1. Title: Deep process-based learning networks in biological, technological and economical systems

2. Summary

We are in a enthralling scientific era. We have the computer power, the open-source tools, the know-how in many specialized fields and the team capabilities to infer complex patterns from highly heterogeneous data. We are in a period where novel analytical methods and data are being fussioned at an incredible speed. Yet, deciphering the strength of process-based feedbacks underlying patterns of big multilayer data across fields is at a very incipient stage (Figure 1). Our proposal aims to fussion data with process-based feedbacks to disentangle the mechanisms underlying complex empirical patterns. First, we will develop an open-source automated research platform accounting for data integration, complexity reduction, inference, validation, visualization and reporting generation (Figure 2). Second, we will test the platform to decipher the strength of the feedbacks underlying Earth Biodiversity patterns accounting for multilayer network data (Figure 3). Our research goals contain two main milestones for a 24 months duration plan (Figure 4): The first milestone will be the deployment of an automated research platform by the end of the first year. The second will test the automated research prototype with Earth biodiversity data as a case study to be developed during the second year to show mechanistic feedbacks inference in complex empirical patterns.

3. Background and objectives

Specialization has produced an immense gain in detailed knowledge at each of the levels and scales studied across many scientific disciplines. However, integrating the information obtained in specialized fields with existing technologies and methods in natural, technological, social and economical systems still present many challenges. Despite rapid advances of automated research platforms facilitating data integration accounting for parts of the research cycle[1] open-source automated research platforms are still at a very incipient stage of development.

One of the reasons automated research platforms are still at a very incipient stage of development is because most methods in data science and other scientific disciplines have been considered classically as distinct fields. This is rapidly changing due to the current scientific ecosystem. We are at an stage where merging methods from distinct fields is radically transforming the discipline boundaries, the reproducibility of science and our predicting-understanding power(Reichstein et al., 2019). Many recent approaches applying deep learning methods in biology, economics, social and technological systems have mostly focused on pattern detection within one level of organization (Sheehan and Song, 2016) (OTHER REFS SYSTEMS). While this might produce additional gain in detailed knowledge at each level for understanding such systems, it remains unknown how many layers are going to be needed for maximizing predictive- and process-based knowledge in biological, economical social and technological systems (Figure 1).

Gaining predictive and understanding power need the merging of distinct databases into hybrid deep process-based learning methods accounting for many layers and the topology of the interactions within and between the layers(Schmidhuber, 2015; Ghaharmani, 2015; Melián et al., 2018; Reichstein et al., 2019) (OTHER REFS HERE). Many methods from data science and biological, economical and technological systems share fundamental properties (i.e., network-like patterns, multiple layers, spatiotemporal dynamics, interdependent hieraqrchies and feedbacks with interacting learning entities within and between the layers, etc), and the full potential of these shared properties have not been sufficiently explored combining big data with deep process-based multilayer networks in automated research platforms (Figure 2). In this regard, science, engineering and technological landscapes require the integration of many layers to facilitate automation, reproducibility, cooperation, new science of science methodologies, and public access to the full research cycle and research findings. Yet, technologies facilitating compactly open-access to the full research cycle accounting for multilayer data is currently not in place. Our research proposal aims to deploy a multilayer automated network accounting fully for the research cycle to provide decentralized

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

and real-time open-access data-rule-knowledge to gain informed decisions to help solve complex ecological, social and technological problems.

Data driven multilayer process-based methods can increase the pool of deep learning models in data science to understand more broadly the connection between predictive power (i.e., pattern detection), and understanding power (i.e., process-based inference). While conceptual frameworks unifying different layers in many research fields is well established (REFS HERE :: 28–31), there is currently a lack of deep process-based learning models accounting for many layers in biological, social, technological and economical systems. Here is where data science can benefit to further developing approaches unifying data driven patterns and process based theory (REFES HERE :: 35,36). The interaction between prediction and understanding power can also advance synthesis in data-science by gaining broader insights from deep pattern- and process-based learning models that can be applied to other many fields.

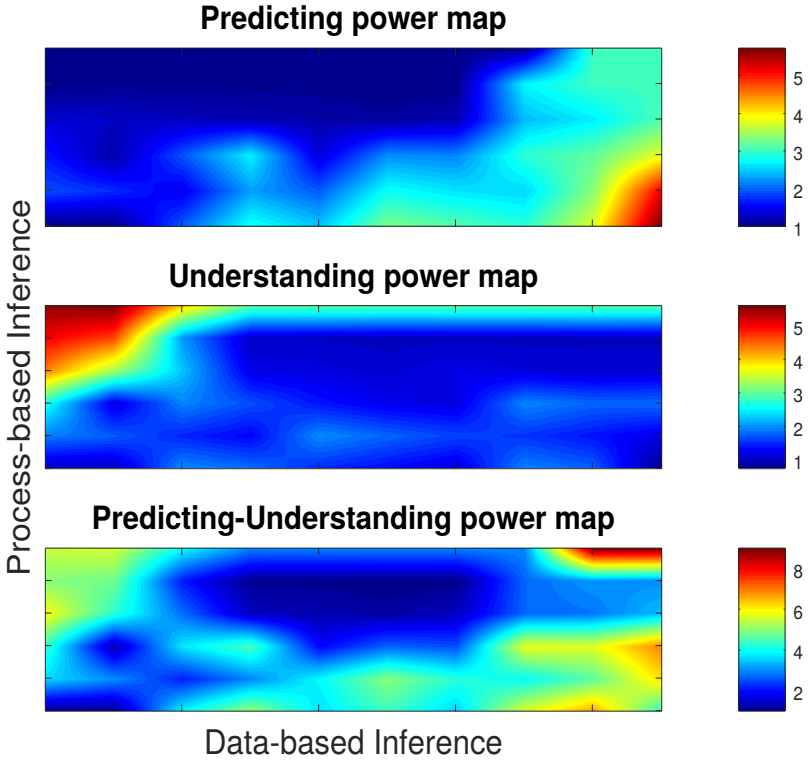


Figure 1: Prediction power (top), understanding (middle), and prediction-understanding power maps (bottom). x-axis represents data-based inference (i.e., gradient of AI methods from low (left) to high (right) predictive power). y-axis represents process-based inference (i.e., gradient of process-based 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 red hot spot 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. The first research goal of this proposal aims to build an automated research platform to maximize the predicting and understanding power highlighted in the red hot spot of the predicting-understanding power map (bottom).

4. Research methodology

INTRODUCE BRIEFLY THE TWO PARTS:

- To develop an automated data-driven platform combined with process-based AI.
- To infer how the interdependencies of the different biological levels of organization impact Earth biodiversity.

4.1. Goal 1: Automated data-driven platform

SUMMARIZE MORE COMPACTLY THIS PARAGRAPH Webb (2018); Heaven (2019) The recent success of Artifical Intelligence can be easily illustrated with applications to many discipline. Examples could include machine learning nowcasts lightening occurrence Mostajabi et al. (2019), ... astounding performance in beating board and video games Mnih et al. (2015); Silver et al. (2016), computer vision that enables self-driving cars Poggio and Bizzi (2004); Krizhevsky et al. (2012), medical diagnosing Ferrucci et al. (2013), ...

economic and social importance Significance of the project for data science The project will help to improve multilayer inference from broad classes of multidimensional data. It will bring a class of deep learning networks, deep process-based learning networks, to facilitate merging biodiversity research and data science. Deep process-based learning networks will serve as a fundamental and applied tool to unfold the interdependence among biological levels for understanding biodiversity dynamics and its current decline with and without feedback. The project will also be communicating in a permanent public exhibition how data science and biodiversity research account for interacting biological levels for predicting future response scenarios to rapid global change.

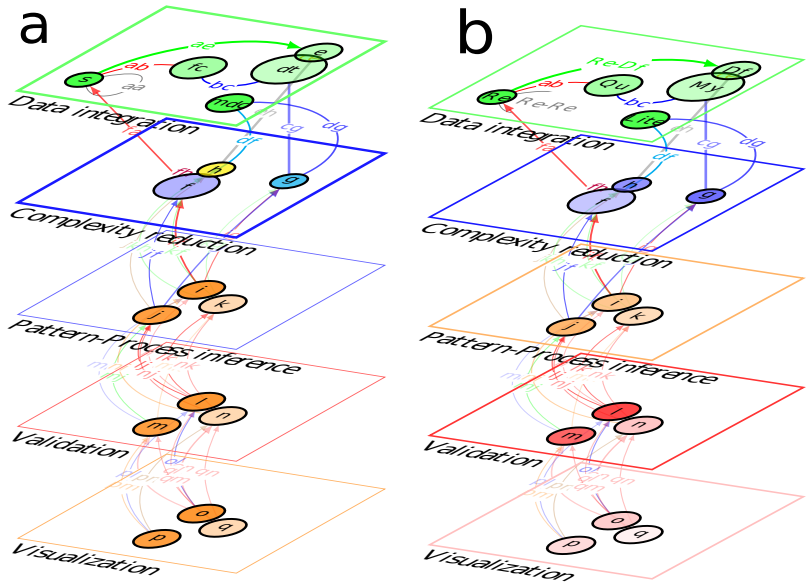


Figure 2: Automated research platform prototype: a) Our prototype contains initially six layers (this is not an exhaustive number): Data Integration, Complexity reduction, Pattern-process inference, Validation, Visualization and Reporting. Nodes and links represent algorithms and interactions between two algorithms, respectively. The inter-layer interactions will be implemented using the open-Renku-Swiss Data Science Center 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.

4.2. Goal 2: Deep-process based learning networks in Earth Biodiversity

We will apply our results to decipher the complexity of the processes and the feedbacks underlying patterns in Earth biodiversity data. We are in a massive human-driven biodiversity extinction with large uncertain consequences for Earth climate, life

[2]Renku

conditions and the stability of Earth (Figure 3). Uncertainty comes mostly from the unknown consequences of feedbacks. contrasting deep process-based scenarios accounting for feedbacks between the Earth system and Biodiversity are at a very incipient stage (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).

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

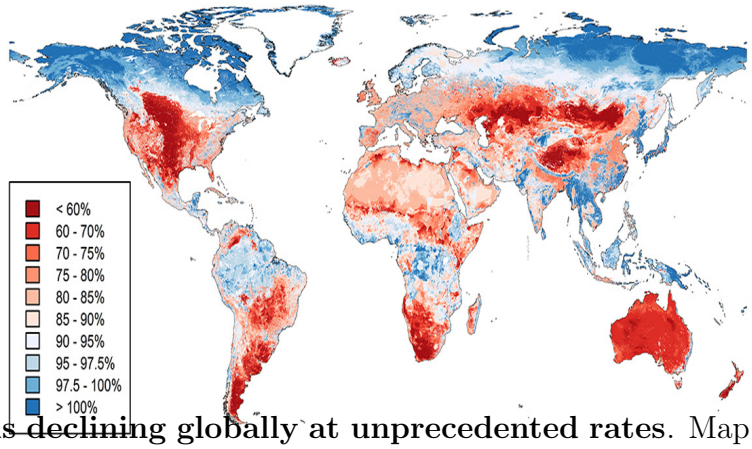
$$\mathrm{W}(\mathbf{z_i})_{jx}^{t} = exp[-\gamma(((\mathbf{z_i}_{jx}^{t} - \theta_{jx})^2)^T \omega^{-1} (\mathbf{z_i}_{jx}^{t} - \boldsymbol{\theta}_{\mathbf{jx}})^{\mathbf{2}})] \,, \tag{1}$$

where $(\mathbf{z_i})_{jx}^{t}$ 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., 2018).

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 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., 2018). 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).



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/breaching-safe-limits-worldwide.html.

5. Experience of the research group

VME is a member of the Group of Interdisciplinary Physics which forms the core of IFISC (Institute for Cross-Disciplinary Physics and Complex Systems), a joint research Institute of the University of the Balearic Islands (UIB) and the Spanish National Research Council (CSIC) created in 2007. IFISC has been awarded in 2018 the "Unit of Excellence María de Maeztu" distinction, entering the selective SOMMa Alliance and thus consolidating IFISC as a reference institute in the research field of complex systems. The award has been granted by the Spanish National Agency (AEI), Ministry of Science, Innovation and Universities. Emerging from a backbone transversal research line of exploratory nature on Complex Systems, Statistical and Nonlinear Physics, IFISC has 5 research lines of transfer of knowledge in the interface with other disciplines (Quantum Technologies, Information and Communication Technologies, Earth Sciences, Life Sciences and Social Sciences). These are: i) Biocomplexity, ii) Dynamics and collective phenomena of social systems, iii) Transport and Information in Quantum Systems, iv) Nonlinear Photonics, v) Nonlinear dynamics in fluids.

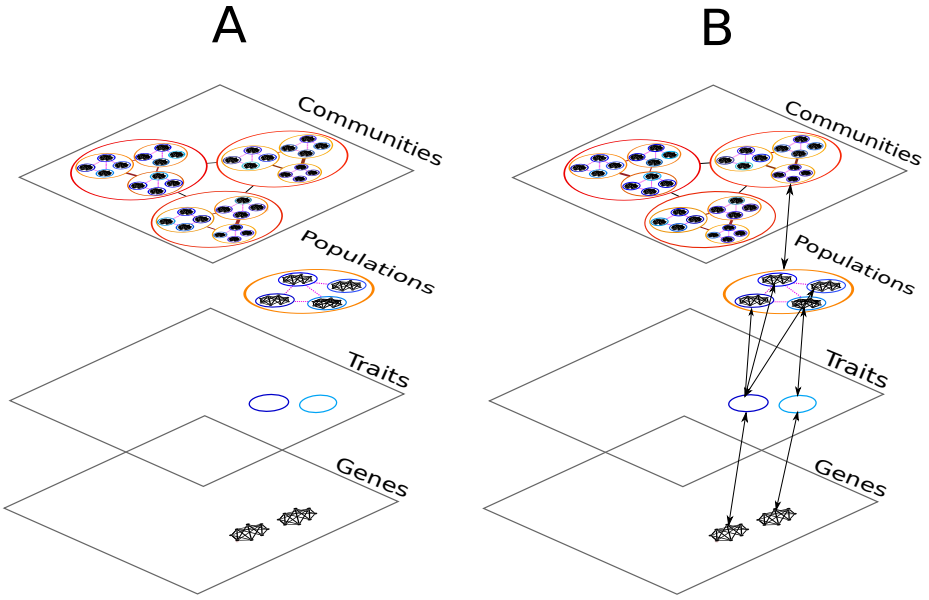


Figure 2: Biodiversity is hierarchically structured yet inferring interdependencies among the levels developing hybrid deep-process based learning approaches to predict the consequences of biodiversity decline remains poorly studied. A) Biodiversity has been studied mostly considering independent levels, 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.

As a member of IFISC, we have access to cutting edge facilities. These include a computer cluster with 46 nodes and a total of 552 cores and 3.1TB of RAM and configured for High Throughput Computing (HTC) and used for intensive numerical calculations; a new cluster being deployed in December 2019 with 20 nodes with next generation AMD Epyc Rome processors with a total of 960 cores and 12TB of RAM configured for High Performance Computing (HPC) to be used for big data analysis and memory intensive simulations; a MongoDB database cluster used for big data storage with a primary node with 42 TB SSD storage and 512GB of RAM a replica node with 40TB HD storage an 256GB of RAM; and a data repository with 80 TB HD storage. This is complemented by general purpose purpose equipment including a private cloud OpenNebula cluster used for virtualization with a total of 180 cores, 1.7TB of RAM and 70TB storage; a NFS disk server with 128GB of RAM and 80 TB storage; a server for data backup with 104 TB HD storage and a 44" plotter. Transparent access to computational clusters and servers is provided through a fully integrated network of about 60 Linux desktops complemented by several windows desktops and iMacs and around 40 laptops. IFISC has also a specific system to life webcast seminars and to distribute the recordings on demand.

ADD MY EXPERIENCE AS A RESEARC GROUP

6. Work plan and calendar

FIGURE 4 WITH MILESTONES

Task 1: Data selection and uploading Task 2: Multilayer inference Task 3: deep process-based learning

M1: multilayer inference package M2: deep processe-based learning method M3: visualization tool

7. Results dissemination and utilization plan

8. Budget

Human resources. We require a two-year post-doc: 38648.95 x 2= 77297.90 Travel: one-week visit every six months (VME + + postdoc to EAWAG; CJM to IFISC) 300 transport + 120*5 accommodation + 40 *5 food= 1300 x 6 = 7800 one-month stay per semester (postdoc) = 2600 Computation resources:

Pre-doc 26775,15

9. References

M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, 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.

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

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

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, 2018.

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.

Sarah Webb. Deep learning for biology. *Nature*, 554(7693):555–557, feb 2018. ISSN 0028-0836. doi: 10.1038/d41586-018-02174-z. URL `http://www.nature.com/doifinder/10.1038/d41586-018-02174-z`.

Douglas Heaven. Why deep-learning AIs are so easy to fool. *Nature*, 574(7777):163–166, oct 2019. ISSN 0028-0836. doi: 10.1038/d41586-019-03013-5. URL `http://www.nature.com/articles/d41586-019-03013-5`.

Amirhossein Mostajabi, Declan L. Finney, Marcos Rubinstein, and Farhad Rachidi. Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. *npj Climate and Atmospheric Science*, 2(1):41, dec 2019. ISSN 2397-3722. doi: 10.1038/s41612-019-0098-0. URL `http://dx.doi.org/10.1038/s41612-019-0098-0http://www.nature.com/articles/s41612-019-0098-0`.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, 2015. ISSN 14764687. doi: 10.1038/nature14236. URL `http://dx.doi.org/10.1038/nature14236`.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, jan 2016. ISSN 0028-0836. doi: 10.1038/nature16961. URL `http://dx.doi.org/10.1038/nature16961http://www.nature.com/articles/nature16961`.

Tomaso Poggio and Emilio Bizzi. Generalization in vision and motor control. *Nature*, 431(7010): 768–774, oct 2004. ISSN 0028-0836. doi: 10.1038/nature03014. URL `http://www.nature.com/articles/nature03014`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`.

David Ferrucci, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T. Mueller. Watson: Beyond jeopardy! *Artificial Intelligence*, 199-200:93 – 105, 2013. ISSN 0004-3702. doi: https://doi.

org/10.1016/j.artint.2012.06.009. URL http://www.sciencedirect.com/science/article/pii/S0004370212000872.