Modeling: The Human Approach to Science (CDT-8)

Preprint · May 2019

DOI: 10.13140/RG.2.2.29240.39683/1

Luciano da F. Costa
University of São Paulo

Modeling: The Human Approach to Science (CDT-8)

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

May 26, 2019

Abstract

The activity of modeling, the basis of the scientific method, is approached in an accessible way, including discussions of its properties such as generality, completeness, and relationship with natural language. The features and limitations of modeling are studied in terms of the mathematical concept of invertible mappings as well as in a framework incorporating network science concepts.

"Quella scienza è piú utile della quale il frutto è più comunicabile."

Leonardo da Vinci.

1 Introduction

Getting up with the first rays of light. We ring the bell as we visit a friend. You read this article... To live is to exchange information with the environment, or in other words, to *communicate* in an ample sense.

Yet, there are preconditions for staying informationally integrated in an environment. The main one is that we should have some kind of *model* of it, so we can recognize the entities in the world, and know how they behave. Informally speaking, a model is a simplified representation of some entity or phenomenon that conveys, to a certain level of accuracy, information about its organization and behavior, so that we can not only better understand it, but also make valuable predictions (e.g. [1, 2, 3]).

So, our modeling of the world includes facts such as that morning starts with the first rays of lights. It also contains the information that people announce visits through bells. And we need to recognize words and their meaning so we can read and communicate with other people. At a more sophisticated level, our model of the world also includes language, and the meaning of words, as well as information about ourselves and other humans. All in all, modeling is perhaps one of the most quintessential abilities of life, and especially human beings.

It is then hardly surprising that *science*, from its earliest days, has relied so much on modeling, giving rise to the *scientific method*, outlined in Figure 1 and explained in more detail along this didactic text.

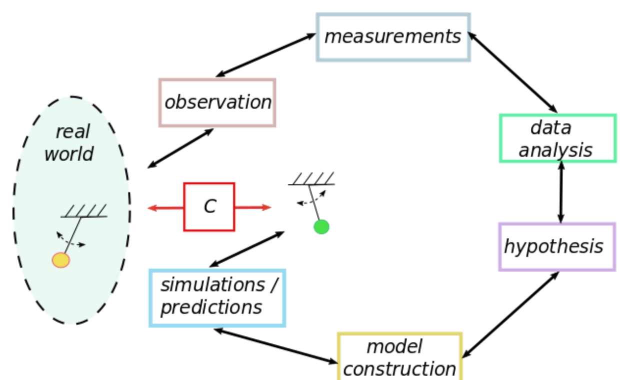

Figure 1: While modeling, one often starts by observing some entity or phenomenon (e.g. pendulum oscillation) and then perform experiments aimed at measuring related properties (e.g. *parameters* such as mass and shaft length; as well as *variables* including time, angular position, etc). By analyzing these measurements, it is possible to derive some hypothesis about the workings of the studied entity, including relationships about its properties (e.g. the angular velocity is related to the gravity), leading to a set of equations and/or rules describing the phenomenon, yielding a preliminary *model*. Through calculations and/or simulations, it is possible to derive predictions about properties of the entity, and the validity of the model can then be checked through new experiments aimed at comparing (box C) the predicted and actual features of the entity

.

Though involving several decisions and choices, and being therefore subjected to mistakes and incompleteness, the scientific method constitutes our best approach to understanding and predicting nature. Thus, in a sense, it was thanks to this recurring modeling approach that the modern world to a great extent has been shaped, including major advances in medicine, food production, communications, and leisure, to name but a few.

In the present work, we address the enticing, essential subject of scientific modeling in an introductory, accessible, and multidisciplinary way. We start by describing its basic framework, and then proceed by illustrating the development of a specific model (regarding a simple pendulum), while discussing the potentials and limitations of modeling.

2 The Scientific Method

Let's consider the overall framework of scientific modeling as outlined in Figure 1. As we have a cyclic overall structure, the construction of a model, in principle, can start at any of its stages. However, for simplicity's sake, consider that it initiates in the physical world with the observation of some phenomenon of interest, e.g. a pendulum.

These observations provide insights for selecting more specific properties to be studied. In the case of a pendulum, these could include the angular position of the bob, the length of the shaft, etc. These properties are then systematically measured, with unavoidable errors and limited accuracy, through well designed experiments in which the entity/phenomenon of interest is isolated the most completely as possible from the remainder of the universe (e.g. keeping the pendulum in a mechanically stabilized platform so as to avoid environment vibrations, producing vacuum, etc.).

Several measurements are repeatedly taken, yielding a *quantitative* description of the studied properties. Two main types of quantitative properties are usually considered: *parameters*, which are kept constant during each experiment and simulation; and *variables*, quantities of interest that can vary during experiments. In the case of the pendulum, we have the mass of the bob, length of the shaft and the gravity acceleration as parameters, while the time and angular position are possible variables.

By analyzing the obtained measurements, e.g. by using statistical methods, it is often possible to derive insights about relationships between the involved variable and parameters, such as the possibility that the angular speed relates to the gravity. These hypotheses are then combined through mathematical relationships/equations, and also possibly through some rules. This represents a preliminary model of the studied entity. By performing calculations and/or simulations (especially in the case of more sophisticated models), it is possible to obtain predictions about the properties of the object, and these can be compared to real objects by performing new experiments. This latter stage, essentially important for modeling, is often called *model validation*.

More often than not, little agreement is observed between the predictions and real properties, which motivates further working on the modeling. This is done by repeating the above described stages, until two situations may arise: (a) a relatively good agreement is obtained, and the model is provisionally validated to a certain accuracy; or (b) no agreement can be reached between prediction and the real properties, and the model requires revision.

Even in the former case, it is important to keep in mind that a virtually infinite set of experiments, covering all possible situations and parameter configurations, would be necessary in order to corroborate the model. On the other hand, if just one counter-example is systematically verified, the model needs to be revised or eventually abandoned. So, we have a first important fact about scientific modeling:

(1) – A scientific model cannot be proved, remaining forever open to revision.

Yet, the above outlined approach, more formally known as the *scientific method* represents our current best way to model the real world. Its essential feature consists in the *validation stage*. It is only thanks to this step that discrepancies and errors in the hypothesis and modeling can be identified, allowing the model to be eventually improved.

More formally speaking, we can understand this validation as a *negative feedback* aspect of modeling, in which identification of discrepancies between predictions and reality can lead to model revisions or eventual abandonment. In case the validation stage is removed, the facts implied by the model can only be taken as unconfirmed hypotheses. In brief, we can conclude that:

(2) – Validation is critically important for scientific modeling.

In the next section, we will discuss how modeling can become an intrinsic ability of living beings.

3 Life, Language, Modeling

In this section we discuss how modeling can become essential for the interaction between living beings.

Let's consider the simple universe represented in Figure 2, which we shall call Ω. This 2D aquatic world has its own geography, with bays under the flow of currents, as well as three species of living beings: a kind of *green alga*, *yellow ellipses*, and *orange rectangles*. Both the latter two species feed on the green alga (attached to rocks and getting energy from light and feeding on minerals), but the rectangles also prey on the yellow ellipses. All these entities are dragged by the currents into and out of bays and, as they drift, feed on the alga when they contact it. Also, the rectangles eat the ellipses when they collide one another. The ellipses survive because they can reproduce faster than the rectangles.

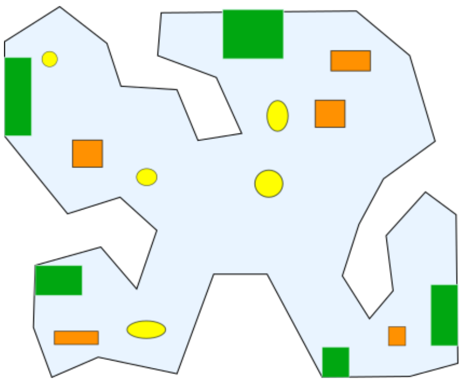

Figure 2: The hypothetic Ω aquatic world contains algae (green), and two ani- mals: yellow ellipses and orange rectangles. The latter two drift freely, under the effect of currents, along bays containing algae and other animals. Ellipses and rectangles feed on algae, while rectangles also prey the ellipses. The acquisition of modeling abilities can confer great advantages to the ellipses and rectangles.

Life in such a world is simple, as each individual has to do virtually nothing other than let to be taken by the currents and feeding by chance. Now, consider the advantages that the ellipses or the rectangles could acquire if they could get more information about their environment, and also being allowed to move intentionally. This could happen, for instance, as a consequence of mutations and evolution. Let's say the ellipses acquire the ability to see around them, and to move. These resources could enable this species to develop means to recognize the algae, and to swim purposefully towards them; or identify rectangles and flee from them.

In other words, now information about the surrounding environment is *communicated* to an individual and processed by it, possibly inducing a respective action-response as a consequence. The importance of this information loop cannot be by any means exaggerated for one very important reason, for it reveals that the individuals have developed a *model* of the environment. This does not mean that these simple species would perform anything like the framework in Figure 1. Nevertheless, each individual now has a *representation* of some of the entities in the surrounding world, allowing their respective *recognition* as well as, possibly, the *prediction* of events involving these entities. These models can be progressively improved by incorporating recognition of new individuals (from the same or other species), which confers substantial advantages to the circles, improving their chances of nourishment, predator avoidance (or prey identification), and finding partners.

With time, the models may even become able to include the models of other individuals/species, and the interaction between these models tend to become more and more complicated. The development of a central nervous system, capable of information processing and memory, may pave the way to even more sophisticated and adaptable modeling. Probably, human intelligence, to a good extent, is the consequence of such interactions between models, mediated by neuronal networks. So, we have a second central aspect of modeling:

(3) – Models are required for interacting with the environment and can improve the survival chances of individuals and species.

To a good extent, humans ended up depending critically on modeling, in conscious or subconscious ways, several aspects of our nature, lives and actions in order to improve our chances of success. Interestingly, the development of language itself was probably a consequence of the need of communication allied to modeling requirements. Indeed, each word in a language is a model of something: actions, entities, and their properties/relationships. Perhaps we could understand language, combined with logic, as the modeling framework underlying human activities, so that:

(4) – Human language represents a modeling approach to the world.

Another interesting way to understand the relationship between language and modeling is that the former is, typically, a prerequisite to the latter. Indeed, societies capable of communicating models and ideas can become more advanced and effective in their interactions with the environment, to the point of including its preservation while continuously advancing in knowledge.

One of the important consequences of achieving communication between modelers is that collaborative systems can arise, capable of dealing with much more sophisticated systems, and leading to more complete models of nature. To a large extent, that has been the case with human science.

4 Modeling as Mapping

Figure 3 illustrates a mathematical construction allowing us to better understand modeling: an entity or phenomenon from a domain A (e.g. nature) is mapped by a function f into a respective description or representation in the domain B (e.g. the mind and/or workspace of the modeler). In this example, the representation corresponds to the English word associated to the original object, namely a pair of glasses, spectacles, or *specs* for short. Though this example considers a human-made object, reflecting the importance of modeling for human behavior, natural entities such as a pendulum could be treated in an analogous way.

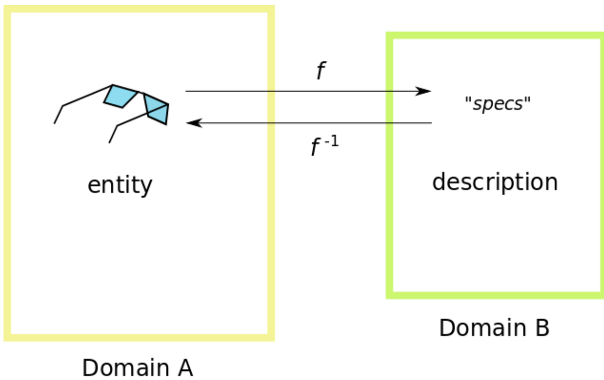

Figure 3: An object (a pair of glasses) in a domain A being mapped, through the function f, into a respective description/representation (e.g. the substantive 'specs') in the domain B. In this particular case, we consider an one-to-one, bijective mapping between the original object and its representation.

This modeling framework implies that, whenever the original object is presented to the modeler, it will be recognized as a pair of glasses, or 'specs'. Since the have assumed that the map f is one-to-one, and therefore invertible through f^{-1}, we have that the original object could be perfectly specified from its respective representation. The mapping f, and its inverse f^{-1} (e.g. [4]) can be understood effectively as corresponding to the *model* of the original entity. If the changes that the object can undergo along time (e.g. folding the temples of the 'specs') are explicitly taken into account while developing the model, it may also become capable of predicting aspects of the dynamical behavior of the object.

Though this first modeling situation was considered to be invertible, this is hardly possible in practice, as such complete mappings would require a lot of information about the object (e.g. the lenses opacity and color, shape size, materials, weight, etc.). Achieving bijective mappings also depends on the set of entities present in domain A. For instance, we can understand that, in the case of the previous example, the mapping was invertible because there was only one object in domain A. Observe that, in this case, the original object cannot be reconstructed from the respective abbreviated description. The inverse only became possible because we imposed the *restriction* that there was only one such an object in domain A. Actually, as we shall argue, *perfectly* invertible mappings (and therefore modeling) of objects from the real world are impossible to be obtained.

Let's consider a more realistic situation, illustrated in Figure 4, in which several instances of the object to be modeled are present in domain A. The relatively small differences in the objects properties (e.g. size, shape and color) allow each of them to be mapped into the same description as a consequence of the intrinsically expected degree of *tolerance* and *generalization*. As a consequence, each of the objects from the same category, including the one we are particularly interested, will be equally mapped by f into the same description *'specs*. In fact, good descriptions of world objects involve some degree of generalization in order to obtain *abstraction*, but this feature at the same time implies some degree of uncertainty in identifying specific instances of that type of object.

Observe that adjectives can be added to the description (e.g. a green specs) in order to make the mapping more restrictive, but it would be completely impossible to fully specify all minute properties of every single specs in the real world. Actually, the essence of obtaining effective linguistic representations depends directly on considering the entity properties that are more typical of that category and more likely to distinguish it from the other existing objects.

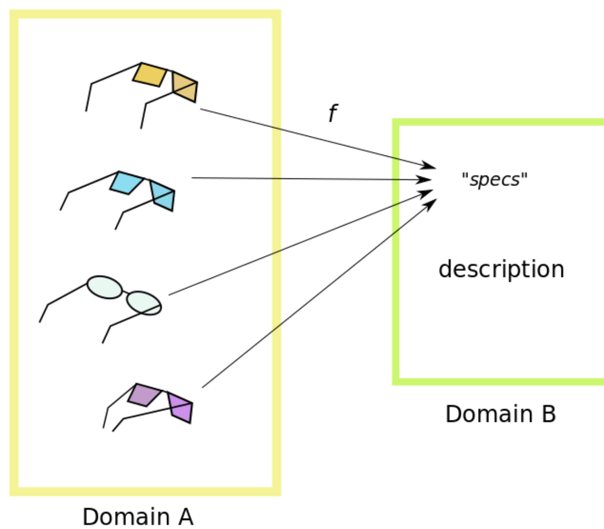

Figure 4: A more realistic situation in which there are multiples instances of the same entity in the domain A.

The problem with such multiple mappings is that, mathematically speaking, they will no longer be *invertible* into a single entity. We can try to recover all the instances of the recognized entity from the respective description, but them we would not be sure about which of them corresponds to the original entity of interest. So, we have a new fact about modeling:

(5) – The effectiveness of a model depends on the degree of similarity and multiplicity of the considered entities.

There are two main possibilities that can be considered in order to solve this problem. One is to derive a more complete model of the entity, including additional properties and behavior. However, in case there is a large number of similar objects in domain A, which is certainly the case for specs in the real world, this possibility becomes unfeasible.

The alternative approach is to try to incorporate some additional *restriction* about the entity of interest. For instance, in case I am interested in specifying my own spectacles, it could be an useful choice to consider the fact that this particular specs would be the one frequently found closer to me. This restriction-based approach is formally studied under the name of *regularization*(e.g. [5, 6]) as a means to solve ill-posed problems. Yet, we should keep in mind that, typically, there is a probability of this approach not working and we recognizing the wrong entity.

All in all, our discussion indicates that the presence of multiple similar objects can substantially make modeling more difficult, often implied in wrong predictions. That is primarily because we are unable to incorporate all minute properties of real entities in our models, including their relationships, which would increase the chance of a more restrictive mapping. In the next section, we will show how the incorporation of network science concepts can allow a much better appreciation of this type of issue and how it can be eventually alleviated.

5 A Network Approach to Modeling

Complex networks (e.g. [7]), provide a versatile means for representing virtually every discrete system in terms of respective networks (or graphs). In particular, every entity to be modeled can be decomposed into smaller parts. For instance, the specs in Figure 3 can be decomposed (simplistically) into terms of two temples, two lenses, and a bridge. This approach allows us to revisit the mapping framework in Figure 3, now representing the entities to be modeled in terms of a respective decomposition in terms of a graph, as illustrated in Figure 5.

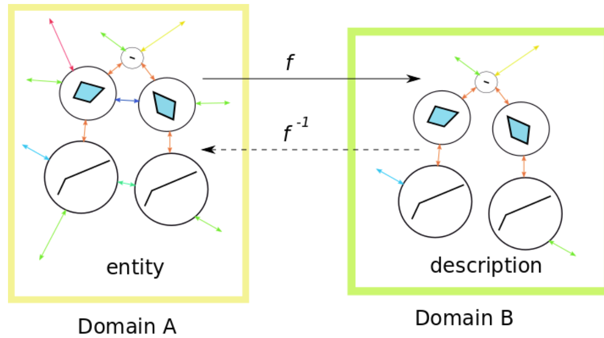

Figure 5: The entity to be mapped can be represented in terms of its parts and relationships between these parts, as allowed by a network or graph. The arrows indicate relationships between the parts of the specs as well as between these and other entities and actions in the domain A. The mapping of the original entity in domain A typically overlooks some of the relationships (or even parts) of the entity being modeled, implying in difficulties in obtaining the inverse f^{-1}.

Note that the relationships in domain A take place not only between the parts of the entity, but also between these and other entities or actions in the real world (indicated by the outgoing edges). For instance, the orange edges indicate attachments between the parts of the specs, while the edge in dark blue may correspond to the fact that both lenses are made of glass. The outgoing edges may indicate that the specs relates to other specs, or that it may be stored into a box, etc.

Modeling the entity in domain A involves identifying its parts and interrelationships, as well as considering relationships with other entities and actions. For instance, in the example in Figure 5, the obtained description misses some of the interrelationships between the parts of the specs, and also with other entities and phenomena. Tough such lacking information may help to generalize a concept, it will also imply difficulties in obtaining predictions of even identifying the original entity.

An example of model that was found to miss some relationship is Newton's second law of motion (i.e. $\vec{f} = m\vec{a}$, where m is the mass, $\vec{f}$ the resulting force, and $\vec{a}$ the acceleration). This important law was found to be not accurate when tested with experiments involving speeds close the the velocity of light, i.e. there was a missing link with this velocity. As a consequence, a relativistic correction needed to be incorporated into the second law, yielding

$$\vec{f} = \gamma m \vec{u} = \frac{d}{dt} \frac{m\vec{u}}{\sqrt{(1-u^2)/c^2}} \quad (1)$$

where γ is the *relativistic momentum* and $\vec{u}$ and u correspond to the speed and respective magnitude of the movement, while c is the speed of light.

We have that the consideration of the many involved components and interrelationships is often a challenge, because relationships that may seem to be little important can turn out to be determinant for prediction. This is a consequence of the fact that several objects and phenomena from the real world are non-linear, so that small perturbations may lead to big consequences (e.g. chaos). Thus, in principle we need to consider all interrelationships, and we have another conclusion about modeling:

(6) – The effectiveness of a model depends on incorporating as many as possible relevant relationships between the parts of the entity of interest, as well as with other existing entities.

The problem of missing information becomes even more complex when we consider that, because of the infinite-range action of fields (e.g. gravitational or electric), every entity and phenomenon in the real world is potentially influenced by every other entity! Indeed, even the moon or a more distant astronomical entity could influence the motion of a chaotic pendulum at the very long term. This defines a striking network of interrelationships between virtually every entity and phenomenon in the real world.

Such facts lead us to another important conclusion:

(7) – A model of a real world entity or phenomenon will always remain uncompleted.

Though this may seem to be a little bit disappoint at first, it can be understood in at least two more positive ways: (i) it motivates an attitude of respect toward nature; and (ii) it paves the way to continuing research and results, keeping scientists occupied and motivated.

6 Prospects and Conclusion

Modeling is a critically important activity underlying science and the scientific model. In this text, we reviewed the concept of modeling from several perspectives. First, we considered the more traditional approach in which modeling is seen as a cycle of activities ranging from phenomenon observation to model validation.

Then, we discussed how modeling abilities can provide great advantages to individuals and species along evolution, to the point of development of languages capable of representing concepts and actions of the real world, allowing us to plan our activities and behavior.

We also approached the activity of modeling in a more formal, mathematical way, in which objects are mapped between two domains, and then complemented this framework by considering networks representing interrelationships. The latter allowed a comprehensive understanding why models of real world entities are unavoidably limited and incomplete, and how these models can be eventually improved by taking more parts and interrelationships into account.

Several important aspects of modeling were revealed as we progressed along this work, several of which indicating the limitations of the modeling approach. Indeed, to accurately model objects and phenomena from the real world often represents a substantial challenge, so that we could say that modeling is difficult, or *complex*. Indeed, the concepts of complexity and modeling have been recently interrelated [8], in the sense that complexity would correspond to the cost of developing the model (f and the inverse f^{-1}) plus the cost implied by the prediction errors. Interestingly, these two costs tend to complement one another, in the sense that more efforts invested into the development of a model tend to yield smaller prediction errors and respective costs.

Thus, in a sense, the important issue of complexity would itself be related to human modeling abilities, and the accuracy of the obtained predictions. We can summarize this interrelationship as:

(8) – Complexity is related to the cost of modeling as well as the cost implied by the prediction errors.

Yet, despite the unavoidable limitations of the modeling approach, which are after all a consequence of the infinite network of interrelationships in which the universe is weaved, modeling remains our best approach to understanding and predicting, to a certain accuracy, the properties of objects and phenomena in the universe.

Though we cannot get fully complete models of reality, recent technological and scientific advances, especially in computing and information sciences, have provided perspectives for obtaining increasingly more complete and accurate models. For instance, development of new materials allow devising more comprehensive and accurate experiments, as well as measurement equipments.

Regarding computational aspects, numerical methods can be used to solve difficult equations involved in modeling (e.g. [9]), while parallel computing can be applied to make these calculations even faster. At the same time, concepts from pattern recognition, deep learning (e.g. [10]), and artificial intelligence (e.g. [11]) are being progressively incorporated into science, allowing automation of several steps in the framework shown in Figure 1. For instance, the stage of data analysis can be automated by incorporating methods capable of recognizing particularly interesting features in large masses of data. Part of these possibilities are being developed in the area of *data science*.

Even the derivation of hypotheses and model construction can be in principle automated. These prospects would allow much larger databases to be analyzed, and provisionally models to be obtained in a faster and more comprehensive way. Complex networks can also provide concepts and methods that can be immediately applied not only to better understanding modeling, but also to enhancing all the involved steps.

So, we reach another interesting conclusion:

(9) – Modeling can substantially benefit from continued incorporation of scientific and technological advances.

Interestingly, the recursive application of results derived from science and technology contribute to the advancement of both these areas, establishing a sustainable *positive feedback* that can substantially boost modeling to unprecedented levels of complexity.

All in all, we believe that the intrinsically human ability of modeling can pave the way to achieving ever improving human standards of leaving while maintaining more balanced and harmonic relationships with nature. To participate in this journey is very motivating and instigating, so that we reach our concluding remark that

(10) – Scientific modeling can be much gratifying.

Acknowledgments.
Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) for sponsorship. This work has benefited from FAPESP grant 15/22308-2 .

Costa's Didactic Texts – CDTs

This is a *Costa's Didactic Text* (CDT). CDTs intend to be a halfway point between a formal scientific article and a dissemination text in the sense that they: (i) explain and illustrate concepts in a more informal, graphical and accessible way than the typical scientific article; and, at the same time, (ii) provide more in-depth mathematical developments than a more traditional dissemination work.

It is hoped that CDTs can also provide integration and new insights and analogies concerning the reported concepts and methods. We hope these characteristics will contribute to making CDTs interesting both to beginners as well as to more senior researchers.

Though CDTs are intended primarily for those who have some preliminary experience in the covered concepts, they can also be useful as summary of main topics and concepts to be learnt by other readers interested in the respective CDT theme.

Each CDT focuses on a few interrelated concepts. Though attempting to be relatively self-contained, CDTs also aim at being shorter than the more traditional scholar article. Links to related material are provided in order to complement the covered subjects.

The complete set of CDTs can be found at: `https://www.researchgate.net/project/Costas-Didactic-Texts-CDTs`.

References

[1] Models and modeling: An introduction. `http://ambitiousscienceteaching.org/wp-content/uploads/2014/09/Models-and-Modeling-An-Introduction1.pdf`. [Online; accessed 26-May-2019].

[2] A. B. Shiflet and G. W. Shiflet. *Introduction to Computational Science: Modeling and Simulation for the Sciences*. Princeton Univ. Press, 2014.

[3] K. Borner, K. W. Boyack, S. Milojevic, and S. Morris. An introduction to modeling science: Basic model types, key definitions, and a general framework for the comparison of process models. In: Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences, 2012.

[4] K. Wortman. Inverse functions. `https://www.math.utah.edu/~wortman/1050-text-if.pdf`. [Online; accessed 26-May-2019].

[5] S. Lu and S. V. Pereverzev. *Regularization Theory for Ill-posed Problems Selected Topics*. De Gruyter, 2013.

[6] A. M. Abidi, A. V. Griboi, and J. Paiki. *Optimization Techniques in Computer Vision: Ill-posed problems and regularization*. Springer, 2016.

[7] L. da F. Costa. What is a complex network? `https://www.researchgate.net/publication/324312765_What_is_a_Complex_Network_CDT-2`, 2018. [Online; accessed 05-May-2019].

[8] L. da F. Costa. Quantifying complexity. `https://www.researchgate.net/publication/332877069_Quantifying_Complexity_CDT-6`, 2019. [Online; accessed 05-May-2019].

[9] The history of numerical analysis and scientific computing. `http://history.siam.org/`. [Online; accessed 26-May-2019].

[10] Prabhat. A look at deep learning for science. `https://www.oreilly.com/ideas/a-look-at-deep-learning-for-science`. [Online; accessed 26-May-2019].

[11] National Academy Press. Artificial inteligence in mathematical modeling. `https://www.nap.edu/read/1909/chapter/6`, 1991. [Online; accessed 26-May-2019].