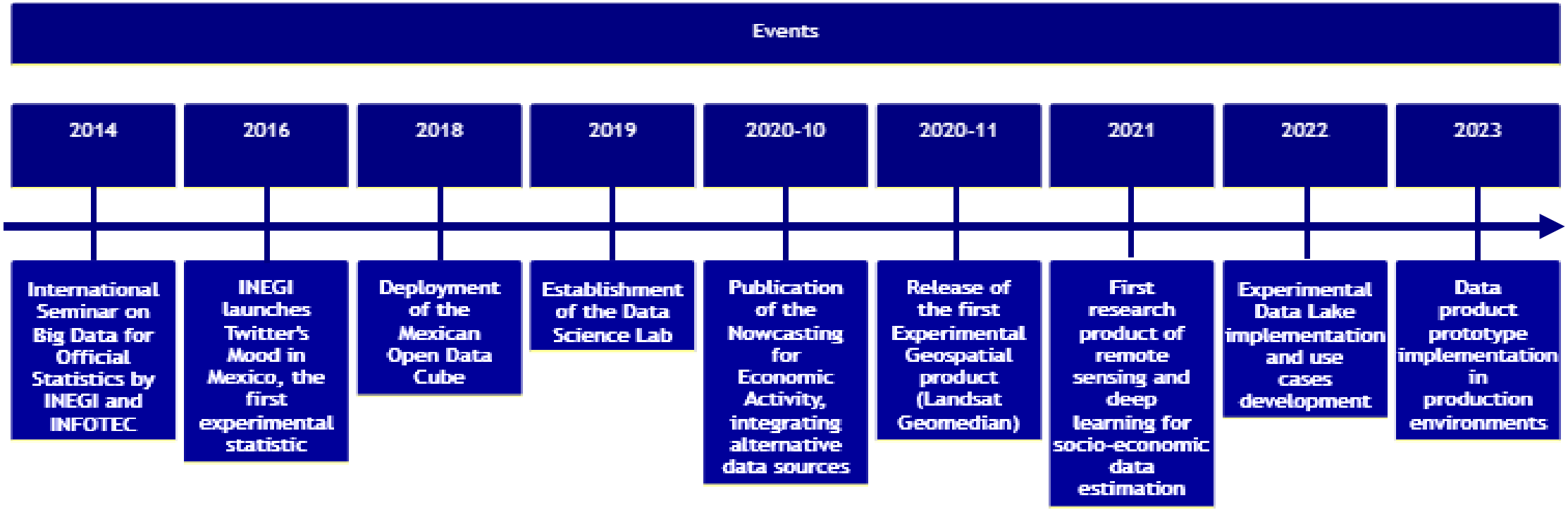
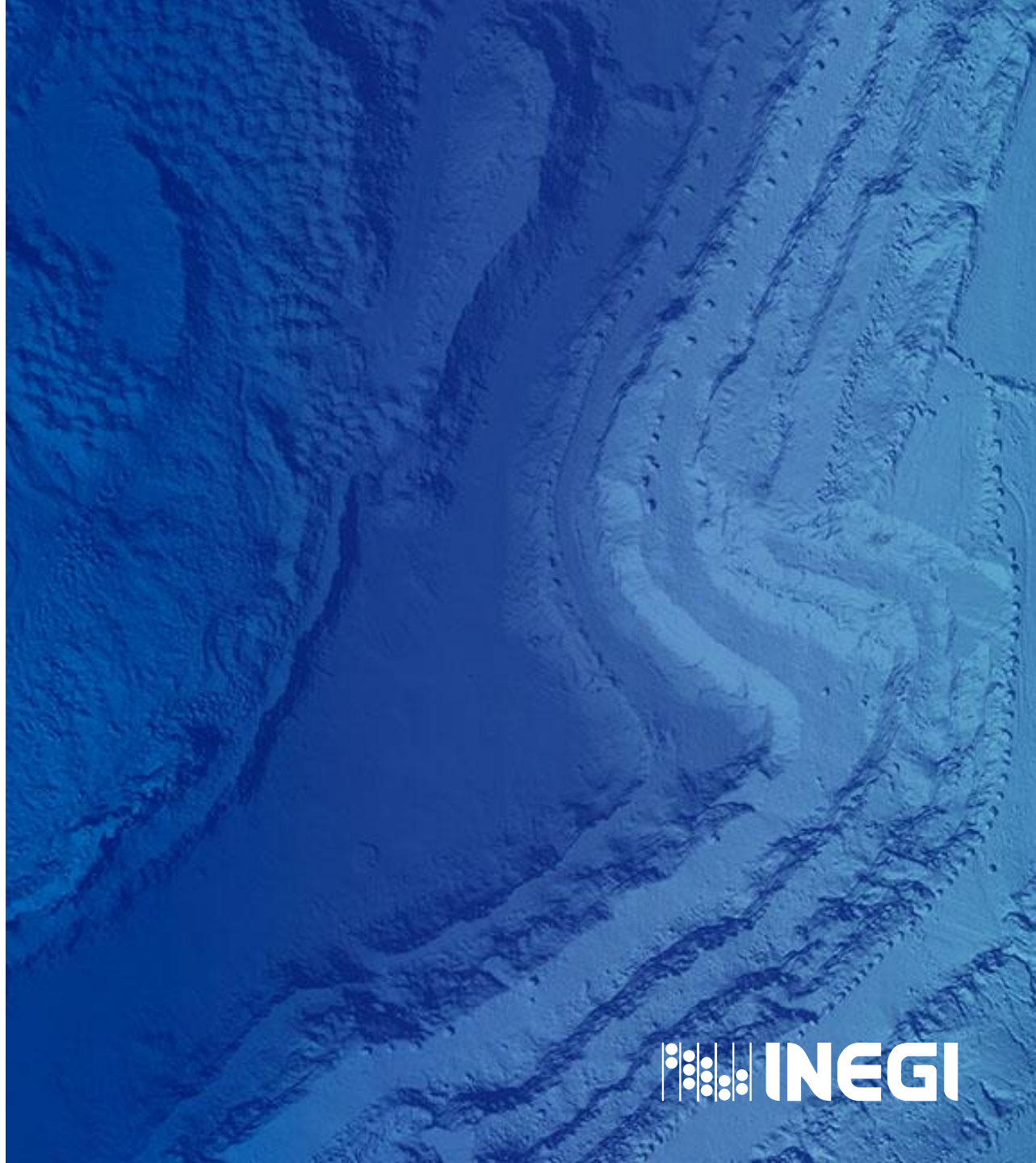

INEGI
Laboratorio de Ciencia de Datos del INEGI

Cronología de la Ciencia de Datos del INEGI

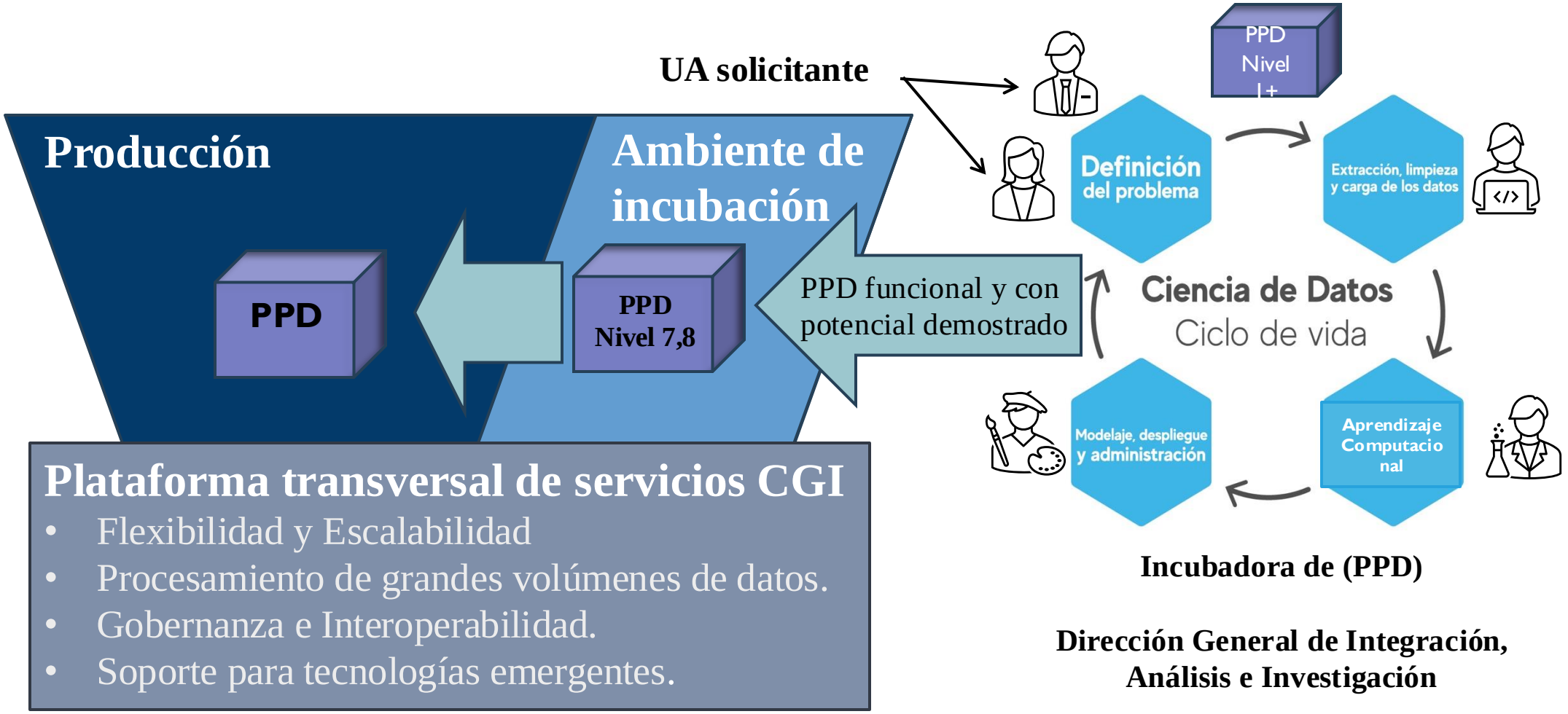


Propósitos del Laboratorio de Ciencia de Datos

- Desarrollar capacidades para aprovechar fuentes de datos alternativas y métodos modernos para la producción de información.
- Generar nuevos productos (análisis estadístico y geoespacial).
- Hacer que los procesos de producción sean más eficientes.
- Brindar un mejor servicio a nuestros usuarios.



Proceso de implementación de Prototipos de Productos de Datos (PPD) en producción

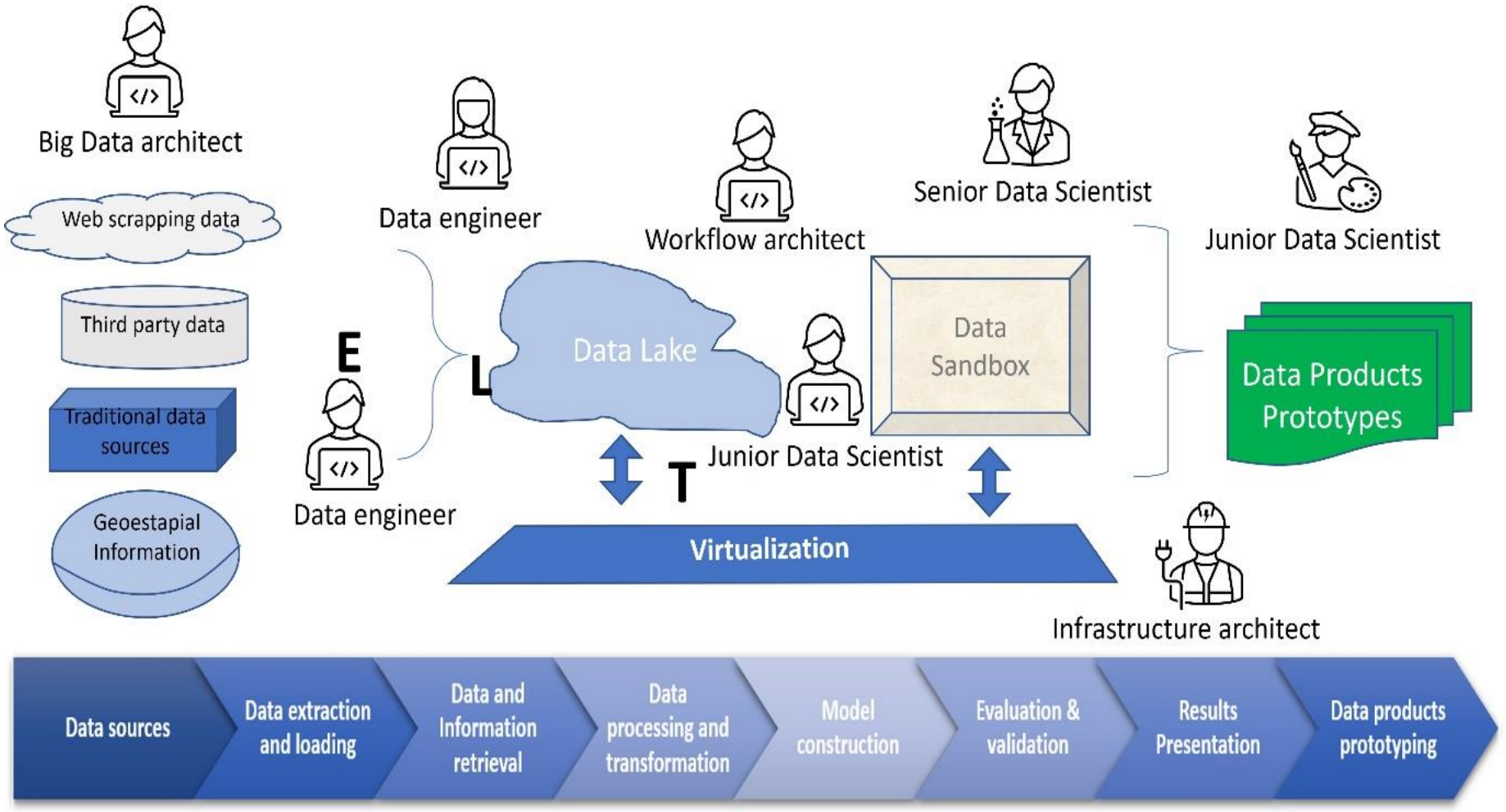


Incubadora de (PPD)

Dirección General de Integración, Análisis e Investigación

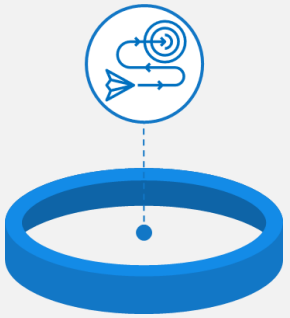
INEGI

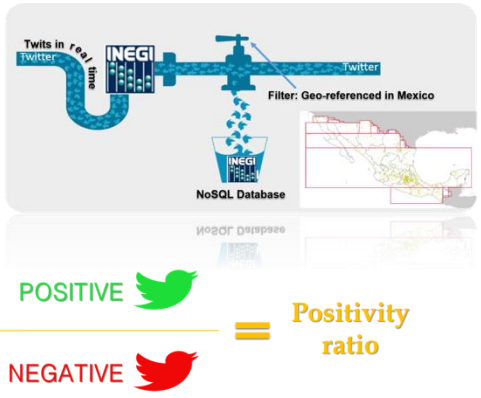
Creación de un equipo multidisciplinar



Lago de datos

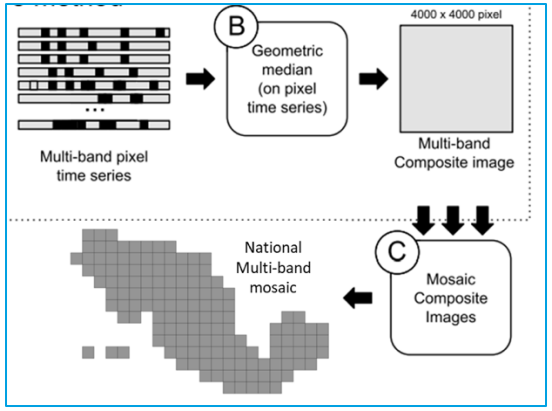
Plataforma de integración de datos



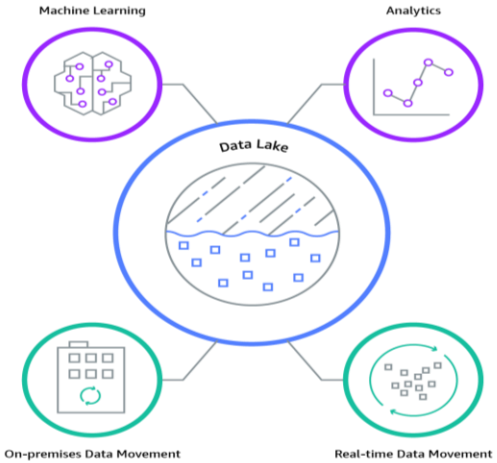


Redes sociales y datos de internet

Productos experimentales estadísticos.



Datos geoespaciales



Información Estadística
Fuentes Estadísticas Tradicionales

Extracción, validación, carga de datos

Almacenamiento – Resguardo de datos

Homologación, Transformación de datos

Integración de datos

Analítica y ciencia de datos

Visualización de datos

python

docker

R

ubuntu

jupyter

CPU

Core1 Core2 Core3 Core4

GPU

Scala

portainer

docker

ubuntu

APACHE Spark

CPU

Core1 Core2 Core3 Core4

GPU

GDAL

QGIS

AutoKeras

TensorFlow

PyTorch

CPU

Core1 Core2 Core3 Core4

GPU

Red Hat

Cluster and Grid Sandbox-ITo (Areneros Preproducción Capacitación – 4 nodos) Procesamiento 160 cores en cpu´s, Memoria Ram 1.5 TB, Almacenamiento 16 TB

Cluster and Grid Sandbox (LLM & SML – 4 nodos) Procesamiento 96 cores en cpu´s, Memoria Ram 2 TB, Almacenamiento 48 TB, GPU´s 1,152 núcleos Tensor (4 NVIDIA RTX 6000 ADA - 48 Gb).

Cluster and Grid HPC (High Performance Computing), Procesamiento 448 cores en cpu´s y 4 gpu´s [Tensor Core + TeraFlops], Memoria Ram 3 TB, Almacenamiento 30 TB, GPU´s 2,560 núcleos Tensor.

Grid Storage Raid (Data Lake | Lago de Datos)

NAS (Network Attached Storage)
Almacenamiento 50 TB

SAN (Storage Area Network)
Almacenamiento 20 TB

INEGI

Infraestructura Tecnológica de Ciencia de Datos

Aplicaciones de aprendizaje profundo y teledetección

Análisis de Expansión Urbana:

- Los algoritmos de aprendizaje profundo analizan las imágenes satelitales para monitorear y mapear con precisión el crecimiento urbano.
- Beneficios implementados: Garantiza información actualizada para la planificación urbana.

Monitoreo de Tierras Agrícolas:

- Los algoritmos procesan imágenes satelitales para identificar y clasificar áreas agrícolas, mejorando la precisión de los datos agrícolas.
- Beneficios implementados: Proporciona estadísticas confiables para la gestión agrícola y la formulación de políticas.

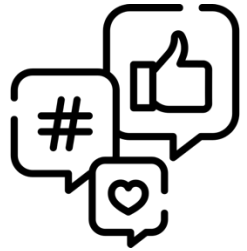
Identificación de Áreas Desfavorecidas:

- Utiliza técnicas de aprendizaje profundo con imágenes satelitales y datos censales para identificar y evaluar áreas urbanas desfavorecidas.
- Beneficios implementados: Ayuda en la formulación de políticas específicas y la asignación de recursos al identificar áreas de vulnerabilidad.

Procesamiento del Lenguaje Natural (PNL)

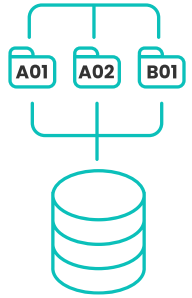
Análisis de Sentimiento en Redes Sociales:

- Las técnicas de PNL analizan el contenido de las redes sociales para monitorear el sentimiento del público y detectar tendencias.
- Beneficios implementados: Mejora la capacidad de respuesta de las políticas públicas a las dinámicas sociales en tiempo real.



Clasificación automatizada de texto en encuestas:

- El NLP se utiliza para clasificar y procesar automáticamente los datos textuales en las encuestas, lo que mejora la eficiencia en el manejo de los datos.
- Beneficios implementados: Reduce el trabajo manual y acelera la disponibilidad de los resultados de las encuestas.



Explorando el Potencial de LLM

Aplicaciones y desarrollos actuales:

Respuestas automatizadas a las consultas de los usuarios:

- Uso de LLM para interpretar consultas complejas y proporcionar respuestas de datos claras y precisas.

Estado: Mejora continua de la precisión de las respuestas y la participación de los usuarios.

Informes de datos personalizados:

- Desarrollo de capacidades de LLM para generar informes y visualizaciones personalizados basados en las entradas del usuario.

Estado: Desarrollo activo para mejorar la personalización y garantizar la relevancia y claridad de los datos.

LINEAMIENTOS ESTRATÉGICOS

- Fomentar la innovación en el uso de nuevas fuentes de datos, metodologías y procesos avanzados en la producción.
- Maximizar la interoperabilidad para promover el uso intensivo y extensivo de la Información Estadística y Geográfica.
- Mejorar y modernizar la infraestructura tecnológica y de datos para procesar e integrar de manera eficiente diversas fuentes de datos, con la capacidad de aprovechar los métodos de IA
- Amplificar la colaboración aprovechando los datos y la información de alta calidad de los sectores privado, académico y social para maximizar el potencial de las fuentes de datos alternativas.

Observaciones finales

La adopción de métodos de ciencia de datos podría implicar un cambio de paradigma en los procesos de producción de información.

La colaboración con las áreas de TI es clave para hacer viables los proyectos de Data Science y Big Data.

La Ciencia de Datos presenta una oportunidad para la modernización de las ONE.

Sin CIENCIA no hay Ciencia de Datos

Para la implementación en la producción, es esencial demostrar valor y garantizar la sostenibilidad.

INEGI