Natural history research as a replicable data science

Rutger Vos

Naturalis Biodiversity Center

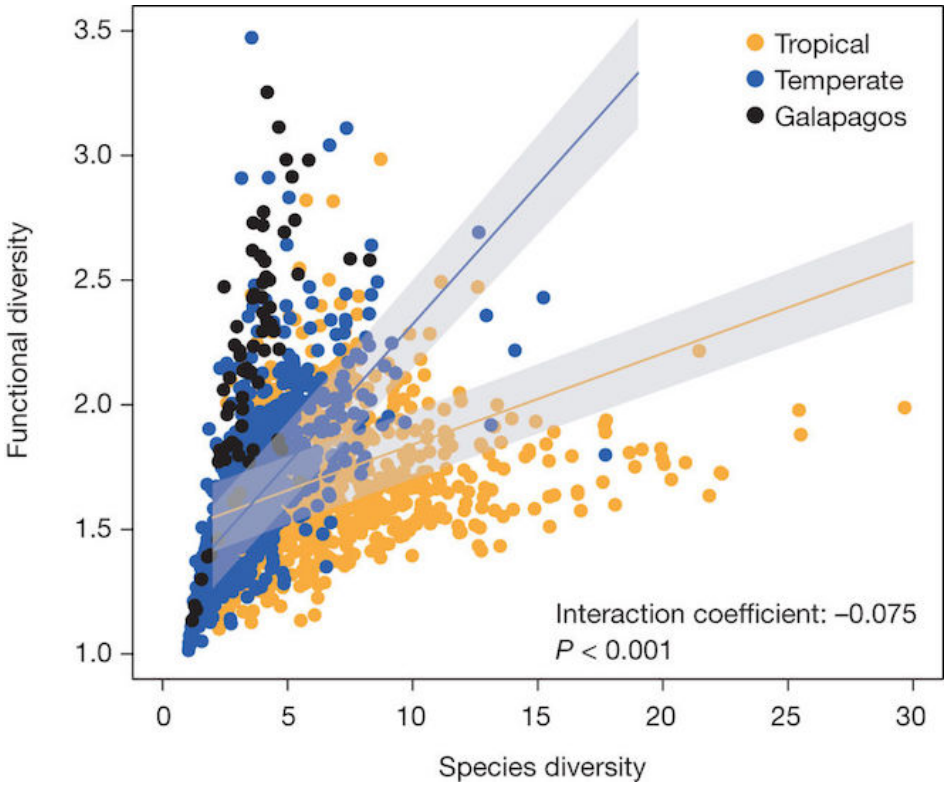
Natural history museums and collections

- Main goal is not to exhibit but to collect and curate specimens
- Usually multiple specimens per species, sometimes many more
- Specimens are research and reference materials

Natural history research

*To understand the patterns and processes of **biodiversity***

Biodiversity is expressed and studied in multiple ways:



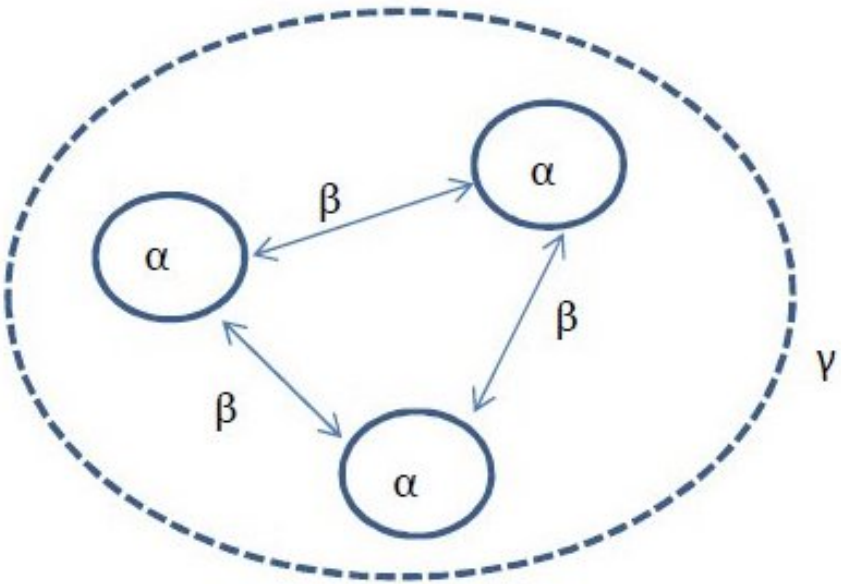
- **Species diversity**, e.g. counts of species, maybe taking abundances into account
- **Phylogenetic** diversity, i.e. the evolutionary distances between species
- **Functional** diversity, i.e. the ecological roles species play, and the characteristics associated with that role

Natural history research

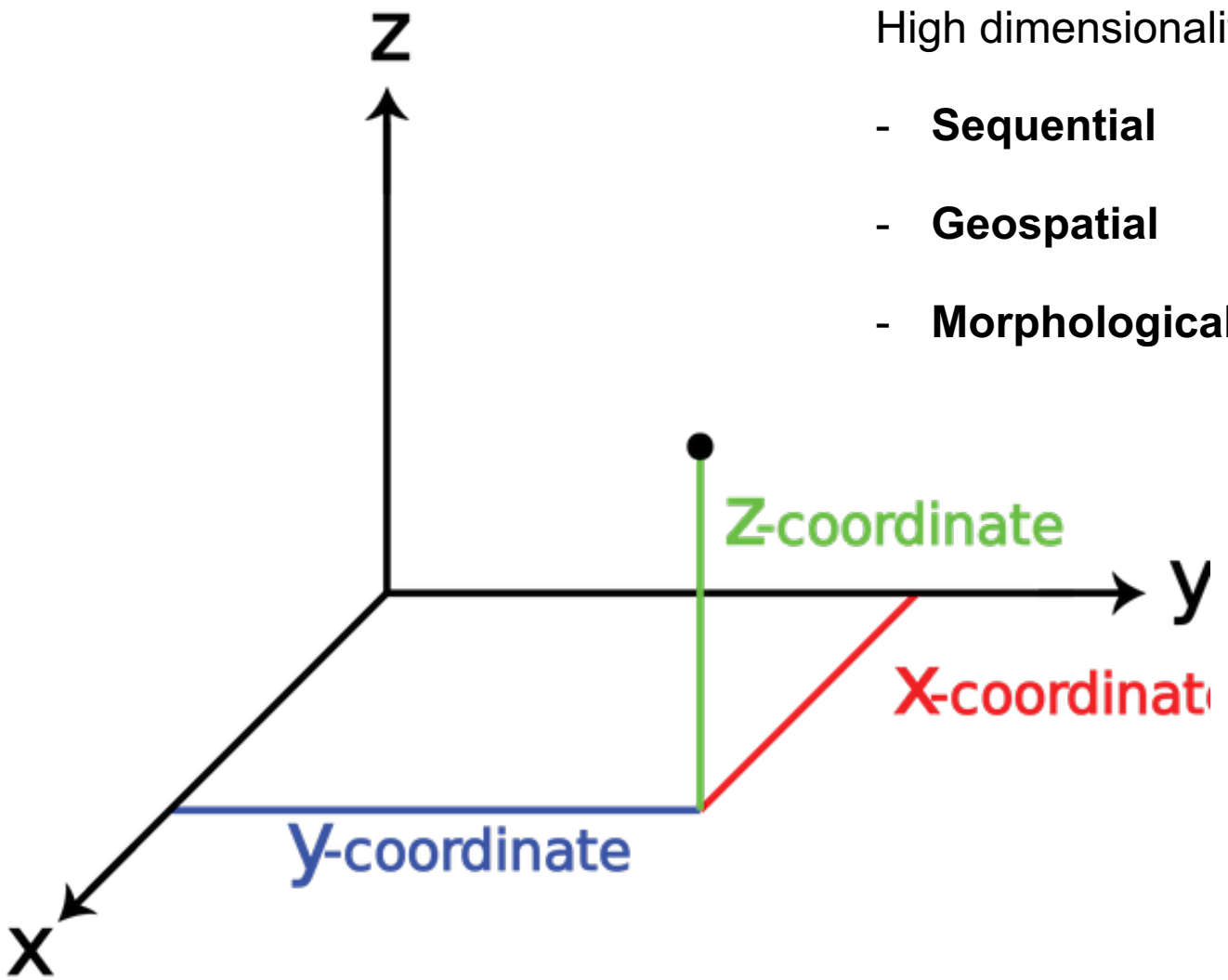
To understand the patterns and processes of ***biodiversity***

The patterns and processes of biodiversity are systematized as taking place:

- Within a given system (**α diversity**), e.g. a biome
- Across systems (**β diversity**, turnover)
- Among systems (**γ diversity**, totality)



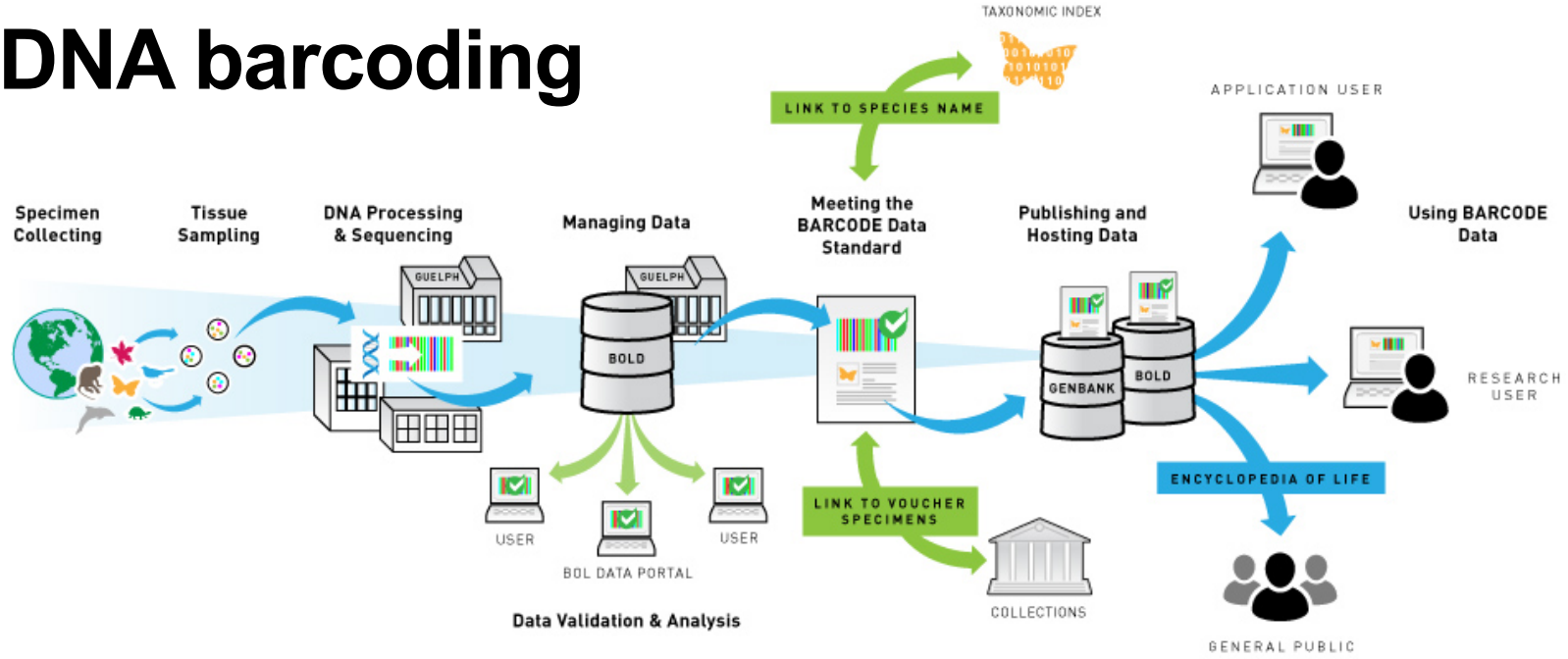
Natural history data

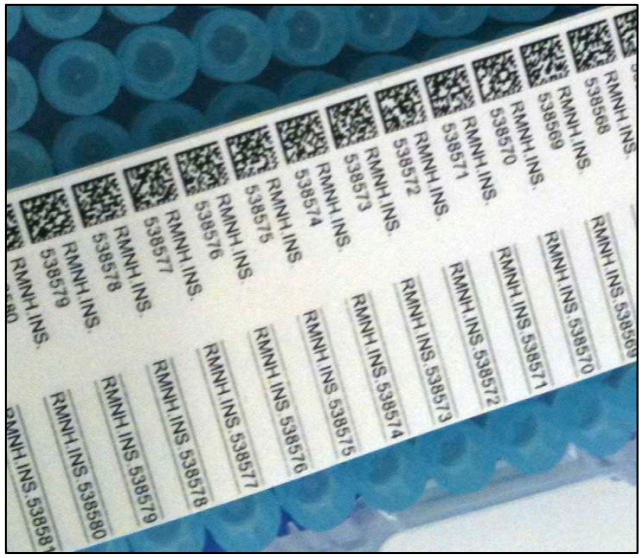


High dimensionality:

- **Sequential**
- **Geospatial**
- **Morphological**

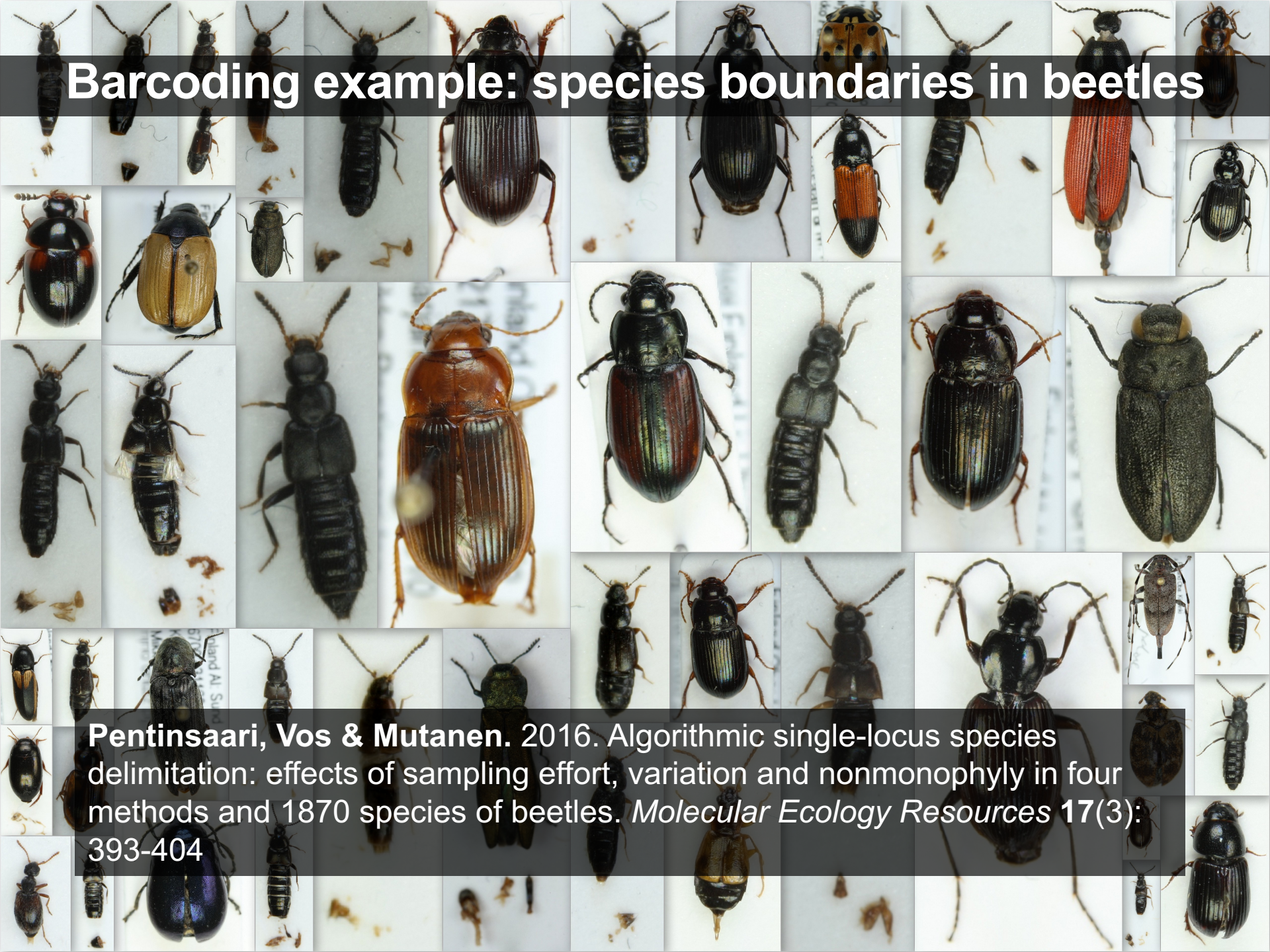
DNA barcoding





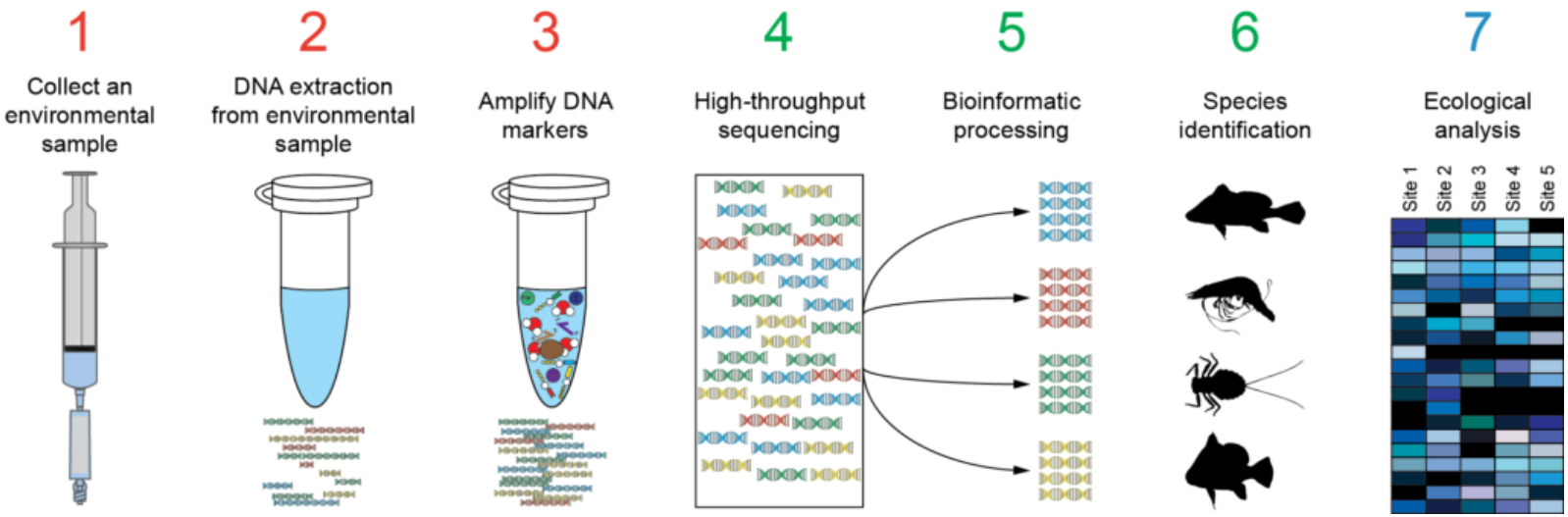
- Some genes are variable so that a few hundred letters suffice to identify species
- In addition, barcodes are useful for studying evolution and phylogeny
- Taking the barcode of a specimen (by Sanger seq) is part of the workflow of indexing collection specimens

Barcoding example: species boundaries in beetles



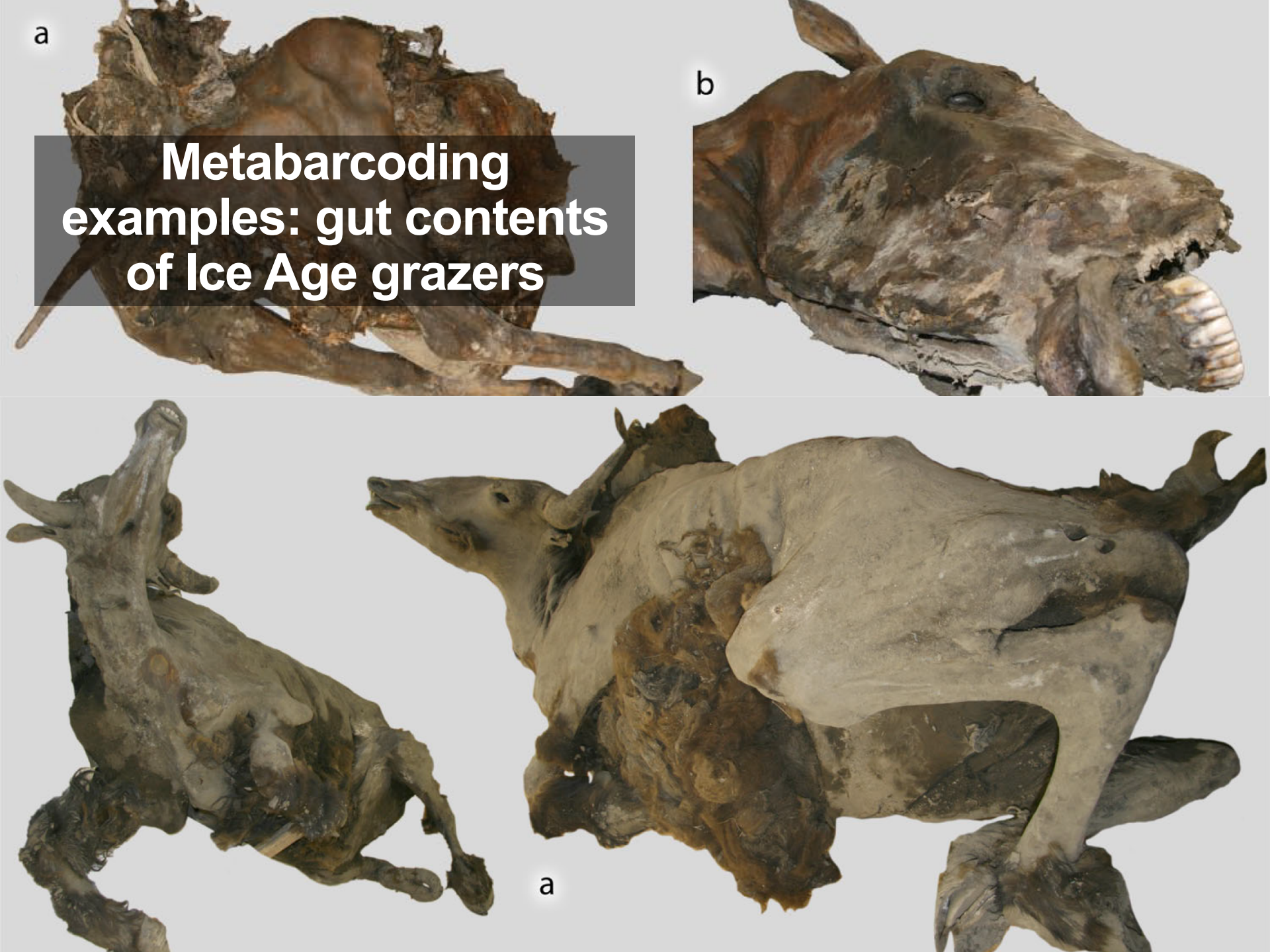
Pentinsaari, Vos & Mutanen. 2016. Algorithmic single-locus species delimitation: effects of sampling effort, variation and nonmonophyly in four methods and 1870 species of beetles. *Molecular Ecology Resources* **17**(3): 393-404

Metabarcoding

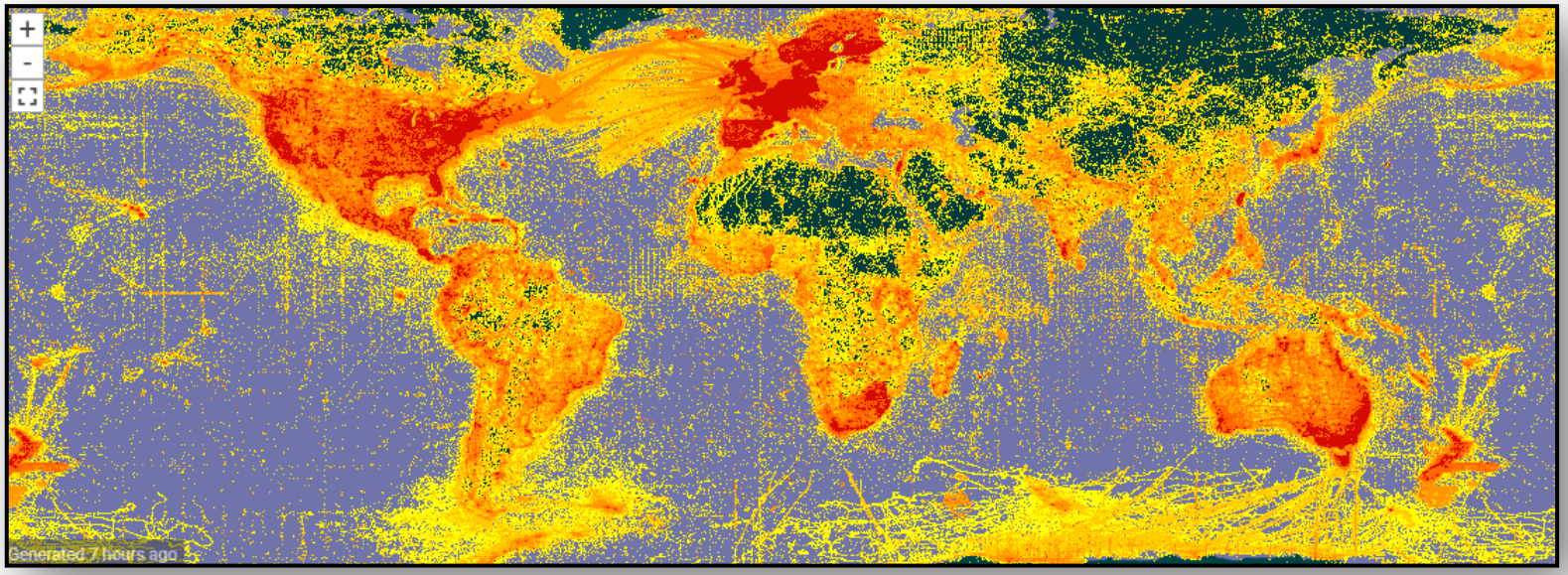


- The species contents of organic mixtures can also be identified using identifiable marker genes
- This is typically done using multiplexed, high-throughput (“next generation”) sequencing
- Consequently, data storage and processing requirements are higher

Metabarcoding examples: gut contents of Ice Age grazers

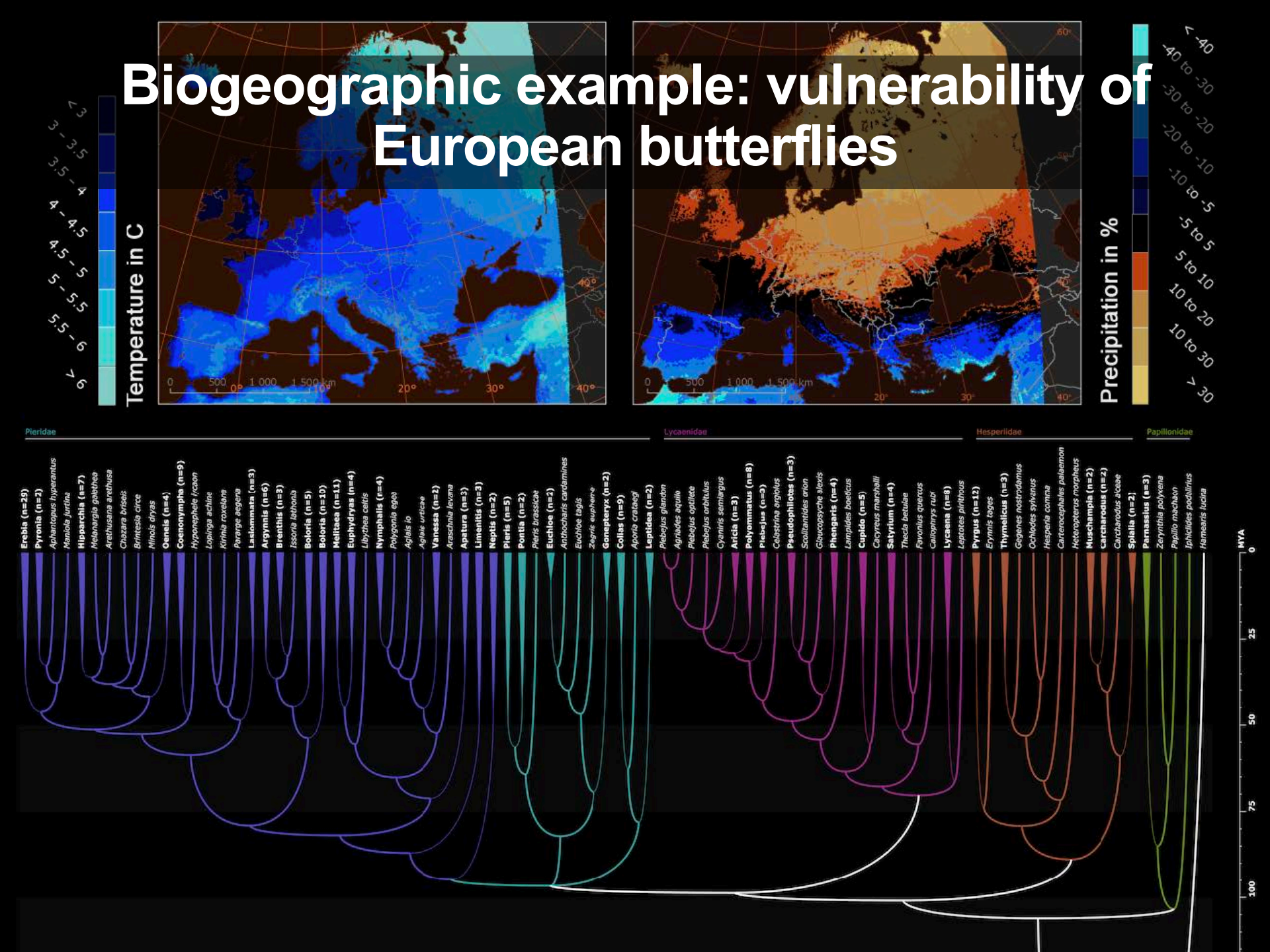


Species distribution modelling

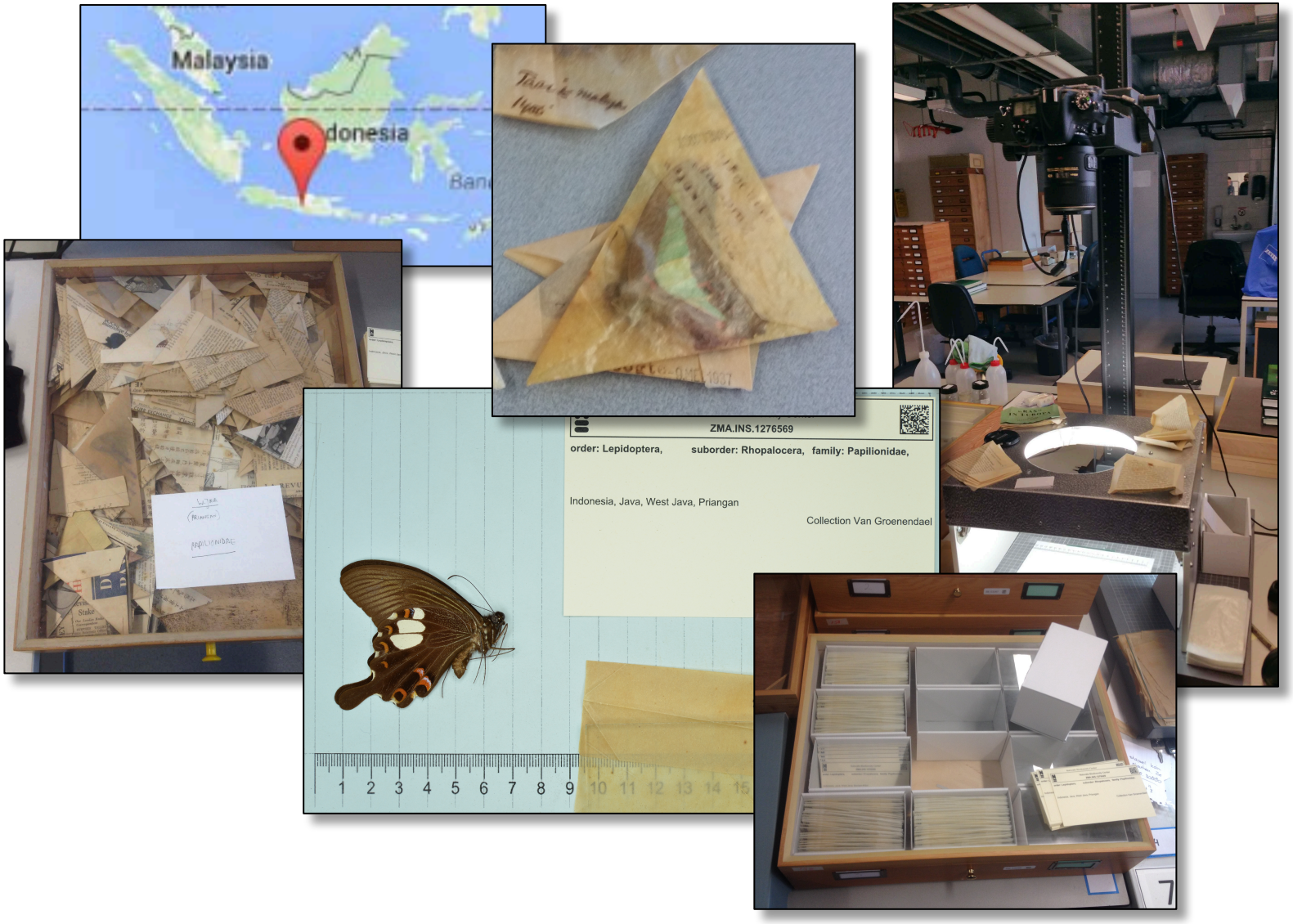


- Collection specimens are (ideally) stored with their collection locality recorded as lat/lon coordinates
- Based on the localities where specimens were found, and geospatial data layers (climate, land use, soil, etc.) a correlative model of the species affinities can be constructed
- With such a model, habitat suitability and predictive scenarios (e.g. climate change) can be projected

Biogeographic example: vulnerability of European butterflies

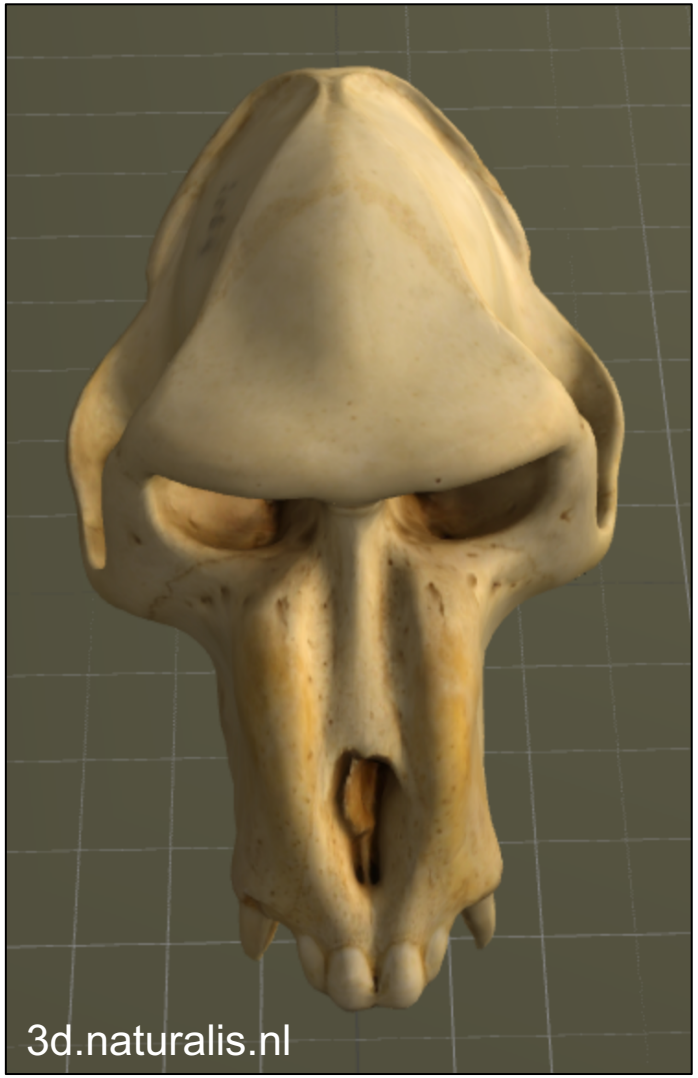


Shapes, traits, and phenotypes

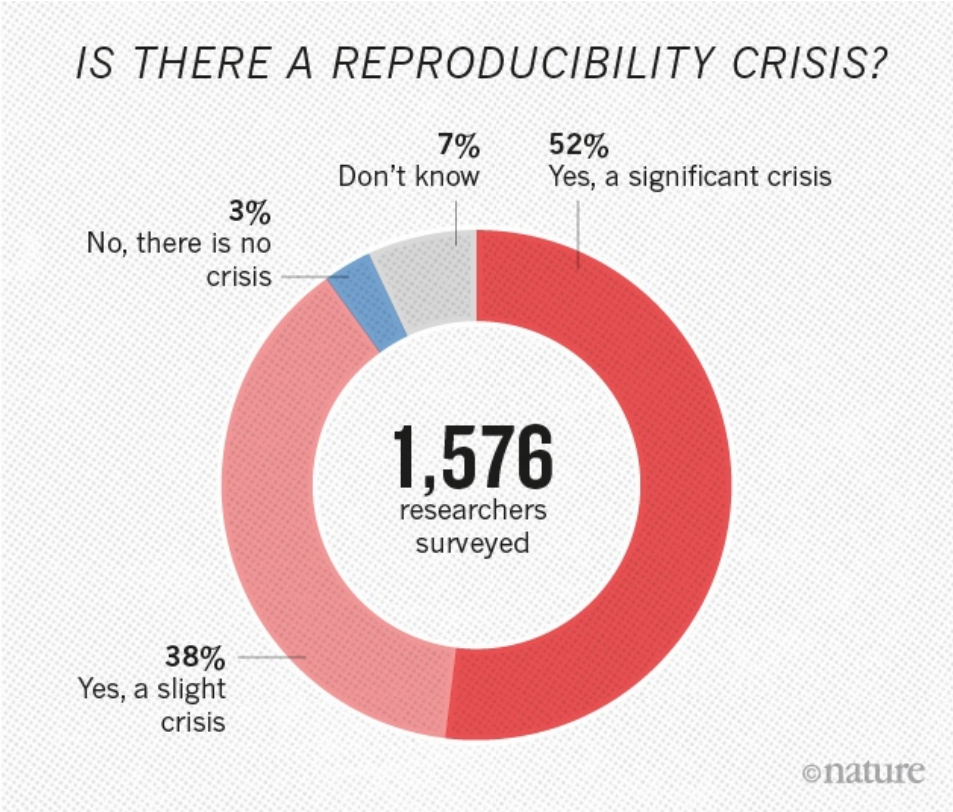


Natural history data

- Highest data volumes are HTS, 3D scanning, images
- High dimensionality at multiple scales
- Many biases in species/locality sampling
- Many axes are messy:
  - Species names have been changing for centuries
  - Likewise place names
  - Trait descriptions are often ambiguous



The Reproducibility Crisis



- More than 70% of researchers (n=1576) have tried and failed to reproduce another scientist's experiments
- More than half have failed to reproduce their own experiments

Reproducible data science and cultural change

1. *“Data available from the author upon request”*
   No: data are open, as FAIR as possible

2. *“Data were processed with custom scripts”*
   No: scripts/workflows are open source

3. *“Data were analyzed on a Pentium III 450 MHz…”*
   No: the environment can be cloned as VM

1. FAIR data management

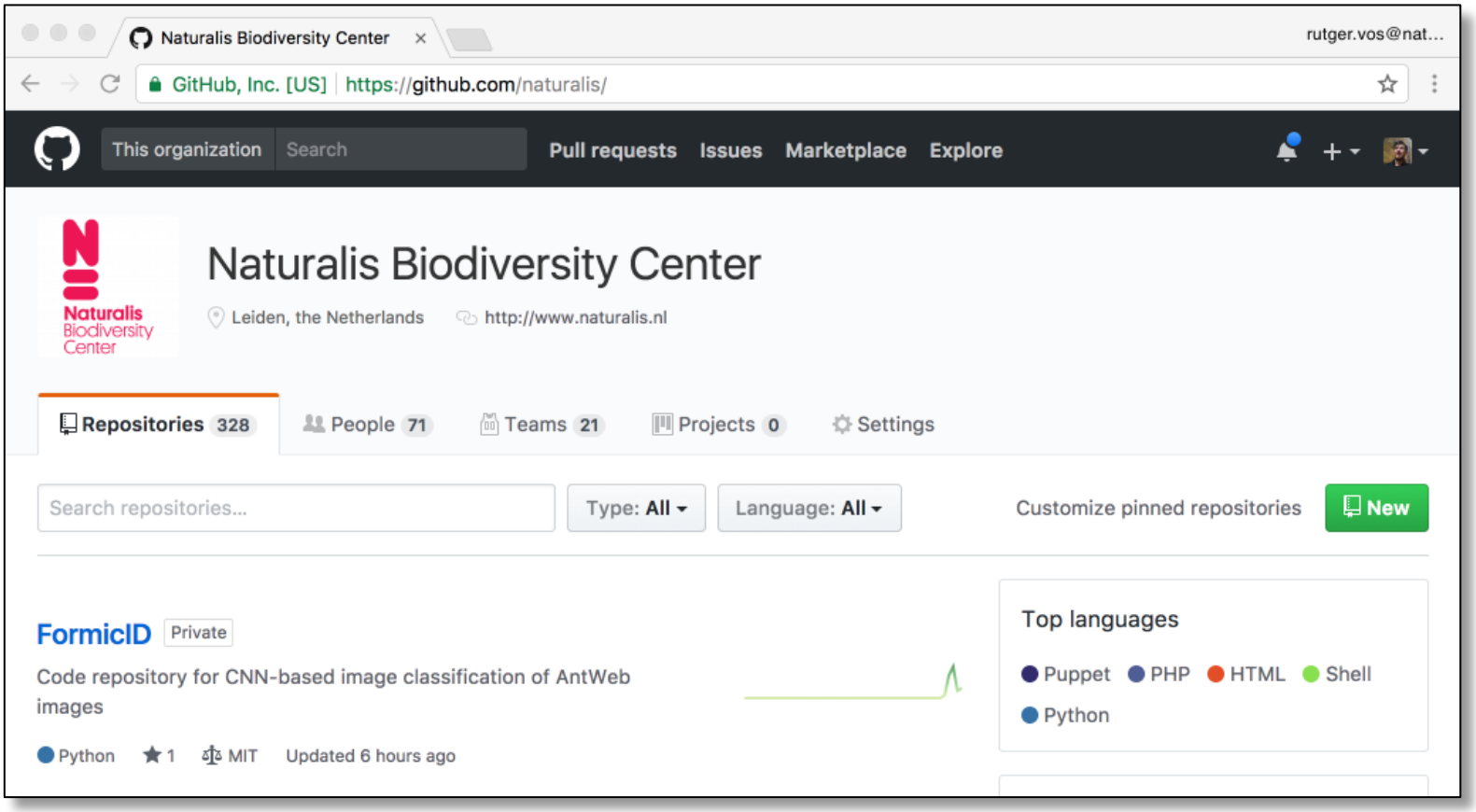
Findable: increasing attention to metadata, and discoverability and indexing of data

Accessible: implementation of resolvable identifiers, e.g. PURLs and DOIs

Interoperable: increasing attention for open community standards (syntax) and semantics

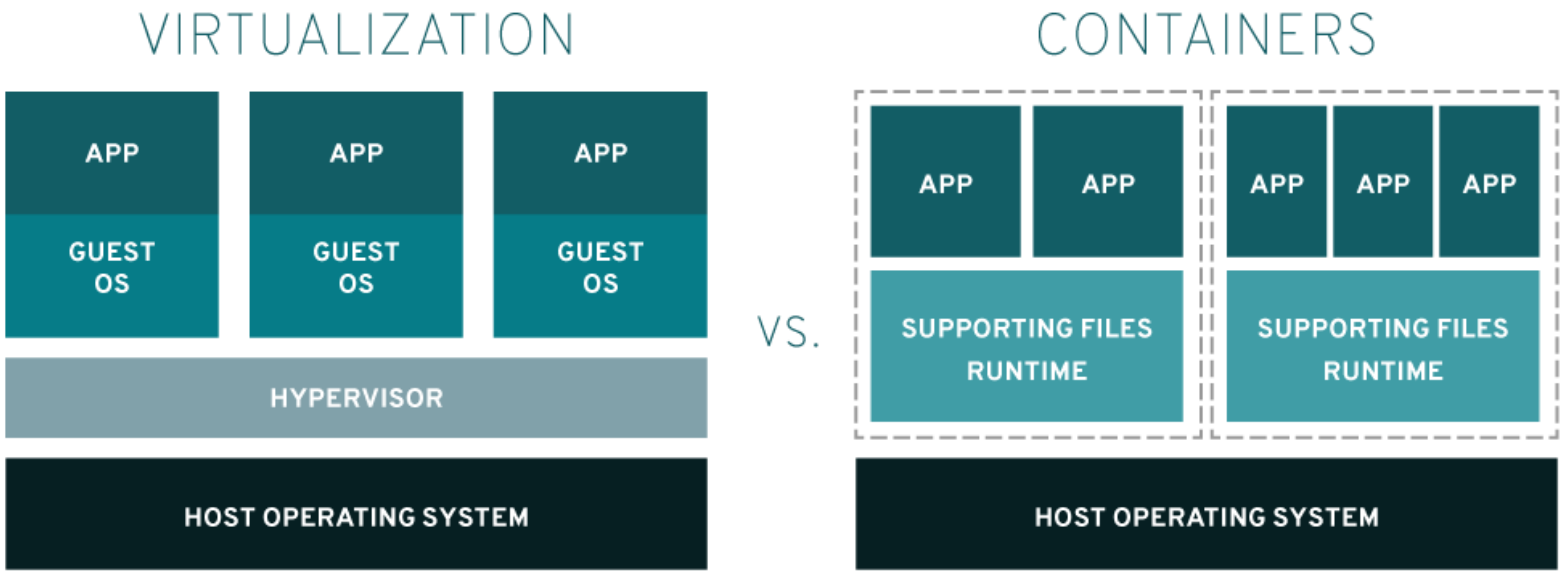
Re-usable: increasing attention for data ownership and licensing

2. Open source



Analytical code is no longer a folder on a postdoc's laptop, it's a code repository with specific versions, documentation, tests, and a license

3. Virtualization



- Analyses are not run on dedicated hardware, e.g. workstations, clusters, but in the (private) cloud

- Complex workflows are distributed as virtual machines, docker containers, or deployed with devops tools

OPEN SCIENCE
OPEN DATA
OPEN SOURCE

21st century research skills for the life sciences

Pedro L. Fernandes & Rutger A. Vos

OSODOS.ORG

Naturalis Biodiversity Center

INSTITUTO GULBENKIAN DE CIÊNCIA

Thank you for your attention

Naturalis
Biodiversity
Center