Introduction to Machine Learning

Finse Alpine Research Centre, Norway

Konstantin A. Maslov
k.a.maslov@utwente.nl

University of Twente,
Faculty of Geo-Information
Science and Earth Observation

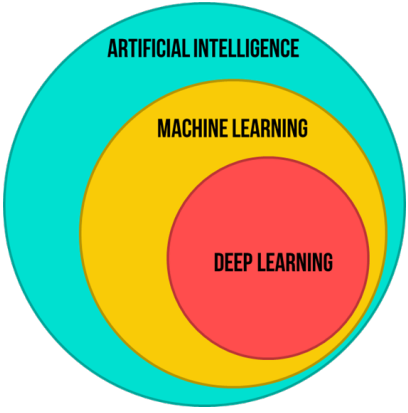
Thomas Schellenberger
thomas.schellenberger@geo.uio.no

University of Oslo,
Faculty of Mathematics and
Natural Sciences

September 2024

Outline

Artificial intelligence, machine learning and deep learning



Source: https://bit.ly/43D5KEP

- Artificial intelligence (AI) is a broad field of computer science that explores machines that can perform tasks that typically require human intelligence, such as perception, reasoning, learning, and problem-solving
- Machine learning (ML) is a subset of AI that focuses on algorithms and statistical models that allow computers to improve their performance on a specific task based on data without being explicitly programmed to do so
- Deep learning (DL) is a subset of ML that uses algorithms designed to learn hierarchical representations of data where each following representation is progressively more abstract and high-level

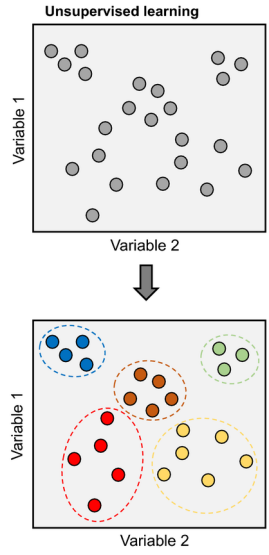
The definitions are given by ChatGPT

Unsupervised learning

Unsupervised learning is a type of ML where an algorithm learns from unlabeled data without the need for explicit supervision or guidance. In unsupervised learning, the goal is to find patterns and structure in the data, without any knowledge of what the correct output should be.

Typically, the following are related to unsupervised learning:

- Clustering
- Dimensionality reduction
- Anomaly detection



Source: https://bit.ly/41l905Y

Unsupervised learning: clustering

Unsupervised road extraction via a Gaussian mixture model with object-based features

Jiayuan Li, Qingwu Hu and Mingyao Ai

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

ABSTRACT

Automatic road extraction from remotely sensed images is an important and challenging task. This article proposes an unsupervised road detection method based on a Gaussian mixture model and object-based features. Our approach has five major stages, i.e. superpixel segmentation, feature description, homogeneous region merging, clustering via the Gaussian mixture model, and outlier filtering. In the third step, we present a graph-based region merging algorithm, in which the nodes of the graph are superpixels and edges are the similarities of intensity, colour, and texture. We also define two shape features, called deviation of parallelism (DoP) and narrow rate (NR), to automatically recognize road layer and filter outliers in the last step. We evaluated the proposed method on a variety of datasets, in which the Vaihingen dataset from the International Society for Photogrammetry and Remote Sensing Test Project is also included. Results demonstrate the power of our approach compared with some state-of-the-art methods.

ARTICLE HISTORY

Received 13 June 2017

Accepted 30 December 2017

Figure 4. Road networks extracted in three patches of the EPFL. First row: original images; second row: Turetken's results; third row: Wegner's results; fourth row: our results; fifth row: ground truth. Green true positives, blue false positives, red false negatives.

Source: Li et al., 2017

Unsupervised learning: dimensionality reduction

RESEARCH ARTICLE

UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts

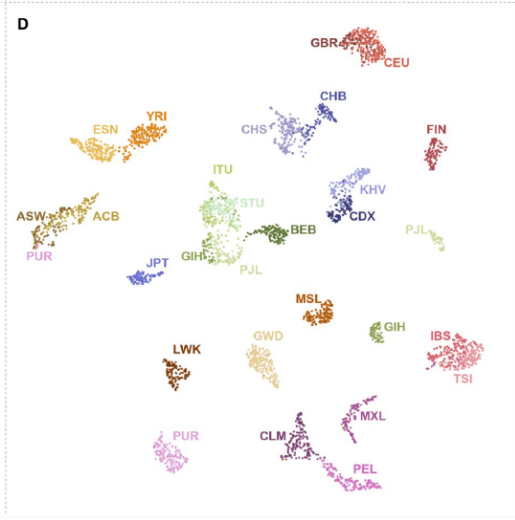
Alex Diaz-Papkovich[1,2], **Luke Anderson-Trocmé**[2,3], **Chief Ben-Eghan**[2,3], **Simon Gravel**[2,3]*

1 Quantitative Life Sciences, McGill University, Montreal, Québec, Canada, **2** McGill University and Genome Quebec Innovation Centre, Montreal, Québec, Canada, **3** Department of Human Genetics, McGill University, Montreal, Québec, Canada

* simon.gravel@mcgill.ca

Abstract

Human populations feature both discrete and continuous patterns of variation. Current analysis approaches struggle to jointly identify these patterns because of modelling assumptions, mathematical constraints, or numerical challenges. Here we apply uniform manifold approximation and projection (UMAP), a non-linear dimension reduction tool, to three well-studied genotype datasets and discover overlooked subpopulations within the American Hispanic population, fine-scale relationships between geography, genotypes, and phenotypes in the UK population, and cryptic structure in the Thousand Genomes Project data. This approach is well-suited to the influx of large and diverse data and opens new lines of inquiry in population-scale datasets.



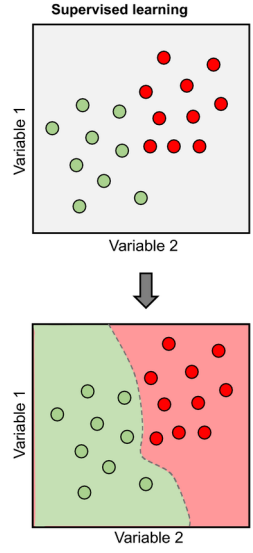
Source: Diaz-Papkovich et al., 2019

Supervised learning

Supervised learning is a type of ML where an algorithm learns from labelled data with explicit supervision. In supervised learning, the input data is paired with corresponding output data, which serves as the 'correct' answer. The algorithm then learns to map the input data to the correct output data, typically by adjusting a set of parameters in the model to minimize the difference between the predicted output and the true output.

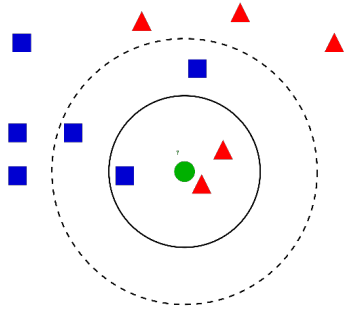
Typical tasks are:

- Classification
- Regression

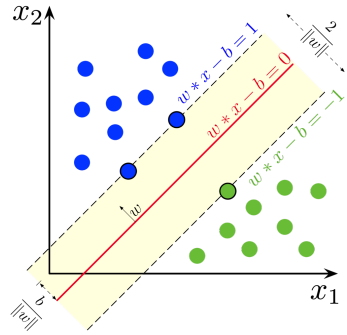


Source: https://bit.ly/41l905Y

Machine learning: k-NN, SVM



Source: https://bit.ly/3KDmAL7

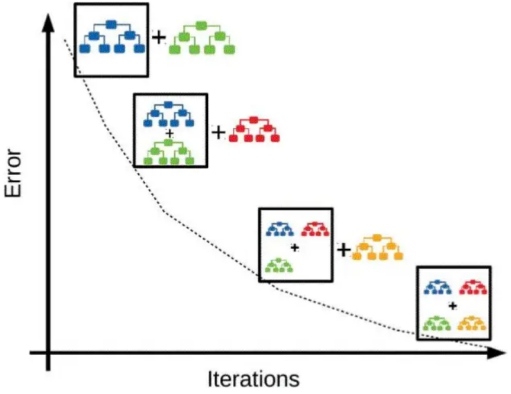


Source: https://bit.ly/3MN4cSW

- Easy to construct and interpret
- Relatively fast

- Require serious feature engineering for some tasks
- Probably, not the choice if one wants the highest performance

Machine learning: xgboost



Source: https://bit.ly/3o742va

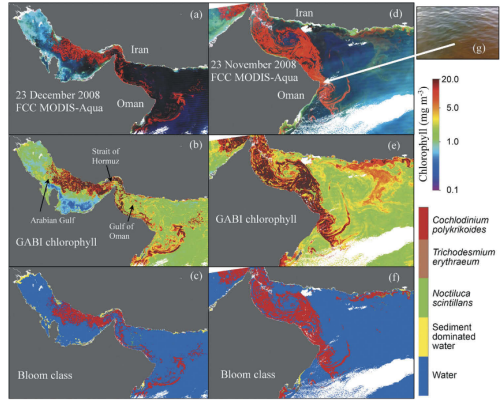


Figure 9. (a) Gulf of Oman and Arabian Gulf captured by MODIS-Aqua on 23 December 2008 in the False Colour Composite (FCC) (R-748nm, G-547nm, B-488nm) image, (b,c) Corresponding chl concentration and bloom classes estimated using the XGBoost models, (d) FCC MODIS-Aqua image of Gulf of Oman captured on 23 November 2008, and (e and f) the corresponding chl concentration and bloom class images. (g) Field photograph reported in Shanmugam, Suresh, and Sundarabalan (2013).

Source: Ghatkar et al., 2019

- Probably, state of the art for tabular data
- Resistant to overfitting
- For some data modalities (e.g. images), requires feature engineering as well

Deep learning: MLP

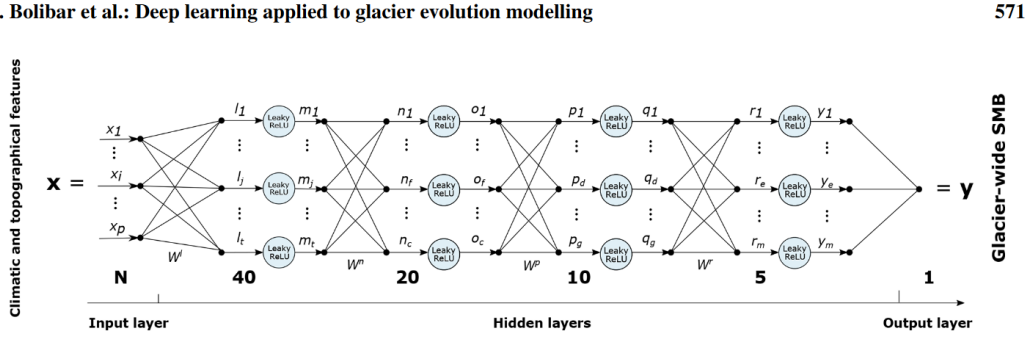
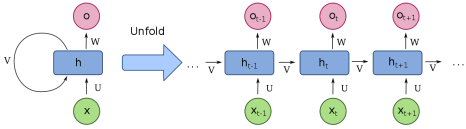


Figure 3. Deep artificial neural network architecture used in ALPGM. The numbers indicate the number of neurons in each layer.

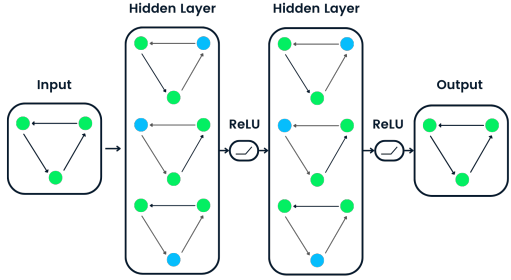
Source: Bolibar et al., 2020

- Simple concept
- Capable of learning from 'raw' features
- If you end up using MLP, maybe have a look at xgboost

Deep learning: RNN, GNN



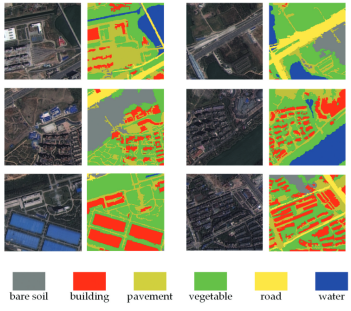
Source: https://bit.ly/3A2goYj



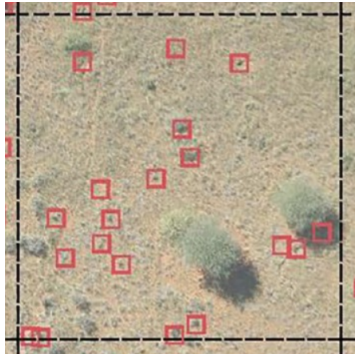
Source: https://bit.ly/40cDmql

- Natural choices for time series and graphs, respectively

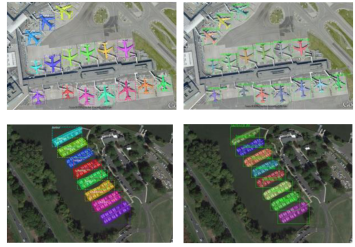
Deep learning: CNN



Source: Li et al., 2021



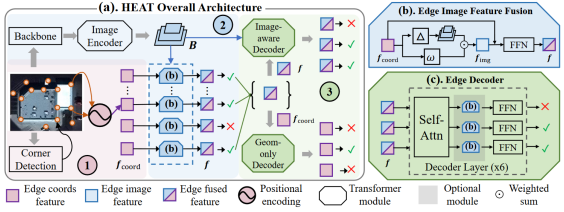
Source: Kellenberger et al., 2018

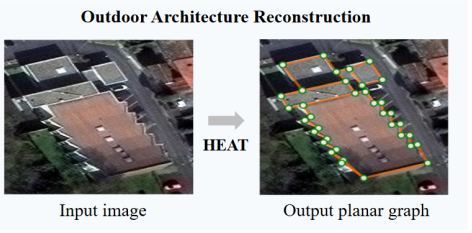


Source: Su et al., 2020

- Probably, state of the art (still) for computer vision
- Too many design choices

Deep learning: transformer





Source: Chen et al., 2021

- Global receptive field from the first layer
- Suitable for any data modality
- Probably, state of the art for NLP and geometry learning

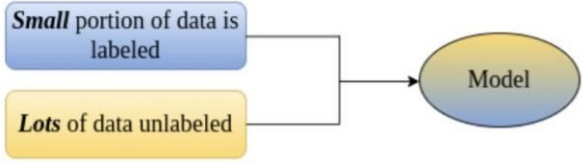
- High time complexity
- Usually, requires a lot of data to train

Semi-supervised learning

Quite often, it is very expensive to get labelled data while easy to access only the input features. For example, we can easily access a lot of satellite imagery, but manually digitised high-quality land-use maps are not available everywhere.

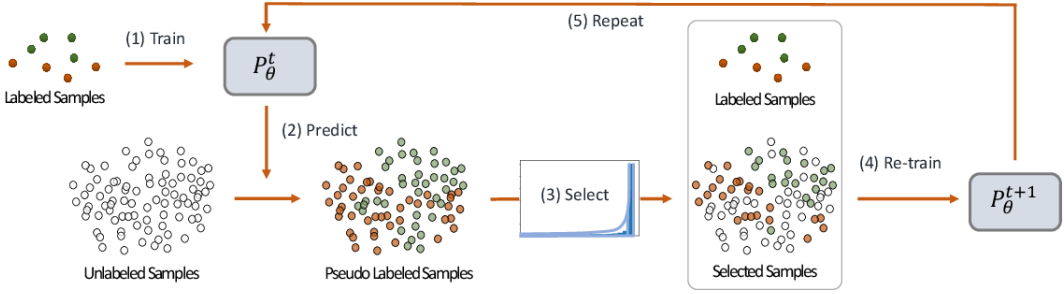
We still can utilise the input features with no labels as they give as valuable information about data distribution.

Here comes semi-supervised learning, which is a type of ML that relies on both labelled and unlabelled data trying to exploit as much data as possible.



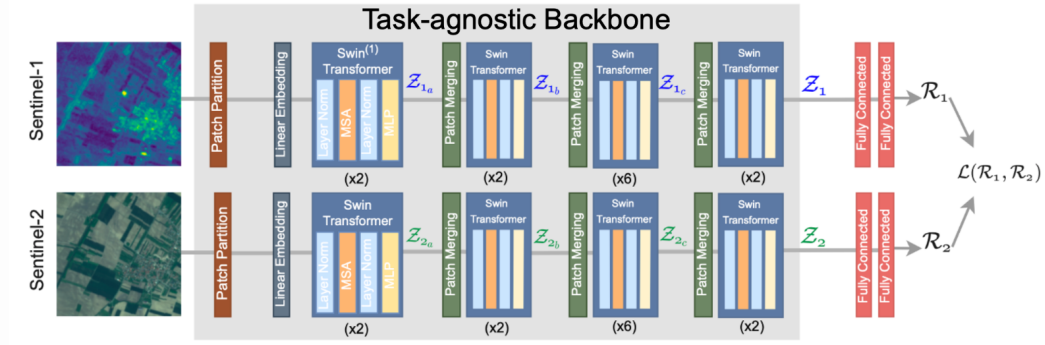
Source: https://bit.ly/3GMtz3o

Semi-supervised learning: pseudo-labelling



Source: Cascante-Bonilla et al., 2020

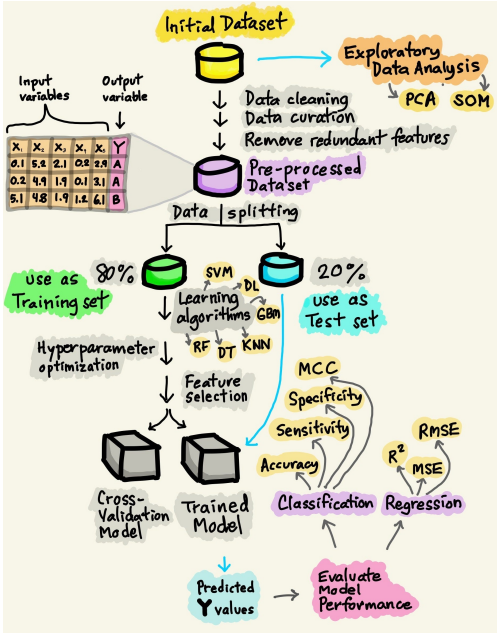
Semi-supervised learning: self-supervised task-agnostic models



Architecture of our task-agnostic backbone. For Sentinel-1 and Sentinel-2 input pairs, we pre-train a unique backbone consisting of two streams of Swin Transformers using a self-supervised contrastive loss.

Source: Scheibenreif et al., 2022

Machine learning pipeline



Source: https://bit.ly/43A5whN

- There are several ways to perform validation
  - Cross-validation
  - Having a fixed subset for validation
- When doing the data split, pay attention to possible cross-correlations between subsets (both spatial and temporal). Avoid them, otherwise, your accuracy estimates are biased
- 80/20 split is just one of the options
- Some practitioners merge training and validation data to produce the final model ready for deployment. It is debatable
- Note that often the models you decided to proceed with dictate how you adapt different steps

Trustworthy machine learning

- A lot of ML, especially DL, algorithms are not interpretable by design
- However, they often provide state-of-the-art accuracies

Important questions arise:

- How can we trust AI systems?
- How to make them transparent (at least to some degree)?
- How to ensure that they are fair?

- **How sure** is a particular model about its predictions?
- **Why** did the model make a particular prediction?

Trustworthy machine learning: uncertainty quantification

How sure?

Typical approaches are:

- Ensembling
- Bayesian machine learning
  - Monte-Carlo dropout
- Deterministic methods
- Test-time augmentation

No matter which approach you use, check the quality of the uncertainty estimates empirically, e.g. with reliability diagrams. Calibrate if needed.

Efficient Uncertainty Estimation in Semantic Segmentation via Distillation

Christopher J. Holder Muhammad Shafique
New York University Abu Dhabi
Abu Dhabi, United Arab Emirates
{chris.holder, ms12713}@nyu.edu

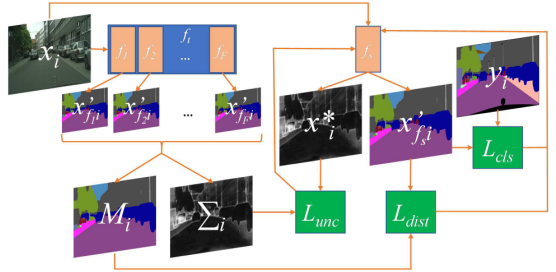
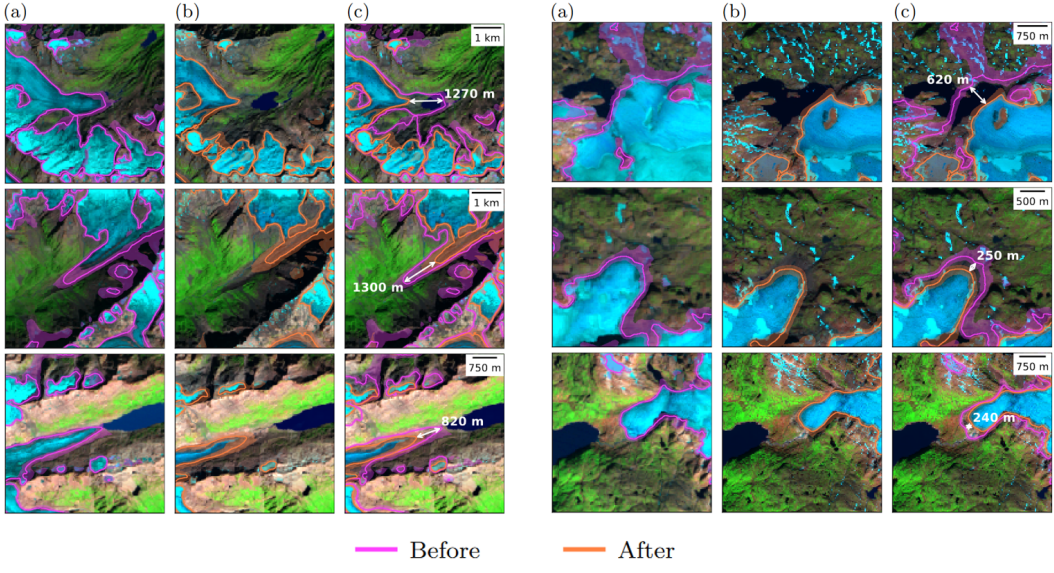


Figure 1. The process for training the student model in Uncertainty Distillation. For input x_i, ensemble models f_1 to f_T comprising teacher f_t output probability maps x'_{f1} to x'_{fn}, the elementwise mean and standard deviation of which populate M_i and $\sum_i$ respectively. Student model f_s outputs class probability map x'_{fsi}, which is compared with ground truth segmentation y_i to compute classification loss L_{cls} and to M_i to compute distillation loss L_{dist}, and uncertainty map x^*_i, which is compared to $\sum_i$ to compute uncertainty loss L_{unc}. These losses are then combined to optimise f_s such that it learns to approximate the output distribution of f_t.

Source: Holder and Shafique, 2021

Trustworthy machine learning: uncertainty quantification



Trustworthy machine learning: explainable AI

scientific reports

OPEN

Predicting sex from retinal fundus photographs using automated deep learning

Edward Korot[1], Nikolas Pontikos[1], Xiaoxuan Liu[1,2,3], Siegfried K. Wagner[1], Livia Faes[1,4], Josef Huemer[1,5], Konstantinos Balaskas[1], Alastair K. Denniston[1,2,3,6], Anthony Khawaja[1✉] & Pearse A. Keane[1✉]

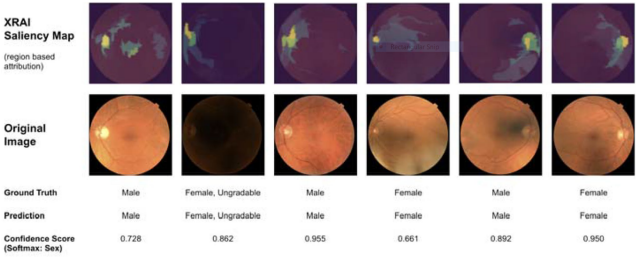


Figure 2. Region based saliency maps for model prediction: colors represent regions in order of decreasing performance: Yellow, Green, Blue. Images sourced at random from validation set, with the addition of an ungradable image.

Source: Korot et al., 2021

Why?

Algorithms depend strongly on the data modality and (partly) on the models.

Similar to p-tests, you may be tempted to choose those algorithms that support your motivation and expectations while ignoring others. Don't do it, be honest.

Machine learning in Python

Luckily, Python has a lot of packages that provide ML and DL algorithms out of the box so we do not need to invent the wheel once and once again.

For ML in Python:

- scikit-learn
- scipy.stats
- xgboost
- pymc3

For DL in Python:

- tensorflow and tensorflow.keras
- pytorch
- JAX

And more...

Some useful repositories

README GPL-2.0 license

nicesnappy

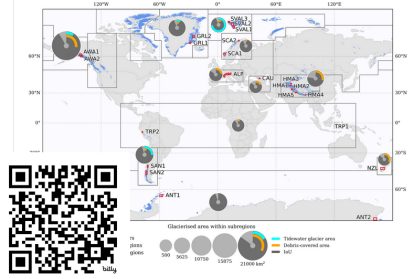
A simple programming interface to snappy.

README GPL-2.0 license

Scalable Glacier Mapping using Deep Learning and Open Earth Observation Data Matches the Accuracy of Manual Delineation

Konstantin A. Maslov, Claudio Persello, Thomas Schellenberger, Alfred Stein

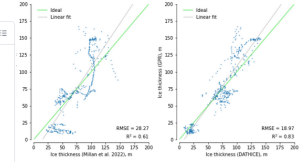
[Paper] [Datasets] [BibTeX]



README GPL-2.0 license

DATHICE: Data-Assisted THickness correction for ICE

[Installation] [Getting started] [BibTeX]



is short exercise demonstrates a simple hybrid-physics-data model (inspired by Daw et al. 2017) that is trained to rect the modelled data from Millan et al. 2022. It adapts and improves upon traditional physical models by orporating ground-penetrating radar data to correct inaccuracies. This method involves training a deep learning odel to adjust the ice thickness measurements based on discrepancies identified between the physically modeled ta and GPR observations from three distinct glaciers in Svalbard—Scott Tur rection relies and two coregistered elevation models (Cop30DEM and Arct odelling (remotely sensed glacier surface velocities) and the modelled thickr slightly) more accurate assessment of ice thickness and glacier volumes (up

is repository is used mainly as a teaching resource and does not claim to h

Introduction to Machine Learning

Finse Alpine Research Centre, Norway

Konstantin A. Maslov
k.a.maslov@utwente.nl

University of Twente,
Faculty of Geo-Information
Science and Earth Observation

Thomas Schellenberger
thomas.schellenberger@geo.uio.no

University of Oslo,
Faculty of Mathematics and
Natural Sciences

September 2024