5. Long-term research agenda

Biodiversity informatics is the science that holds the key to many of the challenges facing humanity, especially when it comes to combating the effects of climate change and biodiversity loss: only when we have described and cataloged the species on Earth and their interactions, can we create realistic models that help us understand the potential impact of protective measures that humanity may undertake. A successful approach in biodiversity informatics comprises of two facets: (a) a data aspect—creating and, more importantly, interlinking primary biodiversity data on species occurrences, interactions, genomic data, etc. and (b) a modeling aspect—using computational models to understand how biodiversity changes with time and interacts with its environment both in the past and in the future.

In the data aspect, the challenge is that the task of describing biodiversity is not finished, let alone cataloging it and properly interlinking it. Ultimately, we cannot hope to do these steps in this sequence and rather should opt for a continual approach that describes and catalogs, and interlinks at the same time.

In the modeling aspect, the challenge is to overcome the current crisis in reproducibility in science, which is evidenced by numerous studies indicating that a significant portion of scientific research is not replicable. This reproducibility crisis stems from various factors including bad quality data, but also pressure to publish positive results, which often lead scientists to try models that they do not fully understand. The second factor can be combated by better modeling frameworks such as probabilistic programming languages.

The Habsburg AI's

The major scientific insight that I would like to highlight and inform this research proposal is known as the Habsburg AI. Philosophers of science since the 1980s have speculated that given exponential improvements in computing (e.g. the so-called *Moore's law*), machines of super-human level intelligence will appear, and once they appear they will start improving on themselves even faster leading to an event known as *singularity*. In reality, however, we observe that even though AI systems of the current generation outperform humans on many tasks and show "sparks of general intelligence", they are not only not capable of self-improvement, but, worse, they deteriorate in capacity if trained on self-generated data[1]. This technological "inbreeding" has been called Habsburg AI.

The implications of Habsburg AI's for the field of AI are that collective *human intelligence* is what makes AI systems smart and that AI systems need to maintain their expertise by relying on human input by curating large data-streams. So for now, human expertise cannot be automated away, but the focus of human researchers, even in fields such as biology will be shifting more towards human-computer interaction.

The many cultures

In my experience as an early-stage researcher I have been convinced that there are many cultures within the science of biology: traditional taxonomists, ecologists and evolutionary biologists, people with mathematical but not computational backgrounds, computational and machine-learning and data-driven researchers. Representatives of the three cultures often work together and have considerable difficulties communicating and exchanging ideas because as C.P. Snow points out they each have their own climate of thought and intellectual approach. We need to find a way to bridge the many cultures gap between humans and find ways to encode taxonomic, ecological, and evolutionary expertise into AI systems. We also need to enable humans with domain expertise to easily create computer models that work.

For this reason, my research vision is to *harness human intelligence for biodiversity informatics* by (a) creating state-of-the-art intelligent (semantic) databases of biodiversity knowledge by processing vast amounts of biodiversity literature and (b) creating cutting-edge statistical models of ecology and evolution by enabling biologists to use probabilistic programming languages (PPLs).

[1] S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, and R. G. Baraniuk. Self-consuming generative models go mad. arXiv:2307.01850, 2023.