Machine Learning in Python for Environmental Science Problems: Introduction

AMS Committee on AI Applications to Environmental Sciences

January 28, 2024

Instructors and Organizers

AMS AI STAC Committee on AI Applications to Environmental Science

Kara Lamb
(kl3231@columbia.ed

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

LEAP

Evan Krell
(ekrell@islander.tamucc.edu)

TEXAS A&M UNIVERSITY
CORPUS CHRISTI

Maria J. Molina
(mjmolina@umd.edu)

DEPARTMENT OF
ATMOSPHERIC &
OCEANIC SCIENCE

Hamid Kamangir
(Hamid.Kamangir@tamucc.edu)

UCDAVIS
UNIVERSITY OF CALIFORNIA

Julia L. Simpson
(jls2391@columbia.edu)

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

LEAP

Training objectives for today's short course

At the end of training, participants will be able to :

- Recognize common machine learning methods used for processing Environmental Science data
- Describe benefits and limitations of machine learning for Environmental Science
- Understand basic machine learning algorithms and techniques as they can be applied to Environmental Science problems
- Learn about stages for developing machine learning pipelines to analyze Environmental Science data sets and evaluation metrics
- Introduction to more advanced ML methods and evaluation metrics

- Have access to resources where you can learn more!

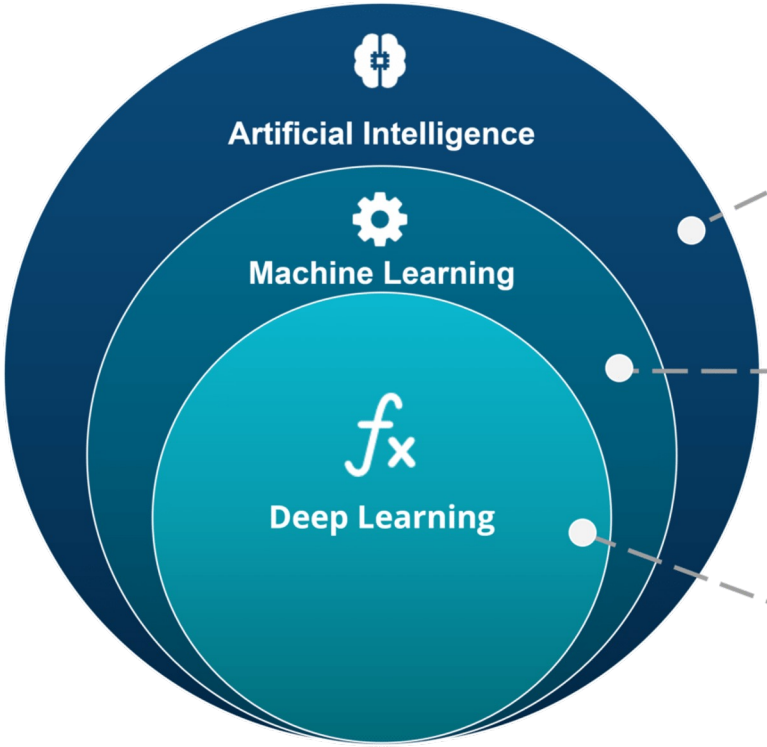
Overview and Theory

Why do you want to do ML?

What ML technique to use?

How to prepare the data?

What do we mean by machine learning?



ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour

MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

Image source: https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/

How does Machine Learning Work?

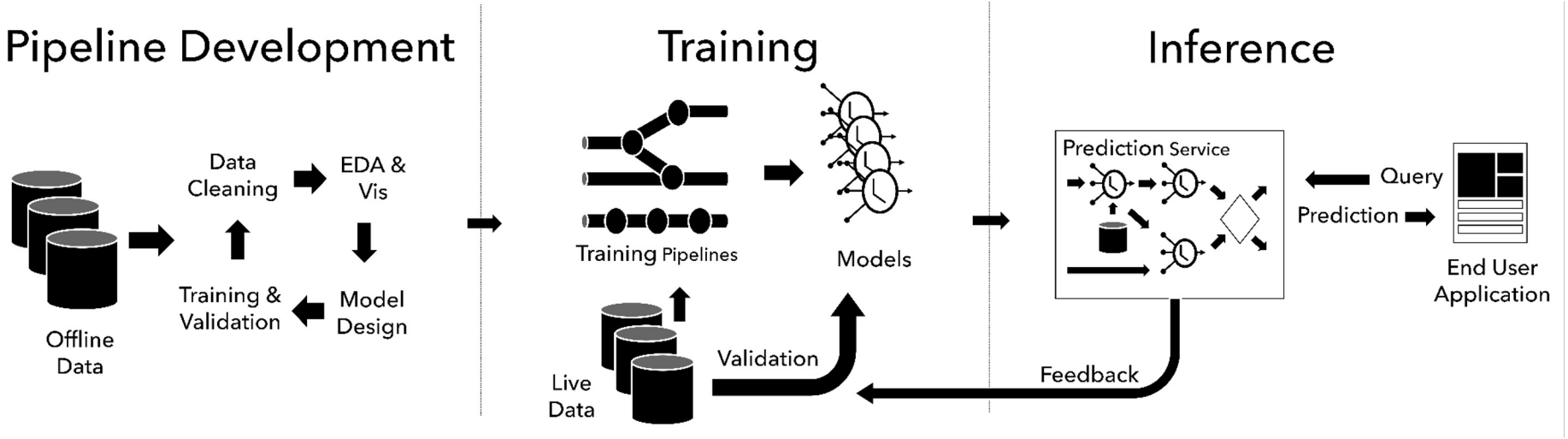


Image source: Daniel Crankshaw (*Short History of Prediction-Serving Systems*)

Machine learning steps

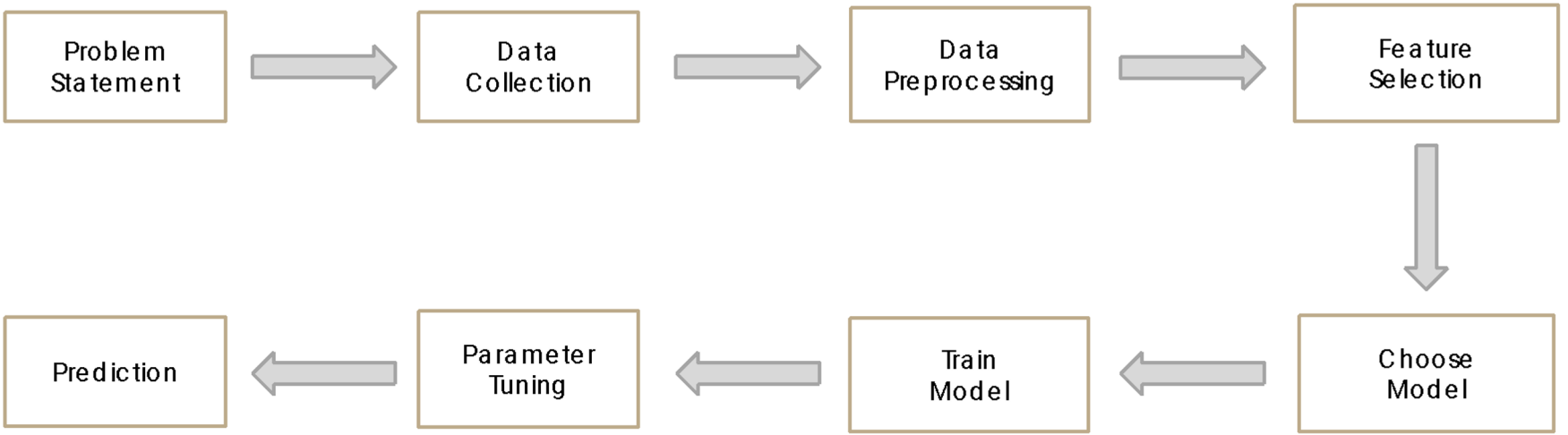


Image source: NASA ARSET ML training course

Machine Learning Algorithms

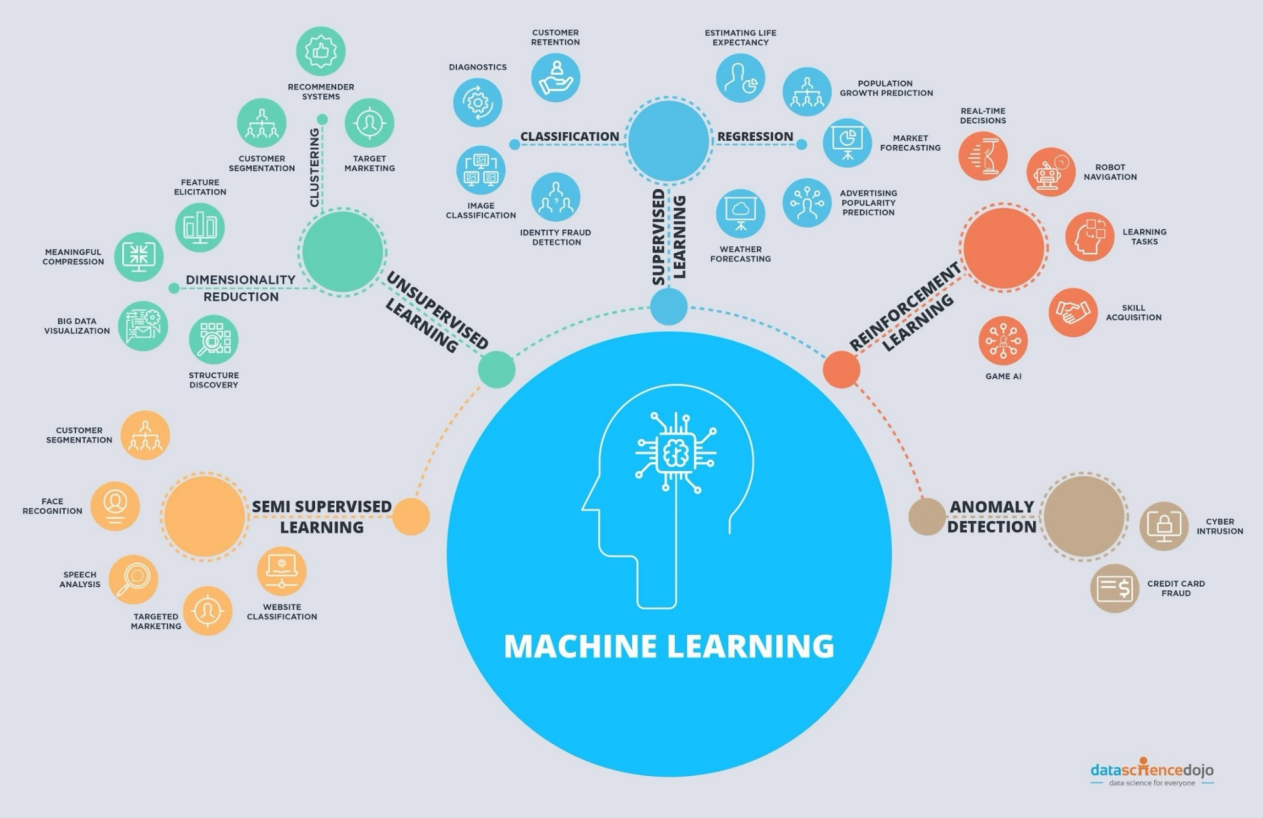
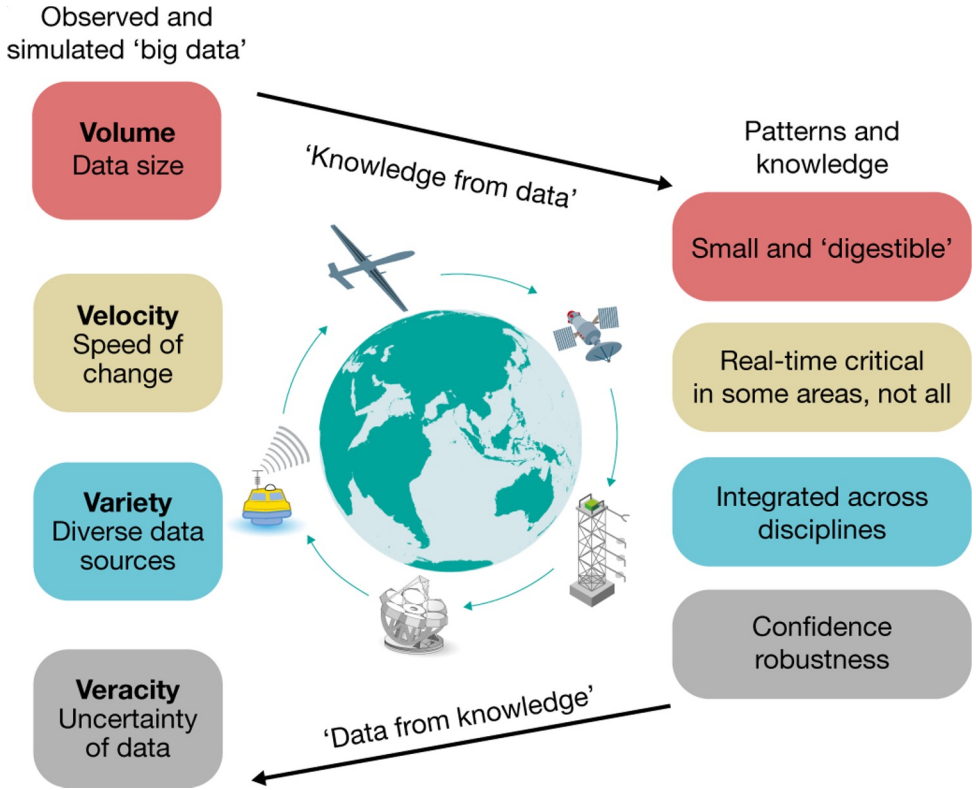


Image source: Data Science Dojo (https://datasciencedojo.com)

Data Driven Earth System Science



Reichstein et al., "Deep Learning and Process Understanding for Data-Driven Earth System Science", *Nature*, 566, 195-204 (2019)

Machine learning in the environmental sciences

- Problems in the environmental sciences are typically complex
- We often don't have a full mathematical model that can describe every aspect of the systems we are interested in

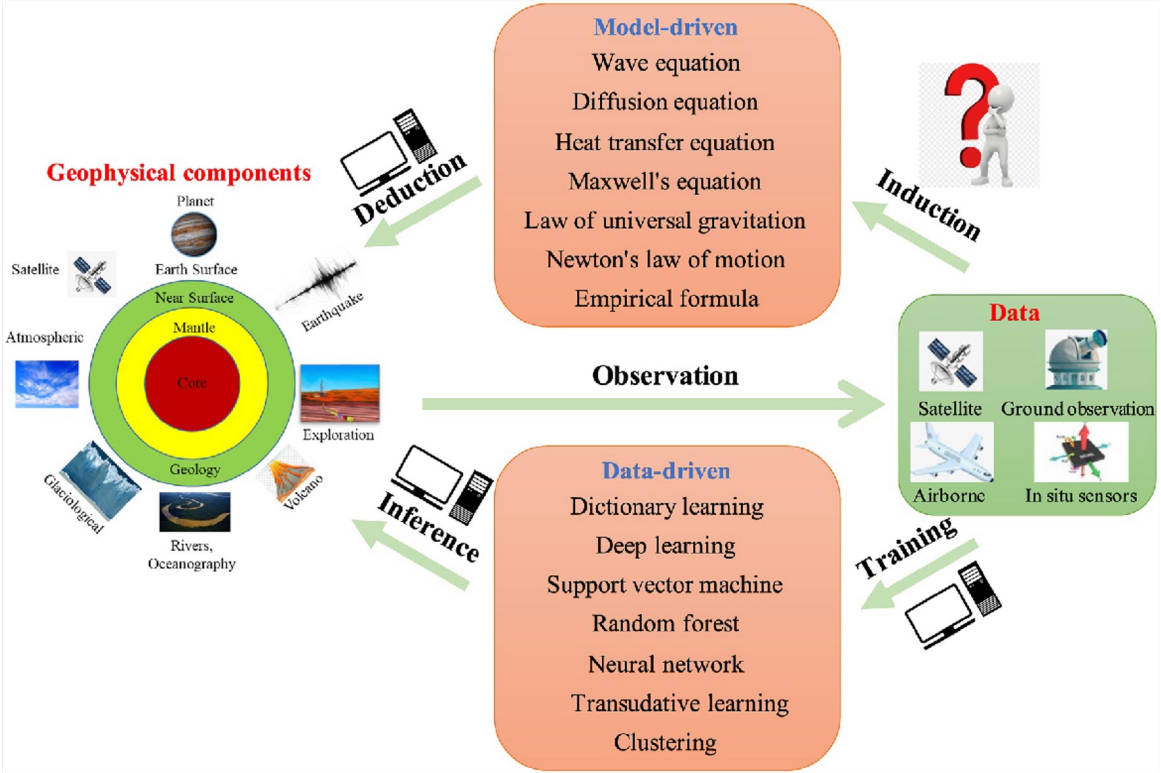
Machine learning is good at:

- Non-linear prediction tasks
- Pattern recognition
- Processing large amounts of data quickly
- Fast prediction (compared to traditional physics-based model)

Challenges for machine learning:

- Biases in data sets/typically need large data sets
- Interpretability/Lack of physical basis for models
- Integration with existing models and data sets

How machine learning is applied in the Environmental Sciences



Yu and Ma, Deep Learning for Geophysics: Current and Future Trends, *Reviews of Geophysics* (2021)

How can machine learning be used to accelerate scientific research?

Increased efficiency:

- Machine learning can automate analysis of large and complex data sets, allowing faster processing and integration of new observations

New insights and discoveries:

- Machine learning can identify patterns and relationships in complex data sets, leading to new insights

Improved predictive modeling:

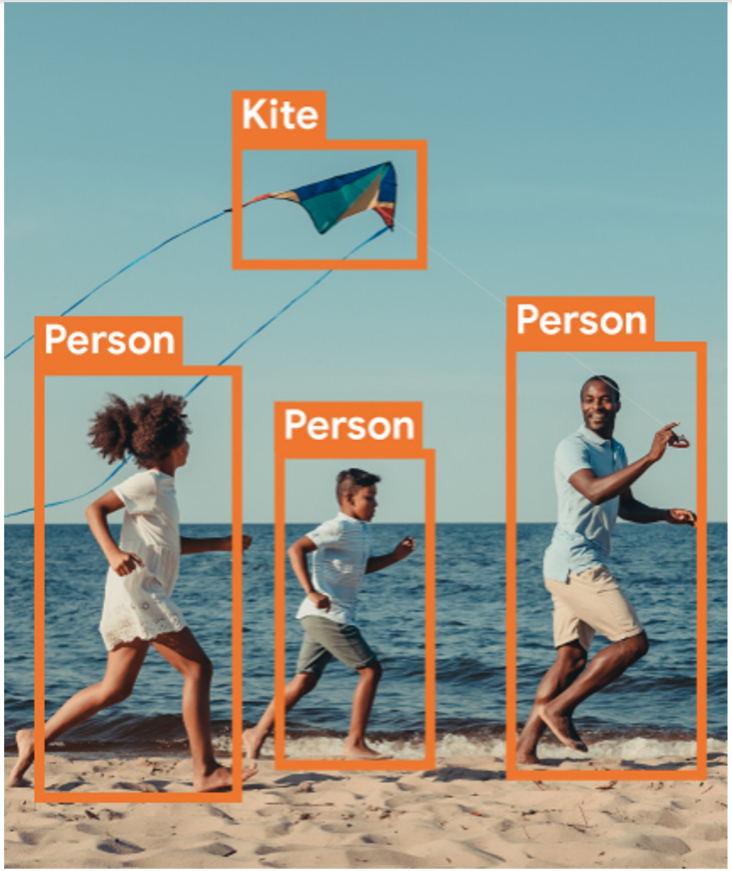
- Machine learning can be used to build accurate predictive models that can help scientists better understand complex environmental systems

Pattern recognition in the environmental sciences

- One major use case where deep learning has demonstrated increasing skill and progress in recent years is **computer vision** - using machine learning to understand and identify objects in images or video

- Pattern recognition lends itself to a number of different use cases (feature detection, processing large amounts of data quickly, anomaly detection)

- Pattern recognition in the geosciences can also extend to the analysis of time series, spatio-temporal patterns, or observational data sets from specialized instrumentation



Pattern recognition in the environmental sciences

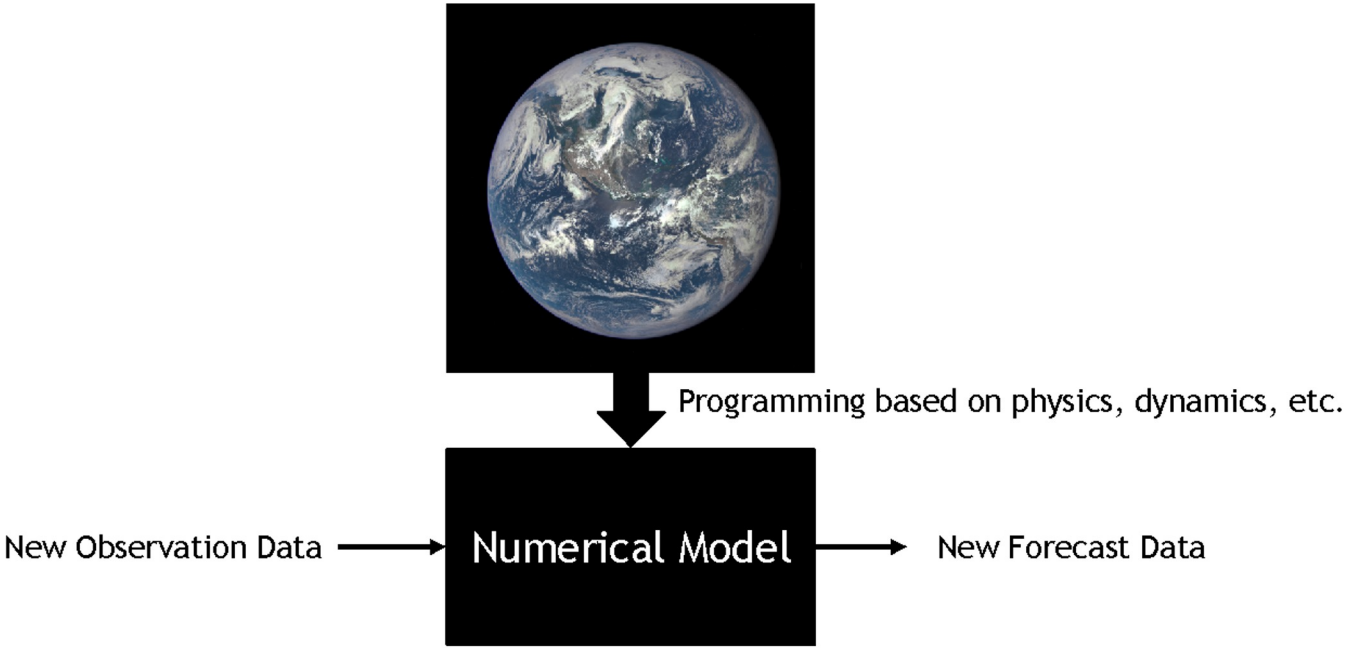
- Analysis of observational data sets – satellite remote sensing
- Examples of use cases: fire detection, wildfire burnt area, coastline and flood mapping, canopy height, glaciers, sea ice, classifying clouds

Classification

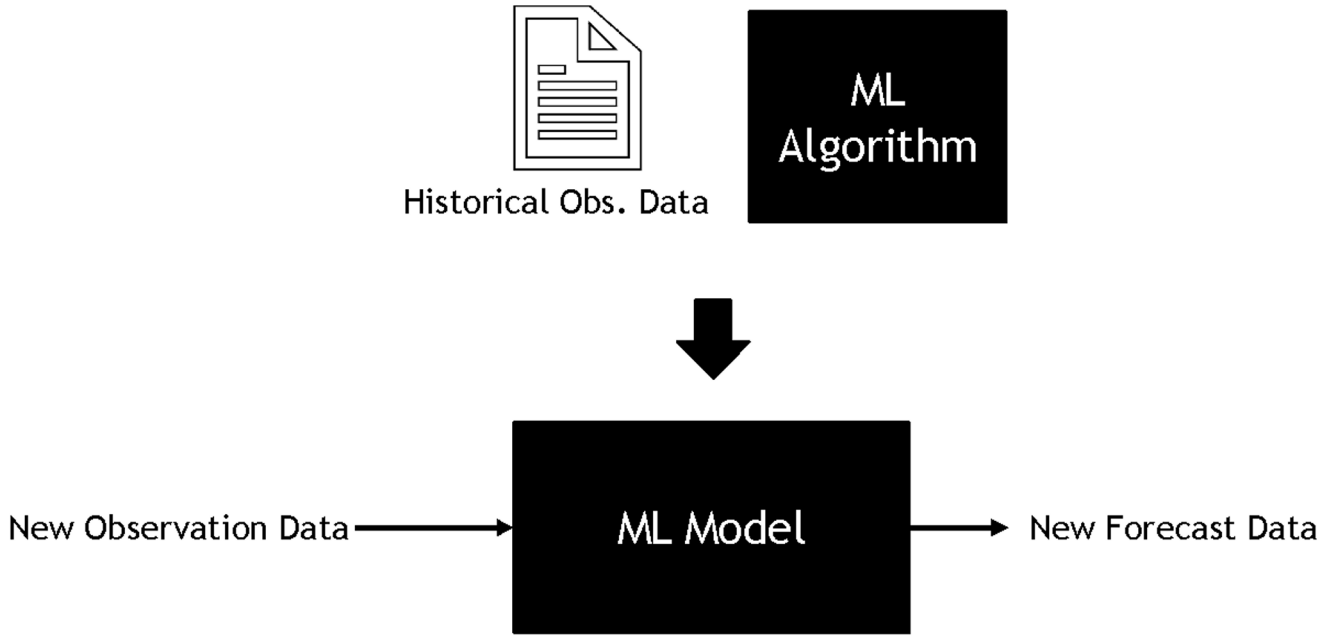
Segmentation

Source: https://github.com/satellite-image-deep-learning/techniques

Improving predictive models using machine learning



Improving predictive models using machine learning



Scientific machine learning to improve model parameterizations

Neural networks are universal function approximators

$$y = f(X)$$

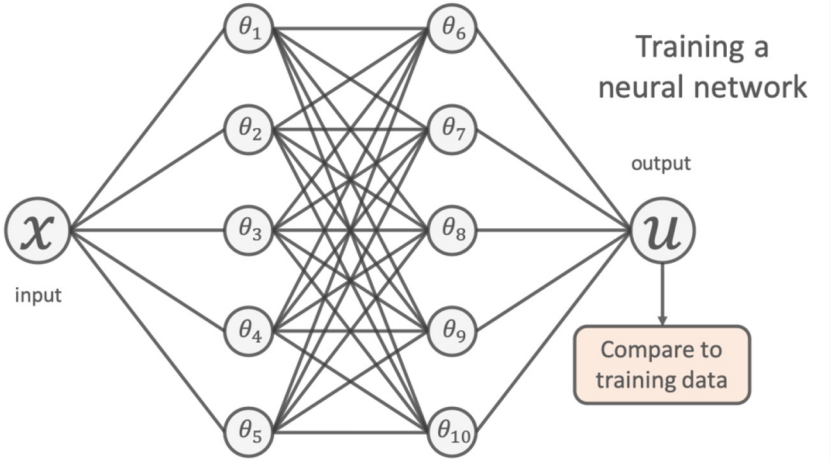
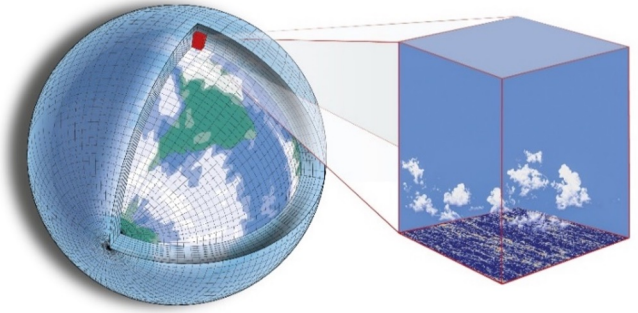


Image credit: B. Moseley



Schneider et al. 2017

Scientific machine learning:

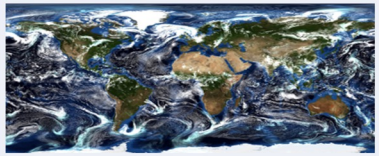
- Spatial & temporal upscaling and downscaling
- Reduced order modeling
- Perturbed Parameter Ensembles for modeling tuning and sensitivity evaluation
- Hybrid physics machine learning models
- Symbolic regression
- …

Development of hybrid-physics machine learning models

- Physics-informed machine learning
- Hybrid-physics machine learning models
- Model parameterization development
- Reduce structural and parametric uncertainty in physics-based models

1. Massive data from Earth observation

2. High-resolution cloud resolving models

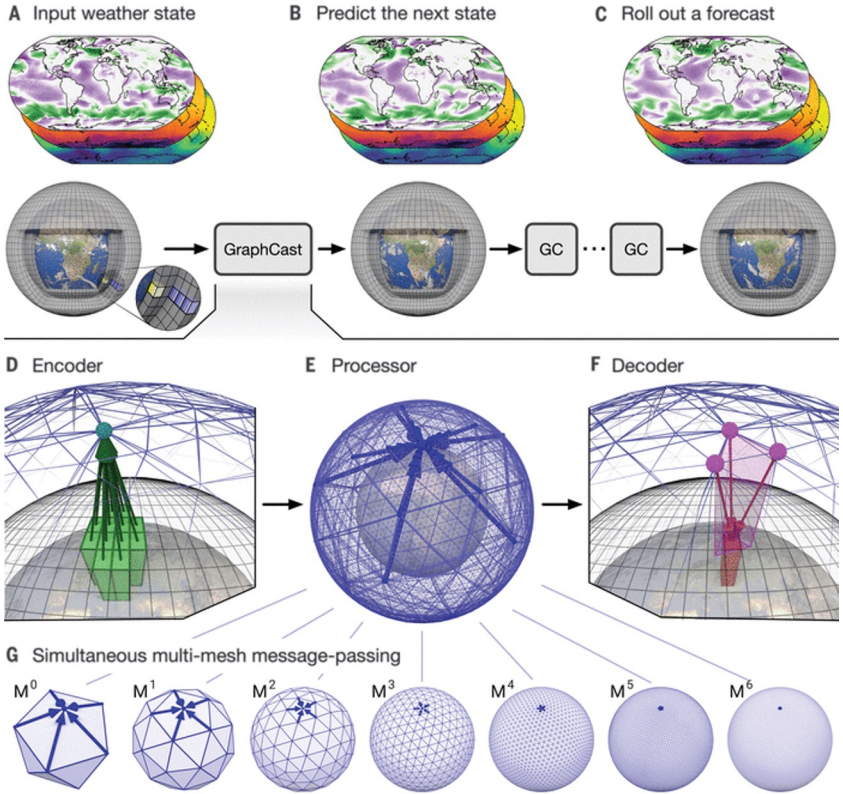


3. Progress in machine learning

Improved Earth System Understanding!

Coupled hybrid model

Improved climate and weather forecasting



Machine learning based forecasting models have recently demonstrated skill on par with state of the weather forecasting models

- FourCastNet [Pathak et al. 2022]
- Pangu-Weather [Bi et al. 2023]
- GraphCast [Lam et al. 2023]
- Neural GCM [Kochkov et al. 2023]
- Etc.

Lam et al. Learning skillful medium-range global weather forecasting, Science, 382, 6677, 1416-1421 (2023)

Schedule for Today's Course

Time	Topic	Track
8:00 – 8:30	Introduction to Machine Learning for the Environmental Sciences	Beginner
8:30 – 9:30	Data Preprocessing	
9:30 – 9:45	*Coffee break*	
9:45 – 11:15	Learning Methods	
11:15 – 12:00	Model Evaluation	
12:00 – 1:00	*Lunch*	
1:00 – 1:45	Physics-informed AI	Intermediate
1:45 – 2:30	Explainable AI	
2:30 – 2:45	*Coffee break*	
2:45 – 3:30	Transformers	
3:30 – 3:45	Conclusions & Additional Resources	

Software to support machine learning

Python: most popular language for machine learning. Many of the most popular libraries for machine learning have been developed in python
Julia: newer language, focused on performance computing in scientific and technical fields
Other languages that support ML: Matlab, R, Java, etc.

Machine learning frameworks in python

- **Scikit-learn** – good for data preprocessing, metrics, model evaluation, traditional machine learning algorithms (easy to use out of the box algorithms, but not deep learning)
- **Tensorflow –** Deep learning library developed by Google (better visualization, more out of the box, more scalable?)
    - **Keras** – Deep learning library (integrated with Tensorflow since 2017)
- **Pytorch –** Deep learning library developed by Facebook AI (generally considered more "Pythonic", more popular in the research community)
- **Etc.**

Tools for development

Jupyter notebooks – rapid prototyping of python code, data pre-processing and visualization

Computational architecture to support machine learning

GPU (graphical processing unit) - Deep learning algorithms are typically trained on GPU's because optimization algorithms need to be trained with large data sets for many epochs, and these operations are highly parallelizable

Many Python libraries used for deep learning (Tensorflow, Keras, Pytorch, OpenCV, etc.) are designed to run on NVIDIA CUDA-enabled graphics cards. Typical users of these libraries won't have to worry about programming in CUDA directly, but do need to be aware of moving data and models to and from the GPU

CPU	GPU
Central Processing Unit	Graphics Processing Unit
Several cores	Many cores
Low latency	High throughput
Good for serial processing	Good for parallel processing
Can do a handful of operations at once	Can do thousands of operations at once

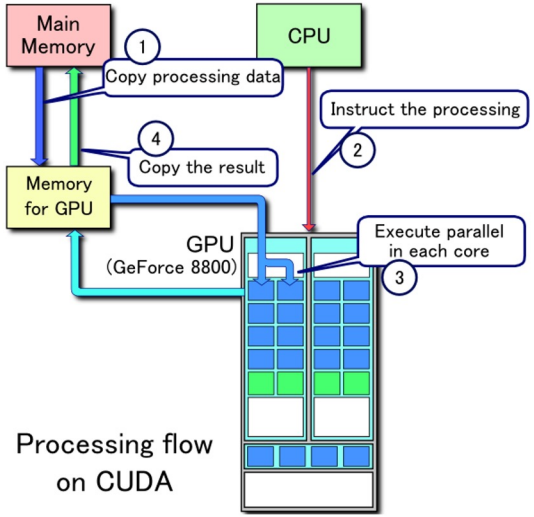


Image credit: https://blogs.nvidia.com/blog/whats-the-difference-between-a-cpu-and-a-gpu/
https://commons.wikimedia.org/wiki/File:CUDA_processing_flow_(En).PNG

Course Resources

Webpage: https://www.ametsoc.org/index.cfm/ams/education-careers/careers/professional-development/short-courses/machine-learning-in-python-for-environmental-science-problems2/

Github repository: https://github.com/ekrell/ams_ai_shortcourse_2024

Google colab:

Machine learning library:

Deep learning library:

Github