Department of Fish Ecology and Evolution,
EAWAG Swiss Federal Institute of Aquatic Science and Technology,
Centre of Ecology, Evolution and Biogeochemistry,
Seestrasse 79, CH-6047 Kastanienbaum,
Switzerland

# 1. Towards deep process-based learning in Biodiversity research

## *1.1.* Summary

We are in a enthralling scientific era. We have the computer power, the open-source tools, the know-how in the many highly specialized fields studying biodiversity, and the team capabilities to break down the disciplinarity barriers to integrate Earth science and Biodiversity research. We are in a period where novel analytical methods and data are being fussioned at an incredible speed for first time 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 take in science the necessary risks to reduce the uncertainty related to the consequences of feedbacks between the Earth system and Biodiversity (Figure 2). For this to happen, we must team up to 1) break down the disciplinarity barriers by merging heterogeneous and independent datasets, and 2) fussioning data-analytics and process-based theory to create synergies between predictive and understanding power (Figure 3).

During my sabbatical I want to pursue two main interdependent goals: 1) fussion modern data analytics and theory in Biodiversity research to delineate future process-based scenarios of biodiversity and function decline. Biodiversity research has been sistematically studied at only one biological level and splitted in many temporal and spatial scales.

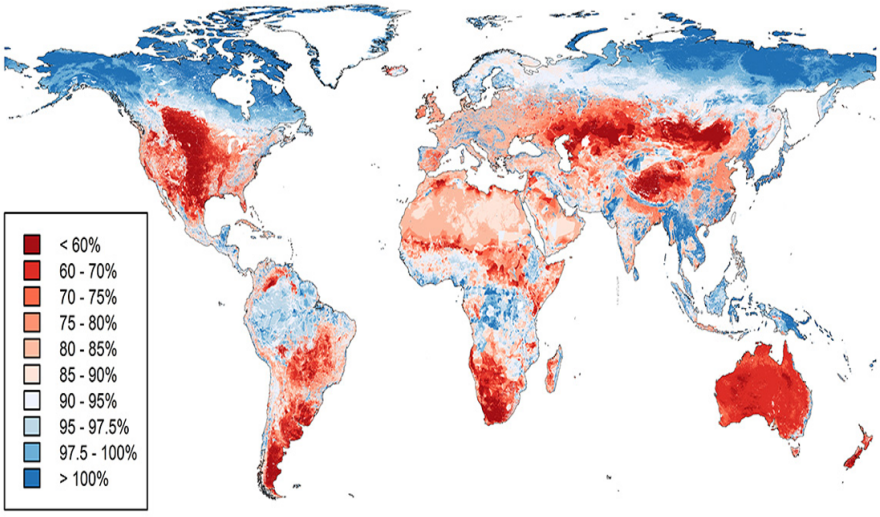


Figure 1: **Biodiversity is declining globally at unprecedented rates**. This map shows the remaining populations of native species across many taxa as a percentage of their original populations. Blue areas are within proposed safe limits, and red areas are beyond these limits (http://www.nhm.ac.uk/discover/news/2016/july/biodiversity-breaching-safe-limits-world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

This has produced an immense gain in detailed knowledge at each of the levels and scales studied, but it might be insufficent to understand the consequences of biodiversity decline in predicting the outcome of feedbacks between Earth system and the diversity of life. We will extend a recent framework to facilitate data- and process-based integration to explore the interdependencies among levels and scales in ecological and evolutionary networks (Figure 2.)[1], and 2) develop an open-source automated research platform to integrate pattern and process-based scenarios of biodiversity and function decline taking into account the interdependencies among levels and scales in ecological and evolutinary networks (Figures 3 and 4). Automated inference is rapidly evolving[2], yet open-source platforms providing insightful patterns and processes integrating databases with analytical frameworks remains challenging. Specifically, open-source research platforms might help to move forward the following five key elements in the scientific process: 1) Finding patterns and processes for the science of science; 2) Identifying bias and uncertainty in inference; 3) Exploring predictions-explanatory gradients to gain sinergy between predictive and explanatory power; 4) Identifying gaps in patterns not explored consequence of lack of integration within and between disciplines, and 5) Facilitating the 4R in open science: reusability, repeatability, replicability, and reproducibility. Below I provide a more detailed description of the tasks to be developed during my sabbatical in merging data science and biodiversity research.

### *1.2.* Deepining process-based learning networks into Biodiversity research

Most methods in AI and ecology and evolution 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[3]. For example, recent approaches in ecology and evolution have introduced deep learning methods for labelled data, from which selection modes and demographic history can be jointly inferred[4]. The nature of biological data is large heterogeneity and a mixture of labelled but also unlabelled data. From one side, there are large databases with labelled DNA sequence or gene network expression data from which deep learning methods can be used to jointly infer selection modes, demographic histories and range dynamics. On the other side, there are many databases with unlabelled ecological data like the patchy distribution of many unidentified species ranges with the corresponding uncertainty associated to quantifying functions like $CO_2$ sources and sinks, for example. This creates many uncertainties and challenges the finding of sufficient enough labelled data for training a machine learning system using deep learning networks from which the feedbacks among species

[1]Melián, C. J.; Matthews, B.; de Andreazzi, C. S.; Rodríguez, J. P.; Harmon, L. J.; Fortuna, M. A. (2018) Deciphering the interdependence between ecological and evolutionary networks, *Trends in Ecology and Evolution*, 33:504-512.

[2]Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*. 521:452-459.

[3]Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*. 566:195-204.

[4]Sheehan, S., Song, Y. S., (2016). Deep learning for population genetic inference. *PLoS Comput. Biol.* 12:e10048452.

distributions, range dynamics, ecosystem functions and species interactions could be predicted for many species across broad spatiotemporal scales.

Many of the recent approaches applying deep learning methods in ecology and evolution have mostly focused at one level of biological organization. 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 and understanding the existing biodiversity patterns. Therefore, the one-level and one-scale approach remains insufficently tested 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 ecology and evolution we are going to need to build hybrid deep process-based learning methods accounting for many layers and the topology of the interactions within and between the layers[5]. Fortunately, many methods from data science and biological systems share fundamental properties, yet the full potential of these shared properties have not been sufficiently explored. Biological systems are composed by many layers (Figure 2), and they can contain interdependent hierarchies and feedbacks with interacting learning entities within and also between the layers. The first step of my sabbatical will consist in merging deep learning networks and multilayer biological networks exploring neural network topologies allowing for feedbacks within and between layers (Box 1).

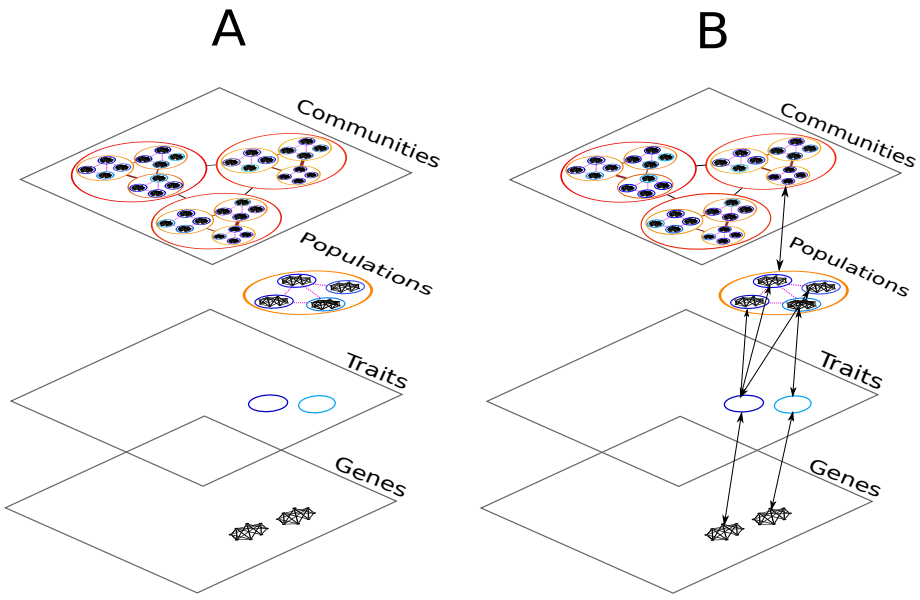


**Figure 2: Biodiversity is hierarchically structured** yet inferring interdependencies among the levels developing hybrid deep-process based learning approaches to predict the consequences of

[5]Melián, C. J.; Matthews, B.; de Andreazzi, C. S.; Rodríguez, J. P.; Harmon, L. J.; Fortuna, M. A. (2018) Deciphering the interdependence between ecological and evolutionary networks, *Trends in Ecology and Evolution*, 33:504-512.

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.**

### Multilayer biological networks

To infer the role of feedbacks between evolutionary and ecological networks on biodiversity decline, we will extend a process-based approach[a] taking into account demography, trait evolution, gene flow and selection to infer the connections between 1) gene interaction networks and landscape trait distribution and 2) landscape trait distributions and biotic and abiotic factors driving multiple-species ranges.
Explain databased to be used ... and connect it to how to use the data as input of the deep learning networks

### Deep learning networks

We will explore a variety of bidirectional recurrent neural networks allowing for the exploration of feedbacks within and between the layers (key figure making a clear connection between multilayer biological networks and deep learning networks)

### Inference

Many large data sets in ecology and evolution are large collections of small data sets (refs). For example, in areas such as species ranges and species interactions, there might be a large amount of data, but there is still a relatively small amount of data for each individual or interaction. To customize predictions for species ranges and accounting for abiotic and biotic factors it becomes necessary to build scenarios accounting for the heterogeneity at individual level – with its inherent uncertainties – and to couple these models together in a hierarchy scaling from genes, to phenotypes, populations, communities and ecosystems, so that information can be borrowed from other similar levels across the landscape. This individualization of models[b], will be implemented using hierarchical Bayesian neural networks approaches such as hierarchical Dirichlet processes therefore accounting for many interdependent layers.

[a] Melián, C. J.; Matthews, B.; de Andreazzi, C. S.; Rodríguez, J. P.; Harmon, L. J.; Fortuna, M. A. (2018) Deciphering the interdependence between ecological and evolutionary networks, *Trends in Ecology and Evolution*, 33:504-512.

[b] Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*. 521:452-459.

In this setting, and in addition to the limited training set contained in many biological and ecological databases, biological and ecological multilayer networks can be trained or explored integrating datasets from many sources. This creates opportunities to infer how the real world ecological systems might be predicted by process-based interactions from complex traits to non-linear ecological models accounting for interdependencies and feedbacks between

levels. There are going to be at least two big group of questions consequence of the fussion between deep learning and multilayer biological networks. Methods driven questions focused in the structural and dynamical properties integrating deep learning networks and multilayer networks. Applied driven questions like inferring future Biodiversity trends projections under different deep process-based learning networks scenarios. Both types of questions would require to explore gradients combining predictive and understanding power to jointly infer the processes and the patterns that can be interacting to produce specific dynamics and topologies. In the applied side, such outputs will produce likelihood scenarios for future biodiversity declines and its consequences for Earth climate and stability (Figure 3). In summary, integrating deep learning and multilayer biological networks accounting for processes within each of the layers, their interaction effects within and between the layers and the effects on biodiversity dynamics and ecosystem functions is full of open challenges and also opportunities to advance our understanding of multidisciplinary data science and biodiversity research.

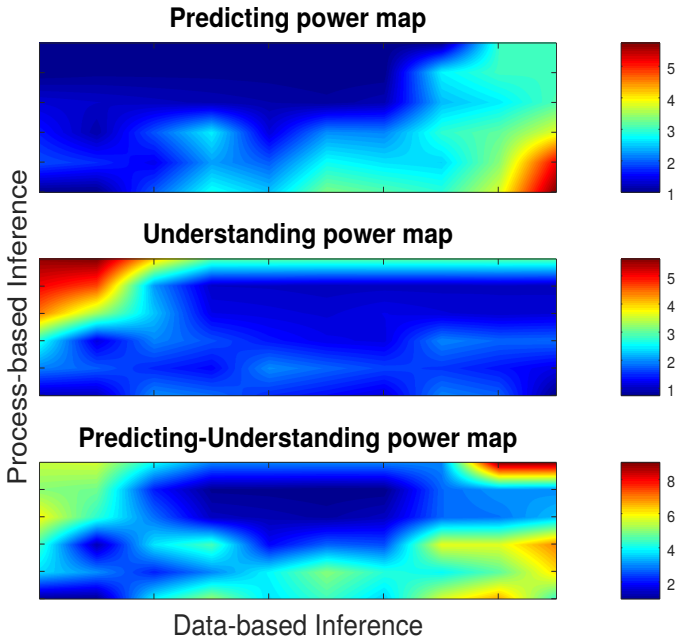


**Figure 3:**
**Interdisciplinarity and synthesis in science** will be needed to join pedicting and understanding power in deep process-based learning networks. This figures shows a cartoon of a predicting power map (top), the understanding power map (middle), and the predicting-understanding power map (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 and interdisciplilnary research joining predicting and understanding power of the empirical data occur.

### *1.3.* **Automated research platforms**

High-resolution and heterogeneous data coming from many sources is standard in science. Yet, automated inference providing insightful patterns and processes integrating databases with analytical frameworks remains challenging. Indeed, automated research platform for automated workflows to integrate data and pattern-process-based inference accounting for many sources of uncertainty are missing in current science ecosystems. In the following lines I will argue that automated research platforms can strongly contribute to the science of science to take better informed decisions in science.

Automation is rapidly occurring in many fronts, from robotics and investments to gaming and ecommerce. What about science? Science is in a era of massive data accumulation, integration and pattern detection. Yet, obtaining insights from such an integration accounting for reproducibility, inference and prediction power is at a very incipient stage (**??**). There are many challenges when aiming to integrate data, inference and prediction. For example, sampling design and experiments (**?**), randomizations to achieve solid statistics , and process- or pattern-based model selection and inference just to name a few require many intermediate decisions that make the scientific process challenging to repeat, replicate, and reproduce. Currently, there are many protocols and platforms automatizing partial steps of the scientific cycle (Table 1). Here, we summarize automated platforms to analize the existing gaps with the aim to automate the whole scientific cycle (Figure 1). Open automated research platforms might play a leading role in addressing at least the five following challenges: 1) Helping in the science of science by providing quantitative statistics (**?**), for example, the many paths with solutions to specific questions; 2) Identifying systematically bias and uncertainty in inference; 3) Exploring prediction and explanatory gradients to gain sinergy between predictive and explanatory power to complex problems; 4) Identifying gaps in patterns not explored consequence of lack of syntesis within and between disciplines, and 5) Allowing for reusability, repeatibility, replicability and reproducibility along the many paths in the scientific enterprise (Figure 1).

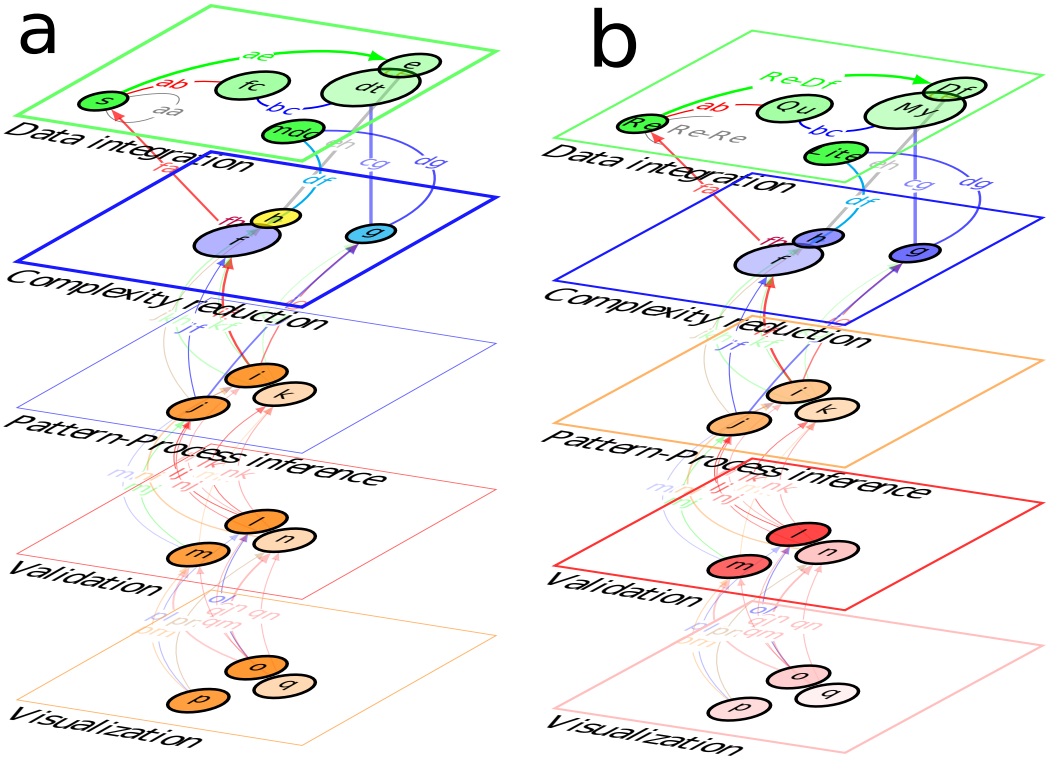


**Figure 4: Five layer automated research platform:** Data Integration, Complexity reduction, Pattern-process inference, Validation, and Visualization. Nodes and links represent algorithms and interactions between two algorithms, respectively. For example, the figure shows five algorithms in the layer Data integration (**a**, **b**, **c**, **d**, and **e**). Algorithm **a** interacts with algorithm **b** and **e** in the same layer (intra-layer connections) and with algorithm **f** from the second layer (inter-layer connection), Complexity reduction. The cartoon represents many intra- and inter-layer connections to solve a problem. The paths can be quantified by many metrics each producing a distribution of automated solutions. This distribution can be analyzed with the ones used for a specific domain in science, the science of science of a domain, to quantify properties as robustness, reproducibility and bias of a domain. **b** A julia prototype of an automated research platform. Nodes and links in each layer represent julia packages and interactions between two packages, respectively. The figure shows julia packages within each layer. For example, the layer Data integration contains the packages "Retriever.jl" (**Re**), "Query.jl" (**Qu**), "MySQL.jl" (**My**), "SQlite.jl" (**lite**), and "DataFrames.jl" (**df**).