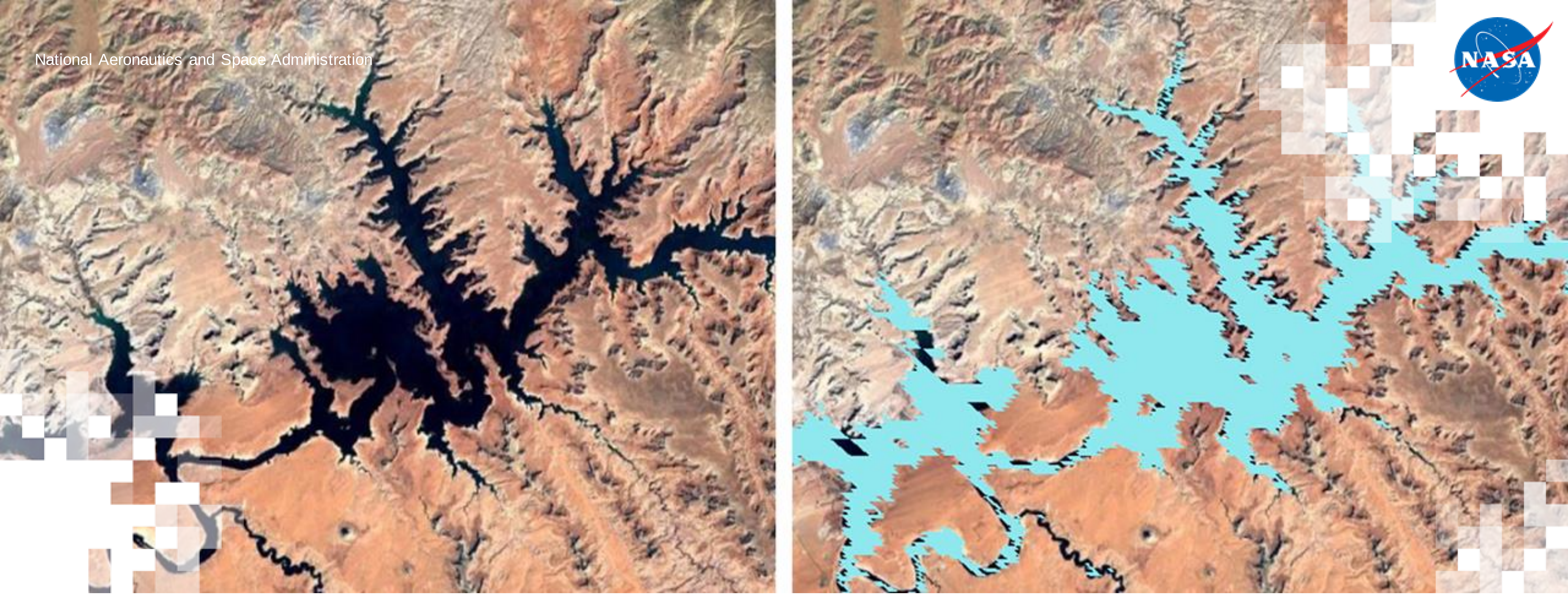


National Aeronautics and Space Administration

Fundamentals of Machine Learning for Earth Science

Overview of Machine Learning

Jordan A. Caraballo-Vega, Caleb S. Spradlin
Data Science Group, 606.3

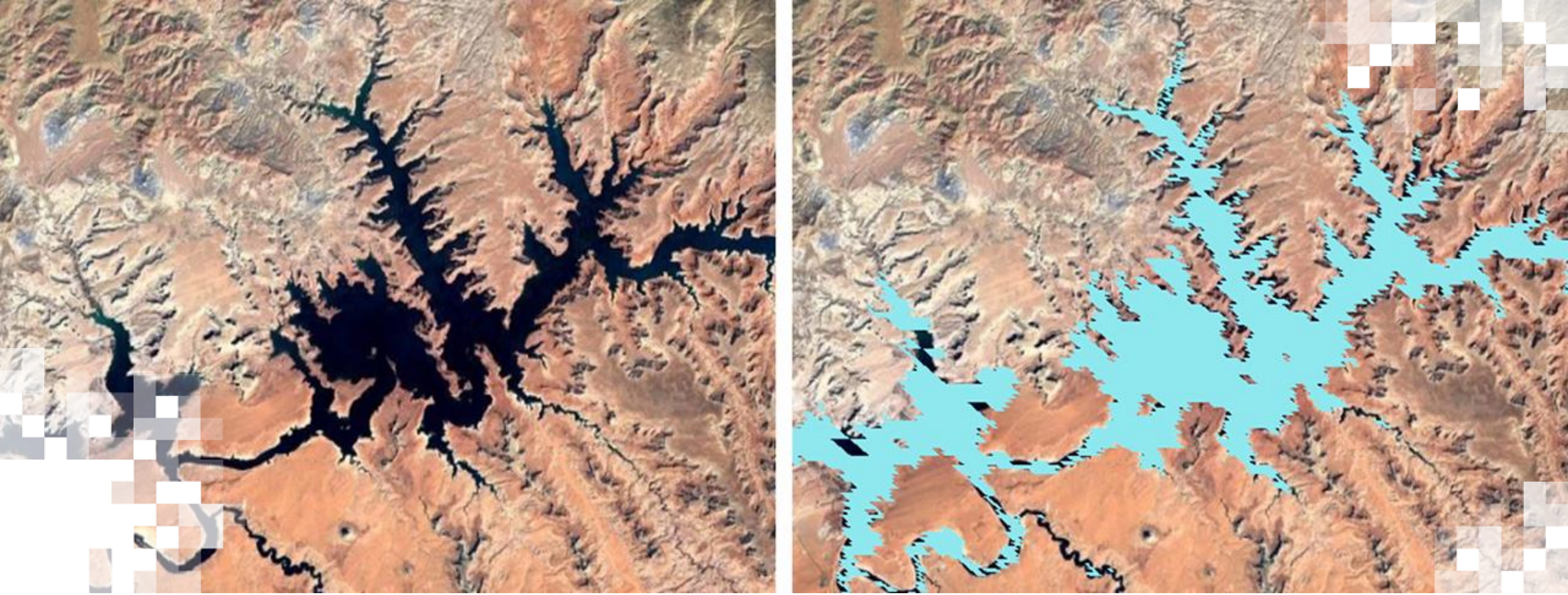
June 29, 2023

Tech Talk Outline

- Overview of Machine Learning
- Importance of Machine Learning targeted towards Earth Science
- Usability of Machine Learning
- Software and hardware to support Machine Learning
- Hands on Jupyter Notebook Exercise: MODIS Water Classification Case Study
- Overview of model explainability and interpretability
- Hands on Jupyter Notebook Exercise: MODIS Water Classification XAI
- Q&A Session

Resources for this Training

https://github.com/NASAARSET/ARSET_ML_Fundamentals



Overview and Theory

Overview of Machine Learning

The following quote from *Arthur Samuel* describes what Machine Learning (ML) is:

> *"Machine learning enables a machine to* ***automatically learn from data,***
> ***improve performance from experiences,*** *and*
> ***predict things without being explicitly programmed.****"*

ML uses techniques from Statistics, Mathematics, and Computer Science to make computer programs learn from data to predict an output.

How does Machine Learning Work?

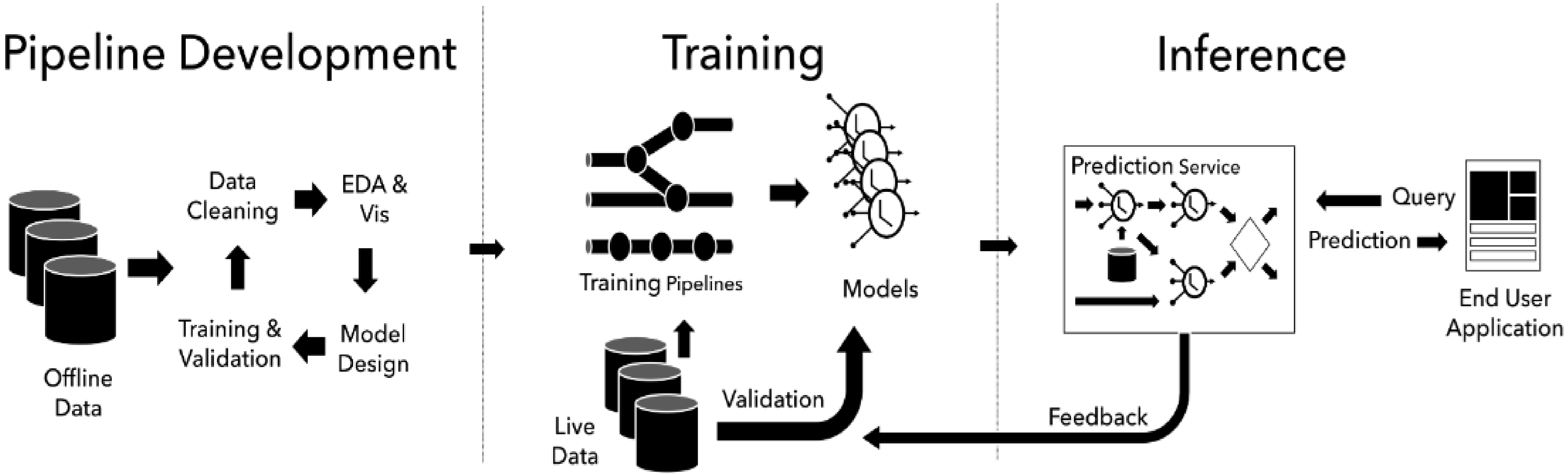


Image Source: Daniel Crankshaw (in a Short History of Prediction-Serving Systems)

Machine Learning Algorithms

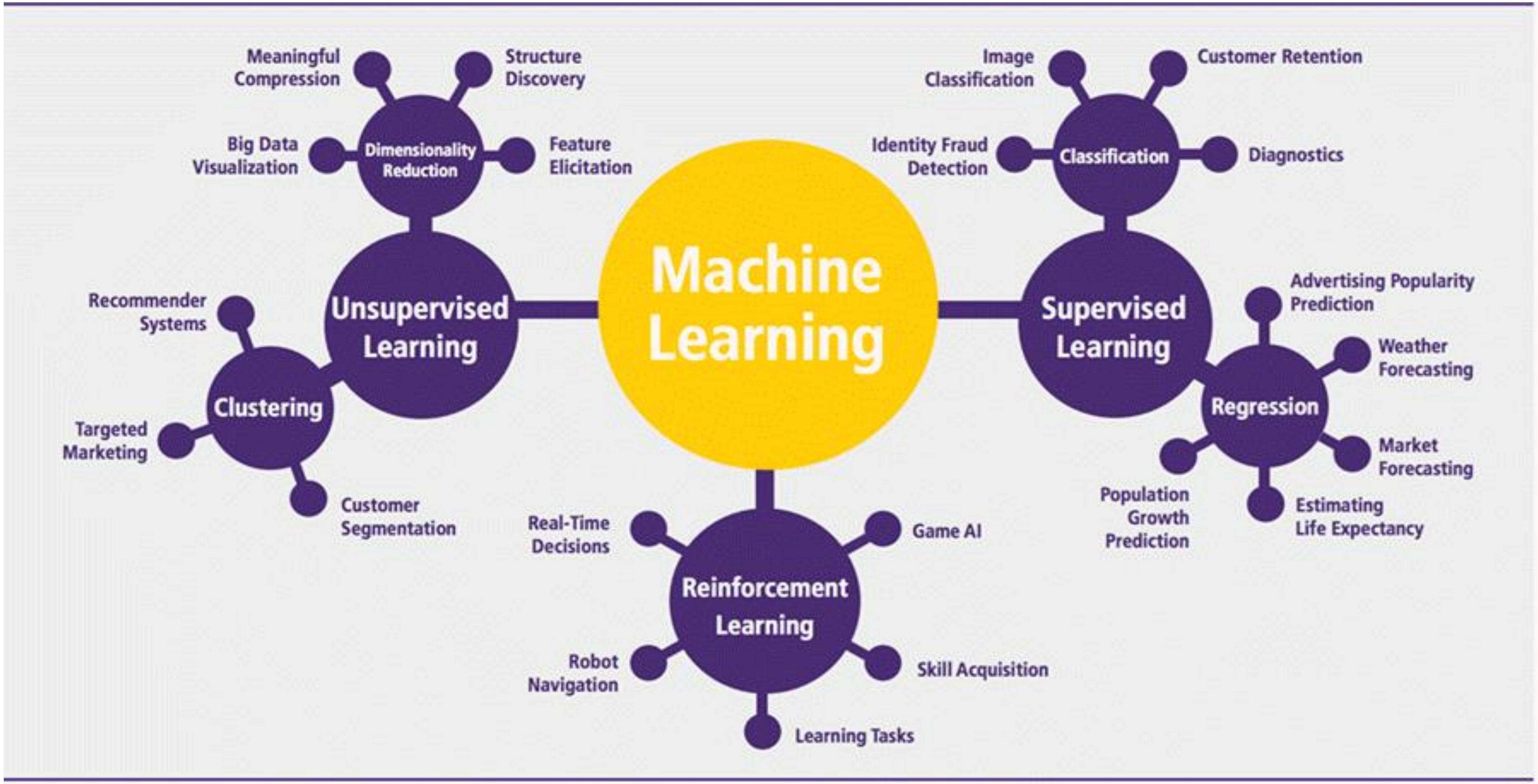
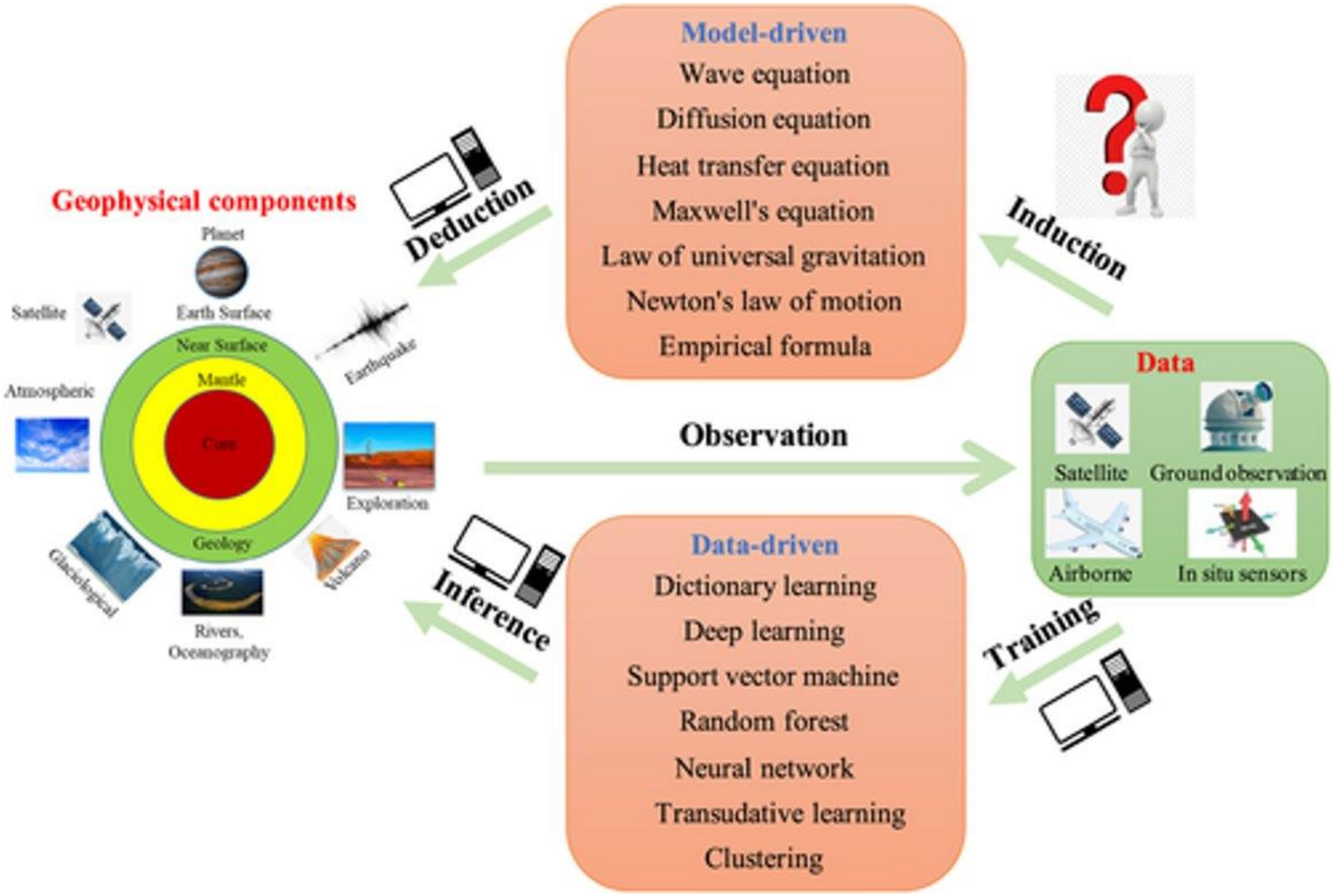


Image Source: guru99.com

How Machine Learning is Applied in Earth Science



Siwei Yu and Jianwei Ma (2021), https://doi.org/10.1029/2021RG000742

Machine Learning Frameworks in Python

- Python-based tools dominate the machine learning frameworks based on *Kaggle's 2021 State of Data Science and Machine Learning survey.*
- Scikit-learn is the top with over 80% of data scientists using it.
- TensorFlow and Keras were each chosen by about half of the data scientists for deep learning.
- Gradient boosting library XGBoost is fourth.

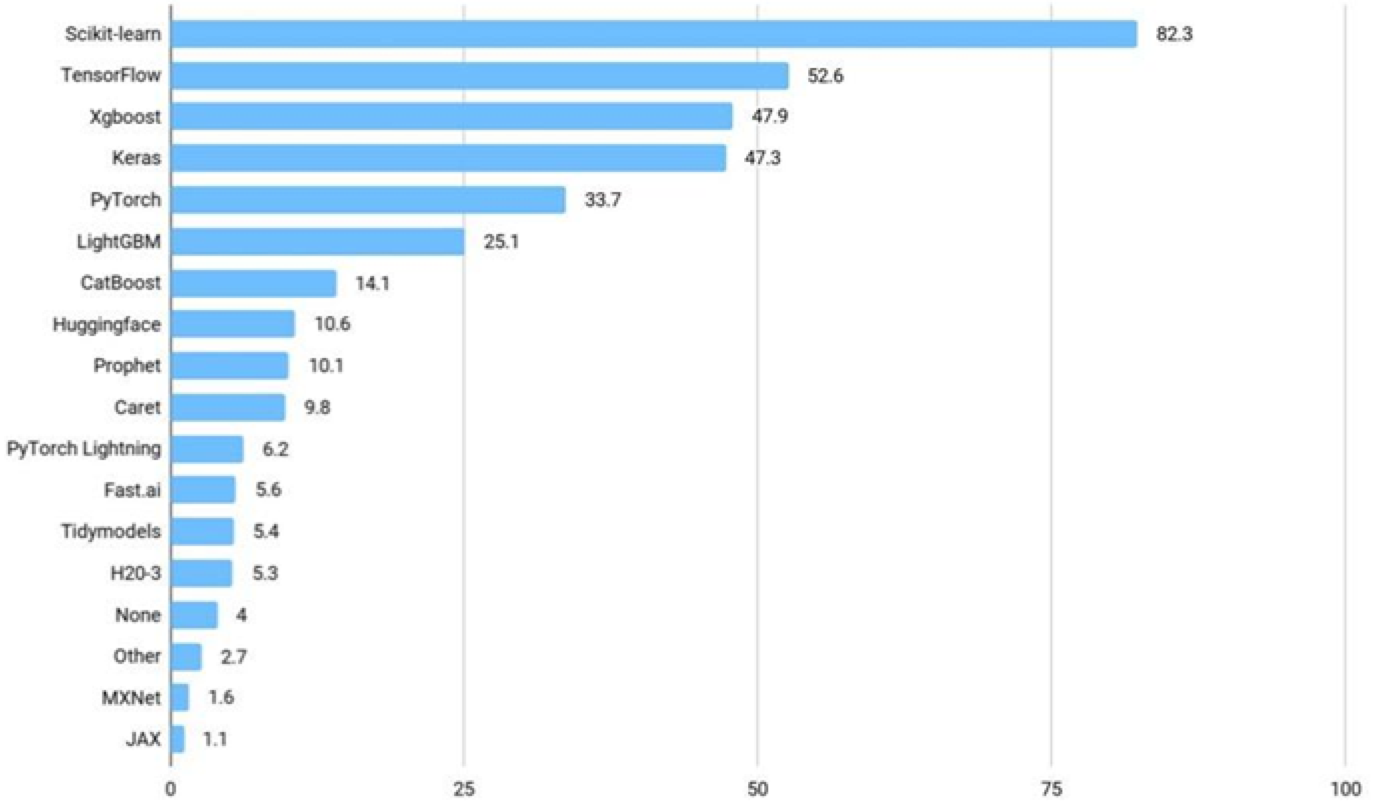


Image Source: Kaggle's 2021 State of Data Science and Machine Learning survey

Graphics Processing Unit (GPU) Role in Machine Learning

- There are many available platforms for parallel computing and programming. Out of them, **CUDA** (by NVIDIA) is the most popular platform due to the following reasons:
  - CUDA runs on both Windows and Linux.
  - Almost all the GPU-supported Python libraries like CatBoost, TensorFlow, Keras, PyTorch, OpenCV, and CuPy were designed to run on NVIDIA CUDA-enabled graphics cards.

- **Popular GPU-Supported Python Libraries:**
  - XGBoost
  - OpenCV
  - cuML (Part of RAPIDS)
  - cuDF (Part of RAPIDS)
  - CuPy (NumPy for GPU)

Prism – On-premises at the NCCS

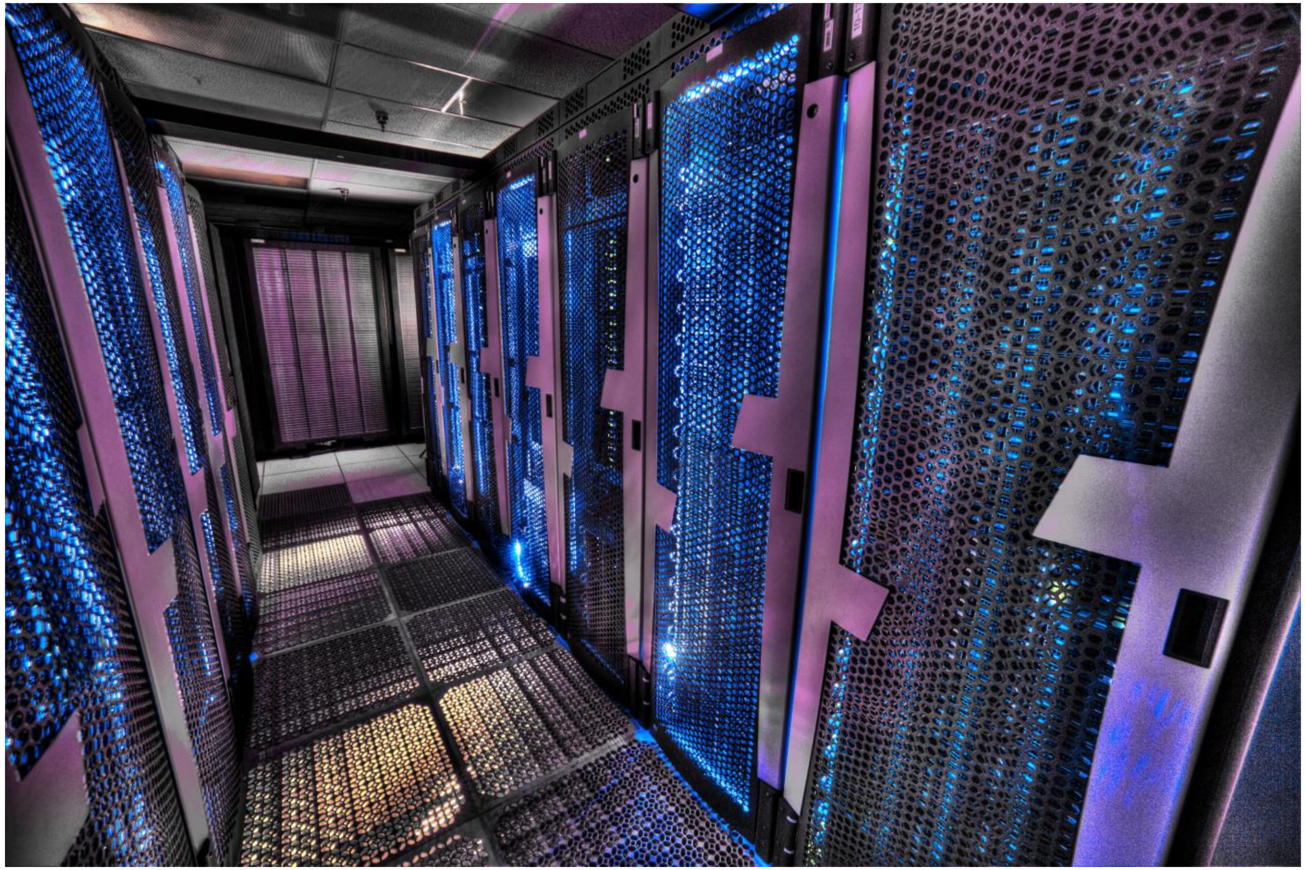
- 22 GPU Compute Nodes
- 4x NVIDIA V100 GPUs with 32 GB of VRAM and NVLink
- Dual Intel Xeon Cascade Lake Gold 6248 CPUs
- 20 cores each at 2.50GHz
- 768 GB RAM
- Dual 25Gb Ethernet network interfaces
- Dual 100Gb HDR100 Infiniband high speed network interfaces
- 3.8 TB RAID protected NVMe drives, mounted as /lscratch

Single DGX

- 8x NVIDIA A100 GPUs with 40 GB of VRAM and NVLink
- Dual AMD EPYC Rome 7742 CPUs
- 64 cores each at 2.25GHz
- 1 TB RAM
- Dual 25Gb Ethernet network interfaces
- Dual 100Gb HDR100 Infiniband high speed network interfaces
- 14 TB RAID protected NVMe drives, mounted as /lscratch

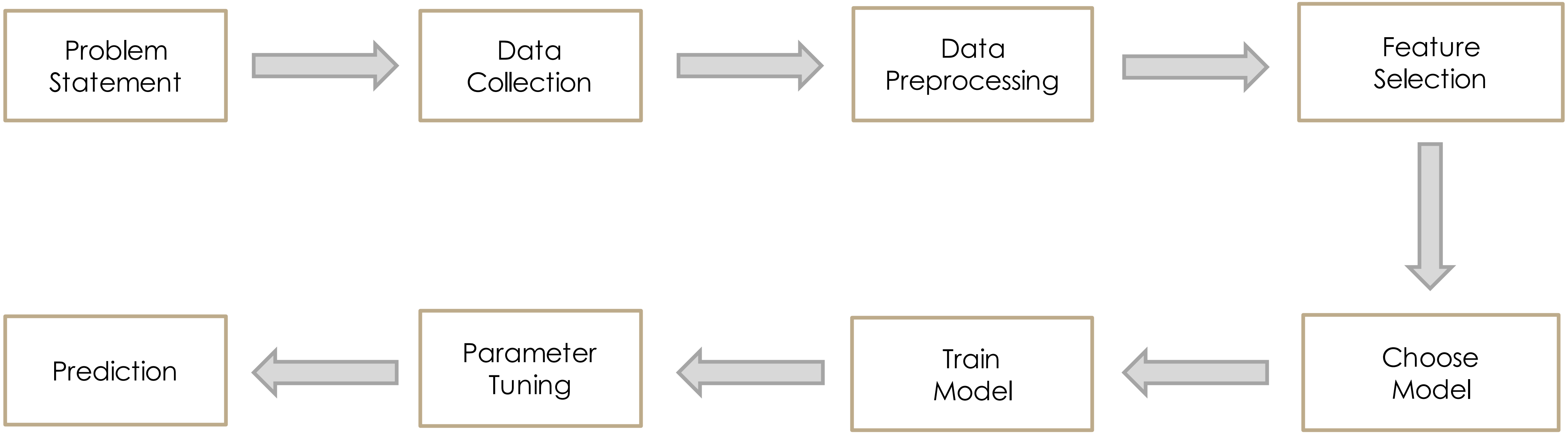
Discover – GPUs in HPC

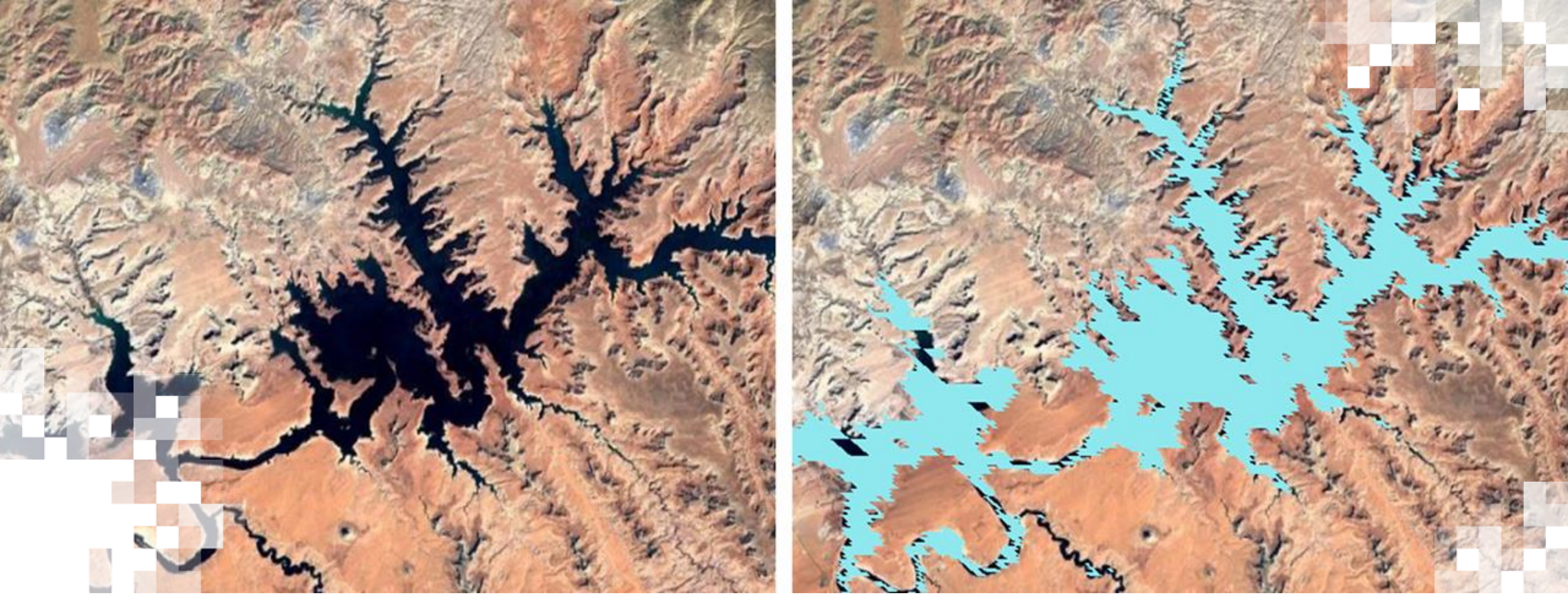
- 12 Supermicro GPU nodes
- 4x NVIDIA A100 GPUs with 40GB of VRAM and NVLINK
- Dual AMD EPYC
- 24 cores each at 2.89 GHz
- Rome 48-core nodes
- 512 GB of RAM
- Dual 25Gb Ethernet network interfaces
- Dual 100Gb HDR100 Infiniband high speed network interfaces
- No Swap Space



Machine Learning Steps

- Depending on the scale of your science problem, each one of these steps can be done using Jupyter notebooks.
- Large scale applications and long running deep learning models can benefit from PRISM Slurm submissions using Singularity containers.

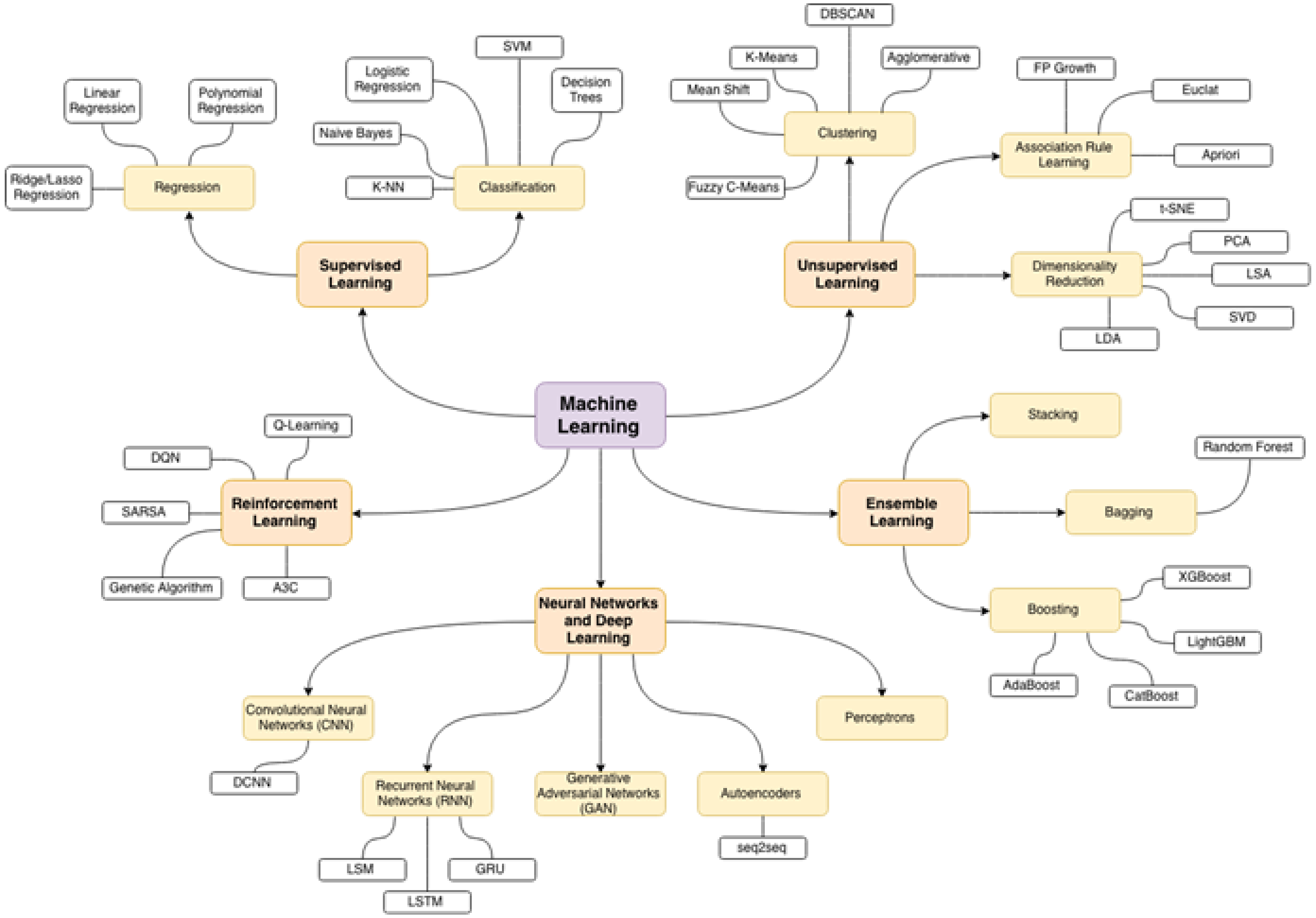




Overview of Machine Learning Algorithms

Machine Learning Algorithms: Which algorithm to choose?

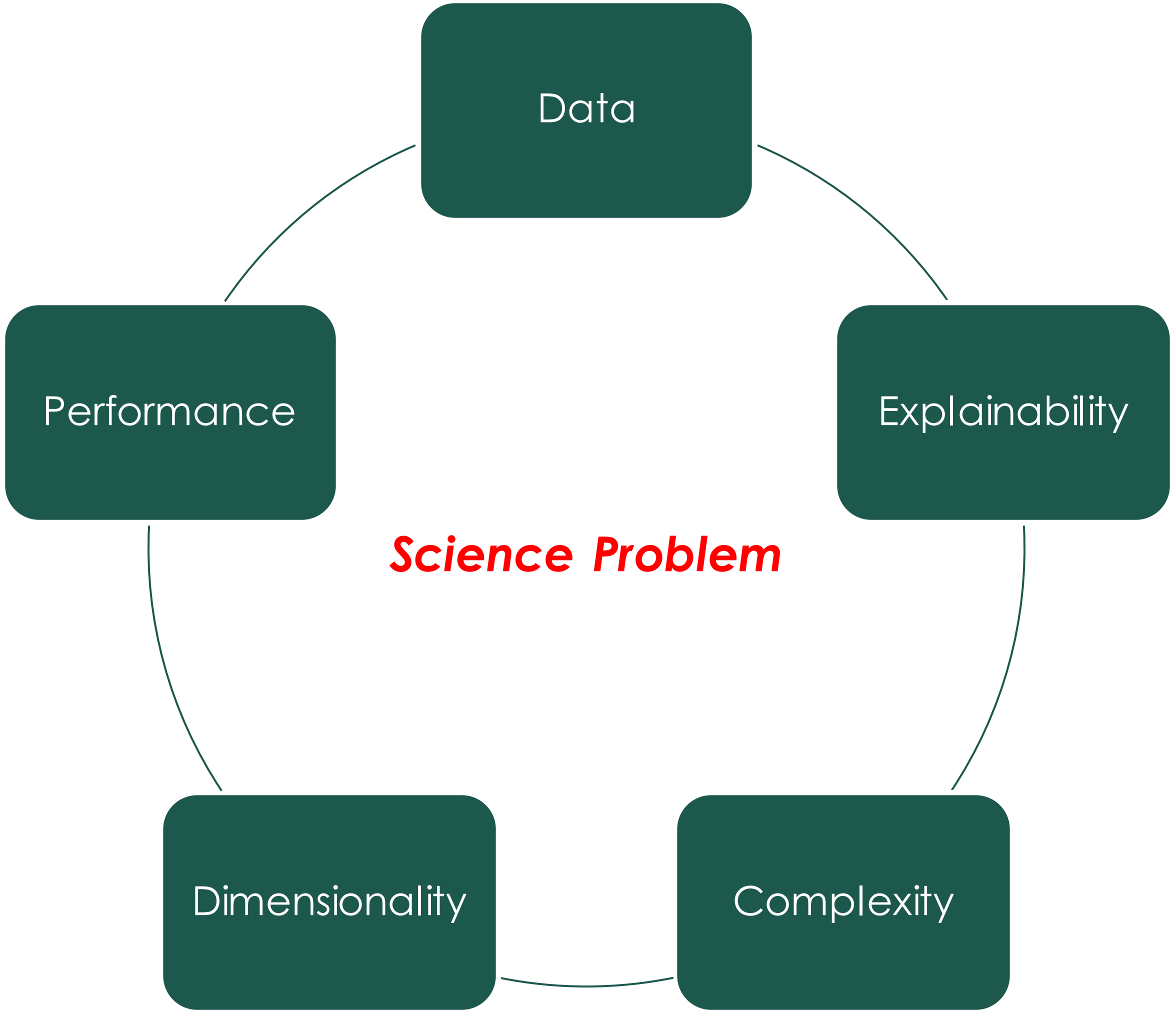
- The development of machine learning algorithms has been exponentially increasing.
- We will not dive into the specifics of each algorithm, but we will give you the tools to aid in the selection of these for your own science problem(s).
- You are not bounded to a single algorithm, but it always saves time to start from a logical base.



Core Machine Learning Algorithms.
Image Source: github.com

Machine Learning Algorithms: Science Problem

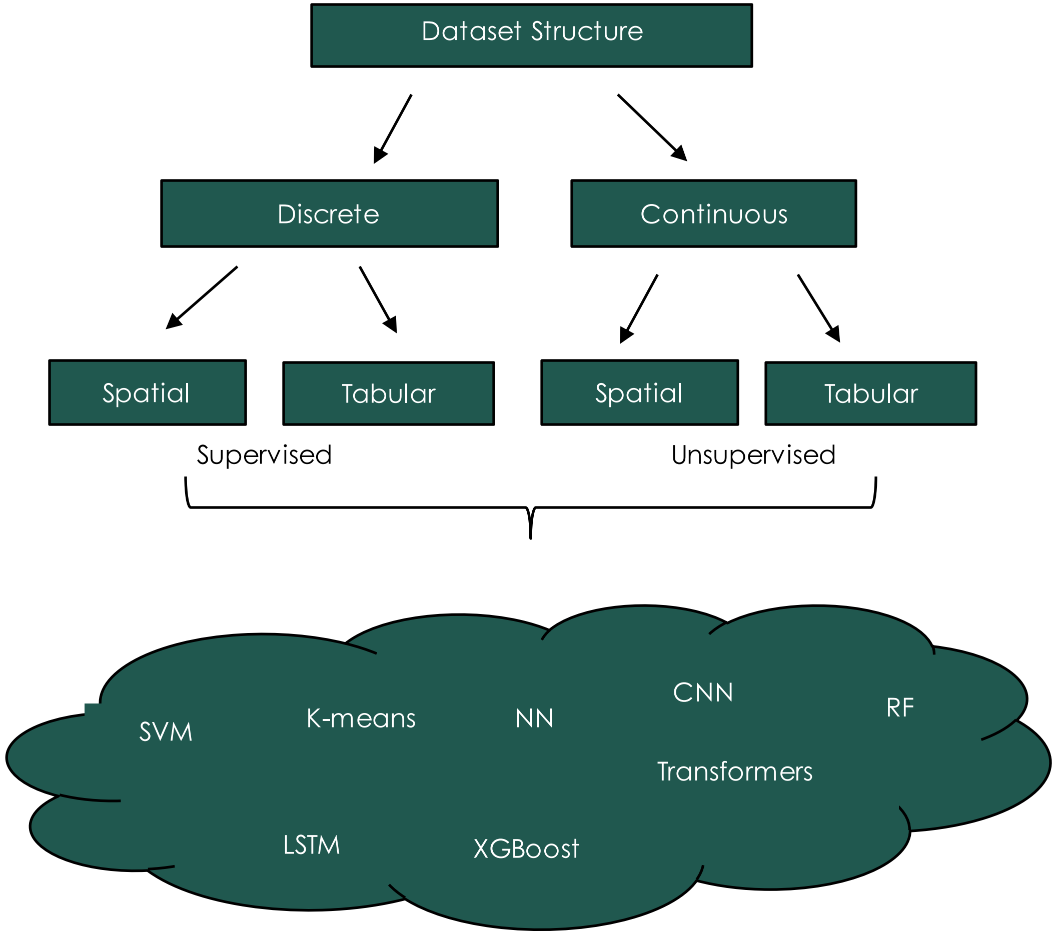
- **Which scientific question would you like to address?** We want to identify the sign, magnitude, and potential drivers of change in surface water extent in X study area.
- **What information is missing to answer this question?** We need surface water extent maps to quantify and analyze these drivers.



Components to aid the selection of your ML algorithm.

Machine Learning Algorithms: Data

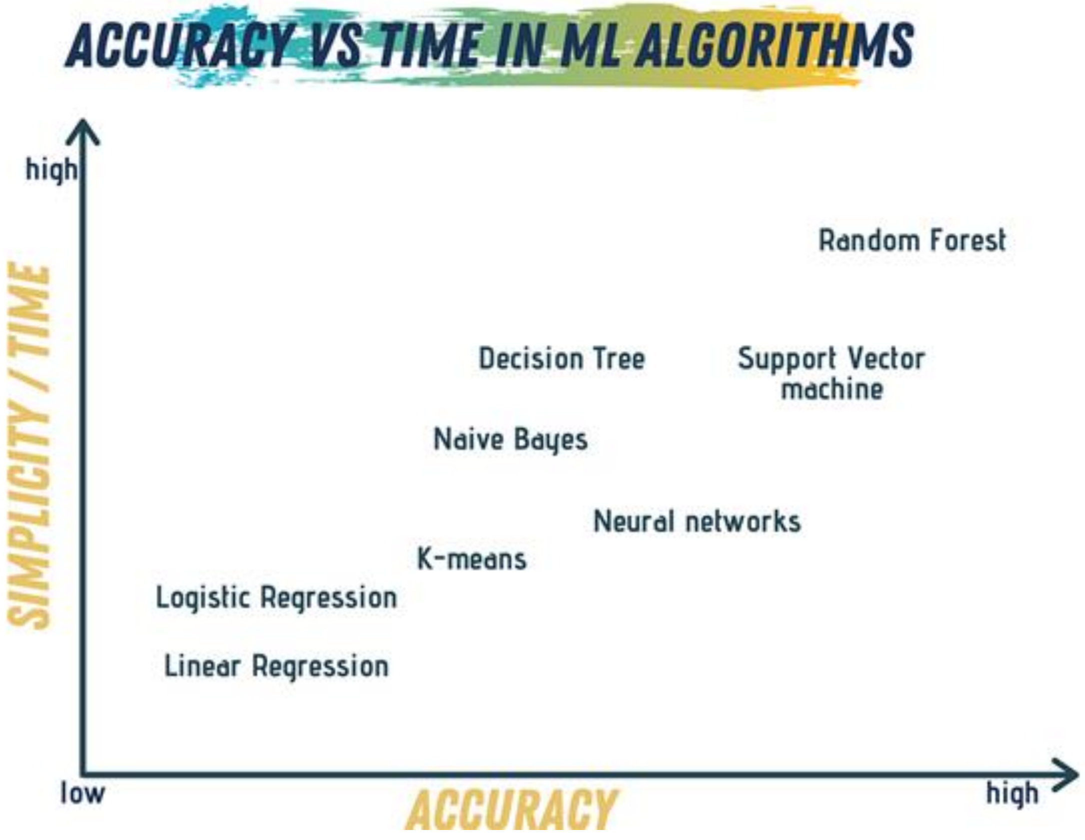
- **What data do you have available?** We have global coverage with data from the MODIS satellite.
- **Do you have training data available?** We have gathered large extents of training data points.
- **What is the data structure of your data?** Our data is in raster format. We can preprocess it to make it tabular.
- **Is your dependent variable a continuous or discrete problem?** Our dependent variable is water pixels, which is discrete (0 – no water, 1 – water)



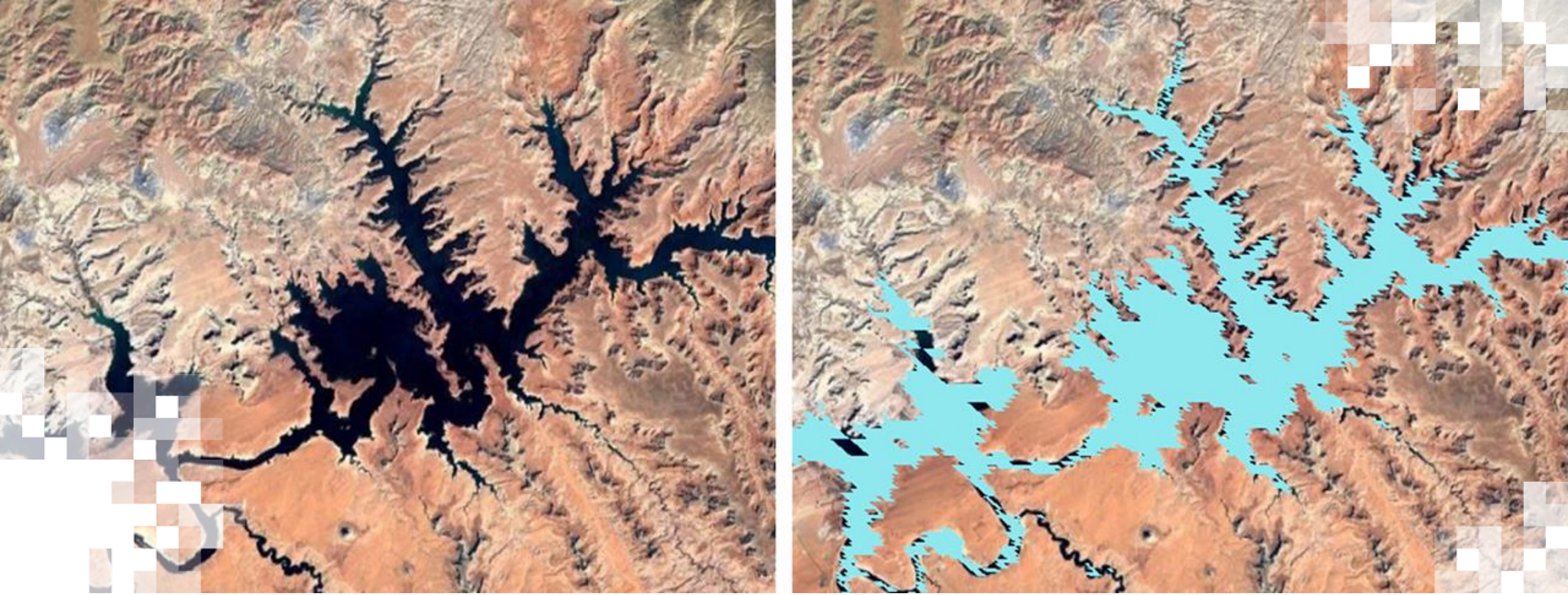
Algorithm decision branch based on data structure.

Machine Learning Algorithms: Performance

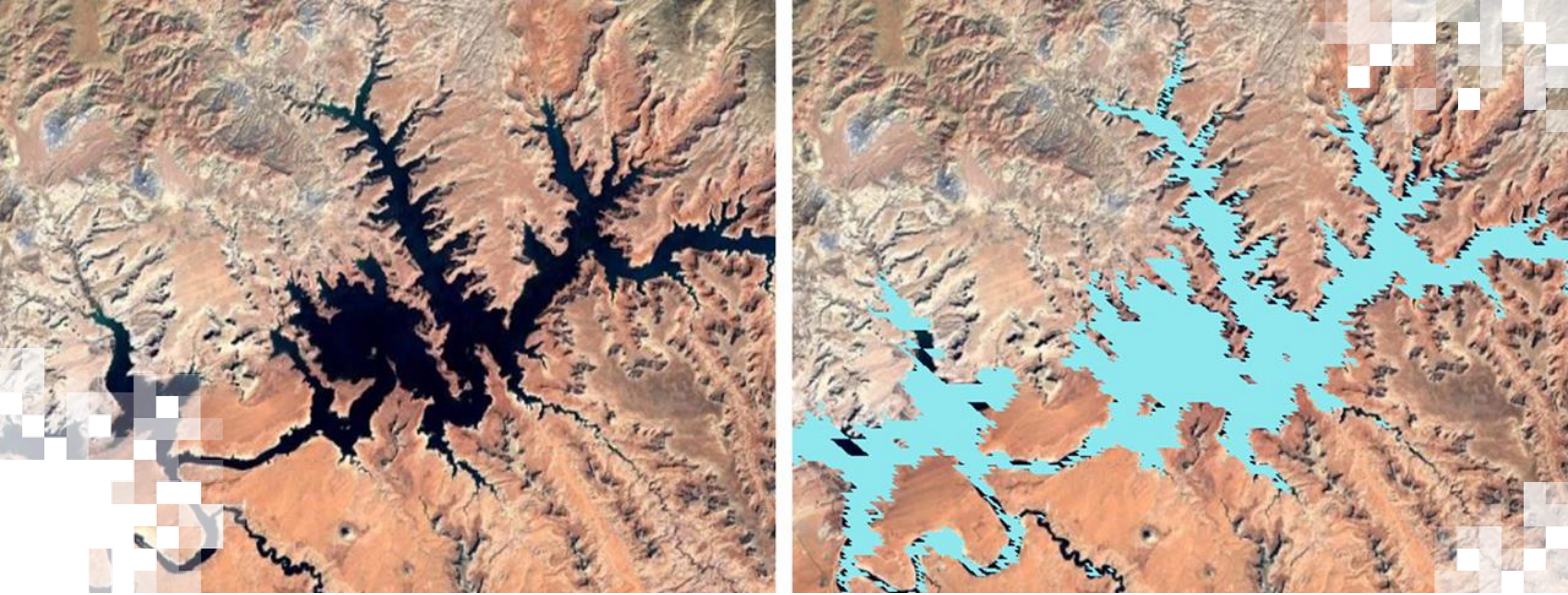
- **Are there any performance requirements based on your science question (e.g., real time vs. static)?** We do not need real time maps (e.g., disaster response teams might need results quickly).
- **Is your software going to run on on-premise, cloud, or embedded hardware?** We want our software to run both on-premise and in the cloud.
- **What is more important for your project: inference time or model performance?** We care more about model performance than inference time.



Tradeoff between speed and accuracy.
Image Source: github.com



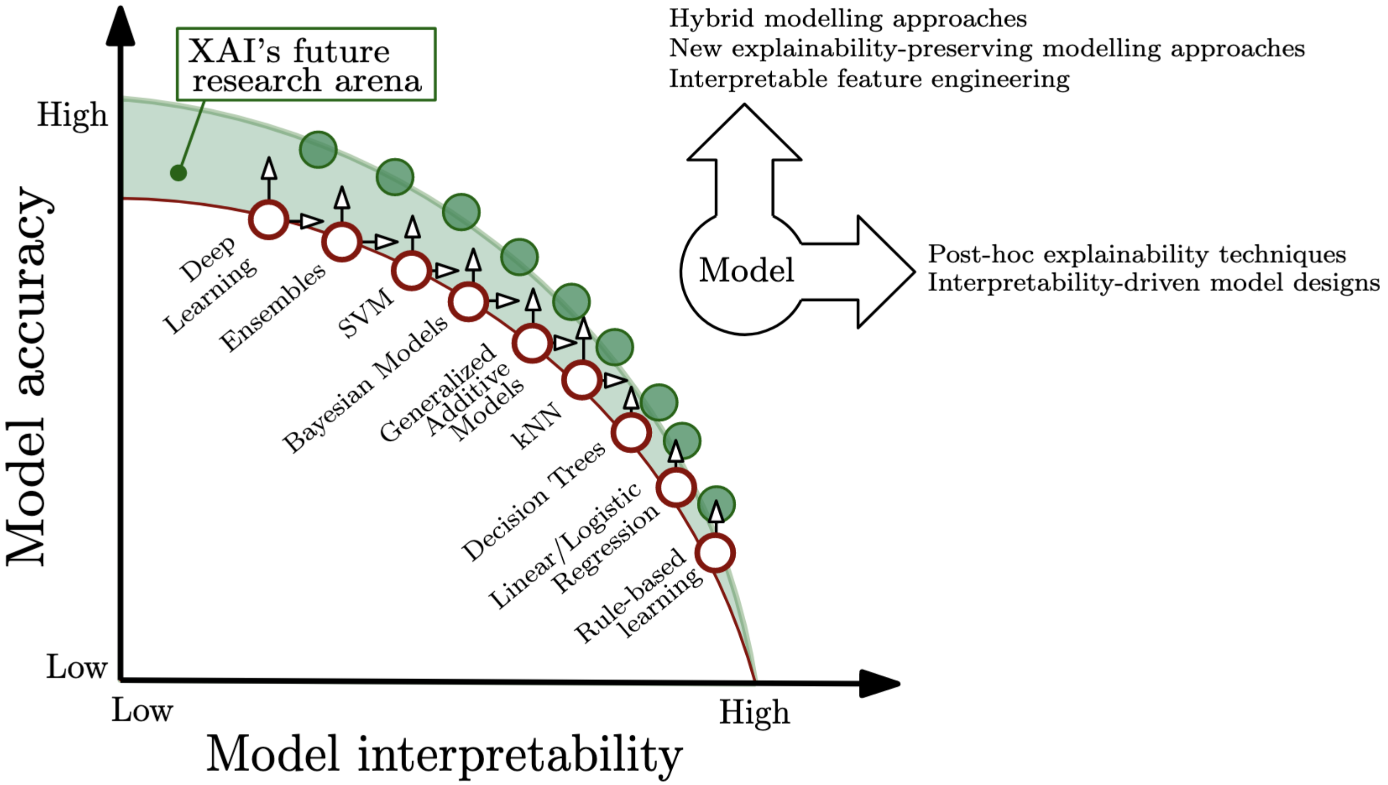
Exercise: Training and Testing of XGBoost Model in JupyterLab



Model Explainability and Interpretability - XAI

Model Explainability and Interpretability – XAI

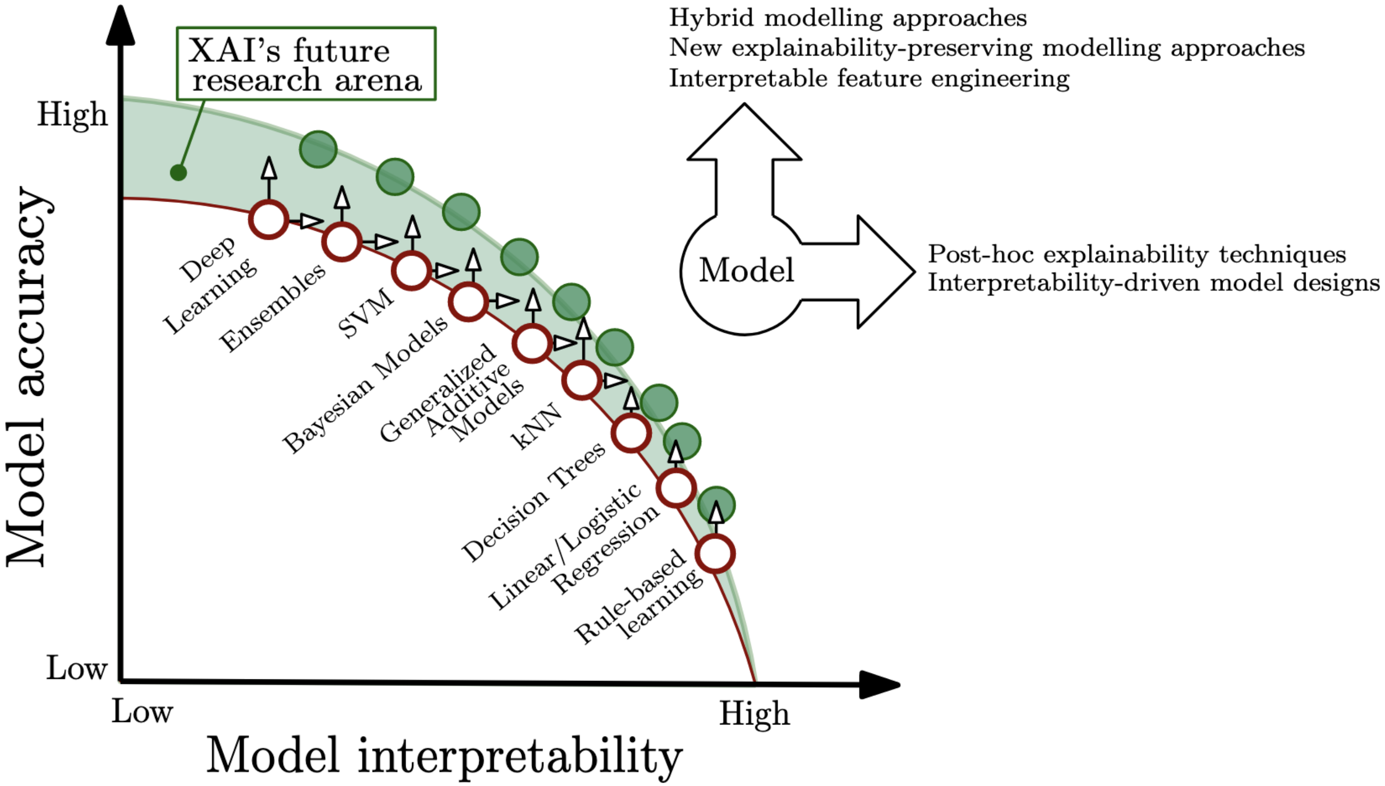
As we come to rely on inferences given by machine learning models, it is important that these models be accurate and interpretable.


XAI's future research arena
High
Low
Model accuracy
Low
High
Model interpretability
Deep Learning
Ensembles
SVM
Bayesian Models
Generalized Additive Models
kNN
Decision Trees
Linear/Logistic Regression
Rule-based learning
Hybrid modelling approaches
New explainability-preserving modelling approaches
Interpretable feature engineering
Model
Post-hoc explainability techniques
Interpretability-driven model designs

Arrieta et al. (2019), https://doi.org/10.3389/fnsys.2021.766980

Why We Need Reliable Models? – XAI

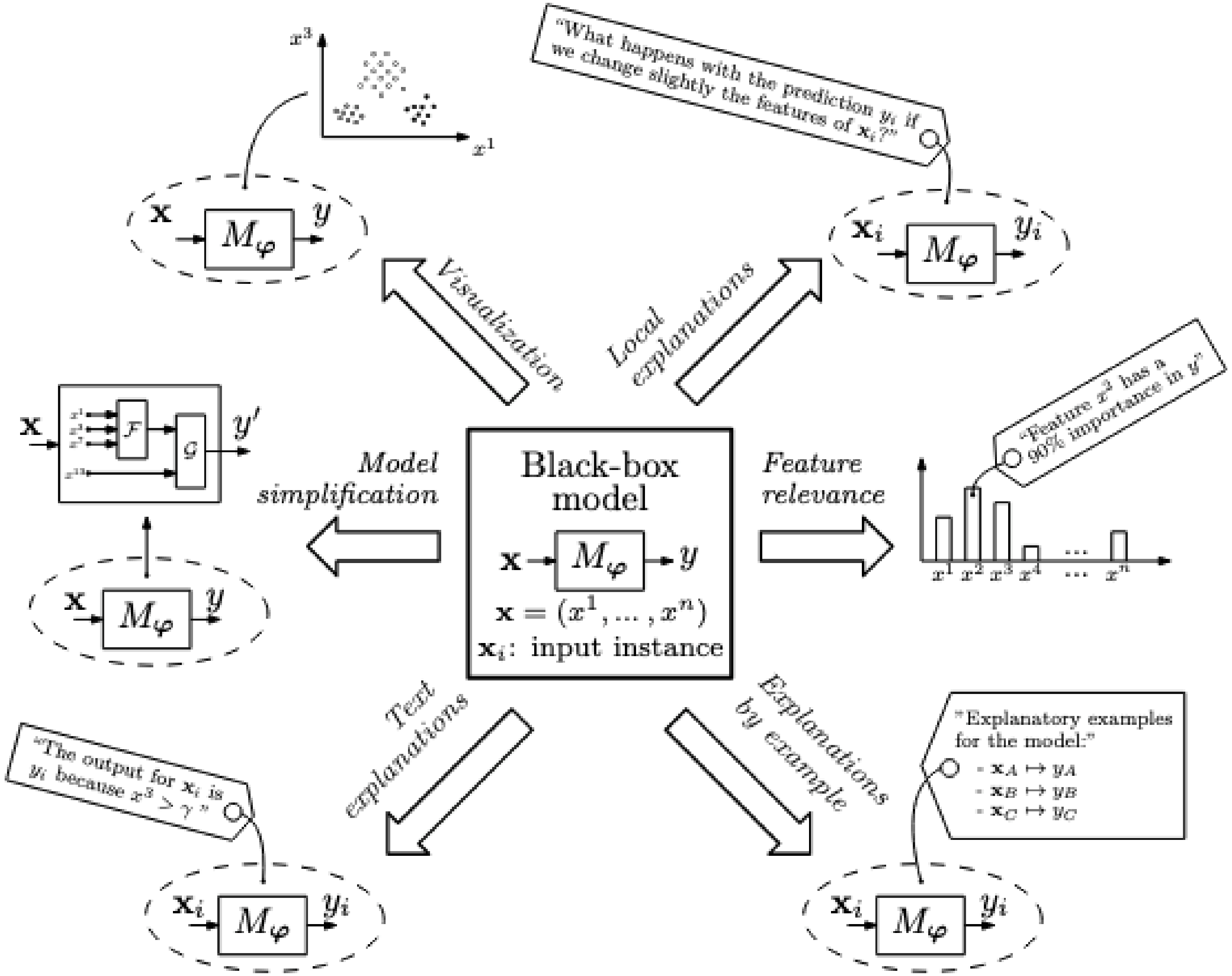
- Accuracy may not be enough.
- Machine learning models need to be reliable.
- Reliability is determined by interpretability and robustness.
- Interpretability: We can explain why a certain outcome was predicted.
- Robustness: Input can be noisy; we still achieve accurate predictions.


XAI's future research arena
High
Low
Model accuracy
Deep Learning
Ensembles
SVM
Bayesian Models
Generalized Additive Models
kNN
Decision Trees
Linear/Logistic Regression
Rule-based learning
Low
High
Model interpretability
Hybrid modelling approaches
New explainability-preserving modelling approaches
Interpretable feature engineering
Model
Post-hoc explainability techniques
Interpretability-driven model designs

Arrieta et al. (2019), https://doi.org/10.3389/fnsys.2021.766980

Post-Hoc Explainability Approaches – XAI

- One of the most common methods of achieving an interpretable ML model is through post-hoc explanation methods (done after the model is trained).
- These methods use the output of the model in conjunction with the inputs to extract information about the model's decisions.



x^3
x^1
"What happens with the prediction y_i if we change slightly the features of $\mathbf{x}_i$?"
$\mathbf{x}$ M_φ y
$\mathbf{x}_i$ M_φ y_i
Visualization
Local explanations
$\mathbf{x}$ $\mathcal{F}$ $\mathcal{G}$ y'
Model simplification
Black-box model
$\mathbf{x}$ M_φ y
$\mathbf{x} = (x^1, \ldots, x^n)$
$\mathbf{x}_i$: input instance
Feature relevance
"Feature x^2 has a 90% importance in y"
x^1 x^2 x^3 x^4 ... x^n
Text explanations
Explanations by example
"The output for $\mathbf{x}_i$ is y_i because $x^3 > \gamma$"
"Explanatory examples for the model:"
- $\mathbf{x}_A \mapsto y_A$
- $\mathbf{x}_B \mapsto y_B$
- $\mathbf{x}_C \mapsto y_C$

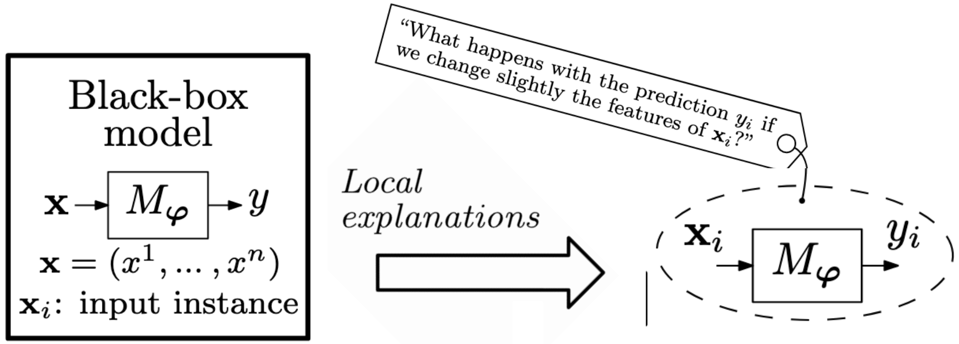
Arrieta et al. (2019), https://doi.org/10.3389/fnsys.2021.766980

Model Explainability and Interpretability – XAI

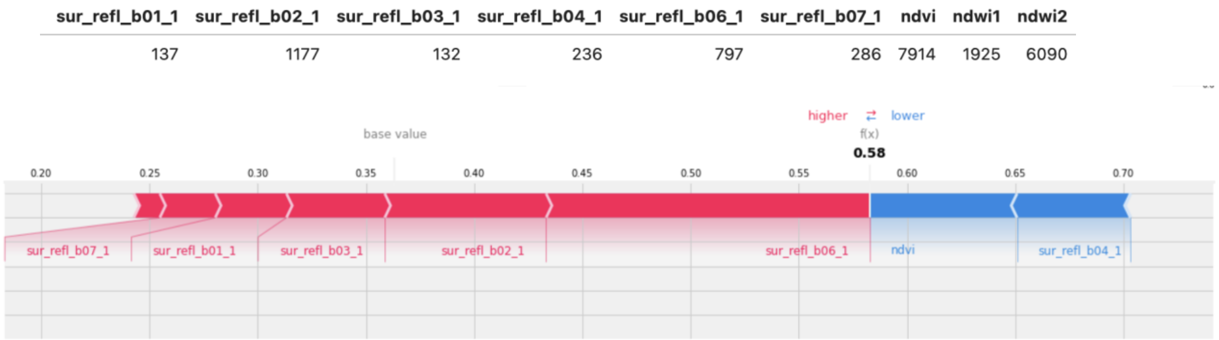
- A tool used commonly is SHAP (SHapley Additive exPlanations).
- SHAP is a model-agnostic approach which can calculate an additive feature importance score for each prediction.

Using SHAP Values for Local Explanations

- Using Shapely values to provide explanations of single decisions for black box models


Black-box model
$\mathbf{x} \rightarrow M_\varphi \rightarrow y$
$\mathbf{x} = (x^1, \dots, x^n)$
$\mathbf{x}_i$: input instance
Local explanations
"What happens with the prediction y_i if we change slightly the features of $\mathbf{x}_i$?"
$\mathbf{x}_i \rightarrow M_\varphi \rightarrow y_i$

Arrieta et al, 2019

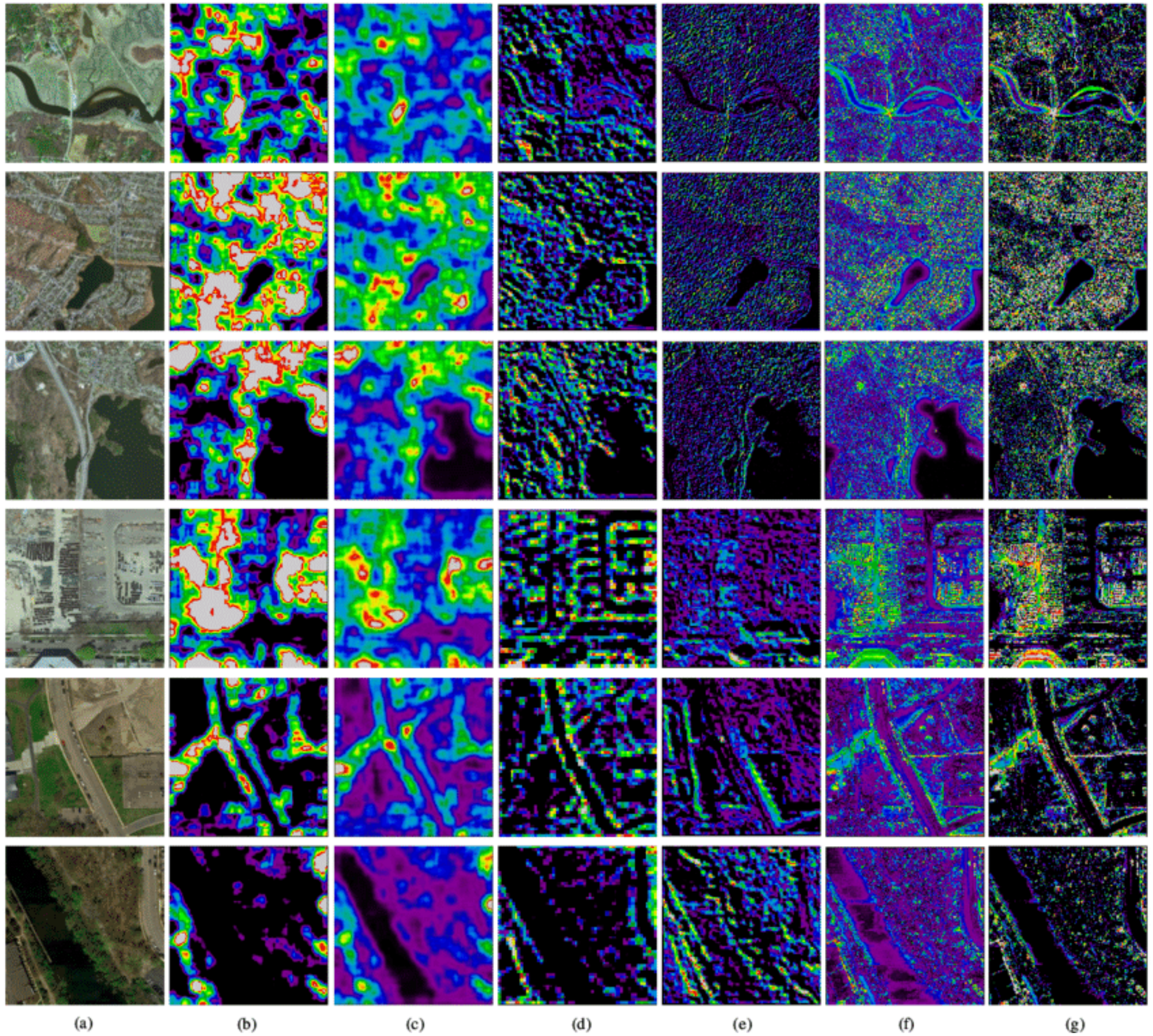


sur_refl_b01_1	sur_refl_b02_1	sur_refl_b03_1	sur_refl_b04_1	sur_refl_b06_1	sur_refl_b07_1	ndvi	ndwi1	ndwi2
137	1177	132	236	797	286	7914	1925	6090

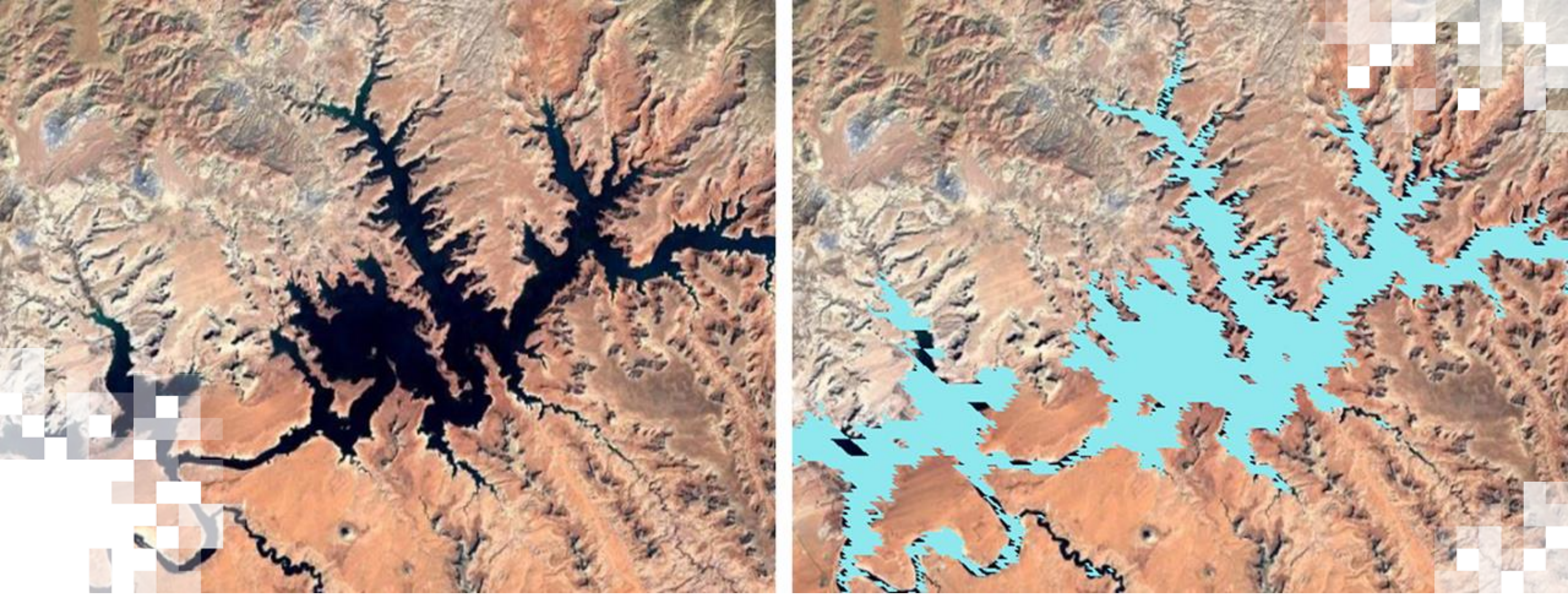
Image Source: https://shap.readthedocs.io/en/latest/index.html

Attention and Explainability – XAI

- Visual Transformers (ViT) can output attention maps.
- Attention maps are the intermediate output of the model that highlights the important region in the image for the target class.
- Visualizing attention maps can lead to a better understanding of how the model is processing the input and which features are most important for the prediction.



Remote sensing images and visualization of attention maps in different moduls. Shamsolmoali et al. (2020), https://doi.org/10.1109/TGRS.2021.3112481



Exercise: Model Explainability and Interpretability - XAI

Closing Remarks

- We have:
  - Provided a base on the fundamentals of Machine Learning for Earth Science using a binary classification problem as an example.
  - Introduced the general concept of machine learning and possible scenarios of its benefits across other domains.
  - Provided the base to produce an effective training, validation, and test dataset from both raster and tabular data sources.
  - Provided the tools to train and perform inference of a XGBoost model, including its fine-tuning and XAI analysis.

This is just an introduction to the very broad field of Machine Learning. The fundamentals learned in this training will provide the basis to understand literature and to know when a specific algorithm might be the most applicable.

References

- Crankshaw, D., & Gonzalez, J. (2018). Prediction-Serving Systems: What happens when we wish to actually deploy a machine learning model to production?. *Queue*, *16*(1), 83-97.
- Elders, A., Carroll, M. L., Neigh, C. S., D'Agostino, A. L., Ksoll, C., Wooten, M. R., & Brown, M. E. (2022). Estimating crop type and yield of small holder fields in Burkina Faso using multi-day Sentinel-2. *Remote Sensing Applications: Society and Environment*, *27*, 100820.
- Fleming, S. W., Watson, J. R., Ellenson, A., Cannon, A. J., & Vesselinov, V. C. (2021). Machine learning in Earth and environmental science requires education and research policy reforms. *Nature Geoscience*, *14*(12), 878-880.
- Prša, A., Kochoska, A., Conroy, K. E., Eisner, N., Hey, D. R., IJspeert, L., ... & Winn, J. N. (2022). TESS Eclipsing Binary Stars. I. Short-cadence Observations of 4584 Eclipsing Binaries in Sectors 1–26. *The Astrophysical Journal Supplement Series*, *258*(1), 16.
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. the National Energy Research Supercomputing Center in Lawrence Berkeley National Laboratory, Berkeley, CA, USA: Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*, 195-204.
- Yu, S., & Ma, J. (2021). Deep learning for geophysics: Current and future trends. *Reviews of Geophysics*, *59*(3), e2021RG000742.

Contributors

- Jordan A. Caraballo-Vega
- Mark L. Carroll
- Jules R. Kouatchou
- Jian Li
- Caleb S. Spradlin
- Brock Blevins
- Melanie Follette-Cook
- Erika Podest
- Brian Powell
- Akiko Elders

Thank You!