The Machine Learning ⇌ Science Colaboratory

Seth Axen, Alexandra Gessner, Hanqi Zhou, Elena Sizana, and Álvaro Tejero-Cantero
Excellence Cluster Machine Learning for Science, University of Tübingen

Cooperation: Integration of Paleoclimatic Models and Proxies (PollenClim) 2

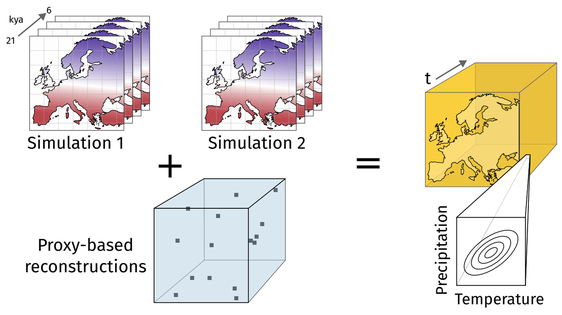
Spatiotemporal interpolation

Evidence about the climate of the past:

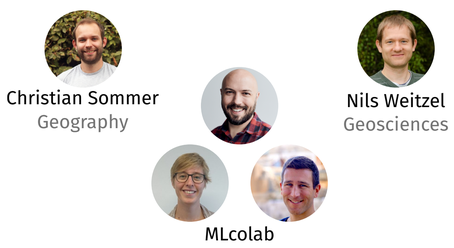
simulations. physical models, at unaligned grids in space/time

proxy data. reconstructions from e.g. fossilized pollen

ℹ There is no consensus model of grid and point evidence with uncertainty



Interdisciplinary team



Christian Sommer
Geography

Nils Weitzel
Geosciences

MLcolab

Scope

model. global temperature and precipitation between 21kya and 6kya (LGM to MH)

data. ~300k pts pollen proxies and ~18m pts from simulations (PMIP4, CCSM4, MIROC-ESM, MPI-ESM-P, CESM, HadCM3, and HadAM3H)

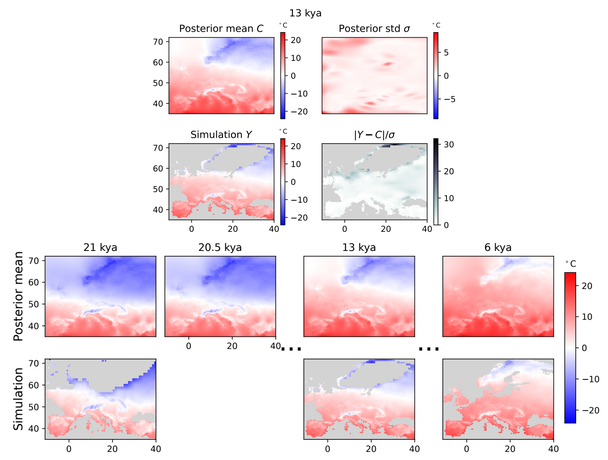
Scalable Gaussian processes

- GPU-backed, spatially and temporally sparse, variational multi-output Gaussian process posterior (S²CVI) [1]
- `tensorflow` → `gpflow` → `markovflow`
- 100 spatial, 6 temporal inducing points
- validation with LOO-CV, h-block CV, and CRPS

$$m(x,t) = C(x,t_m) + \langle w, z(x)\rangle$$
$$C(x,t) \sim \mathcal{GP}(m(x,t), k_x(x,x')k_t(t,t'))$$
$$[Y_{s,p}] \sim \mathcal{N}([C_{s,p}], \Sigma)$$

$Y_{s,p}$: observation at (x_s, t_p)
$C(x,t)$: paleoclimate
$C(x,t_m)$: modern climate
$z(x)$: spatial features
$k_x(x,x')$: anisotropic Matérn 3/2 kernel
$k_t(t,t')$: Ornstein-Uhlenbeck kernel
Σ: sparse covariance matrix

Preliminary results

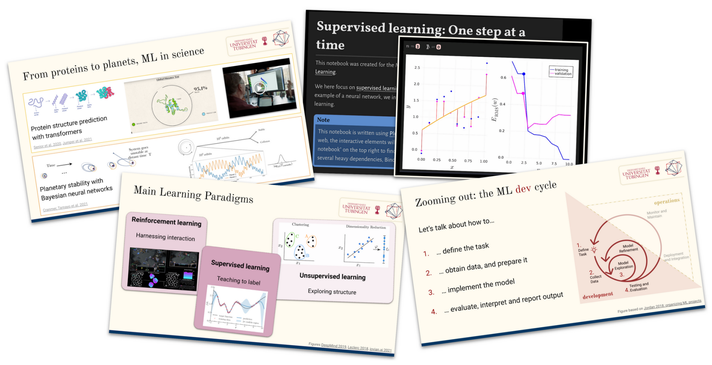


16 time slices take ~2h on a V100

[1] William Wilkinson, Arno Solin, Vincent Adam. *Sparse algorithms for Markovian Gaussian processes*, AISTATS 2021.

Training 1

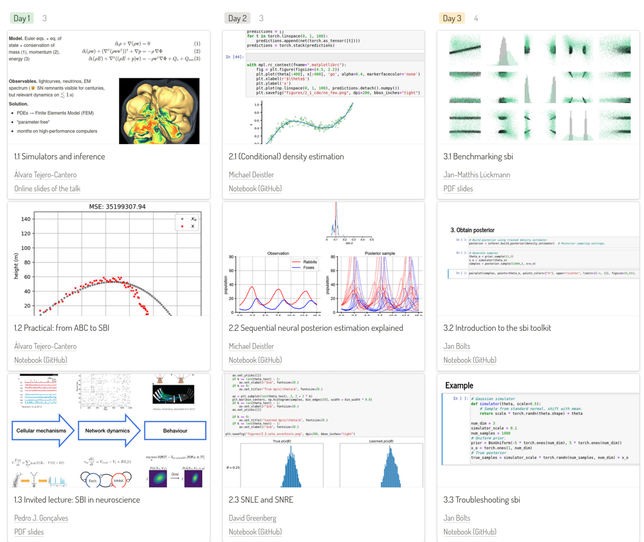
IntroML workshop 3/22, 7/22



A workshop for Tübingen researchers across disciplines that introduces machine learning and how to formulate ML projects for science

SBI workshop 9/21

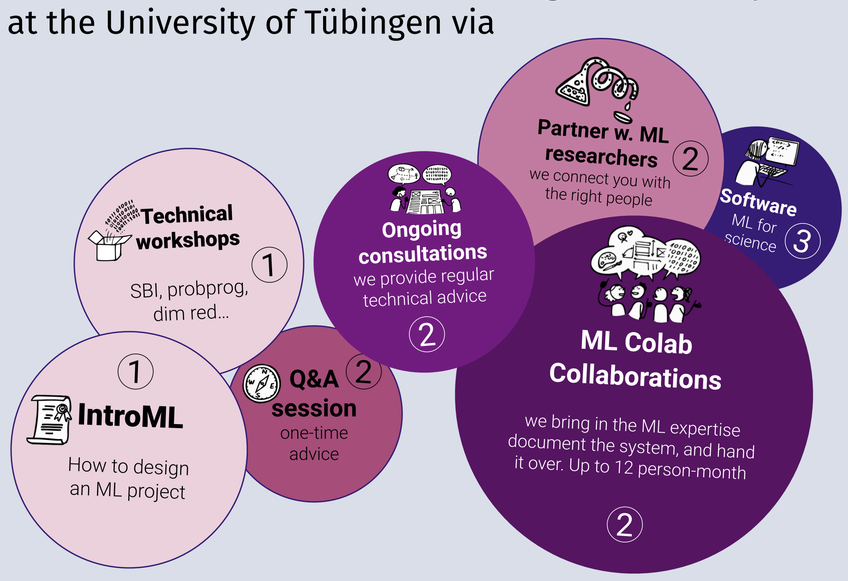
Simulation-based inference: hands-on online workshop co-organized with the Macke Lab and Helmholtz AI



Vortrag: *Was ist künstliche Intelligenz?* 10/21 @Begegnungsstätte Hirsch (with Patrick Klügel)

The ML ⇌ Science Colab

Mission: Establish machine learning across disciplines at the University of Tübingen via



① training — ② joint work — ③ ML software

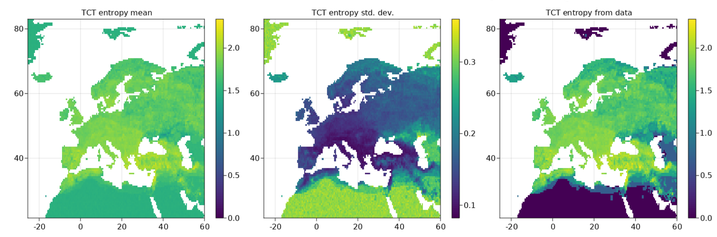
Q&A sessions 2

with geographers, behavioural neuroscientists, archaeologists, sedimentologists, orientalists, microbiologists, evolutionary ecologists, linguists, philosophers, etc.

Consultations 2

Paleobotany *data exploration*

Climate drivers of leaf morphology, w/ C. Traiser, A. Roth-Nebelsick & J. Nebelsick, *MorphoPlant*



Hydrogeology *ML advice & cluster transfer*

Simulation-based infererence for Ammer-valley hydraulic conductivity, w/ J. Allgeier & O. Cirpka

Computational linguistics *ML advice*

Acceleration of linear discriminative learning with H. Baayen & M. Heitmeier

Proposals

- ***AI Combinator*** competence center BMBF 2/22
- **Romance philology** Networks of artistic influence from visuals in Modernismo cultural magazines, w/ H. Ehrlicher & J. Lehmann BMBF 11/21 ϟ)

- **Archaeology** DFG CRC 9/21 ϟ
  1. Evolution of cumulative culture via the stone tool record, with A. Kandel, M. Bolus, C. Tennie
  2. Environmental drivers of cultural differentiation, with C. Sommer, N. Conard, U. v. Luxburg

Software Julia Python 3

Integration of `sbi` ⇌ `ArviZ`	Diagnostics of inference (SBI)
ECDFs w/ confidence bands	Calibrate inference (SBI, PP)
`Pathfinder.jl`: early diagnostics for prob. models	Bayesian regression (MorphoPlant)
Reactive notebook: supervised learning	Workshop series (IntroML)
Bayesian CV functions for `ArviZ`	Validate paleoclimatic GPs, *PollenClim*
Gap analysis: Julia for deep learning (DL)	Unify HPC+DL for sci (AI Combinator)
Mean functions for `markovflow`	Enlarge feature space for GPs (PollenClim)

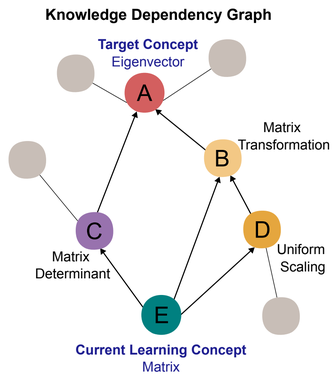
Trigger

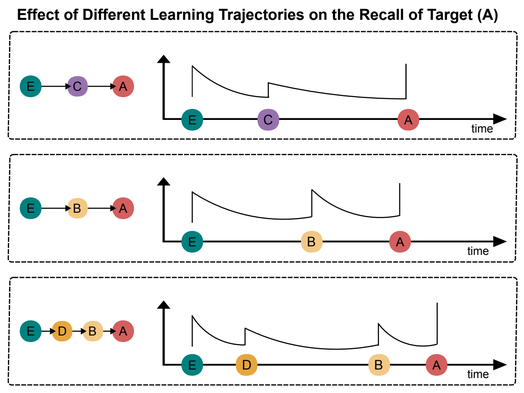
ML to Help Humans Learn Structured Domains

SP2 of ML in Education, PhD Hanqi Zhou, w/ Charley Wu (HMC)

Goal

1. Model students' internal knowledge states
   What is the influence of forgetting and domain knowledge structure?
2. Develop teacher's didactic policy
   What is the optimal time to intervene and best material to present?





Method

1. Simulation Model knowledge graph with probabilistic graph, agents with reinforcement learning
2. Optimization Develop a platform to collect real-user data to improve the model continuously