fully explicit models (that represent too what is false). The accuracy of the world view depends on the accuracy of proxies that are considered and the richness those.

We use pragmatic reasoning schemata (e.g. A causes B; B prevents C; therefore, A prevents C). Evidence-based reasoning thus makes a difference between deterministic conclusions (A cause B to occur: given A then B occurs) and ordered sets of possibilities (A enables B to occur: given A then it is possible for B to occur).

5 Conclusion

Models and model suites are one of the main instruments in science and technology. They support reasoning in various forms, e.g. by systematic revisable modelling based on data and as an associated collection of models. This paper develops an approach for development of fully fledged models (a) with extrinsic parts similar to usual (normal) models and (b) with intrinsic parts which are typically hidden in the modelling approach, in the background and context of the model, and in the intentions behind the model. While making this explicit, we are able to use a model as a problem description and to compute the solution of the problem under consideration directly from the model. The paper presents the first methodological part of this solution. The development of corresponding tools and the implementation are topics of a forthcoming paper.

The presented layered approach should not be applied as a 1-2-3-4-5 waterfall sequence of activities. Rather, model development and model utilisation use an evolutionary approach that returns to previous steps whenever sufficiency characteristics of models become problematic within the application domain. The layers can however be considered as phases of development. We notice that our layered approach also supports model revision and model evolution. It can also be used for model migration and model reengineering.

The layered approach seems to be combinable with modelling cultures, e.g. those that can be observed for our first case study [14]. The approach is based on a separation of concern within an initialisation layer, within an intrinsic and this implicit setup layer, within an extrinsic and thus explicit source reflection layer, within an operational customisation layer, and finally with a model delivery layer.

References

[1] CRC 1266. Scales of transformation - Human-environmental interaction in prehistoric and

archaic societies. Collaborative Research Centre. http://www.sfb1266.uni-kiel.de/en/

[2] L.B. Alberti. *On the art of building in ten books*. MIT Press, Cambridge. Promulgated in 1475, published in 1485, 1988.

[3] R. Berghammer and B. Thalheim. Methodenbasierte mathematische Modellierung mit Relationenalgebren. In [31], pp. 67–106.

[4] M.R. Berthold, C. Borgelt, F. Höppner, and F. Klawonn. Guide to intelligent data analysis. Springer, London, 2010.

[5] C. Blättler. *Das Modell als Medium. Wissenschaftsphilosophische Überlegungen.* In [31], pp. 107–137.

[6] S. Bosco, L. Braucher, and M. Wiechec. *Encyclopedia Britannica*, Ultimate Reference Suite. Merriam-Webster, 2015.

[7] E. Brunsell. Claims, evidence and reasoning. www.explainthatstuff.com

[8] A. Dahanayake and B. Thalheim. Co-evolution of (information) system models. In *EMMSAD 2010*, *LNBIP 50*, pp. 314–326. Springer, 2010.

[9] A. Dahanayake and B. Thalheim. Development of conceptual models and the knowledge background provided by the rigor cycle in design science. In: Models: Concepts, Theory, Logic, Reasoning, and Semantics, pp. 3–28. College Publications, 2018.

[10] D. Embley and B. Thalheim, eds. *The Handbook of Conceptual Modeling: Its Usage and Its Challenges*. Springer, 2011.

[11] B. Van Fraassen. *The scientific image*. Clarendon Press, Oxford, 1980.

[12] J. Gray. eScience: A transformed scientific method. Technical report, Talk given Jan. 11, 2007. http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTBeScience.ppt, MS Research Pub., 2007.

[13] ISO/IEC. Information technology - process assessment - part 5: An exemplar process assessment model. FCD 15504-5:2004, 2004. Not publicly available.

[14] H. Jaakkola and B. Thalheim. Modelling cultures. In: Proc. EJC 2018, pp. 33-52, Riga, TTI, Latvia.

[15] K. Jannaschk. Infrastruktur für ein Data Mining Design Framework. PhD thesis, CAU, Kiel, 2017.

[16] Y. Kropp and B. Thalheim. Data mining design and systematic modelling. In Proc. DAMDID/RCDL'17, pp. 349-356, Moscov, 2017. FRC CSC RAS.

[17] M.S. Morgan and M. Morrison, editors. Models as mediators. Cambridge Press, 1999.

[18] O. Nakoinz and D. Knitter. Modelling Human Behaviour in Landscapes. Springer, 2016.

[19] W. Ockham. Philosophical writings: A selection. Hackett Publishing Company, Indianapolis, translated and edited from writings, early 1300 edition, 1990.

[20] A. Raab-Düsterhöft. Integrating social media information into the digital forensic investigation process. In Models: Concepts, Theory, Logic, Reasoning, and Semantics, Tributes, pp. 29-43. College Publications, 2018.

[21] H.-J. Rheinberger. Experiment - Differenz - Schrift. Basilisken-Presse, Marburg an der Lahn, 1992.

[22] A. Samuel and J. Weir. Introduction to Engineering: Modelling, Synthesis and Problem Solving Strategies. Elsevier, Amsterdam, 2000.

[23] G. Semper. Die vier Elemente der Baukunst. Braunschweig, 1851.

[24] R.J. Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. IEEE Transactions on Information Theory, IT-24:422-432, 1978.

[25] B. Thalheim. The Conceptual Framework to Multi-Layered Database Modelling based on Model Suites, vol. 206 of Frontiers in Artificial Intelligence and Applications, pp. 116-134. IOS Press, 2010.

[26] B. Thalheim. Towards a theory of conceptual modelling. JUCS, 16(20):3102-3137, 2010. http://www.jucs.org/jucs_16_20/towards_a_theory_of.

[27] B. Thalheim. The conceptual model ≡ an adequate and dependable artifact enhanced by concepts. In Information Modelling and Knowledge Bases XXV, Frontiers in Artificial Intelligence and Applications, 260, pp. 241-254. IOS Press, 2014.

[28] B. Thalheim. Conceptual modeling foundations: The notion of a model in conceptual modeling. In: Encyclopedia of Database Systems. Springer, 2017.

[29] B. Thalheim. General and specific model notions. In: Proc. ADBIS'17, LNCS 10509, pp. 13-27, Cham, 2017. Springer.

[30] B. Thalheim. Normal models and their modelling matrix. In: Models: Concepts, Theory, Logic, Reasoning, and Semantics, pp. 44-72. College Publications, 2018.

[31] B. Thalheim and I. Nissen, eds. Wissenschaft und Kunst der Modellierung: Modelle, Modellieren, Modellierung. De Gruyter, Boston, 2015.

[32] Vitruvius. The ten books on architecture (De re aedificatoria). Oxford University Press, London, 1914.