1. Title: Deep process-based learning networks in Biodiversity research

2. Summary

We are in a enthralling scientific era. We have the computer power, the open-source tools, the know-how in many highly specialized fields and the team capabilities to integrate Earth science and Biodiversity research in open and decentralized automated research platforms. We are in a period where novel analytical methods and data are being fussioned at an incredible speed to decipher the complexity and feedbacks between the Earth system and the diversity of life. Yet, we are in a massive human-driven biodiversity extinction with large uncertain consequences for Earth climate, life conditions and the stability of Earth (Figure 1). This combination of an enthralling scientific era and rapid global change put us in an edge to team up to go beyond our disciplinary boundaries to contrast scenarios accounting for feedbacks between the Earth system and Biodiversity (Figure 2). For this to happen we need to connect fundamental and applied science (Figure 3b) and one way to do it is throughout distributed open research platforms to provide informaton for management forums in applied conservation and sustainability centers. This proposal aims to develop a distributed open-source automated research platform to integrate multiple databases into Biodiversity dynamics and function scenarios taking into account the interdependencies among biological levels and scales in ecological and evolutionary networks (Box 1 and Figures 3 and 4).

3. Milestones (Internal)

M1 Submission proposal "Deep Knowledge Ledger Network (*DEEPKLEN* & *ROBHOOT*)"
Funding scheme: FETOPEN-EU Challenging Current Thinking
Deadline: March 2020)
Team:
Switzerland (SDSC, Christine Choirat, and EAWAG (Carlos Melian))
Spain, IFISC (Victor Eguiluz)
Estonia, U. Tartu (Raul Vicente)

M2 Submission proposal "Deep process-based learning networks for Biodiversity research"
Funding scheme: BBVA
Deadline: November 2019

4. Deep process-based learning networks in Biodiversity research

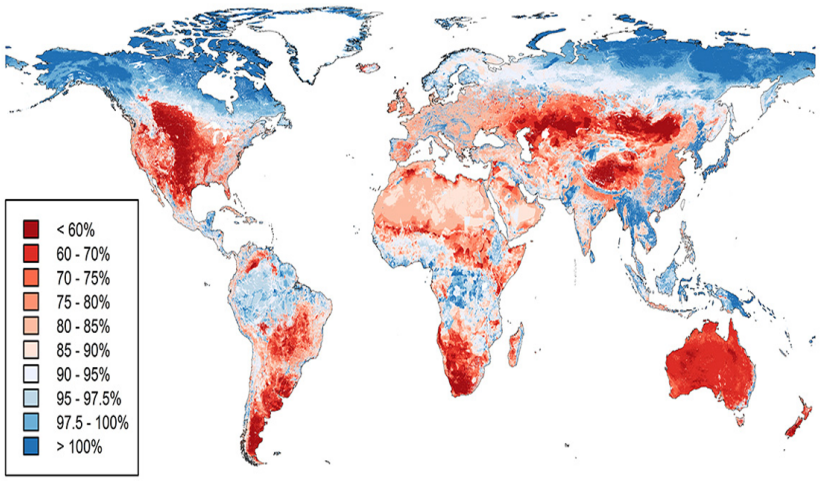

Figure 1: **Biodiversity is declining globally at unprecedented rates**. Map showing the remaining populations of native species across many taxa as a percentage of their original populations. Blue areas are within proposed safe limits, and red areas are beyond these limits. For furhter information please check the original work at http://www.nhm.ac.uk/discover/news/2016/july/biodiversity-breaching-safe-limits-worldwide.html.

Specialization has produced an immense gain in detailed knowledge at each of the levels and scales studied in Biodiversity research. Yet, the information gained and the data obtained in specialized fields are not sufficiently integrated to understand the consequences of biodiversity decline in predicting the outcome of feedbacks between Biodiversity and Earth system. Despite the development of automated research platforms integrating different aspects of the scientific cycle is rapidly advancing[1] distributed open-source automated research platforms in Biodiversity research are still at a very incipient stage.

One of the reasons of still being at a very incipient stage is because most methods in data science and Biodiversity research have been considered classically as distinct fields. However, the current scientific ecosystem is at the stage where merging methods from distinct fields is radically transforming the discipline boundaries, the reproducibility of science and our predicting-understanding power(Reichstein, M. et al., 2019). Many of the recent approaches applying deep learning methods in ecology and evolution have mostly focused at one level of biological organization(Sheehan and Song, 2016). While this might produce additional gain in detailed knowledge at each level, it remains unknown how many layers are going to be needed for predicting the consequences of feedbacks between the Earth system and Biodiversity.

The one-level and one-scale approach might be insufficent to understand the consequences of biodiversity decline in predicting the outcome of feedbacks between Earth system and the diversity of life. To gain predictive and understanding power in biodiversity research we are going

[1]This is by no means an exhaustive list but it gives an indication of the many projects taking place: NakamotoT,BigQuery,Automated statistician,Modulos,Google AI,Iriseaseml

to need to merge distinct databses into hybrid deep process-based learning methods accounting for many layers and the topology of the interactions within and between the layers(Melián et al., 2015). Many methods from data science and biological systems share fundamental properties (i.e., network-like patterns, multiple layers, etc). Yet the full potential of these shared properties have not been sufficiently explored. Biological systems are composed by many layers, and they can contain interdependent hierarchies and feedbacks with interacting learning entities within and between the layers (Figure 2). We will integrate different biological layers into a platform to explore contrasting scenarios of Biodiversity dynamics accounting for interdependencies and feedbacks within and between layers (Box 1 and Figures 3 and 4).

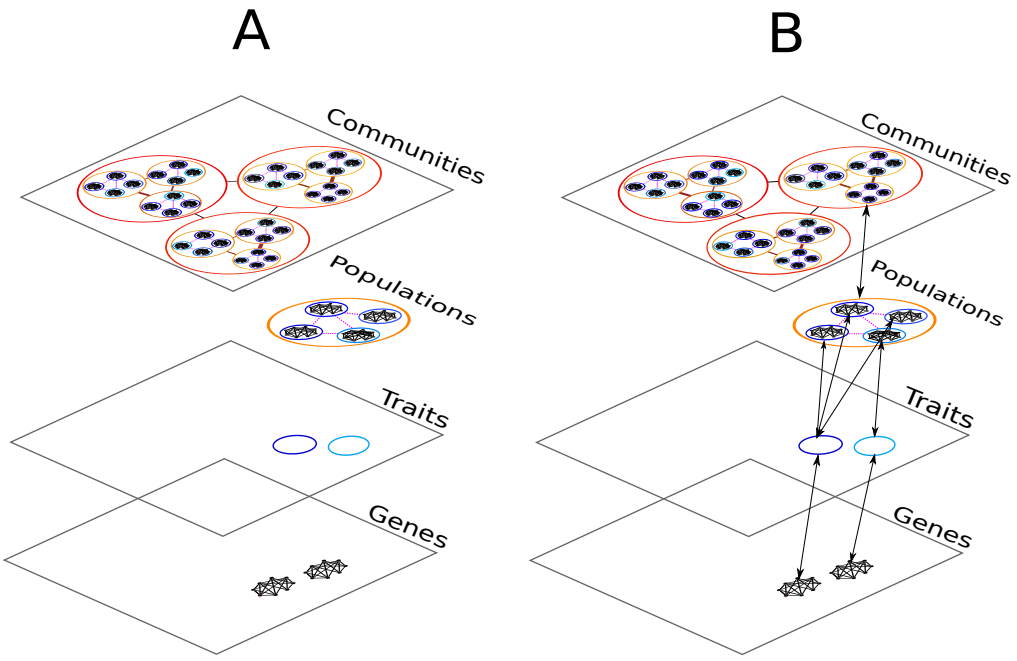

Figure 2: Biodiversity is hierarchically structured yet inferring interdependencies among the levels developing hybrid deep-process based learning approaches to predict the consequences of biodiversity decline remains poorly studied. A) Biodiversity has been studied mostly considering independent levels, from genes, traits and populations to communities and ecological networks. B) Biodiversity represented as interdependent levels accounting for feedbacks from genes and traits, and from traits and populations to communities. It remains unknown which of these two scenarios best predict current trends in Biodiversity decline and its consequences for Earth climate, life conditions and the stability of Earth.

Box 1. Deep process-based learning networks in Biodiversity research

We will infer process-based species distribution maps accounting for biological levels using learning networks. Most datasets in biodiversity are collections of small data. In areas such as species ranges and species interactions, there is a large amount of data, but only a relatively small amount of data match the species ranges or the species interactions with lower biological level data as the gene architecture or the phenotypes. To account for such uncertainty we will use a formalism considering the heterogeneity at individual level(Ghaharmani, 2015) coupling the gene-to-phenotype map to populations, and interactions among phenotypes to communities and species ranges, so that information can be borrowed from other similar levels across the landscape. We will develop our formalism into hierarchical Bayesian neural networks to generate biodiversity distribution maps accounting for biotic, abiotic and migration traits that can be compared against the empirical distribution patterns. We will consider many populations characterized each by individuals containing T normally distributed traits (i.e., biotic, abiotic, and migration traits represented as z_i with i the biotic, abiotic or the migration trait). Populations will be located in a network of discrete/continuous sites guided by long/lat empirical data connected by migration events and the local population demography will be driven by the temporal dependent fitness function accounting for trait architecture following

$$\mathrm{W}(\mathbf{z_i})^t_{jx} = exp[-\gamma(((\mathbf{z_i}^t_{jx} - \theta_{jx})^2)^T \omega^{-1} (\mathbf{z_i}^t_{jx} - \theta_{\mathbf{jx}})^{\mathbf{2}})] \,, \tag{1}$$

where $(\mathbf{z_i})^t_{jx}$ is the vector of trait values of phenotype z at time t for species j and site x, θ_{jx} is the multivariate fitness optimum of species j in site x, ω is the covariance matrix(Lande, 1980; Melo and Marroig, 2014), and γ determines the interaction sensitivity to deviations from the biotic, abiotic and migration optimum. If the covariance matrix, ω, is diagonal, then we are in a no correlated stabilizing selection scenario. Each trait is independently evolving and the connections within and between each biological level are modular and mostly weak. Adding covariation among traits will result in correlated stabilizing selection with strong interactions within and between each biological level. The population dynamics of species j in site x is then given by

$$\frac{dN_{jx}}{dt} = r_{jx}(F(W(\mathbf{z})) + m_{jx}(F(W(\mathbf{z})), \tag{2}$$

where r_{jx} and m_{jx} are the multivariate fitness-dependent intrinsic growth and migration rate, respectively. The first scenario accounting for independently evolving traits will be our proxy for quasi-independent levels considering modularity within- and between-layers (i.e., a highly modular pleiotropy matrix determining the genotype-phenotype map and a highly modular within- and between-species interactions with most interactions weak or zero across the landscape). Such scenario will produce a non- or weakly-interactive species biodiversity map. The second scenario will account for correlated traits and we will consider this scenario as our proxy for feedbacks within and among layers. We will explore a range of topologies from bidirectional recurrent neural networks (BRNN) to feedforward neural networks (FNN) and reinforcement learning (RL) in both static and unknown and dynamic optimum(Schmidhuber, 2015). This scenario will produce an strongly-interactive species biodiversity map. We will disturb both scenarios following random and non-random disturbance regimes (i.e., removing specific interactions, abundances and habitats) and will quantify responses to disturbances using a variety of metrics, from local, regional and global biodiversity metrics(Melián et al., 2015).

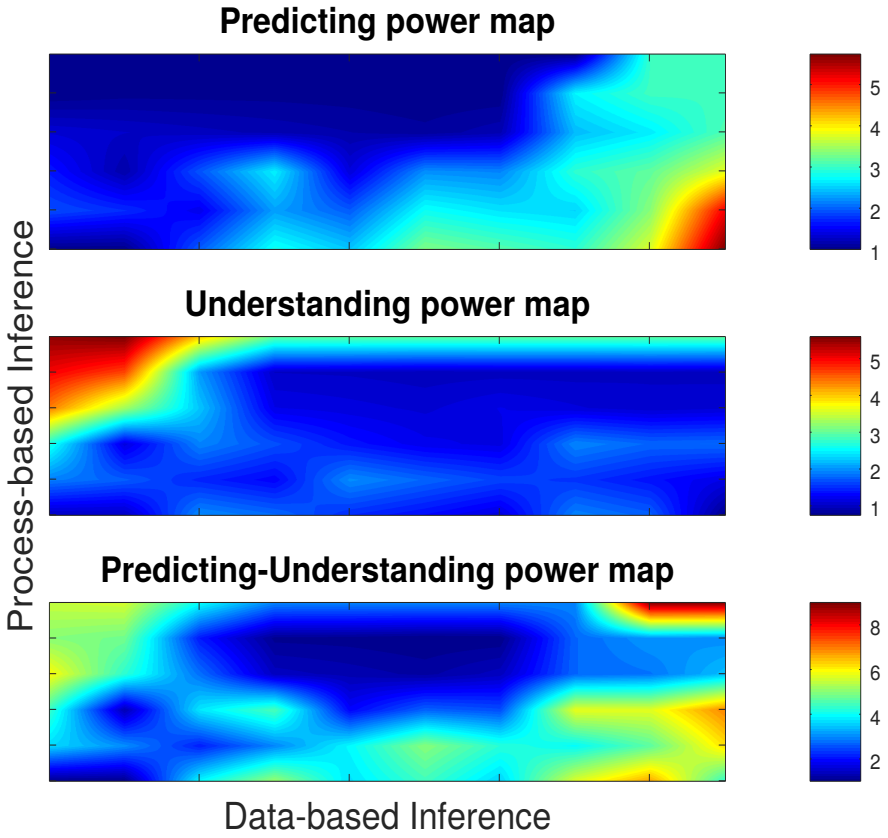

Figure 3: Prediction and understanding power map. This figures shows a cartoon of a prediction power map (top), an understanding power map (middle), and a predicting-understanding power map (bottom). x- and y-axis represent data-based inference (i.e., gradient of AI methods from low (left) to high (right) predictive power) and process-based inference (i.e., gradient of process-based methods from low (bottom left) to high (top left) understanding power). The gradient of predicting power map (top) shows a hot spot red area in the bottom right highlighting the region where AI methods best predict the empirical data. The gradient of understanding power map (middle) shows a hot spot red area in the top left highlighting the region where the best mechanistic understanding occur. The predicting-understanding power map (bottom) shows the sum of the two previous maps highlighting a red hot spot where the best synthesis research joining predicting and understanding power of the empirical data might occur.

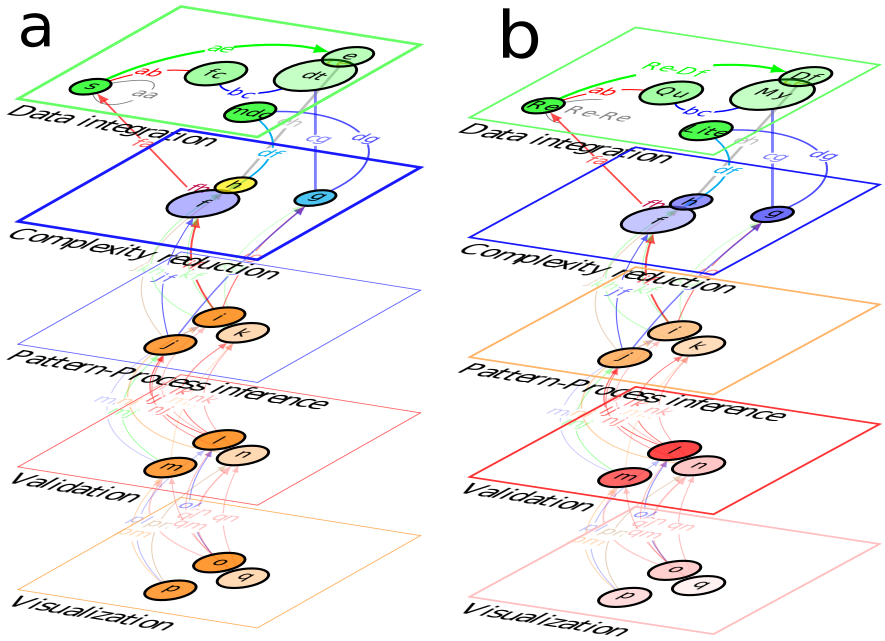

Figure 4: Prototyping a distributed and open automated research platform: a) Our initial prototype will contain five layers (this is not an exhaustive number. Some might be merged and others, like reporting generation, can be introduced): Data Integration, Complexity reduction, Pattern-process inference, Validation, and Visualization. Nodes and links represent algorithms and interactions between two algorithms, respectively. The inter-layer interactions will be implemented using the Renku-SDSC platform[2]. The intra-layer interactions will be developed initially in julia language (other languages will come into play during the development of each layer). **b)** A julia-computing-language prototype of an automated research platform. Nodes and links in each layer represent julia packages and interactions between two packages, respectively. The figure shows the julia packages to be used for the Data integration layer containing the packages "Retriever.jl" (**Re**), "Query.jl" (**Qu**), "MySQL.jl" (**My**), "SQlite.jl" (**lite**), and "DataFrames.jl" (**df**). This cartoon representing many intra- and inter-layer connections might be helpful to show the vision of the platform. For example, the path taken to solve a specific intra- or inter-domain (fundamental or applied) question can be quantified by many metrics each producing a distribution of automated solutions across many nodes in a distributed and open network, the Robhoot Open Network (RON). This distribution can be analyzed to quantify properties as robustness, reproducibility and bias of a fundamental or applied solution.

[2]Renku

5. References

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566:195–2024, 2019. doi: 10.1038/s41586-019-0912-1.

S. Sheehan and Y. S. Song. Deep learning for population genetic inference. *PLoS Comput. Biol.*, 12, 2016.

C. J. Melián, B. Matthews, C. S. Andreazzi, J. P. Rodríguez, L. J. Harmon, and M. A. Fortuna. Deciphering the interdependence between ecological and evolutionary networks. *Trends in Ecology and Evolution*, 33: 504–512, 2015.

Z. Ghaharmani. Probabilistic machine learning and artificial intelligence. *Nature*, 521:452–459, 2015.

R. Lande. The genetic covariance between characters maintained by pleiotropic mutations. *Genetics*, 94: 203–215, 1980.

D. Melo and G. Marroig. Directional selection can drive the evolution of modularity in complex traits. *Proceedings of the National Academy of the Sciences, USA.*, 112:470–475, 2014.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.