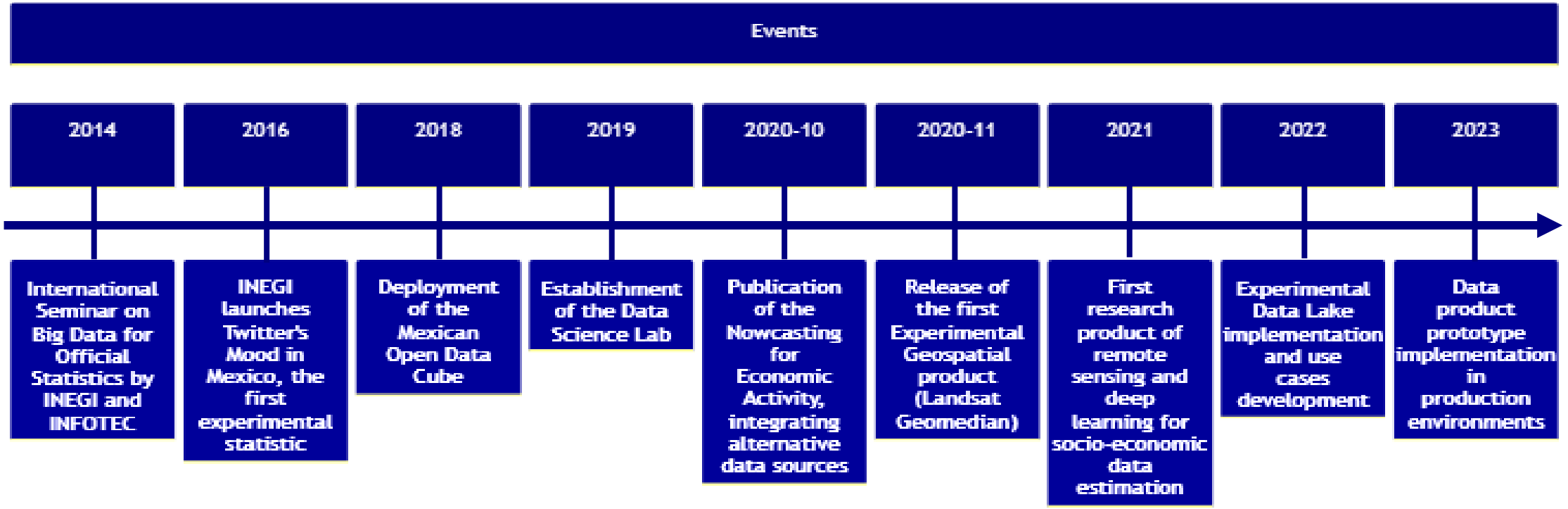
INEGI's Data Science Transformation

DSNL, 2nd Sprint

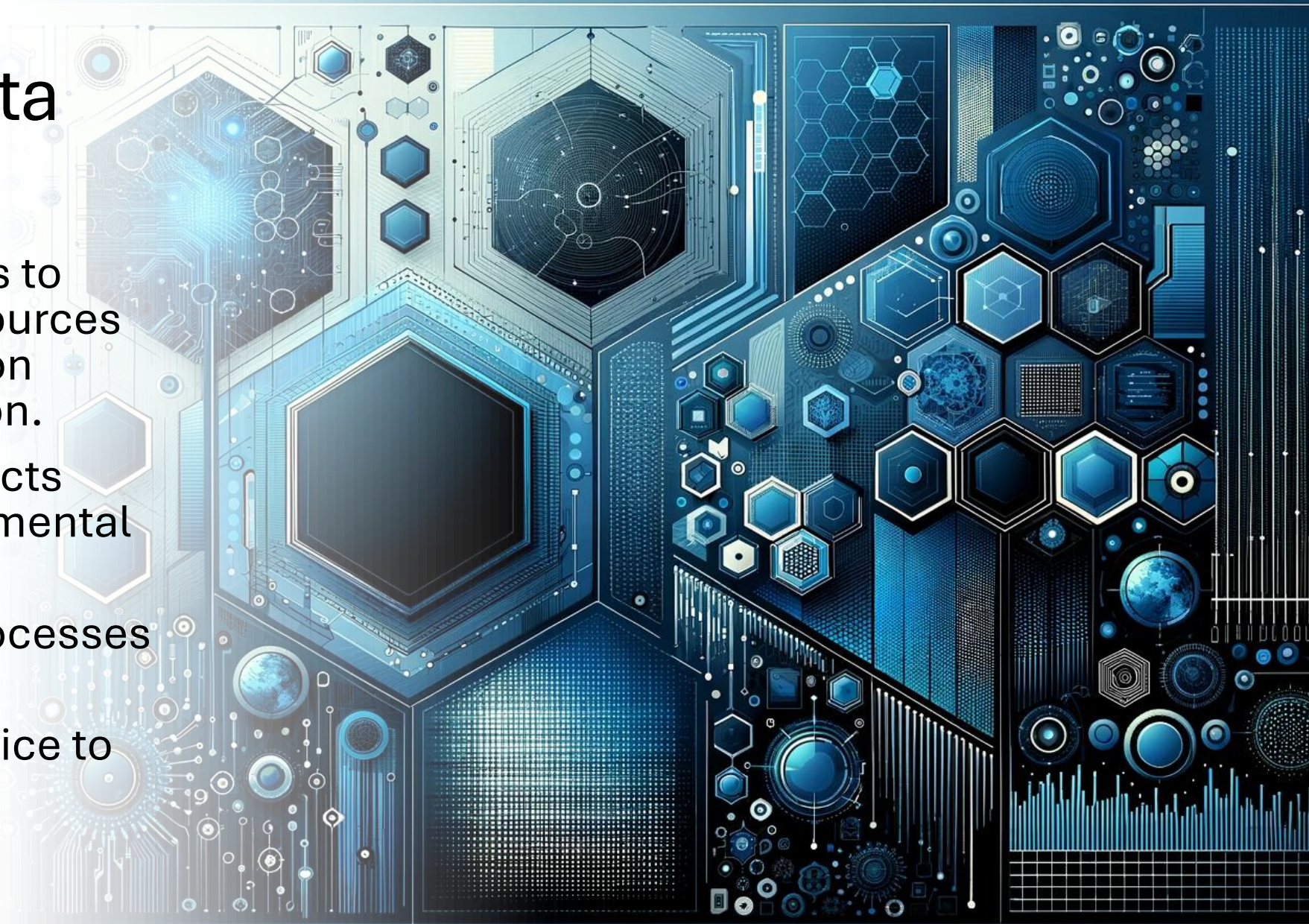
24/01/24

INEGI´s Data Science Transformation



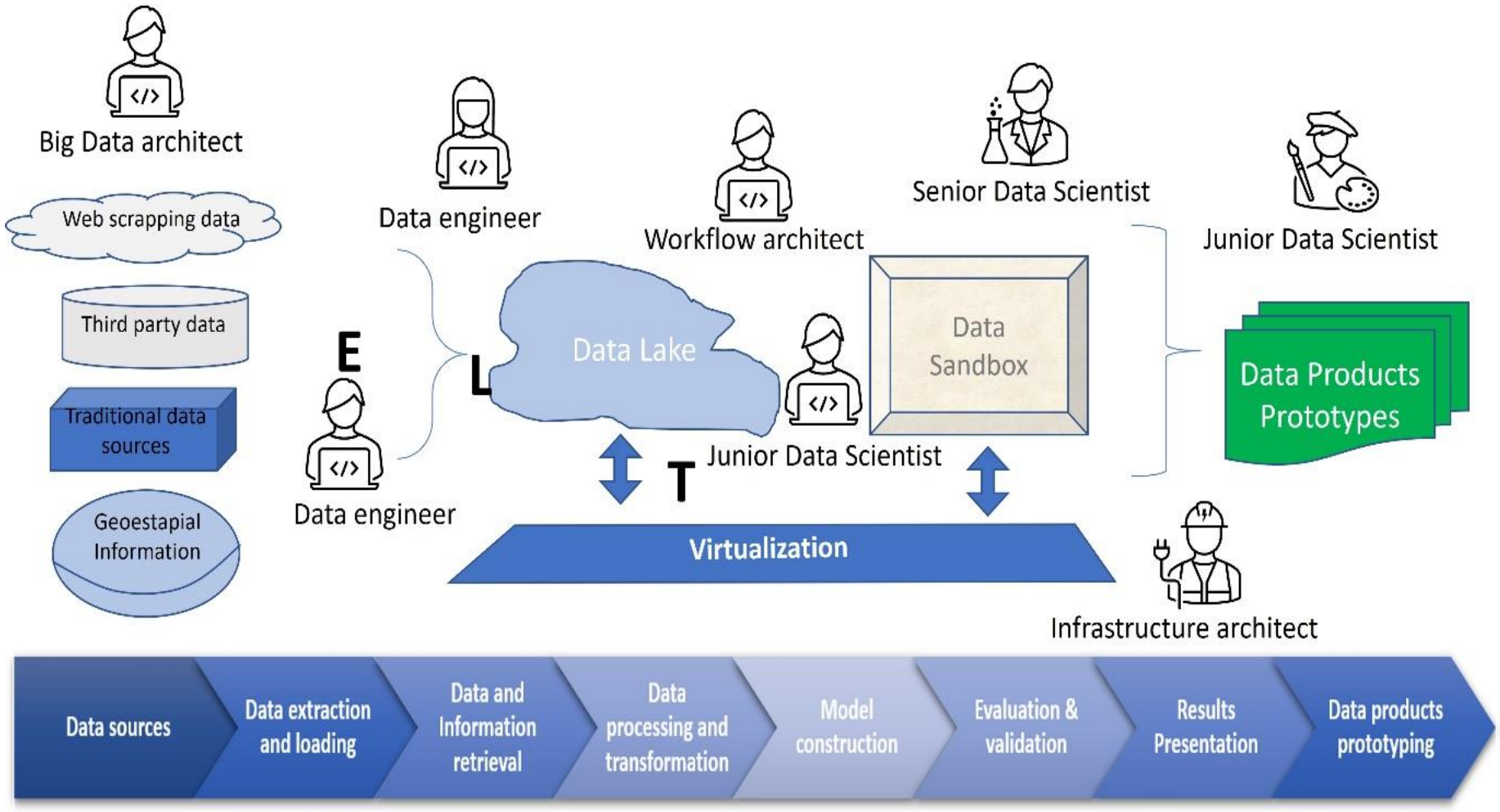
Purpose of Data Science Lab

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



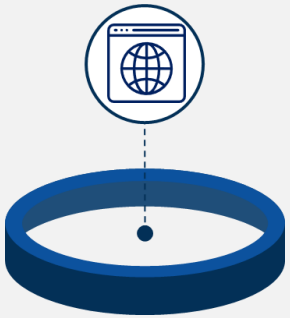
Creating a Multidisciplinary Team

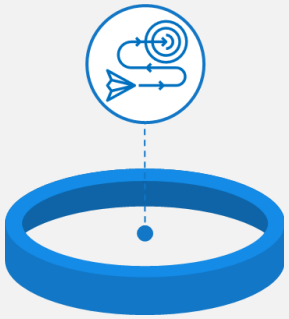
- Recruitment process
- Team structure
- Role of each discipline

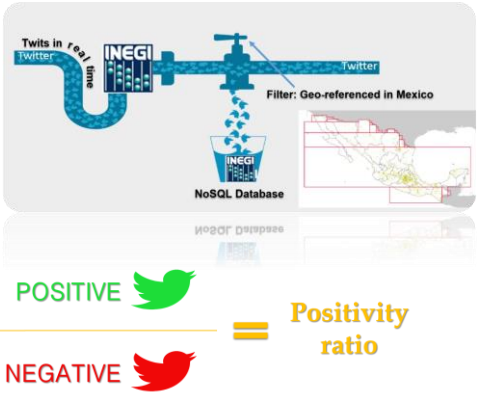


Data Lake

Data integration platform

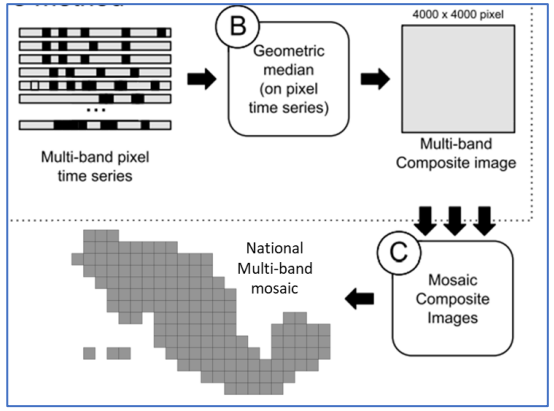






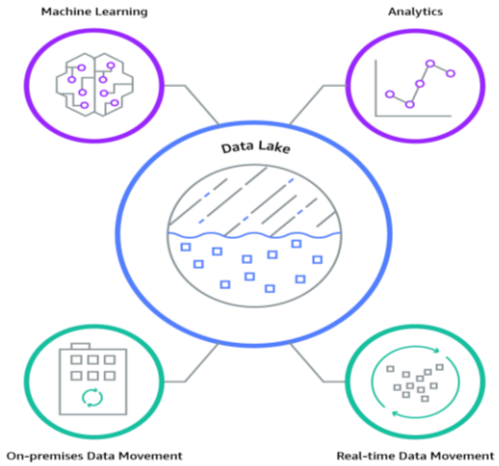
Social Networks and data from internet

Statistical experimental products.



Mexican Geospatial **Data Cube**

Big Earth data platform enabling time series (30 years) pixel level analysis



Statistical Information **Traditional Statistical Sources**

Digital **Map of Mexico**

Diverse datasets in a geospatially enabled visualizer

Cluster and Grid Sandbox-ITo (Areneros Desarrollo – 10 nodos), Procesamiento 80 cores en cpu´s, Memoria Ram 160 GB, Almacenamiento 15 TB,

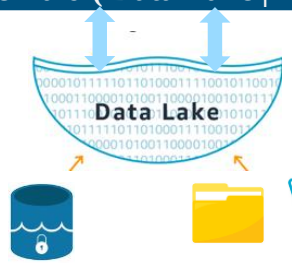
Cluster and Grid Sandbox (Areneros Preproducción Capacitación – 4 nodos) Procesamiento 160 cores en cpu´s, Memoria Ram 1.5 TB, Almacenamiento 16 TB

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

Grid Storage Raid (Data Lake | Lago de Datos)

On-premise

NAS (Network Attached Storage)
Almacenamiento 45 TB



Data [Small | Fast | Big]

SAN (Storage Area Network)
Almacenamiento 20 TB

Data Science Technological Infraestructure

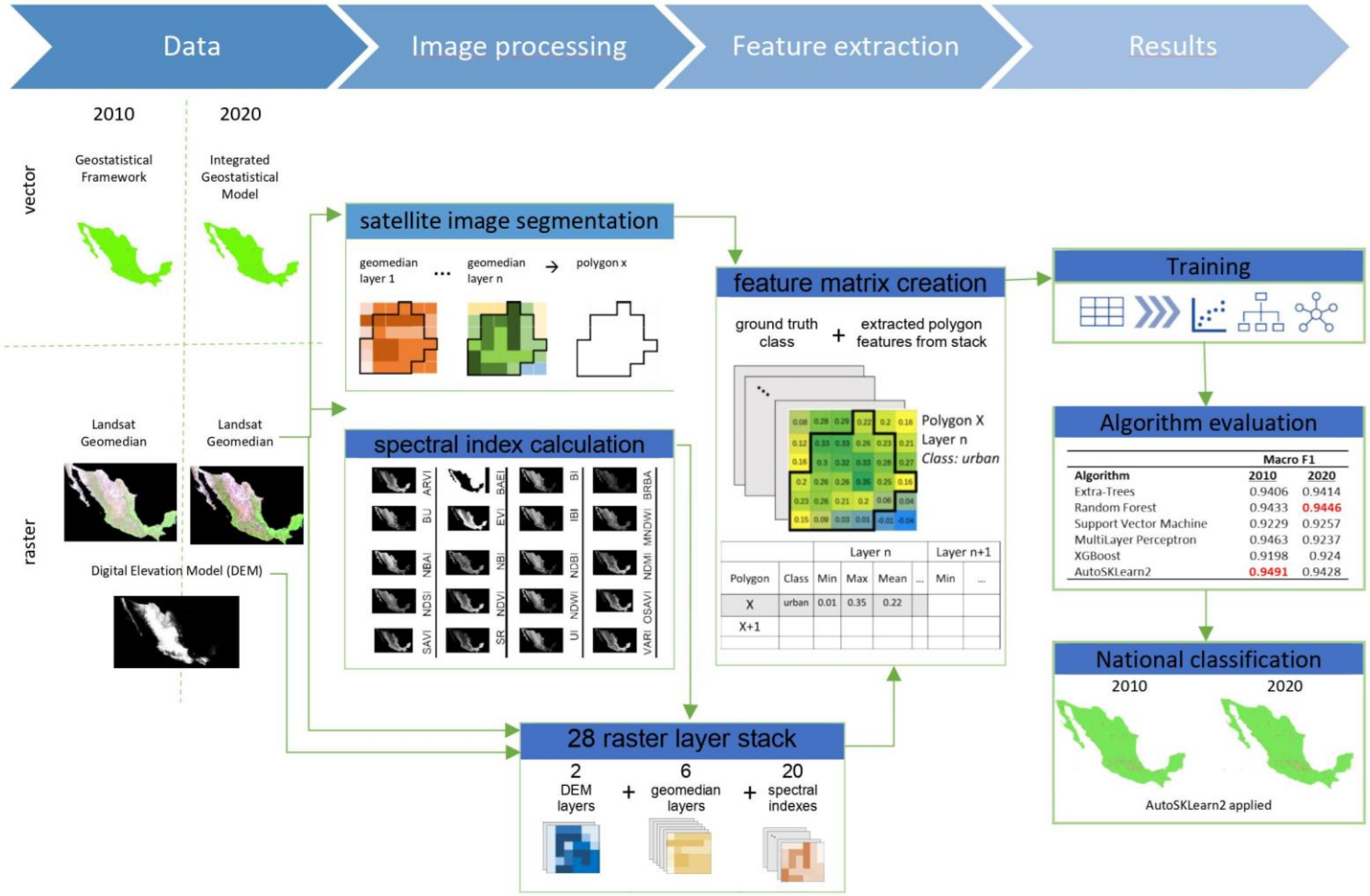
Georeferenced Data Sources	Data Harvesting	ELT (Geo-Datalake)	Data Representation	Machine Learning	Validation and Evaluation	Data Products
• Satellite Imaginary  • Census and Surveys Data 2020 censo de Población y Vivienda • Knowledge Bases  • Web Data	ϕ: *Geographic Mapping*	• Geostatistical Exploration	• Mathematical Modeling of Data.	Unsupervised Supervised	Clustering φ Regression R^2 Classification F_1	Basic Products • Geomedian Advanced Products • Land-use identification • Change footprint pattern discovery • Indicators estimation

Research projects

URBAN CLASSIFICATION

Using **EO, official information** (Geostatistical Framework) and **machine learning** to explore, monitor, and assess **urban growth** and inform SDG 11.

Platforms used
For data: Digital map & MGDC
For analysis: MGDC



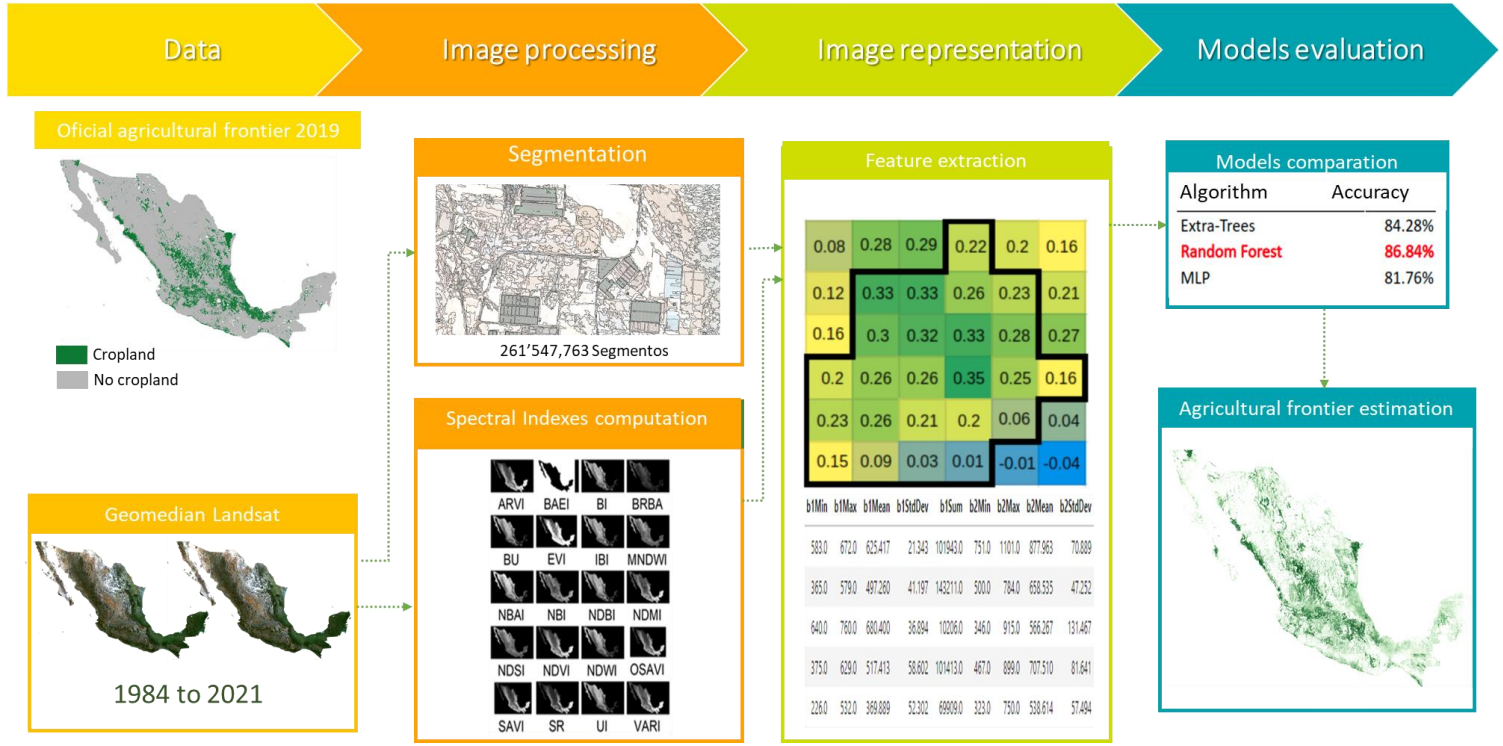
Research projects

AGRICULTURAL STATISTICS

Currently, the identification of agricultural areas is a **complex**, **slow** and **costly** process.

The goal of this research project is to build an annual time series on the **evolution of the agricultural areas** in the Mexican territory, using **satellite imagery** and **in-situ data** (previously generated).

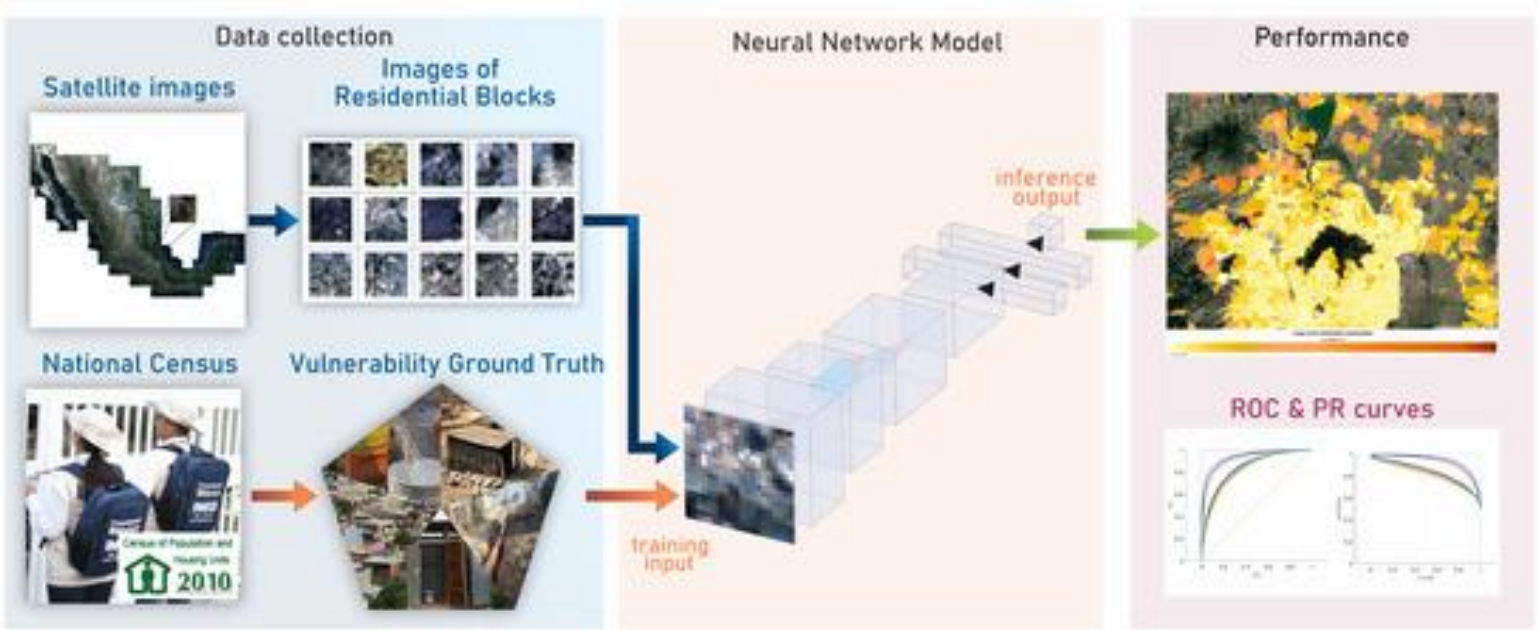
Platforms used
For data: Digital map & MGDC
For analysis: MGDC



Research projects

SLUM SEVERITY ANALYSIS (CENSUS DATA)

Using publicly available information, in the form of **census data** and **satellite images**, along with standard CNN architectures, may be employed as a steppingstone for the **countrywide characterization of vulnerability** at the residential block level.



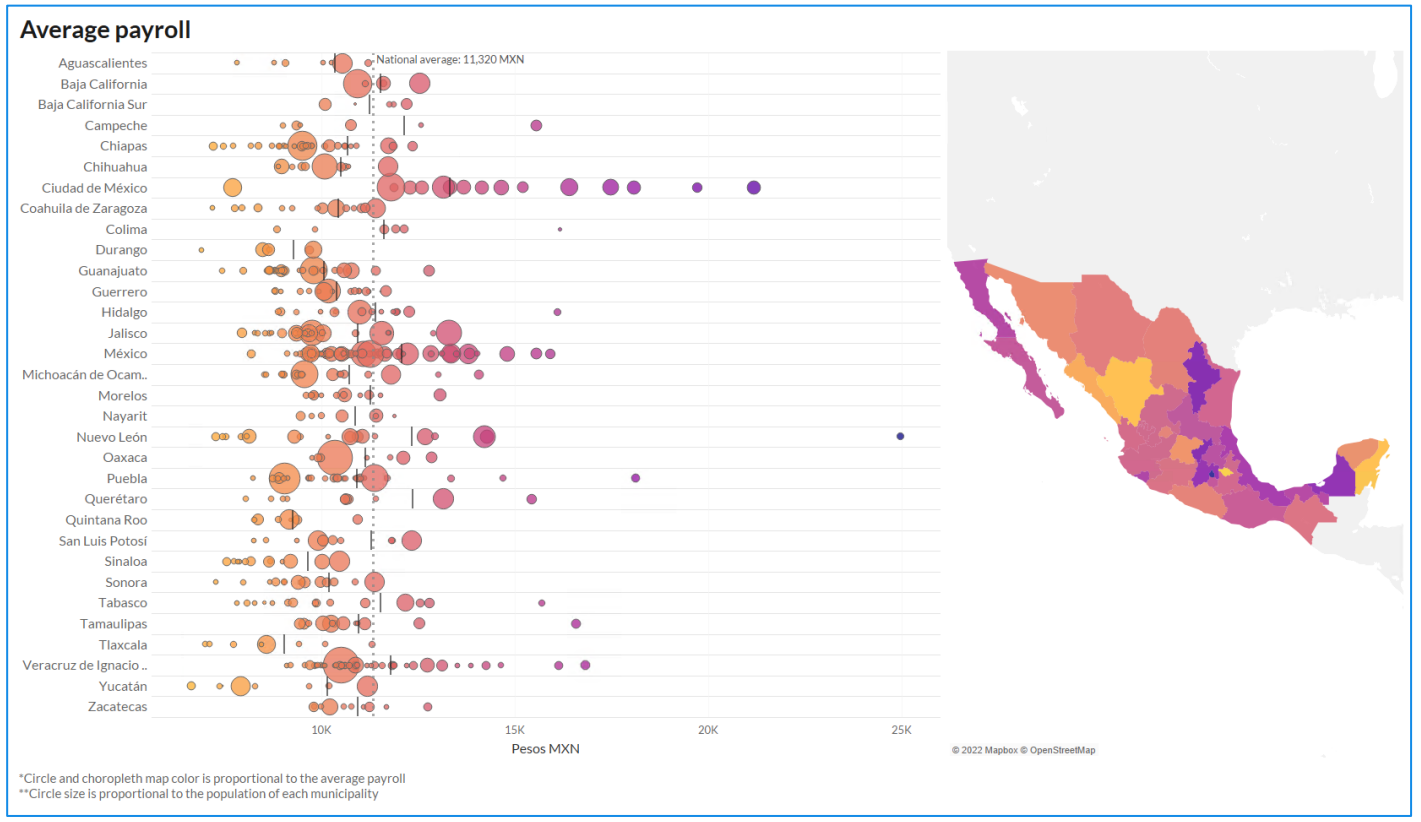
Platforms used
For data: Digital map & MGDC
For analysis: MGDC, Google Earth Engine (GEO credits program)

Research projects

Privately Held Data:
Banking Data.

Bilateral agreements were signed with three of the main financial institutions in Mexico to transfer statistical information **from banking transactions** generated for various microaggregations. Microaggregations are formed by combining **geographical levels** with **demographic characteristics** such as age and sex.

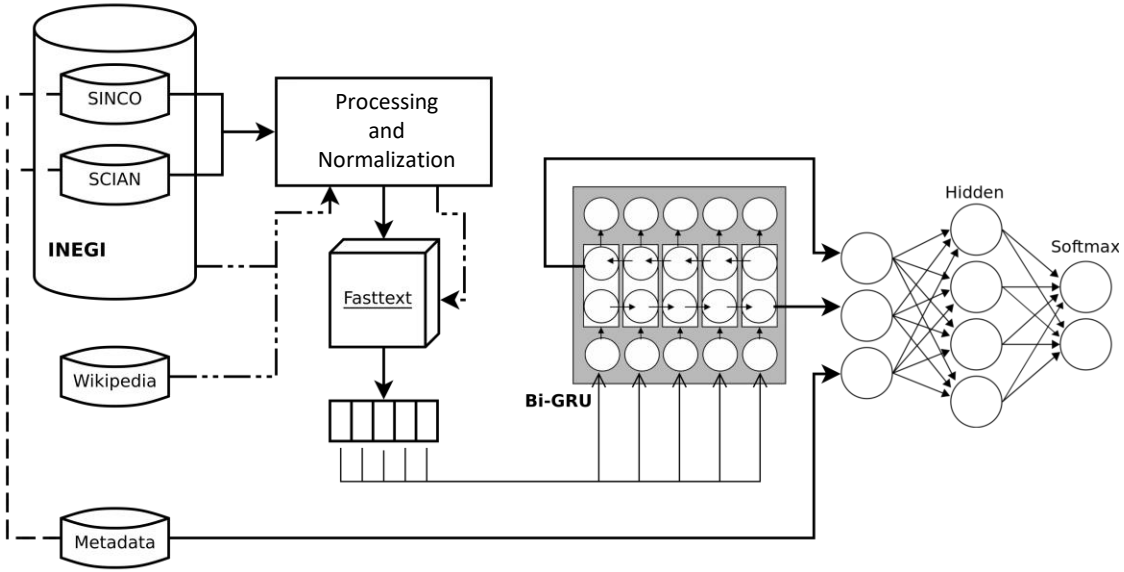
This will enable INEGI to **publish timely monthly information** based on different types of channels related to private consumption, such as cash **withdrawals**, **purchases made physically**, **and purchases made remotely**. Additionally, it will enable information generation based on **payrolls**, which will contribute to greater knowledge of the labor market in Mexico.



Research projects

Automatic Coding:
Economic activity and Occupation.

Before publication, several statistical products require the **coding of variables - a process of assigning an alphanumeric code** from within a thematic catalog. This is the case for Economic Activity and Occupation variables in two of our most important surveys: Employment and Labor Survey, and Income and Expenditure Survey.



To carry out the coding process, text responses provided by interviewees are considered. Currently, two strategies are employed: 1) deterministic computational rules and 2) **manual coding** performed by trained individuals. The latter **requires significant amounts of human resources and time**.

The objective of this project is to design, develop, and implement a **Deep Learning-based methodology** into the production process, **aiming to reduce the burden of manual coding**.

The obtained results show that it **is feasible to reduce the manual workload by 50% for the Economic Activity variable and by 35% for the Occupation** variable while maintaining a similar level of **high quality** to the current processes.

INEGI

Research projects

Item Reclasification:

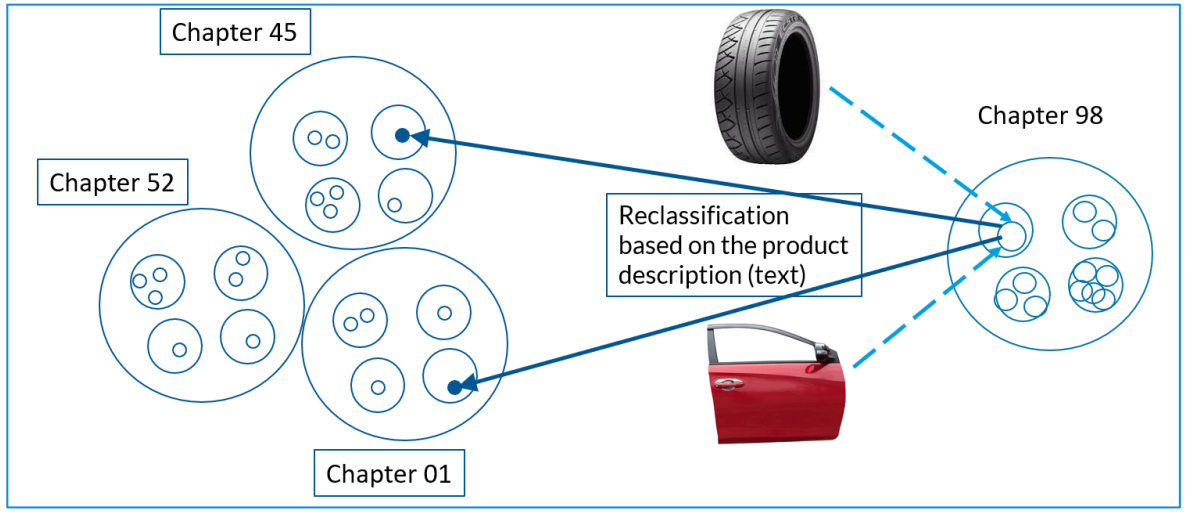
Import and Export General Law, Chapter 98.

Mexican international trade classification system is based on the international Harmonized Commodity Description and Coding System. However, like other countries, we have some **generic chapters** in which we classify items for specific purposes such as **customs duties. Chapter 98 is one of those**. These generic chapters lead to **asymmetries in international trade statistics**, especially with our main trading partners.

This project aims to **reclassify items** originally classified within Chapter 98. These reclassification is **based on the physical characteristics of each item**.

To achieve this, we developed a strategy that takes the textual description of the product provided by customs agents. With over 300 million textual descriptions, **we developed and parameterized a Natural Language Processing and Deep Learning model for product reclassification**.

As a result, **we can generate a new code for 95% of the records originally classified under Chapter 98**, which will help improve international trade statistics.



Final Remarks

The adoption of data science methods could signify a **paradigm shift** in the processes of information production.

Collaboration with IT areas is key to making Data Science and Big Data projects viable.

Data Science presents an opportunity for the **modernization of the NSOs.**

Without SCIENCE there is no Data Science

For production implementation, it is essential to **demonstrate value and ensure sustainability**.

INEGI