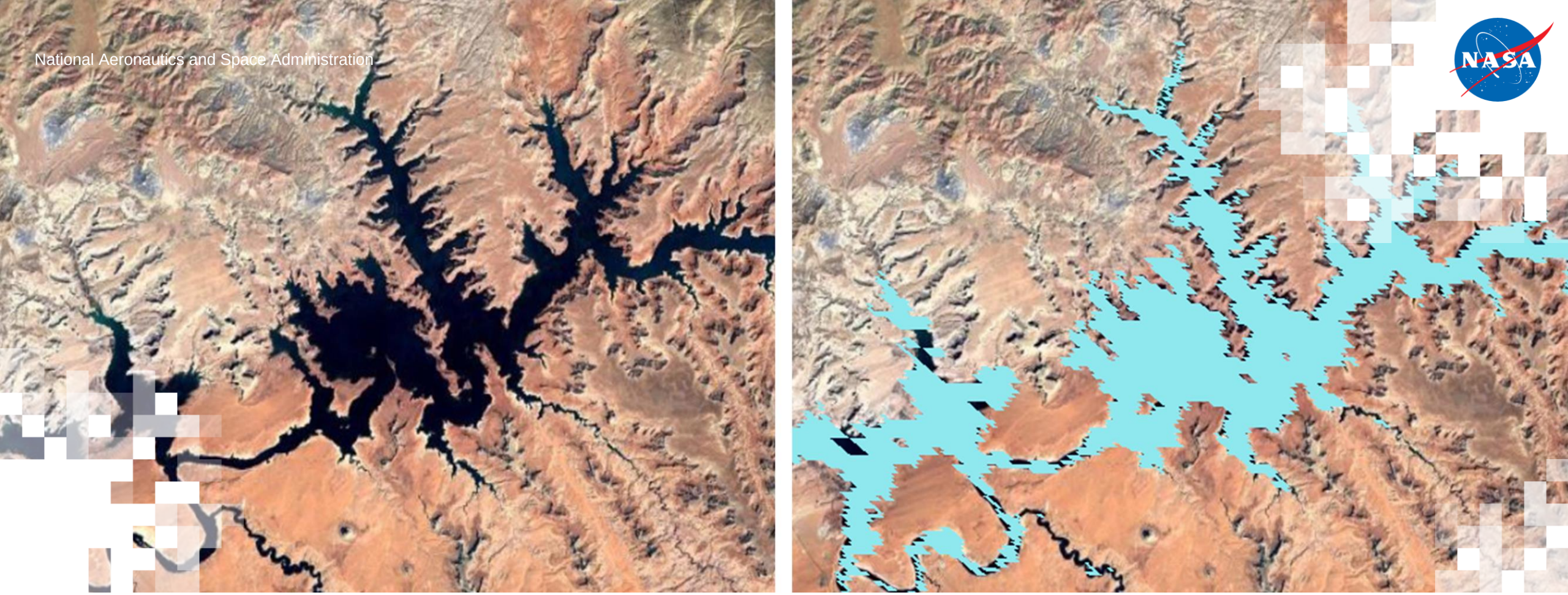


Fundamentos del Aprendizaje Automático para las Ciencias de la Tierra

1era Sesion: Introducción al Aprendizaje Automático

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

20 de abril de 2023

Objetivo del Programa de Capacitación de Teledetección Aplicada de la NASA (ARSET)

Empoderar a la comunidad mundial para incorporar datos de observaciones de la Tierra en la gestión ambiental y en la toma de decisiones


Agricultura


Clima


Desastres


Salud y Calidad del Aire


Ecosistemas Terrestres


Recursos Hídricos

Disponibilidad de las Capacitaciones de NASA ARSET

- Webinarios en línea autoguiados
- Presenciales personalizadas
- Sin costo
- Opciones multilingües
- Variedad de niveles para abordar las diferentes necesidades del público
- Los materiales son gratuitos para usar y adaptar siempre que se le dé crédito a NASA ARSET

Visite la página web de NASA ARSET para ver todas nuestras opciones

Objetivos de Capacitación

Al final de la capacitación, los participantes podrán

- Reconocer los métodos de aprendizaje automático más comúnmente utilizados para procesar datos de observación de la Tierra
- Describir los beneficios y las limitaciones del aprendizaje automático para el análisis de datos de observación de la Tierra
- Explicar cómo aplicar algoritmos y técnicas de aprendizaje automático básicos de manera significativa a datos de teledetección
- Usar datos de entrenamiento para evaluar las condiciones y soluciones para un estudio de caso determinado
- Completar los procedimientos básicos para interpretar, refinar y evaluar la precisión de los resultados del análisis de aprendizaje automático

Recordatorio- Pre-requisitos

- Prerrequisitos:
  - La Sesión 1 de nuestra serie disponible a pedido, Fundamentos de la Percepción Remota (Teledetección) o contar con experiencia equivalente (https://appliedsciences.nasa.gov/sites/default/files/2023-02/Fundamentals_of_RS_Span.pdf).
  - Los participantes necesitarán tener acceso a Google Drive y Google Colab. Para acceder a estos recursos, deben utilizar un correo que termine en 'gmail.com'.
  - Pondremos la grabación de esta sesion a su disposición dentro de 48 horas después de la presentación.

Agenda

Parte 1: Introducción al Aprendizaje Automático

20 de abril de 2023

Parte 2:

Ejemplo de Datos de Entrenamiento y Clasificación de la Cobertura Terrestre

27 de abril de 2023

Parte 3:

Ajustes de Modelos, Optimización de Parámetros y Algoritmos de Aprendizaje Automático Adicionales

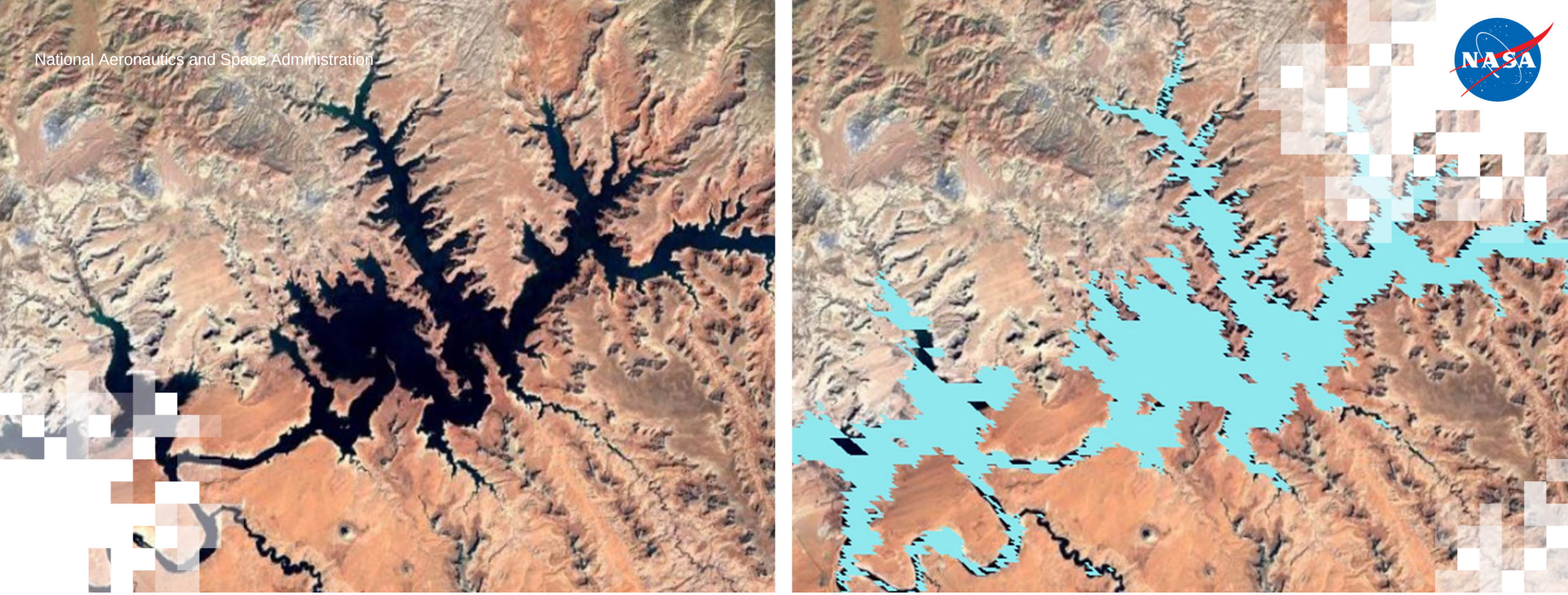
4 de mayo de 2023

Tarea

Práctica y aplicación independientes

Fecha límite: 19 de mayo
Disponible: 4 de mayo

Oportunidad opcional de ganar un certificado de finalización de curso



Fundamentos del Aprendizaje Automático para las Ciencias de la Tierra

Parte 1: Introducción al Aprendizaje Automático

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

20 de abril de 2023

Equipo de Instructores

Jordan A. Caraballo-Vega
Ingeniero Informático

Jules Kouatchou
Programador/Analista Principal

Caleb S. Spradlin
Desarrollador de Software

Mark L. Carroll
Científico Investigador

Jian Li
Ingeniero Principal Sénior de Aplicaciones

Esquema de la Sesión 1

- Visión General del Aprendizaje Automático
- Importancia del Aprendizaje Automático Dirigido a las Ciencias de la Tierra
- Utilidad del Aprendizaje Automático
- Software de Apoyo para el Aprendizaje Automático
- Aplicaciones de Aprendizaje Automático
- Ejercicio Práctico de Jupyter Notebook: Cargar y Visualizar Datos
- Tarea para Después de la Sesión
- Sesión de Preguntas y Respuestas

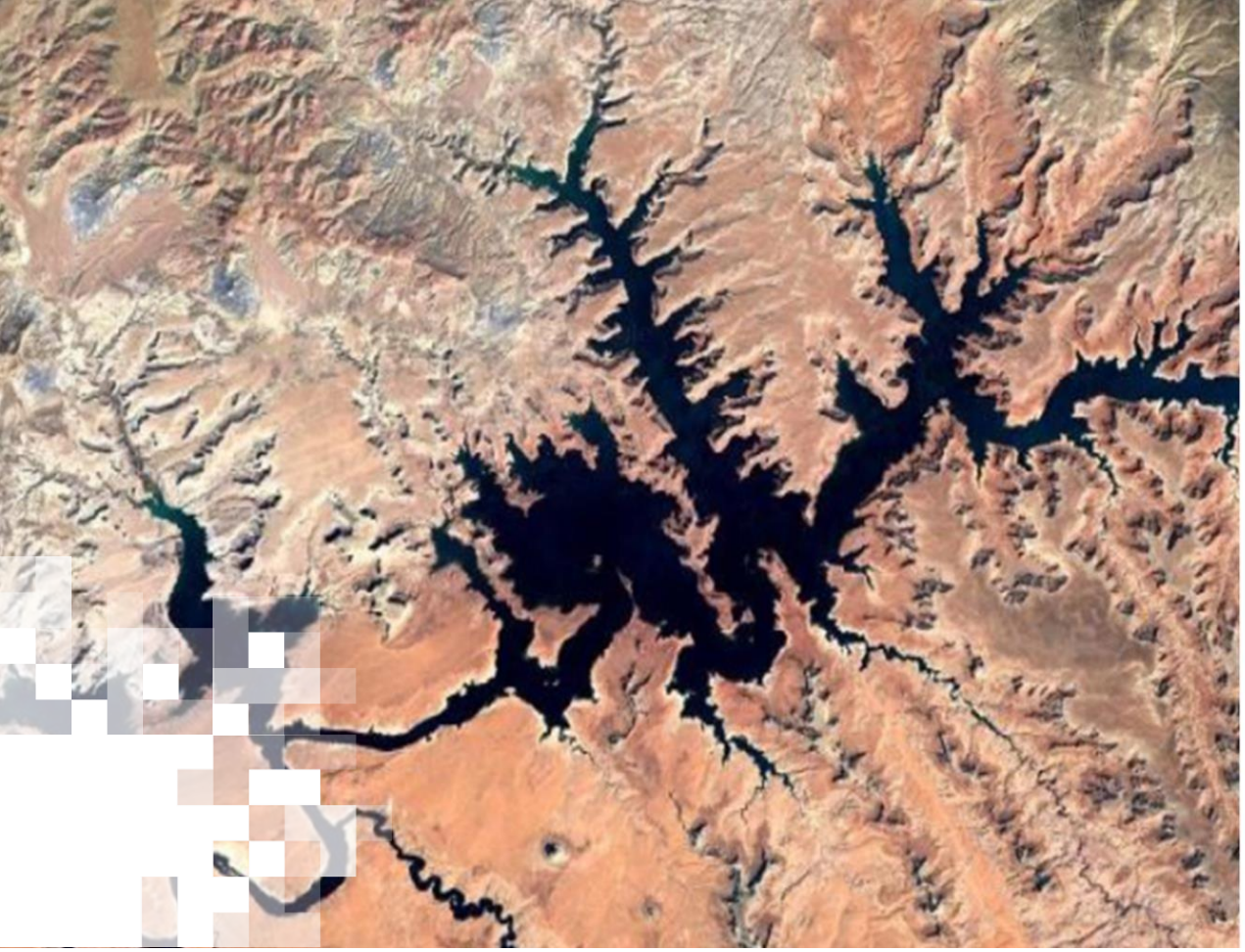
Recursos para Esta Capacitación

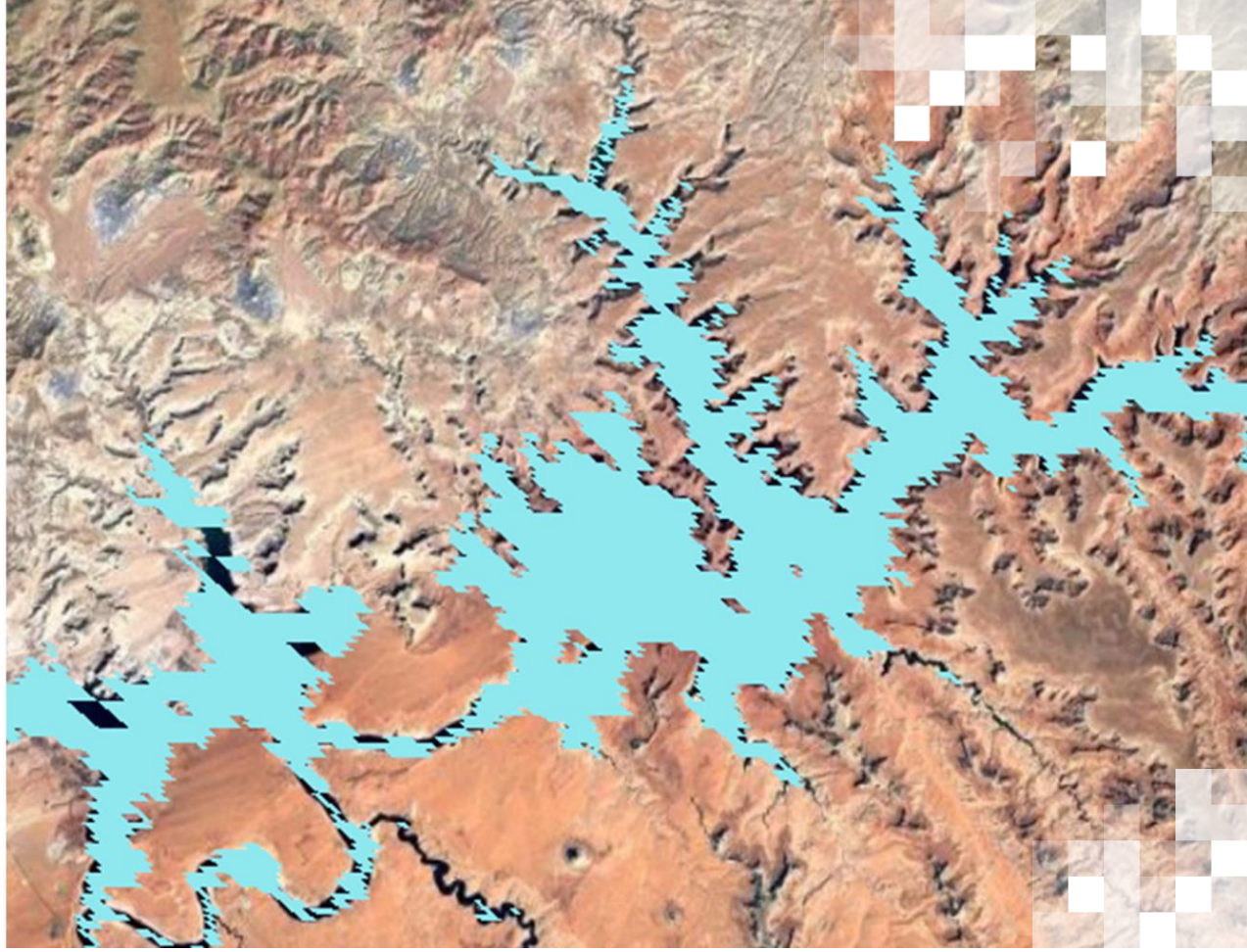
https://github.com/NASAARSET/ARSET_ML_Fundamentals

Objetivos de la Capacitación

Después de participar en esta capacitación, quienes asistan podrán:

- Reconocer los métodos de aprendizaje automático más comúnmente utilizados para procesar datos de Ciencias de la Tierra
- Describir los beneficios y las limitaciones del aprendizaje automático para el análisis de las Ciencias de la Tierra
- Explicar cómo aplicar algoritmos y técnicas de aprendizaje automático básicos de manera significativa a datos de teledetección





Visión General y Teoría

Formador: Jules Kouatchou

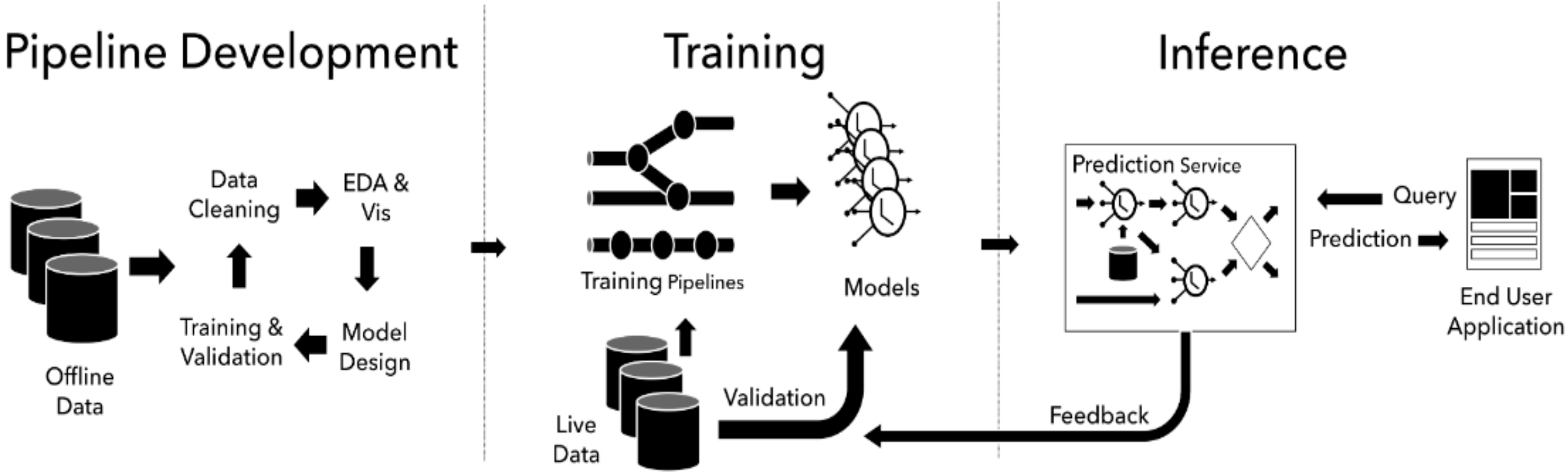
Introducción al Aprendizaje Automático

La siguiente citación de *Arthur Samuel*, describe lo que es el Aprendizaje Automático(Machine Learning o ML):

> *"El aprendizaje automático permite que una máquina* ***aprenda automáticamente de los datos,***
> ***mejore su rendimiento a partir de sus experiencias*** *y*
> ***prediga cosas sin ser explícitamente programada****."*

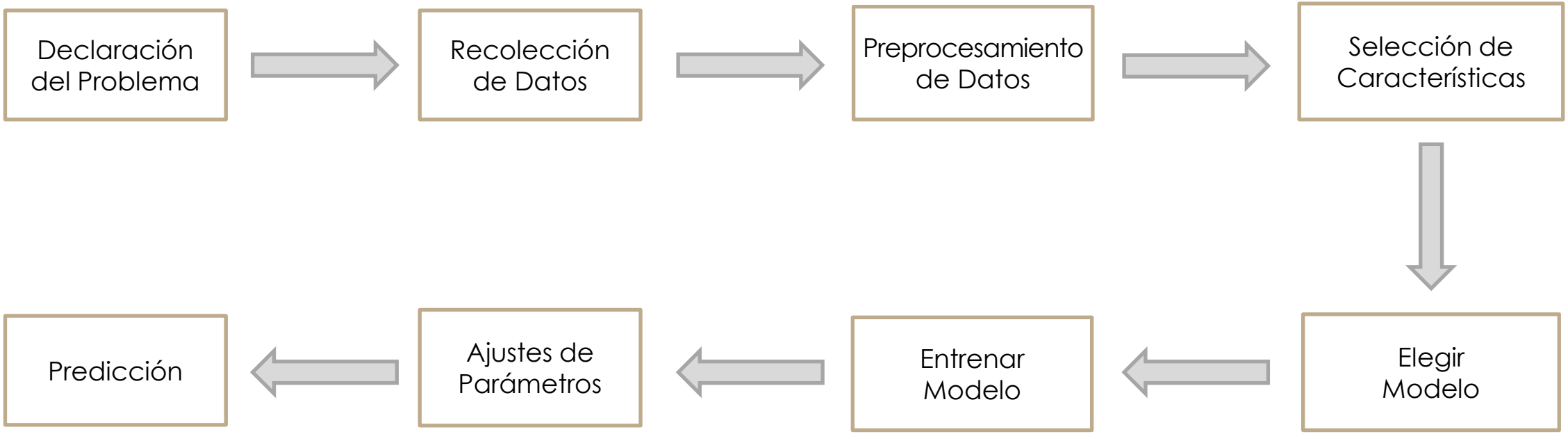
El ML usa técnicas de la estadística, matemática e informática para hacer que los programas de computación aprendan de los datos para predecir una salida.

¿Cómo Funciona el Aprendizaje Automático?

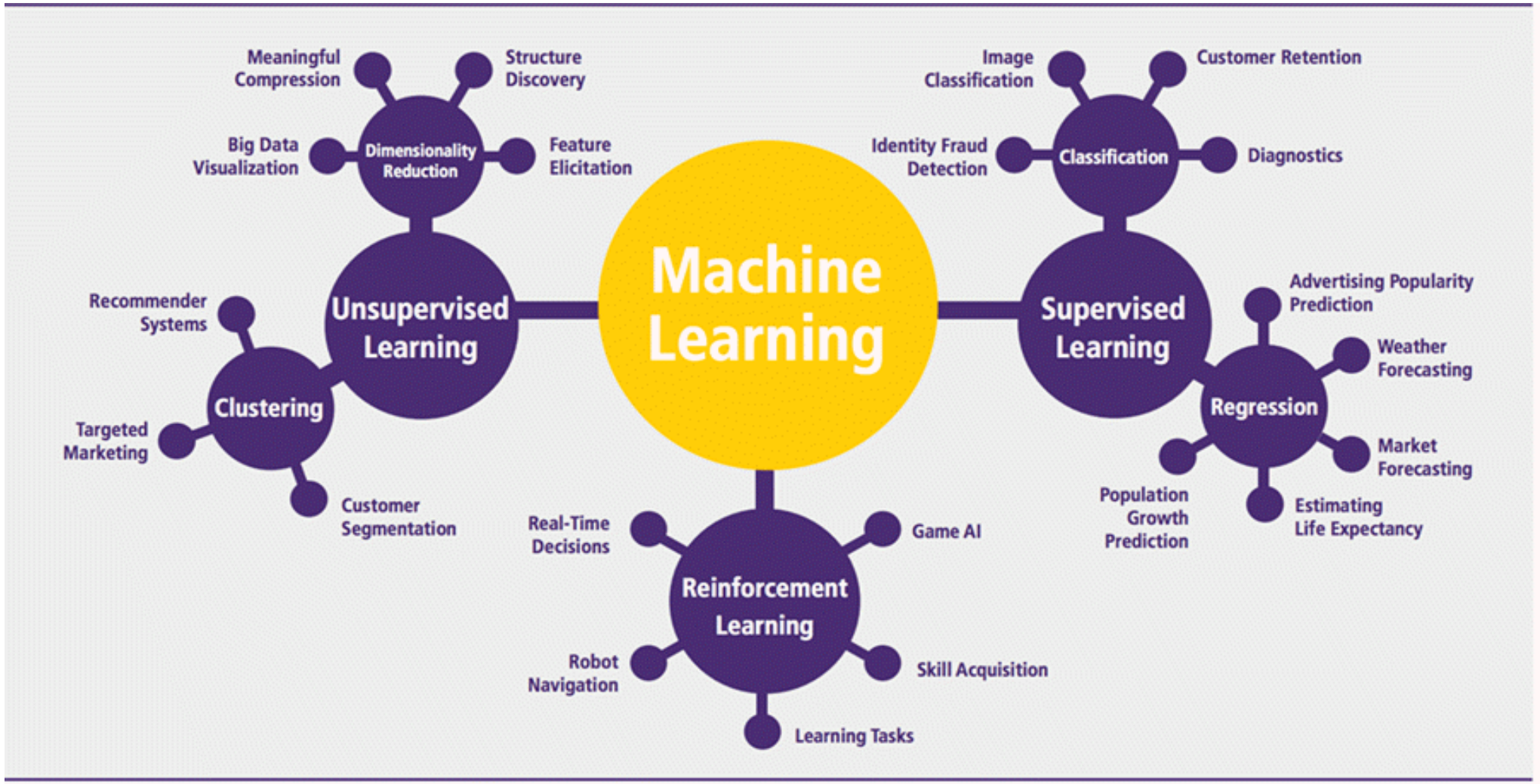


Fuente de la Imagen: Daniel Crankshaw (en *A Short History of Prediction-Serving Systems*)

Pasos del Aprendizaje Automático

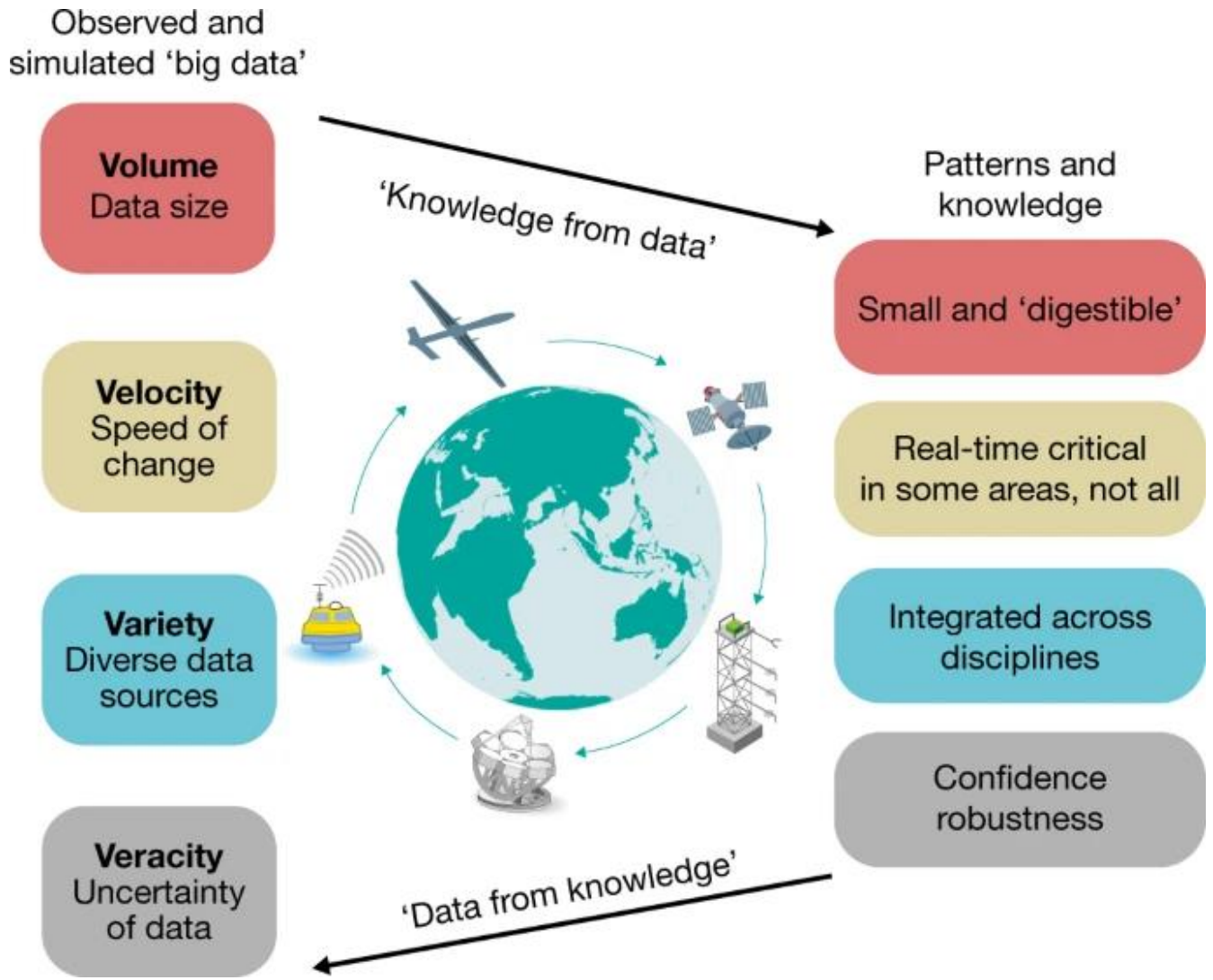


Algoritmos de Aprendizaje Automático



Fuente de la Imagen: guru99.com

Grandes Datos en las Ciencias de la Tierra



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

El Aprendizaje Automático en las Ciencias de la Tierra

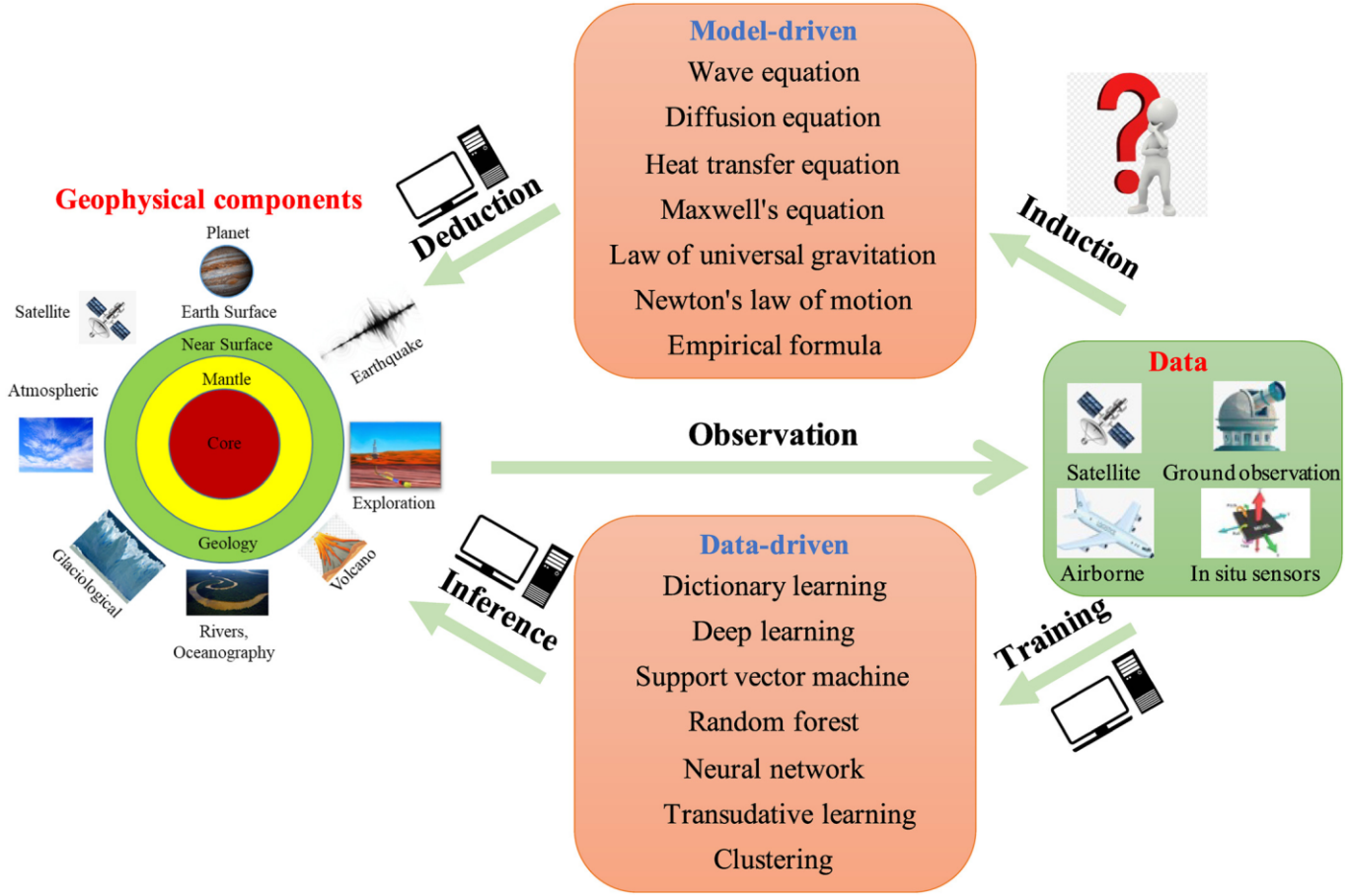
El aprovechar de los avances en la inteligencia artificial podría desencadenar una revolución en las ciencias de la Tierra y del medio ambiente. Debemos asegurarnos que nuestras elecciones en cuanto a financiación y capacitación le den la capacidad para realizar este potencial a la próxima generación de geocientíficos.

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

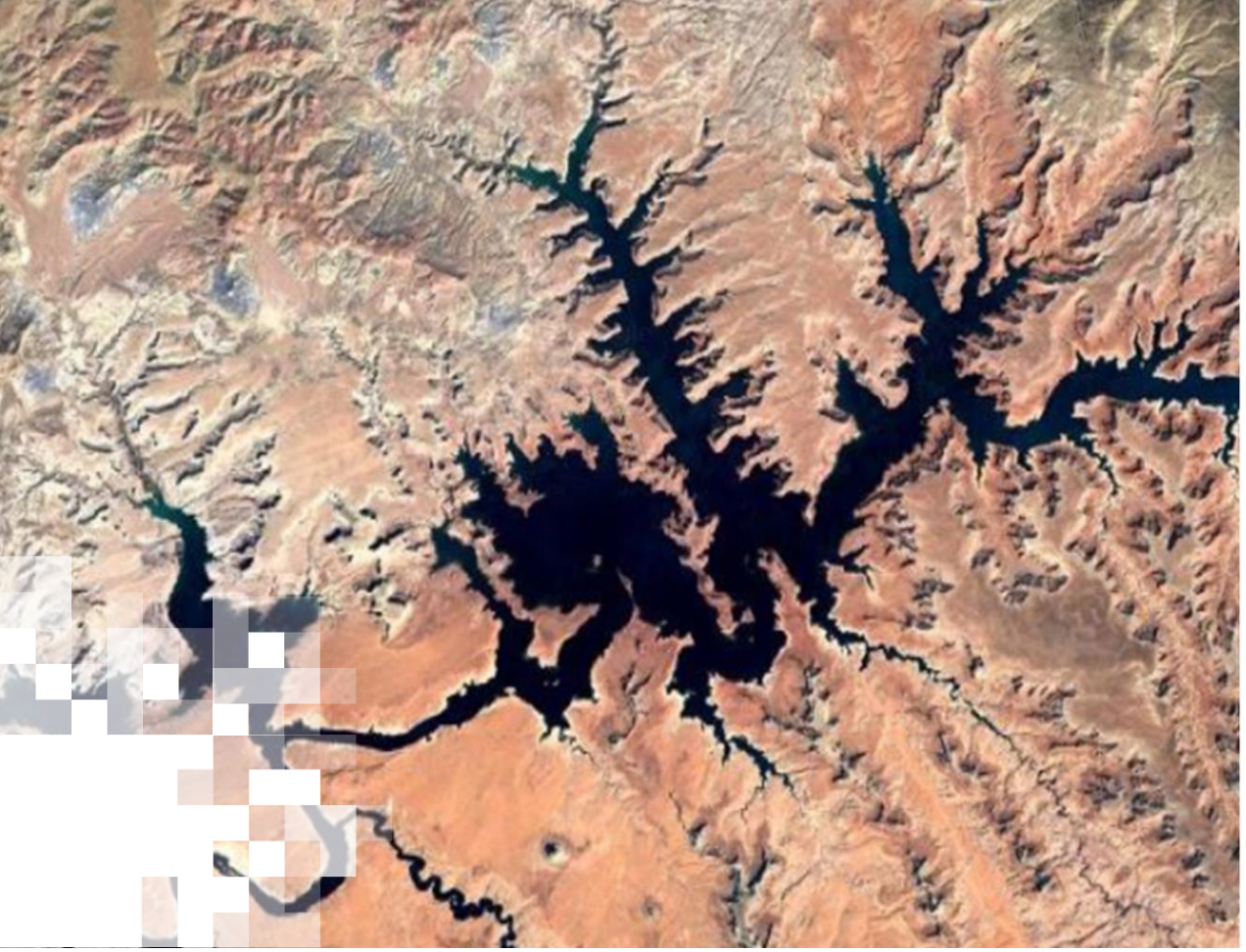
El Aprendizaje Automático en las Ciencias de la Tierra

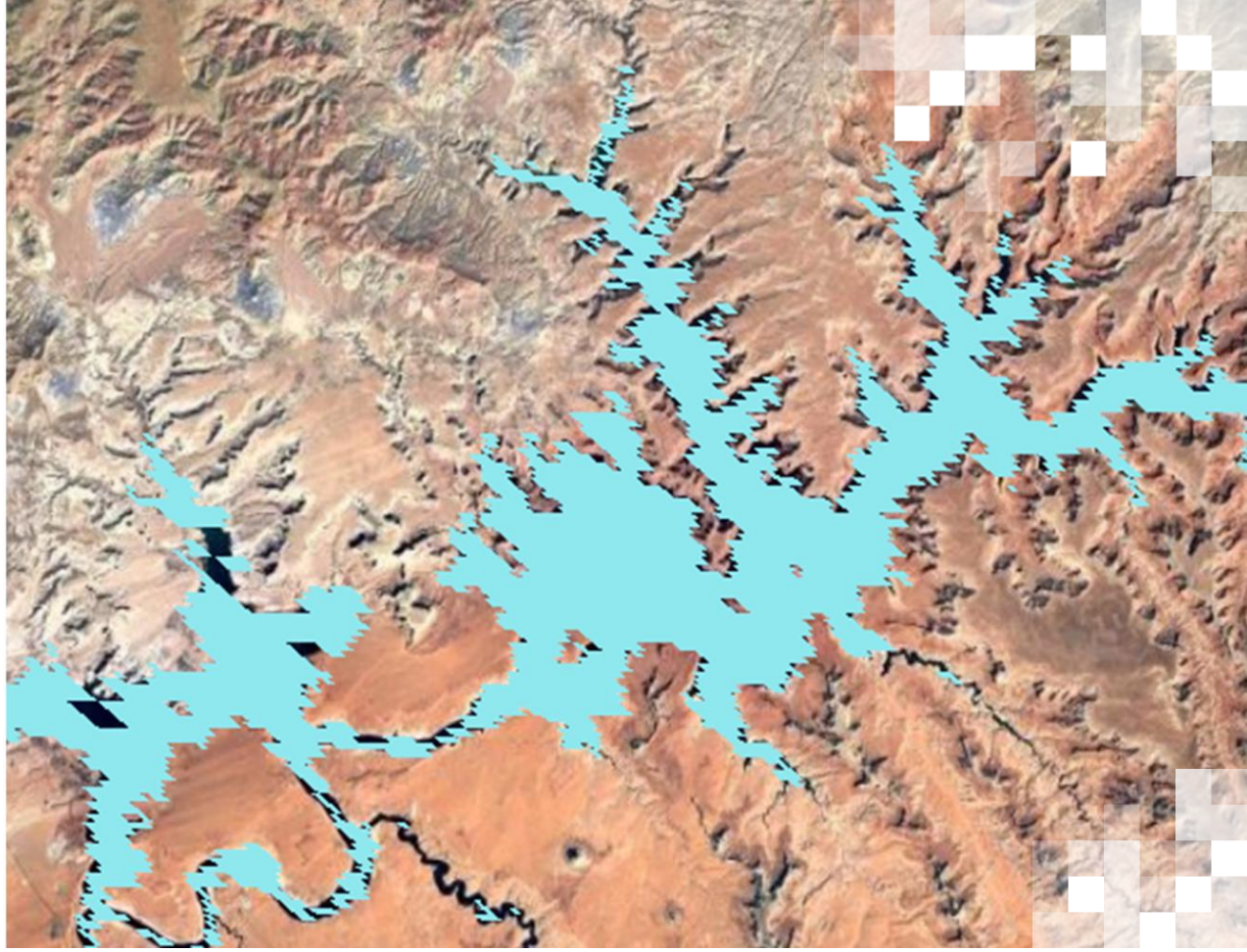
- Los problemas en la ciencias de la Tierra a menudo son complejos.
- Es difícil aplicar modelos matemáticos bien conocidos y descritos al entorno natural:
  - El ML es comúnmente una mejor alternativa para semejantes problemas no lineales.
- Un número de investigadores encontró que el aprendizaje automático tiene mejor rendimiento que los modelos estadísticos tradicionales en las ciencias de la Tierra, por ejemplo en:
  - La caracterización de la estructura del dosel arbóreo,
  - La predicción de cambios de rangos geográficos inducidos por el clima
  - La delineación de facies geológicos.

Cómo se Aplica el Aprendizaje Automático en las Ciencias de la Tierra



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





Aplicaciones de Aprendizaje Automático

Instructor: Jian Li

Beneficios de Utilizar el Aprendizaje Automático

Hay varias formas en las que el ML puede acelerar la investigación científica, como por ejemplo:

- Mayor eficiencia: el aprendizaje automático puede ayudar a automatizar el análisis de conjuntos de datos grandes y complejos, lo que permite a los científicos procesar y analizar grandes cantidades de datos rápidamente.
- Nuevos conocimientos y descubrimientos: el aprendizaje automático puede ayudar a los científicos a identificar nuevos patrones y relaciones en conjuntos de datos complejos, lo que lleva a nuevos conocimientos y descubrimientos en la investigación de las Ciencias de la Tierra.
- Modelación predictiva mejorada: los algoritmos de aprendizaje automático se pueden usar para crear modelos predictivos precisos que pueden ayudar a los científicos a comprender mejor los fenómenos complejos de las Ciencias de la Tierra

Eficiencia, Precisión y Descubrimiento

Identificar nuevas estrellas y sistemas estelares a partir de un número masivo de observaciones

- Una misión de reconocimiento de todo el cielo, llamada el **Transiting Exoplanet Survey Satellite (TESS)**
- Usando herramientas de IA, ML y HPC, los científicos de la NASA han extraído más de 60 millones de curvas de luz para una mayor investigación.
- Los astrónomos de la NASA han identificado
  - **> 50** candidatos a planetas
  - **> 200** posibles estrellas latido de corazón
  - **> 10** posibles sistemas estelares triples
  - **> 20** posibles sistemas estelares cuádruples
  - un posible sistema estelar séxtuple
- Todos anteriormente sin descubrir

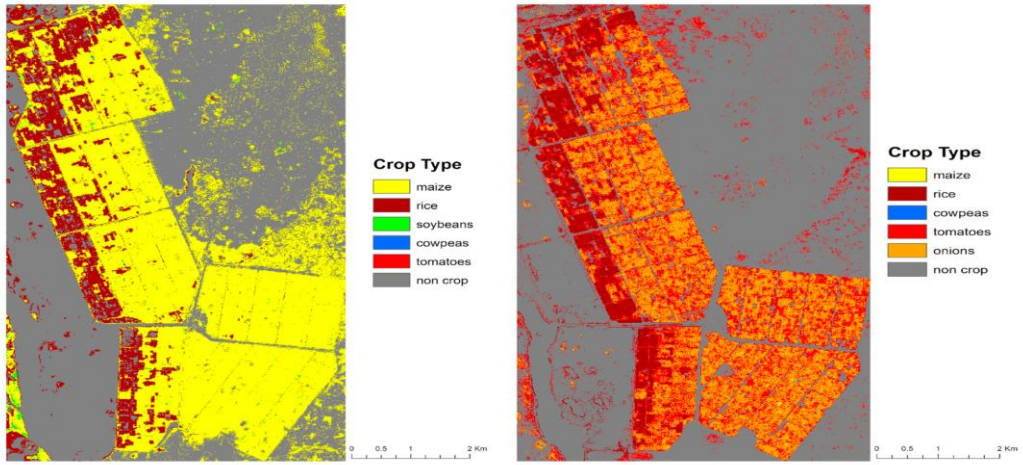
Una proyección bidimensional del espacio de alta dimensión de las representaciones de las curvas de luz TESS. Fuente de la Imagen: Brian P. Powell, NASA Goddard.

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

Eficiencia, Precisión y Descubrimiento

Estimaciones del rendimiento y tipo de cultivo basadas en aprendizaje automático en Burkina Faso, África Occidental

Illustration by ESA.



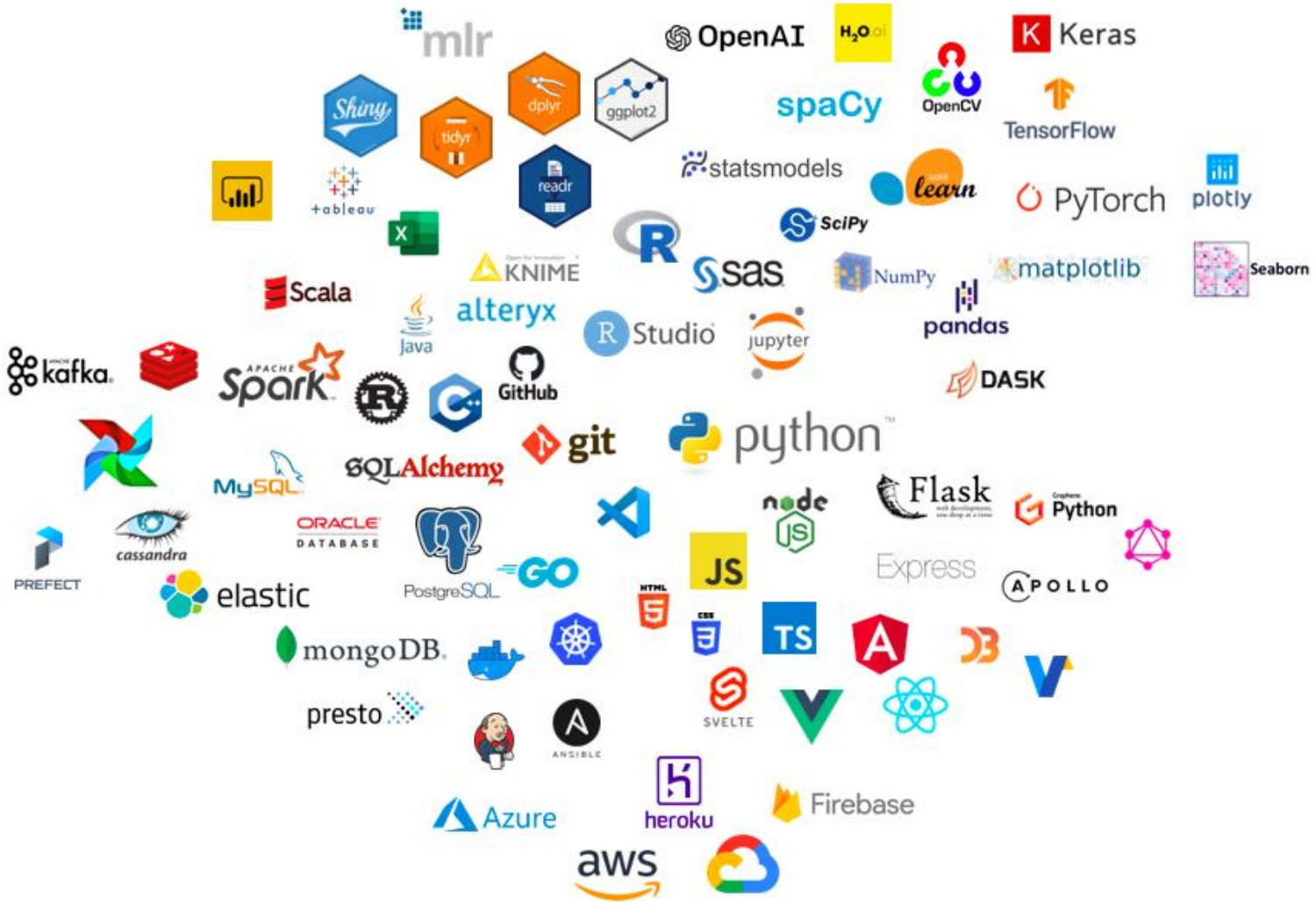
Predicciones del tipo de cultivo ML en la región de estudio de ~2250 hectáreas para las temporadas de lluvia (izquierda) y seca (derecha) de 2019. En la época de lluvia predomina el maíz (amarillo) y el arroz (granate), mientras que en la estación seca predominan la cebolla (naranja), el tomate (rojo) y el arroz (granate).

- NASA Goddard y la Millennium Challenge Corporation (MCC)
- Invierten en proyectos de desarrollo agrícola para empoderar a agricultores locales y combatir la inseguridad alimentaria.
- Datos del satélite *Sentinel-2* e in situ para entrenar y optimizar cinco modelos de aprendizaje automático Random Forest para estimar el tipo y rendimiento de los cultivos a lo largo de la región de estudio
- La precisión del modelo fue de un **88%** para tipos de cultivo, mayor al **64%** para rendimiento de los cultivos durante la época de lluvia del 2019 y de un **64%** para tipos de cultivo y más del **>53%** para rendimiento de los cultivos durante el tiempo seco del 2019
- El modelo de aprendizaje automático formuló predicciones para el tiempo seco del 2020 sin datos de entrenamiento; las precisiones fueron de hasta un **60%** para tipos de cultivo.

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

Software para Apoyar el Aprendizaje Automático

- Lenguajes de programación
- Paquetes de software



Fuente de la Imagen: https://dev.to/minchulkim87/my-data-science-tech-stack-2020-1poa

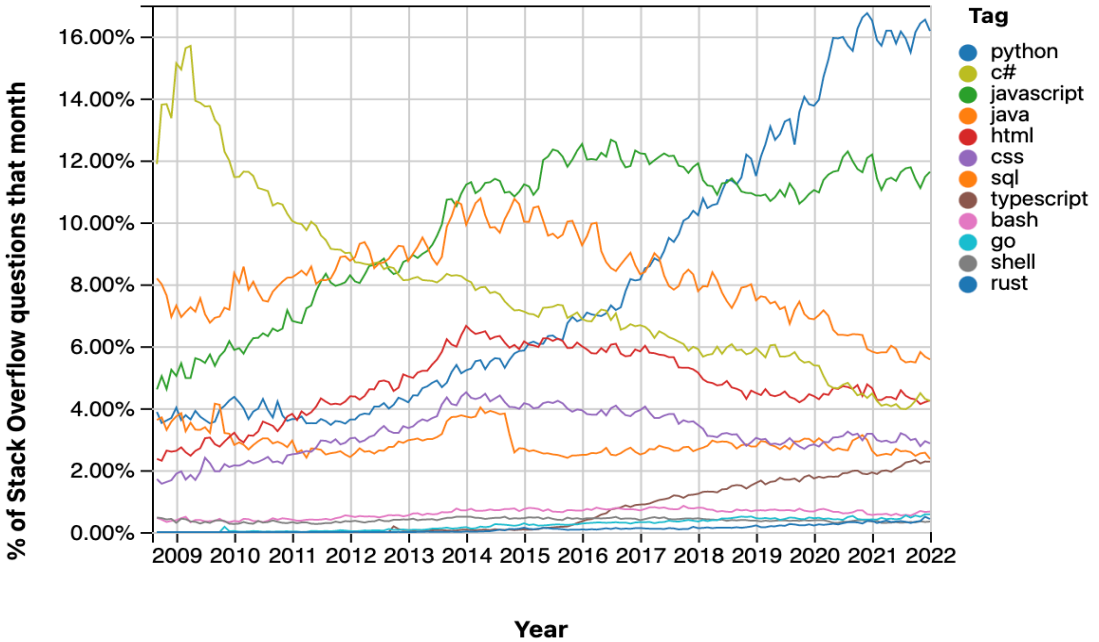
Software para Apoyar el Aprendizaje Automático, Continuación

Python: es el lenguaje más utilizado para el aprendizaje automático. Una de las principales razones por las que Python es tan popular en el desarrollo de la IA es porque se creó como una poderosa herramienta de análisis de datos y siempre ha sido popular en el campo de los grandes datos.

R: R puede que no sea el lenguaje perfecto para la IA, pero es fantástico para procesar números muy grandes, cosa que lo hace mejor que Python a escala. Y con la programación funcional integrada de R, la computación vectorial y la naturaleza orientada a los objetos, se convierte en un lenguaje viable para la IA.

Java: Es un lenguaje importante para la IA. Una de las razones es la prevalencia del lenguaje en el desarrollo de aplicaciones móviles. Y dada la cantidad de aplicaciones móviles que aprovechan la IA, es una combinación perfecta.

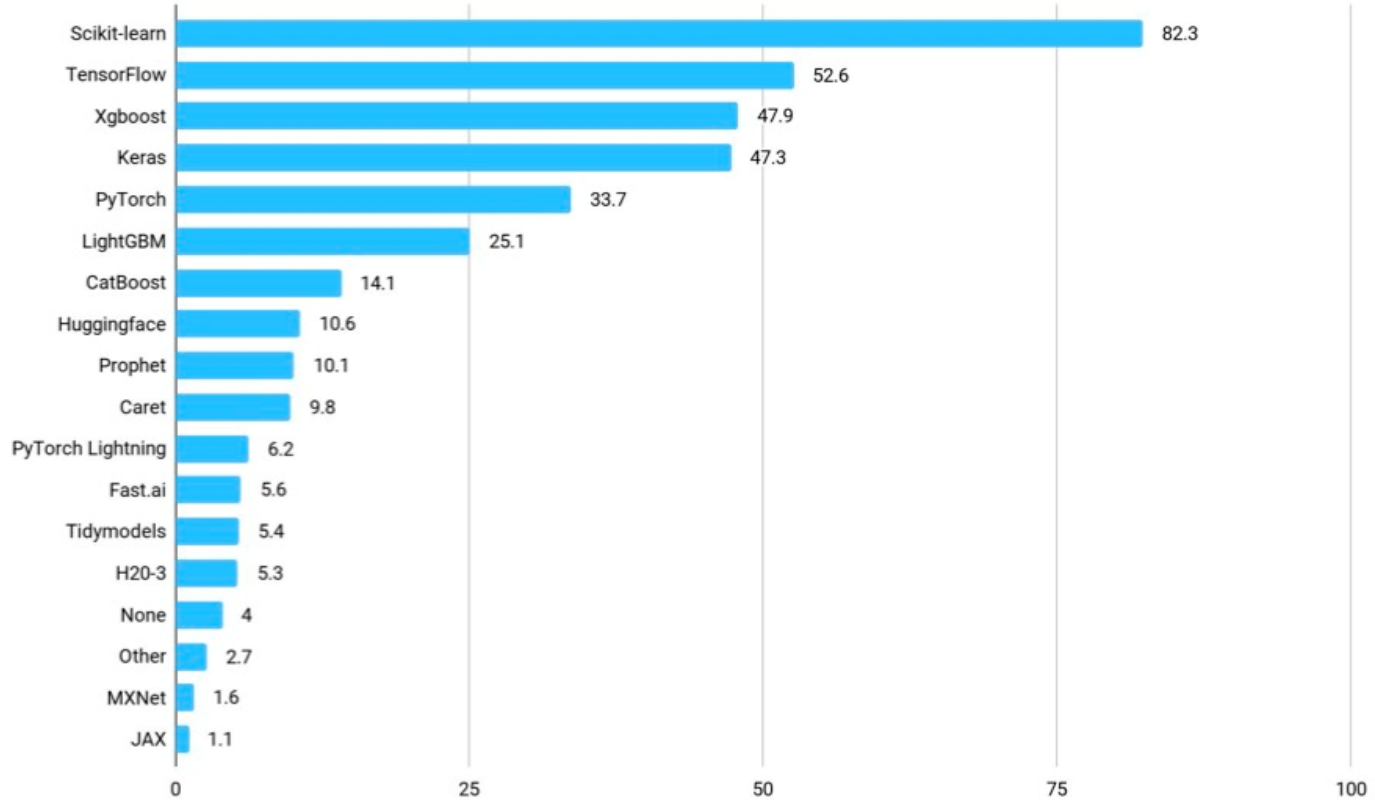
Julia: Julia es uno de los lenguajes más nuevos de la lista y se creó para centrarse en la informática del rendimiento en los campos científico y técnico.



Fuente de la Imagen: Stack Overflow 2022

Marcos de Aprendizaje Automático en Python

- Las herramientas basadas en Python predominan los marcos de aprendizaje automático según la encuesta *Kaggle's 2021 State of Data Science and Machine Learning survey*
- Scikit-learn está en primer lugar, siendo utilizado por más del 80% de científicos de datos
- TensorFlow y Keras fueron elegidos por aproximadamente la mitad de los científicos de datos para el aprendizaje profundo cada uno
- La biblioteca que aumenta el gradiente, XGBoost está en el cuarto lugar



Fuente de la Imagen: Kaggle's 2021 State of Data Science and Machine Learning survey

Marcos de Aprendizaje Automático en Python, Continuación

- **Scikit-Learn**: una de las bibliotecas más importantes (Swiss Knife) para el aprendizaje automático, ya que proporciona una serie de herramientas simples y eficientes para el análisis de datos. Proporciona funcionalidad para la clasificación, regresión, algoritmos de agrupamiento, reducción de dimensionalidad, selección de modelos y preprocesamiento de datos.
- **TensorFlow**: la biblioteca fue desarrollada por ingenieros e investigadores que trabajan en el equipo de Google Brain que lleva a cabo investigaciones sobre el aprendizaje automático y redes neuronales. Permite a los investigadores ampliar los límites en el descubrimiento de resultados de última generación (SOTA por sus siglas en inglés), y también permite a los desarrolladores crear aplicaciones basadas en el ML.
- **Keras**: API de redes neuronales de alto nivel, que se puede implementar sobre TensorFlow o Theano para crear y entrenar modelos de aprendizaje profundo. Permite la creación de prototipos fácil y rápida y es compatible con redes neuronales convolucionales y redes recurrentes.
- **PyTorch**: proporciona una funcionalidad centrada en gran medida en la creación y el entrenamiento de redes neuronales, la columna vertebral del aprendizaje profundo. PyTorch ofrece entrenamiento distribuido escalable de modelos en una o varias CPU y GPU. El primer lanzamiento fue en septiembre de 2016, pero rápidamente ha sido ampliamente adoptado por empresas en la industria como Tesla y Uber.

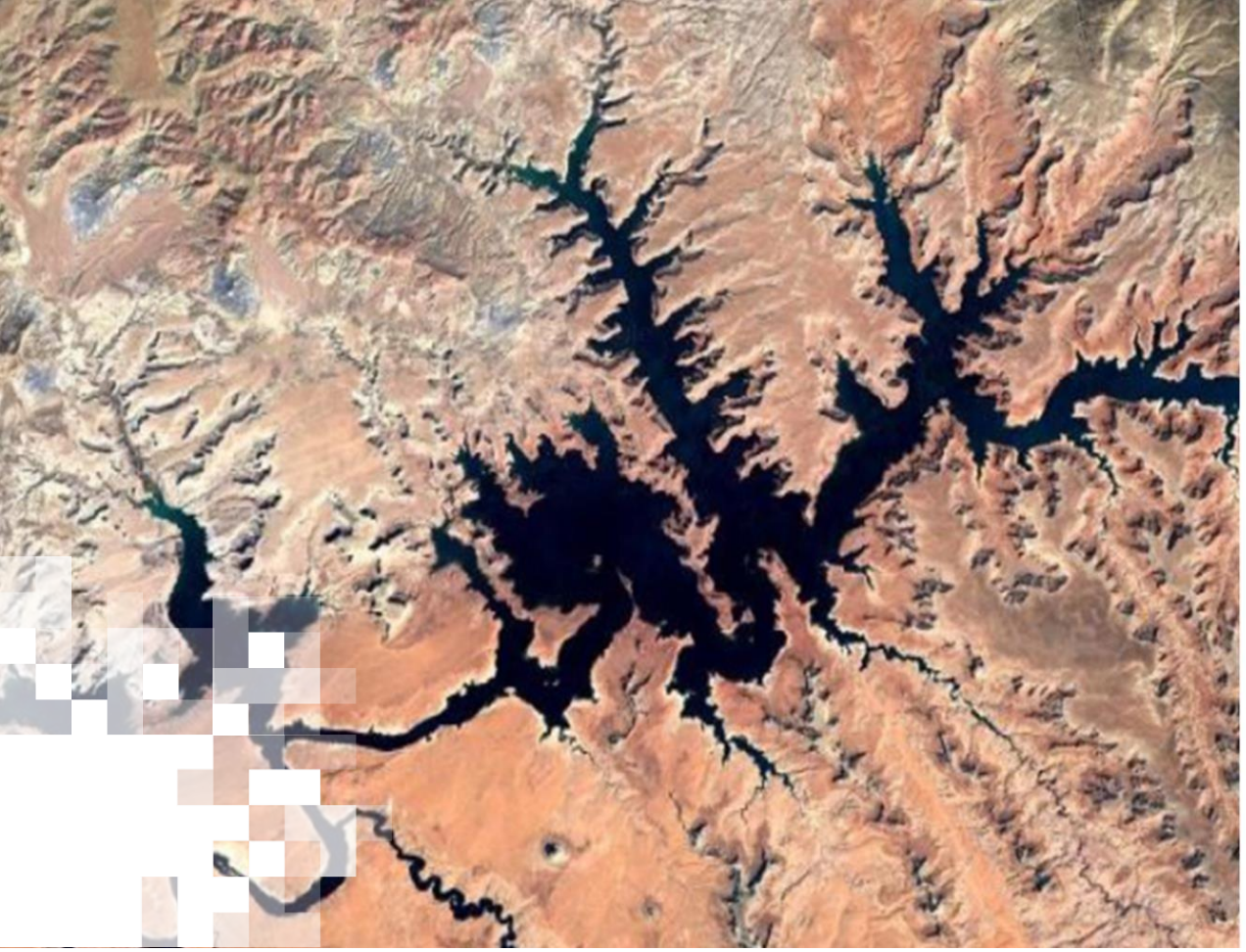
Marcos de Aprendizaje Automático en Python, Continuación

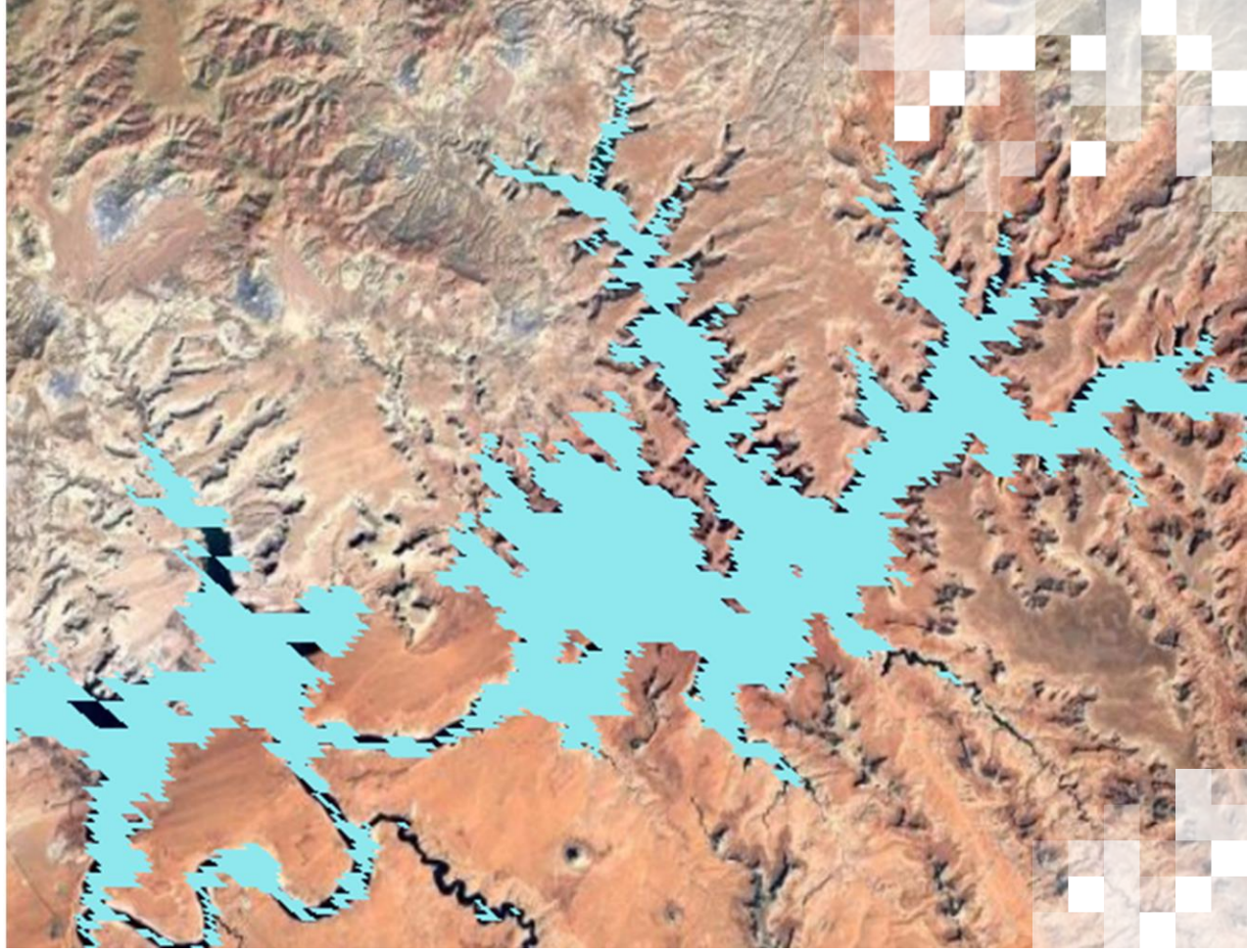
- **Jupyter Notebook** — Aplicación web de código abierto que nos permite crear y compartir documentos que contienen código, ecuaciones, visualizaciones y texto. Sus usos incluyen la limpieza y transformación de datos, modelación estadística, visualización de datos, aprendizaje automático etc.
- **Matplotlib** — Biblioteca de visualización de datos que se utiliza para crear visualizaciones estáticas, animadas e interactivas. Se puede utilizar para crear diagramas de dispersión detallados, histogramas, gráficos de barras, gráficos circulares etc.
- **Seaborn** — Biblioteca de visualización estadística basada en Matplotlib y está integrada con estructuras de datos de pandas. Proporciona una interfaz de alto nivel para gráficos informativos y estadísticos. Puesto que está construido sobre Matplotlib, ofrece gráficos adicionales y puede producir visualizaciones más sofisticadas.

El Papel de la Unidad de Procesamiento de Gráficos (Graphics Processing Unit o GPU) en el Aprendizaje Automático

- Hay muchas plataformas disponibles para computación y programación en paralelo. Entre todos ellos, **CUDA** (de NVIDIA) es la plataforma más popular por los siguientes motivos:
  - CUDA funciona tanto en Windows como en Linux.
  - Casi todas las bibliotecas de Python apoyadas por GPU como CatBoost, TensorFlow, Keras, PyTorch, OpenCV, CuPy fueron diseñadas para funcionar con tarjetas gráficas que utilizan NVIDIA CUDA.

- **Bibliotecas Populares de Python Apoyadas por GPU:**
  - XGBoost
  - OpenCV
  - cuML (Parte de RAPIDS)
  - cuDF (Parte de RAPIDS)
  - CuPy (NumPy para GPU)



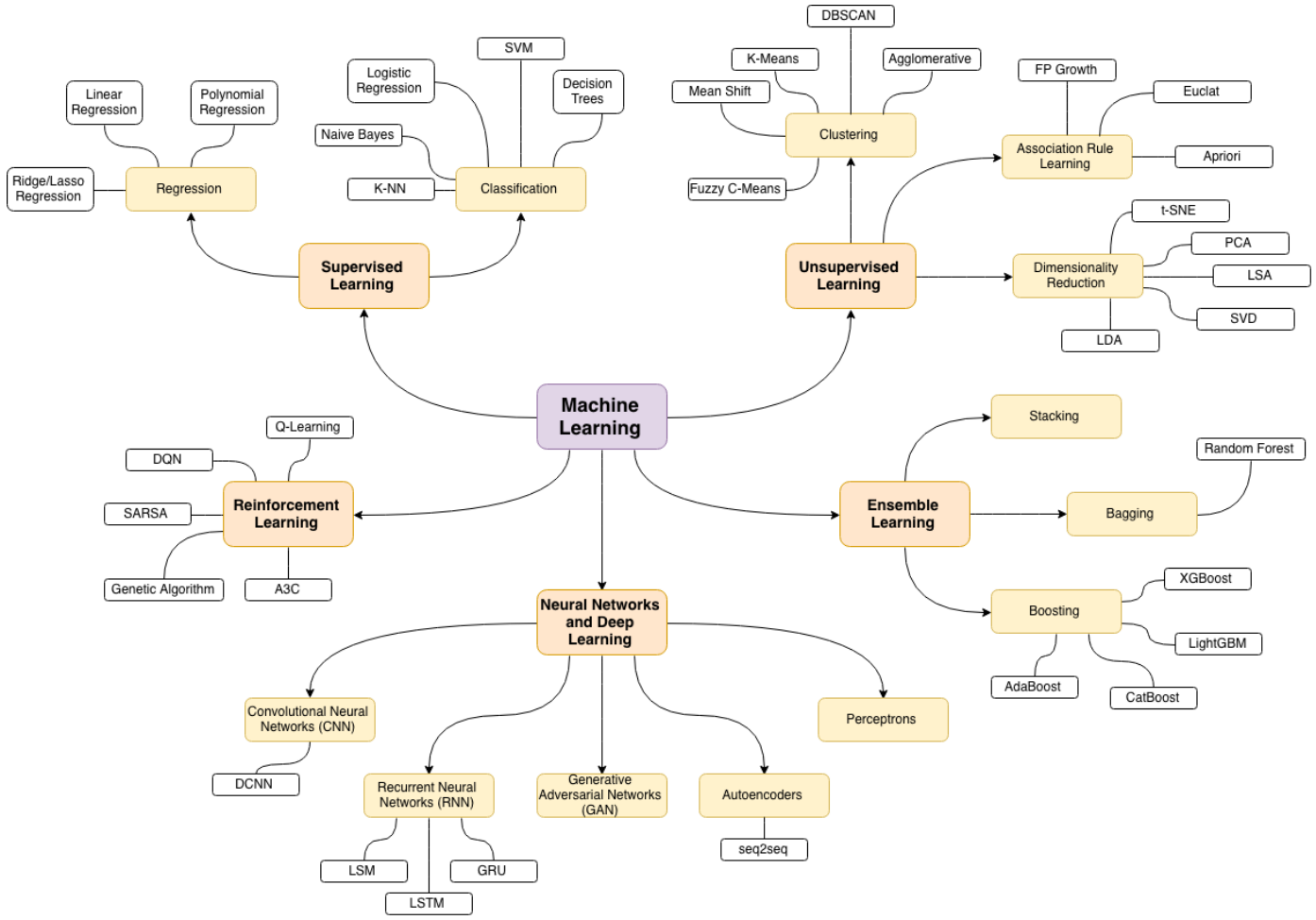


Visión General de los Algoritmos de Aprendizaje Automático

Formador: Jordan A. Caraballo-Vega

Algoritmos de Aprendizaje Automático: ¿Cuál Algoritmo Elegir?

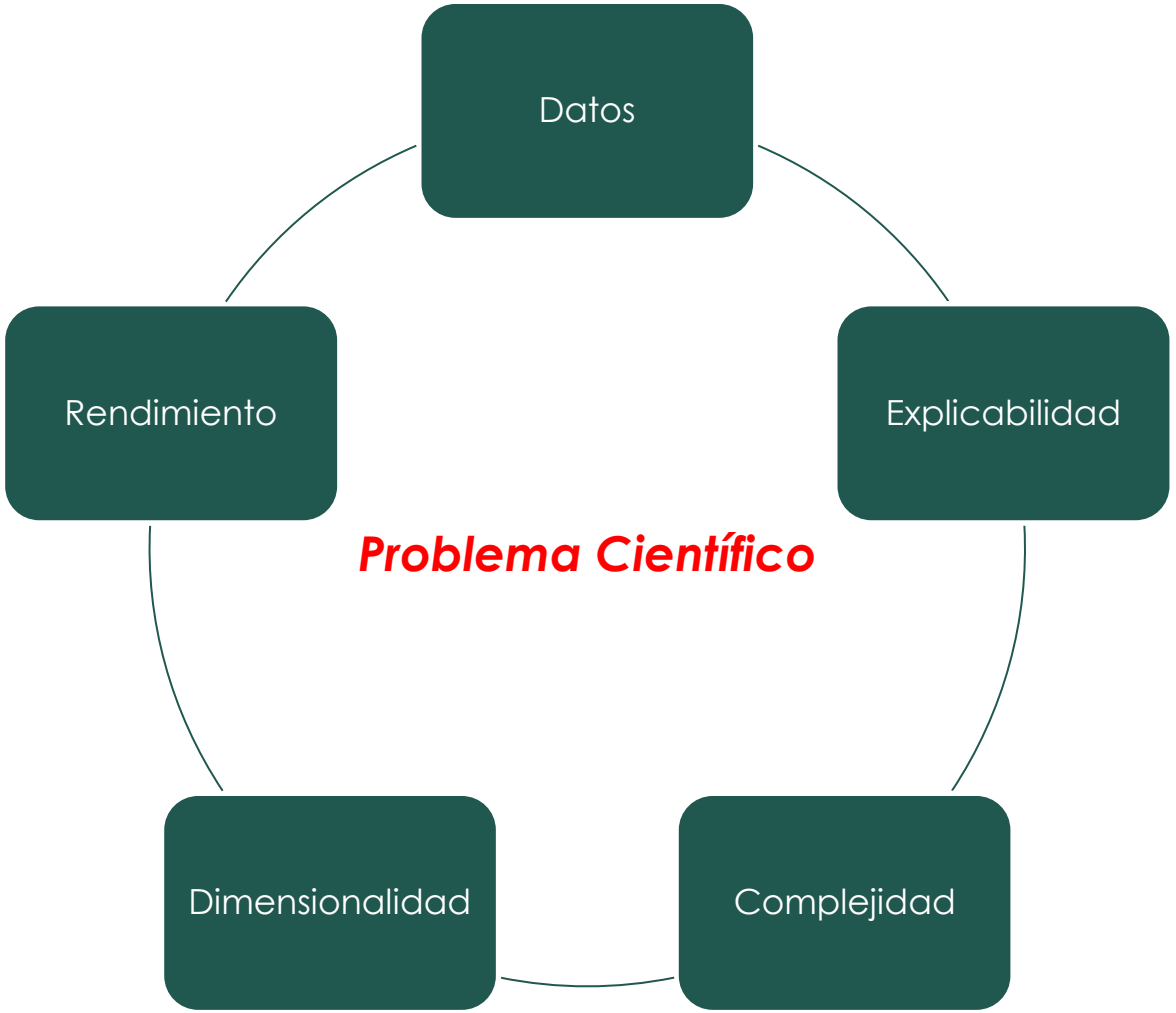
- El desarrollo de algoritmos de aprendizaje automático ha ido aumentando de manera exponencial.
- No profundizaremos en los detalles de cada algoritmo, sino que le daremos las herramientas para ayudarle en la selección de estos para su(s) propio(s) problema(s) científico(s).
- Uno no está limitado a un solo algoritmo, pero siempre ahorra tiempo el comenzar desde una base lógica.



Algoritmos de aprendizaje automático principales. Fuente de la Imagen: github.com

Algoritmos de Aprendizaje Automático: Problema Científico

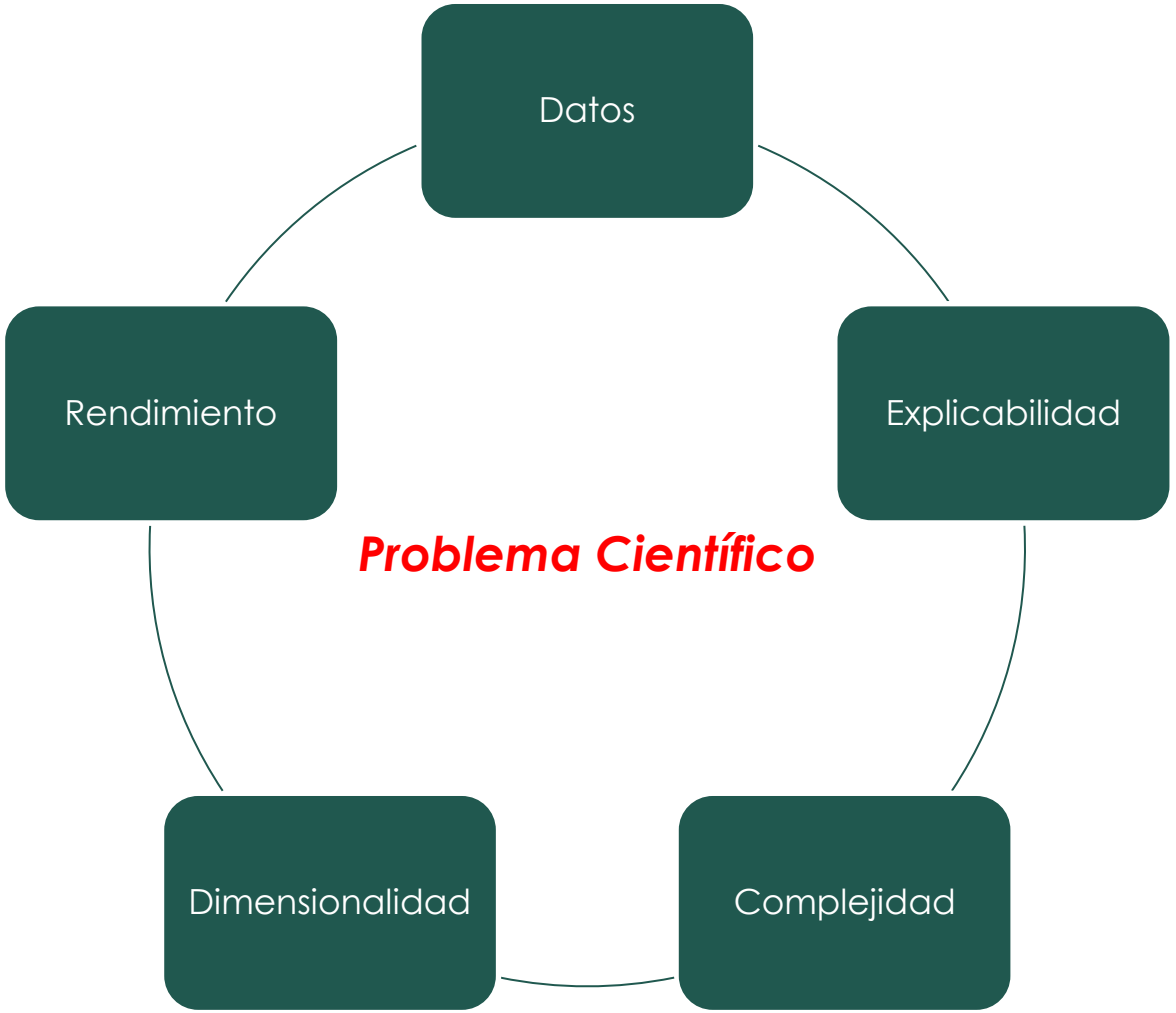
- **¿Cuál pregunta científica le gustaría abordar?**
- **¿Qué información falta para poder responder esta pregunta?**



Componentes para ayudar con la selección de su algoritmo de ML.

Algoritmos de Aprendizaje Automático: Problema Científico

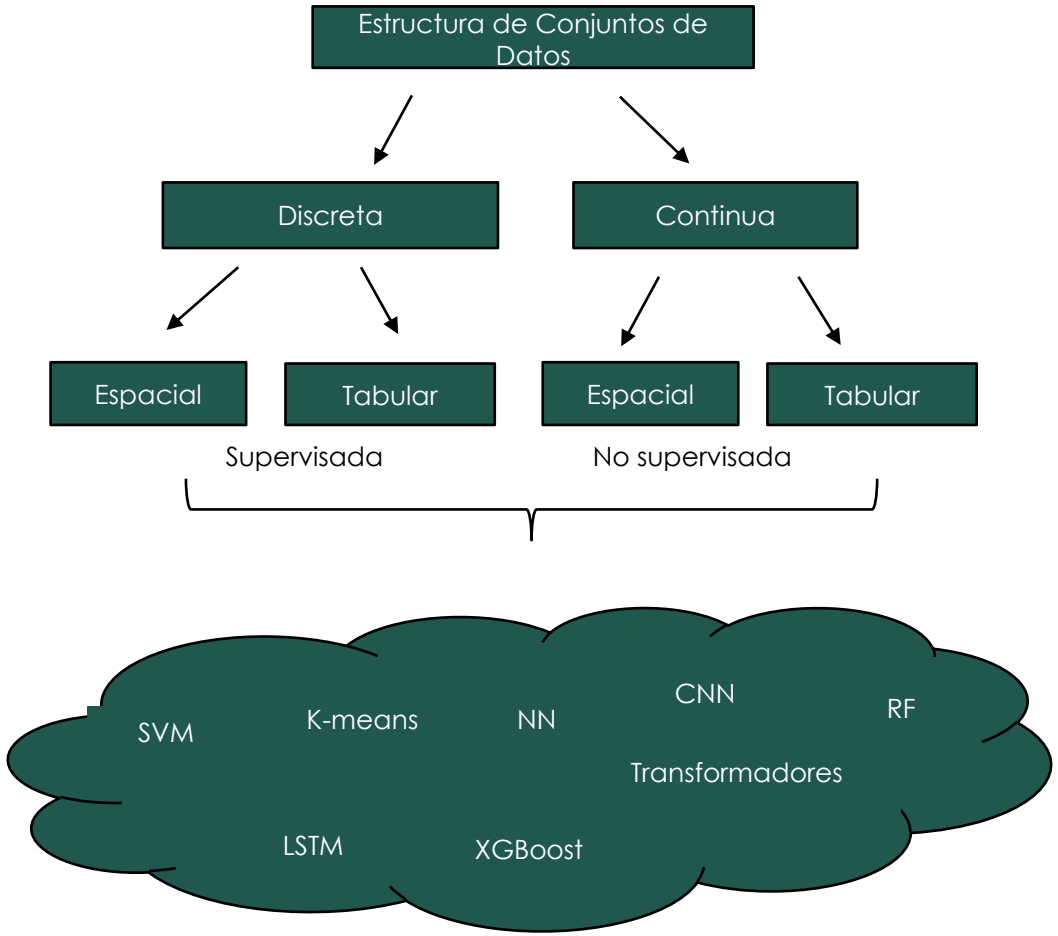
- **¿Cuál pregunta científica le gustaría abordar?** Queremos identificar el signo, la magnitud y los posibles impulsores de cambios en la extensión de aguas superficiales en el área de estudio X.
- **¿Qué información falta para poder responder esta pregunta?** Necesitamos mapas de la extensión de aguas superficiales para cuantificar y analizar estos impulsores.



Componentes para ayudar con la selección de su algoritmo de ML.

Algoritmos de Aprendizaje Automático: Datos

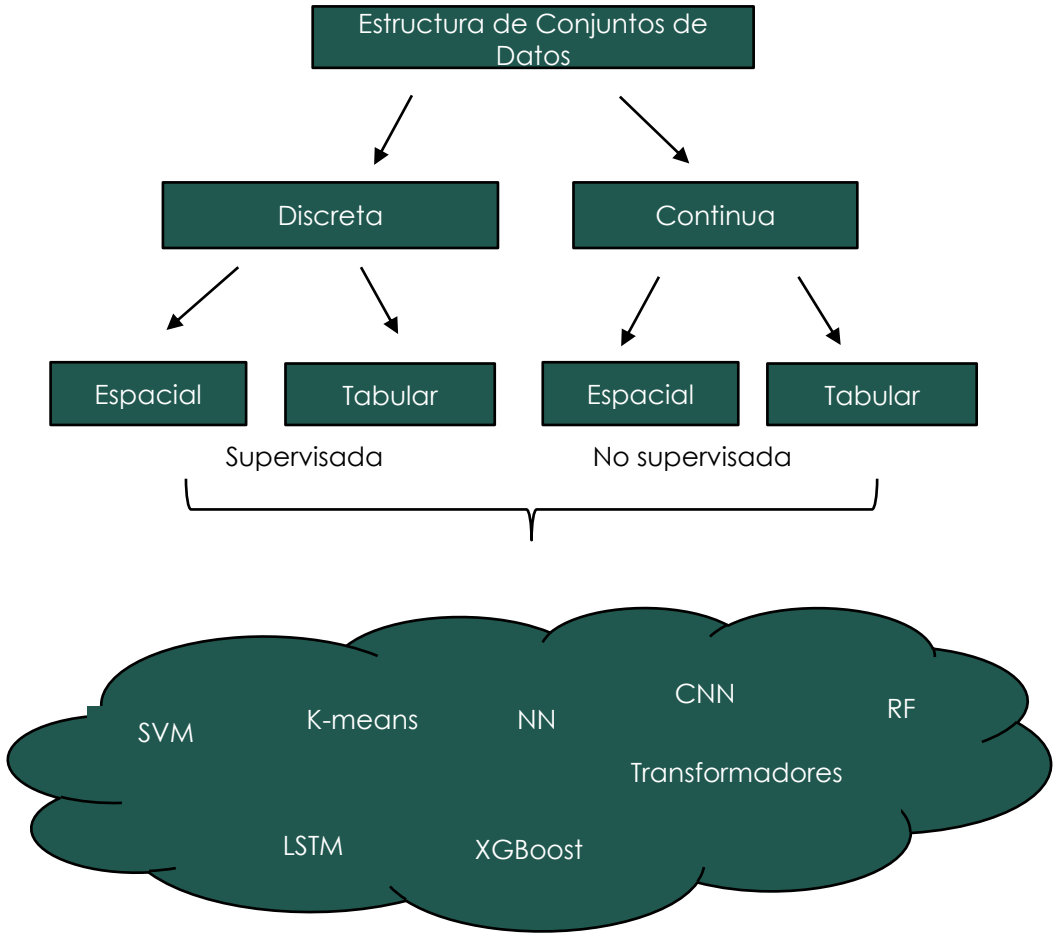
- **¿Qué datos tienen disponibles?**
- **¿Tienen datos de entrenamiento disponibles?**
- **¿Cuál es la estructura de datos de sus datos?**
- **¿Su variable dependiente es un problema continuo o discreto?**



Rama de decisiones del algoritmo basada en la estructura de los datos.

Algoritmos de Aprendizaje Automático: Datos

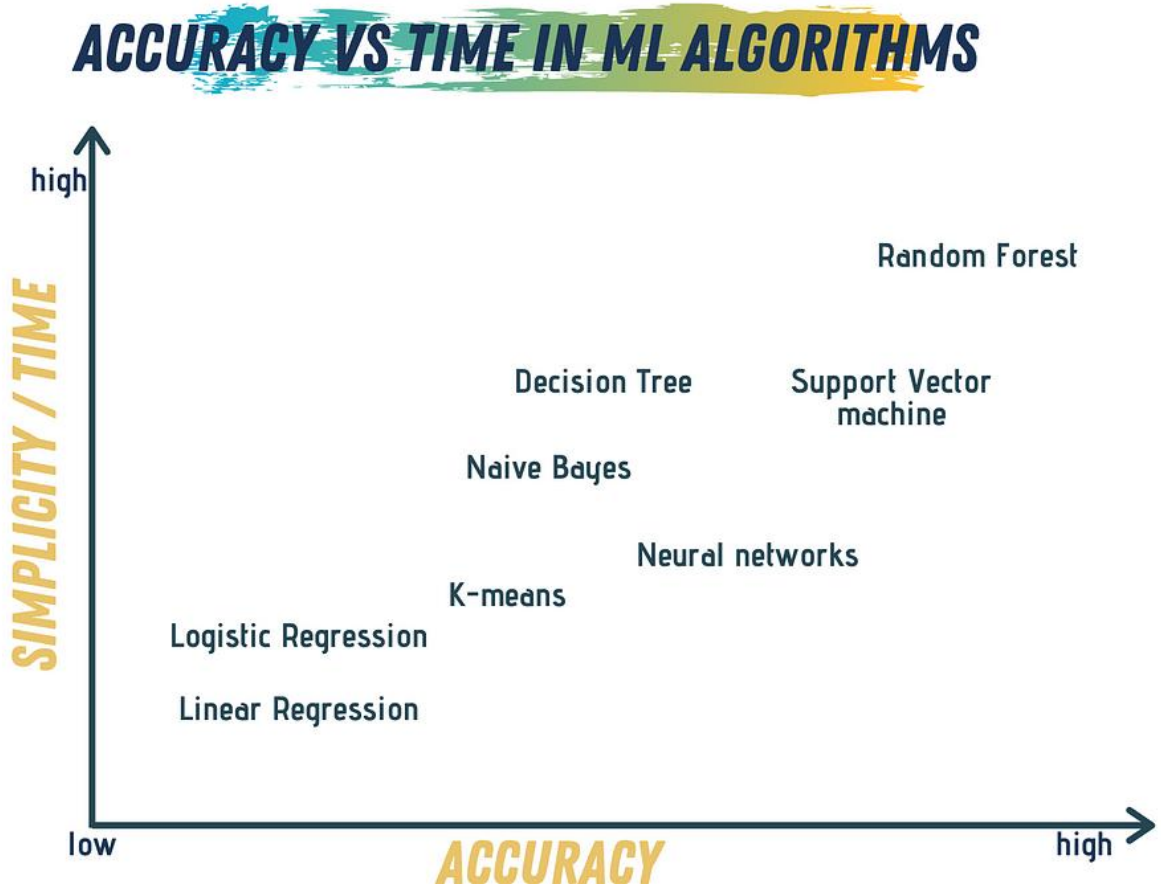
- **¿Qué datos tienen disponibles?** Tenemos cobertura global con datos del satélite MODIS.
- **¿Tienen datos de entrenamiento disponibles?** Hemos recopilados grandes extensiones de puntos de datos de entrenamiento.
- **¿Cuál es la estructura de datos de sus datos?** Nuestros datos están en formato ráster. Los podemos procesar para que sean tabulares.
- **¿Su variable dependiente es un problema continuo o discreto?** Nuestra variable dependiente son los píxeles de agua, la que es discreta (0 – no hay agua, 1 – agua)



Rama de decisiones del algoritmo basada en la estructura de los datos.

Algoritmos de Aprendizaje Automático: Rendimiento

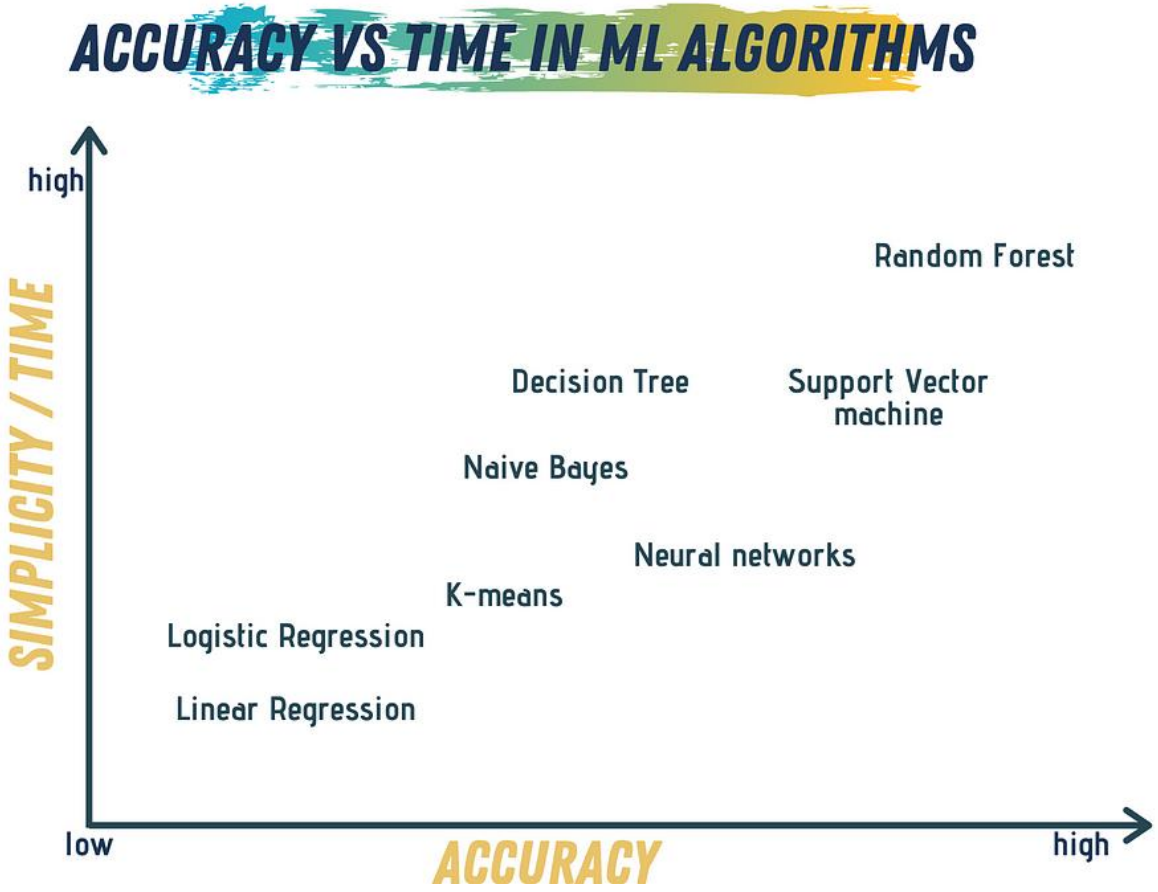
- **¿Hay algún requisito de rendimiento basado en su pregunta científica? (ej., tiempo real vs estático)**
- **¿El hardware en el que su software se va a ejecutar se encuentra en sus instalaciones, en la nube, o está integrado?**
- **¿Qué es más importante para su proyecto: el tiempo de inferencia o el rendimiento del modelo?**



Compromiso entre velocidad y precisión. Fuente de la Imagen: github.com

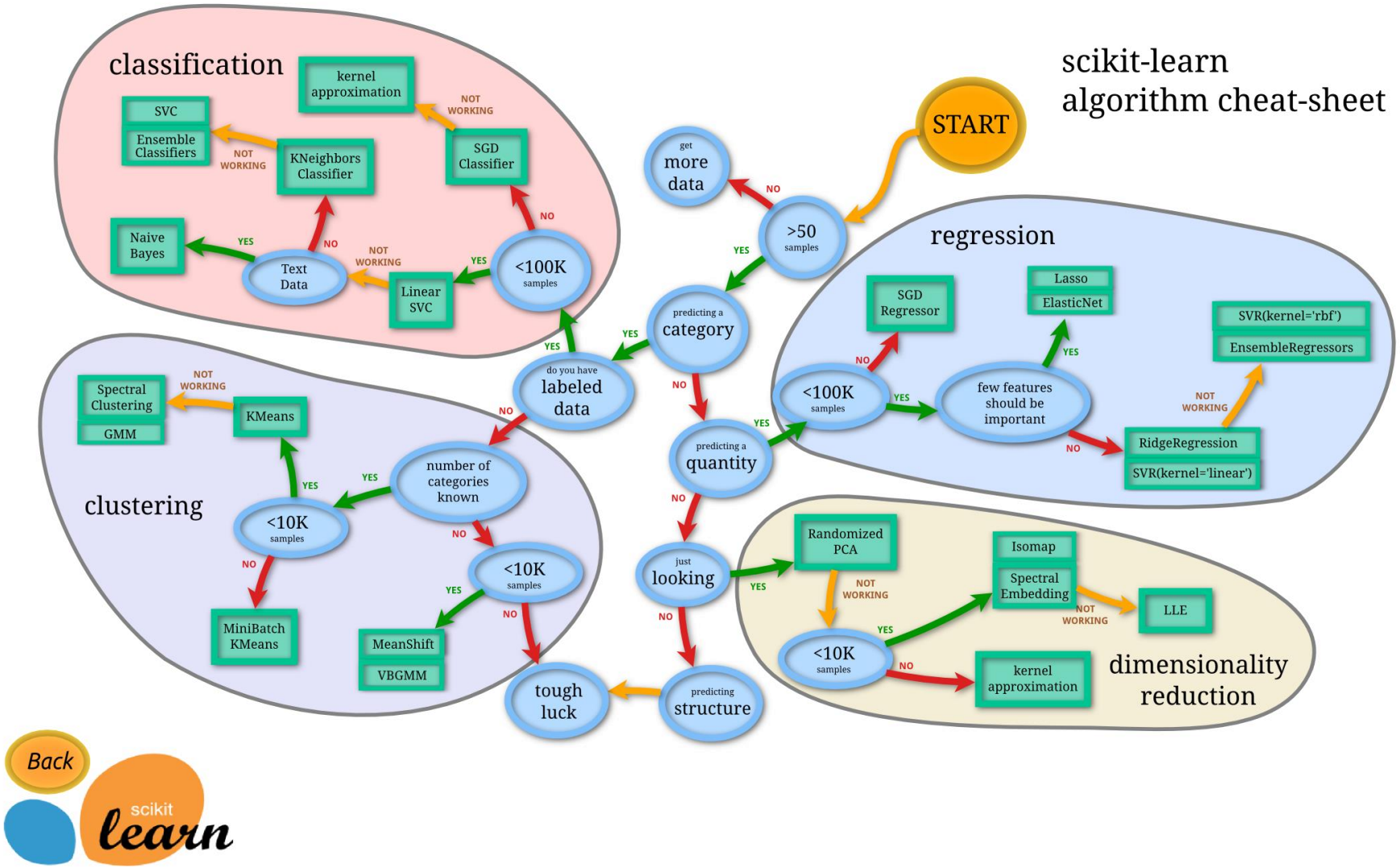
Algoritmos de Aprendizaje Automático: Rendimiento

- **¿Hay algún requisito de rendimiento basado en su pregunta científica? (ej., tiempo real vs estático)** No necesitamos mapas en tiempo real (ej., puede que los equipos de respuesta a desastres necesiten resultados rápidamente).
- **¿El hardware en el que su software se va a ejecutar se encuentra en sus instalaciones, en la nube, o está integrado?** Queremos que nuestro software se ejecute tanto en nuestras instalaciones como en la nube.
- **¿Qué es más importante para su proyecto: el tiempo de inferencia o el rendimiento del modelo?** Nos importa más el rendimiento del modelo que el tiempo de inferencia.

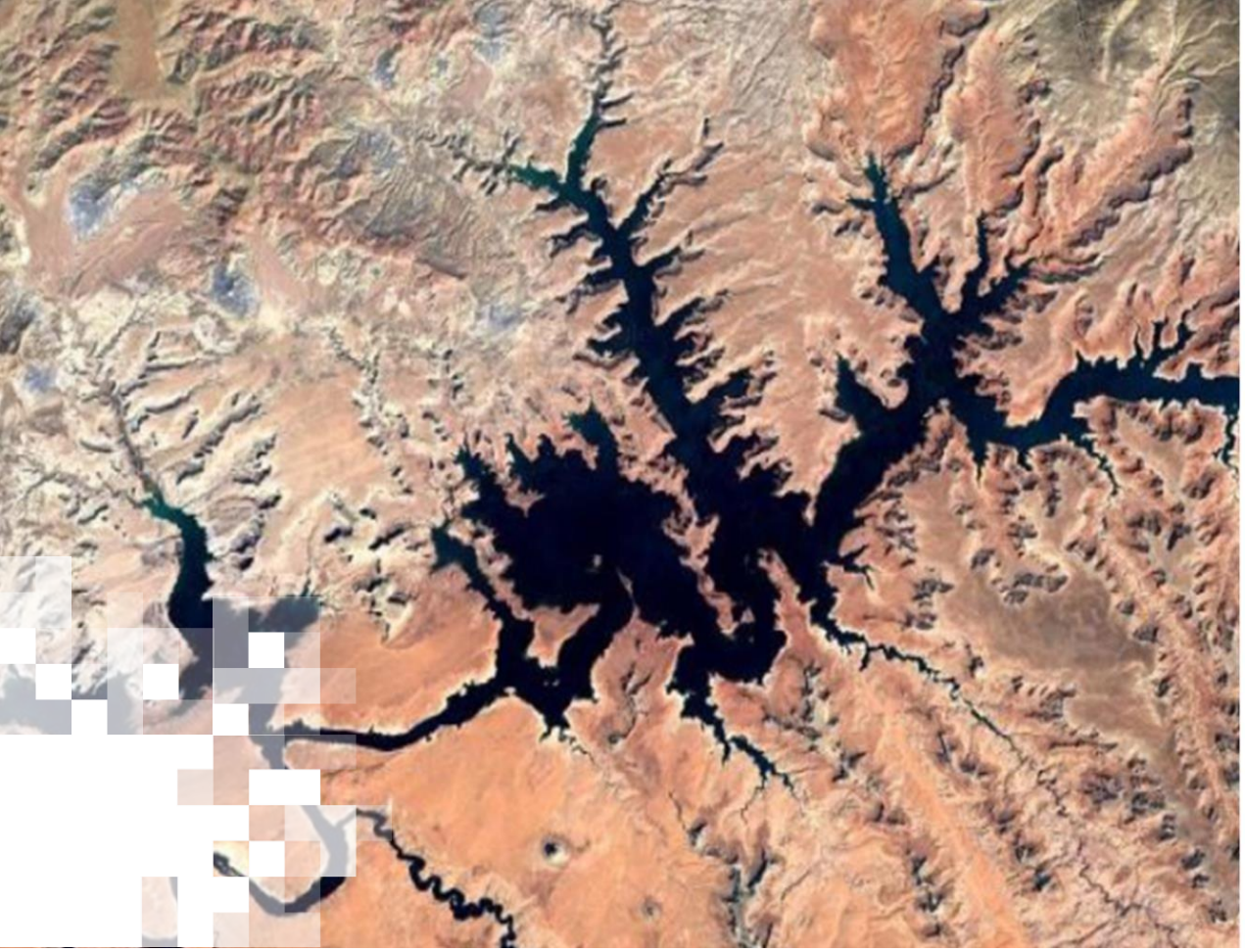


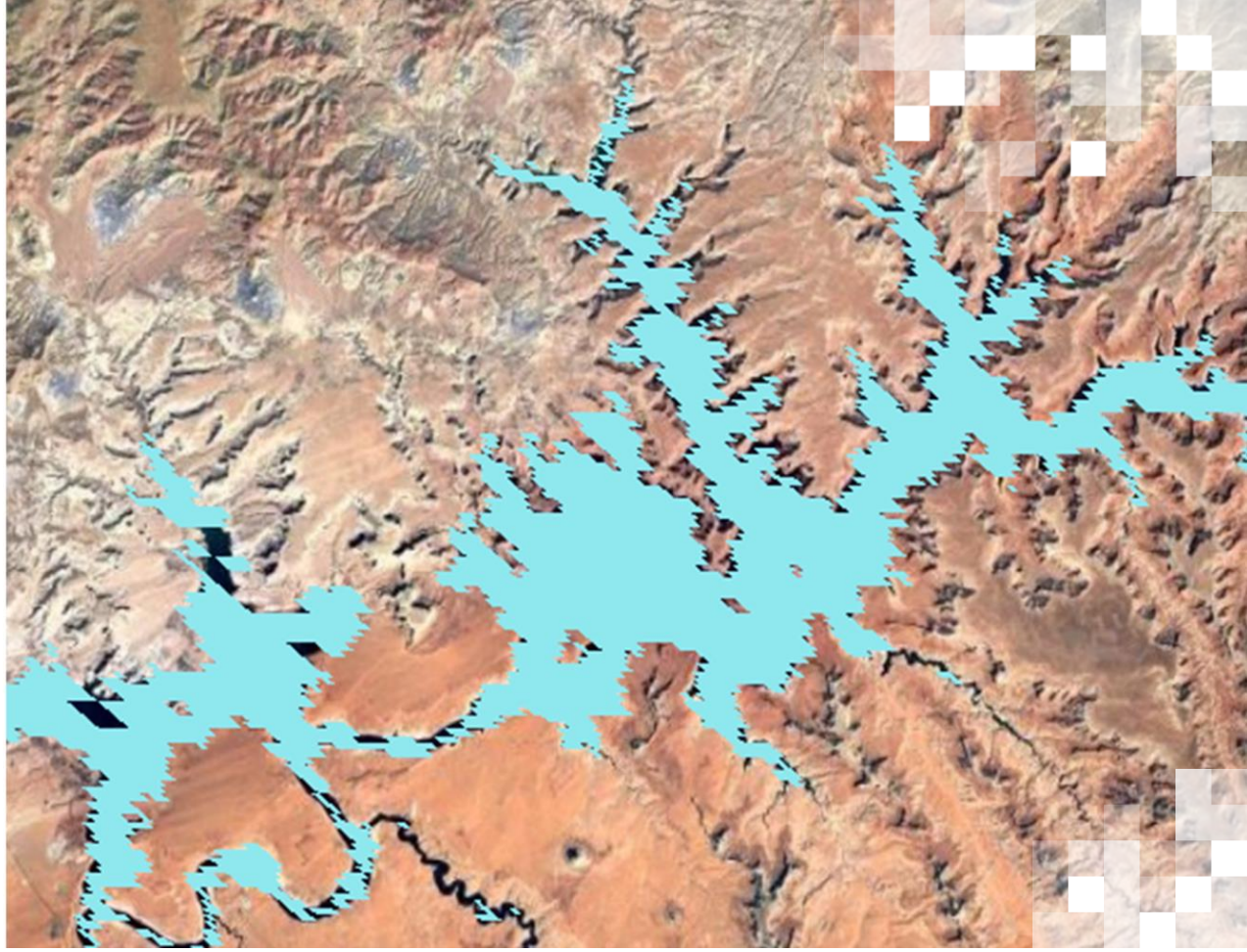
Compromiso entre velocidad y precisión. Fuente de la Imagen: github.com

Algoritmos de Aprendizaje Automático: Operaciones



Flujo de trabajo de escenarios posibles al seleccionar un algoritmo de ML.
Fuente de la Imagen: sckit-learn.org





Ejercicio: Ejecutar cuadernos introductorios en Google Colab

Formador: Jordan A. Caraballo-Vega

Resumen

- Resumen General del Aprendizaje Automático
- Importancia del Aprendizaje Automático Enfocado Hacia las Ciencias de la Tierra
- Utilidad del Aprendizaje Automático
- Software para Apoyar el Aprendizaje Automático
- Aplicaciones de Aprendizaje Automático
- Ejercicio Práctico en Jupyter Notebook: Cargar y Visualizar Datos

Mirando Hacia Adelante

Parte 2: Ejemplo de Datos de Entrenamiento y Clasificación de la Cobertura Terrestre

- Descargar los datos de entrenamiento
- Análisis de datos exploratorios
- Extracción de datos de entrenamiento de un conjunto de datos tabulares
- Extracción de datos de entrenamiento de datos ráster
- Entrenamiento e inferencia de un conjunto de datos tabulares y ráster
- Métricas y evaluación de modelos
- Ejercicio Práctico en Jupyter Notebook: Estudio de Caso de la Clasificación del Agua de MODIS

Contactos

- Instructores:
  - 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

- Página Web de la Capacitación:
  - https://appliedsciences.nasa.gov/join-mission/training/spanish/arset-fundamentos-del-aprendizaje-automatico-para-las-ciencias-de-la

- Página Web de ARSET:
  - https://appliedsciences.nasa.gov/arset

Check out our sister programs:

¿Preguntas?

- Por favor escriban sus preguntas en la ventana de “Questions”. Las responderemos en el orden que las recibimos.
- Publicaremos las preguntas y respuestas en la página de esta capacitación después de la conclusión esta sesión.


APPLIED REMOTE SENSING TRAINING PROGRAM
TRAIN - EMPOWER - ADVANCE

¡Gracias!