A Theory of Intrinsic Bias in Biology and its Application in Machine Learning and Bioinformatics

Juan G. Diaz Ochoa
PerMediQ GmbH
Salzbergweg 18, 85368 Wang
March 31, 2019

Abstract

It is common to consider that a data-intensive strategy is a bias-free way to develop systemic approaches in biology and physiology. At the same time, it is well known that choices concerning collection, sampling, standardization, visualization, and interpretation of data affect the result of its analysis. And seldom a less systemic and more cognitive approach is accepted, according to which organisms' sense and try to predict their trajectories in their environment, which is an intrinsic bias in the sampled data generated by the organism's, limiting the accuracy or even the possibility to define robust systemic models. For this reason, we introduce a theory and methodology to assess this inherent incompleteness in the sampled data, based on the definition of homology groups. We test the use of this methodology in two examples, one from systems biology considering a population of cells with chemotaxis, and other from medicine to analyze heart response to exercise.

Key Words: Internal Bias; Systems Biology; Machine Learning; Evolution; Medicine; Biology; Persistent Homology; distributed cognition; intrinsic bias; modeling bias

Introduction

The notion of systems that engage in the relationship between parts and wholes[1] has guided the development of modern biology and physiology. Thanks to significant technological and mathematical advances it is possible to measure and model complex systems by the interlinking of different components in different scales over time (Aderem, 2005). To this end large amounts of diverse data are collected to identify and extract relevant patterns to develop dynamic models. Neapoletani et al. have defined this as the "microarray" paradigm (Napoletani et al., 2011). This name has origin in the microarray technique in biology to visualize relevant intracellular biochemical reactions by labeling proteins with phosphorescent markers. *A striking and important example of such a trend can be observed in biosciences, where the effectiveness of drugs or the detection of diseases are approached, in practice, by studying clusters, similarities and other structured characteristics of huge arrays of chemical compounds (microarrays) derived by gene, protein or even tissue samples* (Napoletani et al., 2011). Traditional study of biological systems requires reductive methods in which quantities of data are gathered by category. For this purpose computers are required to analyze and model this data (Tarca et al., 2007), in part by using tools from machine learning like deep learning (Webb, 2018), to create

[1] https://plato.stanford.edu/entries/systems-synthetic-biology/index.html#BiolDigiAge

accurate real-time models of a system's response to environmental and internal stimuli, for instance for cancer treatment[2] (Li and Chan, 2017)

However, a significant challenge for the advance of this systemic approach is that data required for modelling is incomplete, contains bias and disbalances, and is continually evolving. Usually the blame falls on the responsible modeler and her methods and technologies used to design and construct the model, which has no more choice but to increase the amount and diversity of the data, perform several model optimizations and model training, and even make use of sloppy parametrization in her models (this is also valid for other models, like weather and economy[3] (Freedman, 2011)).

From this perspective bias in the data is a kind of externality, while it is assumed that organisms behave depending on certain fundamental mechanisms. Seldom it is considered that this bias has also origin in the system. But even when problems in the data originate in the measurement methodology, we must not forget that biological systems and organisms are also observers owning *organized complex interior states that allow the sensing of the environment at different scales as well as sensing and responding to their interior states, while engage these organisms to move within their environment* (Weber, 2018). This implies that constant adaption and evolution continuously challenge historical data used to train models.

Living entities do not react blind to their environment, but also subjectively decide what that response should be. For instance, Denis Noble argue that *life is a symphony that connect organisms and environment across several scales, and is much more than a blind mechanism* (Noble, 2008). This constant sensing across several scales implies that mechanisms and pathways[4] are also adapting and changing as a constant flow (Ross, 2018), and that they are implicitly incomplete, challenging the explicit or implicit accounting of causal explanations based on mechanisms and pathways.

Thus, cognition is an aspect that not only concerns the subjects analyzing the system, but also to the system and organisms subject to this analysis. Such kind of natural distributed cognition represents a challenge to the way how we understand the world and process information, as well as concerns how we can model it using either pure mechanistic models or entropy reduction methods typically used in machine learning and artificial intelligence. By doing this we recognize that "biological systems" are a set of interacting observers that continuously make decisions. Thereby, we are handling with systems with "ears and eyes" that are also capable to model their environment and their corresponding actions, owning elastic-multiscale mechanisms challenging the current mechanistic paradigms (Diaz Ochoa, 2017), particularly due that organisms, like cells, are continuously interacting and deciding about their environment (Laub, 2016), as well as develop intrinsic biases, for instance in grow factor responsiveness(Shah and Anderson, 1997)[5].

Consequently, instead to accumulate ever more data to try to have a better approach of a biological system it is preferable to better acknowledge how an organism senses and "computes"[6] its environment (Hernández-Orozco et al., 2016). To this end we make use of concepts from cognitive computation. As has been referred in an overview about this topic, *cognitively-inspired mechanisms should be investigated in order to make the algorithms more intelligent, powerful, and effective in extracting insightful knowledge, from huge amounts of heterogeneous Big data* (Huang et al., 2018), with topics that span cognitive inspired neural networks to semantic knowledge understanding. Our focus in this

[3] https://www.scientificamerican.com/article/finance-why-economic-models-are-always-wrong/

[4] In what follows we distinguish mechanisms, as biological phenomenon that can be understood from fundamental and invariant physicochemical principles, from pathways, which are a sequence of causal steps that string together an upstream cause to a set of causal intermediates to some downstream outcome.

[5] Intrinsic biases in biology are similar but not equal to the concept of internal bias of agents in psychology and economy, which is a concept applicable only to the human cognition.

[6] "Computing" means in this context sensing more than defining algorithms to perform a given operation

work orients in part to such approaches, while recognize the possibility of a natural distributed cognition in biological systems.

We want to point out that acknowledging this kind of distributed cognition in biological systems is not only relevant from a philosophical, but also from a practical point of view, since it points to minimize the amount of data required to train models. For instance Ghanem et al. employed two metaheuristics, the artificial bee colony algorithm and the dragonfly algorithm to define a metaheuristic optimizer to train a multilayer percepton to reach a set of weights and biases that can yield high performance compared to traditional learning algorithms (Ghanem and Jantan, 2018). Therefore, a clever combination of natural bias with a recognition of heterogeneities in the data used to train models in biology contribute to the development of efficient modelling methodologies using few datasets (Taniguchi et al., 2018).

Thereby, we also recognize that organisms are continuously scanning and adapting to their environment and accordingly are estimating their trajectories; consequently, these organisms are prone to develop intrinsic biases (Diaz Ochoa, 2017). We introduce the theoretical background according to which these biases are the result of a constant rewiring of intrinsic structures responsible for the sensing and "computation" of trajectories according to environmental changes. Based on this concept we make use of persistent topology to characterize these trajectories and detect the *intrinsic biases* of organisms, which are critical to define *modeling bias* aiming to assess the potential generalization of a model from few data sets. We test this concept with a simple model from systems biology and with physiological data obtained from patients wearing IoT devices to measure electro cardiograms (ECGs) while they perform physical exercise.

Internal Bias, and topological methods for its assessment

Definition of internal bias in organisms: computability of trajectories and evolution

In recent years it has become very useful to deal with biological systems from the point of view of theoretical informatics. For example, a bird flying in a forest is "computing" its trajectory Γ according to the tree distribution, which is its environment Env. Honeybees are also computing and creating a model of their environment (Roper et al., 2017). But also, a cancer cell adapting its response also represented by Γ according to changes in its microenvironment, i.e. depending on the acidity of the tissue or even to toxic chemical substances for the cell (Rossetti et al., 2012). But an environment is non-static: a storm can change the distribution of trees in a forest, or the change of a diet can induce substances able to modify microenvironments in tissues affected by cancer cells. In the appendix 1 we introduce with more detail this theoretical background (see also (Diaz Ochoa, 2018)). Using these concepts, we link the variability of the error of a model with the ability of entities to compute their environment and their intrinsic bias.

According to the bias-variance decomposition derivation, the error of a model $\hat{f}$, $Error(\hat{f})$, is composed of three terms: a bias that depends on the definitions of the modeler, a variance term, and an unavoidable *irreducible error* term and is given by[7] $Error(\hat{f}) = E\left[\left(\bar{\Gamma} - \hat{f}\right)^2\right] = Bias(\hat{f}^2) + var(\hat{f}) + \sigma^2$, where $Bias(\hat{f}^2)$ is the modeler bias.

In this concept the bias is assigned to the modeler or external observer. But, what happen when an organism models itself its environment? In this case we are probably confronted to the fact that the bias is not only assigned to the modeler/observer, but also to the organism modeling its environment. The way in which the organism $\boldsymbol{k}$ autonomously explores its environment, and differs respect other

[7] http://scott.fortmann-roe.com/docs/BiasVariance.html

organisms $\boldsymbol{l}$, has an effect in the variability of the estimated error of a model, such that[8] (see the appendix 2 for the derivation of this expression)

$$\Delta^{kl} Error(\hat{f}) = Bias^{kl}[\bar{\Gamma}^2] - 2\hat{f} Bias^{kl}[\bar{\Gamma}], \tag{1}$$

where $Bias^{kl}[\bar{\Gamma}]$ can be defined as (see appendix 1 and 2):

$$Bias^{kl}[\bar{\Gamma}] = E[\bar{\Gamma}^k - \bar{\Gamma}^l] = Bias^{kl}\big(p_i\,(\widehat{M}(\delta_t))\big); \tag{2}$$

here E is the expected value of the difference of the trajectories $\bar{\Gamma}^k - \bar{\Gamma}^l$. This basically is an inherent bias, i.e. it represents how organisms (or system to be modelled) "observe" their environment, which is stored in an internal structure $\widehat{M}$ -for instance a signal transduction network, a neuronal network, etc.- used to compute a trajectory, such that $p_i(M^{\lambda}(\delta_t) = \Gamma^k$ is the computation delivering a trajectory Γ^k (see appendix 1). Simultaneously, this bias is a perturbation of the error respect the trajectories computed by other organisms with a response to the environment.

From equation 2 we concentrate us on the analysis of $Bias^{kl}[\bar{\Gamma}]$, since we want to analyze the relative changes of trajectories between organisms k and l leading to this variability of the error.

We observe two different cases:

- Considering, the trajectory Γ^k can differ from another sampled trajectory Γ^l, the equation 2 represents an internal bias for each sampled trajectory respect other trajectories and is a representation about how the organism k sees the environment, and how this perspective differs from other organisms with their own representation, such that,

  $$Bias^{kl}\big(p_i\,(\widehat{M}(\delta_t))\big) > 0, \tag{3}$$

  i.e. the organism generates incoherent trajectories respect other organisms.

- On the other hand, when all the trajectories of all the organisms are coherent (for instance the trajectory of a birth in a tree labyrinth is similar to the trajectory of other birds), then

  $$Bias^{kl}\big(p_i\,(\widehat{M}(\delta_t))\big) \to 0; \tag{4}$$

  this implies either that the environment is static, or that under high variable environments the organism is rapidly adapting. In the first case the system can be modelled, i.e. there is a function$\hat{f}$ representing a trajectory coupled to an environment (which minimizes a given landscape U_λ). This allows the modelling of the trajectory $\bar{\Gamma} = p_i\,(\widehat{M}(\delta_t))$ using a function $\hat{f}$, assuming that $\bar{\Gamma}$ is an observable that somehow represents a reaction to its environment. The second case can also be modelled only assuming that $\hat{f}(\delta_t)$ can be accessed in each time period δ_t.

Thus, this internal bias is a natural element in the perturbation of $Error(\hat{f})$. This implies that the definition of a valid and stable model does not only depends on the ability of the modeler selecting the best possible model, but also depends on an internal bias of the system to be modelled (i.e. and organism), which implies how this system is modeling its environment, and how plastic is the structure $\widehat{M}$ used by the organisms to compute their environment.

Topological persistence: separation of internal bias from statistical error and modelling

To estimate whether $Bias^{kl}\big(p_i\,(\widehat{M}(\delta_t))\big) \to 0$ (see equation 4) we need to characterize the trajectories, for instance by recognizing harmonic structures in the data. The strategy we propose is to assess the topological structure of the data before a model or regression is performed, ideally combining different trajectories in a phase space.

The strategy selected in this work consists in the construction of point clouds $w_1^k = \big(\Gamma_1^k(t), \Gamma_2^k(t), \Gamma_3^k(t), \dots\dots\big)$ generated from the normalized trajectory sample $\{\bar{\Gamma}^k\}$ of the organism or system k. A point cloud includes a large but finite set of points sampled from a primary form. In the theory t is well known that the combination of the time series of the trajectories $\bar{\Gamma}$, including the time delay of time series, can recover the dynamics of the system (Takens, 1981). Furthermore the presence of harmonic structures in the data, related to this dynamics, can be explored analyzing homological persistence (Emrani et al., 2014). Persistent homology, a tool in algebraic topology, is particularly useful in these situations where the "scale" is not known a priori. It may be viewed as a generalization of hierarchical clustering for higher order topological features, which provides a summary using a bar plot, which essentially validates the different groups discovered using topological invariance (Emrani et al., 2014)[9] [10].

Since Γ^λ owns a topology that reflects the periodic behavior of a signal with Euler characteristics, this means Γ^λ owns a function g with a compact subset of $\mathbb{R}^D$ and $d_K: \mathbb{R}^D \to \mathbb{R}$ the distance function of K. In the following we employ a similar notation as defined by Fasy et. al. (Fasy et al., 2014) for the persistence bars. Consider the *set* $L_t = \{x: d_u(x) \leq t\}$ and $k = L_0$. As t varies from 0 to ∞, the set L_t changes. L_t , defined as a persistent bar, is characterized with persistent topology and summarizes how the topological characteristics L_t change as a function of t. Key topological features include zero and first order topology (see figure 1).

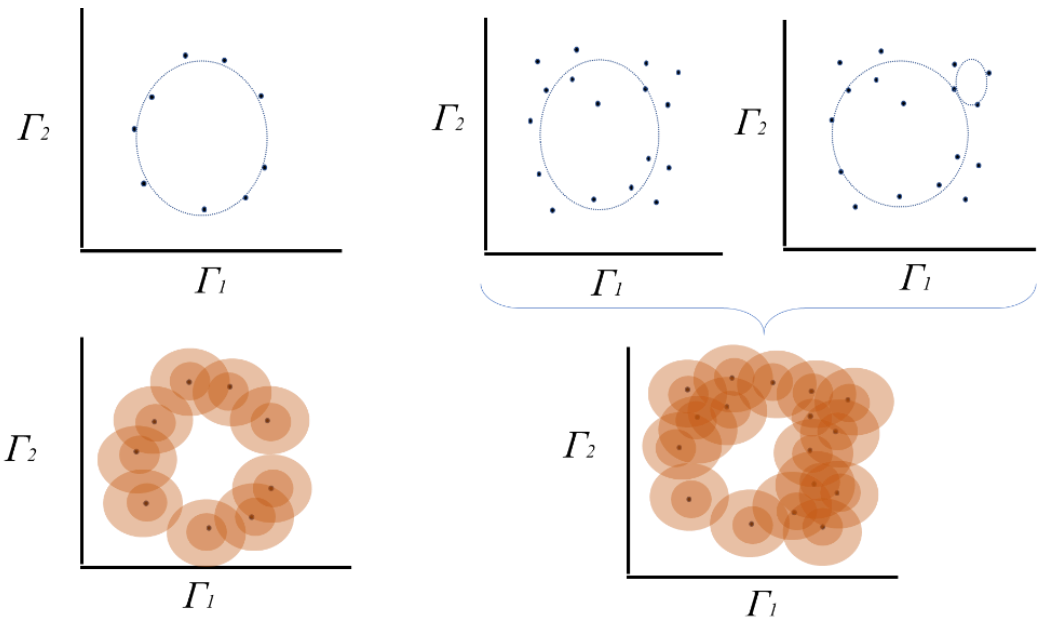


Figure 1 Estimation of topological persistence. In the example left connected points and a surface with homology 0 can be extracted. In the right the same procedure allows the discovery of a second structure with homology 1.

[9] https://cran.r-project.org/web/packages/TDA/index.html
[10] https://cran.r-project.org/web/packages/TDA/vignettes/article.pdf

For $\lambda = 1,2,\cdots \Lambda$, if there is a sample of trajectories generating a point cloud $\{w^{\lambda}\}$ then $\mathrm{L_t}\{\mathrm{w}^{\lambda}\}$ are different topological persistent characteristics of the trajectory Γ^{λ}, such that when equation 3 is combined then,

$$\begin{aligned} &\text{if } Bias^{kl}\left(p_{i(\widehat{M}(\delta_t))}{}^2\right) \geq\ 0, \text{then} \\ &E\left(L_t\{w^{\mathrm{k}}\} - L_t\{w^{\mathrm{l}}\}\right) = \mathcal{M}^{kl} \geq 0, \end{aligned} \tag{5}$$

which is equivalent to low change of Γ^k respect λ implying a preservation of the internal structure $\widehat{M}$ used to compute the trajectory (persistent internal bias). Otherwise,

$$\begin{aligned} &\text{if } Bias^{kl}\left(p_{i(\widehat{M}(\delta_t))}{}^2\right) \rightarrow\ 0, \text{then} \\ &E\left(L_t\{w^{\mathrm{k}}\} - L_t\{w^{\mathrm{l}}\}\right) = \mathcal{M}^{kl} \rightarrow 0, \end{aligned} \tag{6}$$

i.e. organisms show similar patterns or a similar response, i.e. $\widehat{M}$ is equivalent for each organism. The matrix $\mathcal{M}^{kl}$ will be called in what follows distortion matrix.

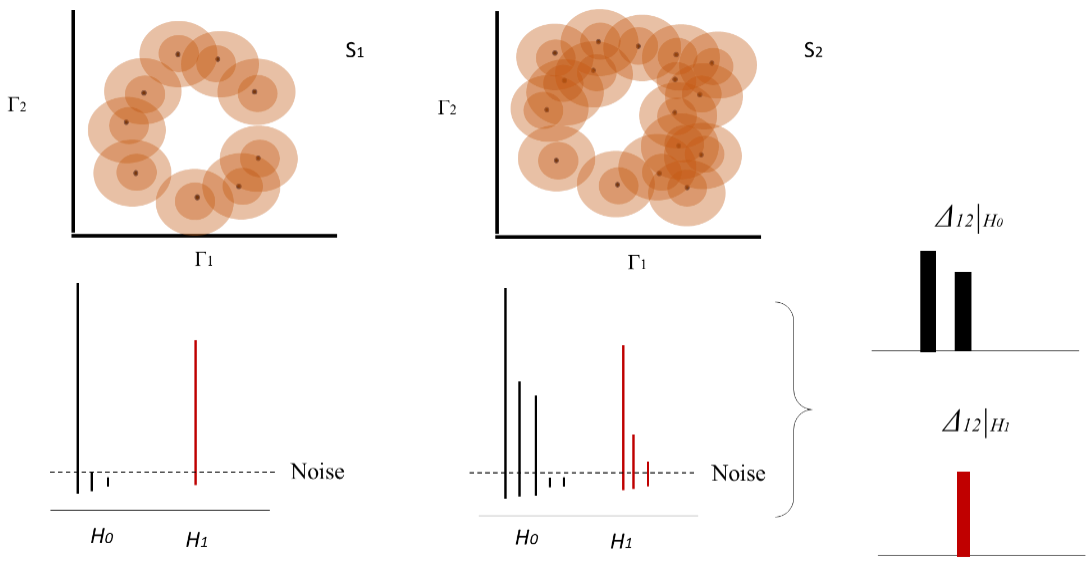


Figure 2 Homology group and bar diagram for two different data sets based on example in figure 1. For S_1 the persistent dyagram delivers a homology group H_1, i.e. a group with genus =1. For S_2 the increase in the number of points allows the defintion of a homology gropup H_0, i.e. a surface with connected points. On the left we schematically represent the difference between both persistent diagrams (for each homology group) to estimate the potential variability $Error(\hat{f})$ due to internal "biological" bias.

The goal is to make a relative assessment of the topological persistence of a cloud of points. Persistent characteristics as well as pure noise are not of central interest. Instead, persistent topological characteristics, computed as the difference of the persistent bars, or equivalently as the difference between homology groups, provide information about similarity between organisms as well as potential internal bias. <u>Accordingly:</u>

> *low relative persistence of data implies a high probability that there is no model $\hat{f}$ fitting the data obtained from a trajectory $\bar{\Gamma}$; this implies high uncertainty with a significative internal "biological" bias that cannot be assessed by a single function $\hat{f}$.*

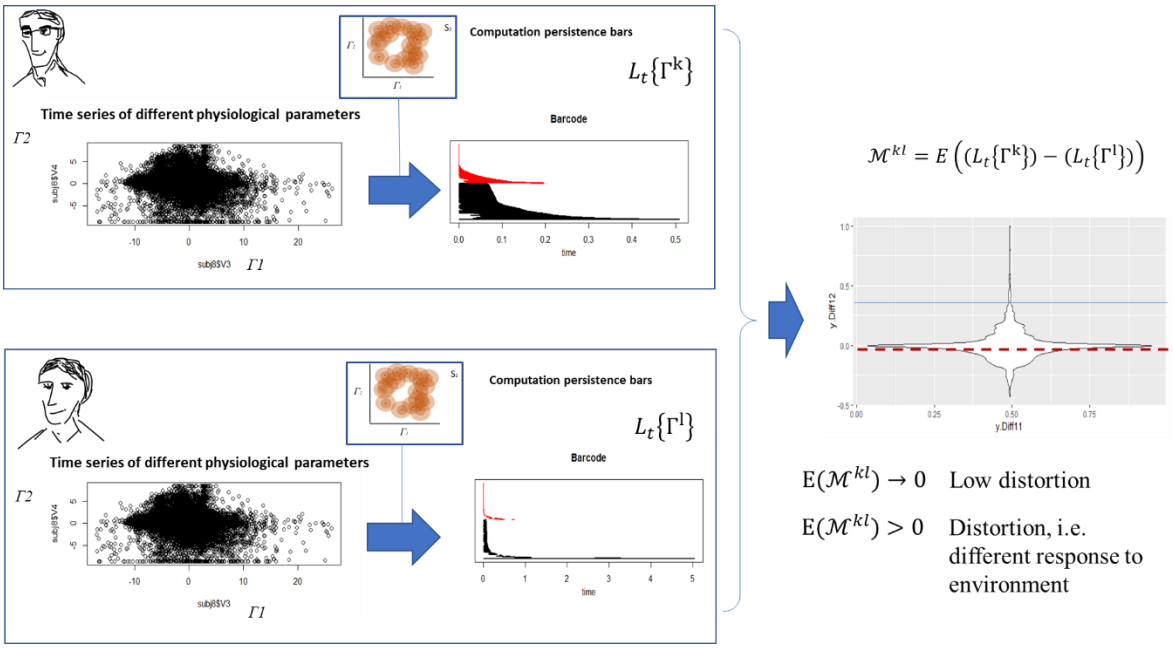


Figure 3 Different steps in the analysis of the matrix $\mathcal{M}^{kl}$ according to equation 8. For example, the data from two individuals is analyzed as a point cloud in a phase space using topological invariance. Once the persistent diagrams are obtained they are normalized respect the highest value of the persistence diagram. The final normalized distribution of the persistence diagram of each individual is subtracted. If the expected value is about 0, then the distortion is minimal, i.e. there is no significant deviation between the patterns of two individuals. Otherwise, a distortion can be identified. Since we deal with the difference between distributions we employ box plots or violin diagrams to estimate the relative difference between two individual patterns.

In figure 3 we describe the main steps of this algorithm: each individual generates her own set of trajectories that produce a point cloud w^{λ}. Using topological homology analysis, the homology is computed for each group, and the persistence is computed using a cluster analysis for each point cloud. This computation essentially extracts the fingerprint from the data generated from each individual. Finally, the relative distance between these fingerprints, i.e. distance between homology groups, $E\left(L_t\{w^{\mathrm{k}}\} - L_t\{w^{\mathrm{l}}\}\right) = \mathcal{M}^{kl}$ is estimated. Due that we compute a difference of the silhouette of the clustering analysis, the final value $\mathcal{M}^{kl}$ is better represented by a histogram. For this reason, a box plot or a violin plot is used to represent the distribution of $\mathcal{M}^{kl}$. The skewness of this distribution provides relevant information about the stability of the perturbation at $\mathcal{M}^{kl}$.

In the following examples we illustrate how to use this methodology to assess the possibility to define a function $\hat{f}$ to model a system.

Examples

Coupled population with chemotaxis

To illustrate this method, we use first the model for a population of organisms with chemotaxis that are predated by a population of predators; these populations consist either on organisms with an adaptable response, non-adaptable response (Figure 3), or an hybrid that can switch between both responses depending on the number of predators, assuming that each predated organism is able to sense the population of predators. Additional details of this model are provided in the original publication (Diaz Ochoa, 2018).

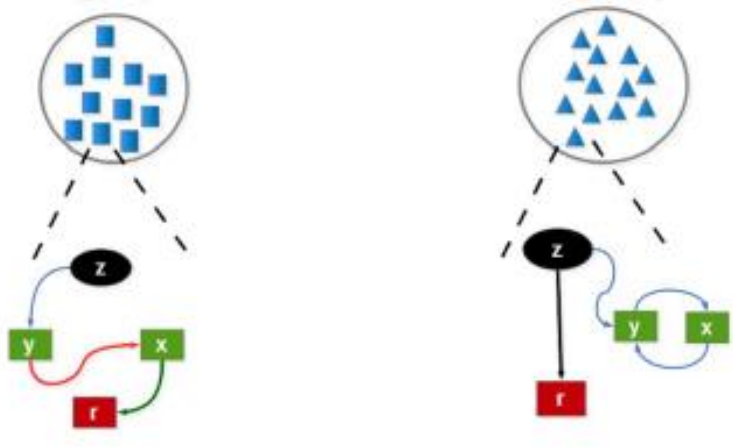


Figure 3 Response mechanism (network: Z input, x and y intermediary mechanisms, r response) of predated organisms with two kind of responses: adaptive (squares) and non-adaptive (triangles)

In this model there are three kinds of organisms: one population with adaptive response, one population with non-adaptive response and one population with an evolving response (between adaptive and non-adaptive) depending on the population of consumers: the consumed organisms in are able to sense the response of the predators and correspondingly change their behavior to influence the size of the population of consumers (please refer to the original publication to obtain the details of the model). Thus, changes and a feed back in the low-level mechanisms are able to induce changes in the whole population.

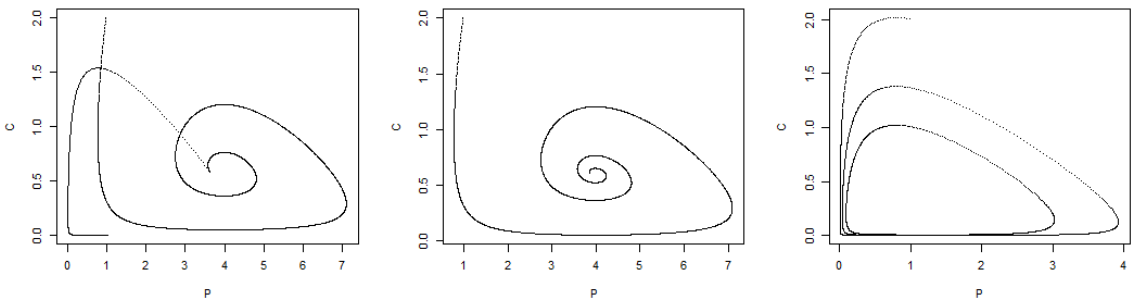


Figure 4 Phase diagram of the predator and predated system with a switching, adaptive, and non adaptive response of the predated individuals.

With the matrix $\mathcal{M}^{kl}$from equation 8 we qualitatively measure the variability of the error respect to internal bias using persistent homology. We show that not only the minimization of the error of a model matters, but also its stability regarding internal variations of the trajectories used to train a model. A modest $\mathcal{M}^{kl}$ ($\mathcal{M}^{kl} \to 0$) is a guarantee that we can derive relative stable models that eventually can be distributed and validated in different situations. On the other hand, an extravagant $\mathcal{M}^{kl} \gg 0$ will probably imply that the model is not able to describe a data set that continuously evolves and probably contains significant internal bias.

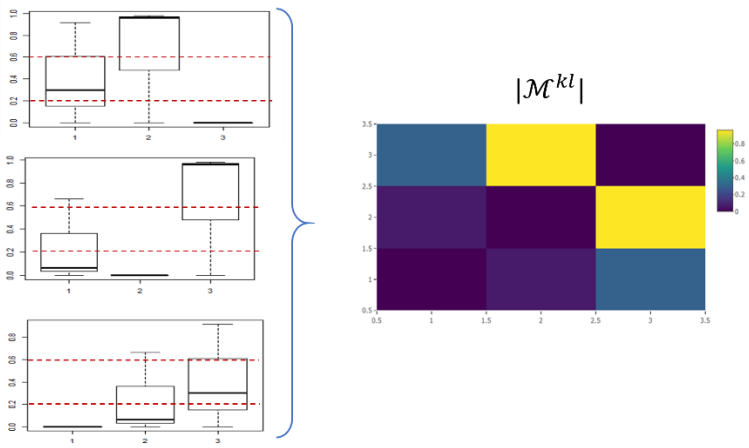


Figure 5 A $\mathcal{M}^{kl}$ computed from the example in figure 3. The red sectors represent strong deformation.

In figure 4 we just observe this case: while the model 1 (a switching response/architecture) approximates to both models (iFFL and NFB architecture), the other two responses (iFFL and NFB) differ from each other, i.e. each one of the adaptive or non-adaptive response represents a considerable internal bias to each other (red sectors in the matrix defined by the box plots).

Assuming that only data of the behavior of adaptive or non-adaptive populations is available, then it could be impossible to derive and validate a single model for the architecture generating a given response, since the internal bias implicit in this data cannot be ignored. On the other hand, the matrix $\mathcal{M}^{kl}$ fulfills its goal informing us that there is still a single response that can be used to approximate both responses, i.e. the data based on the switching response is the best alternative to train both models. In the next example we apply this technique in the analysis of real physiological data.

Analysis of patient data generated from IoT

We selected as a second example the analysis of physiological signals in real patients. As has been suggested in other studies in animals, the locomotor activity is associated to an integrated increase in the activity of different physiological signals (like heart rate, arterial pressure, etc.) (Miyata et al., 2016). Furthermore, the heart response to exercise (macroscopic scale) has origin in complex molecular mechanisms; *for instance in subjects undergoing investigation for angina, individuals with a low chronotropic index (a measure of heart rate response that corrects for exercise capacity) had impaired endothelial function, raised markers of systemic inflammation, and raised concentrations of N-terminal pro-brain natriuretic peptide (NT-proBNP) compared to those with a normal heart rate response* (Routledge and Townend, 2006).

Based on this concept we assume that each individual generates a unique pattern for this integrated response, due to the individual capacity -across several scales- to adapt and/or accommodate to changes in the environment, similar to the case presented in figure 3 (in this case the response to physical exercise). However, in this example we do not have information about these mechanisms, while we assume that they may differ from individuals, i.e. $D^{kl} > 0$. At the end the analysis of this integrated response provides a kind of fingerprint for each individual.

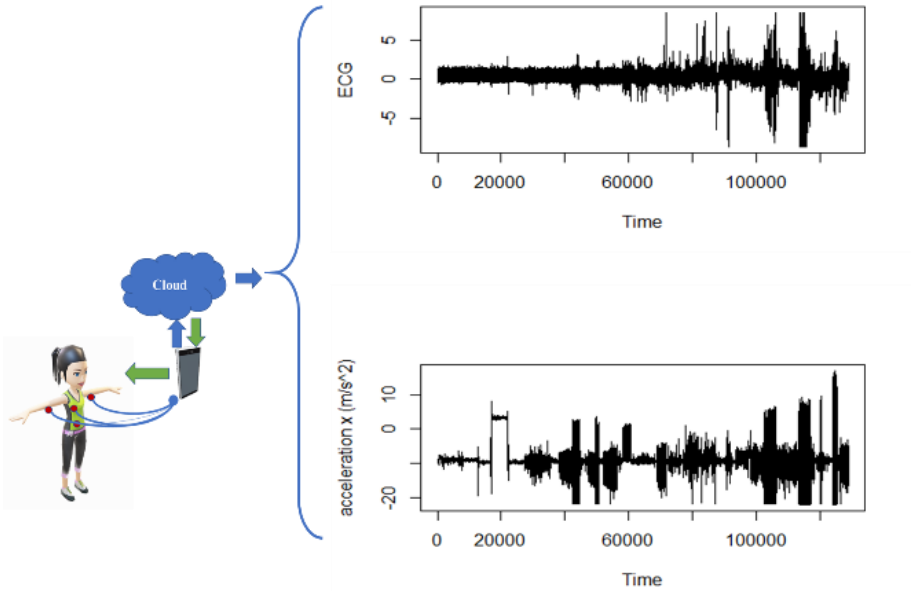
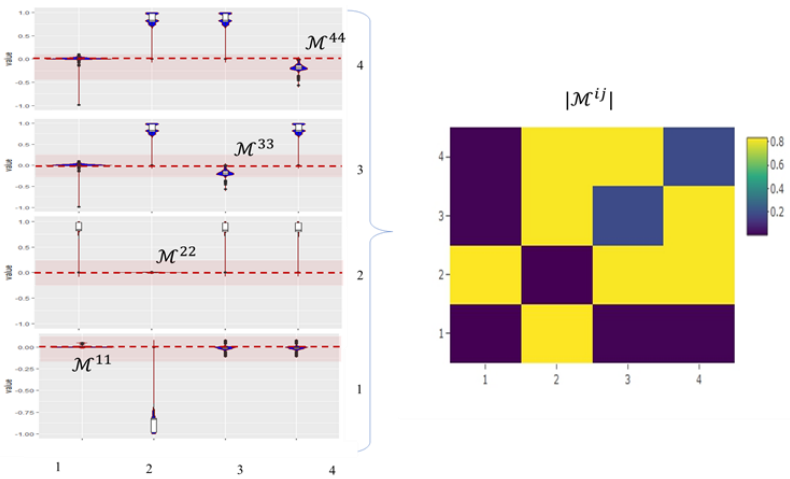


Figure 5 Exemplary data obtained from patients wearing sensors to measure ECG and acceleration in three axes (x, y, z)

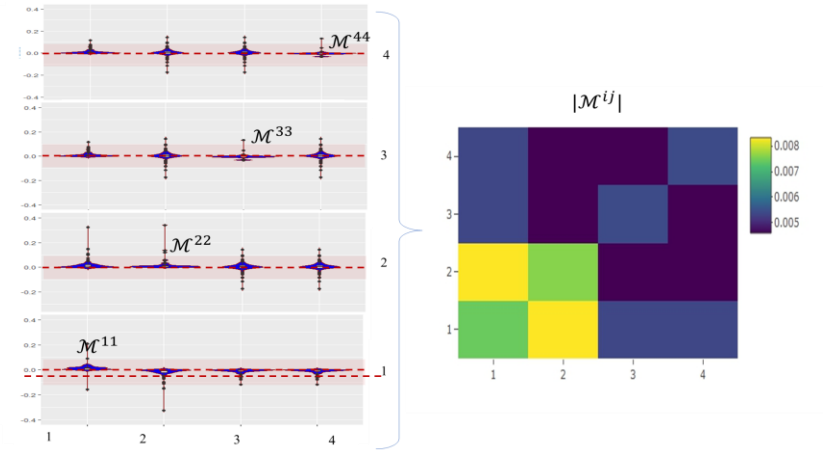
The data used in this analysis has been obtained from the MHEALTH (Mobile HEALTH) dataset, which comprises body motion and vital signs recordings for ten volunteers of diverse profile while performing several physical activities. Sensors placed on the subject's chest, right wrist and left ankle are used to measure the motion experienced by diverse body parts, namely, acceleration, rate of turn and magnetic field orientation. The sensor positioned on the chest also provides 2-lead ECG measurements, which can be potentially used for basic heart monitoring, checking for various arrhythmias or looking at the effects of exercise on the ECG. The activities were collected in an out-of-lab environment with no constraints on the way these must be executed, with the exception that the subject should try their best when executing them[11]. The final raw data is analyzed, i.e. no features are extracted from the time series

A.



[11] http://archive.ics.uci.edu/ml/datasets/mhealth+dataset

B.



C.

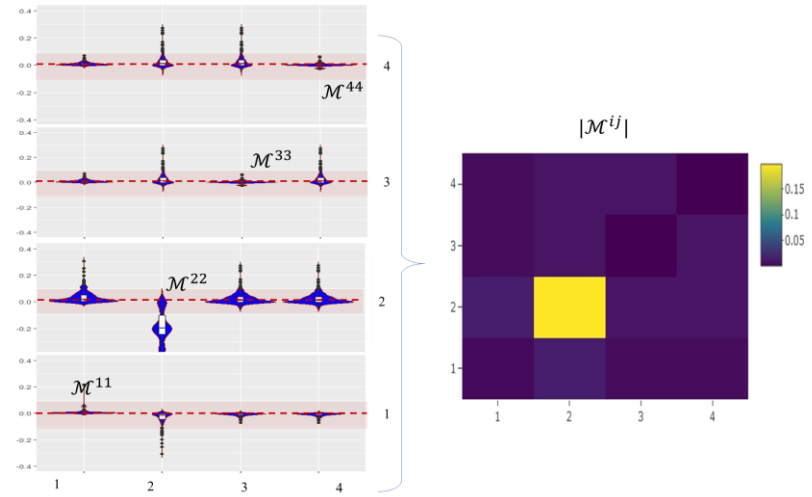


Figure 6 Application of the methodology introduced in fig. 3 for the analysis of physiological data of different patients: once the ECG and acceleration data of a patient doing physical exercise is acquired by IoT devices and stored for instance in the cloud (A.), the relative difference of persistence patterns is analyzed for four different patients considering that the acceleration is measured on the x axis (B.), y axis (C.) and z axis (D).

We basically analyze the acceleration measured with a chest sensor in relation to the ECG, and apply the methodology described in the previous section and in figure 3. We observe some similarities, while a rich structure is revealed when the analysis is extended to each axis where the acceleration is measured: the individual 2 completely deviates from the other three individuals, while the physiological response of individual 1 is similar to both individuals 3 and 4. This result is particularly evident in the axis *x*.

Therefore, this qualitative analysis suggests that the physiology of individual 2 is different to the other individuals (and from itself, according to the result in the *z* axis), probably due to a negative physiological condition (for instance an arrhythmia), or due that the individual 2 moves in an environment different to the other three individuals.

Interestingly, the individual 1 has a pattern similar to individuals 3 and 4, but these last individuals also differ, at least when movement is measured on the x axis. Thus, a more detailed and qualitative model can be trained using only the information from individual 1 to capture and predict similar patterns as individuals 3 and 4.

Discussion

We have introduced a method to assess intrinsic biological bias to organisms modelling their environment, which is an idea that assumes a distributed cognition in biological systems. Our sight

point is that this definition is relevant from a theoretical point of view, to better understand the fundaments of biology, as well as from a practical point of view, for modeling and implement machine learning solutions in life sciences.

Theoretical implications in life sciences:

Life sciences, in particular medicine and physiology, have become prominent in the emerging discipline of e-Science. Up to the present three main data-driven paradigms have guided the development of novel discoveries and technologies in life sciences, namely the empirical, theoretical and computational paradigms. According to the Turing award Jim Grey the e-Science is the fourth data-driven paradigm (Adriaans, 2018). This assumes that with the collection of enough information is possible to access the reality of the world and construct models by observing regularities in the data set. The main goal is therefore to automatically extract relevant patterns from this data using entropy-based information measures to extract meaningful and useful models from large data bases. For this reason in the last years machine learning has become very relevant in biology, see e.g. (Tarca et al., 2007)); such models are for instance neuronal networks (which include all the modern techniques with deep learning, support vector machines, etc.

This paradigm is however complementary to the other three data-driven paradigms, and do not represents a substitute: when patterns in data are discovered there is an implicit assumption that mechanisms produce the observed behavior. Furthermore these causal structures can be represented by a model containing the relevant hypotheses leading to generalized explanations of the phenomenon, and that this model can be further validated with new observations/experiments (see for instance (Groot, 1969)).

In biology data as information is not simply a factor, but is essentially a core element, since it is essential for the organization of life itself and for the propagation of complex organisms. This represents a problem, because all the three paradigms are confronted against the fact that life sciences in particular, and many complex systems in general, owns a particular kind of complexity since entities and systems do not simply react, but also try to predict[12] their future reaction based on the information they own(see e.g. (Roper et al., 2017)). This poses a serious limitation for these paradigms, and in particular for the fourth data driven paradigm, since the automatic model recognition is confronted against bias that are naturally introduced by individual entities and the system itself. Therefore it seems an exception, rather than a norm, to have the possibility to derive "universal" laws or simple mechanisms for complex systems from big-data (see e.g. (Buchanan, 2018)), since we are essentially dealing with entities that also compute and model their environment and disbalance the produced data, introducing a natural limit to any big data approach used for training, due that there is a hopeless accumulation of variations and eventually of biases.

Practical relevance:

By recognizing this limitation, it is possible to avoid the waste of resources: instead of cumulating and managing very large data sets. It seems to be a better strategy to first recognize which data collections are appropriate and stable enough to train models valid for certain groups. This means, an appropriate customization of models including intrinsic bias is more efficient than the training of universal models that inevitably collide against the constant natural evolution and adaption of organisms to their environment (Taniguchi et al., 2018)[13].

To deal with this problem it could be of great advantage to measure the complexity of the data to be analyzed, using for instance a Kolmogorov or a Chaitin complexity measurement (Adriaans, 2018),

[12] To predict is equivalent to compute future events. In computational theory is equivalent to compute future trajectories

[13] Therefore, we think that any research in machine learning do a better job by dealing with the natural symbiosis between information and life sciences, rather than try to simulate or imitate human cognitive capabilities.

together with the design of alternative learning architectures. We think that more comprehensive methods and techniques for machine learning are required. The problem is not only the risk that biases and disbalanced information get into big-data sets used to train machines, but it is also the ever-increasing use of resources: the growing processing and storage of information require a lot of energy and resources that end up in form of waste and green-house gases in the atmosphere[14].

Thus, assuming that organisms continuously compute their trajectories after processing information from the environment, and that they adjust to their environment and eventually evolve (Diaz Ochoa, 2018), we have develop a method to qualitatively detect this complex behavior by measuring the variability of the modelling error. If the data obtained from any organism's trajectory has a persistent structure, then the variability of the modelling error is low, implying that a model can be derived and trained. Otherwise, there is not only a bias that can be assigned to the modeler, but also an intrinsic bias that can be assigned to the organism modelling its environment, challenging the derivation of causal structures, hypothesis formation and model formulation. Simultaneously this intrinsic bias introduces a natural disbalance in the data. Thus, instead to include the data of these individual organisms in the model derivation/training, we suggest ignoring it when universal models are trained, and perform an individual analysis for these individual organisms.

Conclusion

We suggest that by performing this kind of qualitative analysis previous to any data processing helps to recognize and introduce convenient data imbalance that not only contributes to a better understanding of the problem (recognition of entities as modelers, and not simply as a blind reaction mechanisms), but also avoids the expensive training of models. This additionally demonstrates that data without context is not useful, and that this context is not always available.

For the detection of persistent structures in trajectories, we have implemented methods using persistent topology for the analysis of time series (Maletić et al., 2016), which have become a promising way to detect patterns in data different to entropy based methods. Despite the implementation of persistent homology is not straightforward, it shows the possibility to better select the data for training and points to the possibility to introduce clever bias in the models to reduce the amount of training data. In this sense we truly believe that cleverness, and not brute force, will make the real difference in the future of the model derivation in medicine and biology, and in general in all the field of Machine Learning.

Acknowledgments

We acknowledge the critical feedback of Elena Ramirez to this manuscript.

Bibliography

Aderem, A. (2005). Systems Biology: Its Practice and Challenges. Cell *121*, 511–513.

Adriaans, P. (2018). Information. In The Stanford Encyclopedia of Philosophy, E.N. Zalta, ed. (Metaphysics Research Lab, Stanford University), p.

Buchanan, M. (2018). The secret laws of success and status. Nature *563*, 624.

[14] Regarding increasing concentration of greenhouse gases in the atmosphere, among them 6% are generated from computation and It, it results extremely important to develop methods to reduce the ecological impact of performing Machine learning.

Diaz Ochoa, J.G. (2017). Elastic Multi-scale Mechanisms: Computation and Biological Evolution. J. Mol. Evol.

Diaz Ochoa, J.G. (2018). Elastic Multi-scale Mechanisms: Computation and Biological Evolution. J. Mol. Evol. *86*, 47–57.

Emrani, S., Gentimis, T., and Krim, H. (2014). Persistent Homology of Delay Embeddings and its Application to Wheeze Detection.

Fasy, B.T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014). Confidence sets for persistence diagrams. The Annals of Statistics *42*, 2301–2339.

Freedman, D.H. Why Economic Models Are Always Wrong.

Ghanem, W.A.H.M., and Jantan, A. (2018). A Cognitively Inspired Hybridization of Artificial Bee Colony and Dragonfly Algorithms for Training Multi-layer Perceptrons. Cogn Comput *10*, 1096–1134.

Groot, A.D. de (1969). Methodology;: Foundations of inference and research in the behavioral sciences, (Mouton).

Hernández-Orozco, S., Hernández-Quiroz, F., and Zenil, H. (2016). Undecidability and Irreducibility Conditions for Open-Ended Evolution and Emergence. ArXiv E-Prints *1606*, arXiv:1606.01810.

Huang, K., Zhang, R., Jin, X., and Hussain, A. (2018). Special Issue Editorial: Cognitively-Inspired Computing for Knowledge Discovery. Cogn Comput *10*, 1–2.

Laub, M.T. (2016). Keeping Signals Straight: How Cells Process Information and Make Decisions. PLOS Biology *14*, e1002519.

Li, Q., and Chan, M.F. (2017). Predictive time-series modeling using artificial neural networks for Linac beam symmetry: an empirical study. Annals of the New York Academy of Sciences *1387*, 84–94.

Maletić, S., Zhao, Y., and Rajković, M. (2016). Persistent topological features of dynamical systems. Chaos: An Interdisciplinary Journal of Nonlinear Science *26*, 053105.

Miyata, K., Kuwaki, T., and Ootsuka, Y. (2016). The integrated ultradian organization of behavior and physiology in mice and the contribution of orexin to the ultradian patterning. Neuroscience *334*, 119–133.

Napoletani, D., Panza, M., and Struppa, D.C. (2011). Agnostic Science. Towards a Philosophy of Data Analysis. Found Sci *16*, 1–20.

Noble, D. (2008). The Music of Life: Biology Beyond Genes (Oxford: Oxford University Press).

Roper, M., Fernando, C., and Chittka, L. (2017). Insect Bio-inspired Neural Network Provides New Evidence on How Simple Feature Detectors Can Enable Complex Visual Generalization and Stimulus Location Invariance in the Miniature Brain of Honeybees. PLOS Computational Biology *13*, e1005333.

Ross, L.N. (2018). Causal concepts in biology: How pathways differ from mechanisms and why it matters.

Rossetti, S., Esposito, J., Corlazzoli, F., Gregorski, A., and Sacchi, N. (2012). Entrainment of breast (cancer) epithelial cells detects distinct circadian oscillation patterns for clock and hormone receptor genes. Cell Cycle *11*, 350–360.

Routledge, H.C., and Townend, J.N. (2006). Why does the heart rate response to exercise predict adverse cardiac events? Heart *92*, 577–578.

Shah, N.M., and Anderson, D.J. (1997). Integration of multiple instructive cues by neural crest stem cells reveals cell-intrinsic biases in relative growth factor responsiveness. PNAS *94*, 11369–11374.

Takens, F. (1981). Detecting strange attractors in turbulence. In Dynamical Systems and Turbulence, Warwick 1980, D. Rand, and L.-S. Young, eds. (Springer Berlin Heidelberg), pp. 366–381.

Taniguchi, H., Sato, H., and Shirakawa, T. (2018). A machine learning model with human cognitive biases capable of learning from small and biased datasets. Scientific Reports *8*, 7397.

Tarca, A.L., Carey, V.J., Chen, X., Romero, R., and Drăghici, S. (2007). Machine Learning and Its Applications to Biology. PLoS Comput Biol *3*.

Webb, S. (2018). Deep learning for biology. Nature *554*, 555.

Weber, B. (2018). Life. In The Stanford Encyclopedia of Philosophy, E.N. Zalta, ed. (Metaphysics Research Lab, Stanford University), p.

Appendix 1

To assess whether an implicit incompleteness in low-level pathways may compromise the derivation of a model we come back to the idea that organisms behave as computational system: Consider a set of points $\{x_i\}$ which are represented by real values, and which are the inputs encoded in a structure $\widehat{M}$, for instance a network. According to this definition, a trajectory Γ can be estimated as

$$\Gamma = p(\widehat{M}) \tag{A1-1}$$

Where $p(\widehat{M})$ is a program encoding $\widehat{M}$ to compute Γ. Thus, as has been described in a previous work (Diaz Ochoa, 2018), *if the environment is dynamic, i.e.* $Env(\delta_t)$ *for a sequence of times* $\delta_1 \cdots \delta_t \cdots$, *then there exists a computation* p_t, *such that* $p_t(\widehat{M}_{\delta_t}) = \Gamma(\delta_t)$ *is a trajectory representing the environment* $Env(\delta_t)$. *According to the theorem 13 in Hernández-Orozco* (Hernández-Orozco et al., 2016), *if a series of representations of the environment* $p: t \mapsto p_t$ *is computable, the function* $\delta: t \mapsto \delta_t$ *is computable, and the descriptive complexity of the system is bounded. On the other hand, if the sequence of times* $\delta: t \mapsto \delta_t$ *is non-computable, then the system is non-computable, i.e. the environment is non-computable.).*

Computability can be imposed if more than one (dynamical) environments exist and λ are computed. Under this condition, the system explores a family of trajectories $\Gamma^{\lambda}(\delta_t)$ searching for computability conditions in the environment and generating in this way a biological function. This is a plausible condition in multiscale systems (cells, organelles, organs, etc.), such that

$$p(M^{\lambda}(\delta_t)\,) = \Gamma^{\lambda}(\delta_t)\big|_{\{max,min\}U^{\lambda}}, \tag{A1-2}$$

where $\Gamma^{\lambda}(\delta_t)\big|_{\{max,min\}U^{\lambda}}$ is the trajectory of the microstates constrained to the optimization (maximization or minimization) of the control function U^{λ} related to the environment λ, which simultaneously belongs to a family of environments $\lambda = 1 \cdots \Lambda$.

In this equation $p(\widehat{M})$ is not only a program computing a trajectory but is de facto an internal representation of the environment $Env(\delta_t)$ encoded with $\widehat{M}$ and represented by the trajectory $\Gamma^{\lambda}(\delta_t)$, and is an alternative to a conventional definition of dynamical systems, where trajectories are generated by coupled processes, that can be represented for instance by ordinary differential equations. In biology we assume that this internal representation belongs for instance to an organism, or any autonomous element in a biological system or ecology. In social systems this clearly represents an individual. In general, we are describing here any physical system able to store, encode and process information for the estimation of a trajectory representing (the best) response to its environment. This structure is essentially an internal bias (Diaz Ochoa, 2017)[15].

We find here an overlapping between the computational theory and the theory of dynamic systems: the computation of the environment $Env(\delta_t) = Env^{\lambda}$ means the computation of the trajectories Γ representing the environment sampled at λ, i.e. $Env^{\lambda} = \Gamma^{\lambda}$. There are several examples regarding this definition, for example a bird flying in a forest is "computing" its trajectory Γ according to the tree distribution, which is its environment Env. Honeybees are also computing and creating a model of their environment (Roper et al., 2017). But also, a cancer cell adapting its response also represented by Γ according to changes in its microenvironment, i.e. depending on the acidity of the tissue or even to toxic chemical substances for the cell (Rossetti et al., 2012).

But an environment is non-static. A storm can change the distribution of trees in a forest, or the change of a diet can induce substances able to modify microenvironments in tissues affected by cancer cells. If there are changes in the environment in the environment sampled at k and l, i.e. $Env^{k} \mapsto Env^{l}$, and the system/organism searches for a computable solution, then $\Delta^{kl}\Gamma = \Gamma^{k} - \Gamma^{l} > 0$. If changes in the internal structure are associated to $\widehat{M}$, then a distortion factor can be defined as (Diaz Ochoa, 2017)

$$D^{kl} = \frac{\Delta^{kl}\widehat{M}}{\Delta^{kl}\Gamma} \tag{A1-3}$$

iff. the organism is not blind and tries to adapt to its environment[16].

Appendix 2

The following is the detailed derivation of the expression linking the variation of error and internal bias:

$$\begin{aligned}
\Delta^{kl}Error(\hat{f}) &= \Delta^{kl}E\left[\left(\bar{\Gamma} - \hat{f}\right)^2\right] = E\left[\left(\bar{\Gamma}^{k} - \hat{f}\right)^2\right] - E\left[\left(\bar{\Gamma}^{l} - \hat{f}\right)^2\right] \\
&= E\left[\bar{\Gamma}^{k^2} - 2\hat{f}\bar{\Gamma}^{k} + \hat{f}^2\right] - E\left[\bar{\Gamma}^{l^2} - 2\hat{f}\bar{\Gamma}^{l} + \hat{f}^2\right] \\
&= E\left[\bar{\Gamma}^{k^2} - \bar{\Gamma}^{l^2}\right] - 2E\left[\hat{f}\bar{\Gamma}^{k} - \hat{f}\bar{\Gamma}^{l}\right] \\
&= Bias^{kl}[\bar{\Gamma}^2] - 2\hat{f}Bias^{kl}[\bar{\Gamma}]
\end{aligned}$$

[15] From the point of view of a dynamical system, a bias can be understood as a system trapped in a valley in a fitness landscape. This definition is not different, since $p(\widehat{M})$ encodes the optimal representation of a landscape.

[16] Ideally, the estimation of this factor can be helpful to know how strong are changes in the environment to stimulate -or not- the adaptivity of $\widehat{M}$.