Vigesimotercera Reunión del Comité Ejecutivo de la
Conferencia Estadística de las Américas de la CEPAL
28 y 29 de agosto de 2024

Conferencia Estadística de las Américas de la CEPAL | Statistical Conference of the Americas of ECLAC

Innovación basada en la ciencia de datos para la producción de información estadística y geográfica

Science-Driven Innovation for the Production of Statistical and Geographic Information

INEGI, México

Línea de tiempo de la Ciencia de Datos del INEGI
INEGI´s Data Science Timeline

Línea de tiempo de la Ciencia de Datos del INEGI
INEGI´s Data Science Timeline

Fase experimental / Experimental stage			Implementación en producción/ Implementation in production	
2020		2021	2022	2023	2024
Publicación del *Nowcasting* para la Actividad Económica (IOAE) / Publication of the Nowcasting for Economic Activity	Lanzamiento del primer producto Geoespacial Experimental (Geomediana Landsat) / Release of the first Experimental Geospatial product	Primer producto de investigación en percepción remota y aprendizaje profundo para la estimación de datos socioeconómicos / Debut of remote sensing and deep learning research for socio-economic data estimation	Implementación del Lago de Datos Experimental y desarrollo de casos / Experimental Data Lake implementation and case development	Demostraciones de prototipos de productos de datos en entornos de producción / Data product prototype demonstrations in production environment	CDGM se convierte en un programa de información en línea de producción / CDGM becomes a production online information program

Propósito 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 las personas usuarias de nuestra información.

Purpose of Data Science Lab

- Develop capabilities to leverage alternative data sources and modern methods for the production of information.
- Generate new products (statistical and geospatial analysis).
- Make production processes more efficient.
- Provide a better service to our users.

Creación de un equipo multidisciplinario

PERFILES DEL EQUIPO

1 Científico de datos principal

1Arquitecto de Big Data

1 Arquitecto de Infraestructura

2 Científicos de datos senior

2 Científicos de datos junior

2 Ingenieros de datos

Creating a Multidisciplinary Team

TEAM PROFILES

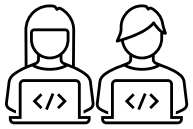
1 Lead Data Scientist

1 Big Data Architect

1 Infrastructure Architect

2 Senior Data Scientist

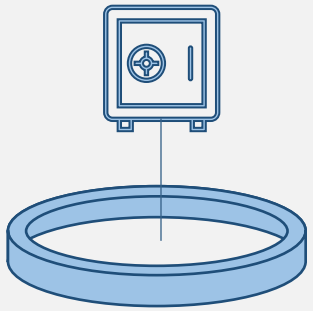
2 Junior Data Scientist

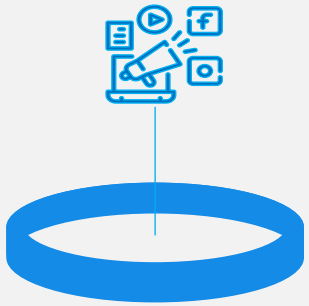


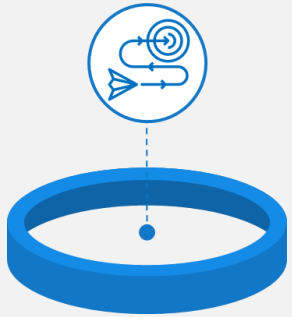
2 Engineer

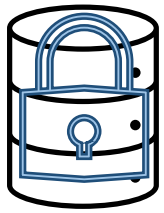
Lago de datos / Data Lake

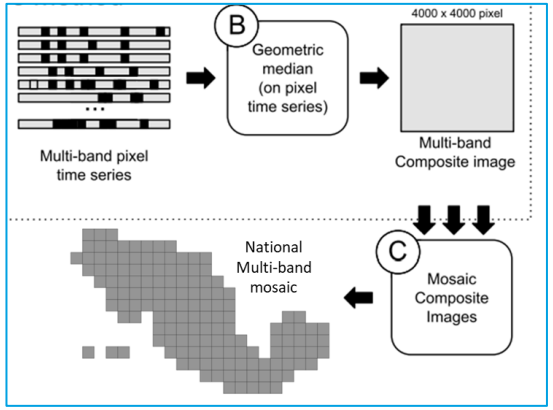
Integración de Datos/Data integration

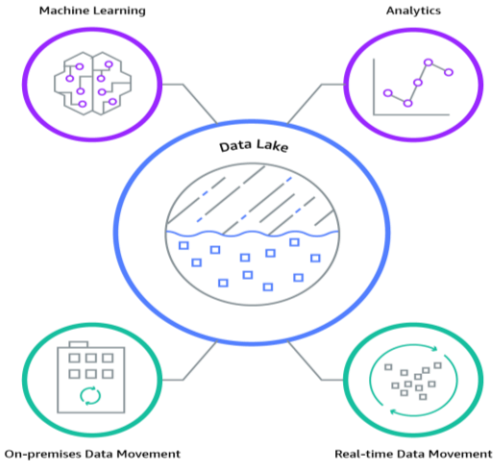












Datos de terceros / Third-party data

Datos disponibles en internet / Data from internet

Datos Geoespaciales / Geospatial Data

Fuentes Estadísticas Tradicionales / Traditional Statistical Sources

Infraestructura Tecnológica de Ciencia de Datos

Data Science Technological Infrastructure

Lenguajes de programación para Ciencia de Datos.

Programming languages for Data Science.

Automatización de flujos de datos (Lago de Datos).

Data Flow Automation (Data Lake).

Cómputo de alto rendimiento.

High-performance computing.

Tecnología de Contenedores.

Container Technology.

Herramientas especializadas de visualización.

Specialized visualization tools.

Aplicaciones de aprendizaje profundo y percepción remota

Análisis de la Expansión Urbana:

Se analizan las imágenes satelitales para monitorear y mapear el crecimiento urbano, proporcionando información actualizada para la planificación de operativos.

Monitoreo de Tierras Agrícolas:

Se procesan imágenes satelitales para identificar áreas agrícolas, lo que puede mejorar la planeación y gestión de censos agrícolas.

Identificación de Áreas Desfavorecidas:

Se identifican áreas urbanas desfavorecidas, lo que puede apoyar la formulación de políticas específicas y la asignación de recursos.

Deep Learning and Remote Sensing Applications

Urban Expansion Analysis:

Satellite images are analyzed to monitor and map urban growth, providing updated information for operational planning.

Agricultural Land Monitoring:

Satellite images are processed to identify agricultural areas, which can improve the planning and management of agricultural censuses.

Identification of Deprived Areas:

Deprived urban areas are identified, which can support the formulation of specific policies and the allocation of resources.

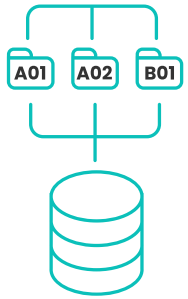
Procesamiento de Lenguaje Natural

Análisis de sentimientos en redes sociales:

- Permitió analizar el contenido de las redes sociales para monitorear el sentimiento público y detectar tendencias.
- Beneficios obtenidos: desarrollo de capacidades para el procesamiento de grandes volúmenes de información en formato de texto.

Codificación automatizada de texto en encuestas:

- Proceso y clasifica automáticamente los datos textuales en las encuestas (Ej. actividades económicas y de empleo).
- Beneficios potenciales: Reduce el trabajo manual y acelera la disponibilidad de los resultados de las encuestas.



Natural Language Processing (NLP)

Social Media Sentiment Analysis:

- It made it possible to analyze social media content to monitor public sentiment and spot trends.
- Benefits obtained: development of capacities for processing large volumes of information in text format.

Automated text coding in surveys:

- Automatically processes and classifies textual data in surveys (e.g. economic and employment activities).
- Potential benefits: Reduces manual work and speeds up the availability of survey results.

Explorando el potencial de los grandes modelos de lenguaje

Respuestas automatizadas a las consultas de las personas usuarias:

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

Informes de datos personalizados:

- Desarrollo de capacidades para generar informes basados en las necesidades de las personas usuarias, con actualización automatizada.

Exploring the Potential of Large Language Models (LLMs)

Automated user inquiry responses:

- Use LLMs to interpret complex queries and provide clear and accurate data responses.
- Status: Continually improving response accuracy and user engagement.

Personalized data reporting:

- Develop LLMs capabilities for generating customized reports and visualizations based on user inputs.
- Status: Active development to enhance personalization and ensure data relevance and clarity.

Direcciones y mejoras futuras

Plataformas de datos interactivas:

- Integración de IA generativa en plataformas que admitan la exploración dinámica de datos, gráficos, mapas, entre otros.

Herramientas inclusivas en captación de información

- Procesamiento de información en lenguas indígenas para una comprensión contextual más profunda durante el proceso de captación.
- Permite ampliar el alcance y la usabilidad en diversos grupos lingüísticos y culturales nativos, promoviendo una alfabetización de datos más amplia.

Future Directions and Enhancements

Interactive Data Platforms:

- Integration of generative AI into platforms that support dynamic exploration of data, charts, maps, and more.

Inclusive tools for information collection

- Information processing in Indigenous languages for deeper contextual understanding during the recruitment process.
- It allows for expanded reach and usability across diverse native language and cultural groups, promoting broader data literacy.

Acciones hacia adelante

- **Fomentar la innovación** y la actualización en el uso de nuevas fuentes de datos, metodologías y procesos avanzados en la producción.
- **Maximizar la interoperabilidad y gobernanza de datos** para promover el uso intensivo y extensivo de la Información.
- 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 conocimientos de alta calidad de los sectores privado, académico y social para maximizar el potencial de las fuentes de datos alternativas.
- Actualizar normativas y regulaciones para **incorporar consideraciones éticas** en el aprovechamiento del Big Data y la IA.

Actions Moving Forward

- **Promote innovation** and updating in the use of new data sources, methodologies, and advanced processes in production.
- **Maximize data interoperability and governance** to promote intensive and extensive use of information.
- Improve and **modernize the technological and data infrastructure** to efficiently process and integrate diverse data sources, with the capability to leverage AI methods.
- **Amplify collaboration** by leveraging high-quality data and insights from the private, academic, and social sectors to maximize the potential of alternative data sources.
- Update norms and regulations to **incorporate ethical considerations** in the use of Big Data and AI.

Gracias

Isabel Islas
isabel.islas@inegi.org.mx