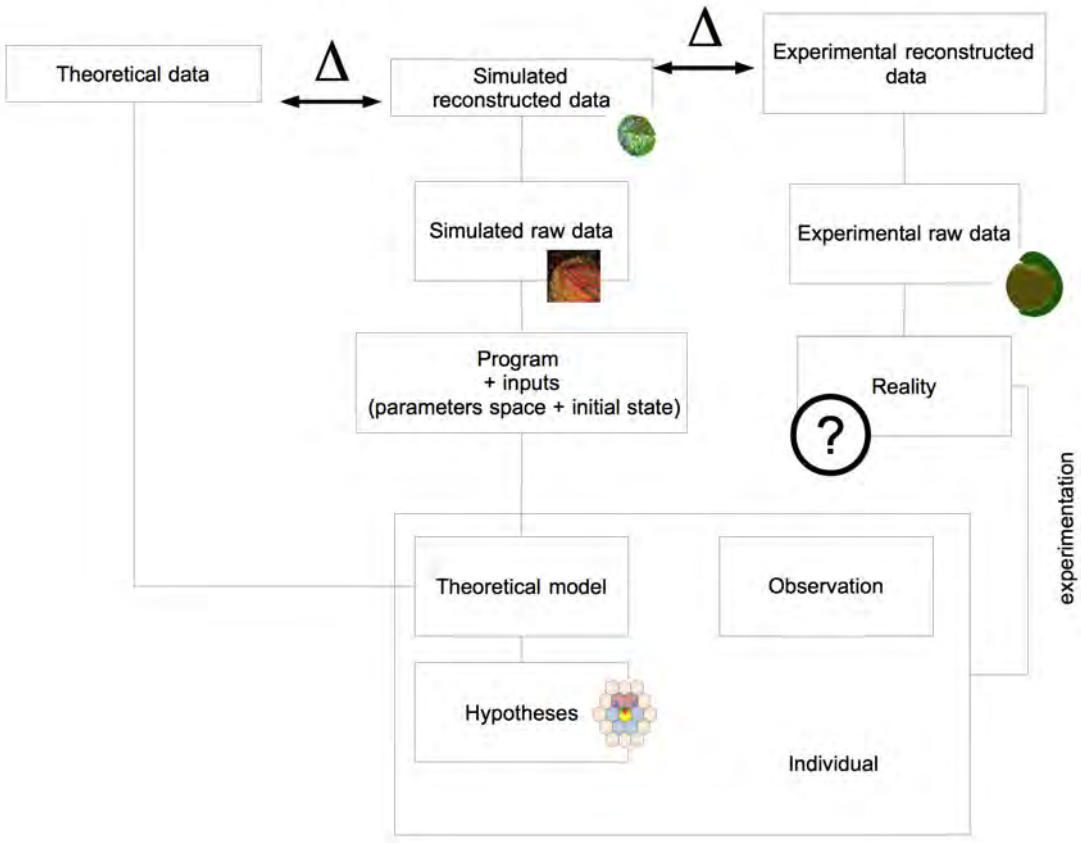


Figure 1.12 :

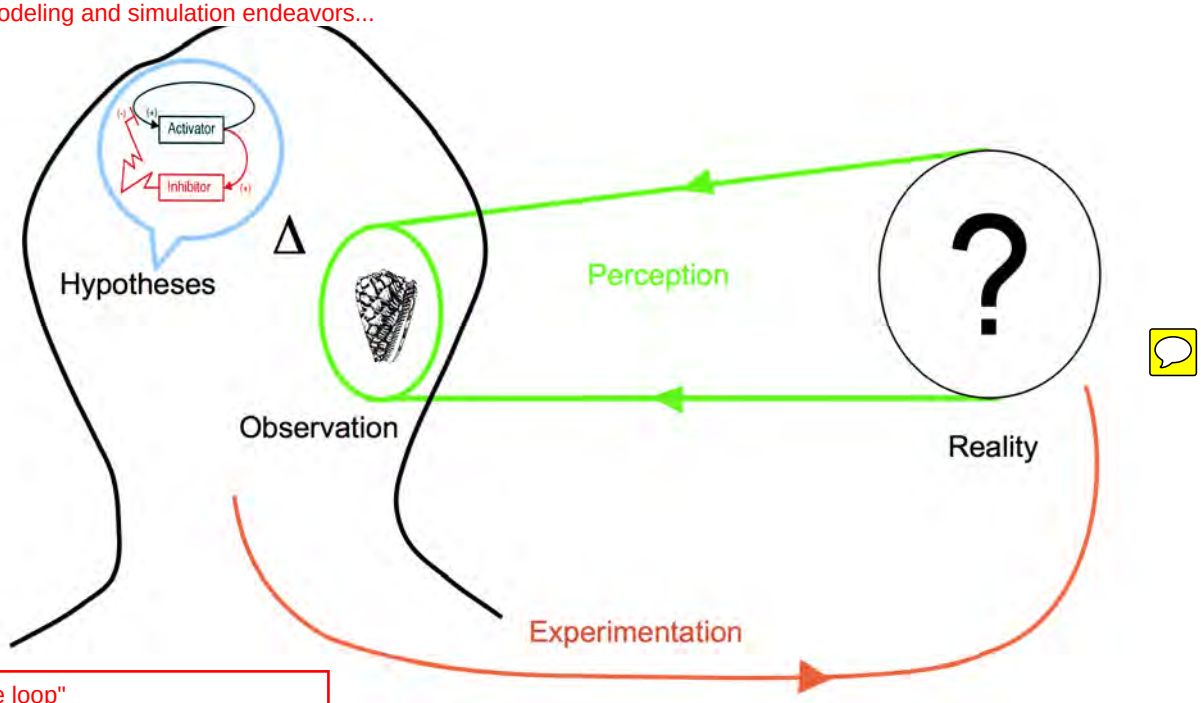
Previous
Next
TOC
References

1.3. Methodology

[195] We will make short but important preliminary remarks about the generation of hypotheses and models by scientists, the tools that they designed to help them in this process, then discuss the notion of "validation" and quality of a model...

Goal of this section : describe how theoretical and more precisely, simulated theoretical hypotheses is integrated in the process of experimental science.

This is not a dissertation in epistemology, but we felt it was crucial to give a clear methodological framework to our modeling and simulation endeavors...



Scientists "in the loop"

Figure 1.2 :

The individual

Experimental science involves an individual and its environment ("Reality"). This activity shares similarities with other of its explanation-seeking activities. Experimental science is characterized by three fundamental processes: the perception of its environment, the generation of new hypotheses and the experimentation on the environment (see figure 1.2).

1. The loop can be entered by the individual perceiving his environment.

2. The observations are then matched with the knowledge of the individual. Most of the time, if the observation conforms to the knowledge, no reactive signal is emitted. However, a significant difference would trigger a signal of curiosity which will challenge the existing set of hypotheses of the individual and lead him to reconsider some of them.

3. The process of how the individual create new hypotheses will not be discussed in this manuscript (analogy, inference, induction, abduction, deduction...). What will be discussed is the conceptual nature of these hypotheses (see below).

4. The "experimental" qualification of sciences like biology or physics is due to their ability to couple the "pure thinking" exercise of generating new hypotheses with interaction on the "real" system in order to test the validity of these new hypotheses.

The environment of the individual is made of multiple potential object of study. The specification of the object of study induces a separation of the object from its own environment which may, or may not, include the individual. In developmental biology, the studied object is the embryo and the individual is excluded from the embryo's environment.

Exchange/Validation by the scientific community

Even if the individual is at the center of experimental science, experimental science is a collective effort. The interaction with the scientific community operates bidirectionally.

All the hypotheses the individual makes are build upon an accumulation of prior scientific works. He can access nearly all knowledge produced by the scientific community thanks to conferences or the various scientific papers databases (Pubmed, arXiv.org, IEEE, ACM, Google Scholar...).

The particularity of science is that validation is made by the approval of the community of scientists. Through the peer-reviewed publication system, each new work is filtered before availability by a panel of individual representing the community. We may distinguish two kind of validation: the validation of the scientific work containing all or some parts of the elements mentionned in figure 1.2, and the validation of the hypotheses contained in the scientific work itself. We will develop the latter in the following (XXXXXX see part III).

Previous
Next
TOC
References

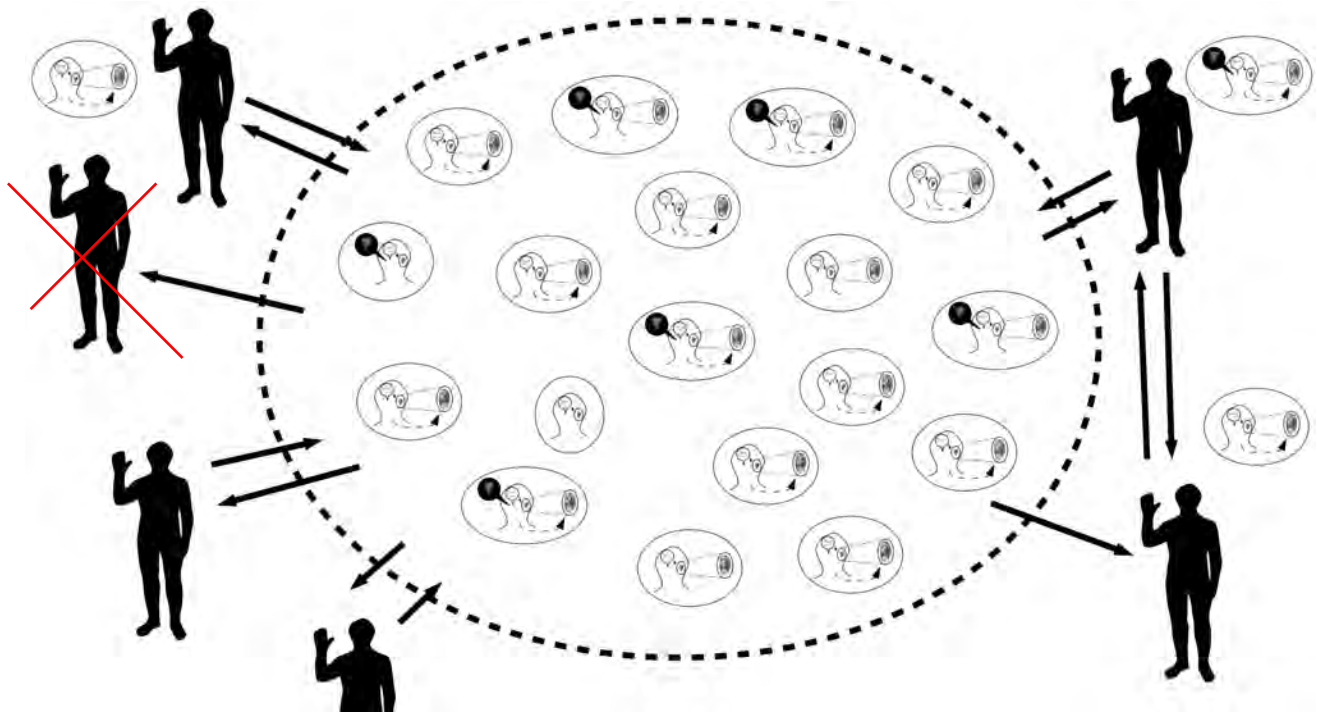


Figure 1.3 : Experimental science is a collective effort. Each member of the scientific community may send or receive scientific work.

Methodology augmented by tools

Experimental science insists on the confrontation of the hypotheses, their consequences and the observation. This confrontation is improved by the means of tools. Tools can be considered as the third actor in experimental science, in addition to the individual and the object of study.

They are also object of study by themselves. In developmental biology, the tools used to observe or perturb are themselves the focus of intensive ongoing research, from other fields of science.

The advances of our understanding are controlled by the advances of theses tools. Always improving microscopy increase the spatio-temporal resolution of the observations and every new microscope triggers a boom of conceptualization of these observations.

The use of new physical tools drives methodological renewals. We will present an augmented version of the figure 1.2 which introduces the methodology developed for this project. We distinguish the tools designed to augment the three fundamental processes of experimental science by the inverse order of usage in this project: tools to manipulate, tools to perceive and tools to conceive.

3. **Tools to manipulate**

These tools are designed to modify the "natural" behavior of the object of study, or its environment, in a controlled manner. These experiments are artificial construction allowing to discriminate the hypotheses ruling the behavior of the object of study. In developmental biology, the embryo can be perturb either genetically or mechanically. Genetic experiments comprise embryo expressing abnormal phenotype either by random mutagenesis or by morpholino injection (knocking down of specific gene). Mechanical experiments can be either lesion applied to some specific tissue to study their evolution (see laser ablation between individual cell-cell boundaries [48], or tissue dissection by laser XXXXXXXXX find ref) or mechanical constraint to measure the response of the tissue. Mechanotranduction mechanism allow to conceive experiments at the border between genetics and mechanics (see genetic regulation by exerting a force with magnetic tweezer and magnetic nanoparticles [38]).

1. **Tools to perceive**

Tools can improve the perception of the object of study, from the interfacing between the real system and the observed data, to the reconstruction of these data which extracts salient features in the observation.

Perception-oriented tools are object whose aim is to produce measures of the object of study. Measures are quantification of the physical quantity of the object of study with ordinary real numbers. It may noted that if these tools add a significant objectivity in the measurement process, allowing interpretation free comparison between measures, the physical quantity *per se* are determined by the current knowledge and thereby are not are subject to interpretation.

Previous
Next
TOC
References

An improvement of the perception-oriented tools is the widening of the nature and scales of the physical quantity measured, accessing dimensions out of reach of the individual senses. In developmental biology, optical devices allowed to reach the sub-cellular scale, sending photons to camera sensors with high spatial resolution.

Sometimes, perception-oriented tools must be coupled with experimentation to allow measurement. For example, the zebrafish, which

has been selected in part for its transparent characteristic at early stages, is modified by injecting fluorescent-protein coding RNA to highlight some structure of the cells such as membranes, or nuclei when exposed to laser stimulation.

The widening of scales induces the increasing of the size of generated data. This tendency is amplified by the coupling of measurement tools with automated recording with computers. When an embryo is measured under microscope device, the size of the data recorded is enormous. For a few hours of development, this measure, which is the spatio-temporally discriminated quantity of light emitted by a laser-excited fluorescent zebrafish, is composed of billions of values (for example, 200 3D-volume of voxels of light quantity obtained each 3 minutes, each volume having a resolution of 512x512x200, gives 10.48576 billions of values). This data size may also be multiply by the number of channel used for excitation (for example if both the cell nuclei and the cell membranes are captured). We call this microscope output data the 4D (3D + time) raw data set (or *raw data* for simplicity's sake).

This kind of extremely large data is not directly accessible by the individual. He can not gain biological insight from these raw data. Processing of the data is needed to allow interpretation and comparison with hypotheses. We call this processing *reconstruction* of the data. The reconstruction is a series of subprocesses organized as a *workflow*. Each subprocess does a specific task which extract some information from the input data sets and generate a new data set. Computers propose visualization software tools which allows the individual to create 2D movies of the captured developmental sequence. However, if a movie permits qualitative insights, it looses the quantitative measurement which would be used for comparison. A reconstruction which is useful for developmental biology question is extraction from the raw data of the lineage tree of captured cells. For a given cell at a given time step, this data set stores the identification number of the same cell at the previous time step. With the lineage tree, cells can be followed through time and, as cell divides during the embryo's development, and form diverging branches of the "tree". In addition, complementary information can be stored for each cell at each time step: the 3D coordinates, the lists of neighbors, the quantity of captured protein signal, or ligand, or any fluorescent labeled molecules... The final data sets of the workflow compose the *reconstructed embryo*. It stores all the biologically relevant information contained in the raw data set at the cellular scale.

Some parts of the aforementioned reconstruction workflow can be realized with commercial softwares. However, the high number of cells involved and the difficulty to interpret and manipulate the 3D volume of data initiated the design of adapted softwares and the automation of most of these subprocesses. These tasks are the past and current results of the European project Bioemergences which developed a generic workflow of reconstruction of the lineage tree in vertebrate embryos. New specific modules has been developed for this project. The detailed presentation of the reconstruction workflow is done in chapter XXXXXXX.

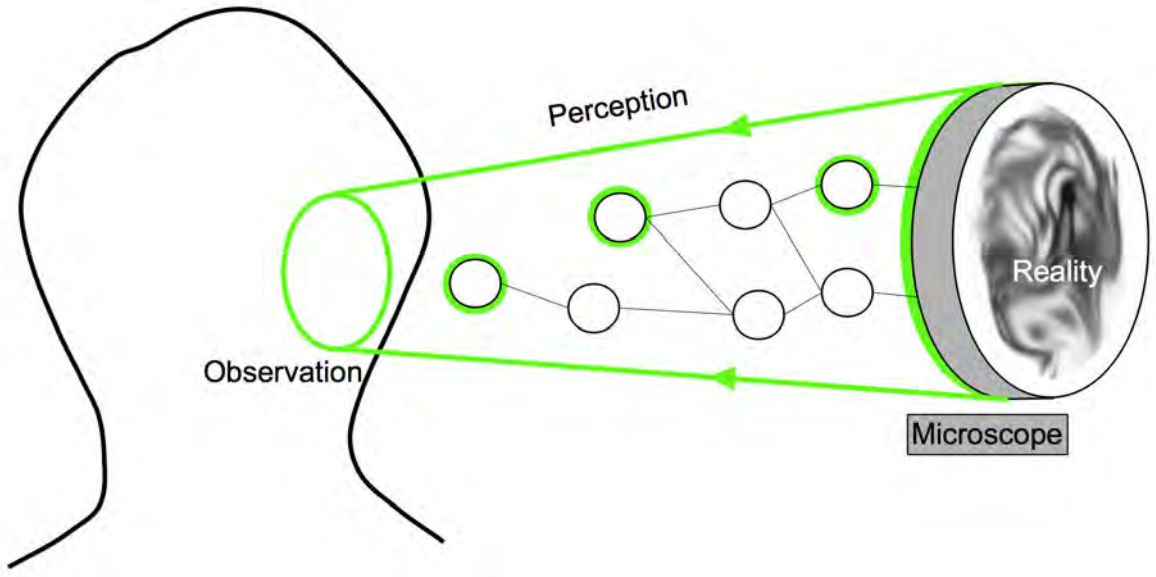


Figure 1.4 : Augmentation of the perception by a reconstruction workflow.

2. Tools to conceive

Models or hypotheses aim at describing the whole picture of the studied phenomenon. They withdraw some supposed details and generalize the underlying mechanisms. They establish relations between the observed data.

Models are constraint by their means of expression. The description of the structures and their interactions varies whether their are expressed through the verbal language, the graphical language or the mathematical language. There exists links between all these means of expression and models are often described through a combination of some of them.

In the context of developmental biology, a distinction is often expressed between "classical" studies and theoretical studies. The former studies often expose "qualitative" models: models highlighting the nature of interaction between elements with no quantitative parameters. The latter studies use the mathematical formalism to express the hypotheses. The elements are represented by variable and their interaction is formulated through equations fine-tuned by parameters. The gain of theoretical studies is an increase of

Previous
Next
TOC
References

predicability of the hypotheses. The consequence can be tested extensively, allowing to automatically reverse engineer the studied phenomenon. This comes at the cost of a "simplification and an idealization, and consequently a falsification" (Turing [192]) of the description of the elements, which may perplexed "classical" biologists who are used to deal with the complexity of living systems.

Either theoretical or "classical", the role played by model remains invariant. However, theoretical models bring some advantages:

- their formalized nature enables the design of precise experimental measures.
- their predictability allow to test new hypothesis by examining the consequences of unobserved phenomenon.
- they allow to integrate a wide range of observation. Without the help of formalization, the consequence of multiple interaction quickly become unpredictable by "pure though" experiment only.

To deepen the previous advantage, we may provide cognitive scientist Marvin Minsky's definition of model in his 1965's text "Matter, mind and models" [118]:

"To an observer B, an object A* is a model of an object A to the extent that B can use A* to answer questions that interest him about A".

This definition is centered around the notion of the question asked by the observer (or the individual in this text). It implies that any attempt of model type categorization should start by categorizing question type. Indeed, a distinction can be made between specific and general questions. In fundamental physics, the distinction is not so clear, as each specific question as generally massive repercussion on everything else. However, in biology, the gap is larger. As an example, we may cite the specific model which studies the shape of cells in a epithelium by Gibson et al. [62], as opposed to generic models aiming at simulating various developmental phenomena like Compucell [85] [25] or Cellerator [171] [172]. The former class asks a question then designs a model to answer it whereas the latter builds an integrative model before answering various potential questions. The latter class of models is proper to the theoretical category and, as we may show in the following, it is even part of a subcategory of theoretical models.

agent-based modeling

As mentioned above, theoretical models formalize interactions with equation linking some selected variable of the studied phenomenon. The solving of these *analytical* formalization is not always doable because of constraint proper to mathematics. But computer is a tool that can help in the resolution of these equations, by converting them into algorithms. These *numerical* solutions, which are approximations of the idealistic solutions, are nowadays used in most fields of research or engineering. They allow scientists to tackle more complex phenomena observed. In 1952, Alan Turing already envisioned the use of computer to help him solve more realistic reaction-diffusion pattern in "The chemical basis of morphogenesis" [192]:

"Most of an organism, most of the time, is developing from one pattern into another, rather than from homogeneity into a pattern. One would like to be able to follow this more general process mathematically also. The difficulties are, however, such one cannot hope to have any very embracing theory of such process, beyond the statement of the equations. It might be possible, however, to treat a few particular cases in detail with the aid of a digital computer. This method has the advantage that it is not so necessary to make simplifying assumption as it is when doing a more theoretical type of analysis. "

Turing emphasizes the fact that the use of computer simulation is not only a practical solution for going over analytically unsolvable mathematical equation but also that it allows the individual to integrate mechanisms that he would refrain from using because of the unsolvability. In this sense, the computer (as a Turing machine) is a tool that augment the ability of the individual to develop mathematical models of the object of study.

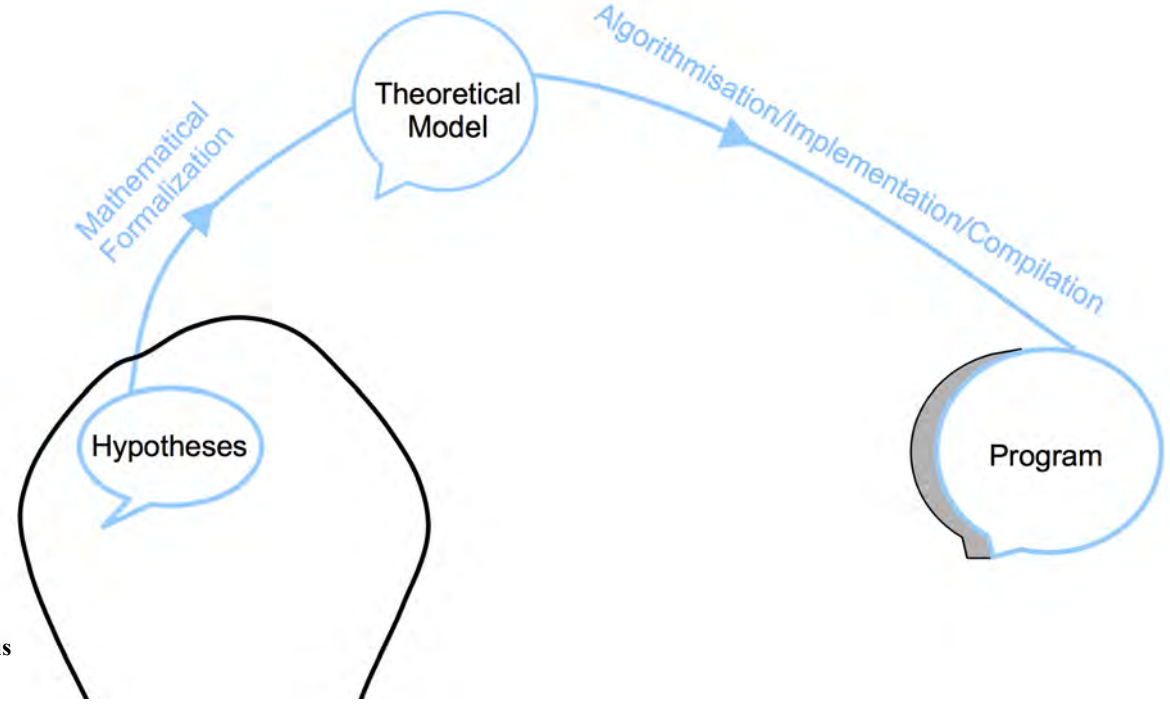


Figure 1.5 :

Previous
Next
TOC
References

In particular, a category of analytically unsolvable model is called the *many-boly problem*. It occurs when a large number of elements are interacting together. As we will present in section (XXXXXXX check), the physical approach we have chosen for our embryogenesis model is based on this assumption, each cell being an elementary particle interacting with its neighbors. Solving this system of equations is highly computationally intensive and requires the use of computers. Computers were originally invented to deal with theses situations (for example, the MonteCarlo simulation performed on the MANIAC computers in the early 1950s [117]).

Figure 1.5 illustrates the process of transformation of the model, from the original hypotheses made by the individual, to its theoretical form and finally to its final form as a computer program.

machine learning

A different kind of theoretical models are the statistical models. We rule out this category as they do not follow the individual-induced hypotheses framework. These models offer ways to simulate observed data without including *a priori* knowledge about these data. They possess predictive capacity but no explanatory value as the model produced is always a *black box* for the individual.

ajouter à la section tools to conceive

Previous

Next

TOC

References

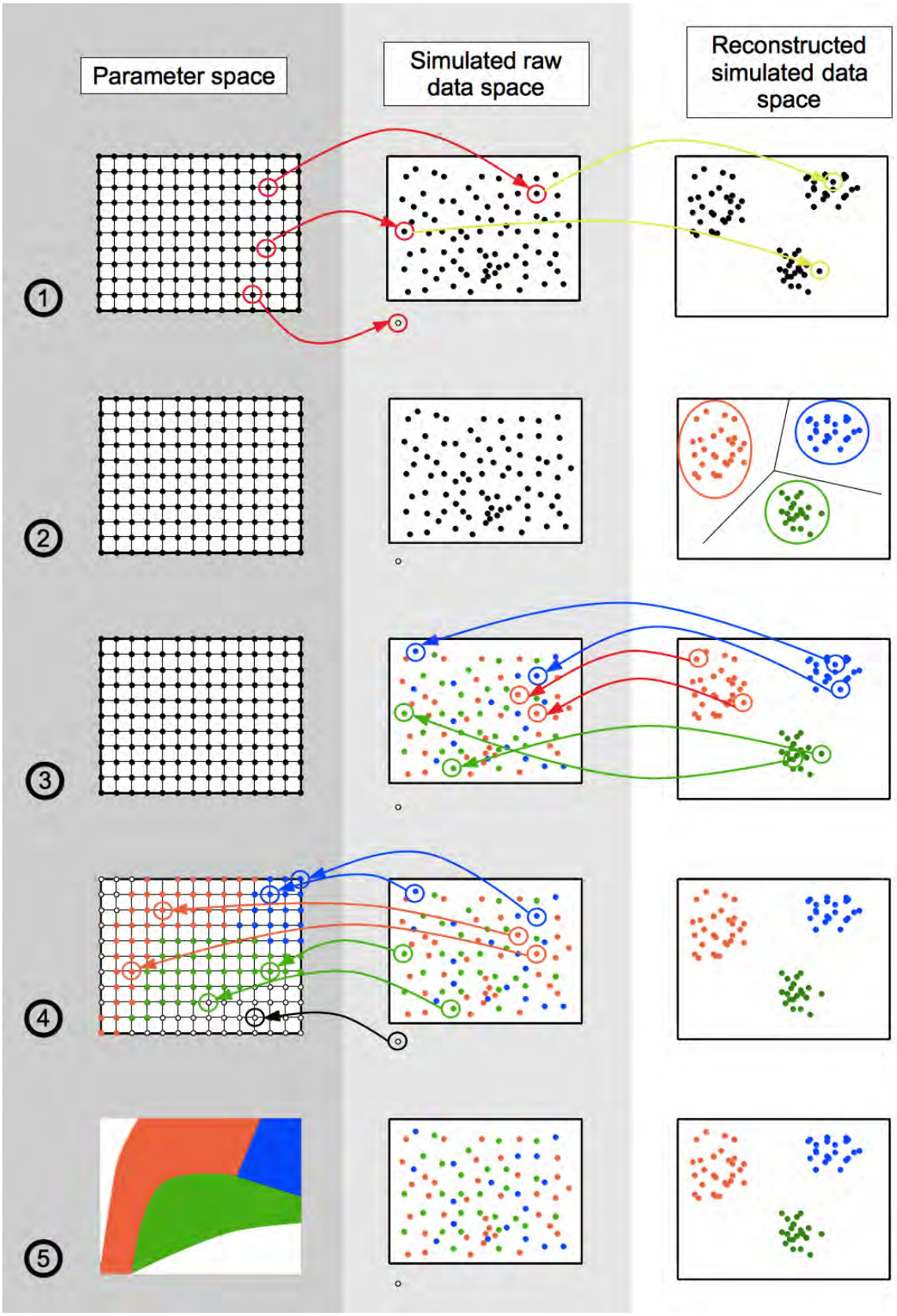


Figure 1.6 :

Previous

Next

TOC

References

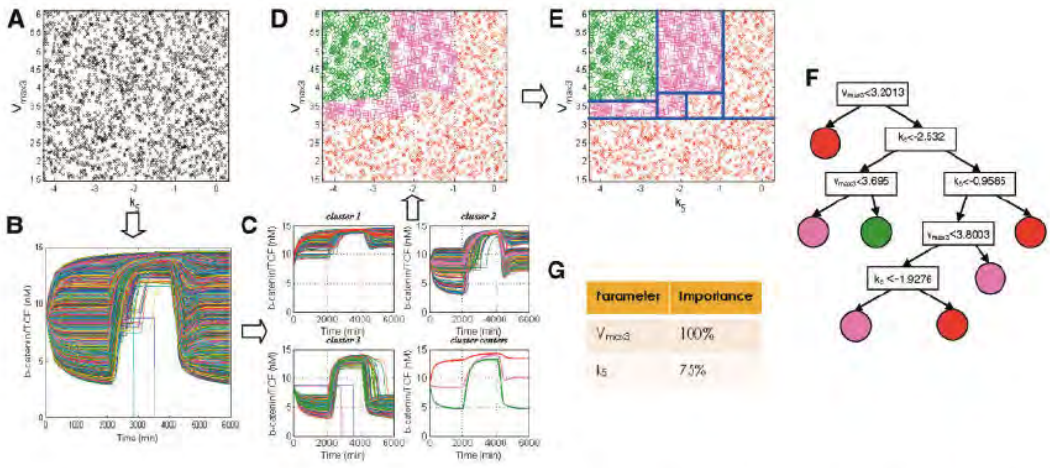


Fig. 1. An example illustrating a 2-parameter perturbation analysis of a pathway model. **(A)** Parameter space is filled with sampling points. **(B)** Plotting of simulation results. **(C)** Clustering shows three distinctive clusters representing model qualitative behaviors. **(D)** Sampling points is colored respecting to clustering results. **(E)** The parameter space is nicely partitioned into sub-regions with different behaviors. **(F)** A tree represents the parameter space partitioning. **(G)** Parameter ranking estimated from the tree.

Figure 1.7 :

Validation of the hypotheses

The term *validation* of an hypotheses employed in this section can not be understood in the sense of stating that an hypothesis is true. Oreskes et al. has demonstrated that establishing the truth of a proposition is possible only in a closed system and that models that used incompletely known input parameters as are models in developmental biology are never closed systems [140]. Popper also advocates that one cannot prove theories and laws and that they can only be falsified [146]. The acceptation of the term *validation* must mean *consistency* between the output of the model and the observation of the object of study. The observations do support the probability of the model [140], or its empirical adequacy [213].

The more observed data are positively confronted to the model, the more adequate it becomes. The diversity of the observed data is also a factor favoring the adequacy of the model. The observed reconstructed data evoked in the previous section are large data sets of various type. The strategy we adopt is to integrate the simulation platform and the reconstruction workflow. It aims at evaluating the adequacy of the model with the assistance of all tools mentioned above. This process of evaluation is equivalent to establishing the fitness of the data generated by the simulation and the observed data, by the means of *fitness function*.

We distinguish two types of fitness function in addition to the original cognitive fitness represented by the symbol Δ in figure 1.8:

- the automated fitness function category, denoted by the symbol Δ_a. These functions require a reconstruction strategy from the data generated by the simulation platform similar to the reconstruction workflow described in the augmented perception part. The simulated raw data are of a different kind than the experimental raw data. The first step of the new reconstruction workflow is to perform transformation of the simulated raw data to match the format of the reconstructed experimental data. This process is represented on figure 1.8, the green dashed line represents the stage on both reconstruction workflow were both reconstructed data are of the same kind. Once data format matches, it becomes possible to design automated *fitness function* to evaluate the discrepancy between both reconstructed data sets and give a quantitative score.
- the visual fitness function category, denoted by the symbol Δ_v. This category exists because of the visual aspect of the data. Each data can be visualized and the individual may intuitively dismissed some hypotheses based on the visualization only.

An orthogonal dichotomy distinguish the fitness function whether the observed data which are matched with the simulated data are originated from:

- the reconstruction of experimental raw data. These fitness function are called *experimental reconstruction fitness* (ERF) functions.
- theoretical data representing idealized phenotypic behavior. We denote theses fitness function by *theoretical fitness* (TF) functions.

Previous

Next

TOC

References

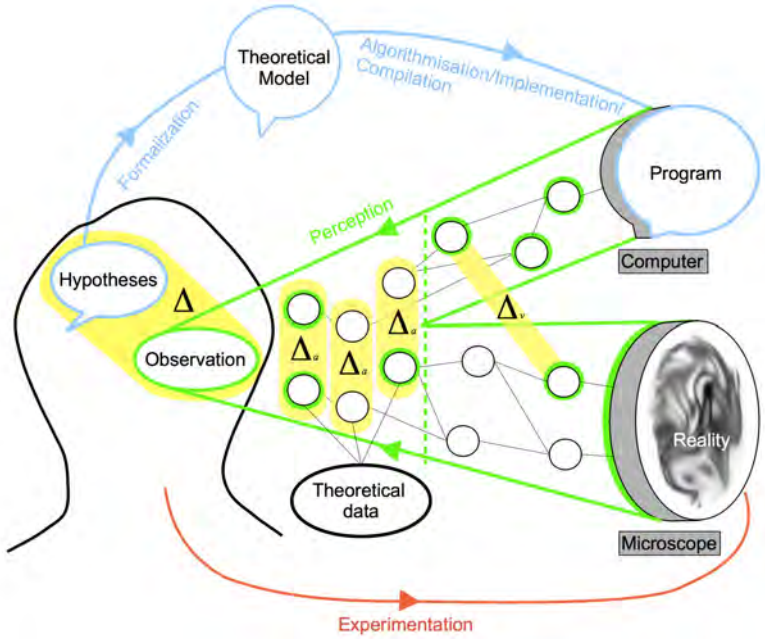


Figure 1.8 :

The automated hypotheses evaluation strategies will be exploited in section 9 (XXXXXXchek) through a series of case studies. The general run of the strategy is to exhaustively map the parameter space with a fitness value in order to :

- assert the probability of the model to show that the hypotheses are sufficient to reproduce the observation in satisfying manner.
- discuss the topology of the map to find different mode of exploitation of the hypotheses.

ancien schéma 1

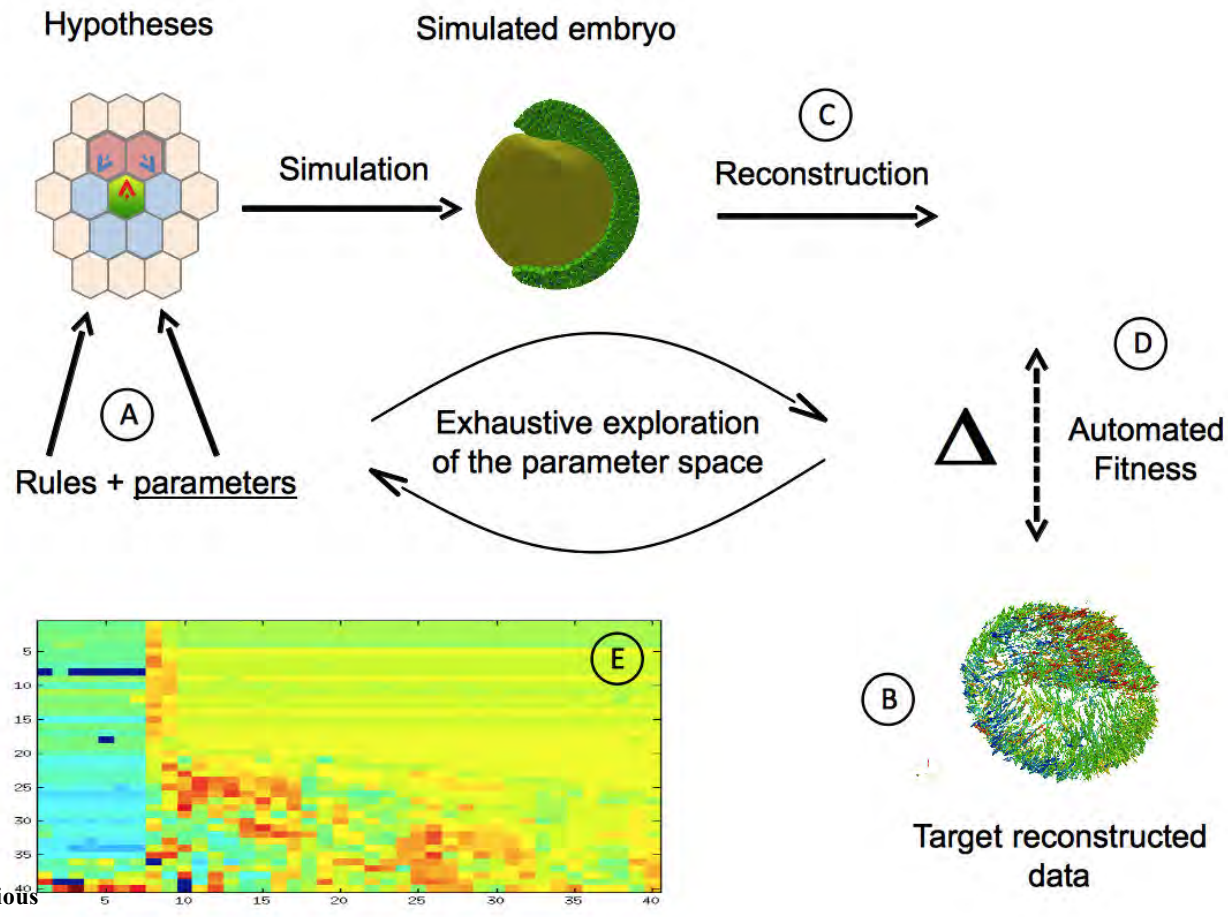


Figure 1.9 :

Previous

Next

TOC

References

ancien schéma 2

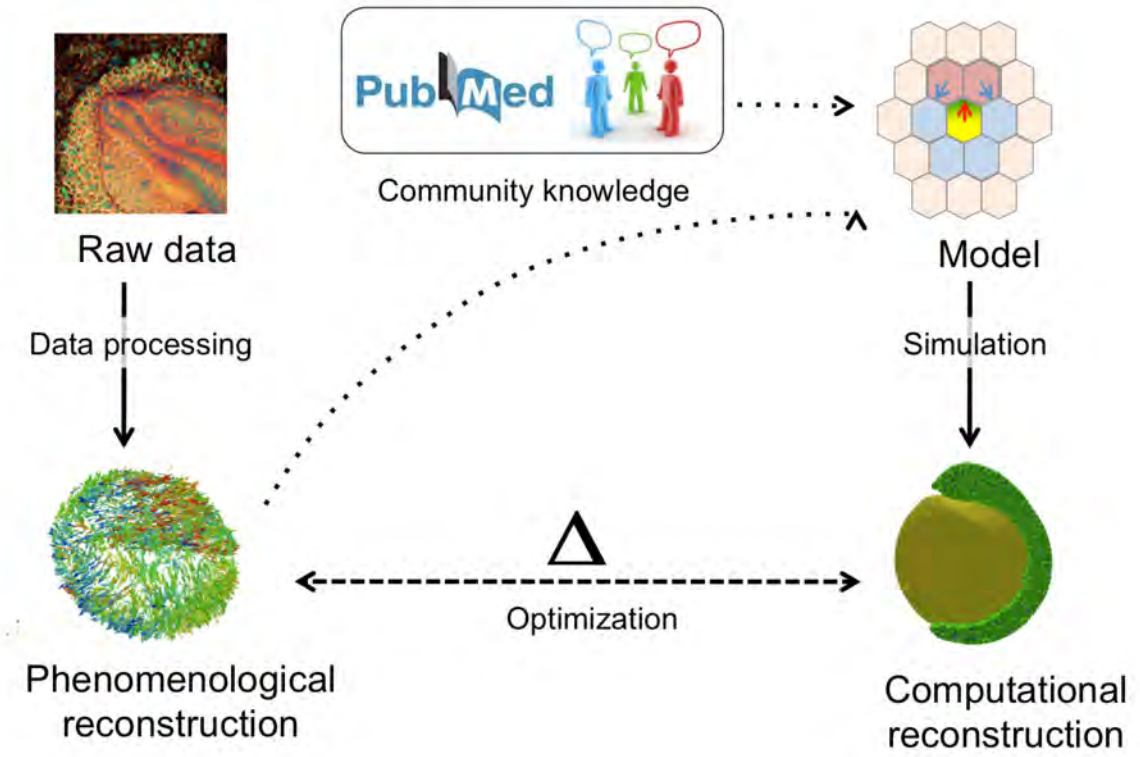


Figure 1.10 :

ancien schéma 3

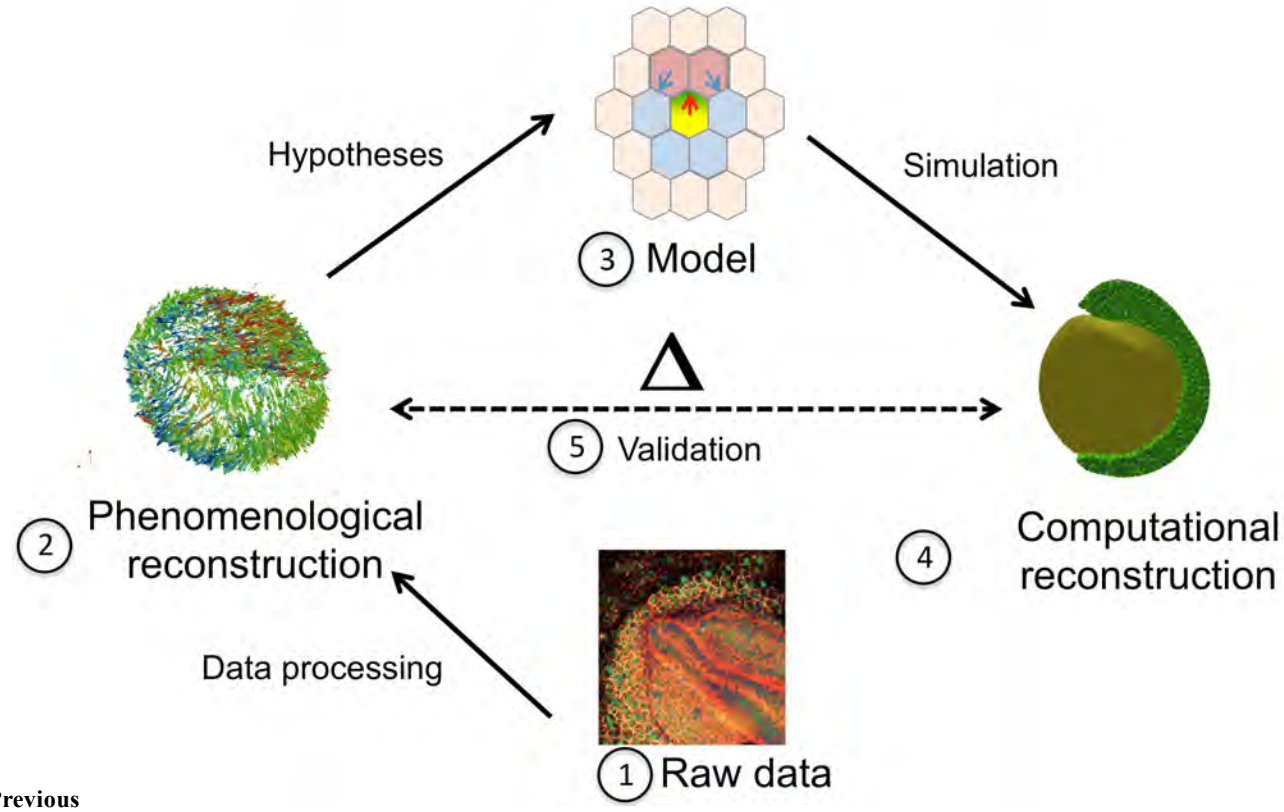


Figure 1.11 :

nouvelle proposition (sketch)

Previous
Next
TOC
References

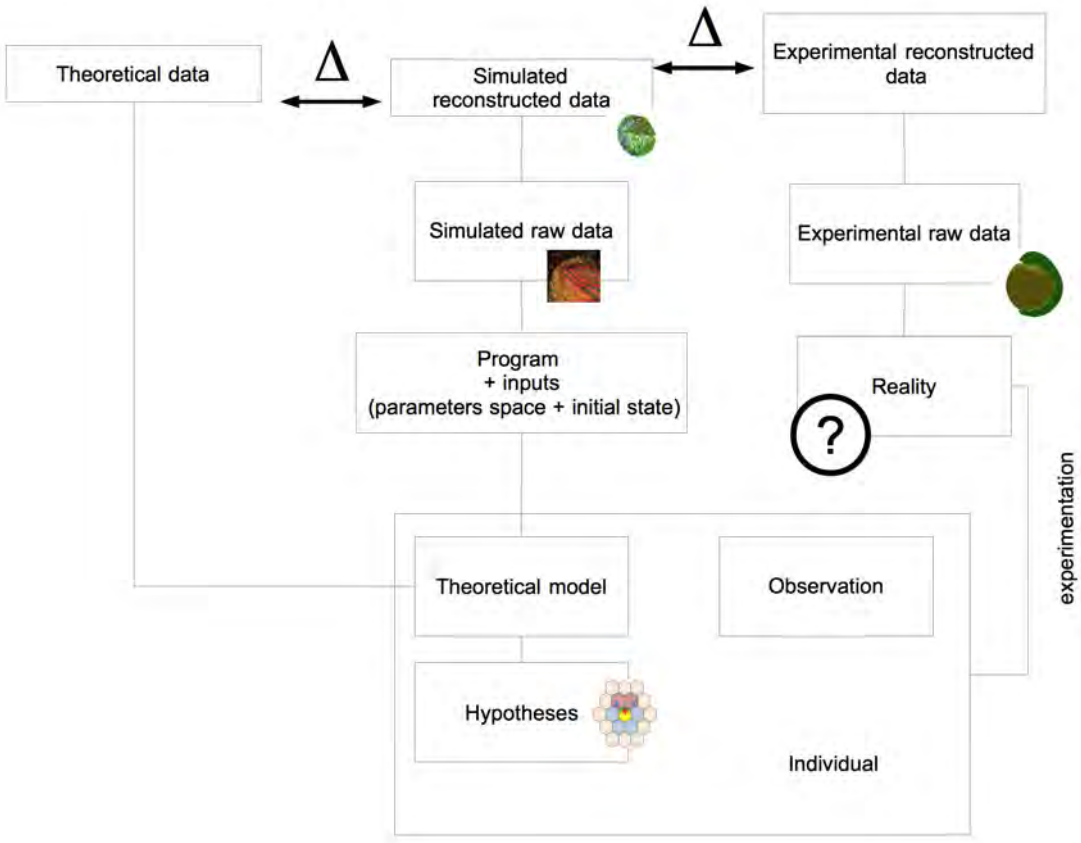


Figure 1.12 :

Previous
Next
TOC
References