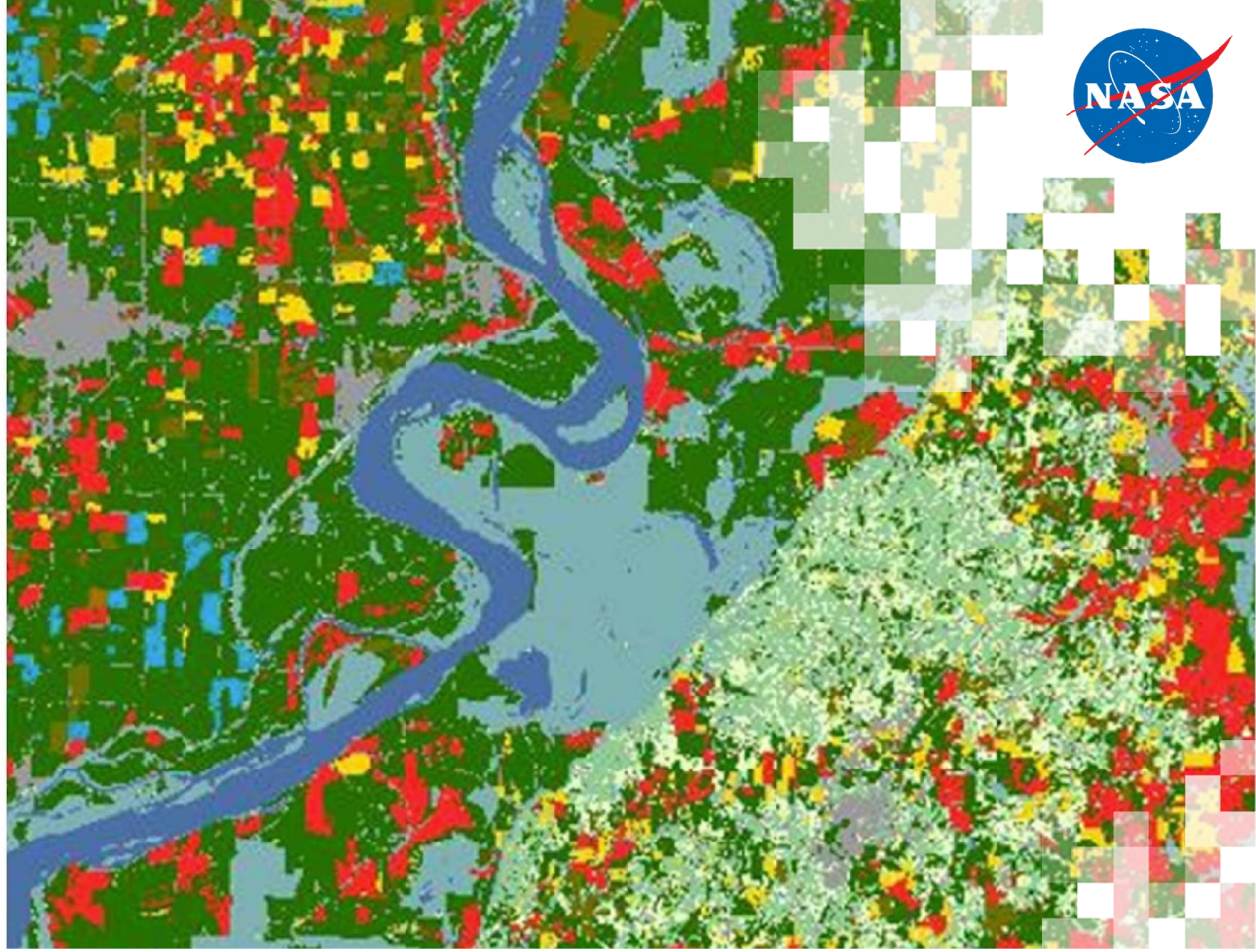
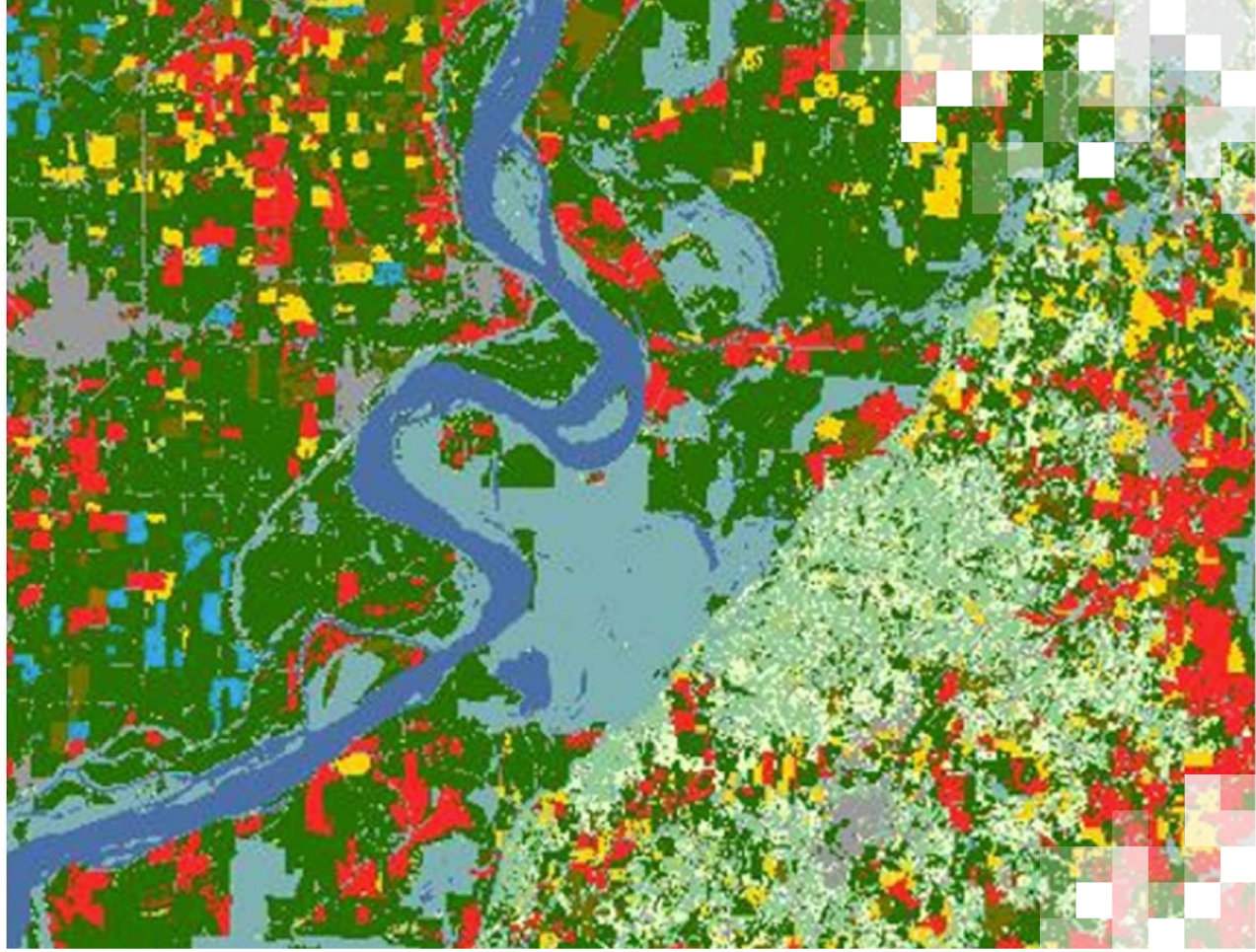



# Aplicaciones de Aprendizaje Automático a Gran Escala usando Teledetección para la Formulación de Soluciones Agrícolas

## 1ra Parte: Preparación de Datos de Imágenes y Etiquetas para la Modelación con Aprendizaje Automático a Gran Escala

John Just (Deere y Cia., Universidad Estatal de Iowa), Erik Sorensen (Deere y Cia.)

5 de marzo de 2024



# Acerca de ARSET

# Acerca de ARSET*

- **ARSET ofrece capacitación accesible, relevante, sin costo sobre satélites, sensores, métodos y herramientas de teledetección.**

- Las capacitaciones incluyen una variedad de aplicaciones de datos de satélite y se personalizan para audiencias con diferentes niveles de experiencia.

AGRICULTURA

CLIMA Y RESILIENCIA

DESASTRES

CONSERVACIÓN ECOLÓGICA

SALUD Y CALIDAD DEL AIRE

RECURSOS HÍDRICOS

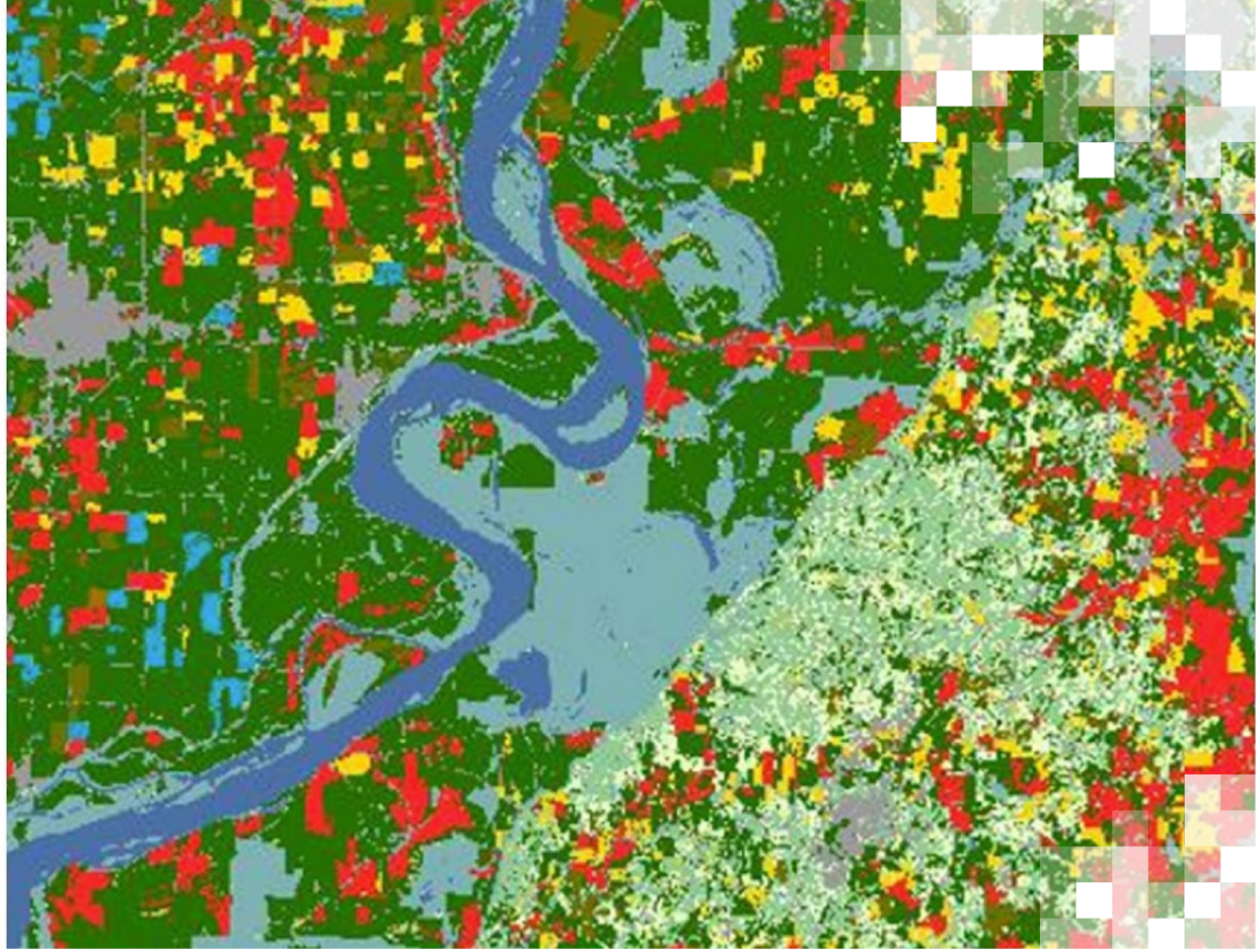
*Siglas de **A**pplied **R**emote **SE**nsing **T**raining Program (Programa de Capacitación de Teledetección Aplicada en inglés)

EARTH SCIENCE APPLIED SCIENCES CAPACITY BUILDING

# Acerca de las Capacitaciones de ARSET

- En línea o presenciales
- En vivo, dirigidas por instructores o autodirigidas por uno a su propio ritmo
- Sin ningún costo
- Opciones bilingües y multilingües
- Solo usan software y datos de fuente abierta
- Acomodan diferentes niveles de experiencia

- Visite la página de ARSET para aprender más.

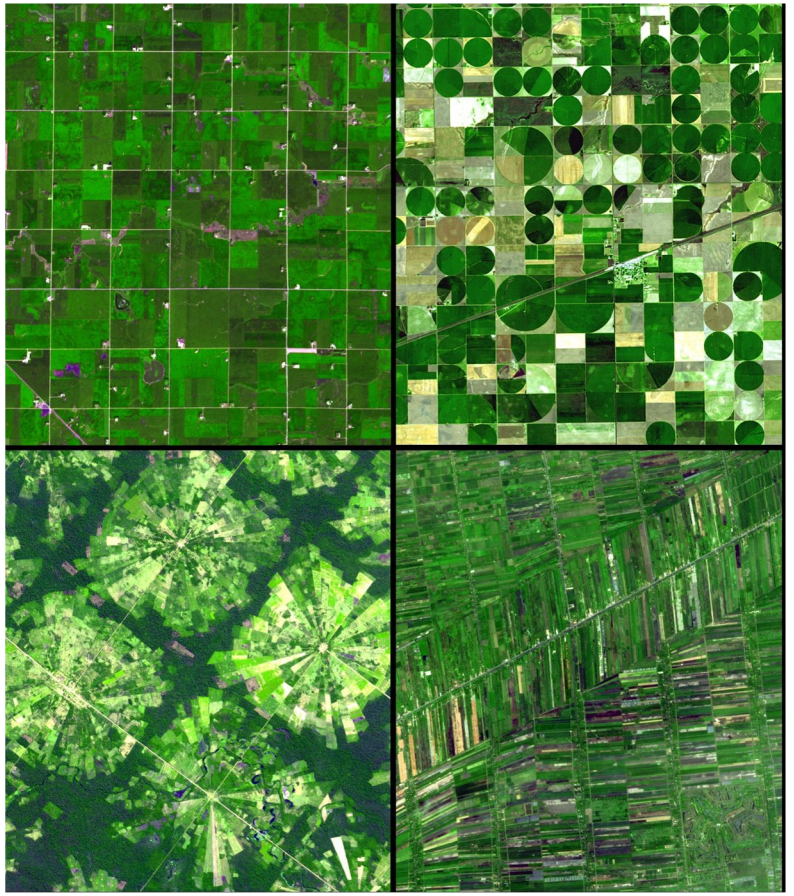


# Aplicaciones de Aprendizaje Automático a Gran Escala usando Teledetección para la Formulación de Soluciones Agrícolas

## **Resumen General**

# Motivación para Esta Capacitación

- Los mapas oportunos y precisos de los cultivos de la temporada, a escala local y regional, son cruciales para la toma de decisiones y la gestión agrícola.
- Las series temporales espaciadas irregularmente son comunes en las imágenes satelitales ópticas.
- El entrenamiento robusto de modelos con datos de teledetección a menudo requiere datos muy grandes, pero el procesamiento y el entrenamiento son complejos.
- La capa Cropland Data Layer* (CDL, USDA–NASS) solo proporciona estimaciones de los tipos de cultivos que se dan a conocer al público unos meses después del final de la temporada de crecimiento, y no su secuencia o cronología (por ejemplo, para cultivos dobles)



Montaje de imágenes mostrando diferencias en la geometría y el tamaño de los campos agrícolas en diferentes partes del mundo. Fuente de la imagen: NASA (Instrumento: Terra – ASTER)

*Capa de datos de tierras de cultivo, en inglés

# Objetivos de Aprendizaje para Esta Capacitación

Al final de esta serie, las/los participantes habrán desarrollado la capacidad para:

- Utilizar las técnicas recomendadas para descargar y procesar datos de teledetección de Sentinel-2 y la capa Cropland Data Layer (CDL) a gran escala (> 5 GB) con herramientas en la nube (Amazon Web Services [AWS] Simple Storage Service [S3], Databricks, Spark/Pyspark, Parquet)
- Producir gráficos interactivos de mapas, tablas, series temporales etc. para la investigación y verificación de datos y modelos.
- Filtrar datos de los dominios medidos (imágenes satelitales) y el objetivo (CDL) para cumplir con los objetivos de modelación basado en factores de calidad, clasificación de tierras, superposición de áreas de interés (AOI, por sus siglas en inglés) y ubicación geográfica.
- Crear canalizaciones de entrenamiento en TensorFlow para entrenar algoritmos de aprendizaje automático en conjuntos de datos geoespaciales/de teledetección a gran escala para el monitoreo agrícola
- Utilizar técnicas de muestreo aleatorio para crear solidez en un algoritmo predictivo y, al mismo tiempo, evitar la fuga de información en las divisiones de entrenamiento/validación/prueba

# Prerrequisitos

- Fundamentos de la Percepción Remota (Teledetección)
- Clasificación de Cultivos con Series Temporales, 2da Parte
- Inscribirse y acceder a Databricks Community Edition

# Esquema de la Capacitación

**1ra Parte**

**Preparación de Datos para Imágenes y Etiquetas para la Modelación con Aprendizaje Automático a Gran Escala**

**5 de marzo de 2024**

**2da Parte**

Cargadores de Datos para Entrenar Modelos de Aprendizaje Automático sobre Series Temporales de Imágenes Espaciadas Irregularmente

12 de marzo de 2024

**3ra Parte**

Entrenamiento y Prueba de Modelos de Aprendizaje Automático para Series Temporales de Imágenes Espaciadas Irregularmente

19 de marzo de 2024

**Tarea**

Abre el 19 de marzo – **Fecha Límite: 1ro de abril** – Publicada en la Página Web de la Capacitación

Se otorgará un certificado de finalización de curso a quienes asistan a todas las sesiones en vivo y completen la tarea asignada antes de la fecha estipulada.

.

# Cómo Hacer Preguntas

- Por favor escriba sus preguntas en la casilla denominada “Questions” y las responderemos al final de este webinar.
- No dude en escribir sus preguntas mientras vayamos avanzando. Intentaremos responder todas las preguntas durante la sesión para preguntas y respuestas después del webinar.
- Las demás preguntas las responderemos en el documento de preguntas y respuestas, el cual será publicado en la página web de la capacitación aproximadamente una semana después de esta.

# 1ra Parte – Formadores

**John Just**

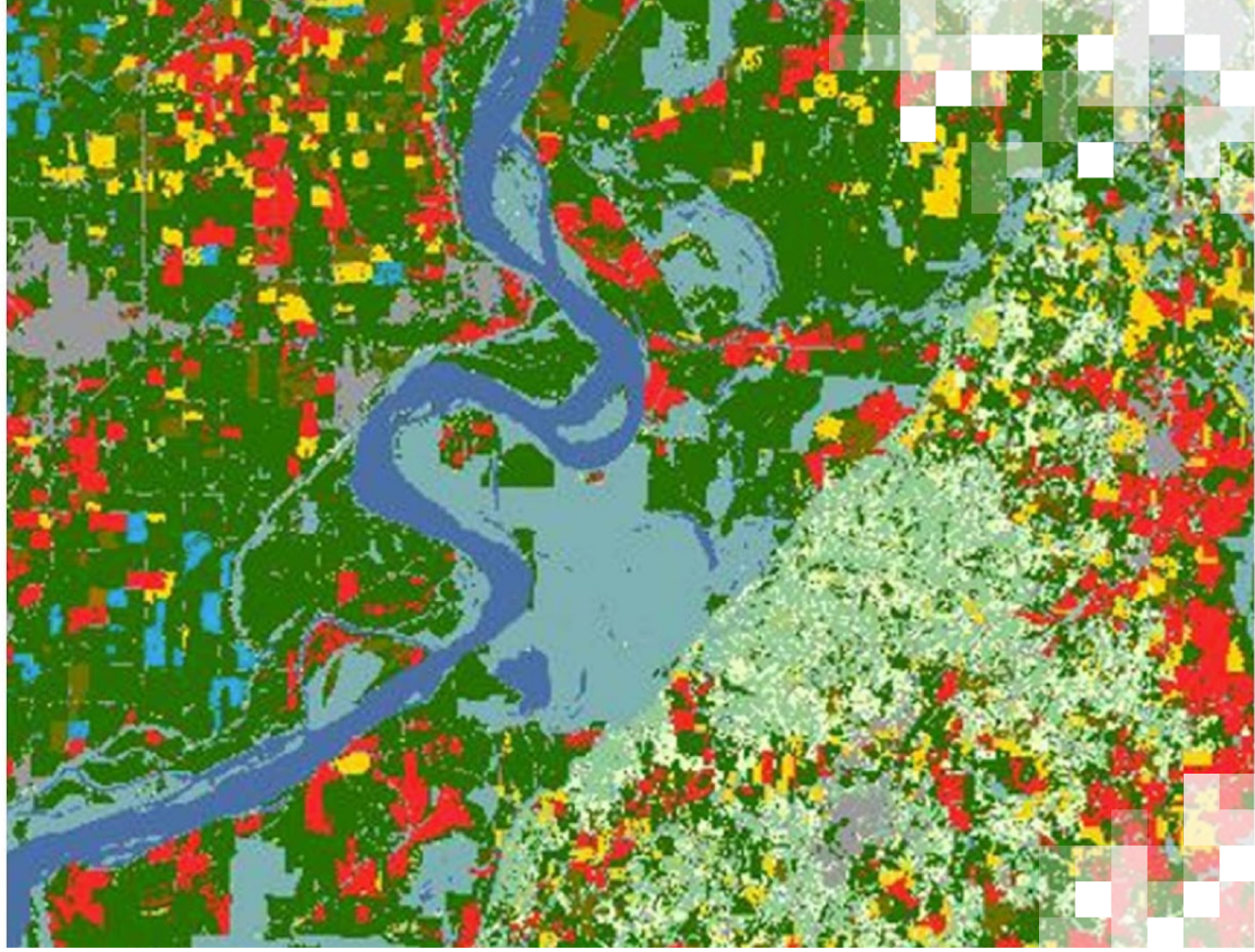
Científico Informático Principal

John Deere

**Erik Sorensen**

Científico Informático Sénior

John Deere



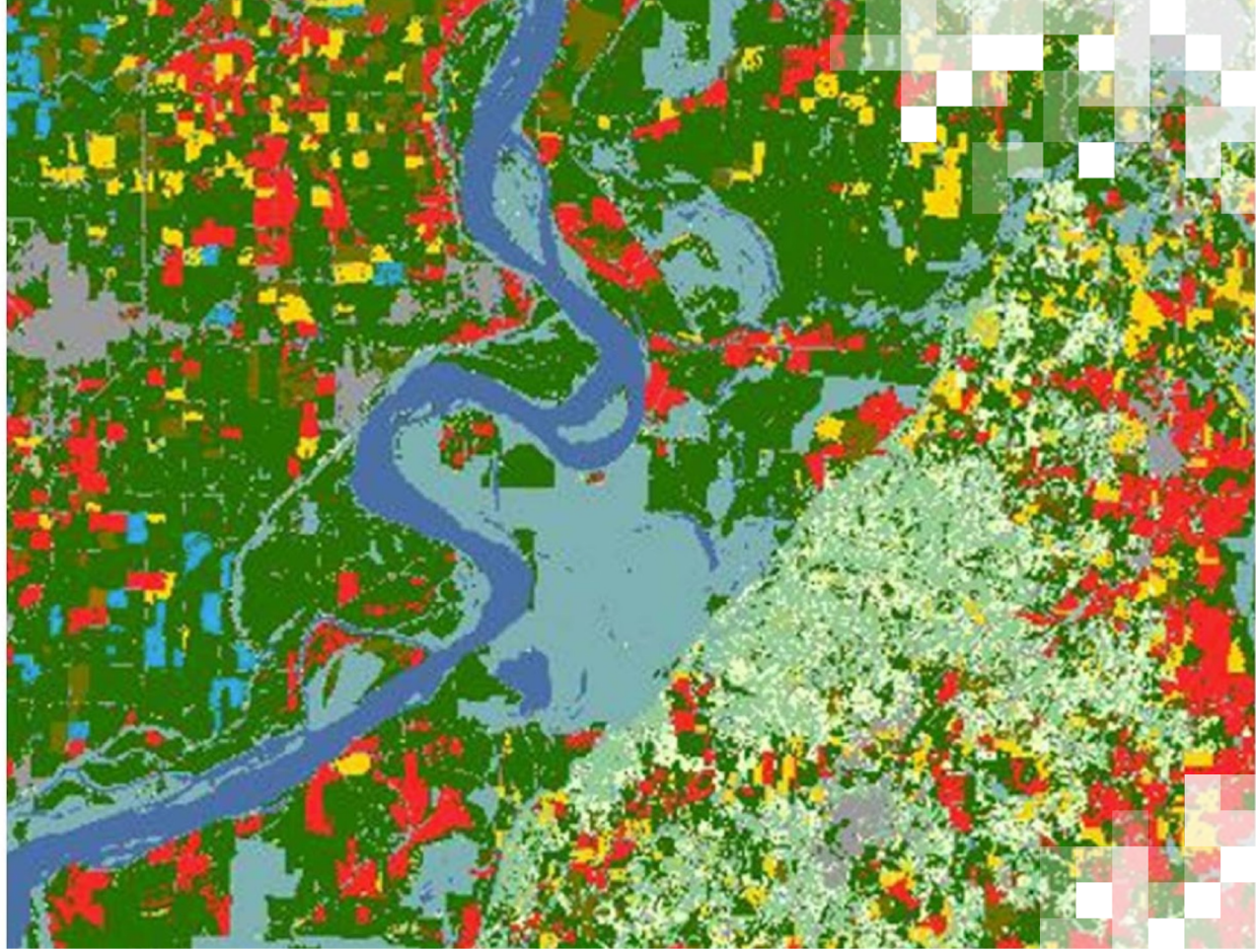
# Aplicaciones de Aprendizaje Automático a Gran Escala usando Teledetección para la Formulación de Soluciones Agrícolas

## 1ra Parte: Preparación de Datos de Imágenes y Etiquetas para la Modelación

# Objetivos – 1ra Parte

Al final de la 1ra Parte, las/los participantes habrán desarrollado la capacidad para:

- Entregar listas de límites al NASS API y extraer los rásteres de la capa CDL de vuelta.
- Sub-muestrear y visualizar datos extraídos de CDL con imágenes espaciales interactivas y otras diagramaciones estadísticas.
- Obtener archivos ráster de Sentinel-2 para un área y un momento determinados correspondiendo con los datos CDL extraídos y manipular los rásteres de Sentinel-2 para crear tablas en preparación para el análisis y entrenamiento de modelos.
- Verificar el procesamiento correcto de datos mediante varias diagramaciones interactivas (p.ej., series temporales de pixeles de varias coberturas terrestres).

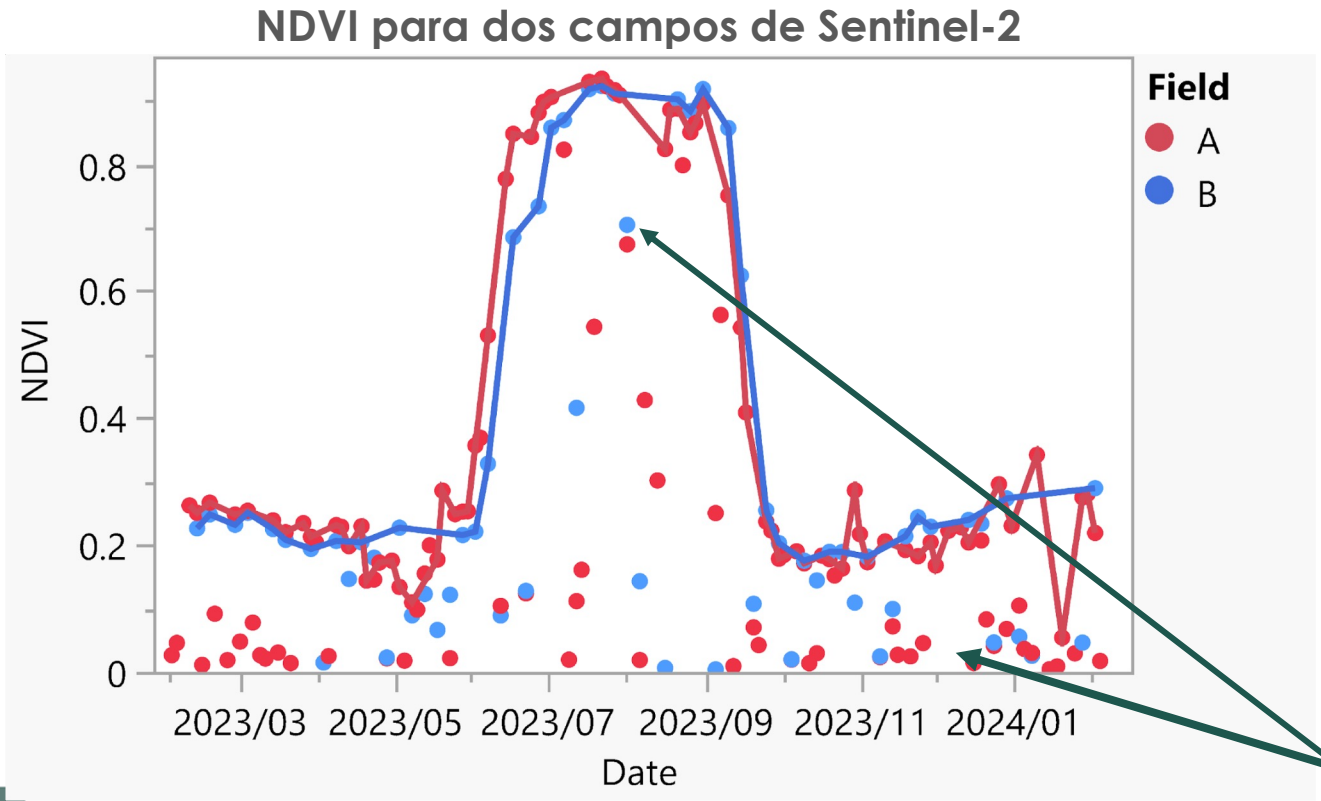


1ra Parte, Sección 1:

# Modelación de Series Temporales Espaciadas Irregularmente

# Modelación de Series Temporales Espaciadas Irregularmente

Es algo común debido a: la geometría orbital, variaciones en la cronología/geolocalización/ extensiones de imágenes de la órbita y disturbios atmosféricos debido a nubes/humo u otros eventos aleatorios.




Cronología/ Espaciado Irregular

| Field | Interval (days) | # Scenes |
|---|---|---|
| A | 2 | 65 |
| A | 3 | 65 |
| A | 5 | 8 |
| B | 5 | 67 |
| B | 10 | 3 |

Los puntos no sobre las líneas son escenas con nubes

A y B están a 2 km el uno del otro
Pero A tiene doble el número de escenas (cobertura) debito al solapamiento de trayectorias orbitales

# Motivación para este Ejemplo

Hemos propuesto una **predicción en tiempo real** para la Capa de datos de tierras de cultivo (Cropland Data Layer o CDL) como el ejemplo operativo para este tutorial.

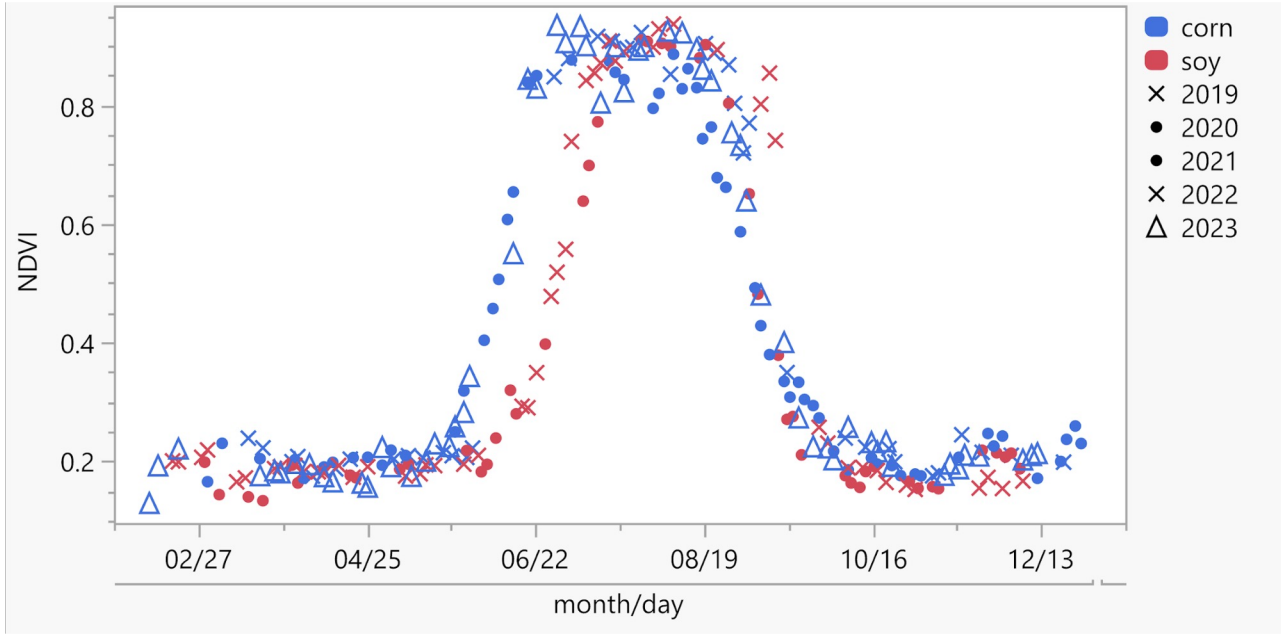
- El algoritmo de la CDL ya utiliza las mismas fuentes de datos satelitales (o fuentes similares) y espaciado/intervalos irregulares para formular predicciones (ejemplo de éxito documentado)
- Las etiquetas están disponibles a través de llamados de API (altamente escalables/disponibles).
- La precisión ha sido muy estudiada y documentada
- ***El código resultante y los métodos son altamente transferibles a otros problemas/ casos de uso***

# Prediciendo la Capa CDL en Tiempo Real

Según las preguntas frecuentes del CDL,

- "El programa CDL utiliza imágenes satelitales de resolución espacial media (30 metros) porque es demasiado costoso usar satélites de mayor resolución para realizar estimaciones de superficie de cultivos en grandes áreas".
- La CDL se considera confidencial y sensible al mercado durante la temporada de crecimiento y no se puede publicar hasta después de que se publiquen las estimaciones oficiales del condado del área de fin de año de la NASS a fines de enero o principios de febrero, después del final de la temporada de crecimiento típica en EE. UU.
- La CDL solo proporciona estimaciones de los tipos de cultivos, pero no su secuencia o cronología (por ejemplo, para cultivos dobles)

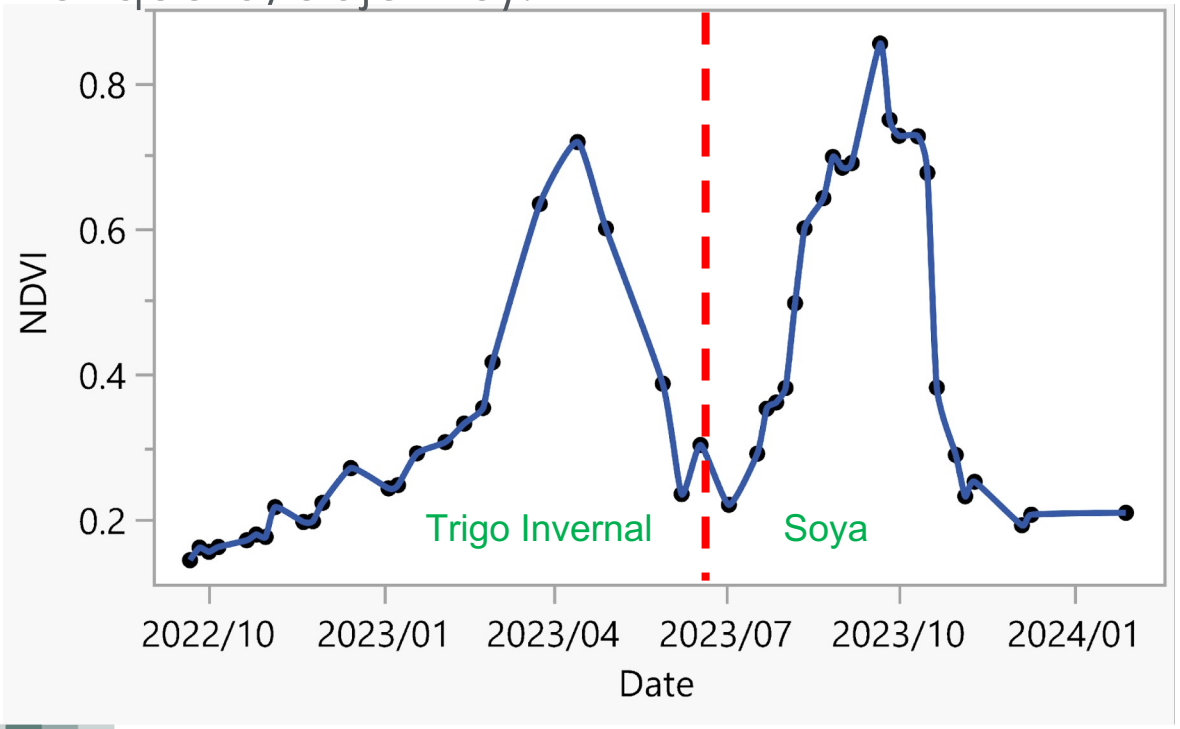
**Sin embargo…¿en verdad hay que esperar hasta el año siguiente para tener estimaciones precisas?**




**"A principios de julio, probablemente estemos bastante seguros del tipo de cultivo solo basado en el NDVI en esta área (probablemente aún más mediante el uso de series temporales multiespectrales). A finales de agosto estamos realmente seguros (6 meses antes de que la publicación de la CDL)."**

# Metodología Generalizable

- Los modelos robustos requieren herramientas y metodologías de gestión de datos a gran escala. Aprovechar la computación paralela multinúcleo y multimáquina es un paso necesario para escalar. Demostramos estas herramientas y metodologías con esta serie.
- Tenga en cuenta que también se podría utilizar una metodología similar a la modelación de cultivos con las series temporales de imágenes para estimar la salud de los cultivos u otros factores que dependen del tiempo (simplemente sustituya la etiqueta/objetivo).

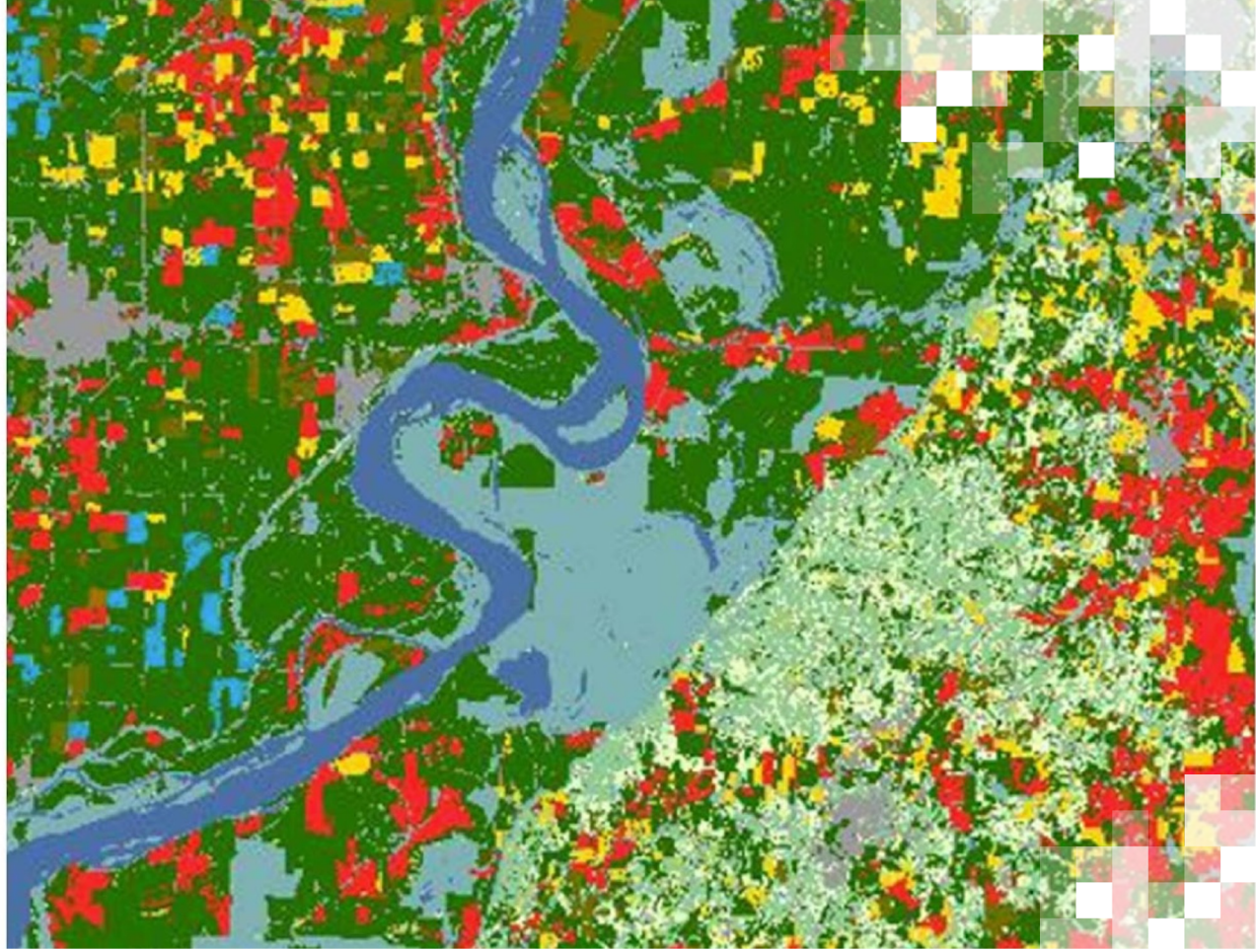



**A principios de la primavera, estamos bastante seguros del tipo de cultivo para las áreas que plantaron cultivos invernales y probablemente tendremos una buena estimación de áreas de doble cultivo a principios del otoño.**

# Modelación de Series Temporales Espaciadas Irregularmente

- Hay una escasez de teoría estadística en cuanto a las series temporales espaciadas de manera desigual y, por lo tanto, los métodos listos para utilizar no ofrecen mucho que se pueda aplicar *directamente* a semejante situaciones.
  - La solución más común es manipular los datos en una serie temporal espaciada regularmente y, a continuación, aplicar métodos estándar. Por ejemplo, interpolación o agrupación de intervalos (*este último es lo que hace el algoritmo CDL*).
  - **Note que el formato de datos resultante de esta demostración de la 1ra parte admitirá cualquier metodología de modelación**.
    - Seguimos una metodología similar al de CDL (intervalos agrupados) para las partes 2 y 3 de esta demostración para cargadores de datos y entrenamiento de modelos debido a la simplicidad.
- Los modelos de secuencia de ML* más nuevos, como los transformadores ("autoatención"), aceptan codificaciones posicionales de insumo/salida y aprenden información de insumo (y salida) absoluta y relativa significativa. Esto podría facilitar la modelación directa de datos satelitales espaciados de manera desigual, pero debido a la mayor complejidad no lo incorporamos aquí.

*ML- siglas de "Machine Learning", aprendizaje automático en inglés



# 1ra Parte Sección 1:
# **La Capa “Cropland Data Layer” (CDL)**

# La Capa “Cropland Data Layer” (Estados Unidos Contiguos)

El mejor lugar para encontrar información sobre esta capa es en las Preguntas Frecuentes del USDA NASS y los metadatos

Algunos datos relevantes:

- **Modelo**: Clasificador con árboles de decisión (maneja datos no contiguos, no continuos, no normales, no lineales, computación eficiente). Salidas probabilísticas (argmax a clase)
- **Insumos**: Landsat 8 y 9 OLI/TIRS, ISRO ResourceSat-2 LISS-3 y ESA SENTINEL-2A y -2B. Las imágenes se descargan a diario con el objetivo de obtener por lo menos una imagen útil libre de nubes cada dos semanas a lo largo de la temporada de crecimiento
- **Verdad en el suelo**: FSA Common Land Unit* (CLU) para cultivos/zonas agrícolas y National Land Cover Database (NLCD) para zonas no agrícolas
- **Precisión**: Generalmente entre el 85% y el 95% para las categorías de cobertura terrestre para cultivos específicos. Resolución de 30m

*Unidad de tierra común

# CDL- Precisión

Como ya mencionamos, la precisión de la capa CDL ha sido muy estudiada y documentada. A continuación hay un extracto de los controles de calidad del Dpto. de Agricultura de EE.UU. (USDA) de 2022 para el estado de Arkansas.

- Hay información de todos los años para todos los estados en el USDA NASS Cropland metadata.
- Puede que algunos cultivos como el sorgo (en este caso) tengan una precisión baja, pero también representan una porción mínima de las tierras agrícolas. Entrenar un modelo enfocado específicamente en ciertas clases podría aumentar la precisión para esas clases perjudicando las demás.

USDA National Agricultural Statistics Service, 2022
Arkansas Cropland Data Layer

STATEWIDE AGRICULTURAL ACCURACY REPORT

| Crop-specific covers only | *Correct | Accuracy | Error | Kappa |
|---|---|---|---|---|
| OVERALL ACCURACY** | 482475 | 87.30% | 12.70% | 0.817 |

| Cover Type | *Correct Pixels | Producer's Accuracy | Omission Error | Kappa | User's Accuracy | Commission Error | Cond'l Kappa |
|---|---|---|---|---|---|---|---|
| Corn | 55159 | 86.40% | 13.60% | 0.855 | 94.10% | 5.90% | 0.937 |
| Cotton | 50682 | 88.00% | 12.00% | 0.873 | 93.20% | 6.80% | 0.928 |
| Rice | 87048 | 90.30% | 9.70% | 0.893 | 96.10% | 3.90% | 0.957 |
| Sorghum | 156 | 22.70% | 77.30% | 0.227 | 77.20% | 22.80% | 0.772 |
| Soybeans | 254301 | 93.60% | 6.40% | 0.91 | 89.60% | 10.40% | 0.857 |

*Correct Pixels represents the total number of independent validation pixels correctly identified in the error matrix.
**The Overall Accuracy represents only the FSA row crops and annual fruit and vegetables

# Cálculos Aproximados del Tamaño de Datos para la Capa CDL sobre EE.UU. Contiguos

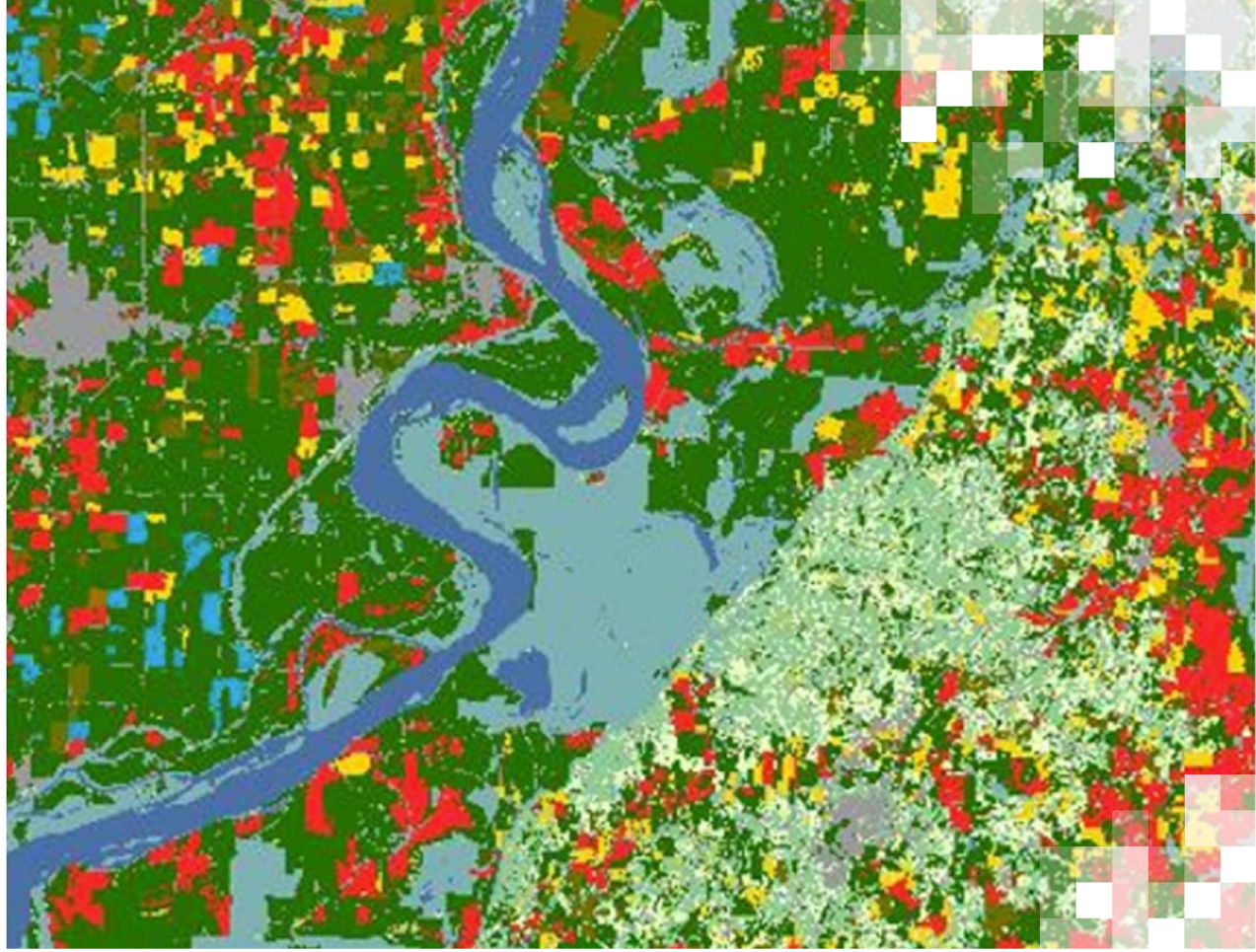
A pesar de que aproximadamente solo el 20% de EE.UU. se utilice específicamente para cultivos, cualquier modelo debe recorrer todo EE.UU. para clasificar las tierras.

Si presumimos:

- El uso de Sentinel-2 para la clasificación de tierras con una resolución de 10m
- Una imagen libre de nubes cada dos semanas (26 imágenes en total por cada 100$m^2$)
- 12 bandas a 2 bytes (16 bits) por píxel (26 imágenes en total por cada 100$m^2$)
- ~7.500.000 $km^2$ de suelo

≈44 TB de datos fueron posprocesados para ejecutar un modelo predictivo. Aunque no es una cantidad trivial, las drives de 10TB solo cuestan USD $200. Usar la resolución de 30$m^2$ reduce esto a ~5TB para poder trabajar.

Si solo usamos Sentinel-2, hay que descargar y procesar aproximadamente 28TB de datos rástер sobre esas tierras porque las teselas de Sentinel-2 tiles son de110 x 110 $km^2$ y tienen 12 bandas, lo que viene a ser aproximadamente 640MB por escena, que ocurre cada 5 días.



# 1ra Parte, Sección 2:
# **Datos Ópticos de Sentinel-2**

# Factores Afectando la Calidad y el Espaciado Temporal de los Datos Satelitales

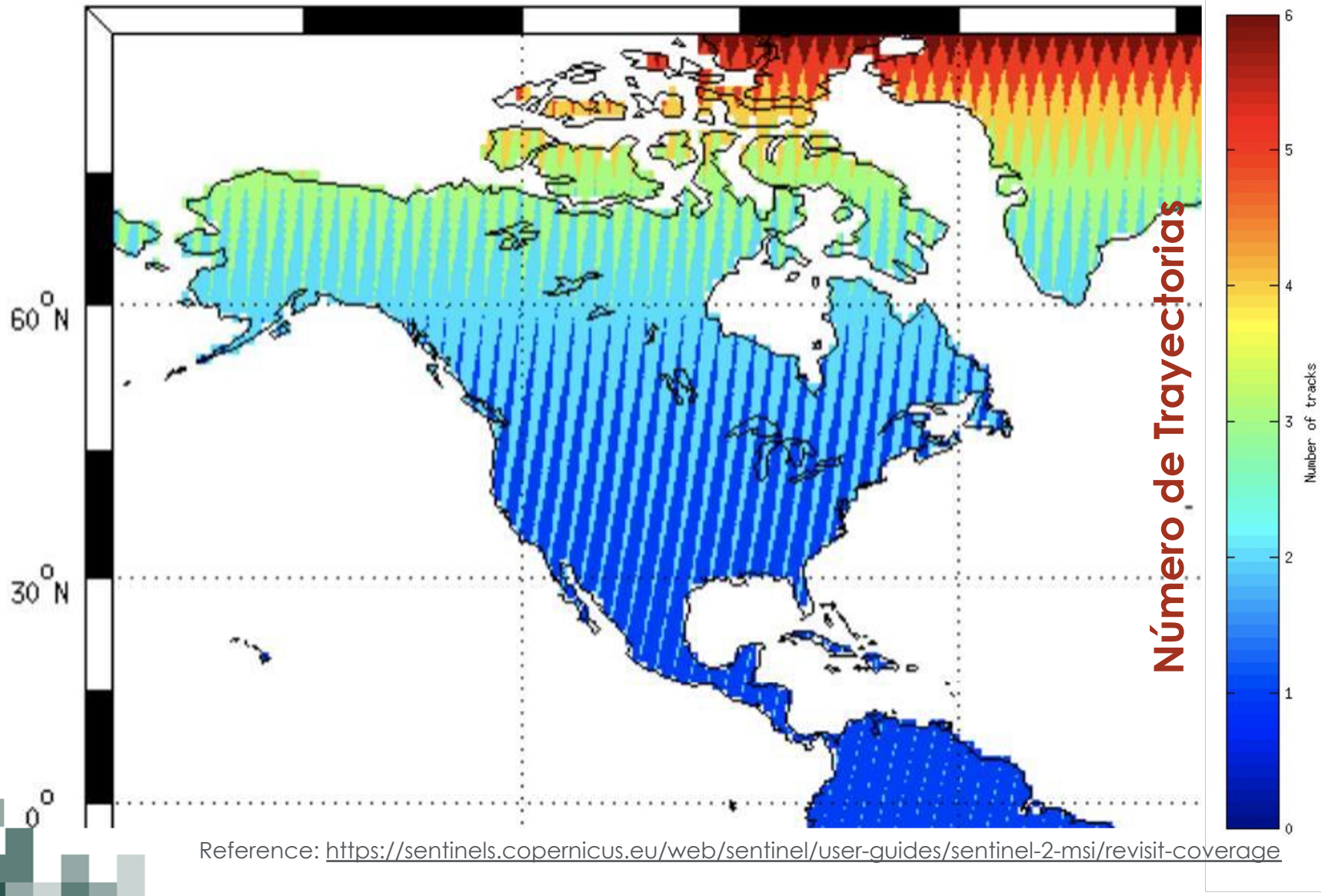
Ejemplos de factores que afectan a los datos:

- Irregularidad
  - **Solapamiento de trayectorias orbitales** – sucede más cerca de los polos y resulta en una mayor cobertura
  - **Repetibilidad de captura de imágenes/trayectoria orbital** – la posición exacta de la imagen puede variar respecto al este/oeste y causar que las zonas al borde de las escenas tengan más incertidumbre en su cobertura.
  - **Cobertura nubosa espesa** – cuando está presente e identificada (y así ignorada) hay una laguna en la cobertura.
  - **Errores de SCL*** – cuando ignoramos los datos de escenas debido a categoría SCL, pero está mal, introducimos lagunas innecesarias en la cobertura.

- Calidad
  - **Bruma de nubes ralas** – la cobertura nubosa no es booleana, es un gradiente. A veces cuesta identificarla cuando es leve (altera los valores de reflectancia y puede que no se documente)
  - **Solapamiento de sistema de teselado** – al borde de las teselas (que son cómo los datos se almacenan y buscan) hay solapamiento y pequeñas diferencias entre los valores para la misma escena y ubicación en diferentes teselas.
  - **Geolocalización/georreferenciación** – puede que la ubicación de los píxeles esté incorrecta y varíe por más del tamaño de un píxel (dando como resultado información equivocada para una ubicación puntual)

*SCL- Siglas de Scene Classification Layer , Capa de Clasificación de Escena en inglés.

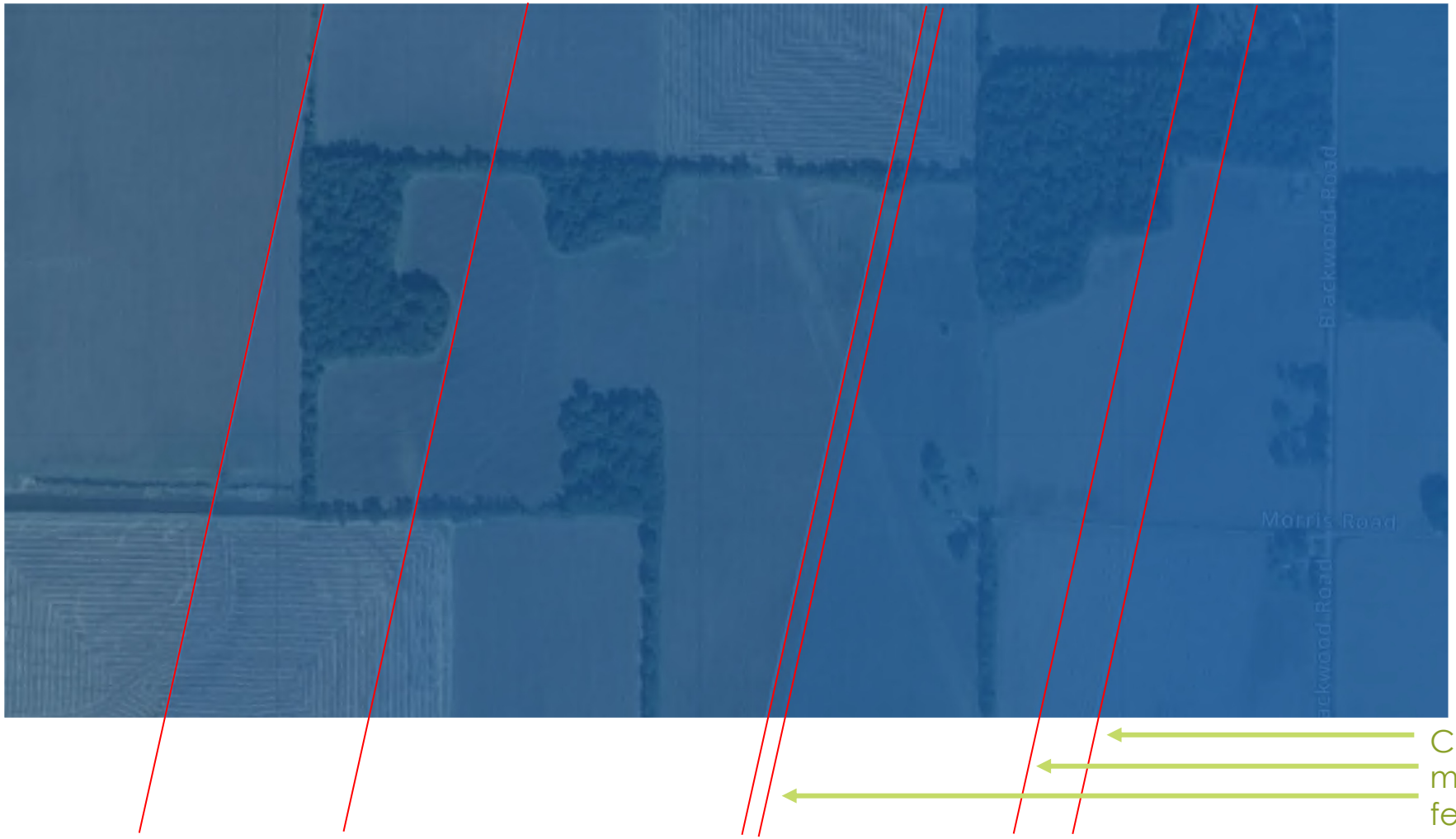
# Sentinel-2 – Solapamiento de Trayectorias Orbitales




Reference: https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/revisit-coverage

La disponibilidad nominal de Sentinel-2 es de 1 imagen cada 5 días. Sin embargo, la superposición entre órbitas adyacentes aumenta más allá del ecuador. Por lo tanto, ciertas partes de EE. UU. en los bordes de las órbitas obtienen una cobertura de hasta 2 veces por satélite y da como resultado intervalos de 2 o 3 días en lugar de 5.

# Repetibilidad de Trayectoria Orbital



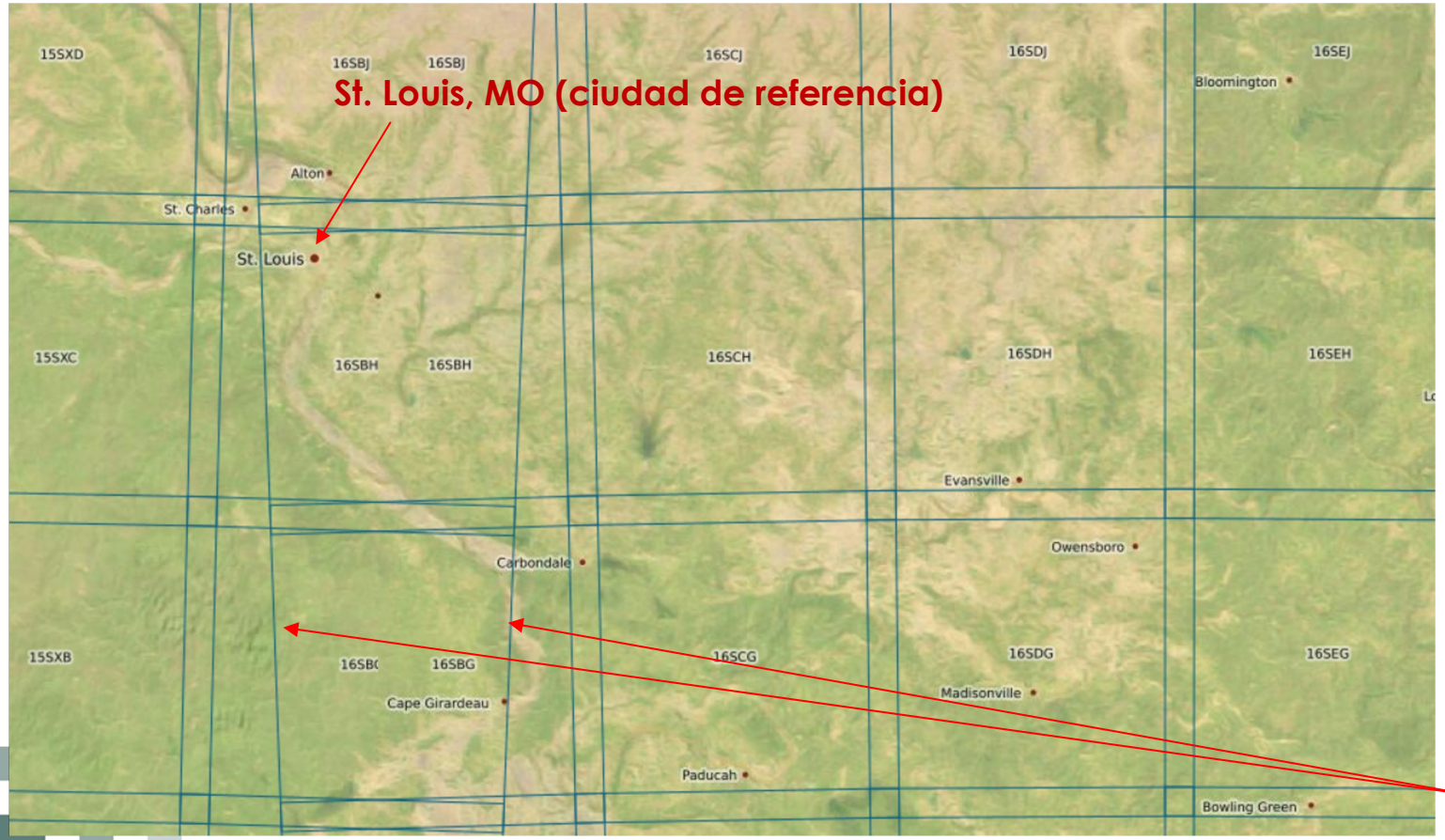

Cada línea diagonal es el límite del mismo recorrido orbital en diferentes fechas

Dentro de una distancia de 1km, la misma trayectoria orbital nominal tiene numerosas variaciones reales de su trayectoria orbital. Otra vez, esto puede afectar la disponibilidad de imágenes.

# Sinopsis del Sistema de Teselación de Sentinel-2

Algunas consideraciones al procesar datos S2 para análisis y modelación:

- Las escenas capturadas de Sentinel-2 se procesan y se ponen a disponibilidad en un sistema de teselación único que es una versión levemente modificada del sistema de referencia de cuadrícula militar (military grid reference system o MGRS).

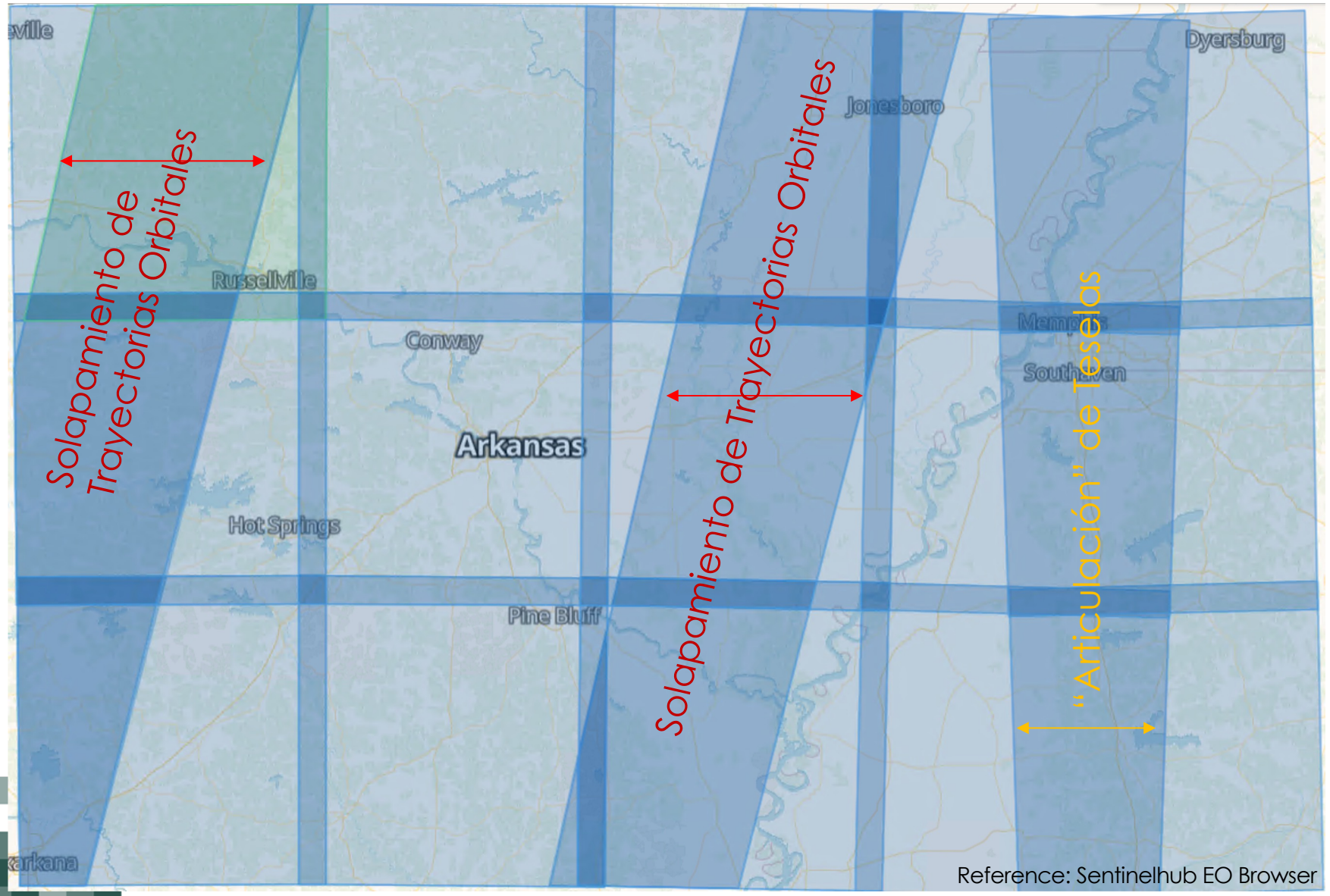



El solapamiento de teselas puede causar que haya valores de la misma escena en hasta cuatro teselas diferentes. También existen "articulaciones" en la cuadrícula para unirlas debido a la superposición de una cuadrícula sobre una esfera.

*Por cualquier motivo, (tal vez por la referenciación o calibración un poco diferente para cada tesela)* ***los valores de banda para la misma escena/ ubicación puede ser un tanto diferente ("falsas" diferencias).***

Referencia: https://maps.eatlas.org.au/

# Trayectorias Orbitales de Sentinel-2 VS Cuadrícula Teselada




Reference: Sentinelhub EO Browser

Las escenas de Sentinel-2 son mucho más grandes que las teselas (ancho de franja de 290 km, frente a 110 km de ancho de una tesela). Por lo tanto, la superposición en los bordes de las trayectorias orbitales puede cubrir una tesela completa en ciertas latitudes.

# Capa de Clasificación de Escena (Scene Classification Layer o SCL)

La banda SCL es útil para la identificación rápida de datos de interés. Si bien no es un clasificador de cobertura terrestre adecuado como el CDL, facilita la clasificación rápida de píxeles por escena en 12 categorías [en su mayoría potencialmente transitorias]. Algunos atributos destacados:

- El caso de uso más común es la identificación de la cobertura de nubes y hay una máscara de nubes separada disponible con probabilidades (usando el algoritmo Sen2Cor). También lo usamos para identificar la vegetación en esta demostración
- Resolución de 60 m, utilizando un solo píxel de una sola escena para la predicción
- Puede ser propenso a errores

Tasas de detección de nubes y cirros y tasas de clasificación errónea de tierras, agua, nieve y sombra como nubes, según se determinó utilizando 108 escenas de Sentinel-2 etiquetadas a mano por Hollstein et al. Referencia

Fraction of classifications as clouds

| True Label | Fmask | Sen2Cor | Sentinel Hub |
|---|---|---|---|
| Cloud | 89.0% | 97.5% | 99.4% |
| Cirrus | 88.3% | 87.7% | 83.8% |
| Land | 7.2% | 5.7% | 2.2% |
| Water | 2.0% | 0.0% | 0.1% |
| Snow | 39.2% | 30.7% | 13.5% |
| Shadow | 3.9% | 3.9% | 5.8% |

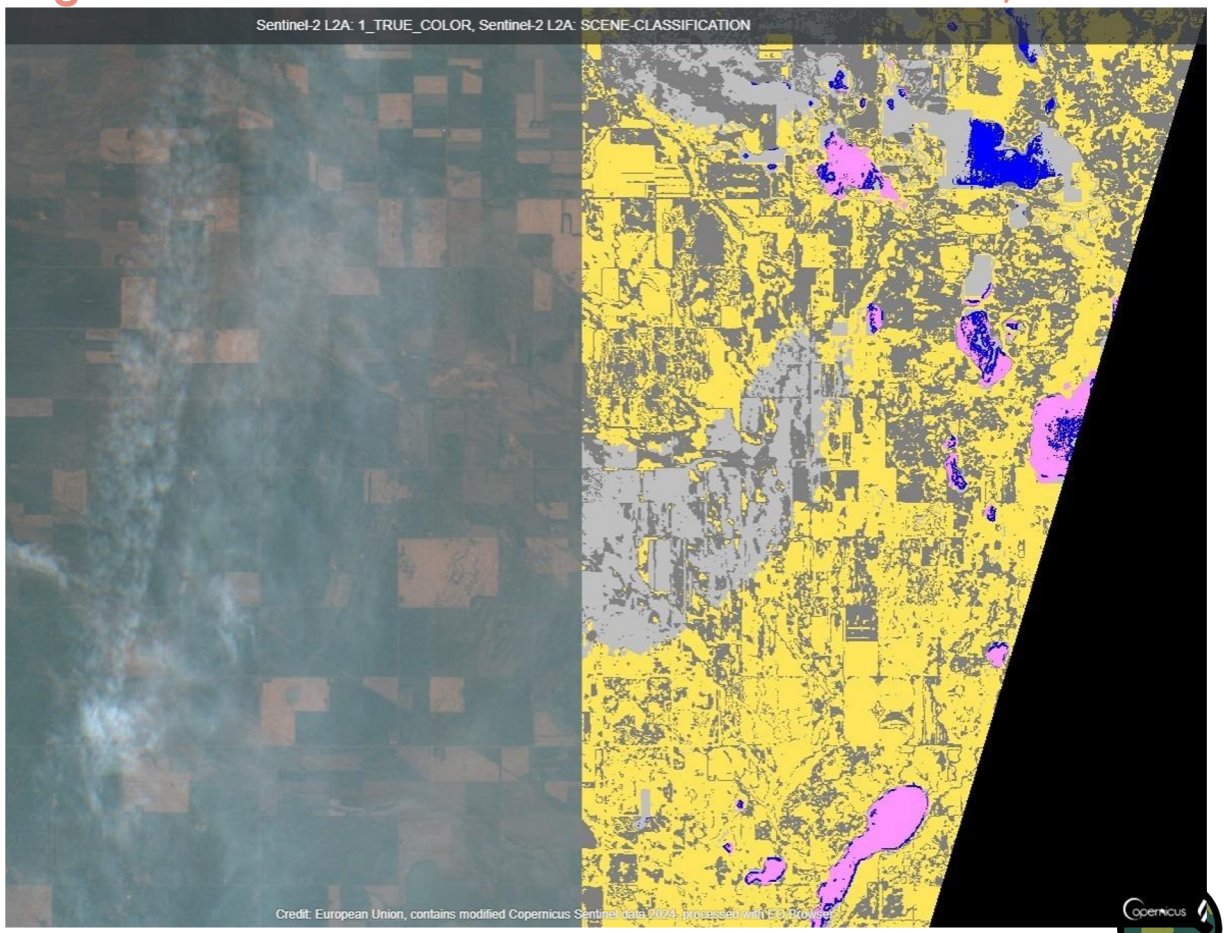
Algoritmos →

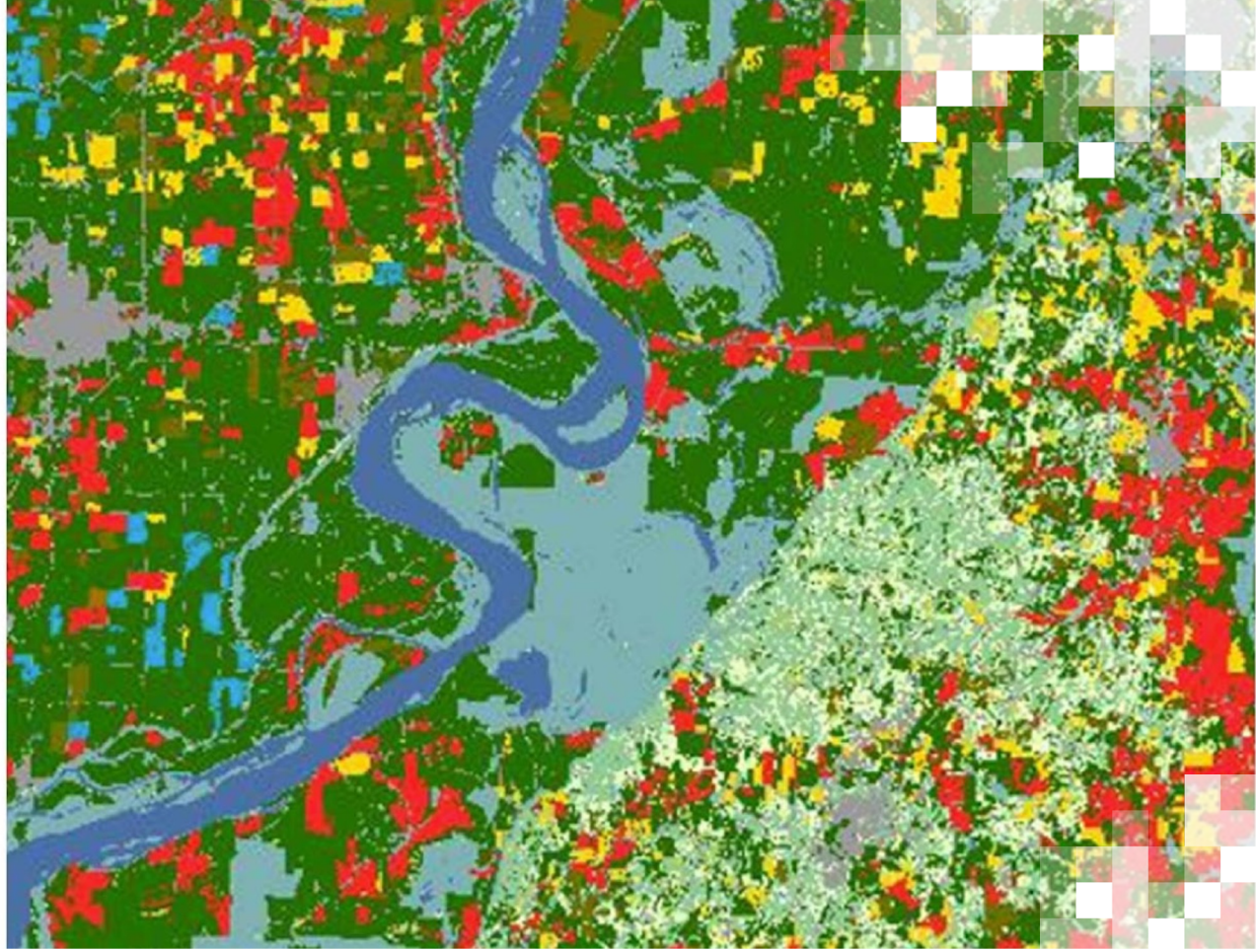
# Problemas/Limitaciones Comunes

La geolocalización/georreferenciación incoherente de píxeles y las etiquetas de clasificación de escenas predeterminadas de los proveedores (por ejemplo, la capa SCL de Sentinel-2) no siempre son precisas.

**Geolocalización inconsistente (desacuerdo de 10-15 m entre imágenes posteriores con 5 días de diferencia)**


**Mala clasificación de la escena, se etiquetaron erróneamente las nubes como suelo desnudo y el agua como nieve debido a la bruma de nubes, etc.**





1ra Parte, Sección 3:

# Demostración del Procedimiento con Databricks (Ejecución del código)

# Acerca de Databricks Community Edition

Enlace a las instrucciones para registrarse para Databricks Community Edition

- Codificación tipo Jupyter Notebook
- Databricks Community Edition permite hasta 10GB de almacenamiento persistente en el "FileStore"
  - Puede almacenar archivos genéricos, tablas y código.
  - Los cuadernos (Notebooks) se almacenan en la parte "workspace"
- Puede poner en marcha pequeñas instancias con 2 cpus,15GB RAM, 130GB de almacenamiento local, Spark está habilitado desde el primer uso
- Cualquier cosa almacenada localmente en la máquina se pierde cuando la instancia se cierra
- Ejecutar el código del cuaderno durante más de aproximadamente 60 minutos hará que cierre. Sin embargo, con tal de estar interactuando con el cuaderno (escribiendo y ejecutando código) normalmente se queda ahí por más tiempo

# Datos y Apuntes sobre la Demostración

Material disponible para esta demostración:

- Tres scripts para el procesamiento de datos de la 1ra Parte (Adquisición de CDL, adquisición de Sentinel-2, manipulación de datos finales)
- Generalización: La tabla CDL puede ser cualquier etiqueta de verdad en el suelo + ubicación del punto + marco de tiempo, y el resto de la adquisición y la modelación de datos pueden seguir siendo los mismos. Por ejemplo, la salud de los cultivos, la etapa vegetativa u otros tipos de clases de cobertura del suelo.
- Cualquier error sistemático de la CDL probablemente pasará a los modelos entrenados en ella.
- Para obtener los datos tal y como existirían después de ejecutar los scripts de esta demostración, descargue los archivos zip que se encuentran con los demás materiales de la capacitación.

Cómo descargar los archivos resultantes de su Databricks FileStore:

- Es poco intuitivo. Para descargar, debe navegar a la ubicación de **SU** archivo usando el formato a continuación.
  - https://community.cloud.databricks.com/files/path/to/folder/filename.extension
    - 'path/to-folder' es la ruta del directorio donde su archivo existe en FileStore.

# Pasos para el Código

Estrategia para procesar y almacenar estos datos:

- 1er Paso: Definir áreas de interés (AOI, por sus siglas en inglés)
- 2do Paso: Adquirir datos CDL correspondientes
- 3er Paso: Buscar y filtrar para encontrar datos satelitales disponibles
- 4to Paso: Adquirir datos satelitales correspondientes
- 5to Paso: Reacomodar archivo Parquet de datos de valores de bandas satelitales en una sola fila por ubicación de píxel y estación, con columnas para componentes de series temporales en forma de listas de valores (p.ej., valores de bandas, fechas de escenas) para apoyar la modelación

# Acerca de los APIs

En lugar de recuperar datos manualmente, las API hacen que la adquisición y el procesamiento de datos sean significativamente más escalables al proporcionar una interfaz consistente para buscar y recuperar grandes cantidades de datos a través de solicitudes web. En esta demostración, nos basamos principalmente en dos API para la adquisición de datos.

- CDL API de NASS geo data (link)
- AWS STAC API para búsquedas de imágenes de Sentinel-2.
  - Los datos ráster de imágenes de Sentinel-2 se pueden descargar a través de enlaces de descarga de URL web a los que podemos acceder directamente una vez conocidos desde la búsqueda de imágenes. Sin embargo, dado que estos son grandes y ralentizan el procesamiento, haremos todo lo posible para minimizar la descarga de escenas superfluas o de bajo impacto.

# Creación de Áreas de Interés (AOI) y Límites

El único paso previo que se necesita normalmente antes de esta parte es definir las AOIs. Para este trabajo se utilizó el nassgeodata web gui para dibujar 7 cuadros y exportarlos como shapefiles ESRI. A continuación, conviértalos en límites (izquierda, abajo, derecha, arriba) en el CRS EPSG:5070 (según lo requiera la API nassgeodata). Ya proporcionamos límites en el código de adquisición de CDL.

Ejemplo de código Python para obtener límites de archivos de forma ESRI [comprimidos] de NASSGEO:

```
import geopandas as gpd
from pyproj import CRS
import pandas as pd
root_path = 'C:/Users/myname/Downloads/'
# List of file paths for esri shapefile boundaries (exported into zips from nassgeo)
paths = [root_path + 'CDL_12345.zip', root_path + 'CDL_6789.zip']
gdf_list = [] # Create a list to hold the GeoPandas dataframes
# Read each shapefile into a GeoPandas dataframe and append it to the list
for path in paths:
    gdf = gpd.read_file("zip://" + path)
    gdf_list.append(gdf)
# Concatenate all the dataframes into a single GeoPandas dataframe
combined_gdf = gpd.GeoDataFrame(pd.concat(gdf_list, ignore_index=True))
target_crs = CRS("EPSG:5070") # Define the target CRS (EPSG:5070)
gdf_5070 = combined_gdf.to_crs(target_crs) # Convert the GeoDataFrame to the target CRS
print(gdf_5070.bounds.apply(lambda row: ', '.join(map(str, map(int, row))), axis=1).to_string(index=False))
```

# Resumen del Código de Adquisición para la Capa CDL

Los resultados de esta primera parte son una versión de la CDL con muestreo espacial reducido para los AOI y años especificados por el usuario.

- Esta parte del código se ejecuta con bastante rapidez (pocos minutos) y da como resultado una tabla de Parquet con un único píxel de 30 $m^2$/año por fila (con la estimación de CDL asociada).
- La siguiente tabla resume las 5 categorías principales de CDL en todas las AOI y el % de todo el conjunto de datos por año que representa cada categoría de CDL. Por ejemplo, la soya de 2021 representó el ~36,6% de la cobertura del suelo para ese año.

| year | CDL | count | total_count | percentage |
|---|---|---|---|---|
| 2021 | Soybeans | 29416 | 80427 | 36.57 |
| 2020 | Soybeans | 28063 | 80151 | 35.01 |
| 2019 | Soybeans | 25422 | 80695 | 31.5 |
| 2019 | Woody Wetlands | 11916 | 80695 | 14.77 |
| 2020 | Woody Wetlands | 11725 | 80151 | 14.63 |
| 2021 | Woody Wetlands | 11765 | 80427 | 14.63 |
| 2020 | Rice | 10809 | 80151 | 13.49 |
| 2021 | Corn | 9331 | 80427 | 11.6 |
| 2021 | Rice | 8599 | 80427 | 10.69 |
| 2019 | Cotton | 8510 | 80695 | 10.55 |
| 2019 | Rice | 8360 | 80695 | 10.36 |
| 2019 | Corn | 7775 | 80695 | 9.64 |
| 2020 | Corn | 6930 | 80151 | 8.65 |
| 2020 | Cotton | 6907 | 80151 | 8.62 |
| 2021 | Cotton | 6847 | 80427 | 8.51 |

# Resultados del Código de Adquisición para Sentinel-2

Resumen: Para cada píxel/año del código de adquisición de CDL, este código adquiere los datos de Sentinel-2 asociados para todo ese año y los guarda en una tabla de Parquet. ***Tenga en cuenta que esta parte tarda mucho tiempo en ejecutarse.*** Se agotará el tiempo de espera de la versión gratuita de Databricks después de una hora.

Los datos "duplicados" tienen la misma fecha y tesela, pero tal vez debido al procesamiento actualizado tienen valores ligeramente diferentes. Estos se dejan entre los datos para este trabajo y se eliminan en el cargador de datos, pero **probablemente sea mejor mantener solo el que tiene el mayor número al final de la tesela (¿último procesamiento?).** Cuando los valores duplicados se deben a la superposición de teselas, elegir uno al azar puede estar bien.

| lon | lat | CDL | scl | coastal | blue | green | red | rededge1 | rededge2 | rededge3 | nir | nir08 | nir09 | swir16 | swir22 | bbox | year | tile | scene_date |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -90.6448304 | 36.46194644 | Corn | 5 | 371 | 604 | 802 | 1024 | 1266 | 1396 | 1536 | 1770 | 1764 | 1784 | 1988 | 1261 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_1 | 1/5/2021 |
| -90.6448304 | 36.46194644 | Corn | 5 | 114 | 422 | 678 | 951 | 1214 | 1344 | 1495 | 1744 | 1739 | 1770 | 2011 | 1267 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_0 | 1/5/2021 |
| -90.6448304 | 36.46194644 | Corn | 5 | 381 | 489 | 670 | 904 | 1037 | 1169 | 1245 | 1420 | 1444 | 1421 | 1864 | 1267 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_1 | 1/13/2021 |
| -90.6448304 | 36.46194644 | Corn | 5 | 230 | 364 | 568 | 855 | 997 | 1136 | 1221 | 1412 | 1436 | 1400 | 1883 | 1273 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_0 | 1/13/2021 |
| -90.6448304 | 36.46194644 | Corn | 3 | 422 | 483 | 438 | 493 | 542 | 567 | 628 | 696 | 695 | 633 | 855 | 680 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_1 | 1/15/2021 |
| -90.6448304 | 36.46194644 | Corn | 3 | 255 | 350 | 319 | 411 | 489 | 519 | 585 | 646 | 651 | 589 | 858 | 683 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_0 | 1/15/2021 |
| -90.6448304 | 36.46194644 | Corn | 5 | 570 | 782 | 1068 | 1376 | 1741 | 1881 | 2014 | 2092 | 2188 | 2177 | 3433 | 2999 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_2 | 1/23/2021 |
| -90.6448304 | 36.46194644 | Corn | 5 | 580 | 800 | 1100 | 1434 | 1758 | 1892 | 2024 | 2132 | 2197 | 2174 | 3466 | 3022 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_0 | 1/23/2021 |
| -90.6448304 | 36.46194644 | Corn | 5 | 1093 | 1023 | 1222 | 1584 | 1917 | 2068 | 2166 | 2220 | 2428 | 2476 | 2400 | 1853 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_1 | 1/28/2021 |
| -90.6448304 | 36.46194644 | Corn | 3 | 785 | 948 | 1138 | 1400 | 1676 | 1775 | 1936 | 2162 | 2134 | 1941 | 2729 | 1925 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_0 | 2/2/2021 |
| -90.6448304 | 36.46194644 | Corn | 5 | 539 | 668 | 847 | 1064 | 1343 | 1427 | 1553 | 1676 | 1780 | 1800 | 2412 | 1670 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_1 | 2/24/2021 |
| -90.6448304 | 36.46194644 | Corn | 5 | 476 | 623 | 827 | 1050 | 1327 | 1403 | 1526 | 1648 | 1753 | 1770 | 2410 | 1670 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_0 | 2/24/2021 |
| -90.6448304 | 36.46194644 | Corn | 5 | 565 | 706 | 933 | 1214 | 1449 | 1557 | 1678 | 1874 | 1863 | 1936 | 2656 | 1986 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_1 | 3/4/2021 |
| -90.6448304 | 36.46194644 | Corn | 5 | 577 | 718 | 946 | 1210 | 1453 | 1558 | 1684 | 1866 | 1862 | 1940 | 2658 | 1993 | 465073, 1479393, 583994, 1504168 | 2021 | 15SYA_0 | 3/4/2021 |

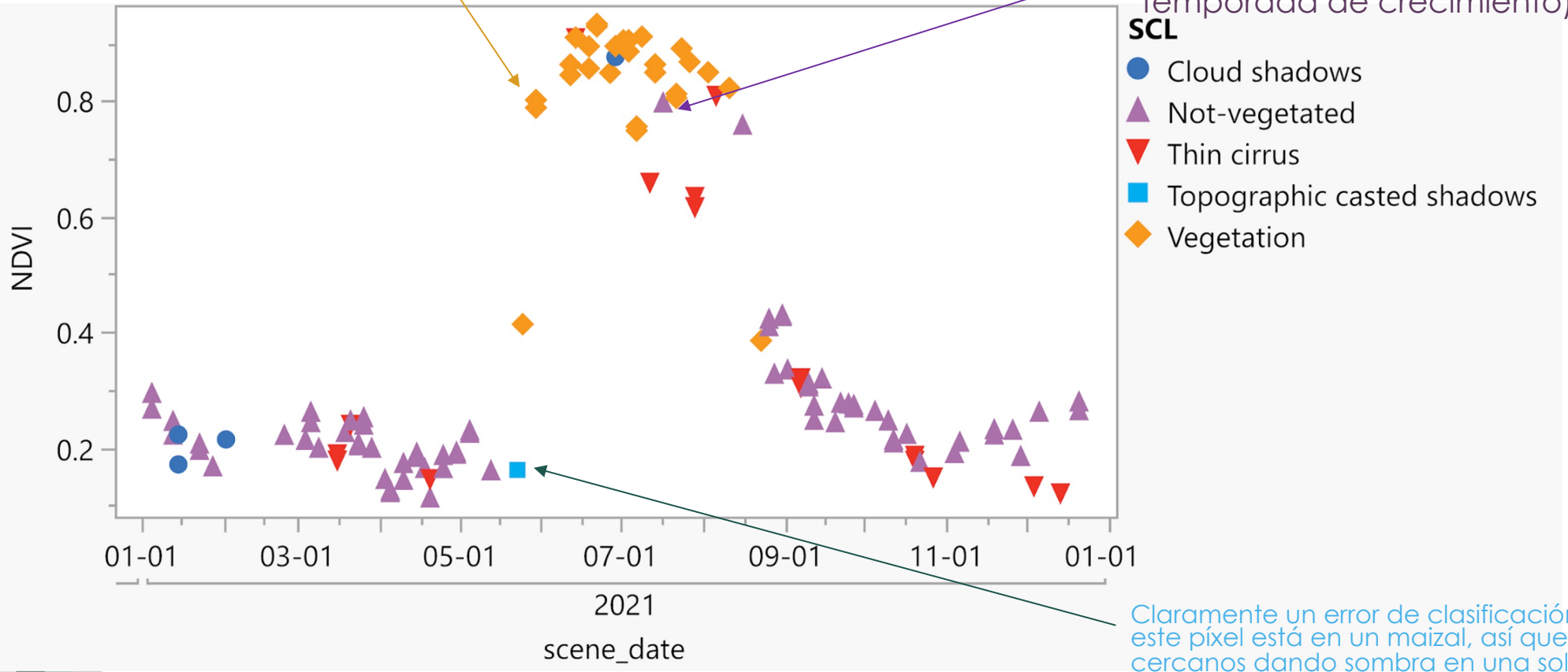
**…123 filas en total disponibles para este píxel/año en particular. Cada fila incluye los valores de banda para esa ubicación de una sola escena/fecha.**

# Sentinel-2 – Código de Adquisiciónón (Diagramado)

Ejemplo de datos de un solo píxel/año de maíz. Incluye datos duplicados.

Ejemplo de datos "duplicados" (valores de banda ligeramente diferentes para la misma escena)

Claramente un error de clasificación (solo sucede una vez y encima durante el medio de la temporada de crecimiento).




Claramente un error de clasificación (solo ocurre una vez y este píxel está en un maizal, así que no hay objetos cercanos dando sombra en una sola instancia).

# Manipulación de Datos Finales

Una manipulación rápida final de los datos combina todas las escenas disponibles en una sola fila para cada píxel/año.

- Varias columnas (bands, tiles, img dates, scl_vals) son todas las listas de valores de las escenas para cada fila
- Por ejemplo, un píxel con 123 escenas tendría valores de 123*12 en la lista de la columna "bands".
- Las listas de valores se convierten en cadenas binarias para un almacenamiento eficiente (eliminando el tipo de datos "lista" y las comas para todo excepto la columna "tiles" que es teselas)

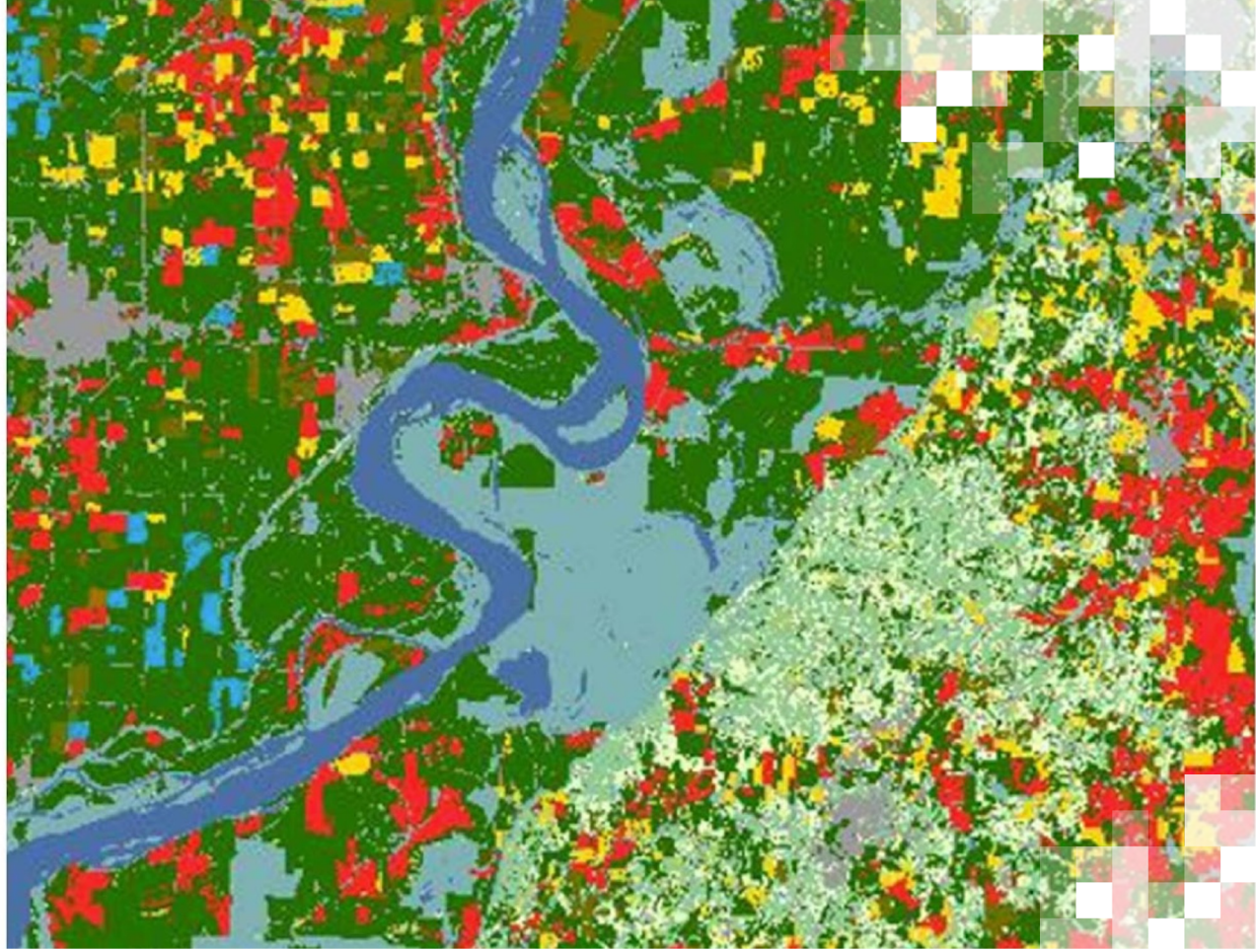
## Tabla final

| lon | lat | # scenes | bands | tiles | img dates | scl_vals | bbox | year | CDL |
|---|---|---|---|---|---|---|---|---|---|
| -89.2973 | 36.85473 | 38 | AUECmAOEBM | MTZTQkZfMCw | Re5GAkYgRjRG | BQUKBQUFCgoF | 549309, | 2019 | Corn |
| -89.3087 | 36.95602 | 71 | AEYBMQHnAq4 | MTZTQkZfMCw | Re5F7kYCRgJGI | BQUFBQUFBQUF | 549309, | 2019 | Corn |
| -89.3206 | 36.88018 | 36 | ATACXAOZBCw | MTZTQkZfMCw | Re5GAkYgRjRG | BQUKBQUFCgUF | 549309, | 2019 | Corn |
| -89.3322 | 36.86472 | 33 | ARcB4QKxA4M | MTZTQkZfMCw | Re5GIEY0RjIGP | BQoFBQUKBQUF | 549309, | 2019 | Corn |
| -89.3382 | 36.79661 | 37 | AQQBugJiA1ID5 | MTZTQkZfMCw | Re5GAkYgRjRG | AgIKBQUFCgoFB | 549309, | 2019 | Corn |
| -89.3394 | 36.84096 | 37 | ACgBcwHOArk | MTZTQkZfMCw | Re5GIEY0RjIGP | BQoFBQUKCgUF | 549309, | 2019 | Corn |
| -89.3493 | 36.9019 | 37 | AO4B5gJSAvgD | MTZTQkZfMCw | Re5GAkYgRjRG | BQUKBQUFCgUF | 549309, | 2019 | Corn |
| -89.3552 | 36.95054 | 72 | Ak8DYAQ4BbgC | MTZTQkZfMCw | Re5F7kYCRgJGI | BQUFBQUFBQUF | 549309, | 2019 | Corn |
| -89.3577 | 36.74937 | 37 | ANIAmQFTAfoC | MTZTQkZfMCw | Re5GAkYgRjRG | BQUKBQUFCgoF | 549309, | 2019 | Corn |
| -89.3632 | 36.97514 | 70 | ACsBSwIVAlUC | MTZTQkZfMCw | Re5F7kYCRgJGI | BQUFBQUFBQUF | 549309, | 2019 | Corn |

## Ejemplo de columnas binarias decodificadas

| decoded band vals | decoded tiles | decoded img dates | decoded scl_vals |
|---|---|---|---|
| 321,664,900,1228,1415 | 16SBF_0,16SBF_0,16SE | 2019-01-06,2019-01-26,201 | 5,5,10,5,5,5,10,10, |
| 70,305,487,686,843,98 | 16SBF_0,16SBG_0,16S | 2019-01-06,2019-01-06,201 | 5,5,5,5,5,5,5,5,5,5, |
| 304,604,921,1068,1473 | 16SBF_0,16SBF_0,16SE | 2019-01-06,2019-01-26,201 | 5,5,10,5,5,5,10,5,5 |
| 279,481,689,899,1075, | 16SBF_0,16SBF_0,16SE | 2019-01-06,2019-02-25,201 | 5,10,5,5,5,10,5,5,5 |
| 260,442,610,850,999,1 | 16SBF_0,16SBF_0,16SE | 2019-01-06,2019-01-26,201 | 2,2,10,5,5,5,10,10, |
| 40,371,462,697,744,81 | 16SBF_0,16SBF_0,16SE | 2019-01-06,2019-02-25,201 | 5,10,5,5,5,10,10,5, |
| 238,486,594,760,872,9 | 16SBF_0,16SBF_0,16SE | 2019-01-06,2019-01-26,201 | 5,5,10,5,5,5,10,5,5 |
| 591,864,1080,1464,174 | 16SBF_0,16SBG_0,16S | 2019-01-06,2019-01-06,201 | 5,5,5,5,5,5,5,5,5,5, |
| 210,153,339,506,738,9 | 16SBF_0,16SBF_0,16SE | 2019-01-06,2019-01-26,201 | 5,5,10,5,5,5,10,10, |
| 43,331,533,597,755,11 | 16SBF_0,16SBG_0,16S | 2019-01-06,2019-01-06,201 | 5,5,5,5,5,5,5,5,5,5, |

Cada fila es un píxel de 30m$^2$ de CDL de un año



# 1ra Parte:
# **Resumen**

# Resumen

- Las API nos permiten automatizar y escalar canalizaciones de procesamiento de datos muy grandes en preparación para el análisis y la creación de modelos.
- El almacenamiento de datos en formato Parquet y el uso de Spark/Databricks para consultar/dinamizar o manipular los datos permite una rápida investigación y transformación.
  - El formato Parquet tiene abstracciones útiles como particiones, que también son directorios
- Una forma conveniente para modelar datos de imágenes de series temporales implica almacenarlos en formato de tabla de Parquet, donde cada fila representa un píxel para un intervalo de tiempo determinado y tiene columnas de:
  - Valores de banda, fechas de escena, valores de clasificación de escena en ese intervalo de tiempo
  - Escalares para lat., lon. que representan el punto central del píxel (podría sustituir una celda hexadecimal Uber H3 o Google S2 en su lugar)
  - Un objetivo de predicción (verdad sobre el terreno)

# Mirando Hacia la 2da Parte

- Procesaremos datos en preparación para el entrenamiento de modelos con TensorFlow
- Dividiremos correctamente los datos en grupos train/val/test* para evitar la "fuga de datos"
- Convertiremos imágenes de series temporales espaciadas irregularmente en series temporales agrupadas en preparación para el entrenamiento del modelo
- Modificaremos etiquetas de CDL para alinearlas con nuestros objetivos de entrenamiento

*entrenamiento/validación/prueba

# Tarea y Certificados

- **Tarea:**
  - Habrá una tarea asignada
  - Abre el 19 de marzo de 2024
  - Acceso desde la página web de la capacitación
  - Debe enviar sus respuestas vía Formularios de Google
  - **Fecha límite: 1ro de abril de 2024**

- **Certificado de Finalización de Curso:**
  - Asista a las tres sesiones en vivo (la asistencia se registra automáticamente)
  - Complete la tarea dentro del plazo estipulado
  - Recibirá un certificado por correo electrónico aproximadamente dos meses después de la conclusión del curso.

# Datos de Contacto

Formadores:

- John Just (John Deere)
  - JustJohnP@JohnDeere.com
- Erik Sorensen
  - SorensenErik@JohnDeere.com
- Sean McCartney
  - Sean.McCartney@nasa.gov

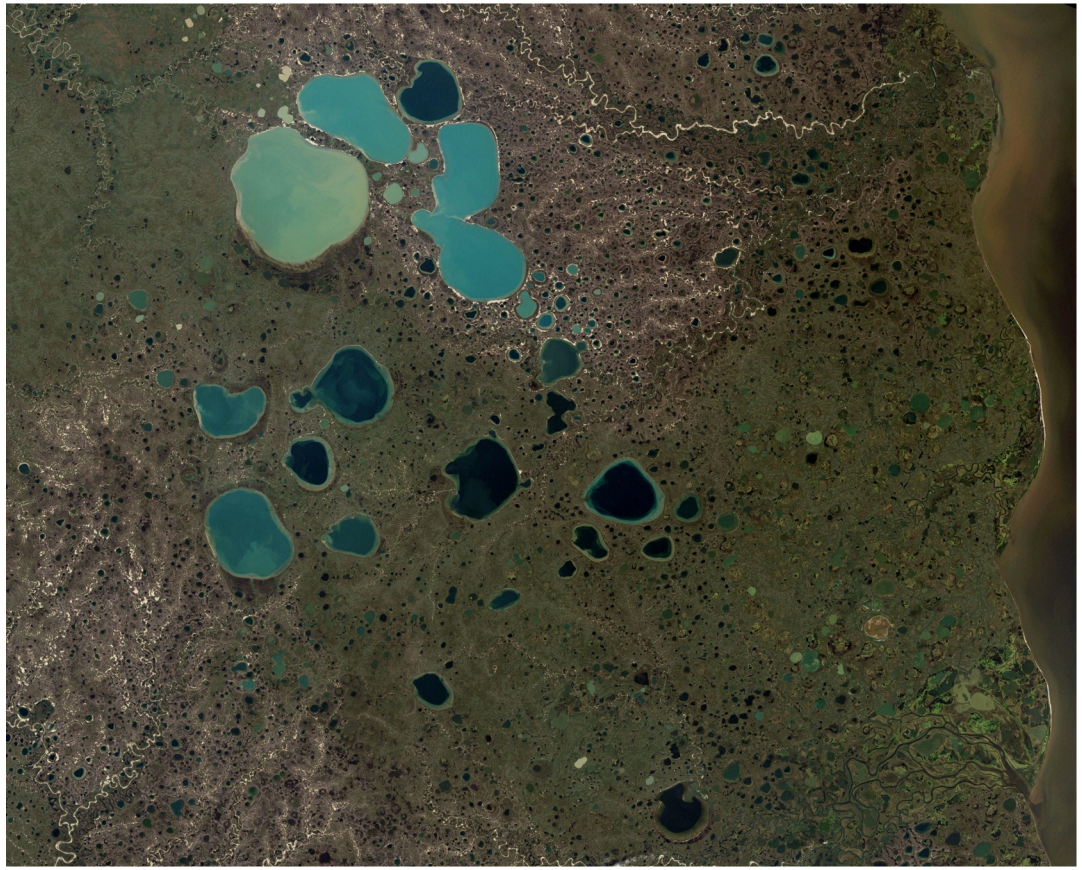
- Página web de ARSET
- ¡Síganos en X (antiguamente Twitter)!
  - @NASAARSET
- ARSET YouTube

Visite nuestros Programas Hermanos:

- DEVELOP
- SERVIR

# ¿Preguntas?

- Por favor escriba sus preguntas en la casilla denominada “Questions”. Las responderemos en el orden en que fueron recibidas.
- Publicaremos las preguntas y respuestas a la página web de la capacitación después de la conclusión del webinar.



https://earthobservatory.nasa.gov/images/6034/pothole-lakes-in-siberia


APPLIED REMOTE SENSING TRAINING PROGRAM
TRAIN - EMPOWER - ADVANCE

# ¡Gracias!