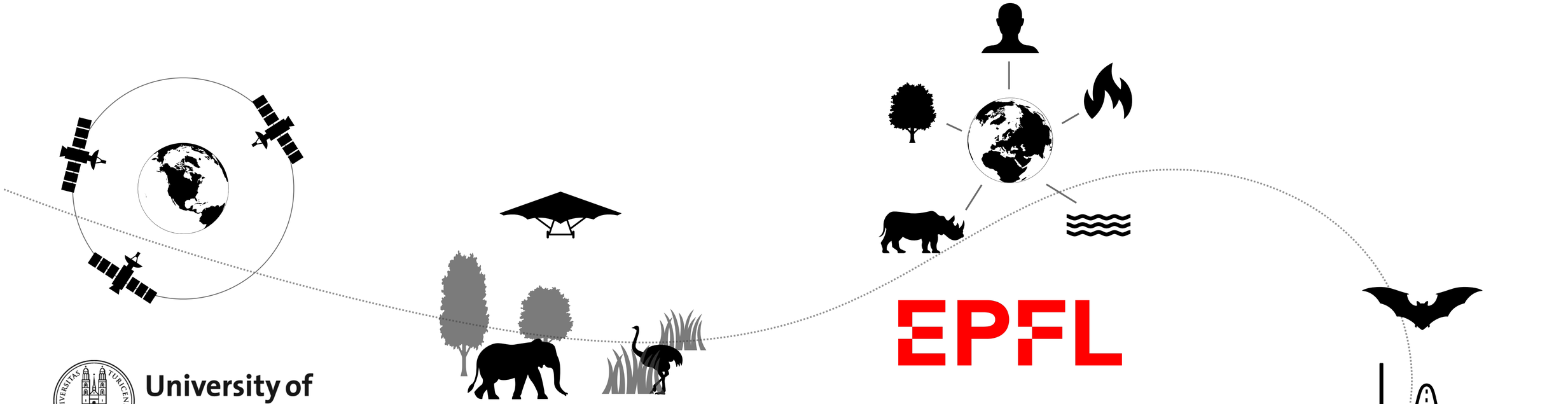
Yale

(Deep) Species Distribution Modelling: Challenges and Pitfalls

BIOS0032: AI4Environment

Benjamin Kellenberger | March 19, 2024

Yale



University of Zurich[UZH]

WAGENINGEN
UNIVERSITY & RESEARCH

EPFL

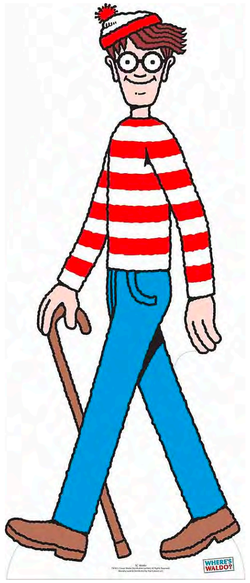
?

UCL

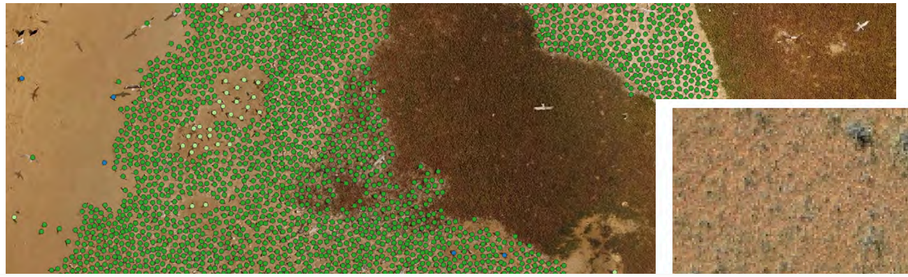
Yale

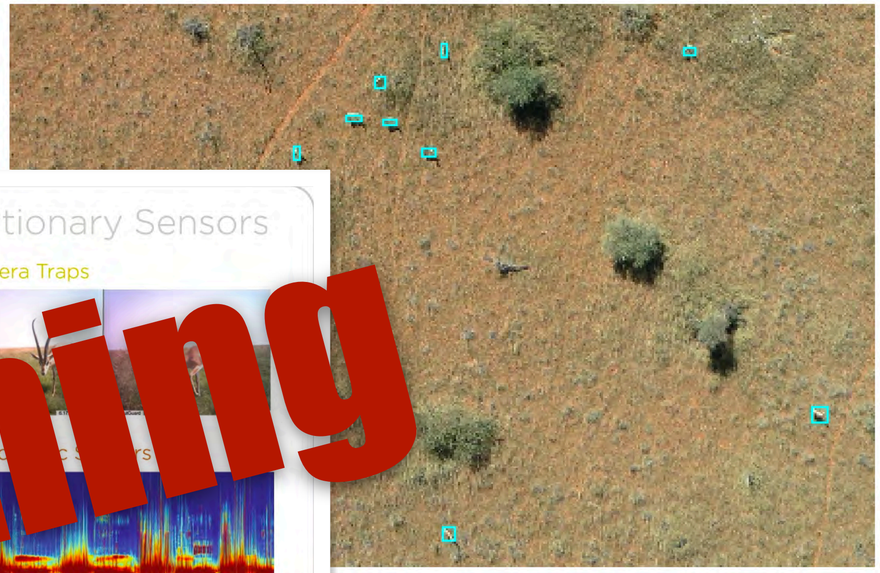
(August 2024)

Yale

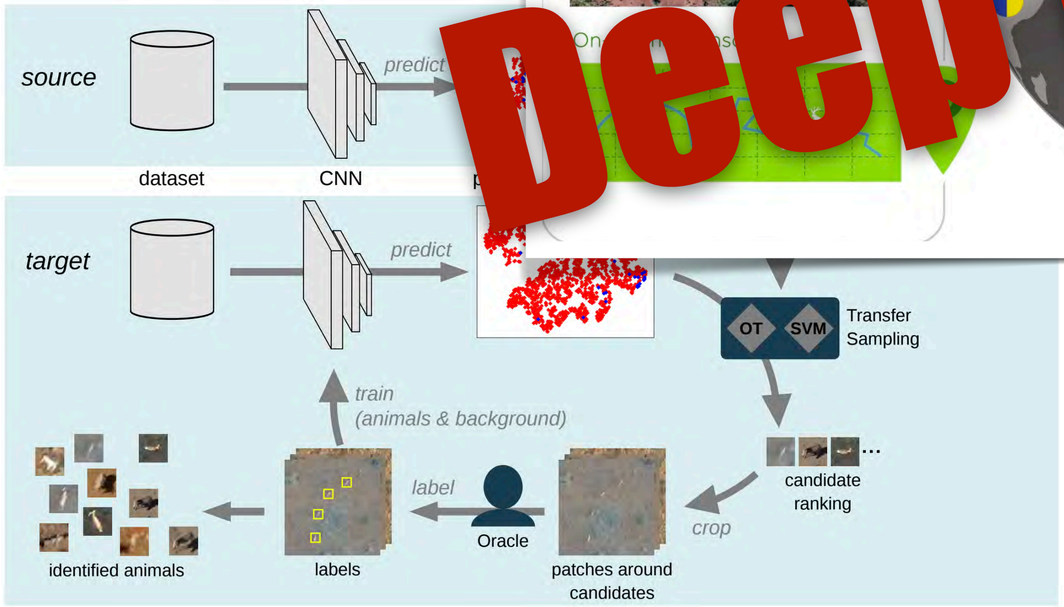


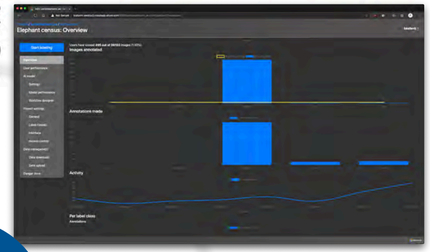
Yale

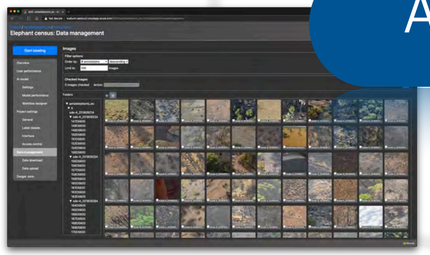




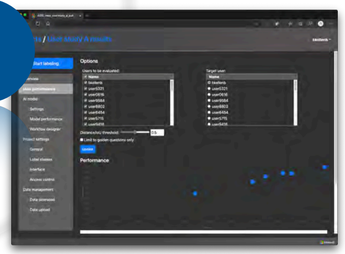
Deep Learning



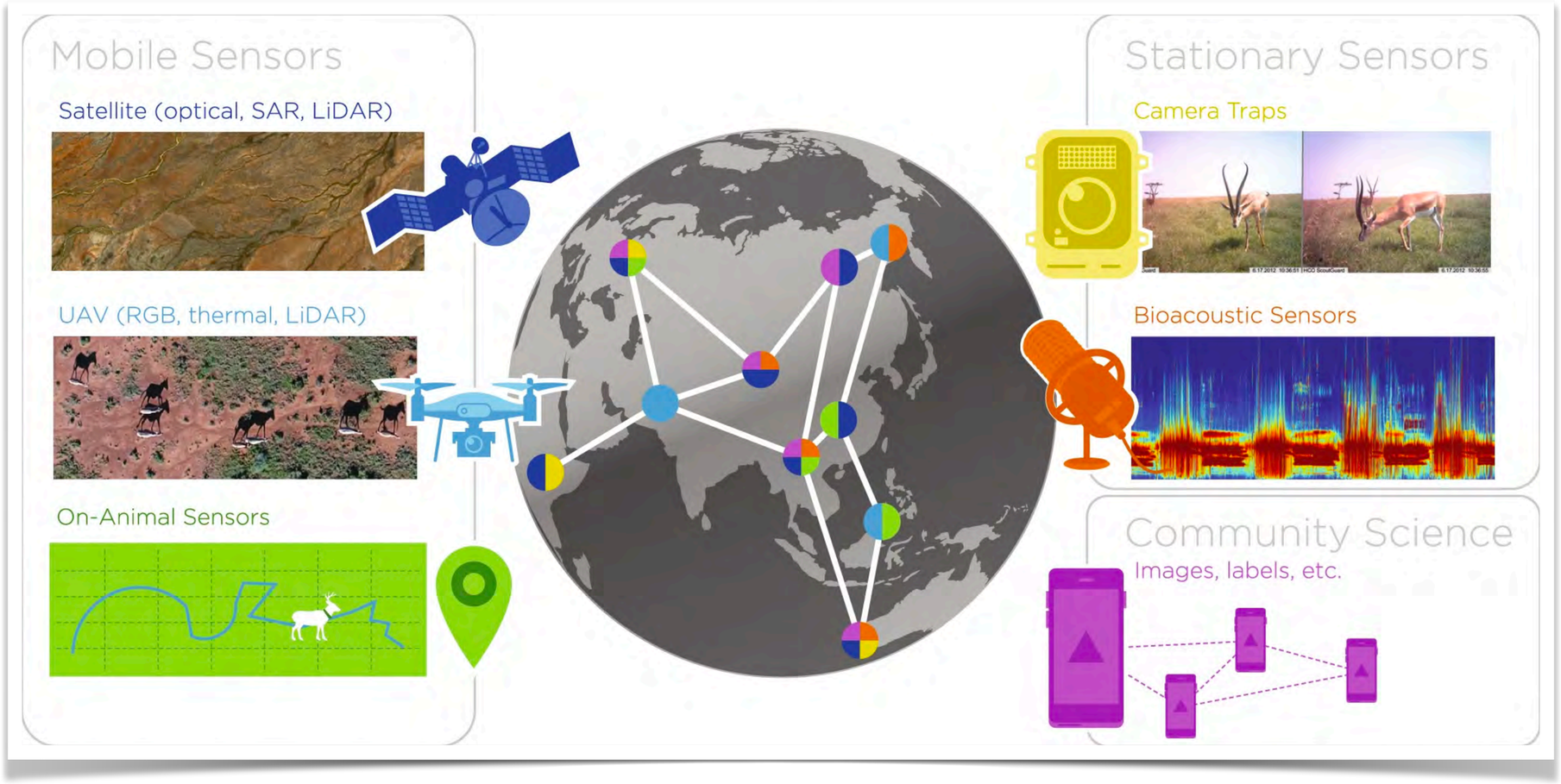




AIDE



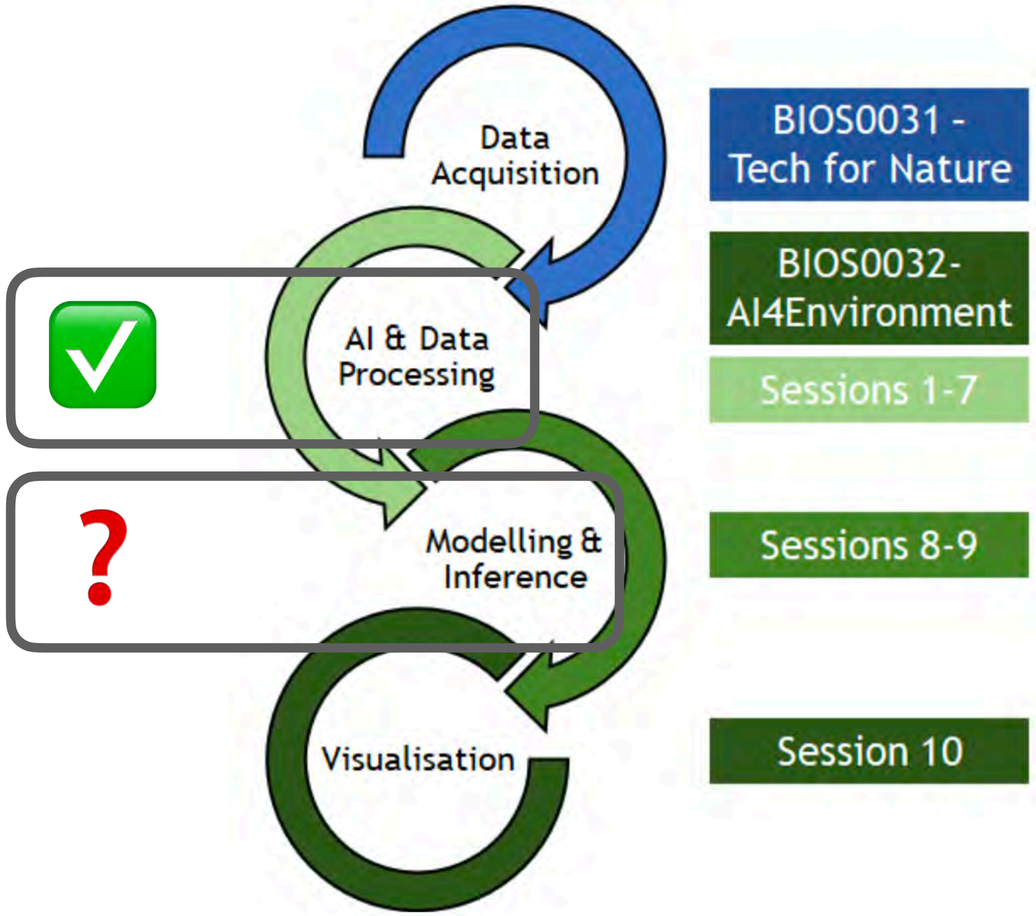
Deep Learning in Ecology?



→ computer vision, signal processing

Tuia et al., 2020

Ecological *Project* Workflow



Deep Learning in Ecology?

nature communications

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature communications > articles > article

Article | Open access | Published: 19 December 2023

Undiscovered bird extinctions obscure the true magnitude of human-driven extinction waves

Rob Cooke ✉, Ferran Sayol, Tobias Andermann, Tim M. Blackburn, Manuel J. Steinbauer, Alexandre Antonelli & Søren Faurby

Nature Communications **14**, Article number: 8116 (2023) | Cite this article

18k Accesses | **1** Citations | **2971** Altmetric | Metrics

Would you blindly rely on an all-deep learning model for this claim?

But what about LLMs?

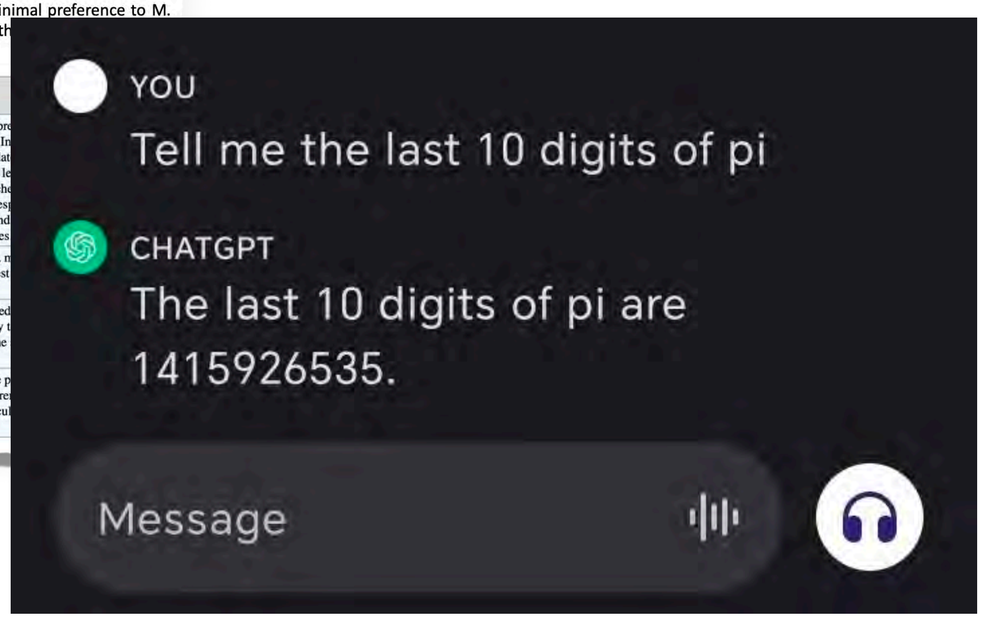
week 5: slide by Daan Scheepens

Title: Approaching and rostrum protrusion behaviours of Rhynocoris marginatus on three prey chemical cues

Abstract: Rhynocoris marginatus (F.) (Heteroptera Reduviidae) is a polyphagous predator predominantly found in agroecosystems, and their bordering ecosystems like scrub jungles, semi-arid zones and forests in India. Although R. marginatus is a polyphagous predator, it exhibited a certain degree of host specificity. Due to its predatory potential, R. marginatus has been used as an important biological control agent in India. Laboratory and field trials showed that R. marginatus feeds mainly on lepidopteran pests followed by coleopteran pests. R marginatus locates the preys by the chemical cues emanating from them. Approaching and rostrum protrusion behaviours of R. marginatus life stages on hexane extract of three groundnut pests, Helicoverpa armigera (Hubner) (Lepidoptera Noctuidae), Spodoptera litura (F.) (Lepidoptera Noctuidae) and Mylabris pustulata (Thunberg) (Coleoptera Meloidae). Significantly R. marginatus adult was found to be more responsive to the chemical cues of S. litura (62.5%) followed by H. armigera (60%) and M. pustulata (40%). R. marginatus showed minimal preference to M. pustulata chemical cues as compared to H. armigera and S. litura chemical cusses. The prey's chemical cues elicited a quicker approaching behaviour of the predator th

GPT-4:

Class	Order	Family	Genus	Species	Role	Generalist /Specialist	Pest Controller	Pest Names	Pest Type	Associated With	Affects	Description
Insecta	Hemiptera	Reduviidae	Rhynocoris	marginatus	Predator	Generalist	TRUE	Helicoverpa armigera, Spodoptera litura, Mylabris pustulata	Invertebrate	Agriculture	Groundnuts	R. marginatus is a polyphagous predator found pr… in agroecosystems and bordering ecosystems in In… used as a biological control agent due to its predat… potential and host specificity. It feeds mainly on le… pests followed by coleopteran pests located by ch… emanating from them. The adult stage is more resp… the chemical cues of S. litura and H. armigera and… minimal preference to M. pustulata chemical cues
Insecta	Lepidoptera	Noctuidae	Helicoverpa	armigera	Pest		FALSE		Invertebrate	Agriculture	Groundnuts	H. armigera, a lepidopteran pest, is a prey for R. m… which is attracted by its chemical cues. It is a pest… agriculture industry and affects groundnuts.
Insecta	Lepidoptera	Noctuidae	Spodoptera	litura	Pest		FALSE		Invertebrate	Agriculture	Groundnuts	S. litura is another lepidopteran pest that is preyed… marginatus. R. marginatus responds significantly t… chemical cues of S. litura. S. litura is a pest in the… industry and affects groundnuts.
Insecta	Coleoptera	Meloidae	Mylabris	pustulata	Pest		FALSE		Invertebrate	Agriculture	Groundnuts	M. pustulata is a coleopteran pest and one of the p… marginatus. R. marginatus shows minimal prefere… chemical cues. M. pustulata is a pest in the agricul… industry and affects groundnuts.

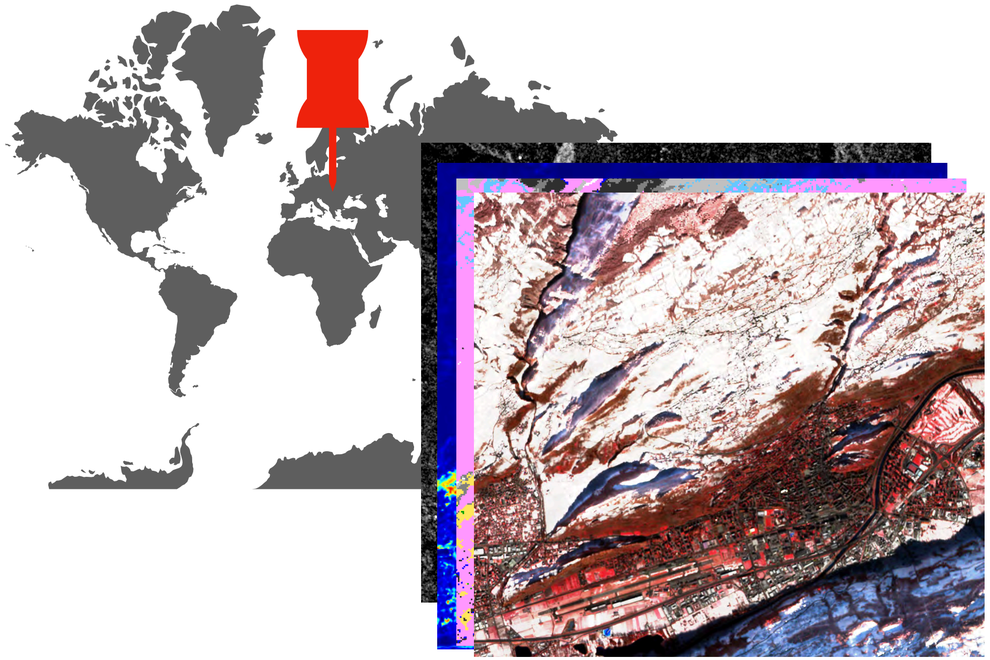


Tremendously useful! However...

https://www.reddit.com/r/ChatGPT/comments/18alu07/how/

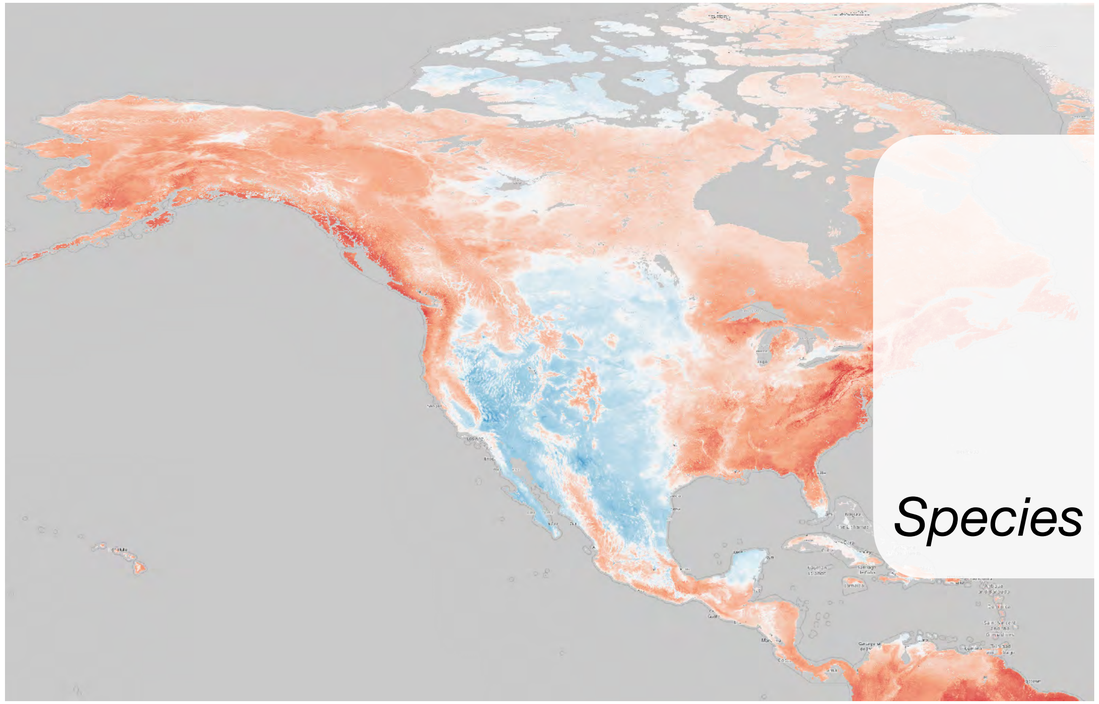
In a nutshell

"Is this a cat or a dog?"

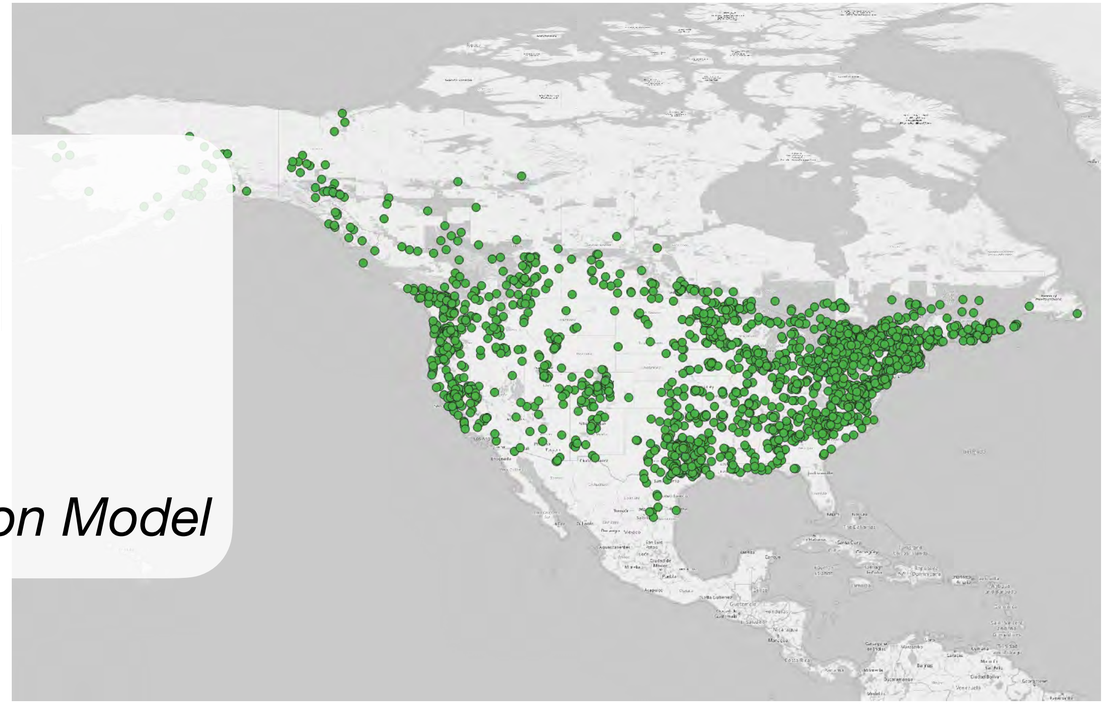


"Is this area suitable
for *Bythiospeum alpinum*?"

SDM: Species **D**istribution **M**odelling



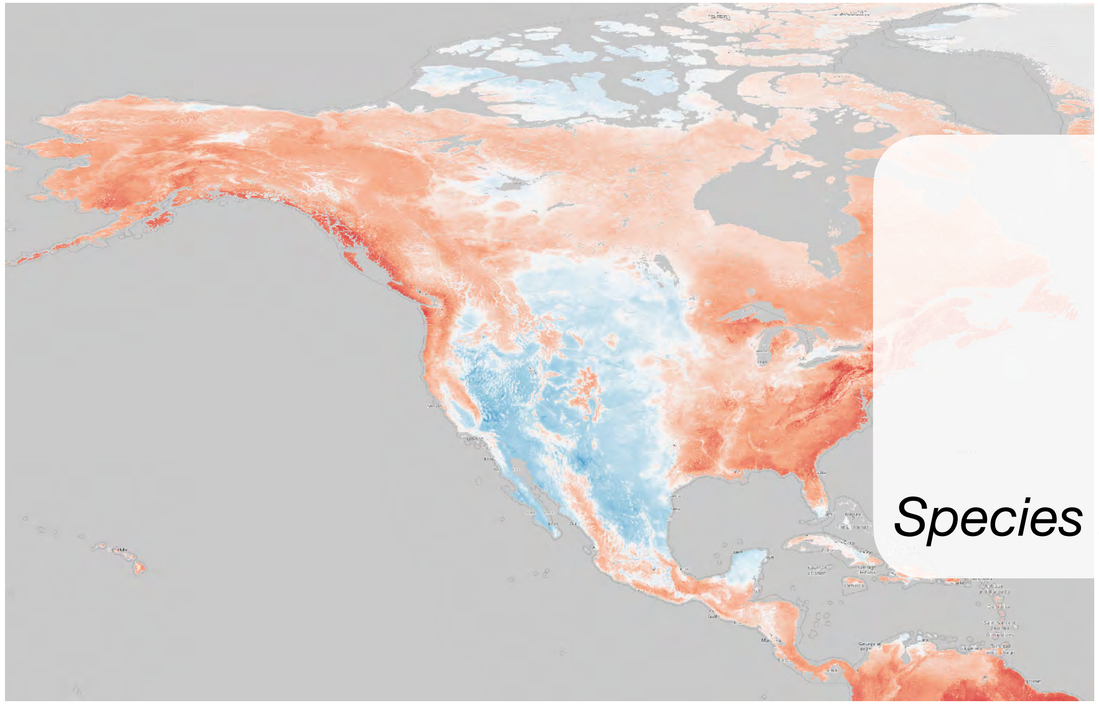
SDM
Species Distribution Model



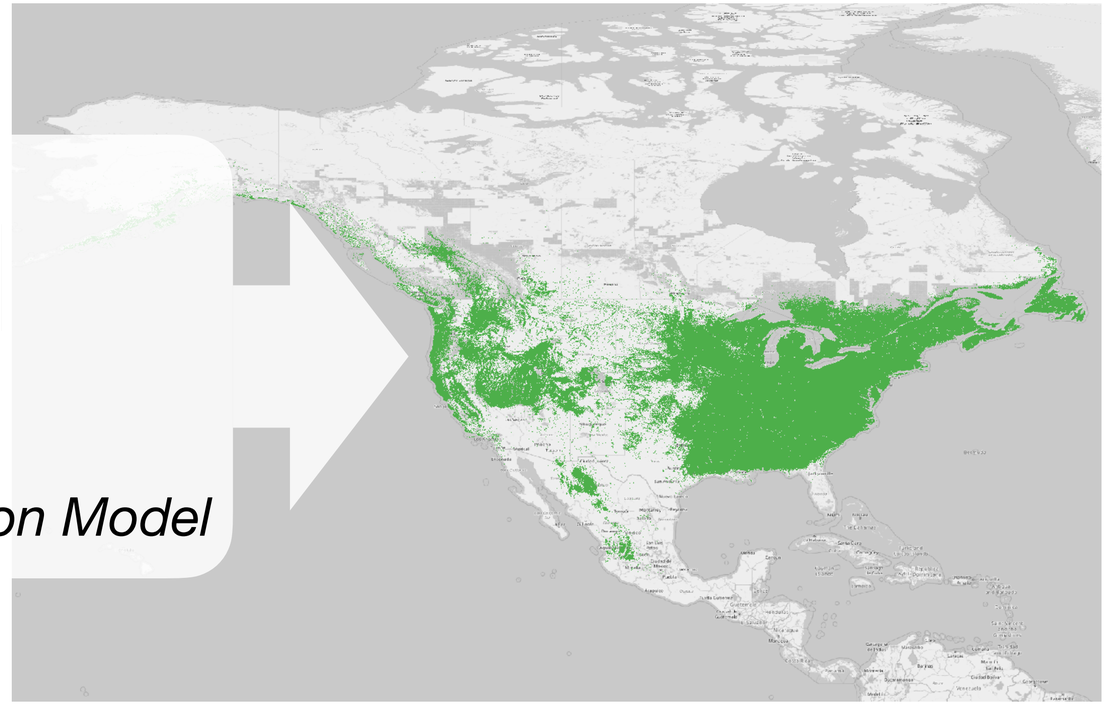
environmental **covariates**
- annual temperature
- precipitation
- soil pH
- ...

species **observations**
- field surveys
- expert range maps
- iNaturalist, eBird, *etc.*

SDM: Species **D**istribution **M**odelling



SDM
Species Distribution Model

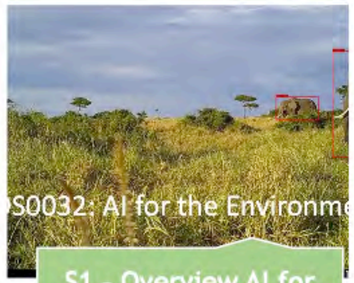


environmental **covariates**

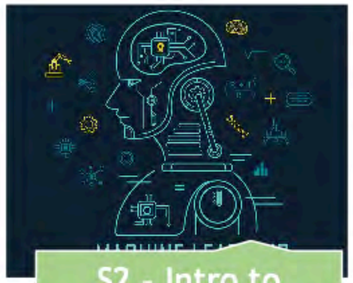
- annual temperature
- precipitation
- soil pH
- ...

predicted **suitability score**

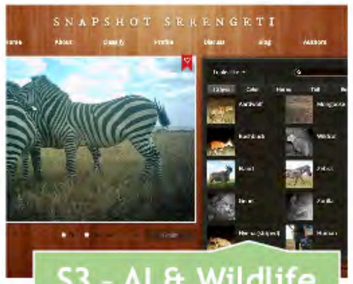
from Session 1



S1 - Overview AI for the Environment

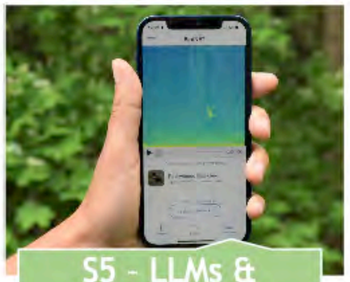


S2 - Intro to Machine Learning



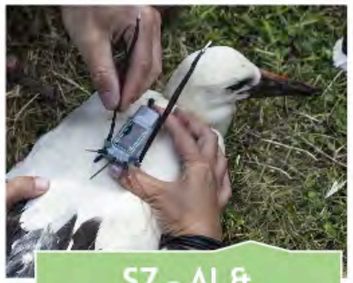
S3 - AI & Wildlife Images

S4 - AI & Bioacoustics

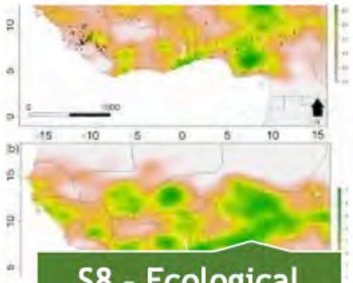


S5 - LLMs & AI Project

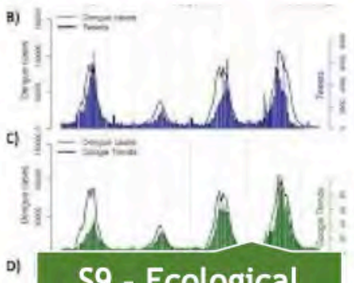
S6 - AI & Remote Sensing



S7 - AI & Movement Data



S8 - Ecological Models I



S9 - Ecological Models II

S10 -Wrap Up Session

from Session 1

Species Distribution Modelling (SDM)

SDM is all this...

S1 - Overview AI for the Environment

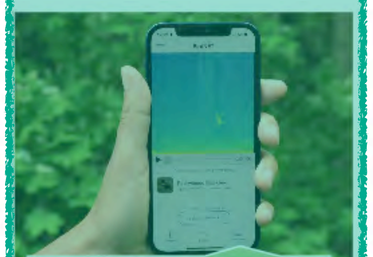
S2 - Intro to Machine Learning

...it often is this...

S3 - AI & Wildlife Images

S4 - AI & Bioacoustics

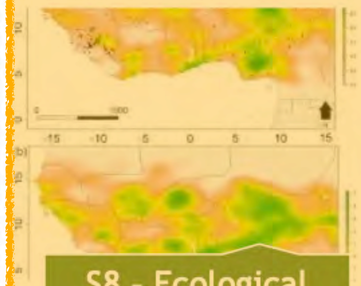
...and it can be that, too.



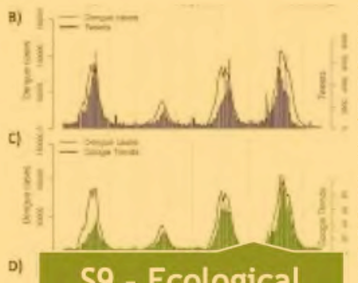
S5 - LLMs & AI Project

S6 - AI & Remote Sensing

S7 - AI & Movement Data



S8 - Ecological Models I



S9 - Ecological Models II

S10 -Wrap Up Session

Learning Objectives

- Understand the **motivation** for, and **working principle** of, *correlative* SDMs.
- Can outline a machine learning **pipeline for SDMs** from data to inference.
- Know the **pitfalls and biases** of such pipeline and can identify them.
- Understand the place of **deep learning in ecology**:
  a. at the moment;
  b. in the future (potentially).
- Can **create, and test, SDMs** in practice (*cf.* exercise).

SDMs are everywhere

Example: COP 15

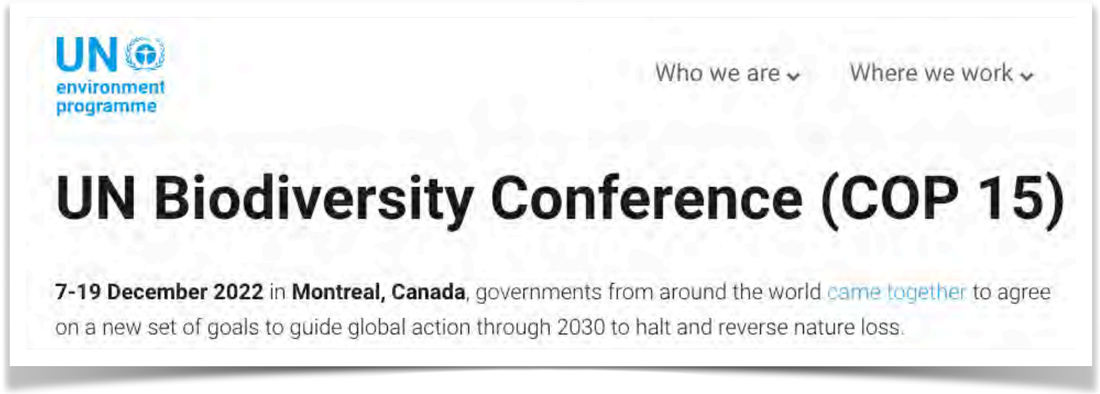


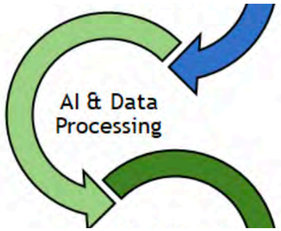
Table 2.

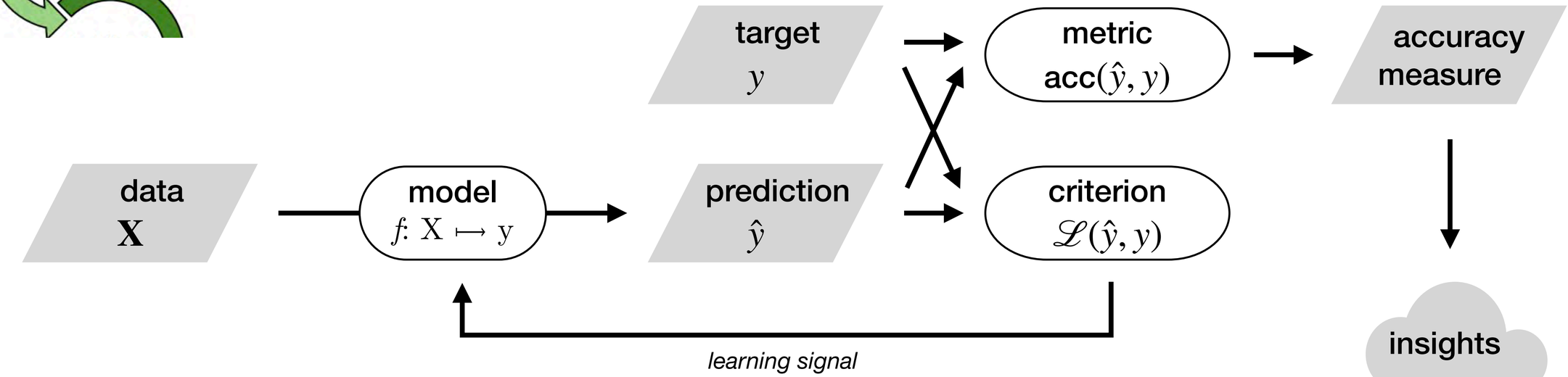
Proposed indicators for the Kunming-Montreal global biodiversity framework

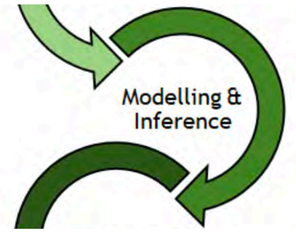
Draft Goal/ Target	Headline indicator	Component indicator	Complementary indicator
A	A.1 Red List of Ecosystems A.2 Extent of natural ecosystems A.3 Red List Index A.5 The proportion of populations within species with an effective population size > 500	Ecosystem Intactness Index Ecosystem Integrity Index Species habitat Index Biodiversity Habitat Index Protected Connected (Protconn) index Parc connectedness EDGE Living Planet Index Change in the extent of water-related ecosystems over time	Forest area as a proportion of total land area Forest distribution Tree cover loss Grassland and savannah extent Mountain Green Cover Index Peatland extent and condition Permafrost thickness, depth and extent Continuous Global Mangrove Forest Cover Trends in mangrove forest fragmentation Trends in mangrove extent Live coral cover
B[b]	B.1 Services provided by ecosystems*	Red List Index (for utilized species) Living Planet Index (for used species)	Levels of poverty in biodiversity depended communities Ecological Footprint
3	3.1 Coverage of protected areas and OECMs	Protected area coverage of key biodiversity areas Protected Area Management Effectiveness (PAME) ProtConn	Protected area downgrading, downsizing and degazettement (PD) Status of key biodiversity areas IUCN Green List of Protected and Conserved Areas
Draft Goal/ Target	Hea		
		Protected Area Connectedness Index (PARC-Connectedness) Red List of Ecosystems Connectivity Indicator (in development) The number of protected areas that have completed a site-level assessment of governance and equity (SAGE) Species Protection Index	Number of hectares of UNESCO designated sites (natural and mixed World Heritage sites and Biosphere Reserves) Protected area and OECM management effectiveness (MEPCA) indicator Protected Area Isolation Index (PAI) Protected Areas Network metric (ProNet) Extent to which protected areas and other effective area-based conservation measures (OECMs) cover Key Biodiversity Areas that are important for migratory species Coverage of Protected areas and OECMS and traditional territories (by governance type)

https://www.cbd.int/doc/c/179e/aecb/592f67904bf07dca7d0971da/cop-15-l-26-en.pdf

Typical ML workflow

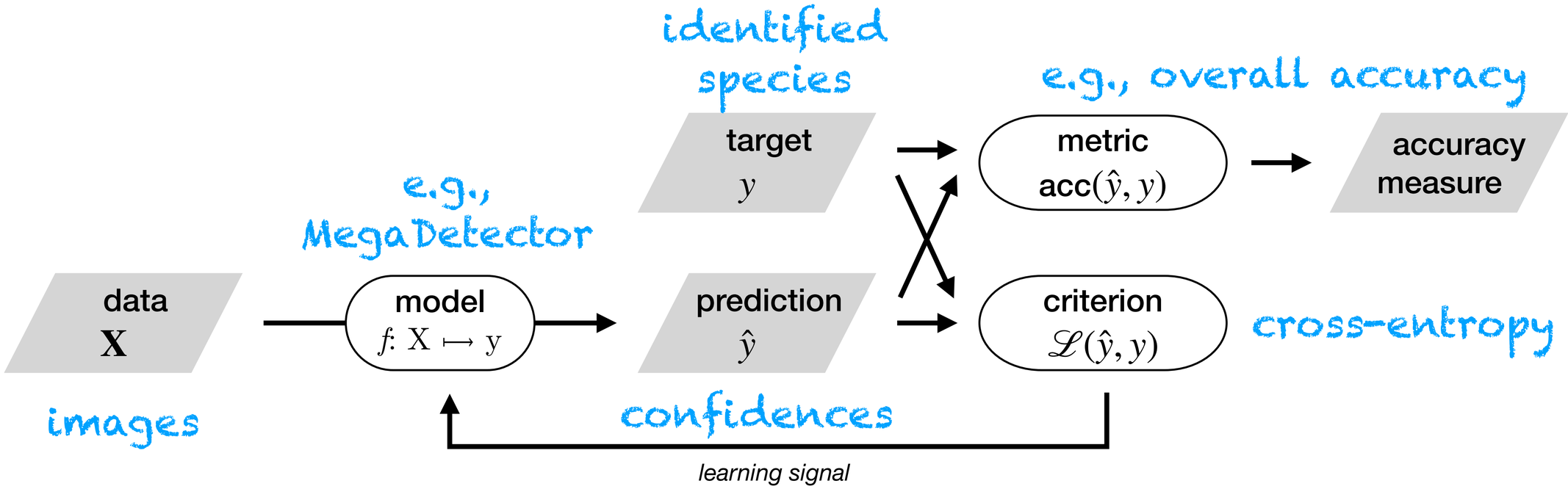






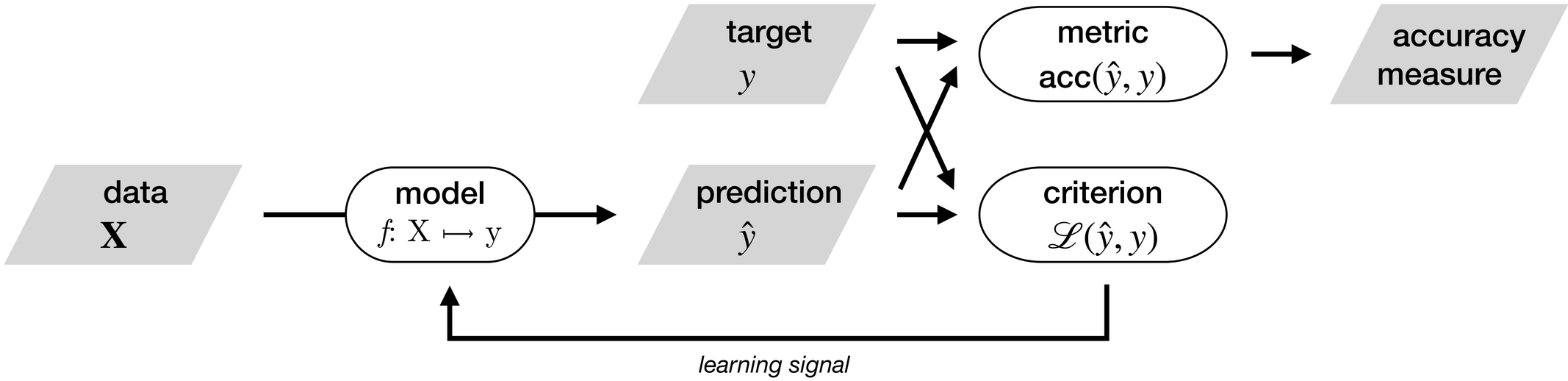
Typical ML workflow

Example: camera trap species ID



Typical ML workflow

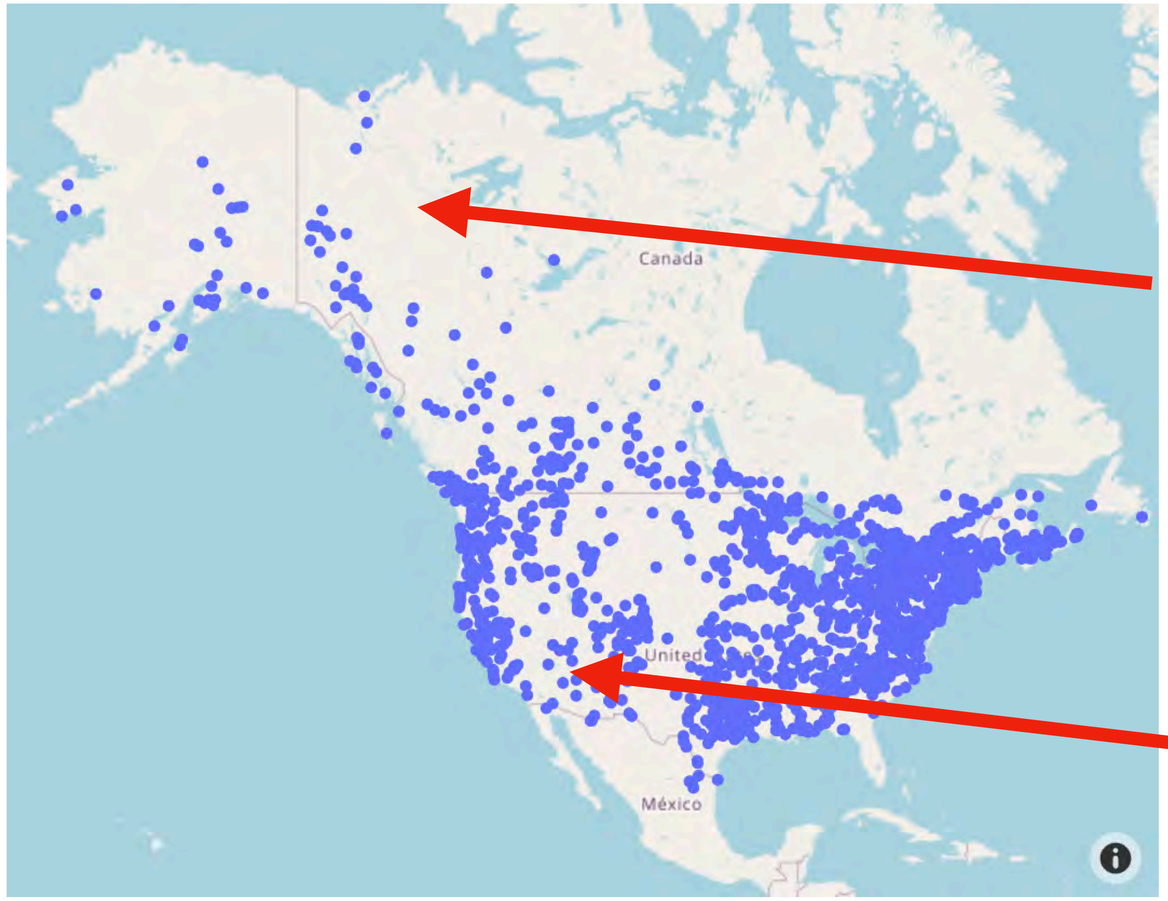
SDM?



SDM challenges

I. Observations

American beaver (*Castor canadensis*)



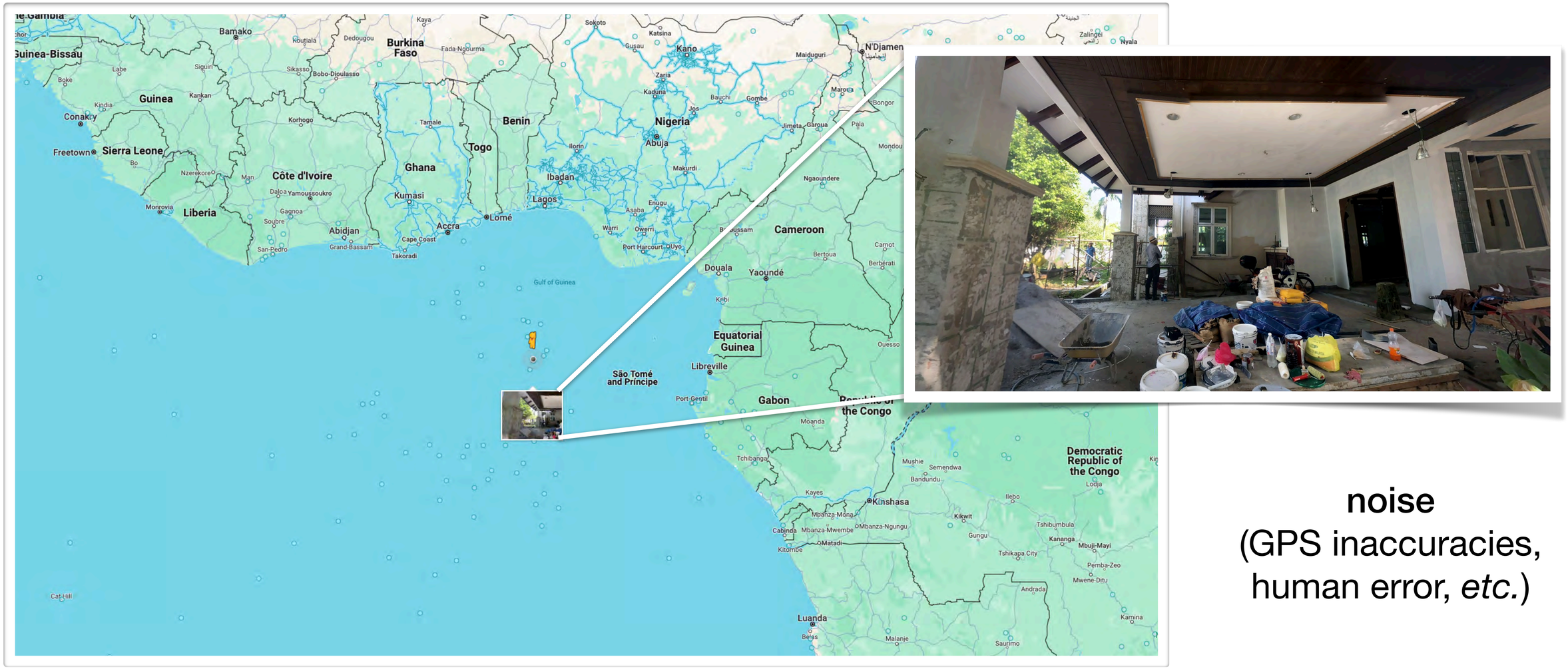
iNaturalist observations

expert range map

MOL MAP OF LIFE

SDM challenges

I. Observations

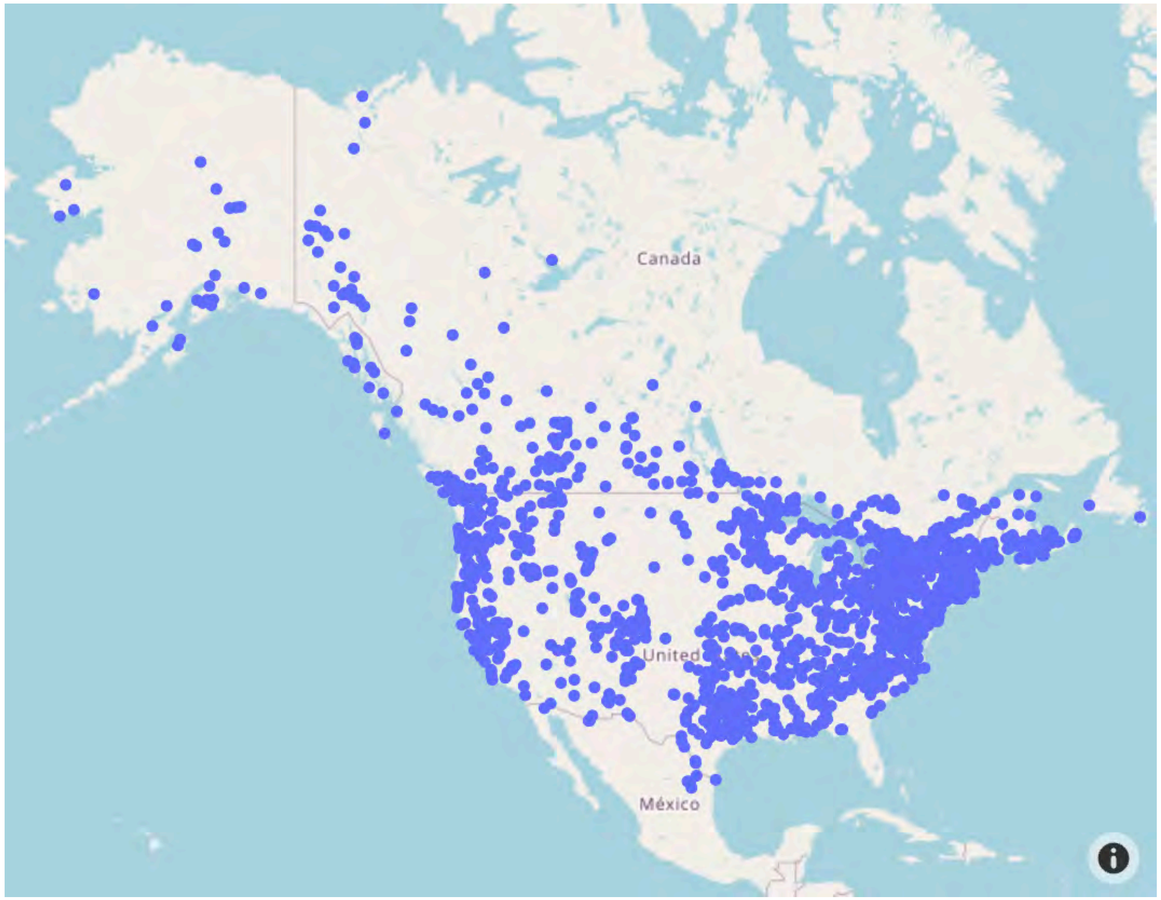


noise
(GPS inaccuracies, human error, *etc.*)

SDM challenges

I. Observations

American beaver (*Castor canadensis*)



iNaturalist

We "know" where the species is.

But where is it **not**?

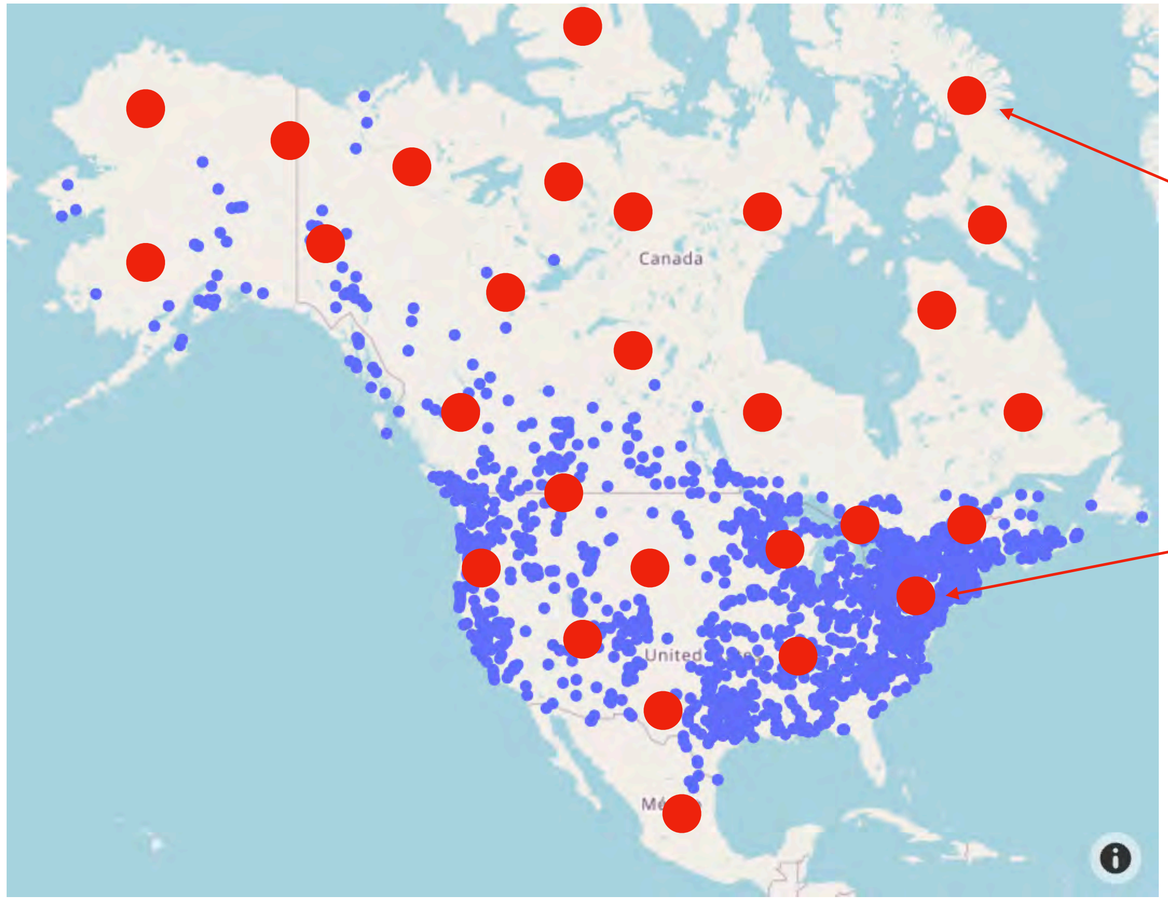
Species **absences** are hard to obtain.

→ pseudo-absences:

- Where?
- How many?
- How trustworthy?

SDM challenges

I. Pseudo-absence sampling strategies



Uniform random

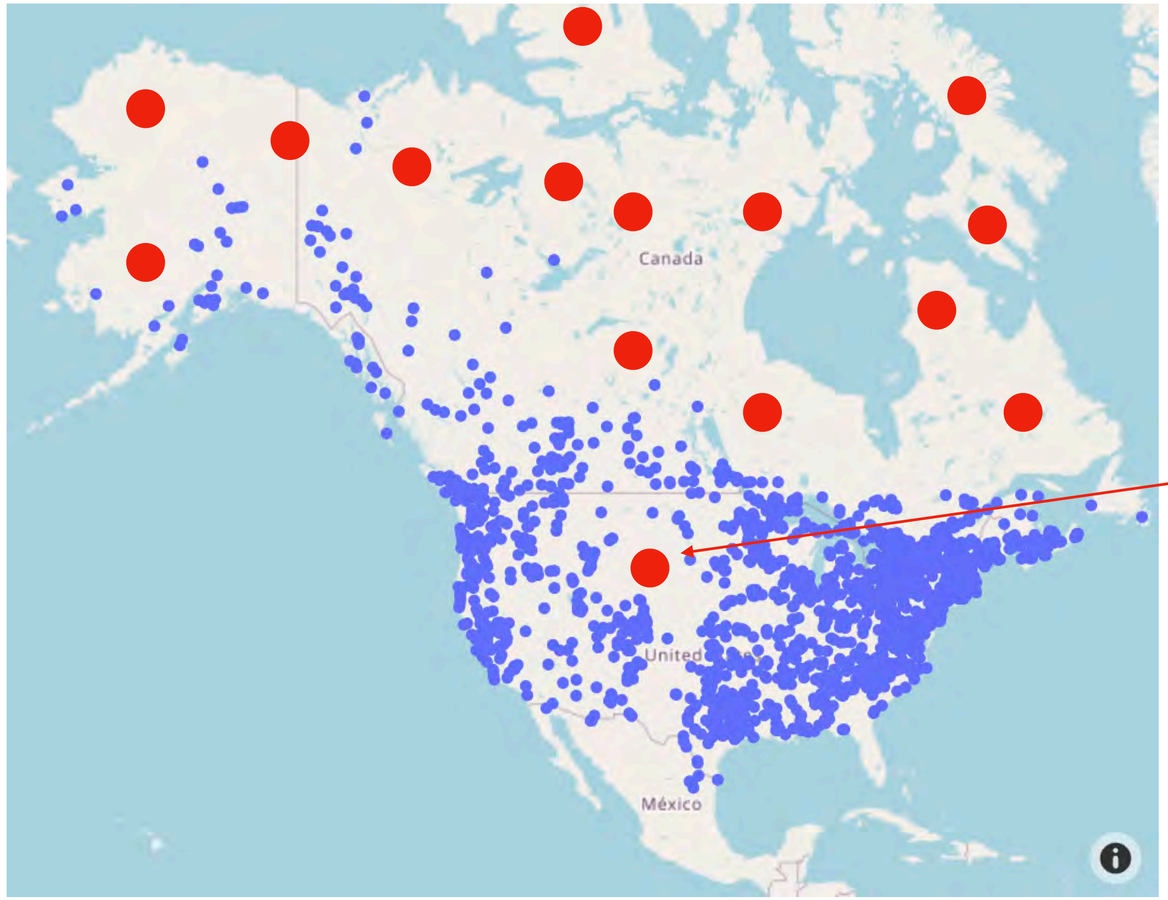
how useful are points like this one?

what does the model do in such a case?

iNaturalist

SDM challenges

I. Pseudo-absence sampling strategies



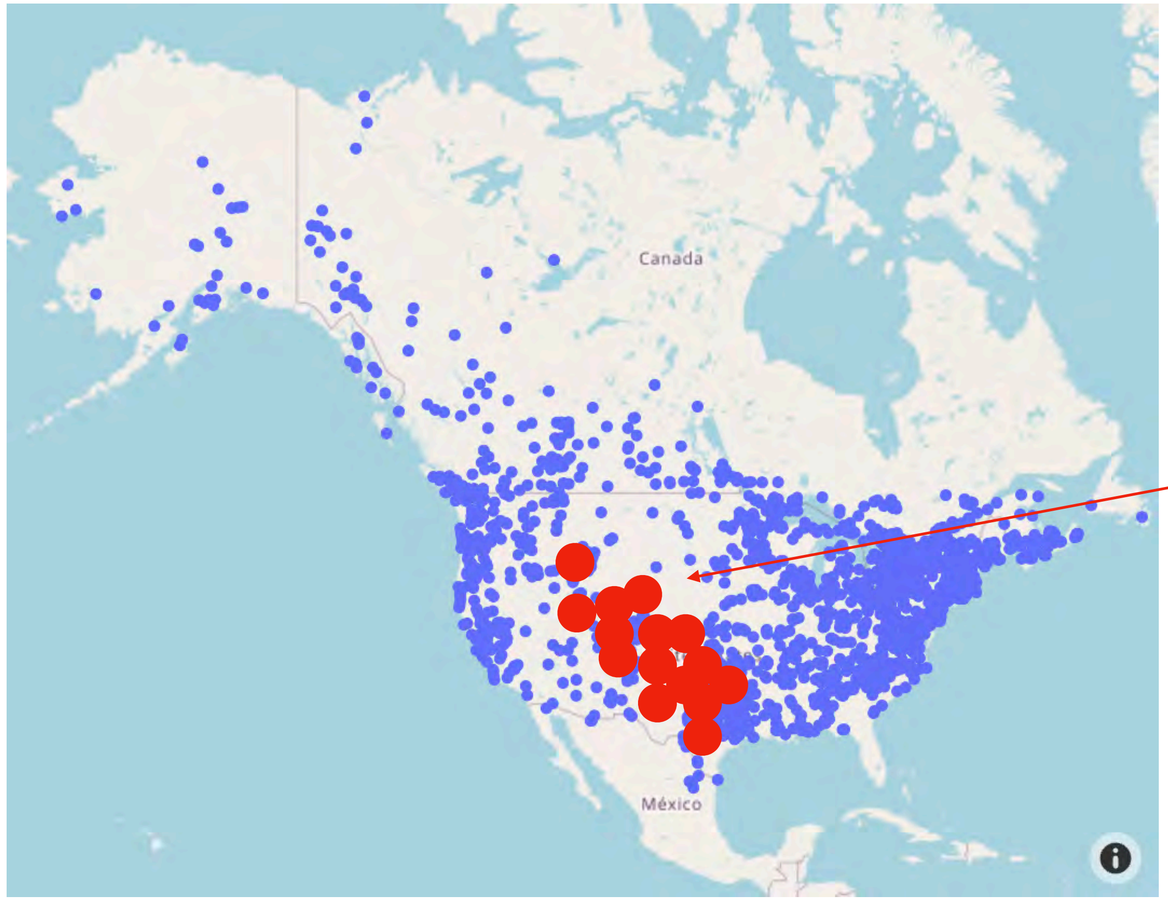
iNaturalist

Uniform random, outside "species domain"

how far outside is "far enough"?

SDM challenges

I. Pseudo-absence sampling strategies



iNaturalist

Target-group background

→ absences = presences of other species

Corrects sampling bias!
But: what if species habitat ranges don't overlap?

SDM challenges

II. Covariates

esa.int

SDM challenges

II. Covariates

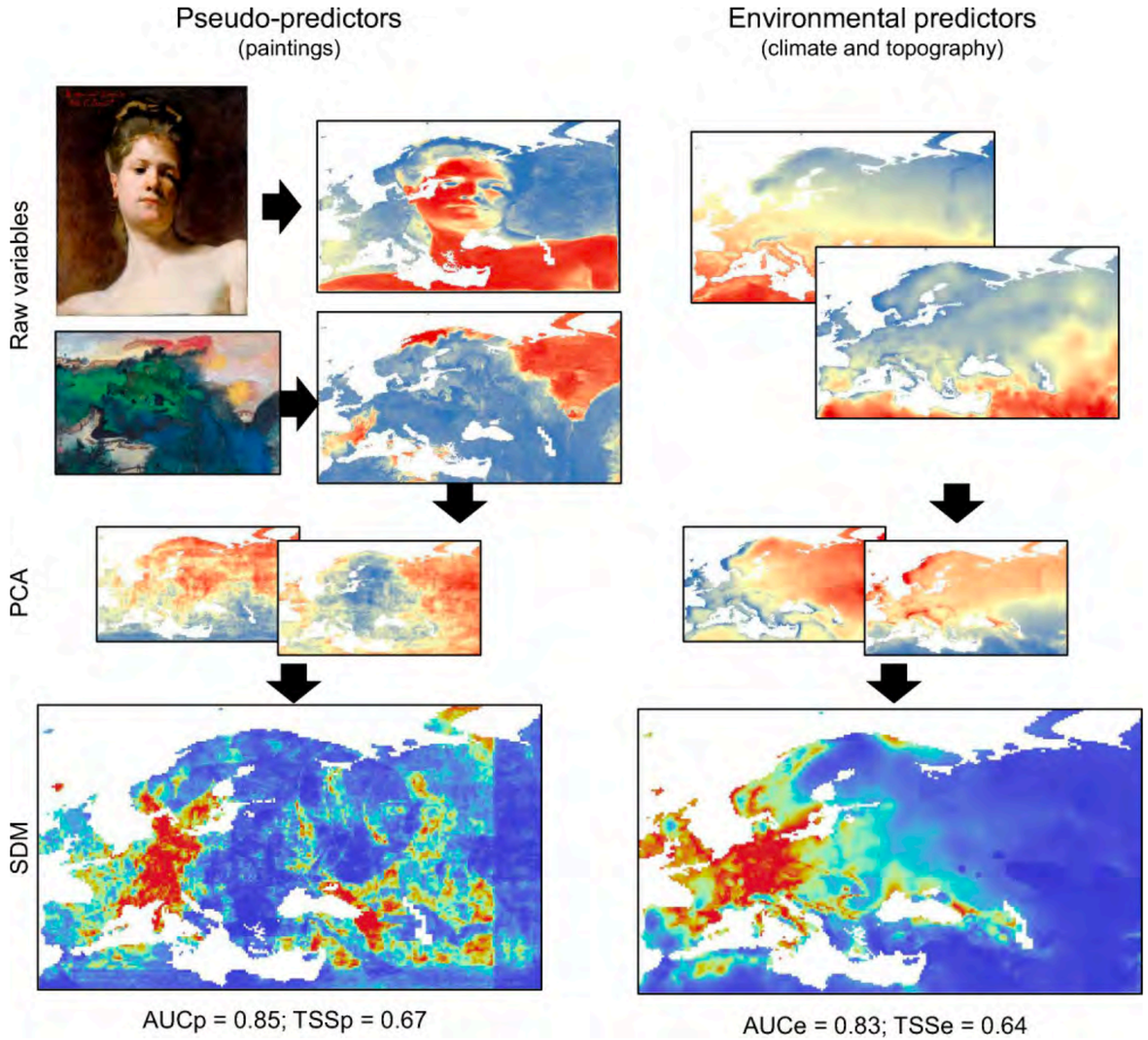
annual Enhanced Vegetation Index (EVI)

topographic wetness index

"mean monthly precipitation
of the wettest quarter in 1984-2016"

SDM challenges

II. Covariates

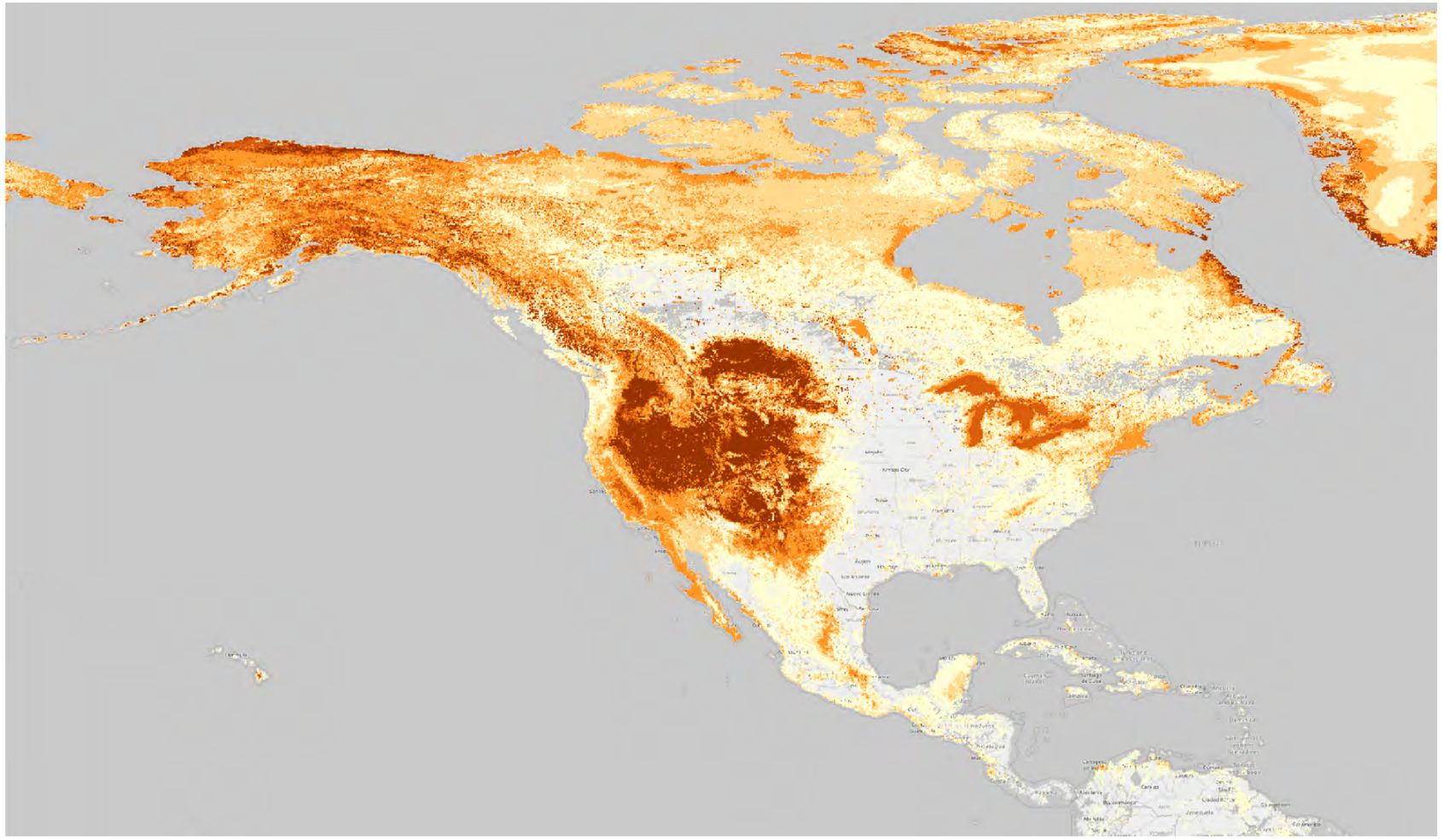


Fourcade et al., 2018

SDM challenges

III. Validation

Is this a good prediction map?



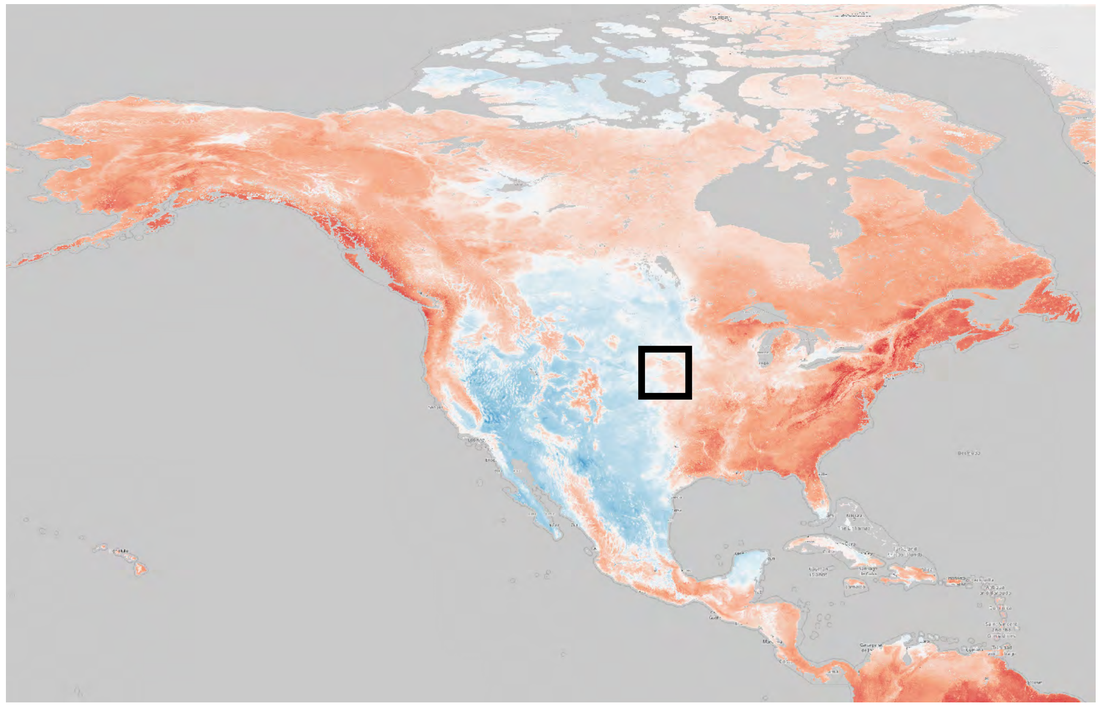
SDMs are ill-posed

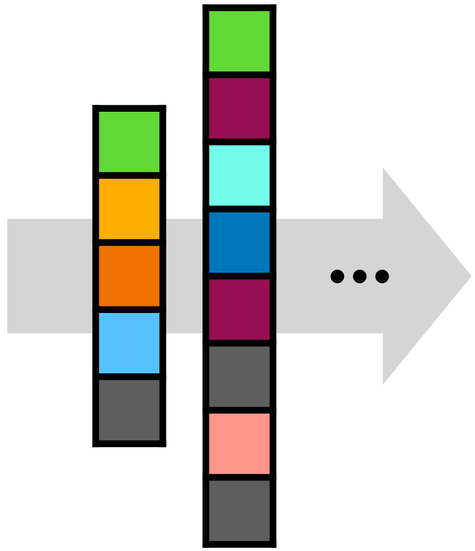
In SDMs, we:

- cannot trust our observations (biases, unconfirmed absences)
- never fully know which covariates to choose
- work with limited models
- have no "objective" evaluation criterion
- have a slew of dangerous pitfalls (data splitting, imbalances, *etc.*)
- deliberately ignore important factors (*e.g.*, biotic interactions)

So why use deep learning?

DL models are *universal function approximators*





Lepus americanus 0.98

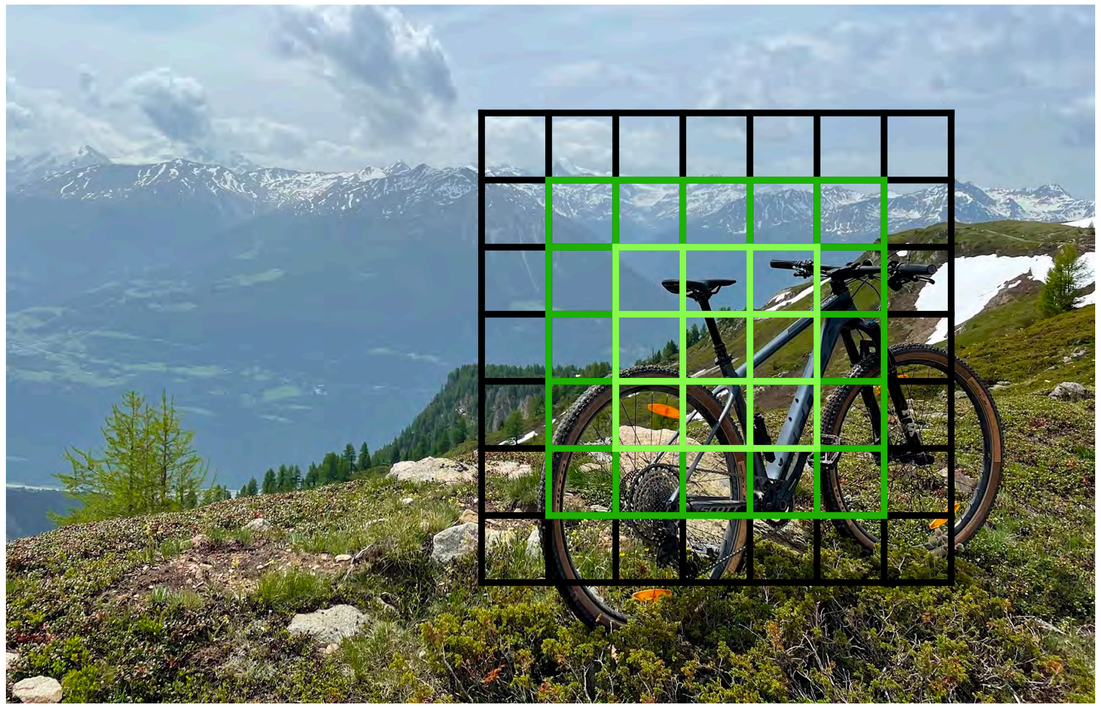
Ovis canadensis 0.04

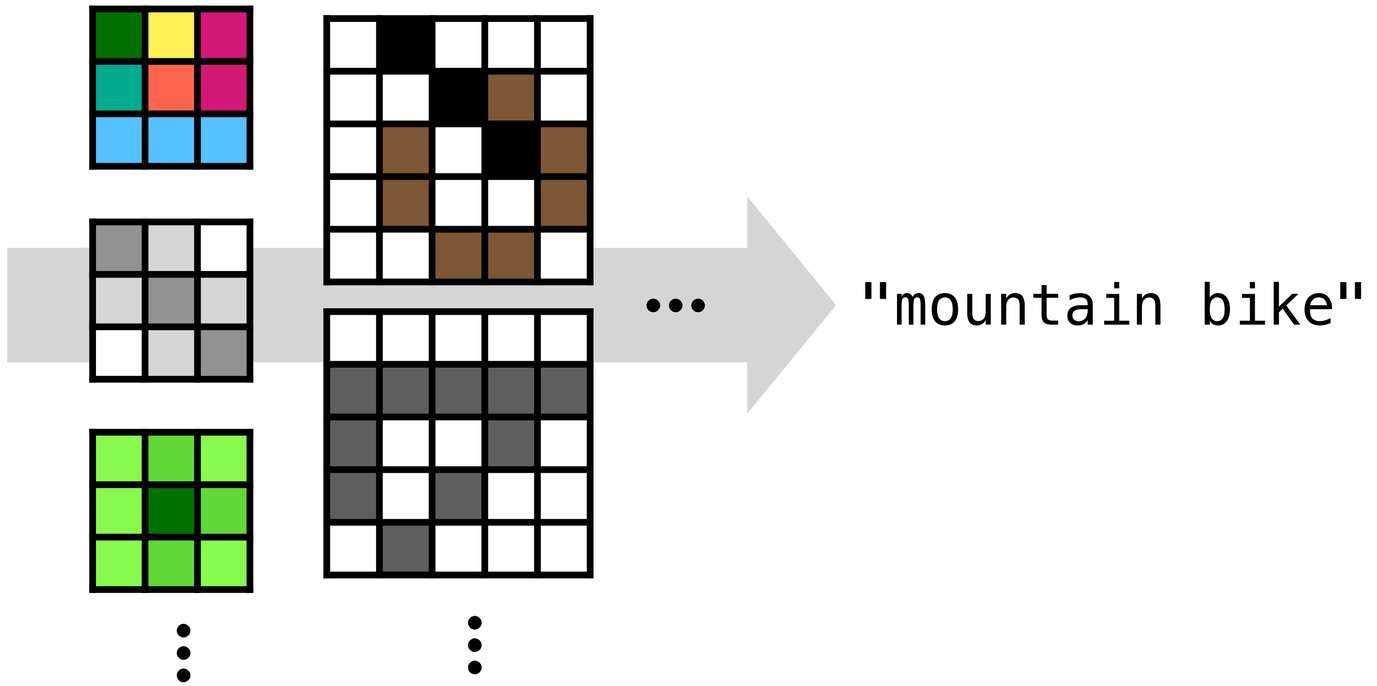
Tyto furcata 0.45

...

So why use deep learning?

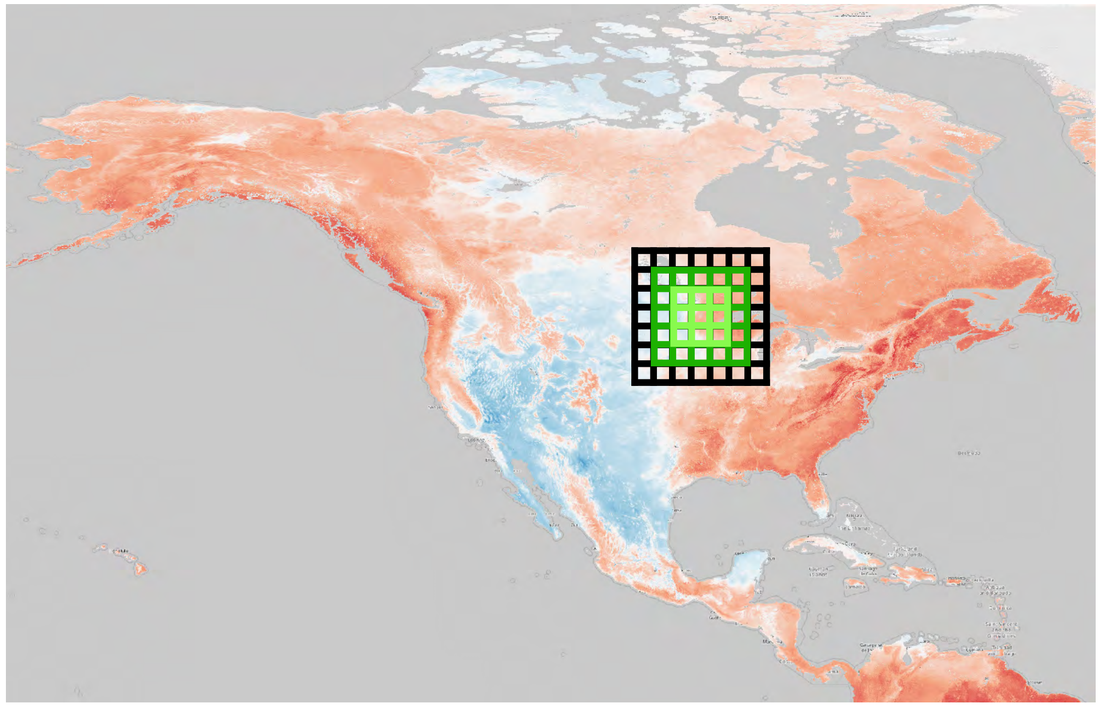
DL can *include spatial context*

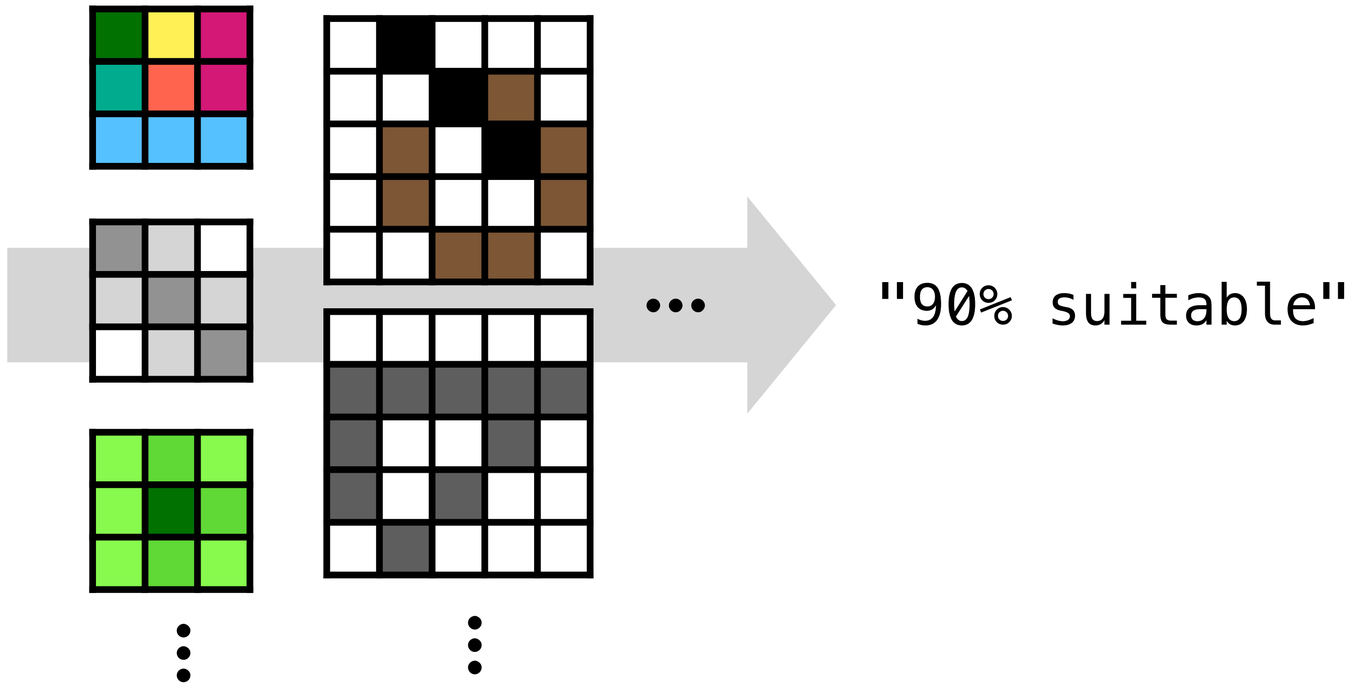




So why use deep learning?

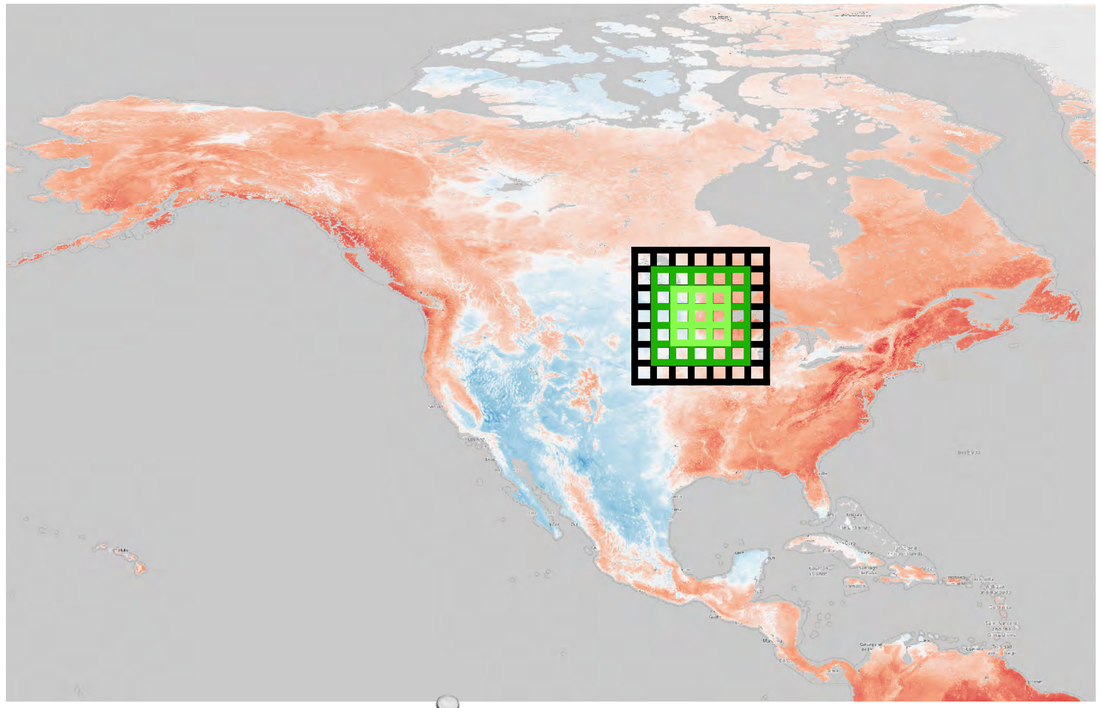
DL can *include spatial context*





So why use deep learning?

DL is easily *multi-modal*

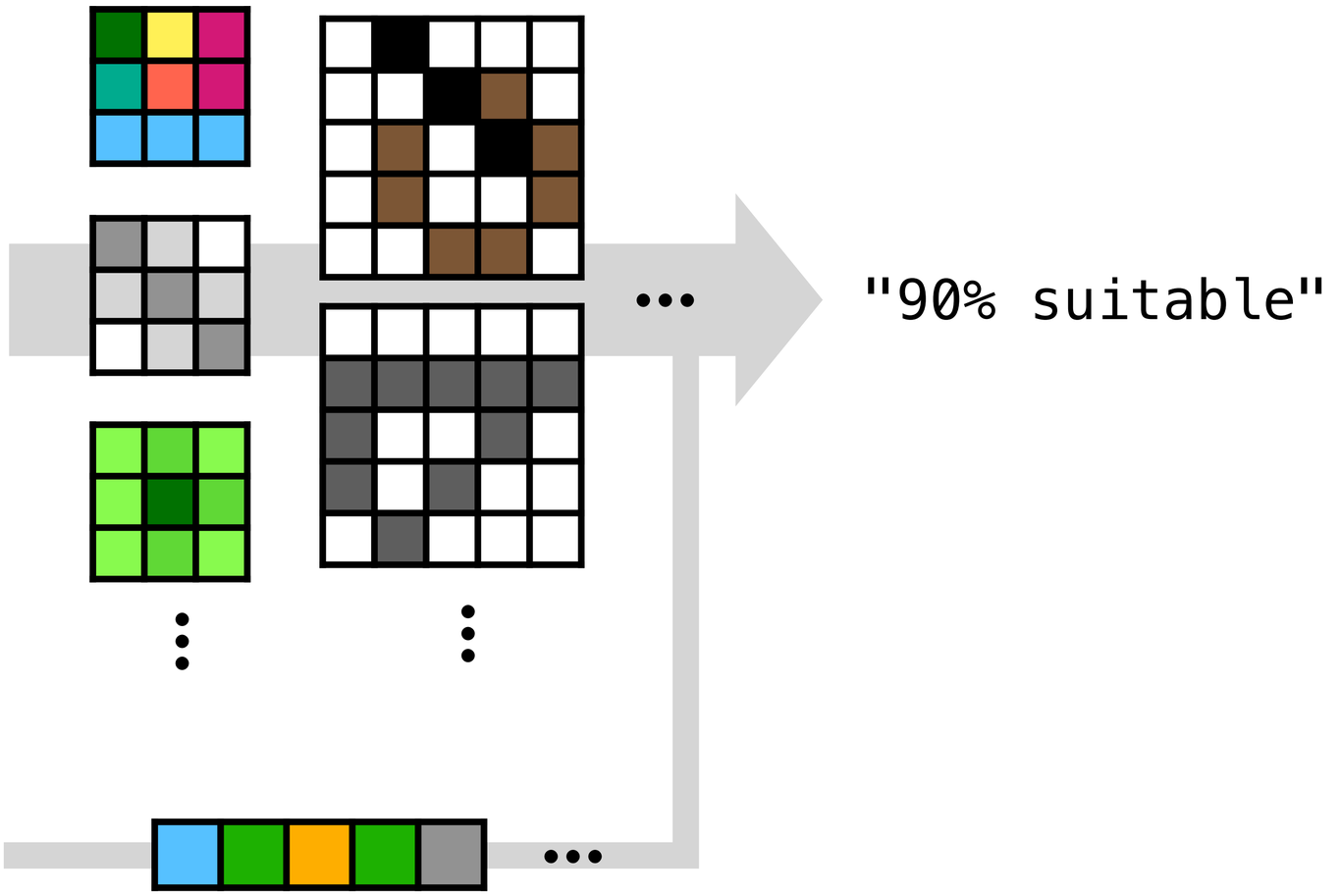


Predators include *Vulpes macrotis* (Egoscue, 1956), *Canis latrans* (Johnson and Hansen, 1979a, 1979b), *Tyto* and Long, 1961;

Resting meta
creases by Januar
ering thermogenes
Photoperiod and c

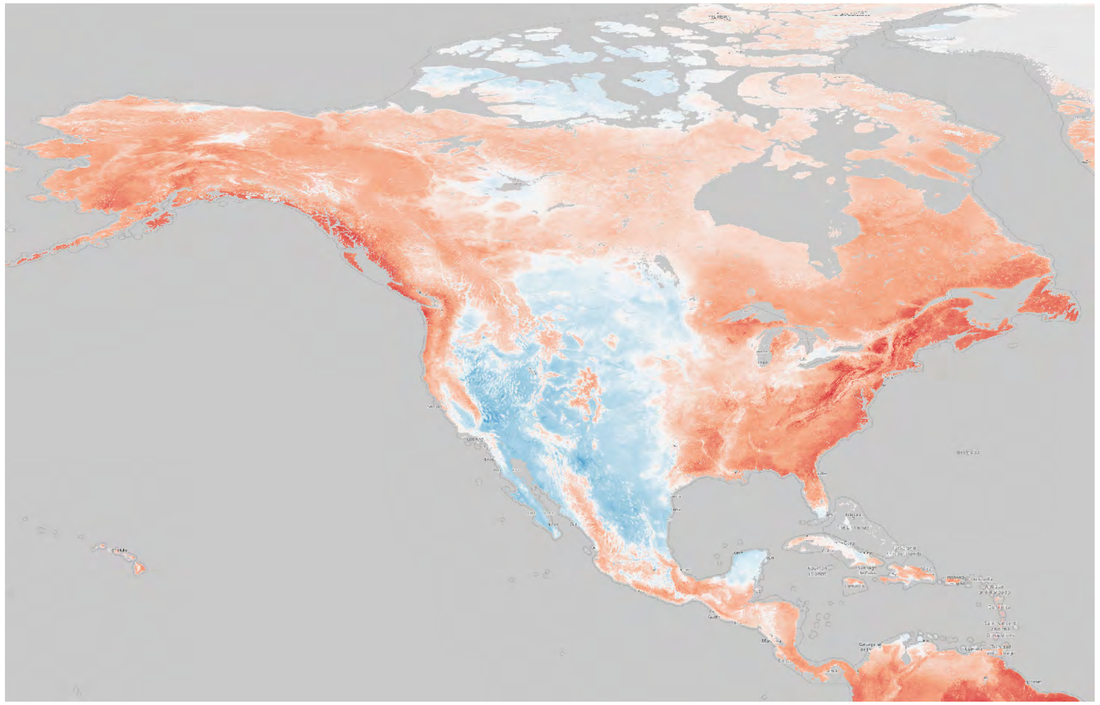
ECOLOGY. *Dipodomys ordii* lives in a diversity of habitats. In Idaho, it is found in several vegetative communities (Groves and Keller, 1983), but is most abundant in the *Juniperus* community with *Chrysothamnus* and *Eurotia* as the understory (Allred, 1973).

resting rates of metabolism and nonshivering thermogenesis, which vary with season. Shifts in thermogenesis do not appear to result

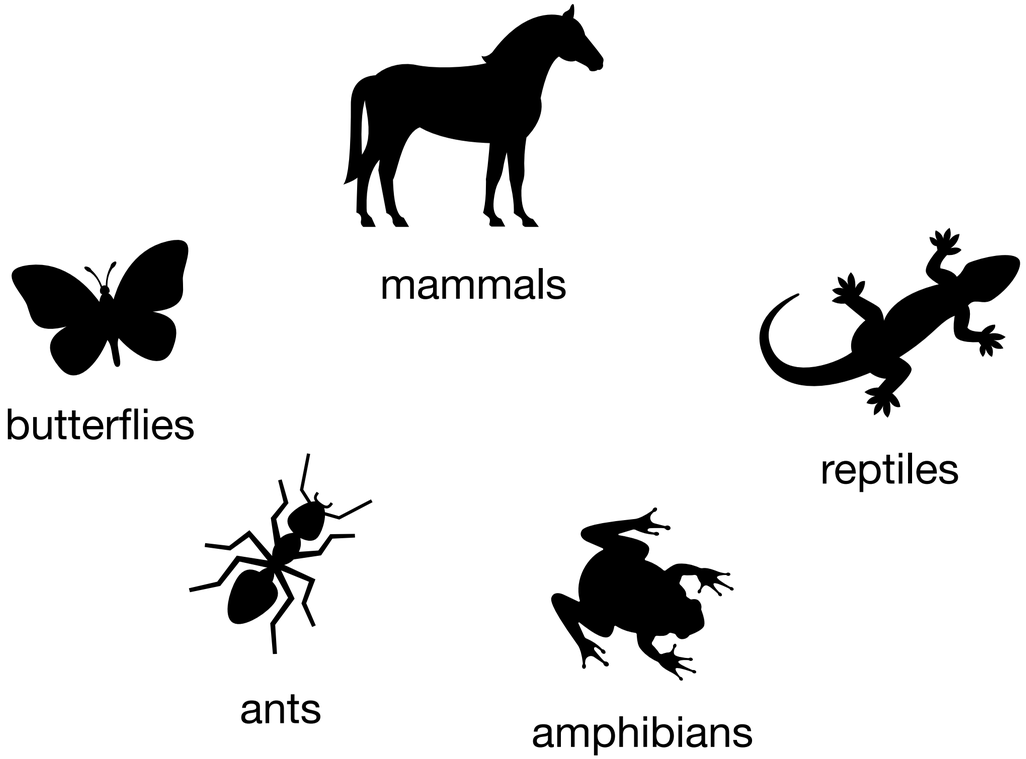


Experimental setup

Data



17 covariates, 1km^2 resolution

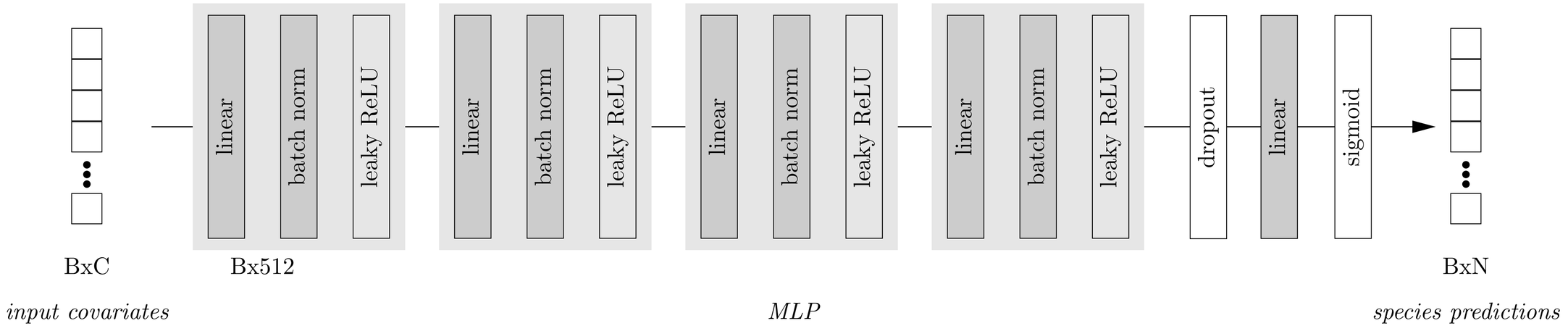


5 taxonomic groups
2,200 species
1.8M observations
5.1M data points (w/ pseudo-absences)

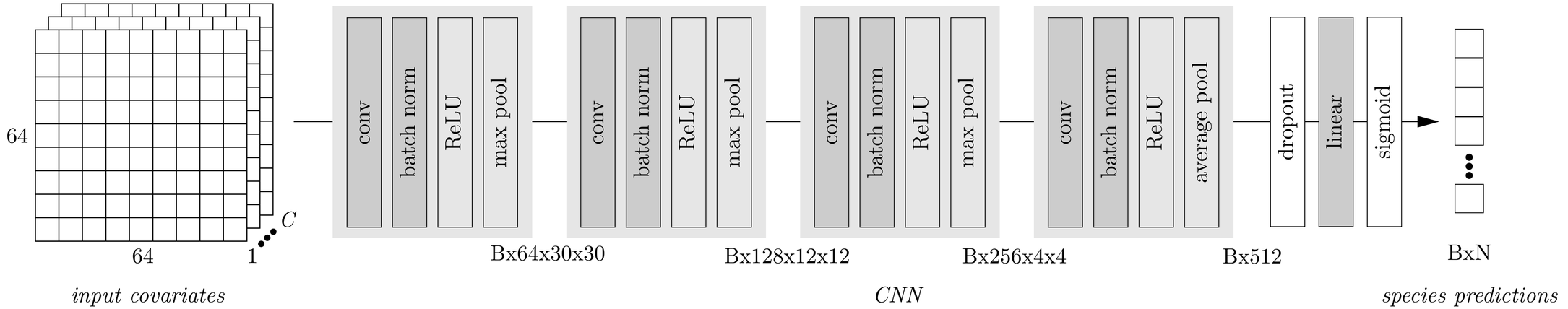
Experimental setup

Models

MLP
(point-based)



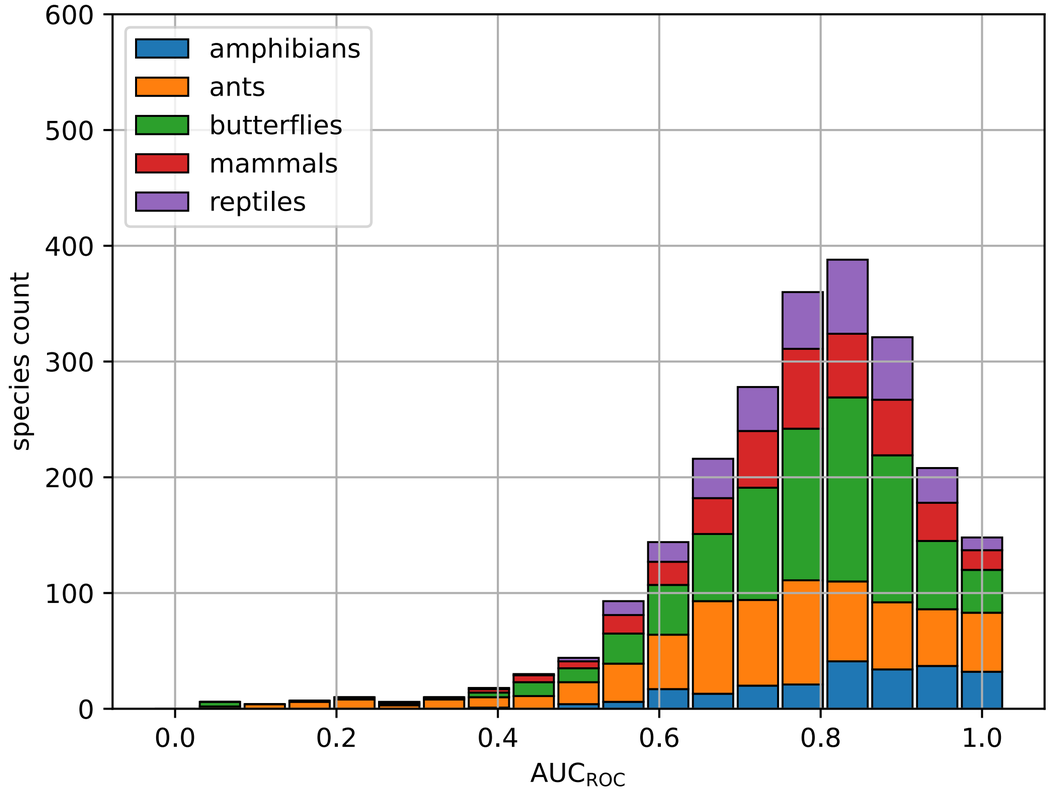
CNN
(spatial; 64x64)



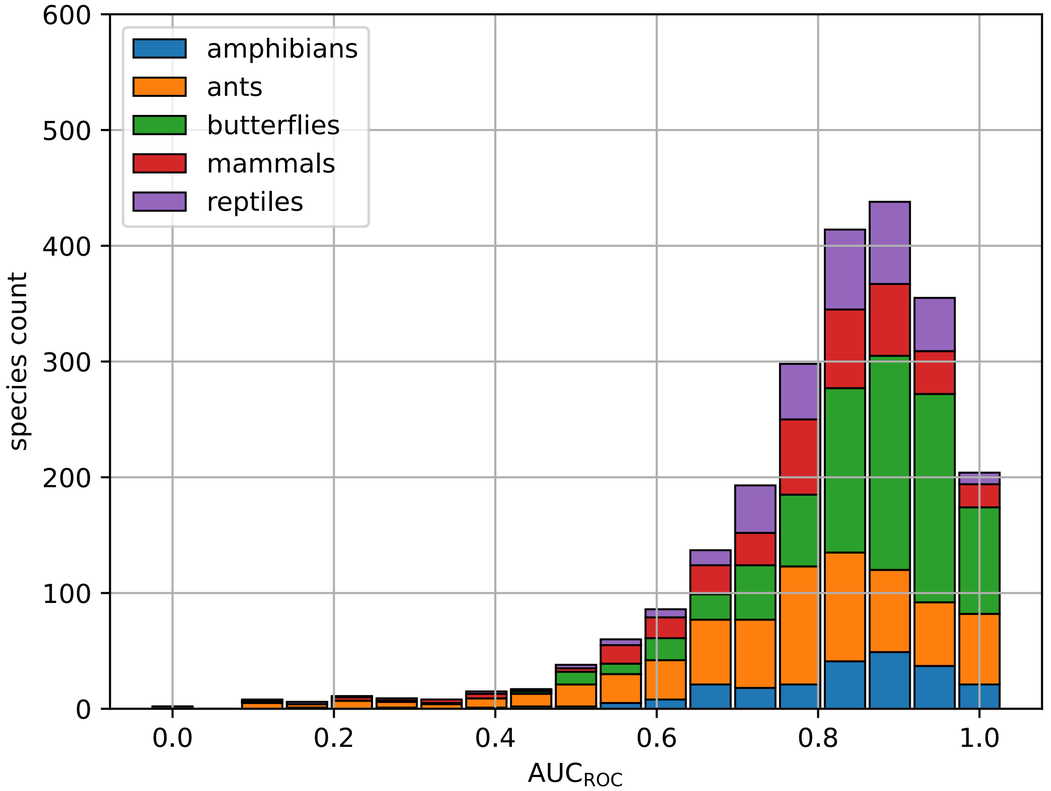
Results

on test set

Random Forest



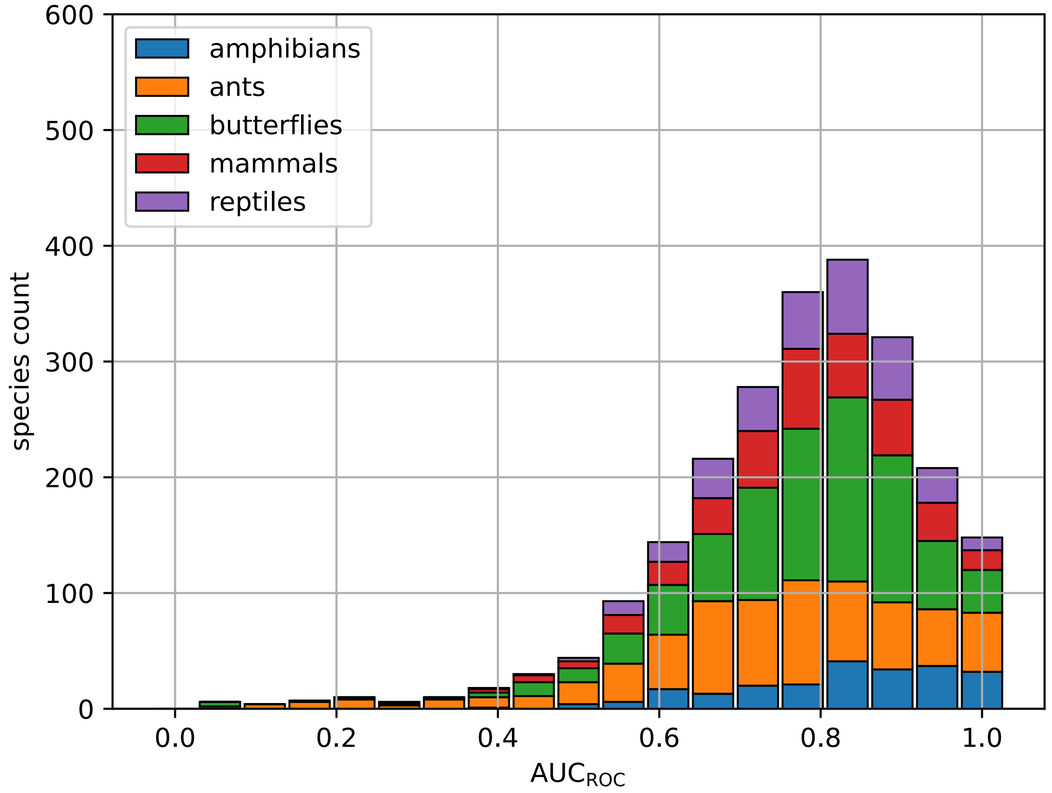
DL: point-based



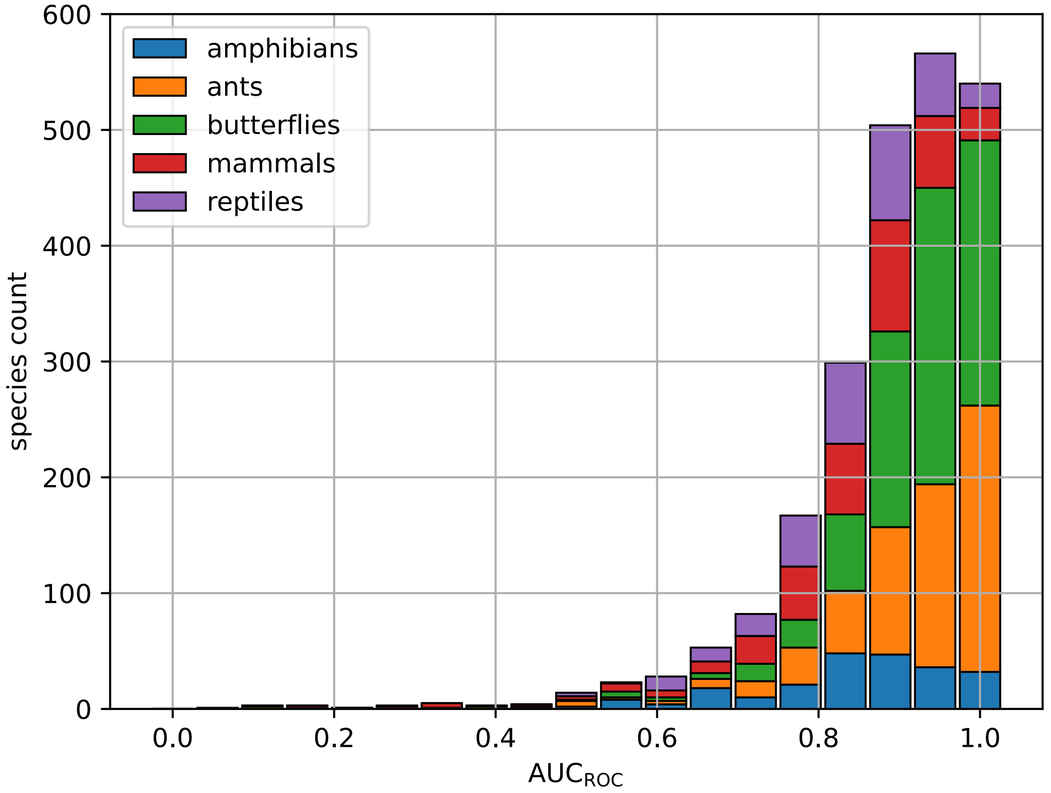
Results

on test set

Random Forest

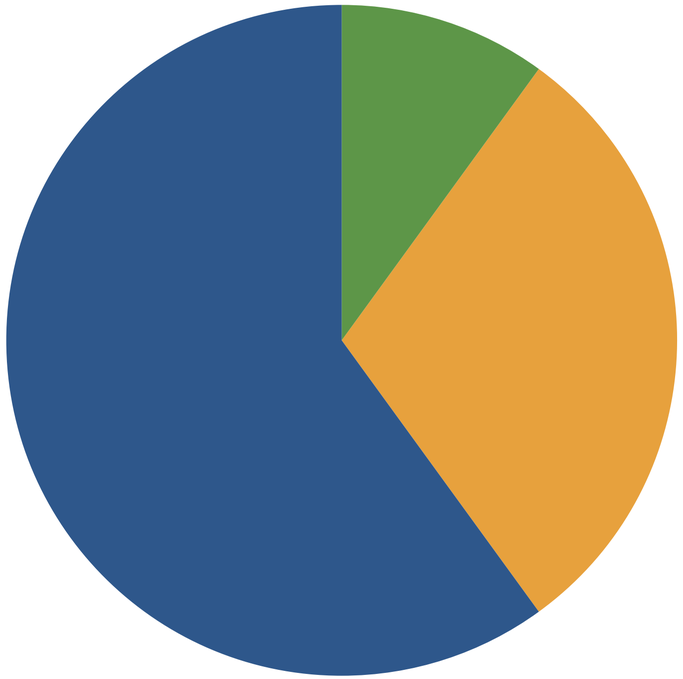


DL: spatial



Recap: data splitting

cf. week 2



- training set

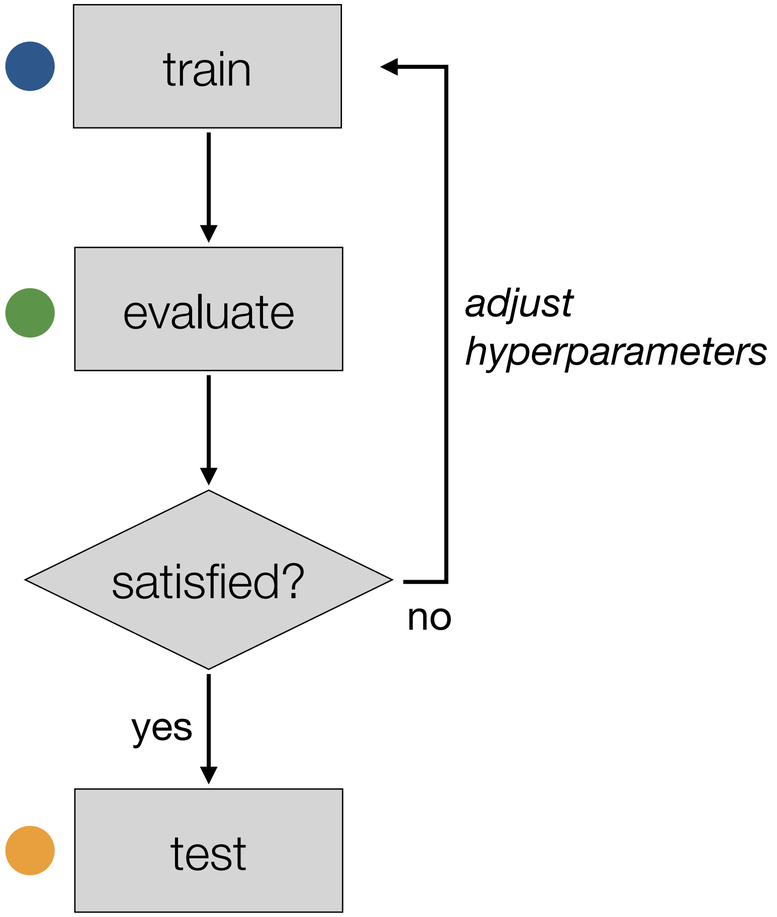
  fit model
  *(set **parameters**)*

- validation set

  evaluate model
  *(set **hyper**parameters)*

- test set

  *test **final** model*



Recap: data splitting

What we originally used:

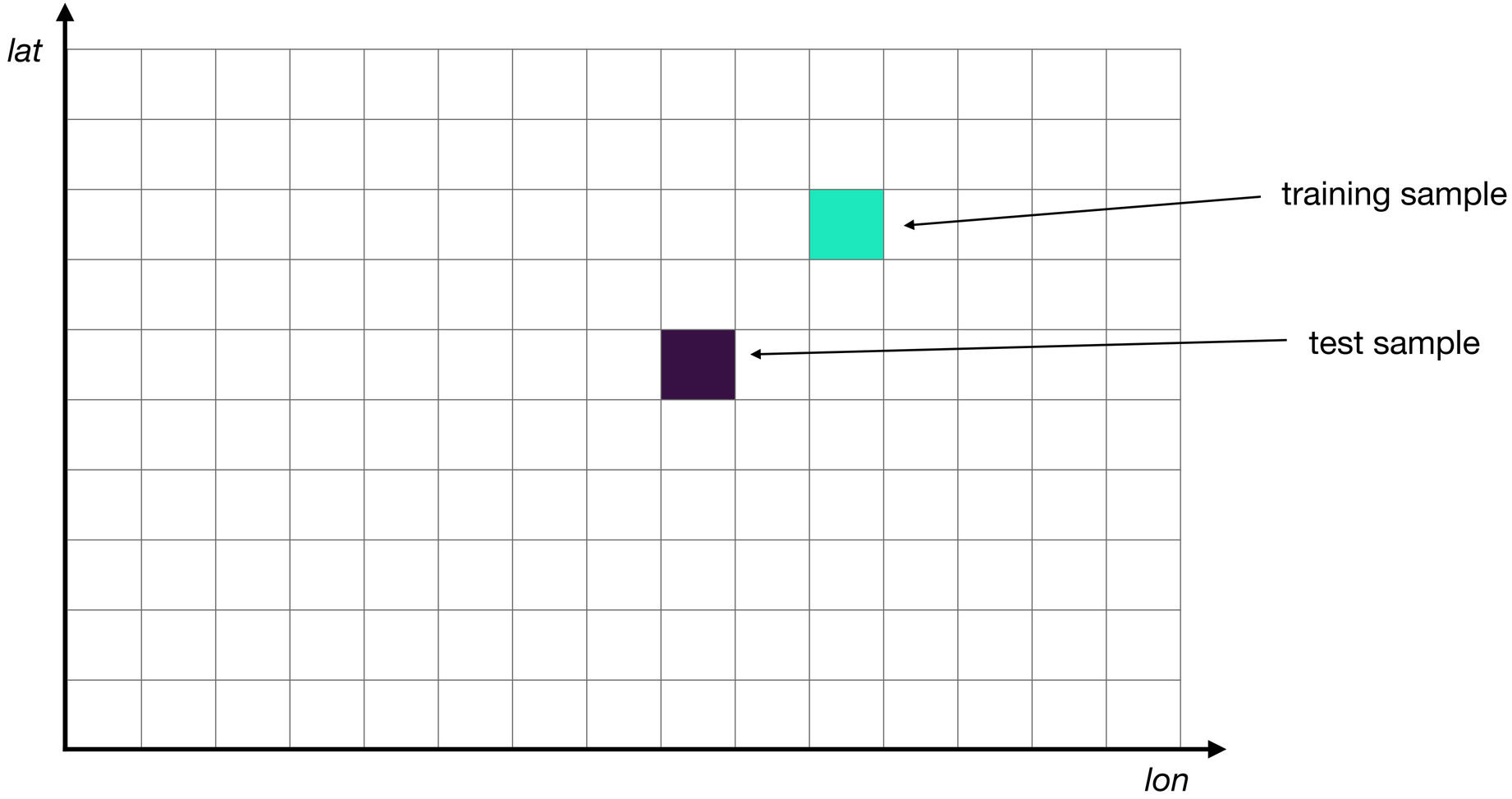
```
for all species; do
    cluster observations
    assign clusters to train, val, test

    randomly assign pseudo-absences to train, val, test
done
```

overlap between species

Recap: data splitting

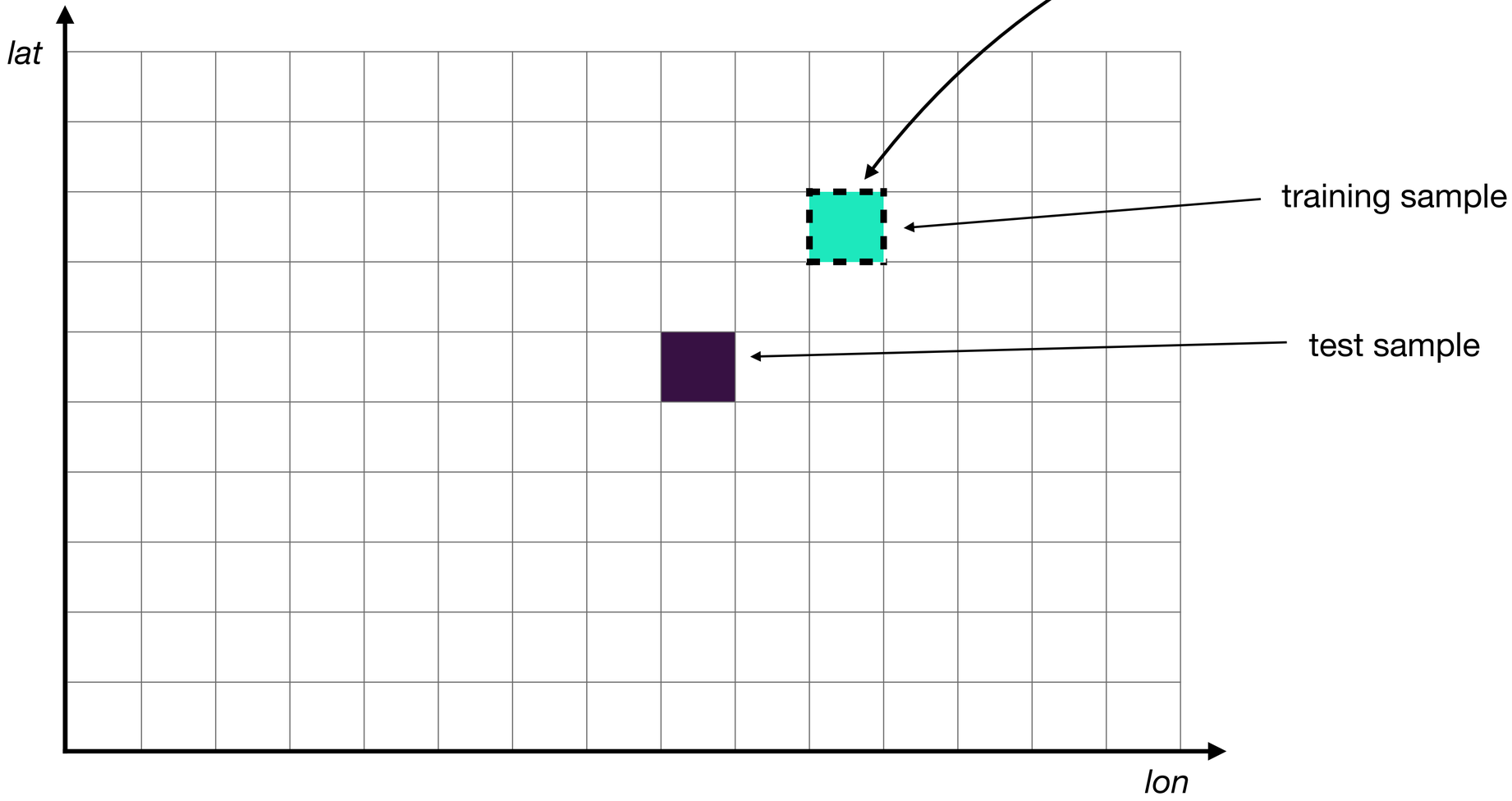
about those spatial models...



Recap: data splitting

about those spatial models...

Point model (Maxent, MLP):
- spatial autocorrelation: small bias
- otherwise "ok"

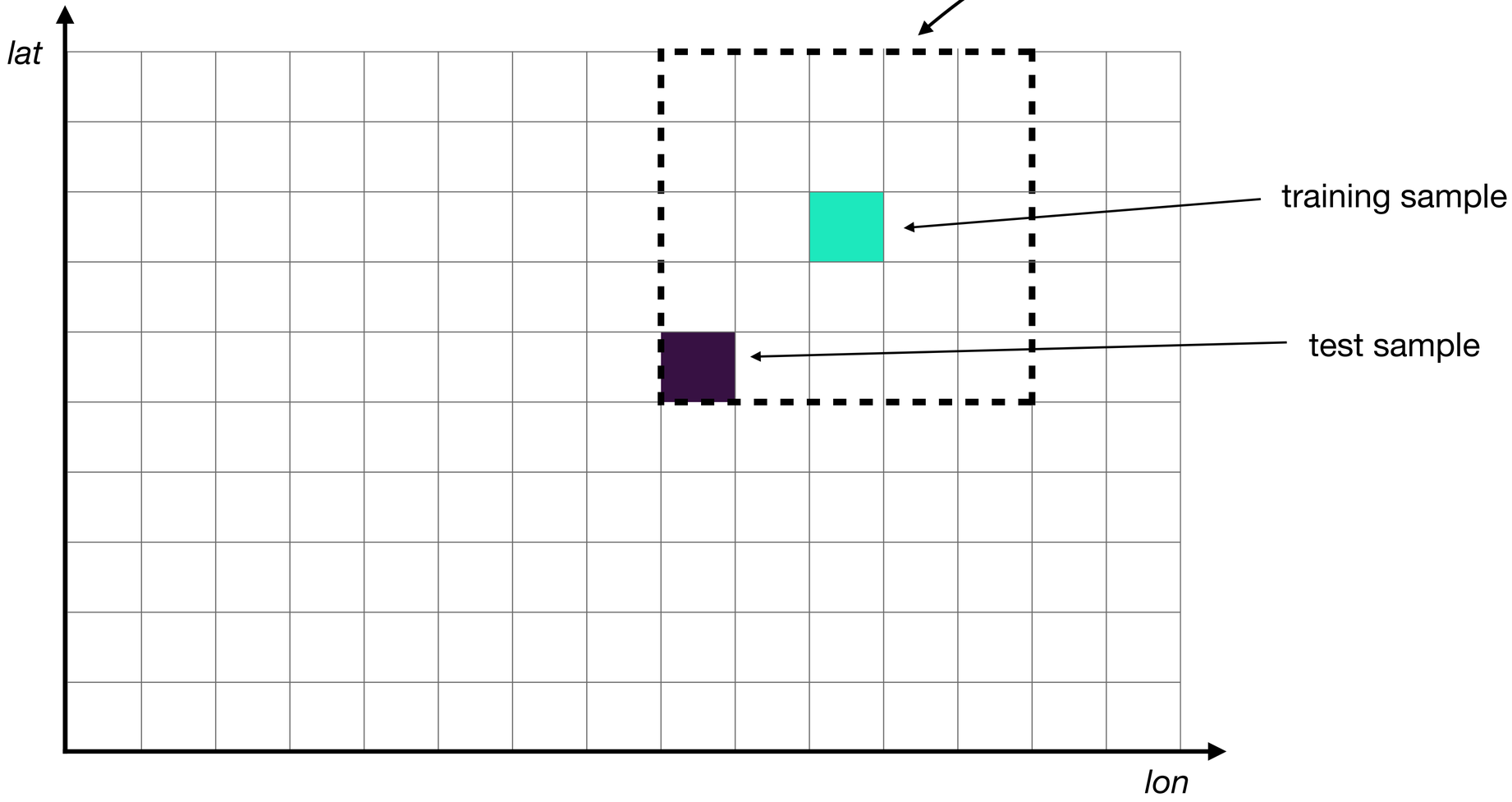


Recap: data splitting

about those spatial models...

Spatial model (CNN):

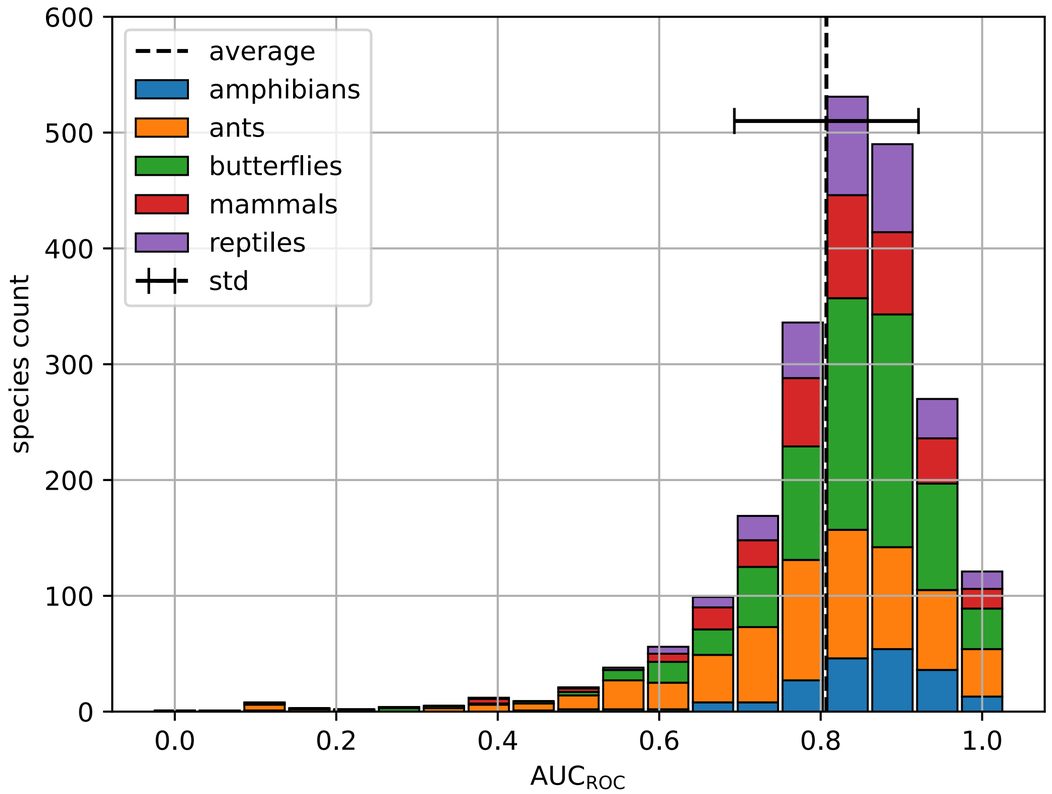
- sees **actual test sample** during training



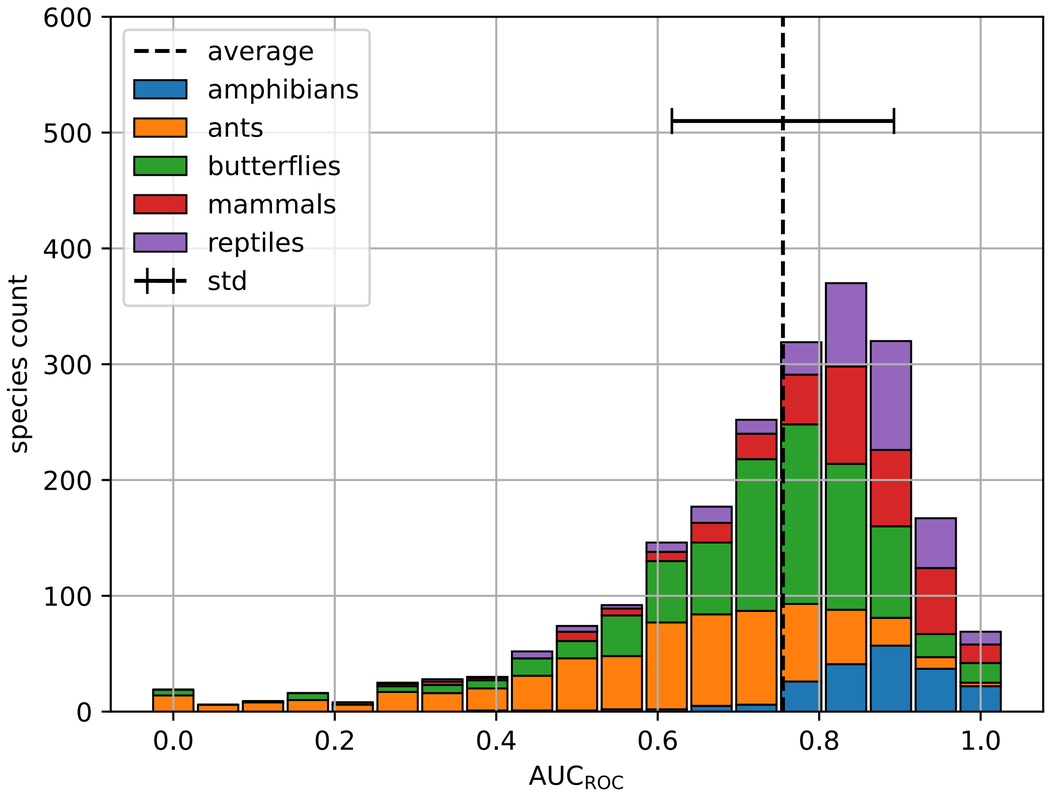
Results @ improved split

on test set

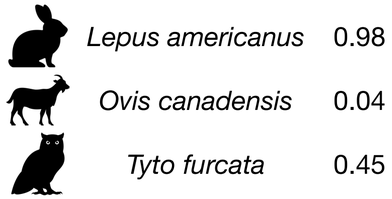
Random Forest



DL: spatial



Why it may still make sense



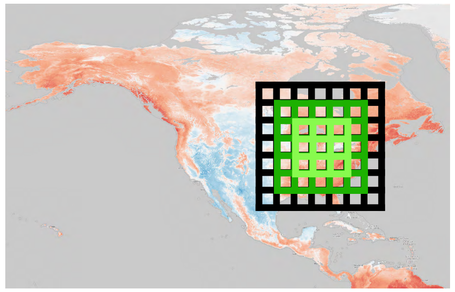
Lepus americanus 0.98

Ovis canadensis 0.04

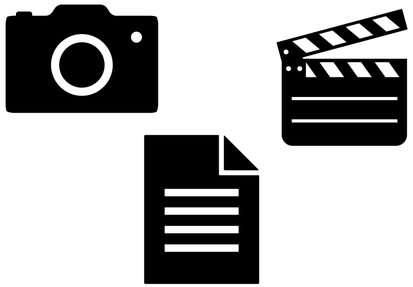
Tyto furcata 0.45

...

inherently **multi-species**



include **spatial context**



support for **multimodal inputs**

Summary

- Species Distribution Modelling (SDM):
  - riddled with challenges
  - hence: lots of leeway in interpretation
    but: statistical assumptions must be met!

- Deep Learning:
  - has mostly taken over in **prediction** tasks
  - is now picking up pace for **inference**
  - has a long, long way to go in ecology

Literature

Fourcade, Y., Besnard, A.G. and Secondi, J., 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. Global Ecology and Biogeography, 27(2), pp.245-256.

Tuia, D.*, Kellenberger, B.*, Beery, S.*, Costelloe, BR*, Zuffi, S., Risse, B., Mathis, A., Mathis, MW, van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, ID, van Horn, G., Crofoot, MC, Stewart, CV, Berger-Wolf, T., 2022. Perspectives in Machine Learning for Wildlife Conservation. Nature Communications.

Winner, K., Ingenloff, K., Sandall, E., Sica, YV, Marsh, C., Cohen, J., Ranipeta, A., Killion, A., Jetz, W., *in preparation.* High Resolution Species Distribution Models of North American Biodiversity.

Zbinden, R., van Tiel, N., Kellenberger, B., Hughes, L., Tuia, D., *in revision*. On the selection and effectiveness of pseudo-absences for species distribution modeling with deep learning. Ecological Informatics.

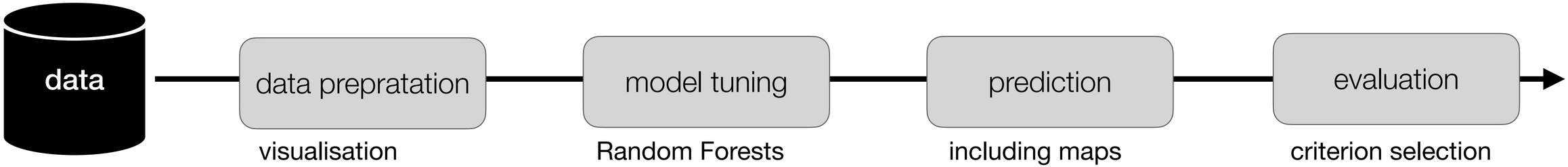
Exercise 10

observations

3 species over North America
observations from iNaturalist

environmental covariates

8 covariates @ 1km^2 resolution



visualisation

different
pseudo-absences

train/val/test split

Random Forests

selection of:
- evaluation metric
- hyperparameters

bonus: a bit of
Deep Learning

including maps

criterion selection

reflection:
do results make sense?

Exercise 10

- Most of the code is provided
- There's some questions to answer

Main consideration:

"*Would I publish these results in a paper?*"
"*Can I stand behind them?*"