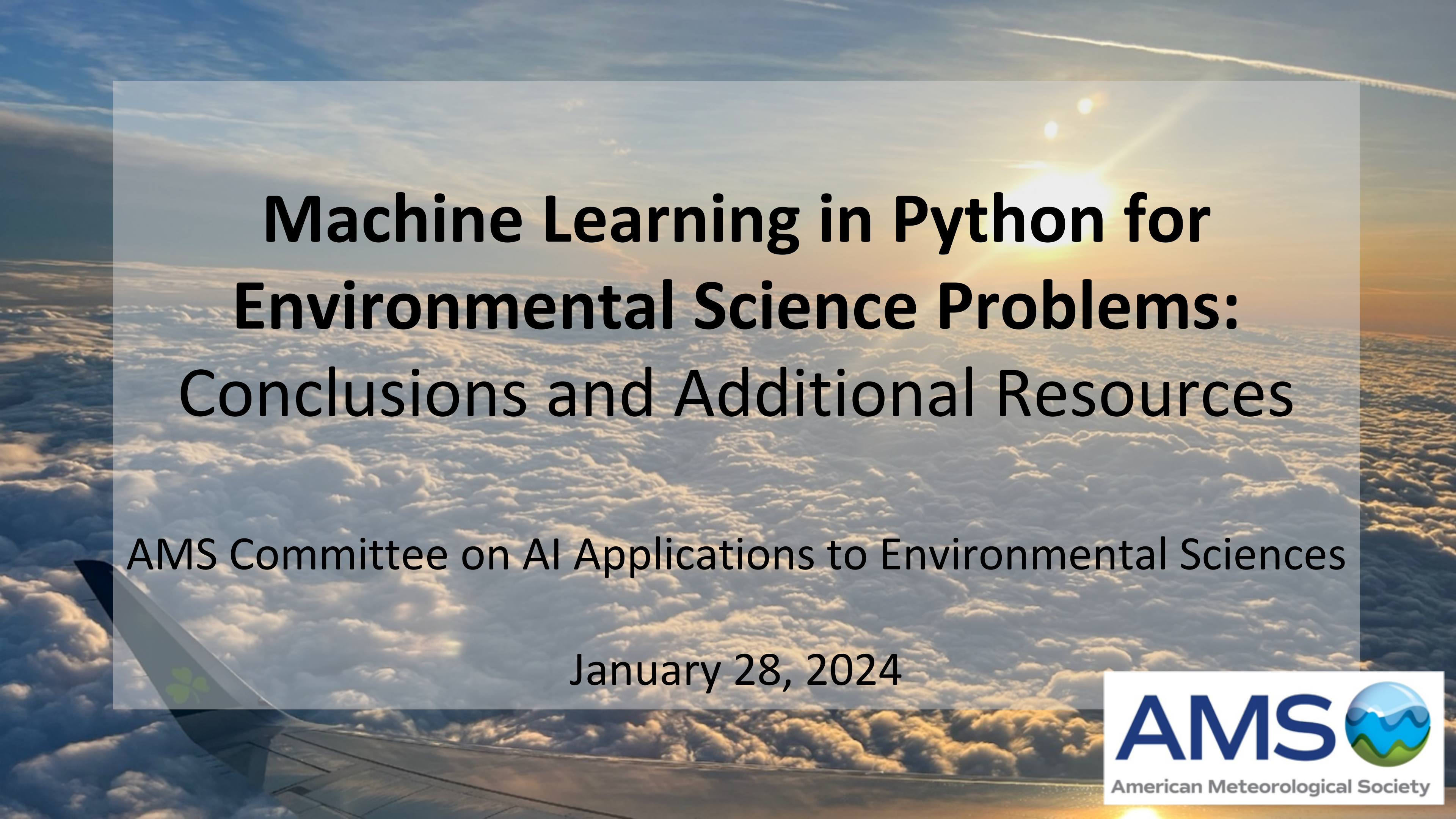


# **Machine Learning in Python for Environmental Science Problems:** Conclusions and Additional Resources

AMS Committee on AI Applications to Environmental Sciences

January 28, 2024

AMS
American Meteorological Society

# 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!**

# 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

**Github**

**Google colab:**

**Machine learning library:**

**Deep learning library:**

# 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)

UNIVERSITY OF MARYLAND 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

# Module 1: Data Preprocessing & Exploring

The module covered:

- Common file formats for handling meteorological data
  - Used Pandas to load, manipulate, and analyze tabular data
  - Used Xarray to load, manipulate, and analyze gridded raster data
- Basic data manipulations
- Simple plots to explore tabular and raster data
- Spatial libraries (e.g. Cartopy) to plot data on a map
- Displaying arbitrary multi-channel raster data
- Combining tabular and raster data for analysis

# Module 1: Data Preprocessing & Exploring - Additional Materials

**Handling Missing Data:** Real-world datasets are commonly messy, with sometimes large amounts of missing or invalid data values. There are many tutorials on c cleaning, but for generic data. For environmental datasets, domain knowledge is crucial selecting appropriate cleaning steps.

- **A survey on missing data in machine learning (Emmanuel et al., 2021)**
  https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00516-9
- **How to Perform Data Cleaning for Machine Learning with Python**
  https://machinelearningmastery.com/basic-data-cleaning-for-machine-learning/

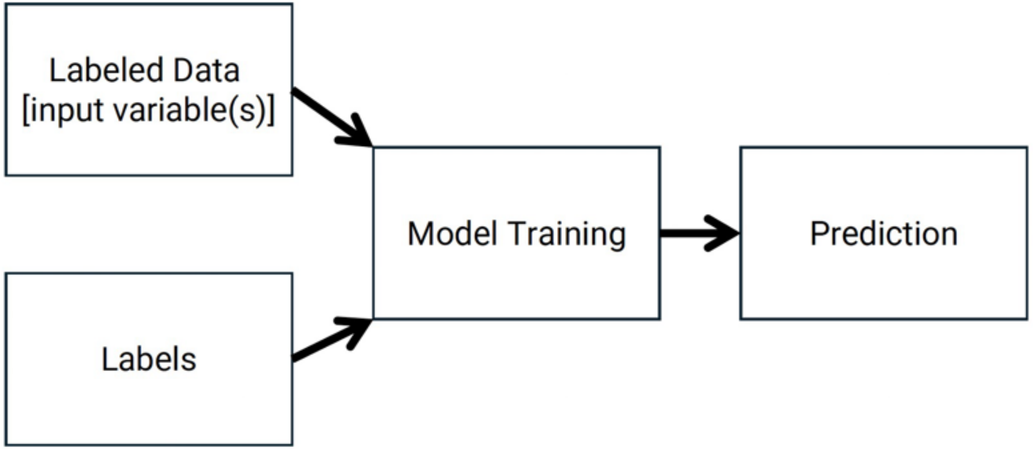
  Using Pandas for common tasks such as detection & deletion of duplicate rows
- **Pythonic Data Cleaning With pandas and NumPy**
  https://realpython.com/python-data-cleaning-numpy-pandas/ - Focuses on common tasks for improving the usability of for Pandas DataFrame
- **A framework for exploration and cleaning of environmental data–Tehran air quality data experience**
  https://journalaim.com/Article/752 - A case study data cleaning pipeline for an environmental application

**Advanced Plotting**

- **Python for Geospatial Plotting**
  https://github.com/ekrell/python-geo-plots/blob/main/GeoPlotting.ipynb - Interactive plots, mesh data, animated time series maps, etc
- **Chapter 4 Visualization | Introduction to Environmental Data Science**
  https://bookdown.org/igisc/EnvDataSci/visualization.html - In R rather than Python, but has many good plotting example
- **Analysing and Visualising Environmental Data with Python and Matplotlib**
  https://www.youtube.com/live/6sB-nu8QIP8?si=C8kU2Sckf1TiY0aT
  Video tutorial

# Module 2: Learning Methods

## Supervised Learning



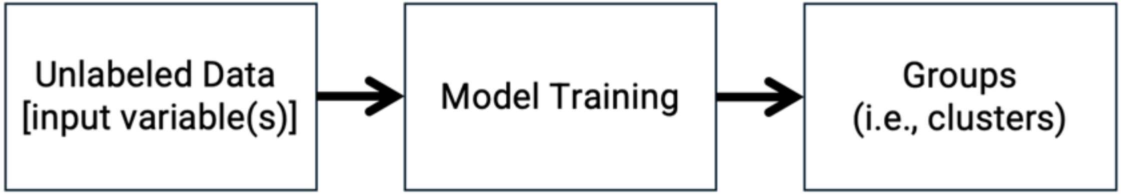
Labeled Data [input variable(s)]

Labels

Model Training

Prediction

## Unsupervised Learning

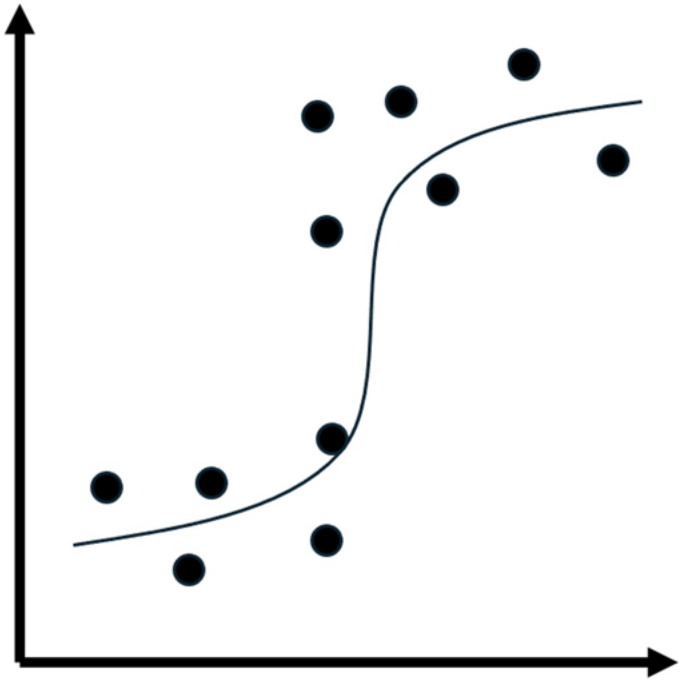


Unlabeled Data [input variable(s)]

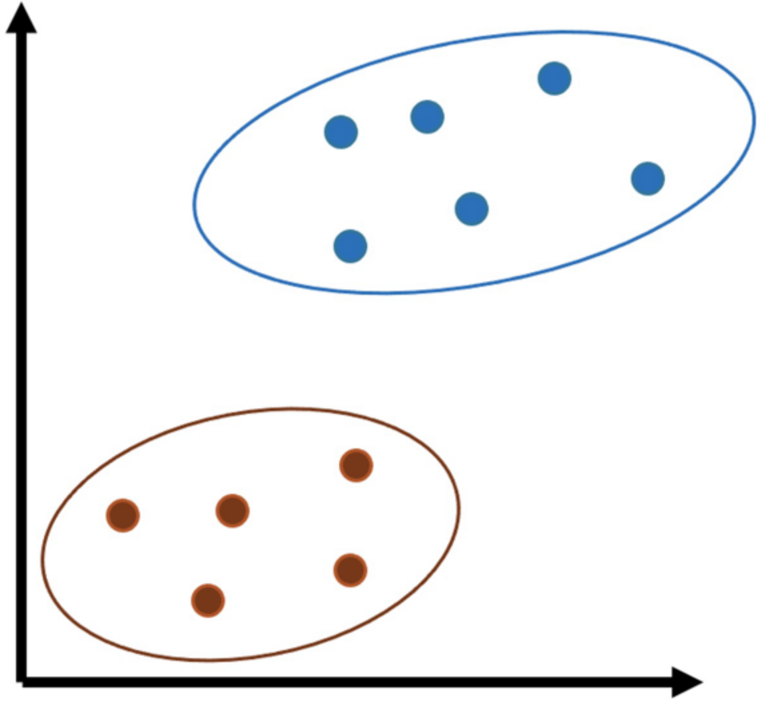
Model Training

Groups (i.e., clusters)

# Module 2: Learning Methods



**Regression**



**Classification**

# Module 2: Learning Methods - Additional Resources

## Introduction to Machine Learning Methods:

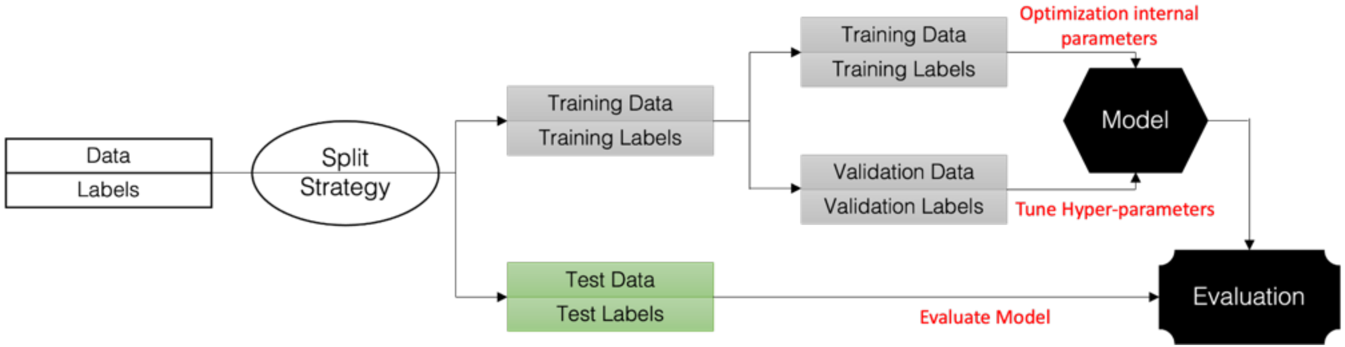
- Chase, R.J., Harrison, D.R., Burke, A., Lackmann, G.M. and McGovern, A., 2022. A machine learning tutorial for operational meteorology. Part I: Traditional machine learning. *Weather and Forecasting*, *37*(8), pp.1509-1529.
- Chase, R.J., Harrison, D.R., Lackmann, G.M. and McGovern, A., 2023. A Machine Learning Tutorial for Operational Meteorology, Part II: Neural Networks and Deep Learning. *Weather and Forecasting*.

## Reviews on Machine Learning Applications in Environmental Sciences:

- Molina, M.J., O'Brien, T.A., Anderson, G., Ashfaq, M., Bennett, K.E., Collins, W.D., Dagon, K., Restrepo, J.M. and Ullrich, P.A., 2023. A Review of Recent and Emerging Machine Learning Applications for Climate Variability and Weather Phenomena. *Artificial Intelligence for the Earth Systems*, pp.1-46.
- de Burgh-Day, C.O. and Leeuwenburg, T., 2023. Machine learning for numerical weather and climate modelling: a review. *Geoscientific Model Development*, *16*(22), pp.6433-6477.

# Module 3: Model Evaluation

- Analysis of Diverse Validation Scenarios in Environmental Modeling
- Evaluation of Learning Metrics for Regression and Classification in Environmental Models
- Calibration Techniques for Environmental Forecasting Models
- Addressing Imbalanced Learning in Environmental Modeling



# Module 3: Model Evaluation - Additional Resources

Raschka, S., 2018. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

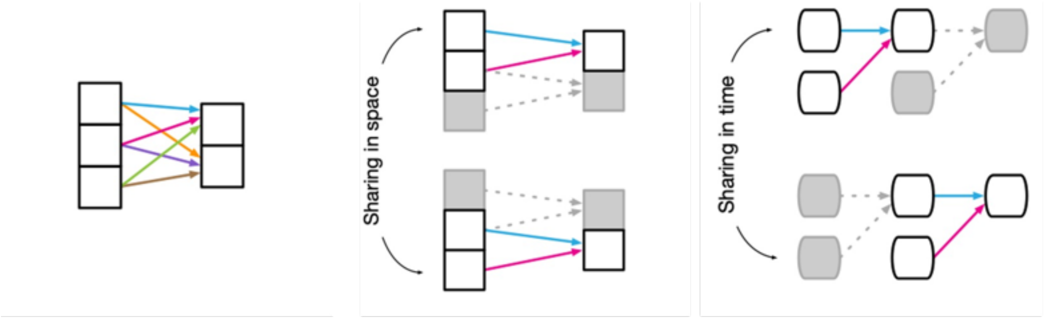
https://towardsdatascience.com/model-selection-and-evaluation-57701ff13c2b

Marzban, C., 2004. The ROC curve and the area under it as performance measures. *Weather and Forecasting*, *19*(6), pp.1106-1114.

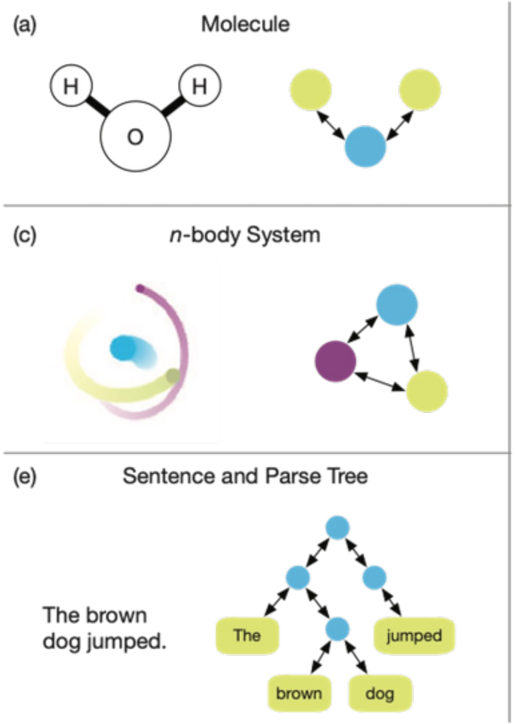
# Module 4: Physics-Informed AI

We can integrate physical knowledge with machine learning in a variety of ways:

- Soft constraints (adding custom terms to our loss functions)
- Hard constraints (changing the network architecture to enforce physical constraints)
- Imposing translation or rotational invariance during the training process
- Model architecture with an inductive bias consistent with system we are modeling
- Hybrid-physics machine learning approaches that integrate both physical knowledge and machine learning algorithms



(a) Fully connected (b) Convolutional (c) Recurrent

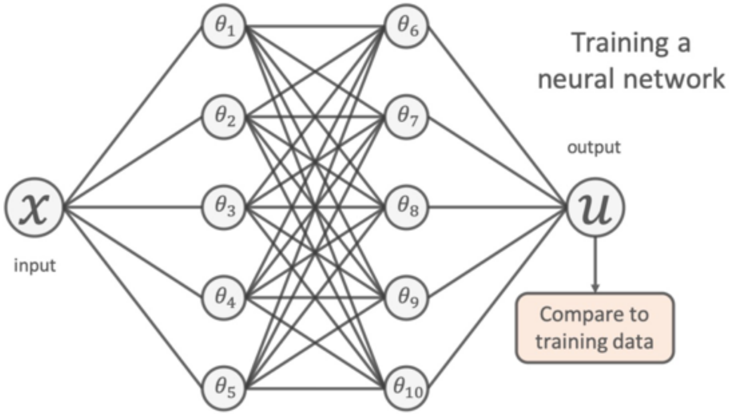


Battaglia et al. 2018

# Module 4: Physics-Informed AI

**Physics Informed Neural Networks (PINN's**) are one example of a hybrid physics machine learning model, where we regularize our loss function with a differential equation.

- PINN's use collocation points calculated during the training process to determine a physics loss term to enforce that the neural network's predictions are physically consistent
- PINN's demonstrate significant improvements over non-physics constrained neural networks in terms of their extrapolation to unseen data regimes
- PINN's can also be used to simultaneously optimize for unknown parameters



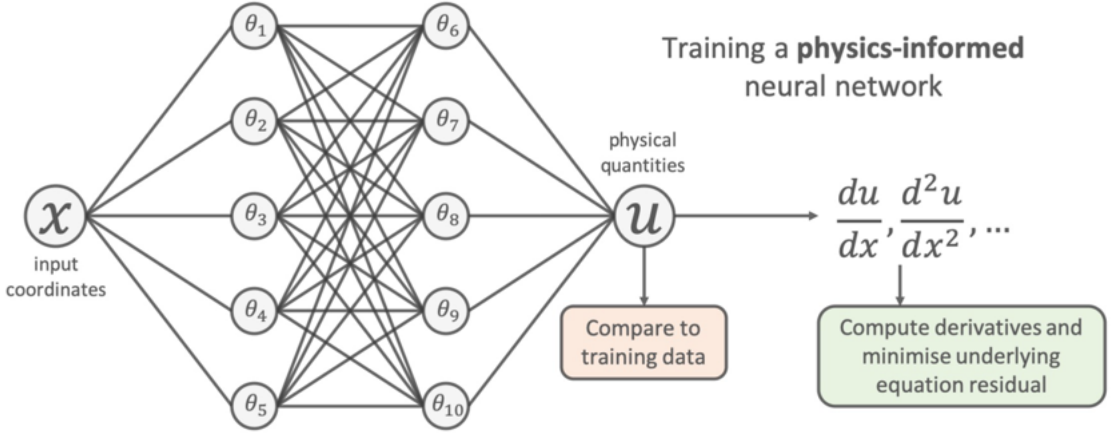


Image credit: B. Moseley

# Module 4: Physics-Informed AI - Additional Resources

Integration of machine learning and physics based knowledge has been a rapidly growing field in the past few years.

**General resources and review papers are given below:**

- Thuerey et al., Physics-based Deep Learning, 2021. https://physicsbaseddeeplearning.org/intro.html
- Cuomo et al. “Scientific Machine Learning through Physics-Informed Neural Networks: Where we are and What’s next” (2022) https://arxiv.org/pdf/2201.05624.pdf
- Karniadakis et al. “Physics-informed machine learning” Nature Reviews Physics, 3, 422-440 (2021). https://www.nature.com/articles/s42254-021-00314-5
- Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, *378*, 686-707.
- Wang et al. “An expert’s guide to training Physics-Informed Neural Networks” https://arxiv.org/abs/2308.08468 (2023)

**In the context of Earth System Science and Climate:**

- Kashinath et al. Physics-informed machine learning: case studies for weather and climate modelling https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0093
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat, F. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195-204.

# Module 5: Explainable AI

- As we use complex ML models more and more for decision-making, we are very interested in what these models learned
- By investigating models, we can potentially:
  - Debug models by identifying error sources
  - Extract novel scientific information by revealing what a high-performing model learned
- However, XAI is still a developing field and there are many pitfalls
- It is quite easy to apply XAI methods, but get disagreement among explanations
- Care should be taken when using XAI:
  - Understand what the XAI method reveals about the model
  - Apply multiple methods and identify commonalities
  - Use repetitions to achieve statistical significance
  - Characterize the correlations in your data, use methods that take them into account

# Module 5: Explainable AI - Additional Resources

- **Molnar's *Interpretable Machine Learning* book**
  https://christophm.github.io/interpretable-ml-book/

- **AI2ES/CIRA Short Course on Explainable Artificial Intelligence for Environmental Science**
  https://docs.google.com/document/d/1lqpABwDl3kPe6ThE-NIDR64PimnltJEuKNkysDZuWKQ/edit

- **A Machine Learning Explainability Tutorial for Atmospheric Sciences (Flora et al., 2024)**
  https://journals.ametsoc.org/view/journals/aies/3/1/AIES-D-23-0018.1.xml

- **Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning (McGovern et al., 2019)**
  https://journals.ametsoc.org/view/journals/bams/100/11/bams-d-18-0195.1.xml

## Challenges & Strategies

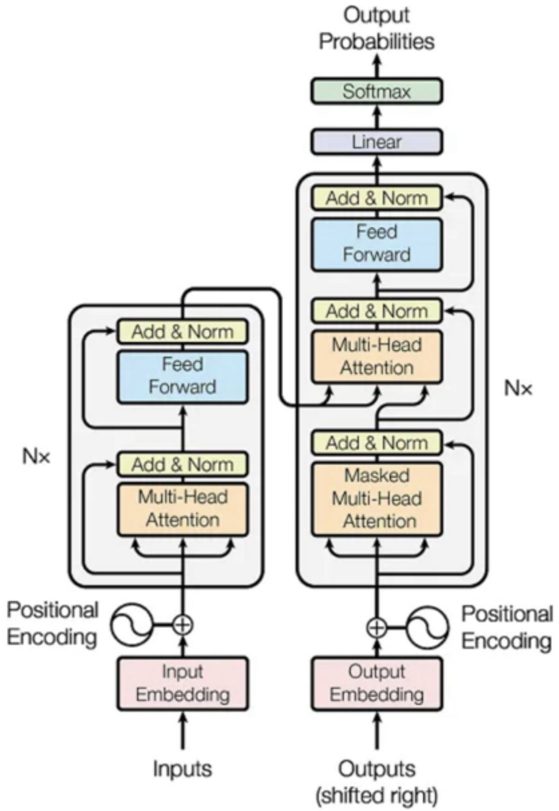
- **General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models (Molnar et al., 2022)**
  https://link.springer.com/chapter/10.1007/978-3-031-04083-2_4

- **Aggregation strategies to improve XAI for geoscience models that use correlated, high-dimensional rasters (Krell et al., 2024)**
  https://www.cambridge.org/core/journals/environmental-data-science/article/aggregation-strategies-to-improve-xai-for-geoscience-models-that-use-correlated-highdimensional-rasters/F6017A23BEF0BD48969225D68DF819A2

- **Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities (Saeed & Omlin, 2023)**
  https://www.sciencedirect.com/science/article/pii/S0950705123000230

# Module 6: Transformers

- An Introduction to Transformers: Exploring the Next Generation of Deep Learning Models
- Comparing Transformers and CNNs: A Deep Dive into Model Architectures.



# Module 6: Transformers - Additional Resources

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.

Cordonnier, J.B., Loukas, A. and Jaggi, M., 2019. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*.

# AMS AI STAC Committee on AI Applications - AMS Meeting Events

- We have a core science keynote presenter **Monday afternoon at 2:30pm** given by Srinivasan Parthasarathy titled - in-the-Loop Explainable AI : A Case Study in Flood Response.

- We have our second core science keynote presenter in the last session on **Monday at 4:30pm** given by Imme Ebert-Uphoff on A Research Agenda for the Evaluation of AI-Based Weather Forecasting Models.

- **On Monday evening starting at 6:30pm**, we invite you all to join a large joint mixer hosted by AI, HPC, Python, and ProbStat after the last sessions conclude. It will be hosted at the Key Ballroom (salons 3-4) of the Hilton Baltimore Inner Harbour.

- Later this week, we would like to encourage you to attend the presidential session “Bridging Weather and Climate: Advancing on All Fronts” on **Wednesday at 10:45am** where Mike Pritchard will be participating as a chair and providing expertise on the evolving state of data driven AI models.

- Lastly, there is a special collection being hosted by ProbStat on “Advances in and applications of statistical and machine-learning methods in weather, climate, and hydrology.” Papers can be submitted to Weather and Forecasting, Hydrometeorology, Climate, Monthly Weather Review, and Artificial Intelligence for the Earth Systems. Please reach out to Yu Zhang, Tara Jensen, Julia Jeworrek, and/or Andy Poppick with questions if interested.

# Thanks for your attention and have a great meeting!

AMS
American Meteorological Society