PIs: Otso Ovaskainen, Tomas Roslin, David Dunson
Poster by: Bess Hardwick, Deirdre Kerdraon and Michele Peruzzi

LIFEPLAN

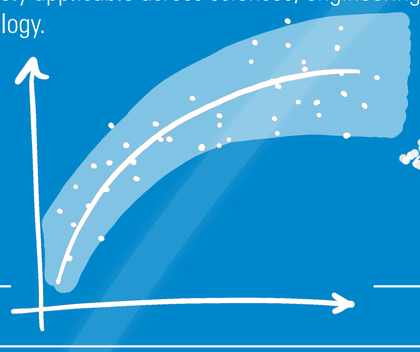
A PLANETARY INVENTORY OF LIFE

visit the LIFEPLAN website!

AMBITIONS

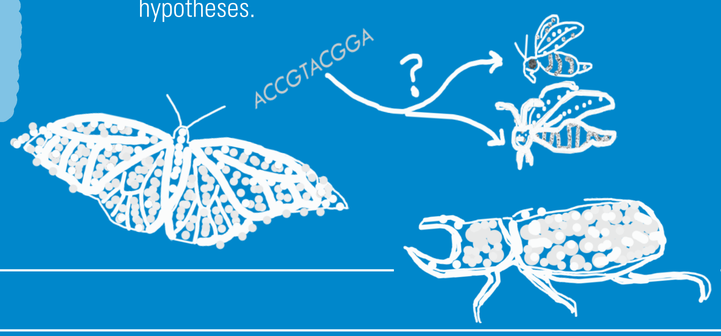
A *new generation* of statistical methods

- For huge dimensional, highly structured, sparse and imbalanced data.
- Widely applicable across sciences, engineering and technology.



A *new understanding* of biodiversity

- Global Joint Species Distribution Models will move community ecology towards a predictive science.
- Addresses long-standing unsolved ecological hypotheses.



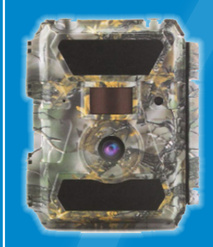
A *transformative* data resource

- Robust data across the globe on millions of species.
- Spatial resolution from 100m to the global scale.

Training *Scientific Data Scientists*

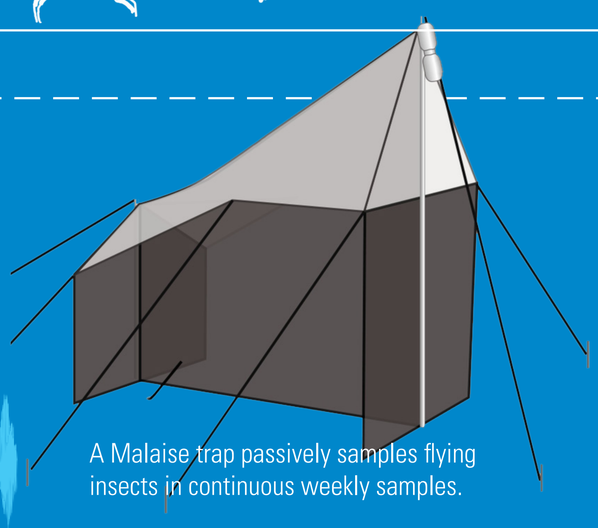
- Tools to build & implement realistic models & algorithms driven by scientific knowledge.
- Fundamentally different from "industry" data scientists.

Semi-automated sampling methods ...



Five camera traps at each plot capture images of animals day and night.

Five AudioMoths at each plot record at bird and bat frequencies on a programmed schedule.



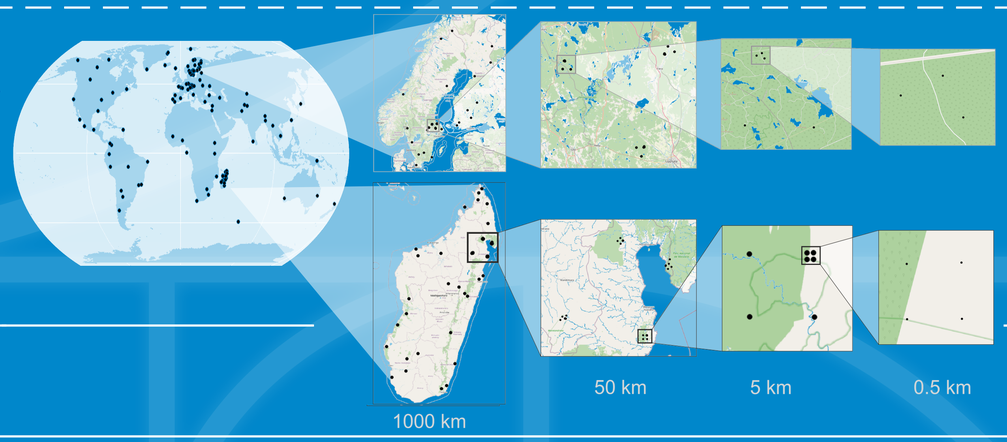
A Malaise trap passively samples flying insects in continuous weekly samples.

A cyclone sampler collects fungal spores from the air in weekly 24-hour samples.

Soil samples are taken from five places at each plot four times a year and sequenced for fungal DNA.

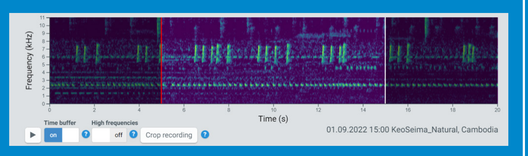
... at multiple spatial scales

- We sample at spatial scales from global to local from 2020 to 2025.
- Spatial scales are relevant for different taxa and ecological processes; sampling at such a range of scales allows us to study these relationships.
- Our highest-density sampling happens in the Nordic countries and in Madagascar. Plots in the global design switch between a natural and an urban environment once a year to study the effects of urbanisation on different taxa.
- Global Lifeplan sampling is done by teams around the world who collect data on their field sites. Each team owns their own data and their results. We commit to the principles of FAIR and open data.



Training ML methods via crowdsourced annotation

- We collect example images and sounds of the species in our data. If you can recognise birds from their vocalisations, you can help us by going to *bsg.laji.fi* and identifying bird sounds recorded in Lifeplan.



- The example images and sounds are used to train machine learning (ML) models. With these models, we can go through the millions of images and years of sound we collect. Active learning methods optimise the amount of human annotation required.
- Our models for identifying bird species based on their vocalizations now have a higher accuracy than previous algorithms.

Lauha P et al. 2022. doi: 10.1111/2041-210X.14003

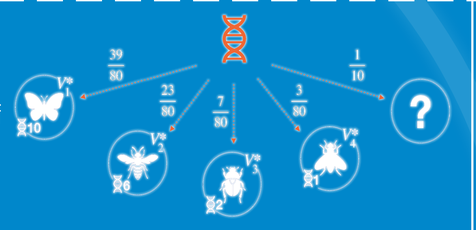
DNA metabarcoding meets image recognition

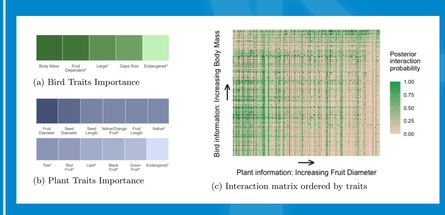
- We metabarcode all our physical samples. As we use non-destructive methods, we can bulk photograph the arthropods in the Malaise samples after metabarcoding.
- For a subset of individuals, we also generate individual images and barcodes, machine learning training data, and bolster taxonomic reference libraries on *boldsystems.org*.

New statistical methods for LIFEPLAN data analysis

Lifeplan data is motivating the development of novel statistical models and algorithms for ecological applications, including:

- Flexible community data analysis within the Hierarchical Modelling of Species Communities (HMSC) framework
- A framework for uncovering latent interactions from biased networks





- A new class of models for discrete processes characterized by self-similarity and long range dependence such as bird vocalizations
- Scalable models for massive multivariate data as well as nonstationary processes in spatial domains constrained by barriers and boundaries.
- DNA barcoding algorithms that account for novel species of insects and fungi.

References:

Ovaskainen O et al. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. doi:10.1111/ele.12757

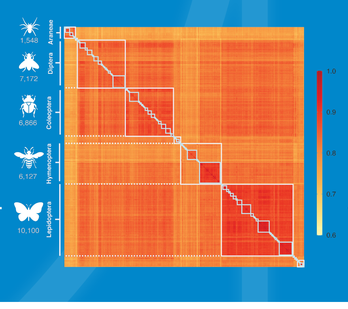
Papadogeorgou G, et al. 2021. Covariate-informed latent interaction models: Addressing geographic & taxonomic bias in predicting bird-plant interactions. arXiv: 2103.05557

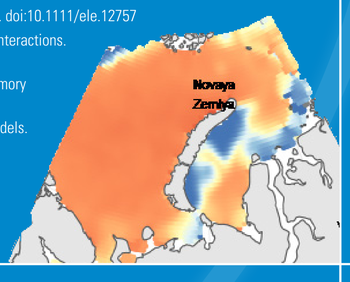
Chakraborty A, Ovaskainen O, Dunson DB. 2022. Bayesian semiparametric long memory models for discretized event data. doi:10.1214/21-aoas1546

Peruzzi M, Dunson DB. 2022. Spatial meshing for general Bayesian multivariate models. arXiv:2201.10080

Jin B, Herring AH, Dunson DB. 2022. Spatial predictions on physically constrained domains: Applications to Arctic sea salinity data. arXiv: 2210.03913

Zito A, Rigon T, Dunson DB. 2022. Inferring taxonomic placement from DNA barcoding aiding in discovery of new taxa. doi:10.1111/2041-210X.14009





This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement No 856506