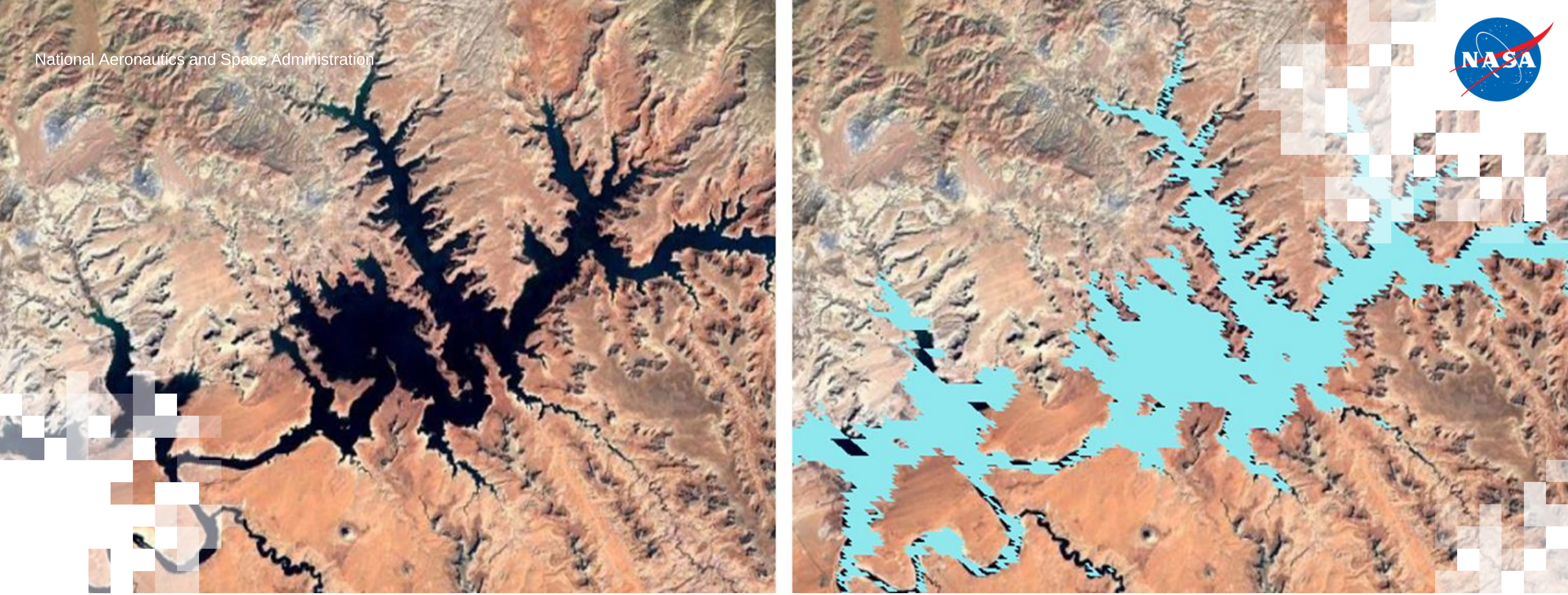
National Aeronautics and Space Administration

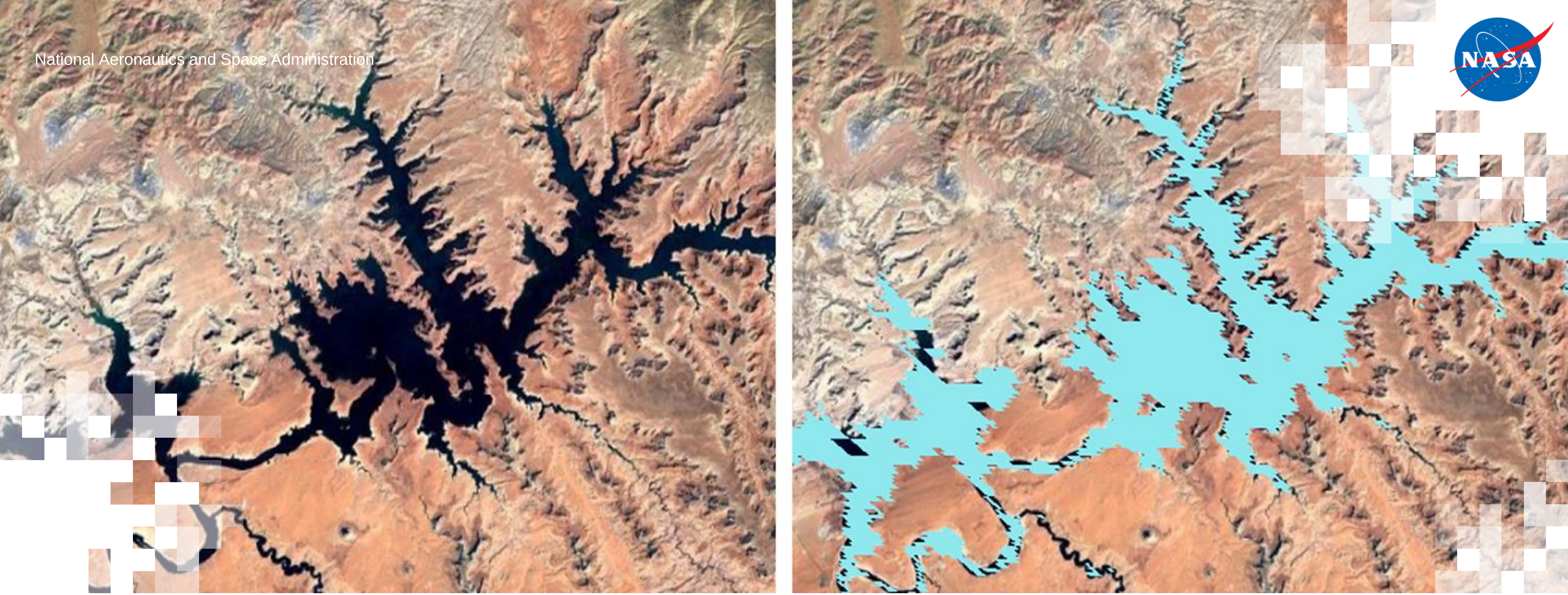


# Fundamentals of Machine Learning for Earth Science

## Part 1: Overview of Machine Learning

Trainers: Jordan A. Caraballo-Vega, Mark L. Carroll, Jules Kouatchou, Jian Li, Caleb S. Spradlin

April 20, 2023



National Aeronautics and Space Administration

# NASA Applied Remote Sensing Training (ARSET)

Brock Blevins, Training Coordinator

# NASA Applied Remote Sensing Training (ARSET) Goal

**Empower the global community to incorporate Earth-observing data into environmental management and decision-making**

- Agriculture
- Climate
- Disasters
- Health & Air Quality
- Land
- Water Resources

# NASA ARSET Training Availability

- Online webinar and self-paced
- Custom in-person
- cost-free
- Multi-lingual options
- Range of levels to meet diverse audience needs
- Materials are free use and adapt with credit to NASA ARSET

Visit NASA ARSET website to view all of our options

# Training Objectives

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

- Recognize the most common machine learning methods used for processing Earth Science data
- Describe the benefits and limitations of machine learning for Earth Science analysis
- Explain how to apply basic machine learning algorithms and techniques in a meaningful manner to remote sensing data
- Use an analysis-appropriate training dataset to evaluate conditions and solutions for a given case study
- Complete basic procedures to interpret, refine and evaluate the accuracy of the results of machine learning analysis

# Reminder of Prerequisites

- Prerequisites:
  - Session 1 of our on-demand Fundamentals of Remote Sensing series or have equivalent experience.
  - Attendees will need access to Google Drive and Google Colab. To access these resources, users must use an email ending in ‘gmail.com’.
  - We will have the video of this demonstration within the training recording available within 48 hours after the presentation for you to go through at your own pace.

# Training Schedule

**Part 1: Overview of Machine Learning**

**April 20, 2023**

Part 2:

Training Data and Land Cover Classification Example

April 27, 2023

Part 3:

Model Tuning, Parameter Optimization, and Additional Machine Learning Algorithms

May 4, 2023

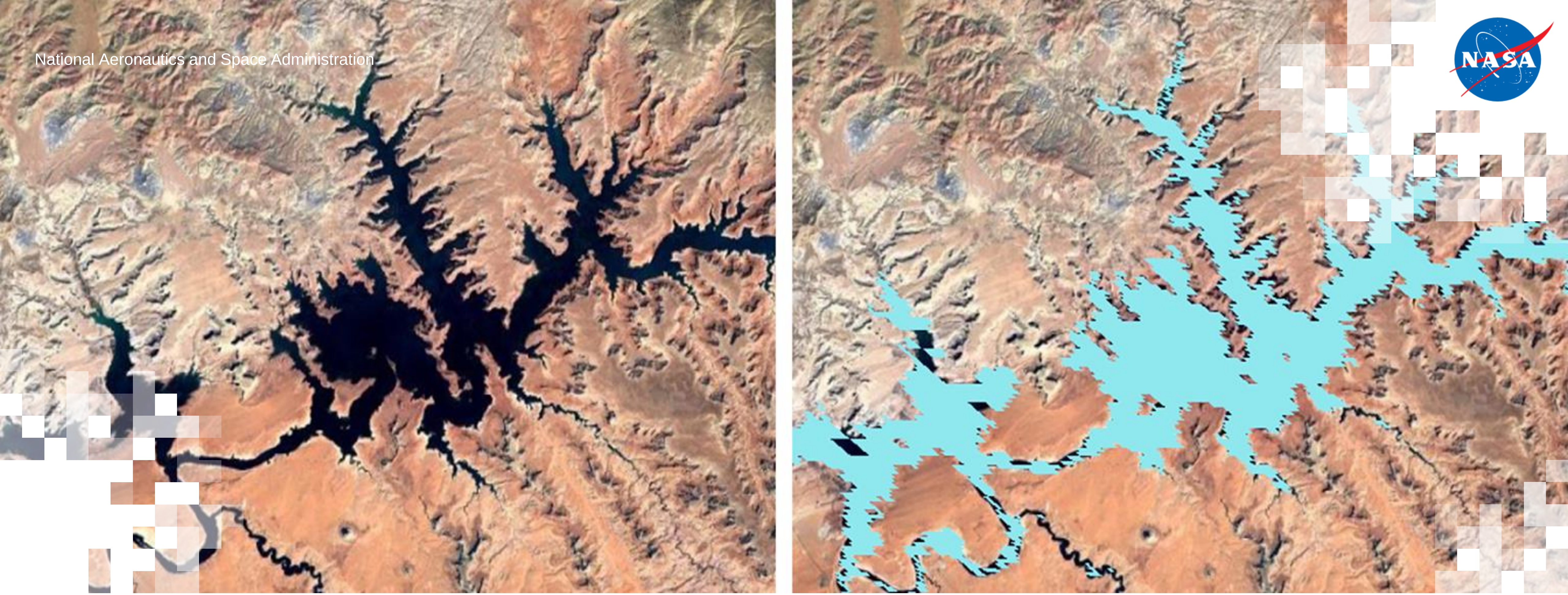
Homework

Independent practice and application

Due May 19
Opens May 4

Optional opportunity to earn a certificate of completion

National Aeronautics and Space Administration



# Fundamentals of Machine Learning for Earth Science

## Part 1: Overview of Machine Learning

Trainers: Jordan A. Caraballo-Vega, Mark L. Carroll, Jules Kouatchou, Jian Li, Caleb S. Spradlin

April 20, 2023

# Instructor Team

Jordan A. Caraballo-Vega
Computer Engineer

Jules Kouatchou
Chief Programmer/Analyst

Caleb S. Spradlin
Software Developer

Mark L. Carroll
Research Scientist

Jian Li
Senior Principal
Applications Engineer

# Session 1 Outline

- Overview of Machine Learning
- Importance of Machine Learning targeted towards Earth Science
- Usability of Machine Learning
- Software to Support Machine Learning
- Machine Learning Applications
- Hands on Jupyter Notebook Exercise: Load and Visualize Data
- Post-Session Assignment
- Q&A Session

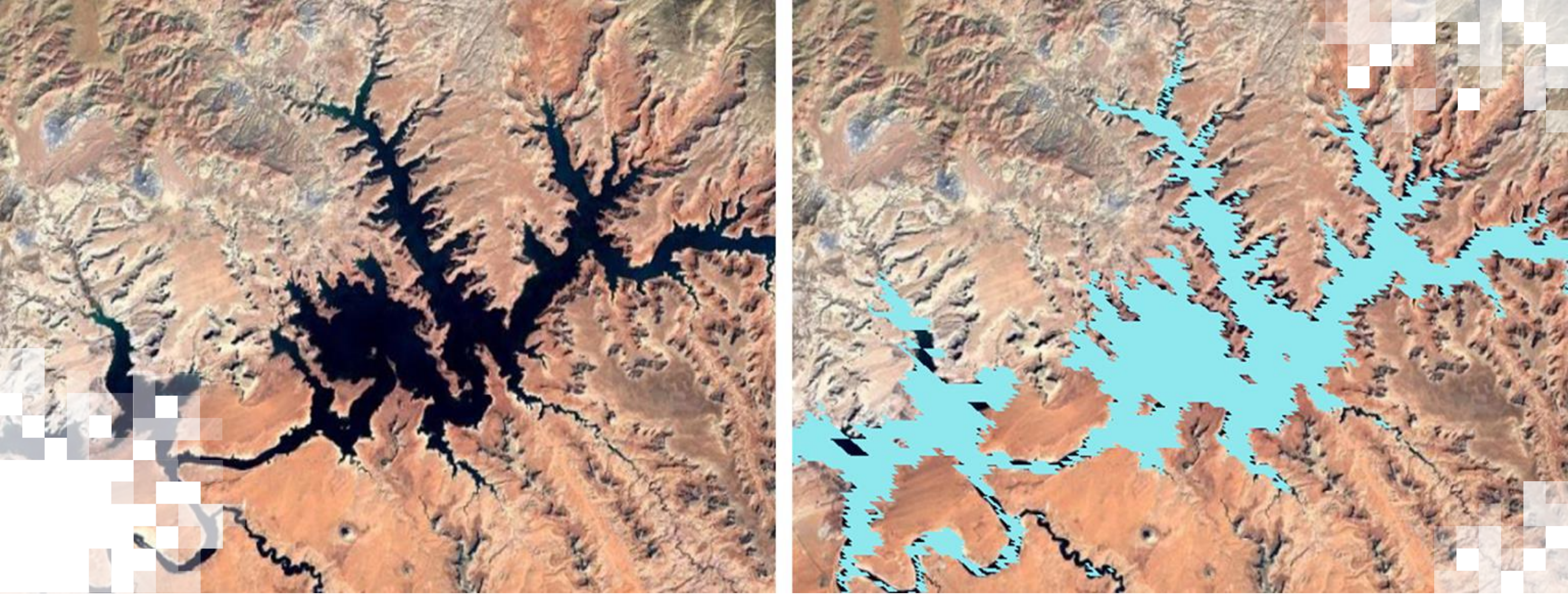
**Resources for this Training**

https://github.com/NASAARSET/ARSET_ML_Fundamentals

# Training Objectives

After participating in this training, attendees will be able to:

- Recognize the most common machine learning methods used for processing Earth Science data
- Describe the benefits and limitations of machine learning for Earth Science analysis
- Explain how to apply basic machine learning algorithms and techniques in a meaningful manner to remote sensing data



# Overview and Theory

Trainer: Jules Kouatchou

# 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?

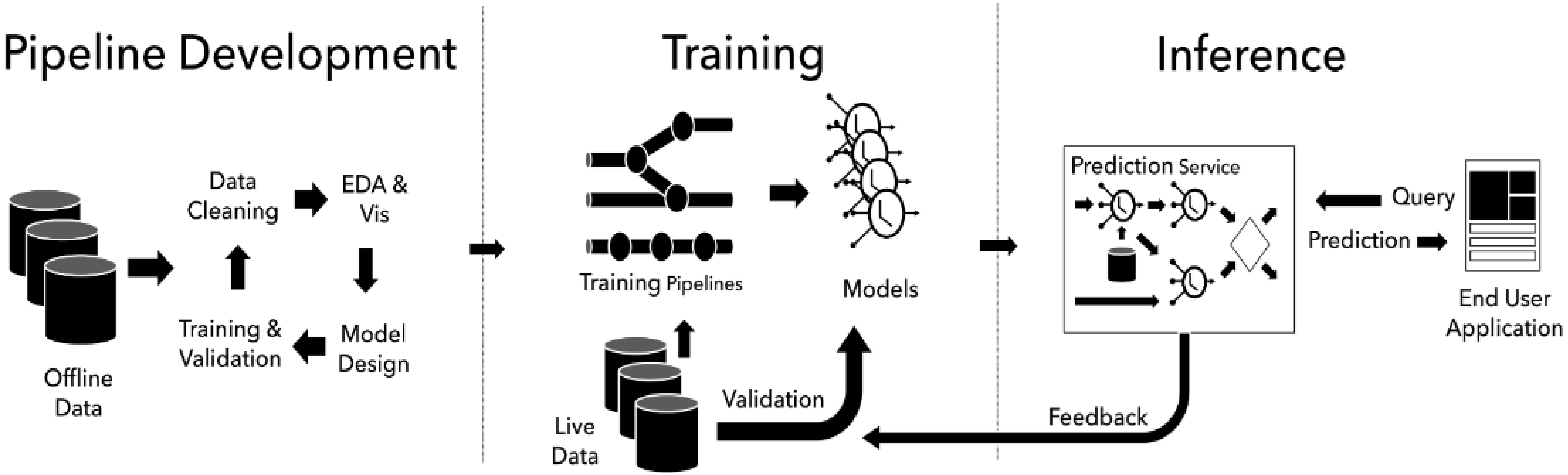
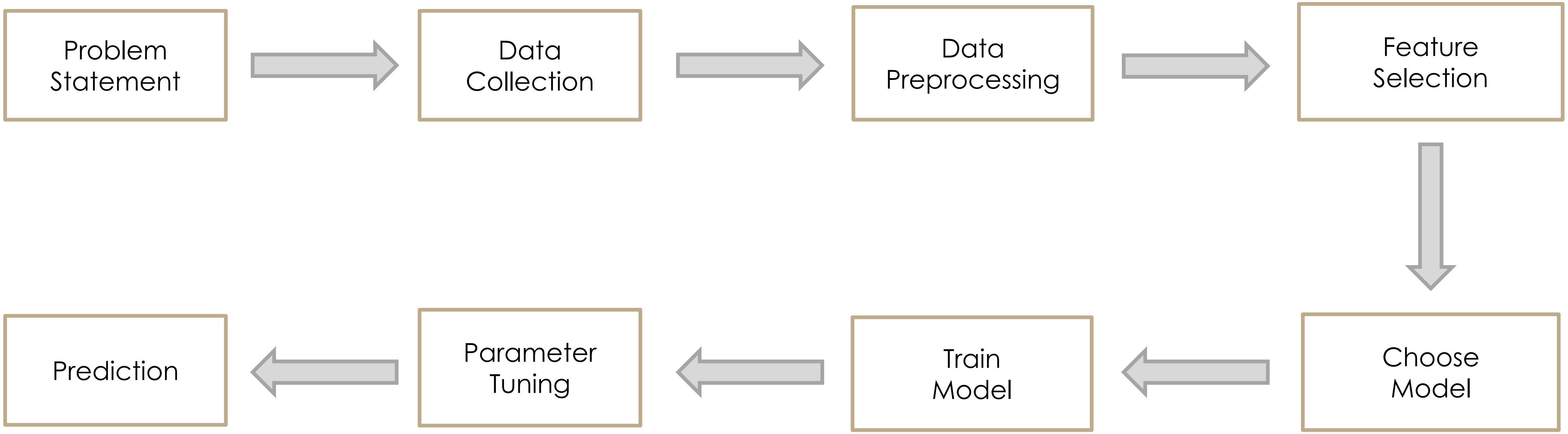


Image Source: Daniel Crankshaw (in a Short History of Prediction-Serving Systems)

# Machine Learning Steps



# Machine Learning Algorithms

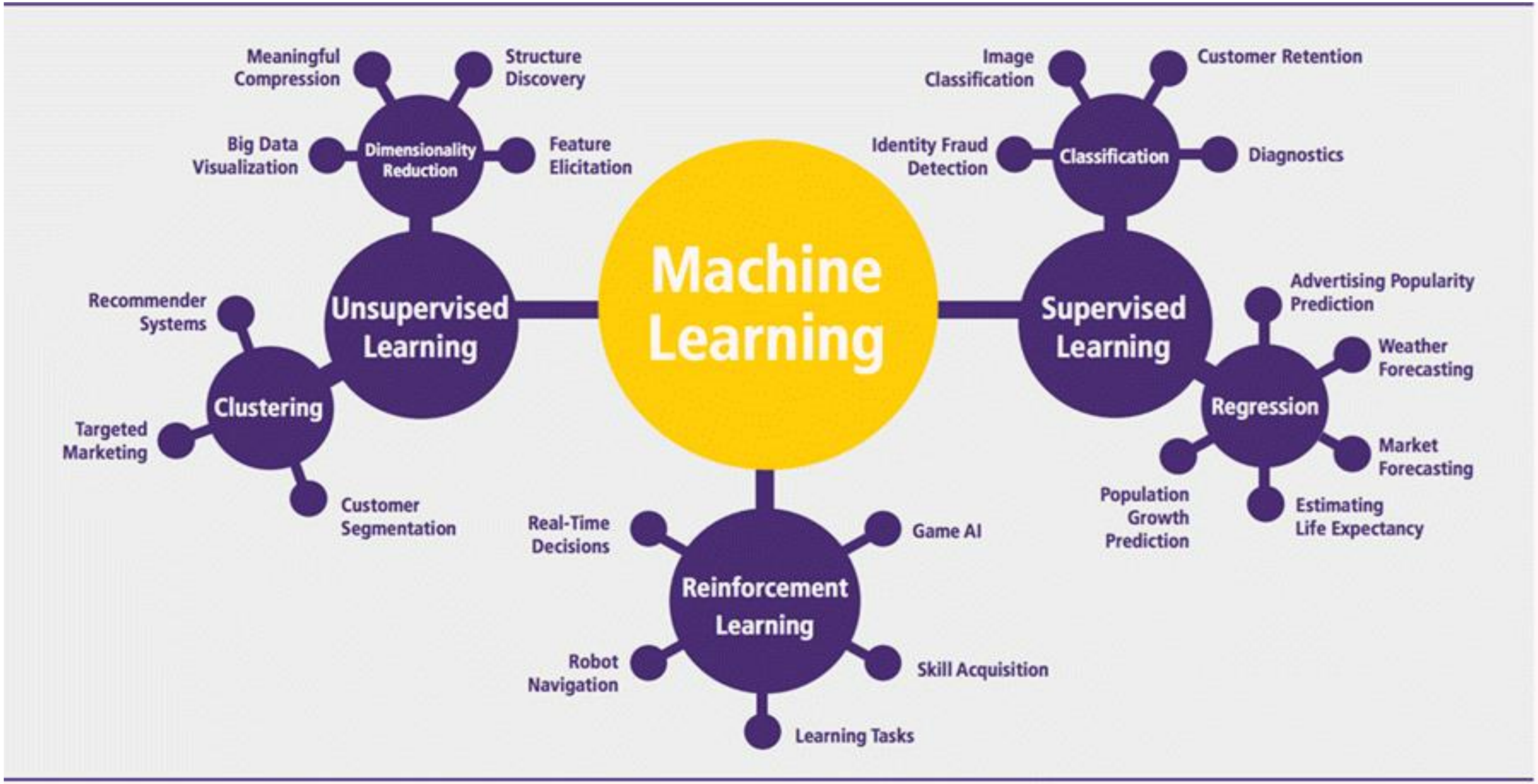
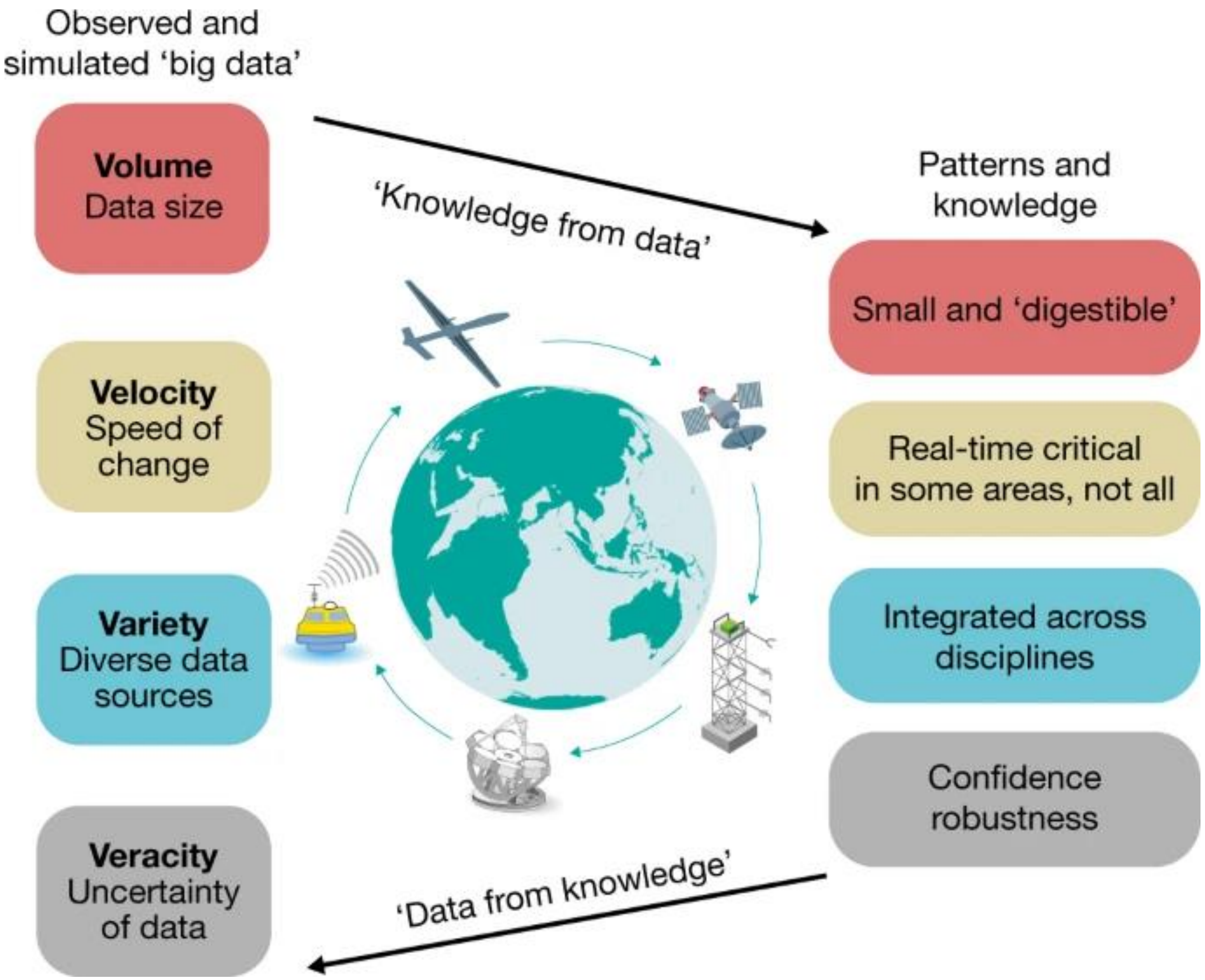


Image Source: guru99.com

# Big Data in Earth Science



Observed and simulated 'big data'

**Volume** Data size

**Velocity** Speed of change

**Variety** Diverse data sources

**Veracity** Uncertainty of data

'Knowledge from data'

Patterns and knowledge

Small and 'digestible'

Real-time critical in some areas, not all

Integrated across disciplines

Confidence robustness

'Data from knowledge'

Reichstein et al. (2019), https://doi.org/10.1038/s41586-019-0912-1

# Machine Learning in Earth Science

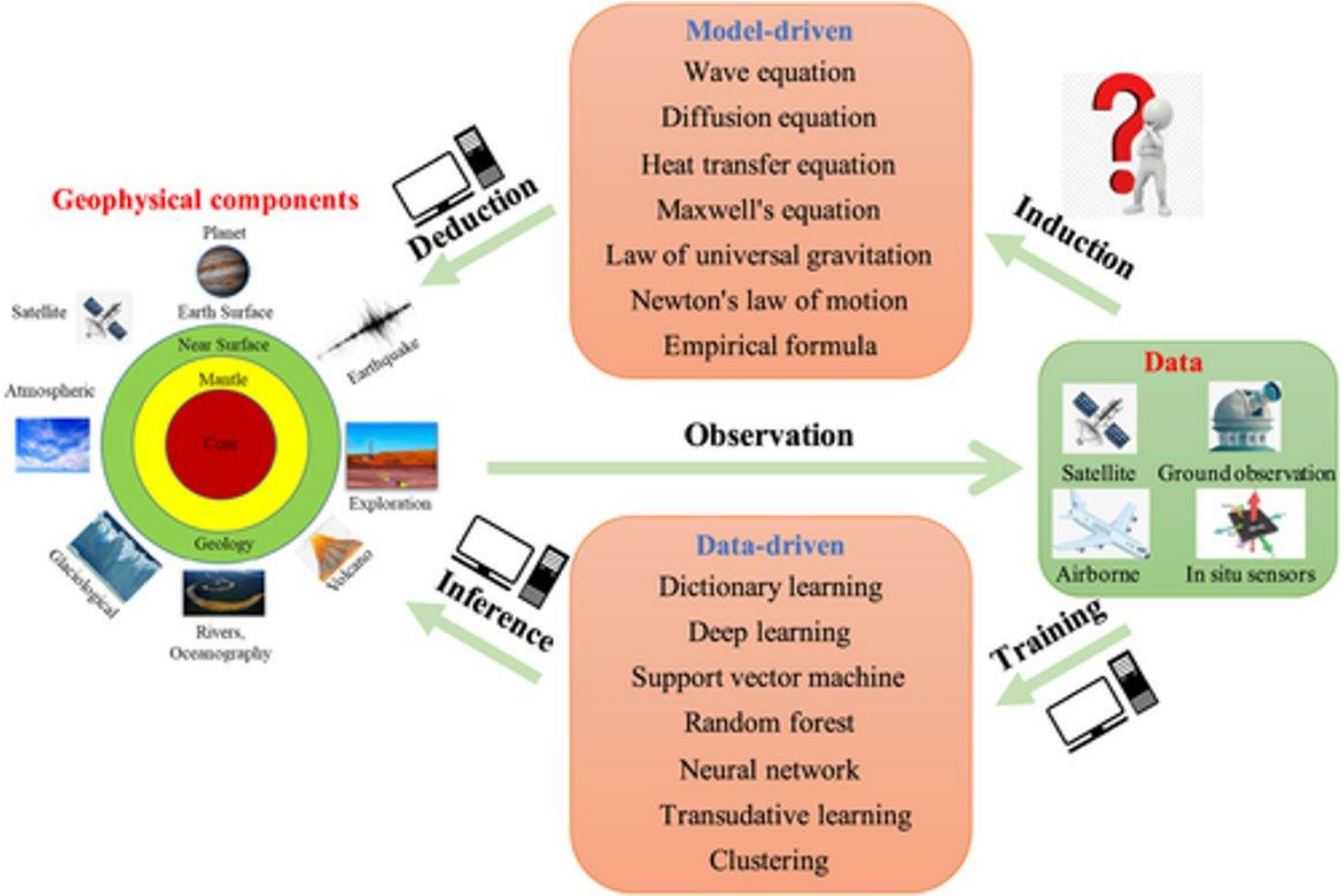
**Leveraging advances in artificial intelligence could revolutionize the Earth and environmental sciences. We must ensure that our research funding and training choices give the next generation of geoscientists the capacity to realize this potential.**

Fleming *et al.* (2021), https://doi.org/10.1038/s41561-021-00865-3

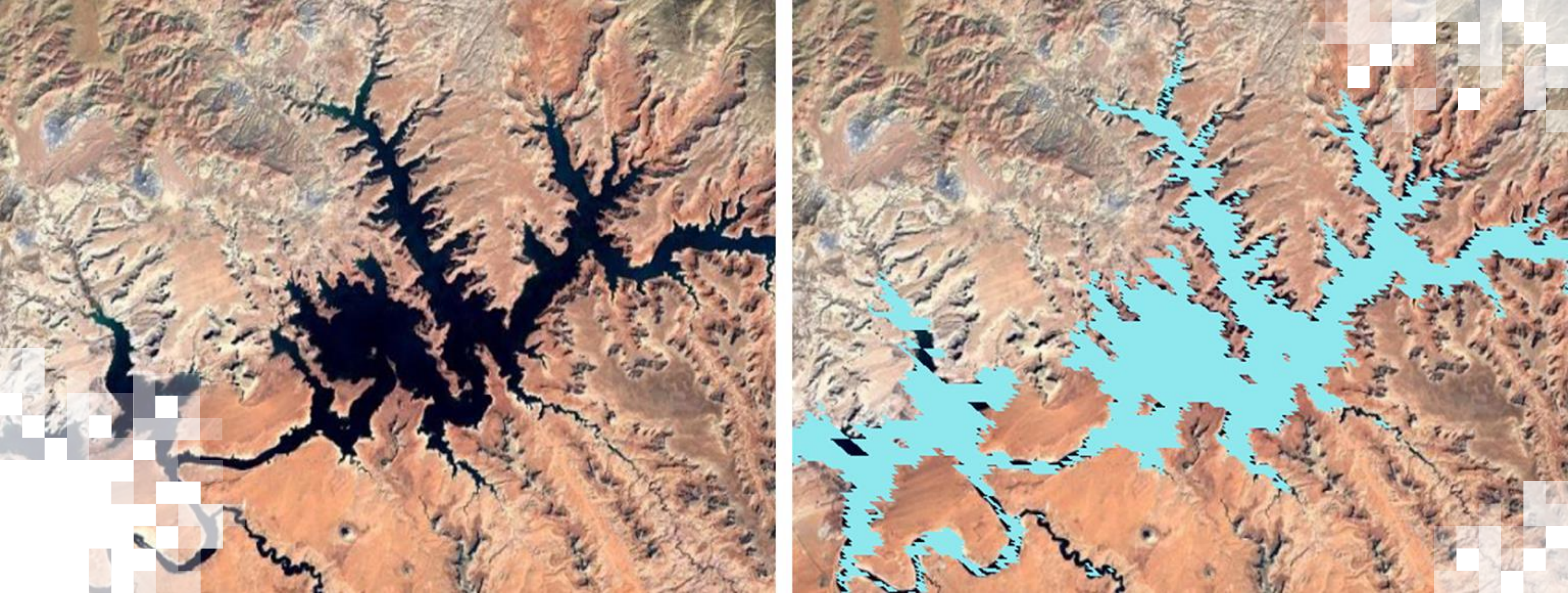
# Machine Learning in Earth Science

- Problems in Earth science are often complex.
- It is difficult to apply well-known and described mathematical models to the natural environment:
  - ML is commonly a better alternative for such non-linear problems.
- A number of researchers found that machine learning outperforms traditional statistical models in Earth science, such as in:
  - Characterizing forest canopy structure,
  - Predicting climate-induced range shifts, and
  - Delineating geologic facies.

# How Machine Learning is Applied in Earth Science



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



# Machine Learning Applications

Trainer: Jian Li

# Benefits of Utilizing Machine Learning

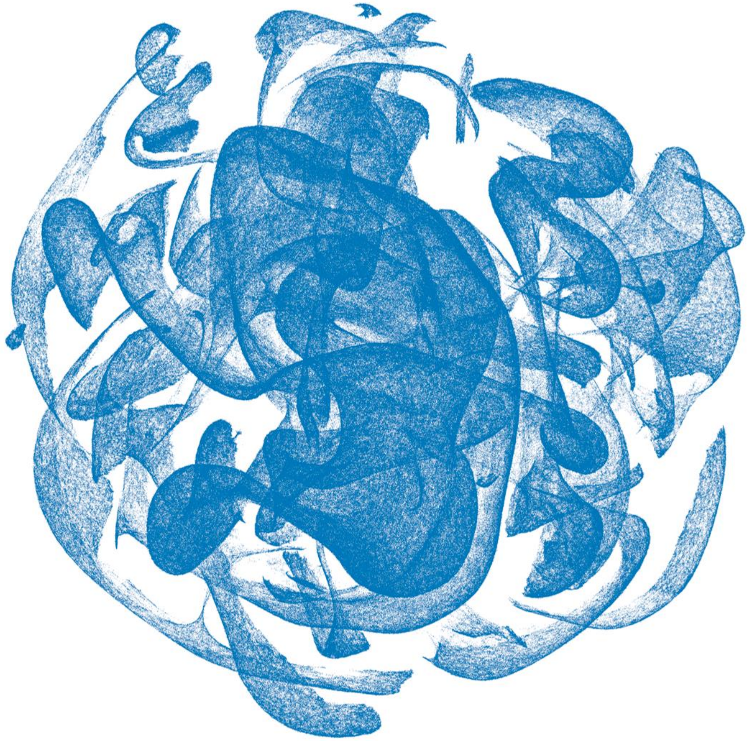
There are numerous ways in which ML can accelerate scientific research, such as:

- Increased Efficiency: Machine learning can help automate the analysis of large and complex datasets, allowing scientists to quickly process and analyze large amounts of data.
- New Insights and Discoveries: Machine learning can help scientists identify new patterns and relationships in complex datasets, leading to new insights and discoveries in Earth Science research.
- Improved Predictive Modeling: Machine learning algorithms can be used to build accurate predictive models that can help scientists better understand complex Earth Science phenomena.

# Efficiency, Accuracy, and Discovery

Identify new stars & star systems from a massive number of observations

- An all-sky survey mission, called the Transiting Exoplanet Survey Satellite **(TESS)**
- Using AI, ML, and HPC tools, NASA scientists have extracted more than **60 million** light curves for further investigation.
- NASA astronomers have identified:
  - > **50** planet candidates
  - > **200** potential heartbeat stars
  - **> 10** potential triple star systems
  - **> 20** potential quadruple star systems
  - A potential sextuple star system
- All previously undiscovered



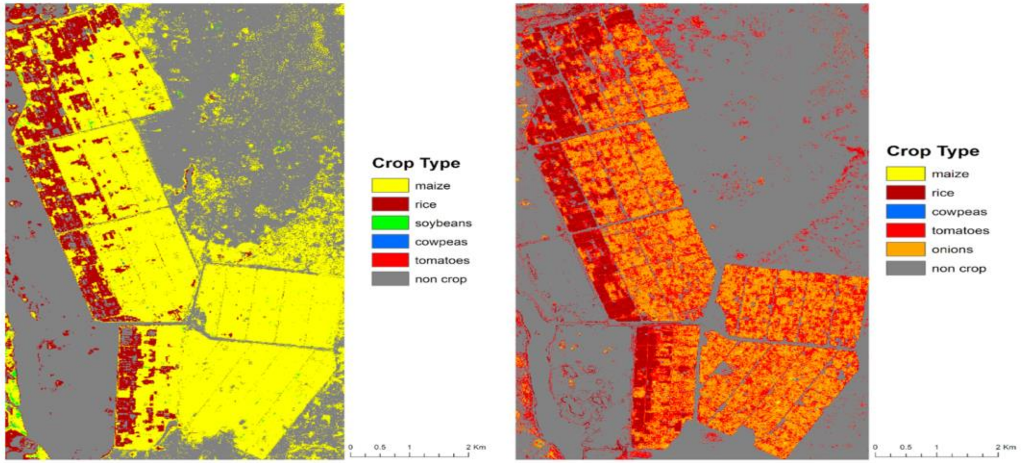
*A two-dimensional projection of the high-dimensional space of TESS light curve representations. Image Credit: Brian P. Powell, NASA Goddard.*

Prša et al. (2022), https://doi.org/10.3847/1538-4365/ac324a

# Efficiency, Accuracy, and Discovery

## Machine Learning-Based Crop Type and Yield Estimates in Burkina Faso, West Africa

*Illustration by ESA.*



*ML predictions of crop type across the ~2,250-hectare study region for the 2019 rainy (left) and dry (right) seasons. The rainy season has maize (yellow) and rice (maroon) predominating, while the dry season has onions (orange), tomatoes (red), and rice (maroon) predominating.*

- NASA Goddard and the Millennium Challenge Corporation (MCC)
- Invest in Agricultural Development Projects to empower local farmers and combat food insecurity
- *Sentinel*-2 satellite and in-situ data to train and optimize five Random Forest machine learning models to estimate crop type and yield across the study region
- Model accuracy was **88%** for crop type and **>64%** for crop yield during 2019’s rainy season and **64%** for crop type and **>53%** for crop yield during 2019’s dry season.
- The machine learning model made interannual predictions for 2020’s dry season without training data; accuracies were up to **60%** for crop type.

*Elders et al. (2022),* https://doi.org/10.1016/j.rsase.2022.100820

# Software to Support Machine Learning

- Programming Languages
- Software Packages

mlr
OpenAI
H2O.ai
Keras
Shiny
tidyr
dplyr
ggplot2
spaCy
OpenCV
TensorFlow
tableau
readr
statsmodels
learn
PyTorch
plotly
SciPy
R
KNIME
sas
NumPy
matplotlib
Seaborn
Scala
Java
alteryx
R Studio
jupyter
pandas
kafka
Spark
GitHub
DASK
git
python
SQLAlchemy
MySQL
Flask
Graphene Python
node JS
ORACLE DATABASE
cassandra
PREFECT
PostgreSQL
GO
JS
Express
APOLLO
elastic
HTML
CSS
TS
A
mongoDB
SVELTE
presto
ANSIBLE
heroku
Firebase
Azure
aws

Image Source: https://dev.to/minchulkim87/my-data-science-tech-stack-2020-1poa

# Software to Support Machine Learning, Cont.

**Python:** Python is the most used language for Machine Learning. One of the main reasons Python is so popular within AI development is that it was created as a powerful data analysis tool and has always been popular within the field of big data.

**R:** R might not be the perfect language for AI, but it's fantastic at crunching very large numbers, which makes it better than Python at scale. And with R's built-in functional programming, vectorial computation, and Object-Oriented Nature, it does make for a viable language for AI.

**Java:** Java is an important language for AI. One reason for that is how prevalent the language is in mobile app development. And given how many mobile apps take advantage of AI, it's a perfect match.

**Julia:** Julia is one of the newer languages on the list and was created to focus on performance computing in scientific and technical fields.

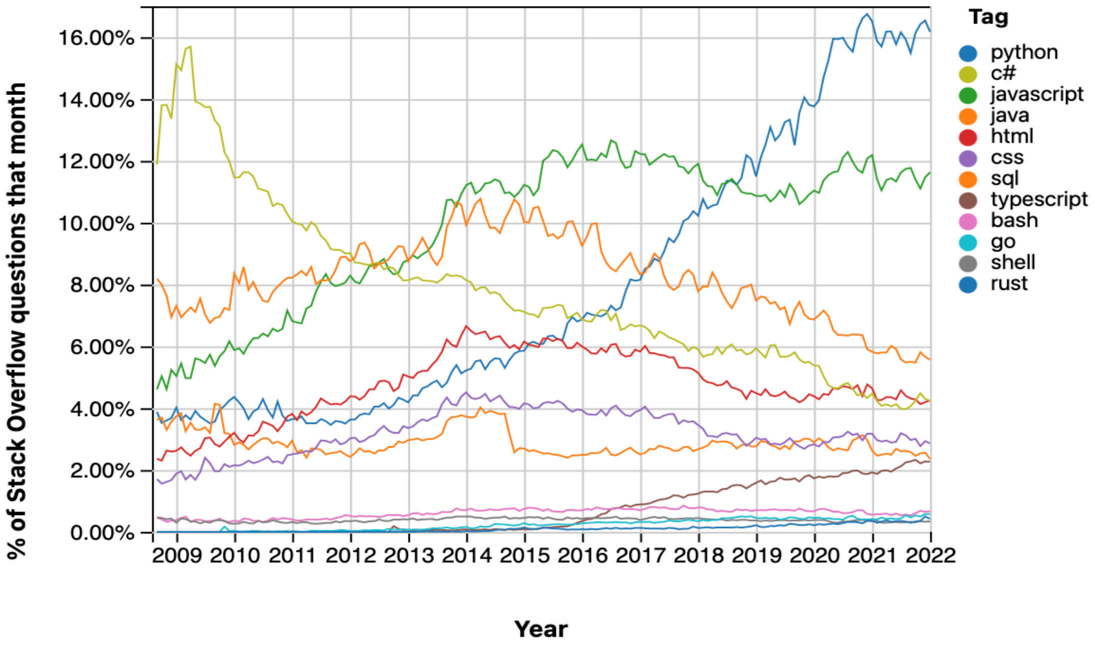


Image Source: Stack Overflow 2022

# 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.

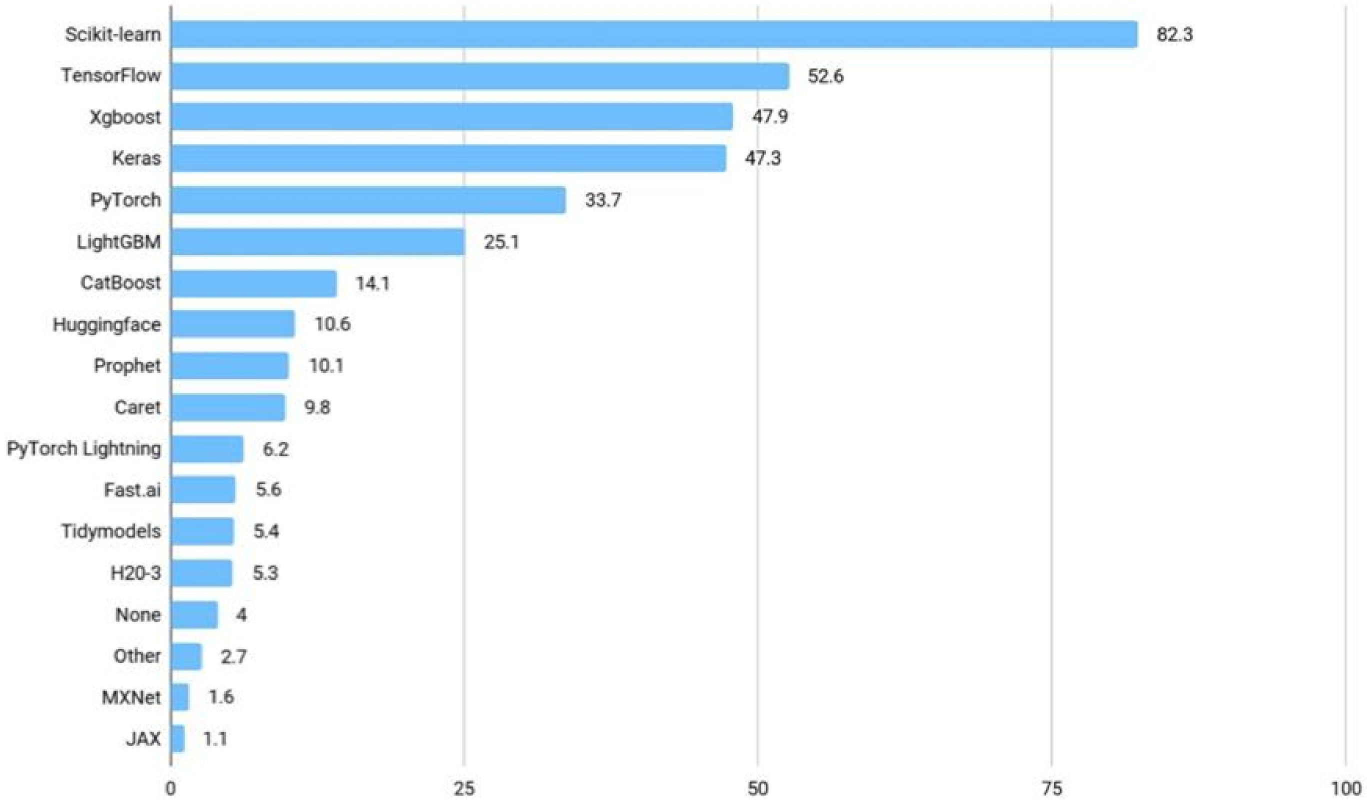


*Image Source: Kaggle's 2021 State of Data Science and Machine Learning survey*

# Machine Learning Frameworks in Python, Cont.

- **Scikit-Learn** — One of the most important libraries (Swiss Knife) for Machine Learning as it provides a number of simple and efficient tools for data analysis. It provides functionality for classification, regression, clustering algorithms, dimensionality reduction, model selection, and data preprocessing.
- **TensorFlow** — Library was developed by engineers and researchers working on the Google Brain team that conducts machine learning and neural networks research. It allows researchers to push boundaries in discovering state-of-the-art (SOTA) results, and also allows developers to create ML-powered applications.
- **Keras** — High-level neural networks API, which can be implemented on top of TensorFlow or Theano used for building and training deep learning models. It allows for easy and fast prototyping and supports both convolutional neural networks and recurrent networks.
- **PyTorch** — Provides functionality largely centered around building and training neural networks—the backbone of deep learning. PyTorch offers scalable distributed training of models across single or multiple CPUs and GPUs. the first release was in September 2016, but it has quickly been widely adopted by industry such as Tesla & Uber.

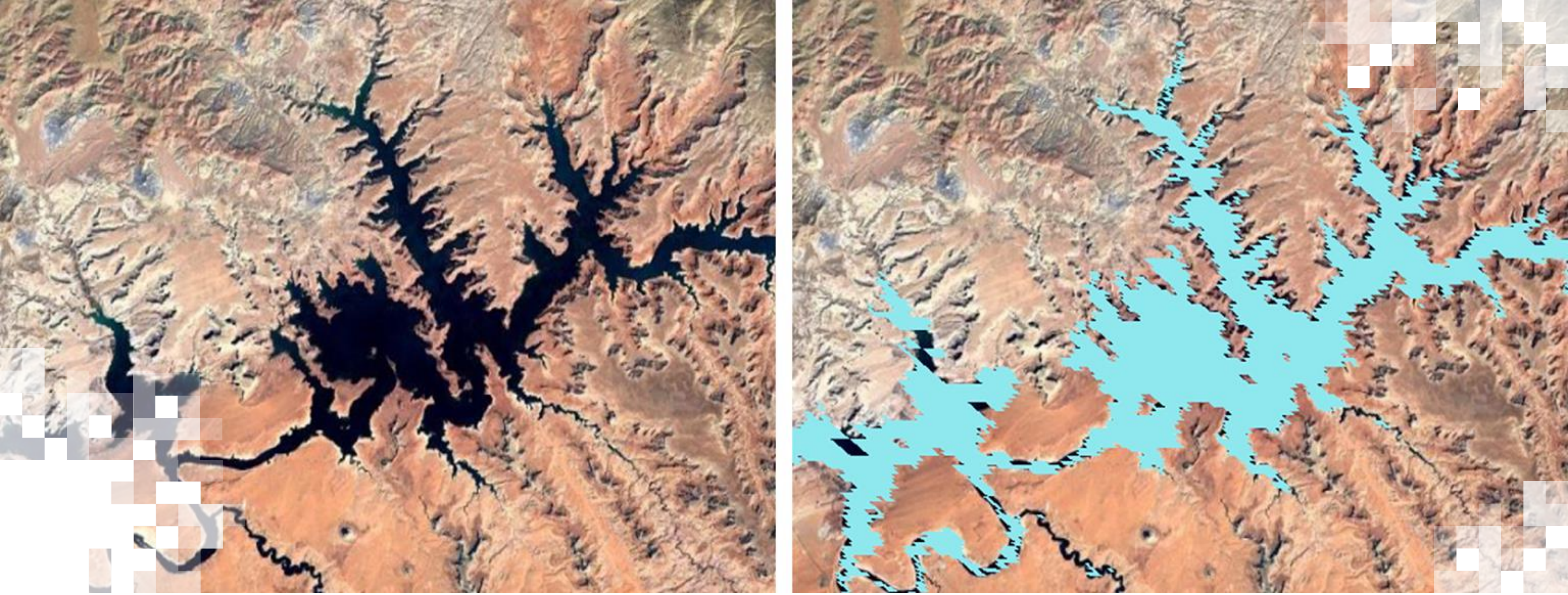
# Machine Learning Frameworks in Python, Cont.

- **Jupyter Notebook** — Open-source, web-based application which allows us to create and share documents that contain code, equations, visualizations, and text. Its uses include data cleaning and transformation, statistical modeling, data visualization, machine learning, etc.
- **Matplotlib** — Data visualization library that is used to create static, animated, and interactive visualizations. It can be used to create detailed scatterplots, histograms, bar charts, pie charts, etc.
- **Seaborn** — Statistical visualization library based on matplotlib and is integrated with pandas data structures. It provides a high-level interface for informative and statistical graphics. Since it is built on top of Matplotlib, it offers extra plots and can produce more sophisticated visualizations.

# 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)

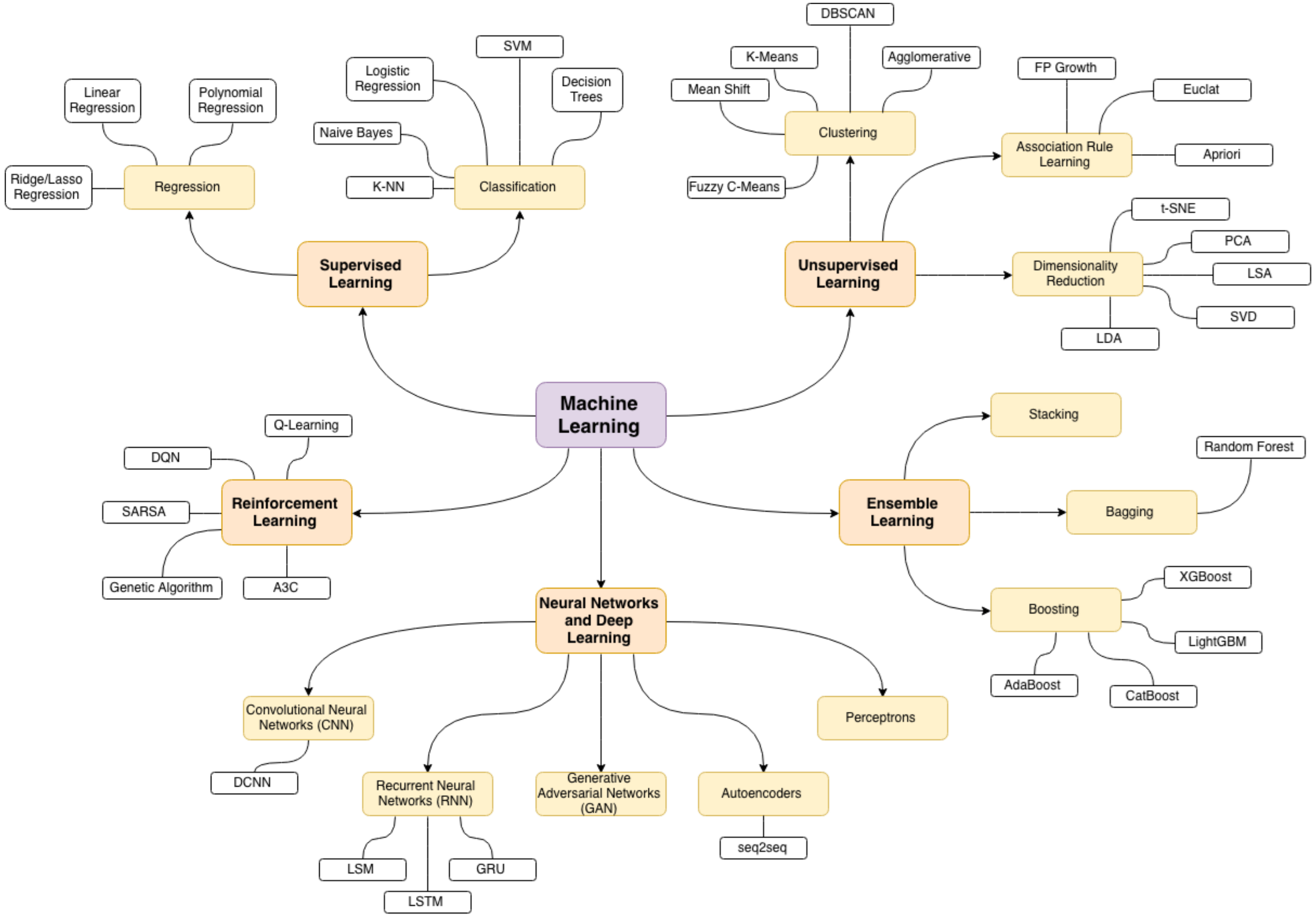


# Overview of Machine Learning Algorithms

Trainer: Jordan A. Caraballo-Vega

# Machine Learning Algorithms: Which algorithm to choose?

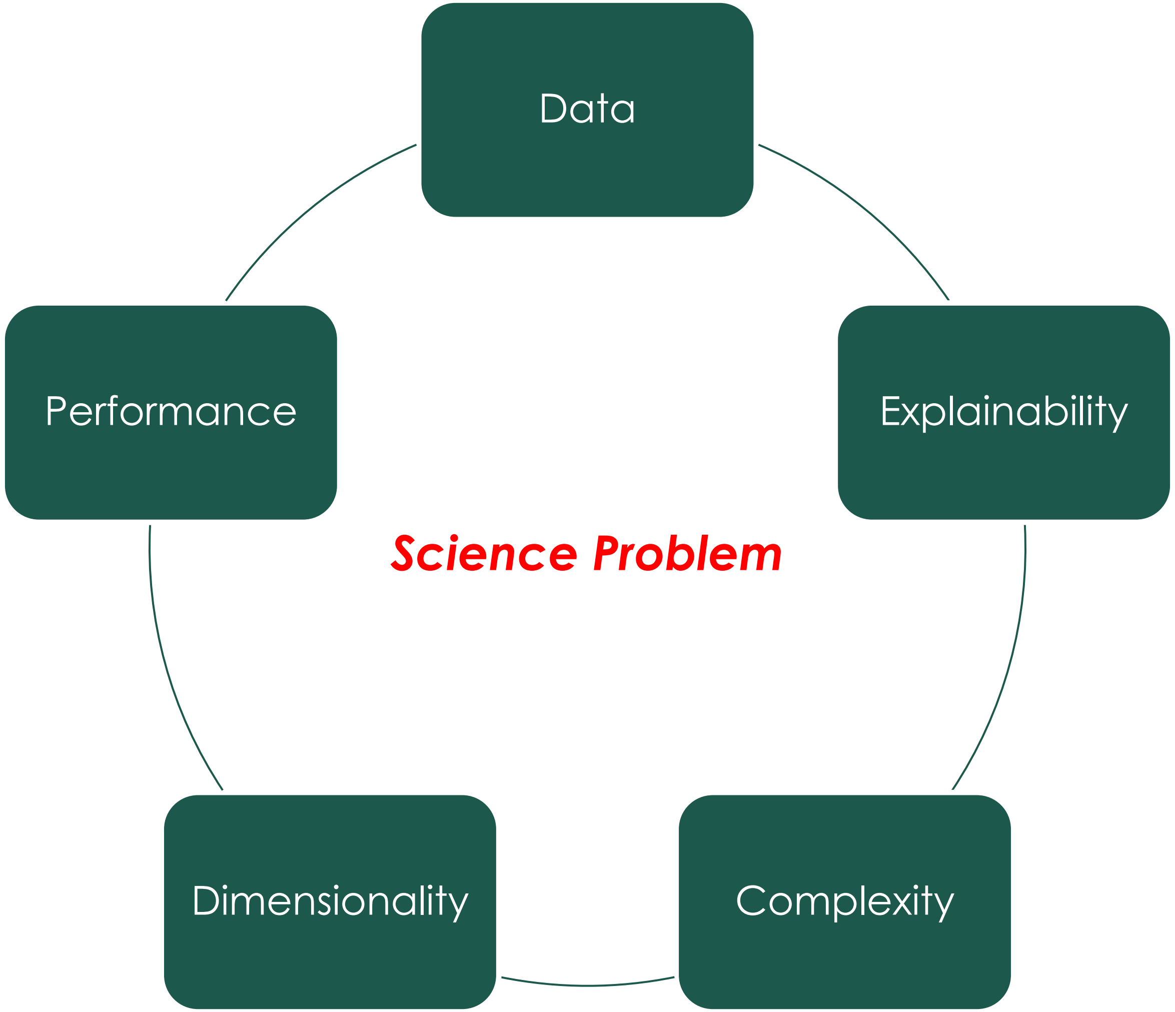
- The development of machine learning algorithms has been exponentially increasing.
- We will not dive into the specifics of each algorithm, but we will give you the tools to aid in the selection of these for your own science problem(s).
- You are not bounded to a single algorithm, but it always saves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



*Core Machine Learning Algorithms.*
*Image Source: github.com*

# Machine Learning Algorithms: Science Problem

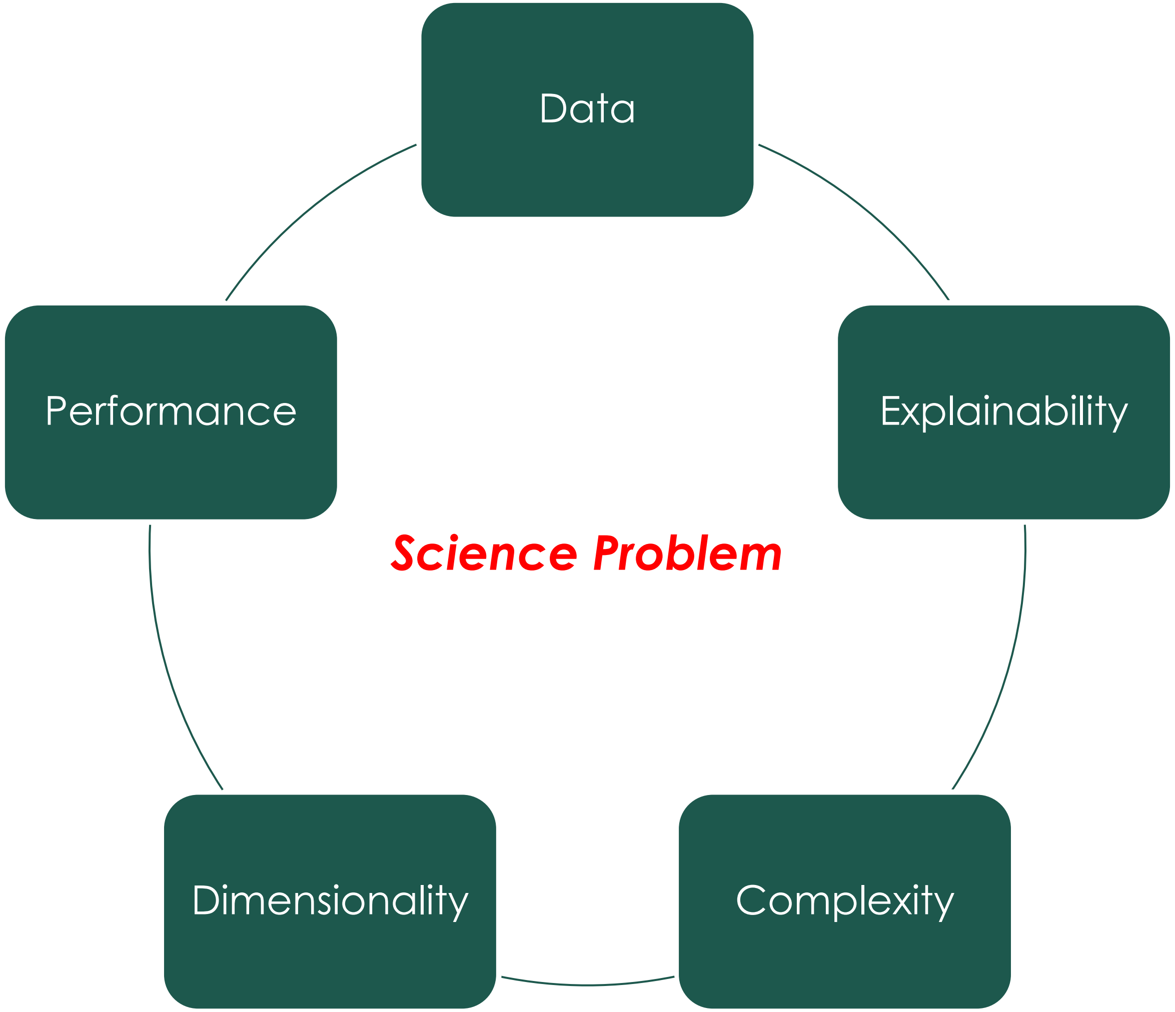
- **Which scientific question would you like to address?**
- **What information is missing to answer this question?**



*Components to aid the selection of your ML algorithm.*

# Machine Learning Algorithms: Science Problem

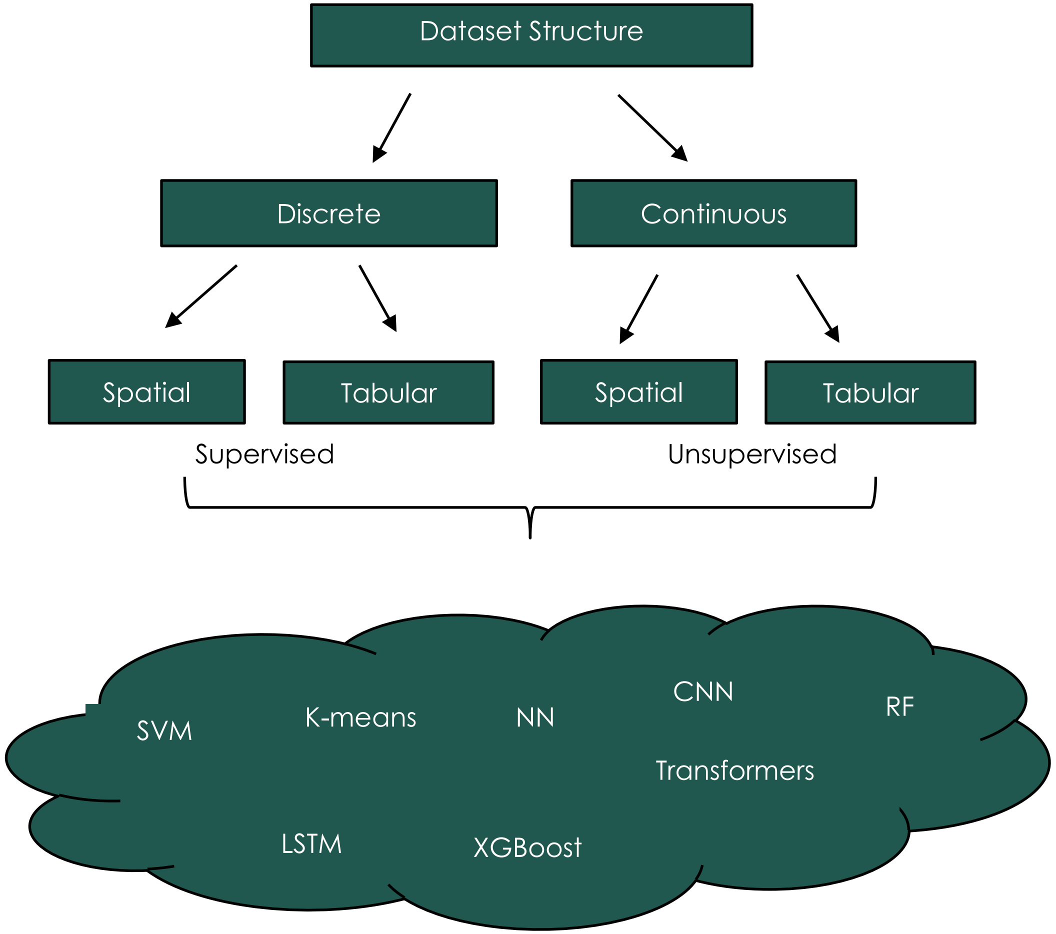
- **Which scientific question would you like to address?** We want to identify the sign, magnitude, and potential drivers of change in surface water extent in X study area.
- **What information is missing to answer this question?** We need surface water extent maps to quantify and analyze these drivers.



*Components to aid the selection of your ML algorithm.*

# Machine Learning Algorithms: Data

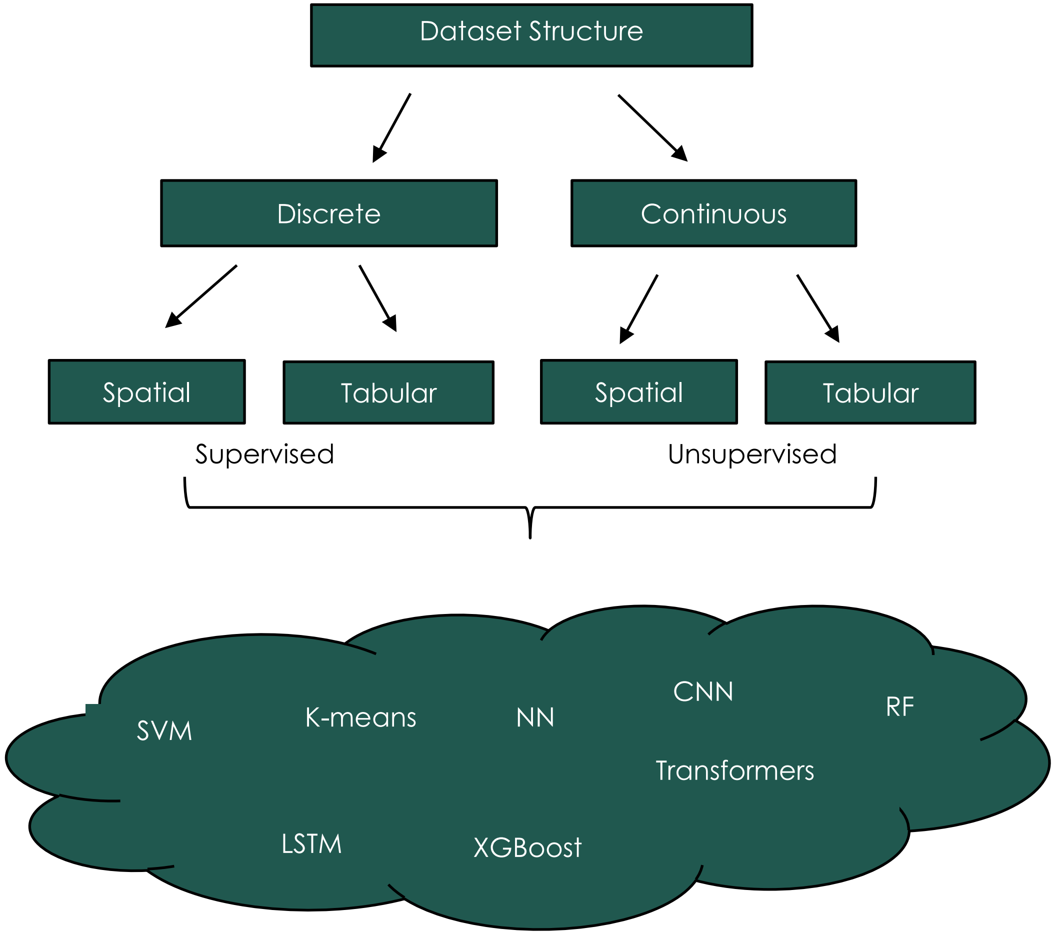
- **What data do you have available?**
- **Do you have training data available?**
- **What is the data structure of your data?**
- **Is your dependent variable a continuous or discrete problem?**



*Algorithm decision branch based on data structure.*

# 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)



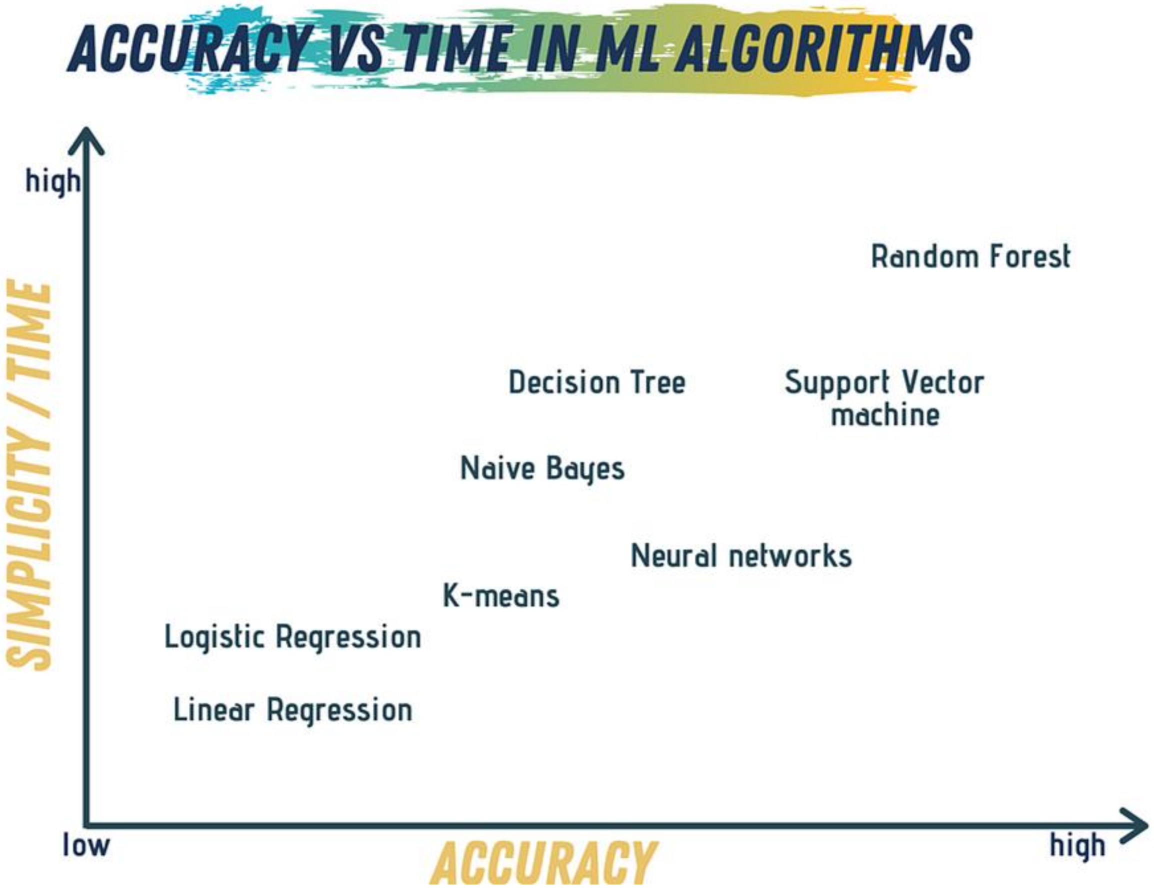
*Algorithm decision branch based on data structure.*

# Machine Learning Algorithms: Performance

- **Are there any performance requirements based on your science question (e.g., real time vs. static)?**

- **Is your software going to run on on-premise, cloud, or embedded hardware?**

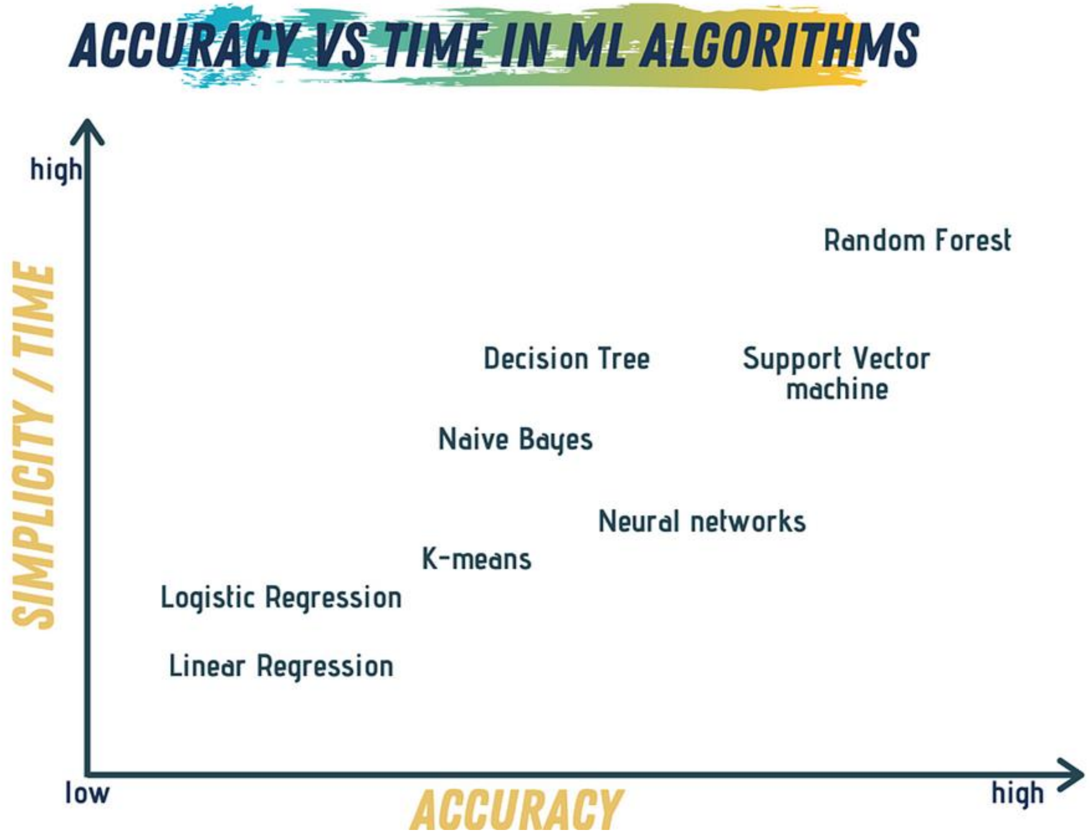
- **What is more important for your project: inference time or model performance?**



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

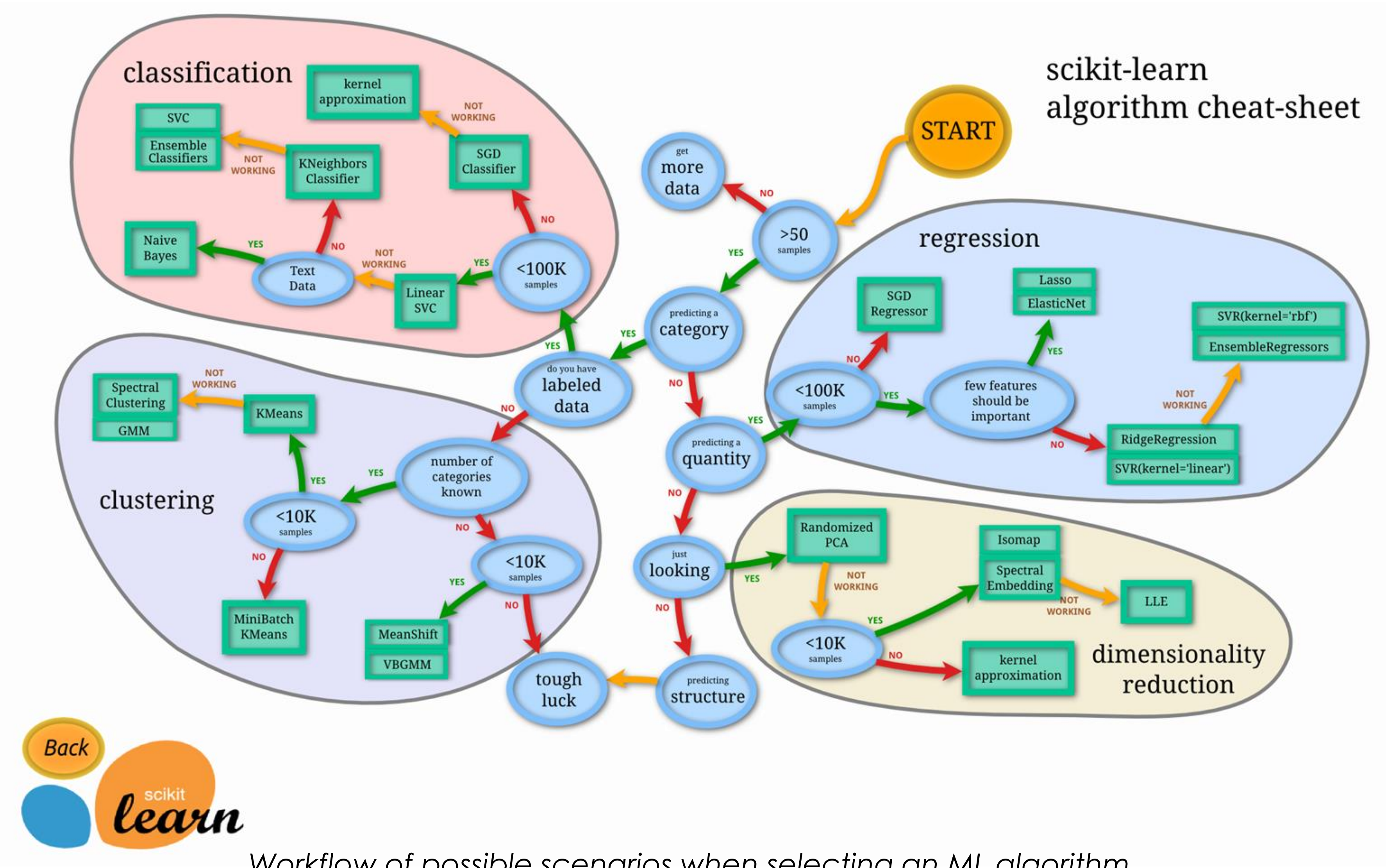
# Machine Learning Algorithms: Performance

- **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.

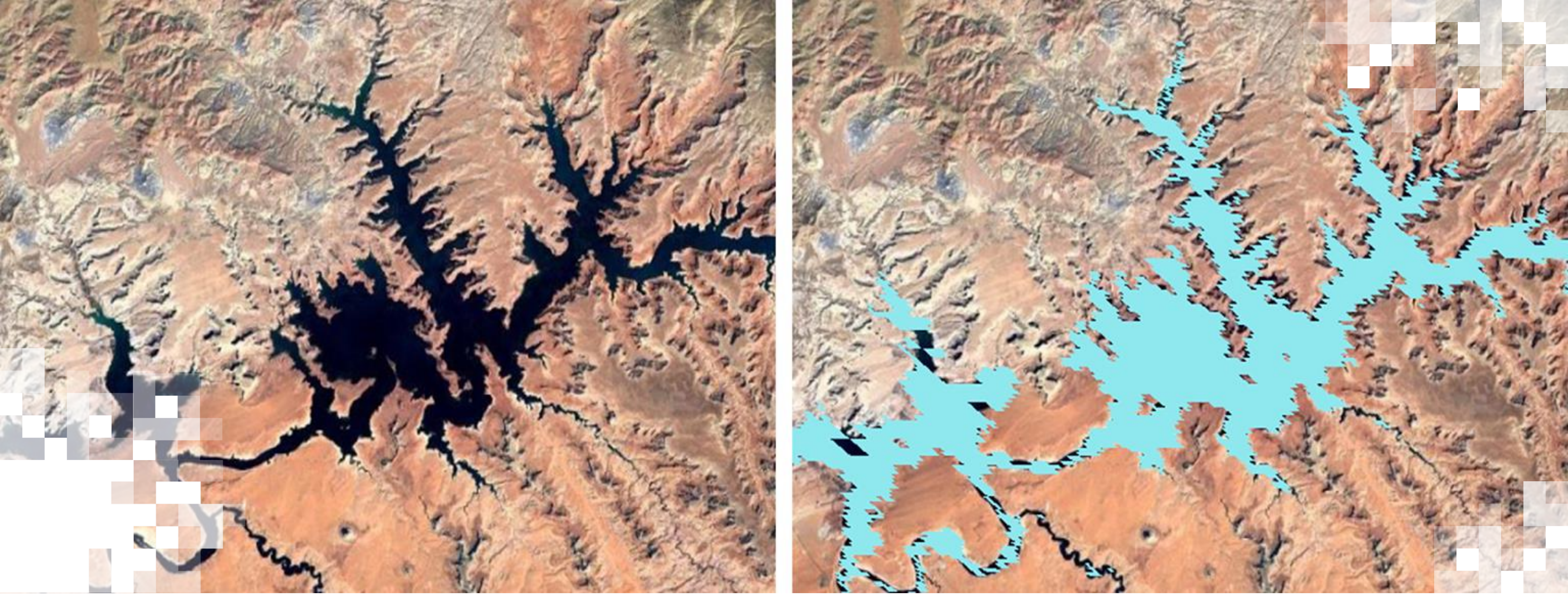


*Tradeoff between speed and accuracy.*
*Image Source: github.com*

# Machine Learning Algorithms: Operations



*Workflow of possible scenarios when selecting an ML algorithm.*
*Image Source: sckit-learn.org*



# Exercise: Running Introductory Notebooks in Google Colab

Trainer: Jordan A. Caraballo-Vega

# Summary

- Overview of Machine Learning
- Importance of Machine Learning targeted towards Earth Science
- Usability of Machine Learning
- Software to Support Machine Learning
- Machine Learning Applications
- Hands on Jupyter Notebook Exercise: Load and Visualize Data
- Post-Session Assignment

# Looking Ahead

## Part 2: Training Data and Land Cover Classification Example

- Download the training data
- Exploratory data analysis
- Extracting training data from a tabular dataset
- Extracting training data from a raster dataset
- Training and inference of a tabular and raster dataset
- Metrics and model evaluation
- Hands on Jupyter Notebook Exercise: MODIS Water Classification Case Study

# Contacts

- Trainers:
  - Jordan A. Caraballo-Vega: jordan.a.caraballo-vega@nasa.gov
  - Jules Kouatchou: jules.kouatchou-1@nasa.gov
  - Caleb S. Spradlin: caleb.s.spradlin@nasa.gov
  - Jian Li: jian.li@nasa.gov
  - Brock Blevins: brock.Blevins@nasa.gov

- Training Webpage:
  - https://appliedsciences.nasa.gov/join-mission/training/english/arset-fundamentals-machine-learning-earth-science

- ARSET Website:
  - https://appliedsciences.nasa.gov/arset

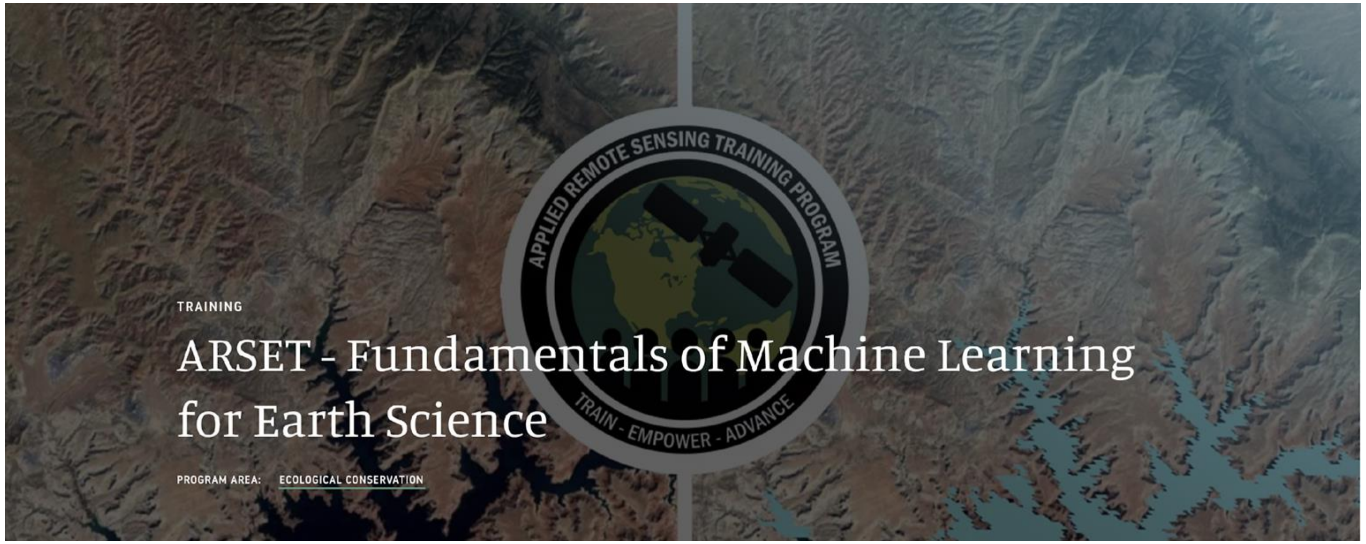
Check out our sister programs:

DEVELOP

SERVIR

# Questions?

- Please enter your question in the Q&A box. We will answer them in the order they were received.
- We will post the Q&A to the training website following the conclusion of the webinar.



TRAINING

ARSET - Fundamentals of Machine Learning for Earth Science

PROGRAM AREA: ECOLOGICAL CONSERVATION

APPLIED REMOTE SENSING TRAINING PROGRAM
TRAIN - EMPOWER - ADVANCE

APPLIED REMOTE SENSING TRAINING PROGRAM

TRAIN - EMPOWER - ADVANCE

# Thank You!

# 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