The Role of Artificial Intelligence in the Production of Official Statistics: Now and in the Future

NOTE: This session was cancelled due to unforeseen circumstances.

Instructions: Click on the link to access each author's presentation.

Organiser: Linda J. Young

Chair: Alejandro Ruiz

Discussant: Linda J. Young

Participants:

Elio Villaseñor

Kenneth Haas

IAOS-ISI 2024
MEXICO CONFERENCE

INEGI's Data Science & AI Transformation

DSNL, 2nd Sprint

24/01/24

INEGI

International Statistical Institute

IAOS
IMPROVING OFFICIAL STATISTICS

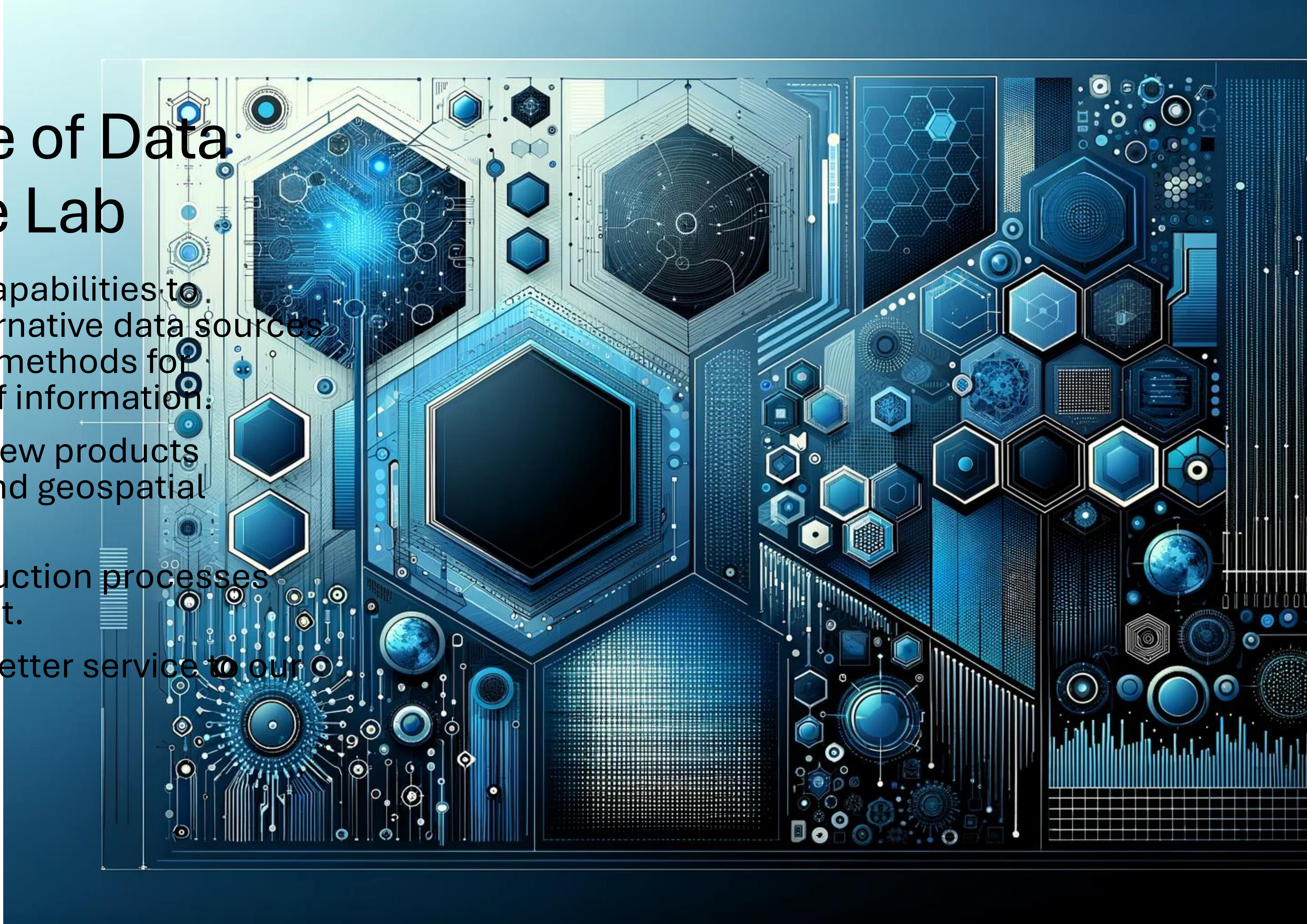
INEGI´s Data Science Transformation

Events

2014	2016	2018	2019	2020-10	2020-11	2021	2022	2023
International Seminar on Big Data for Official Statistics by INEGI and INFOTEC	INEGI launches Twitter's Mood in Mexico, the first experimental statistic	Deployment of the Mexican Open Data Cube	Establishment of the Data Science Lab	Publication of the Nowcasting for Economic Activity, integrating alternative data sources	Release of the first Experimental Geospatial product (Landsat Geomedian)	First research product of remote sensing and deep learning for socio-economic data estimation	Experimental Data Lake implementation and use cases development	Data product prototype implementation in production environments

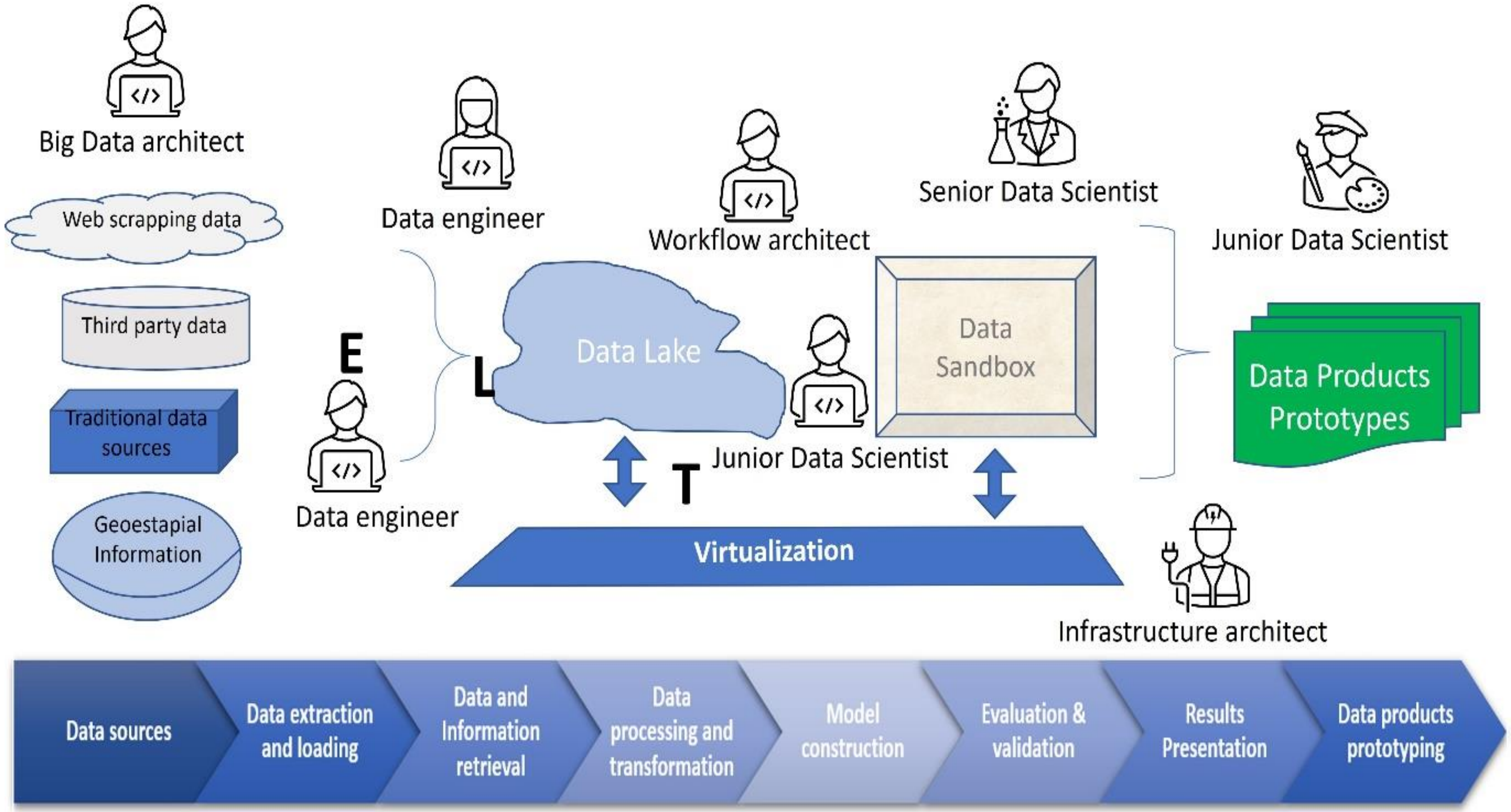
Purpose of Data Science Lab

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



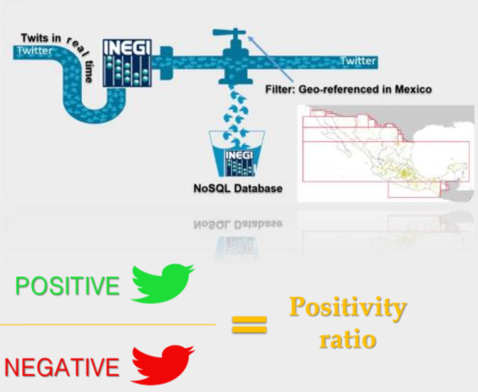
Creating a Multidisciplinary Team

- Recruitment process
- Team structure
- Role of each discipline



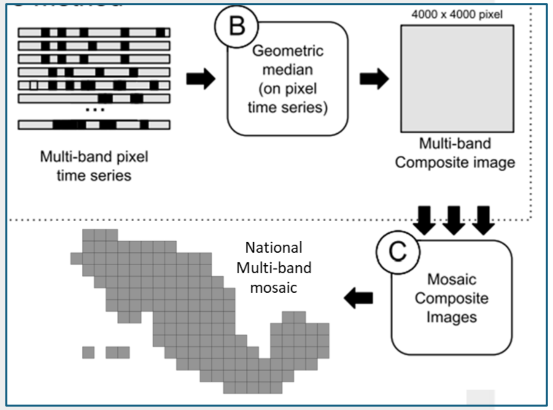
Data Lake

Data integration platform



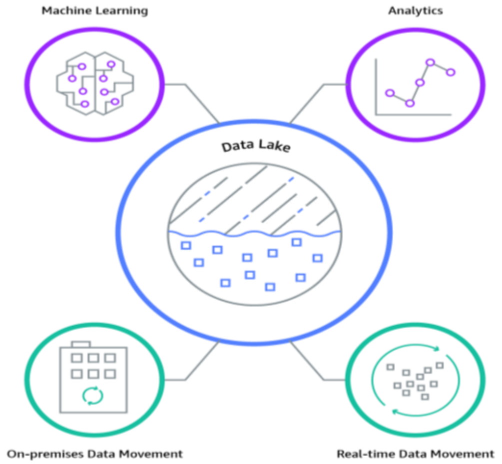
Social Networks and data from internet

Statistical experimental products.



Mexican Geospatial **Data Cube**

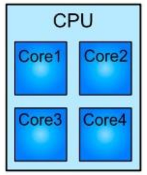
Big Earth data platform enabling time series (30 years) pixel level analysis



Statistical Information **Traditional Statistical Sources**

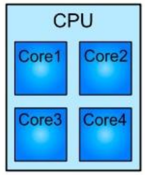
Digital Map of Mexico

Diverse datasets in a geospatially enabled visualizer

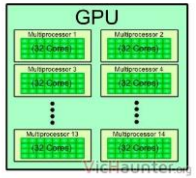


Cluster and Grid Sandbox-ITo (Areneros Desarrollo – 10 nodos), Procesamiento 80 cores en cpu´s, Memoria Ram 160 GB, Almacenamiento 15 TB,

SWARM



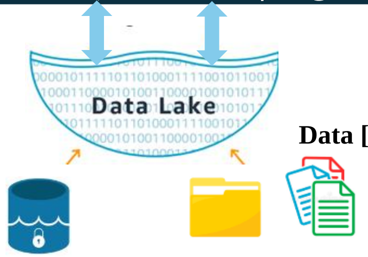
Cluster and Grid Sandbox (Areneros Preproducción Capacitación – 4 nodos) Procesamiento 160 cores en cpu´s, Memoria Ram 1.5 TB, Almacenamiento 16 TB



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)

NAS (Network Attached Storage)
Almacenamiento 45 TB

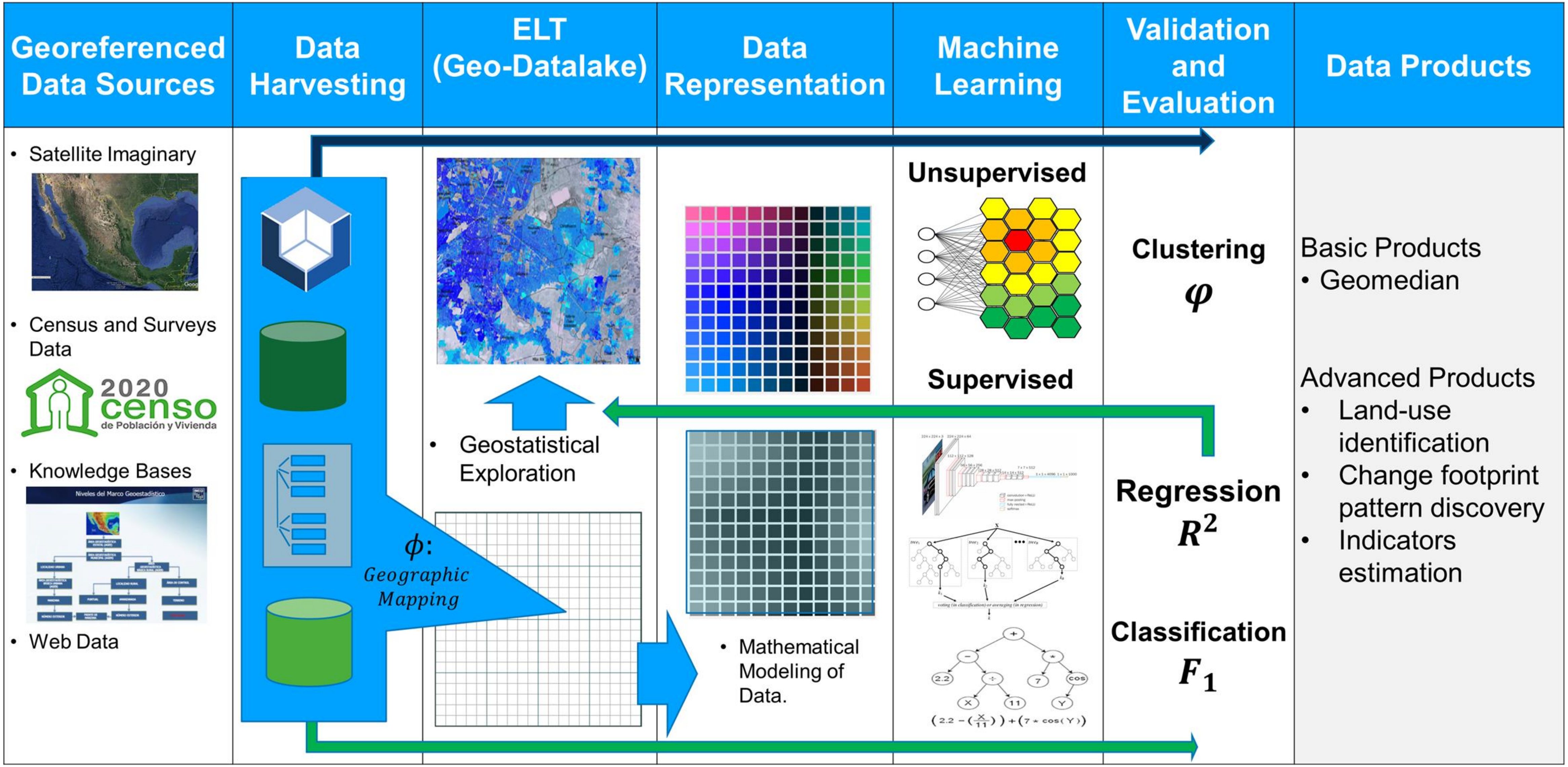


Data [Small | Fast | Big]

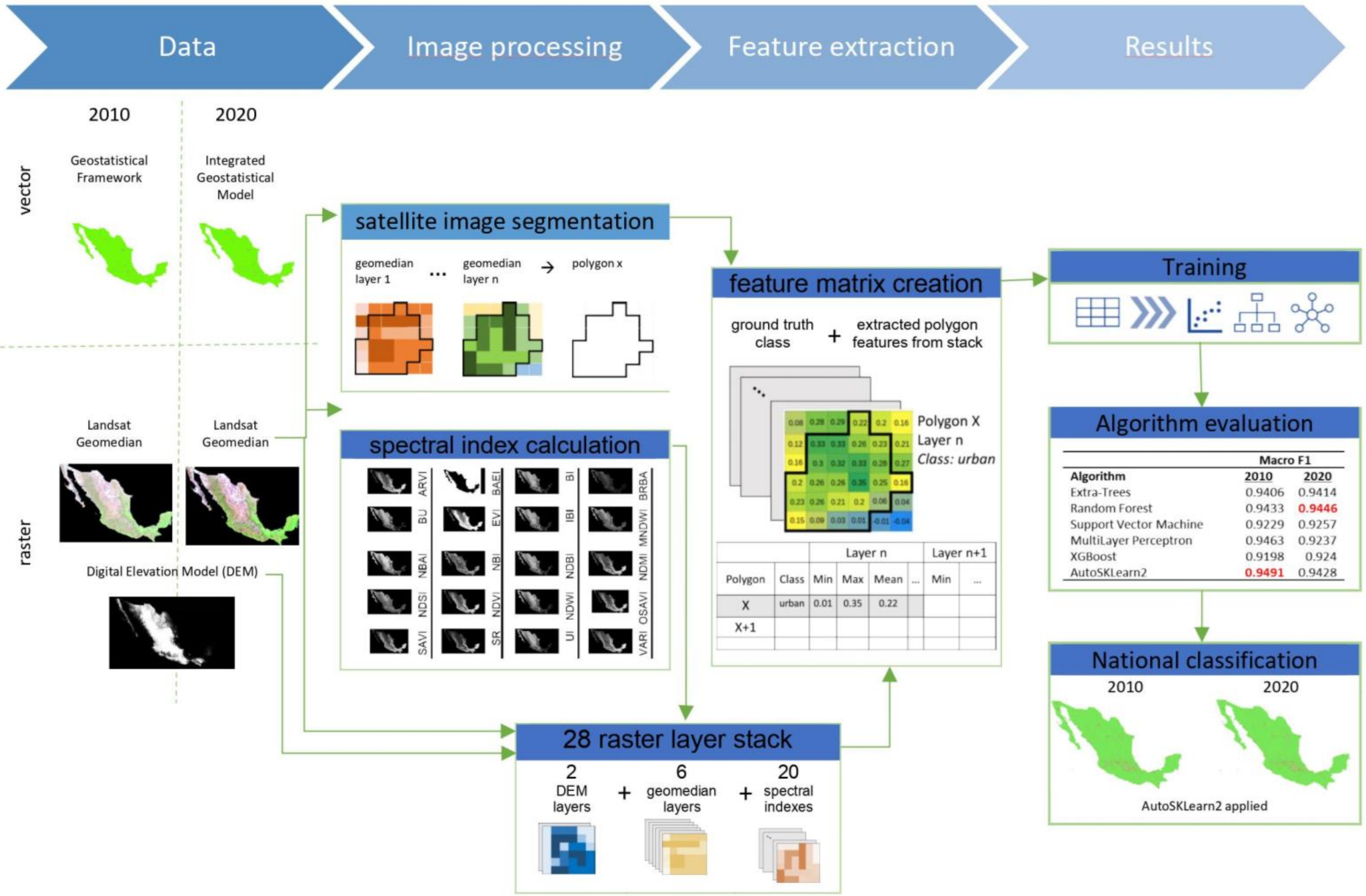
SAN (Storage Area Network)
Almacenamiento 20 TB

On-premise

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
Niveles del Marco Geoestadístico
• 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

11 SUSTAINABLE CITIES AND COMMUNITIES

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

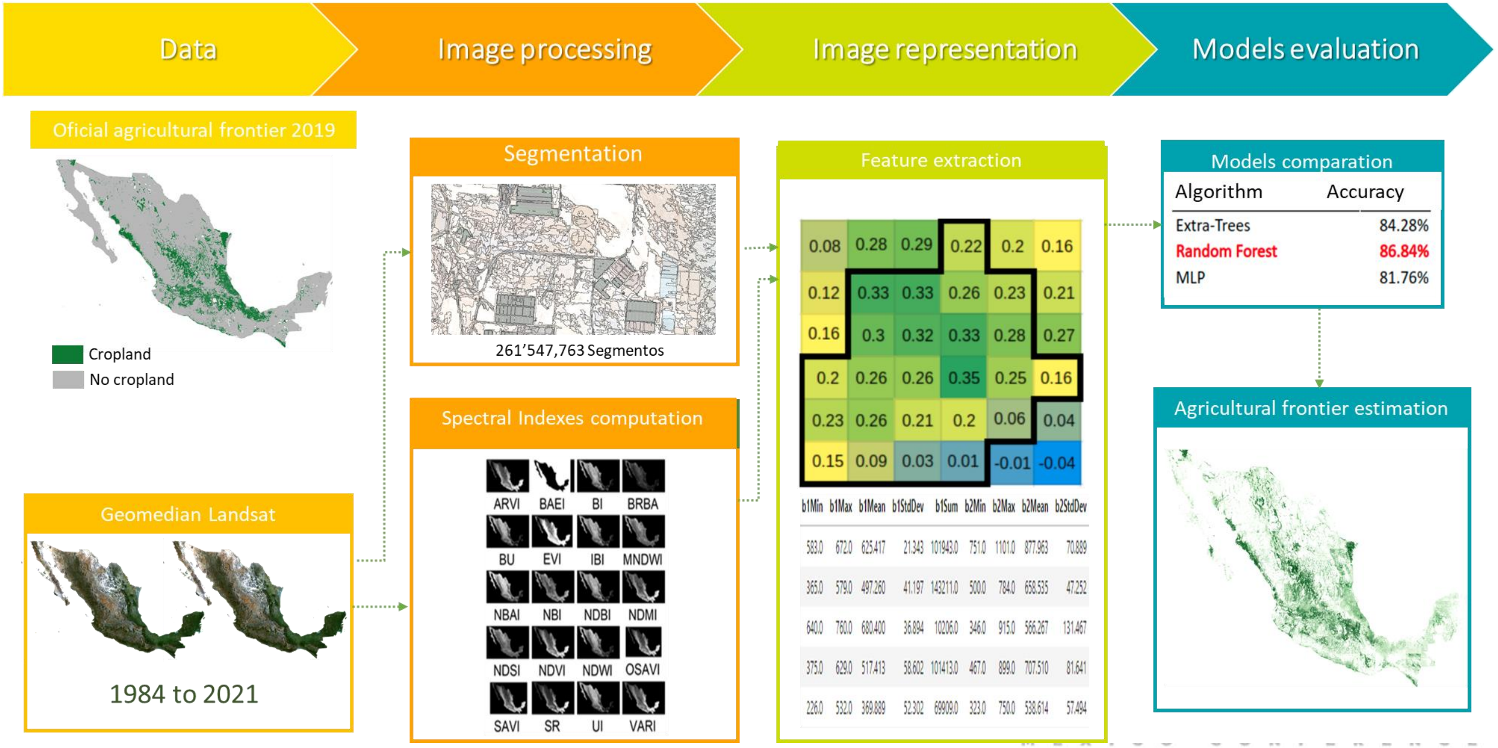
AGRICULTURAL STATISTICS

2 ZERO HUNGER

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



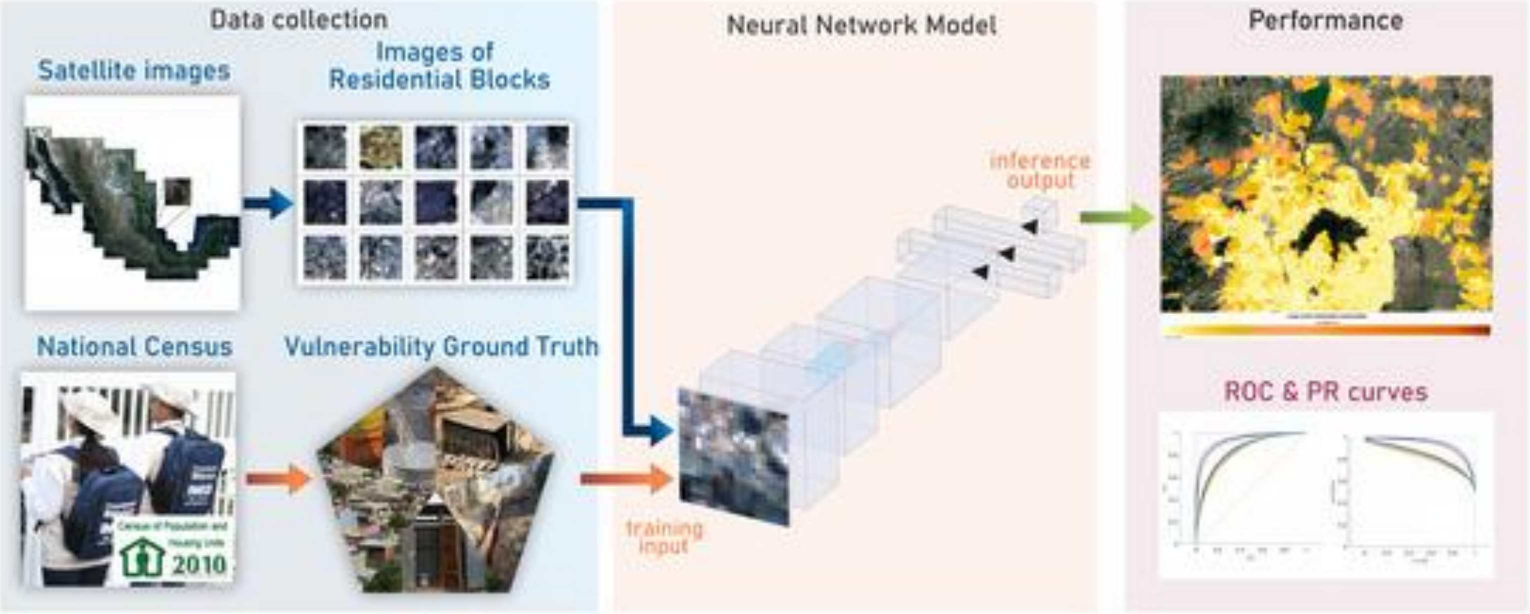
INEGI

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.

11 SUSTAINABLE CITIES AND COMMUNITIES



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

IAOS-ISI 2024 MEXICO CONFERENCE

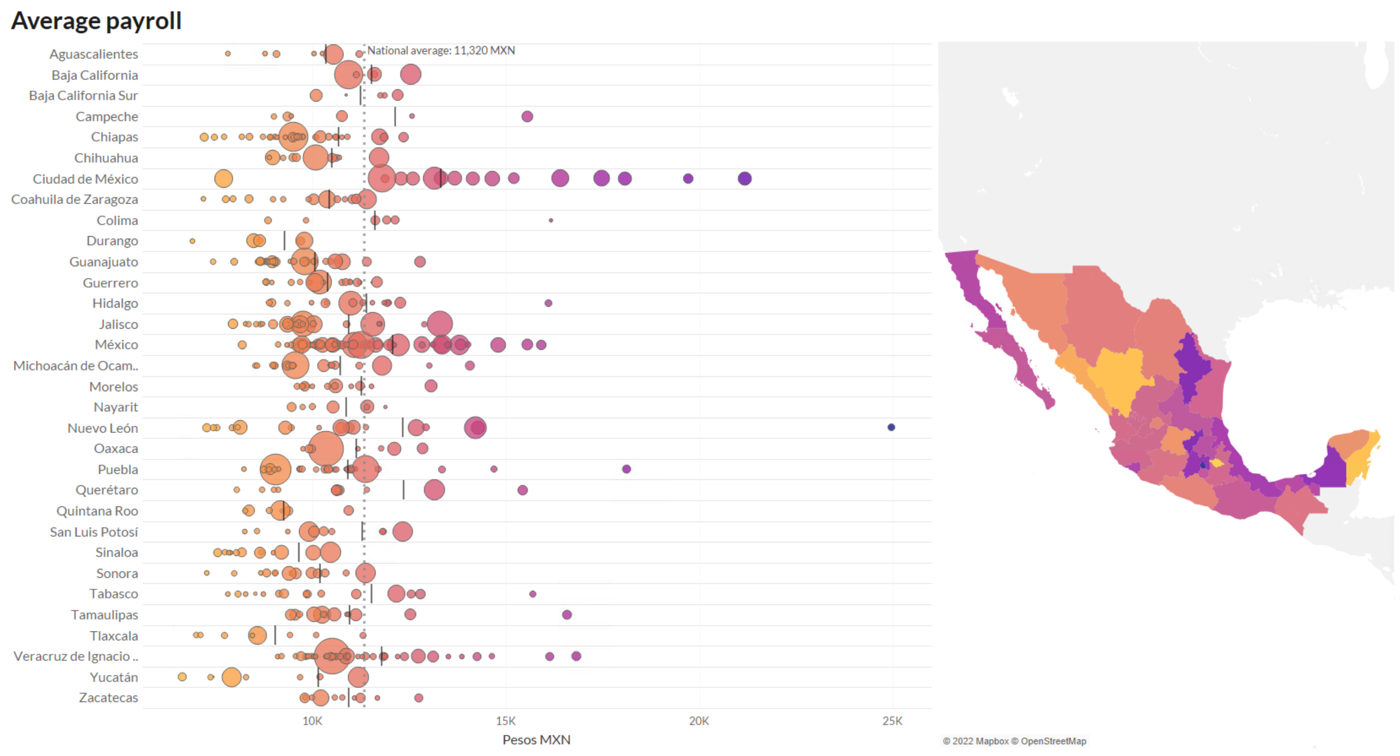
Research projects

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.

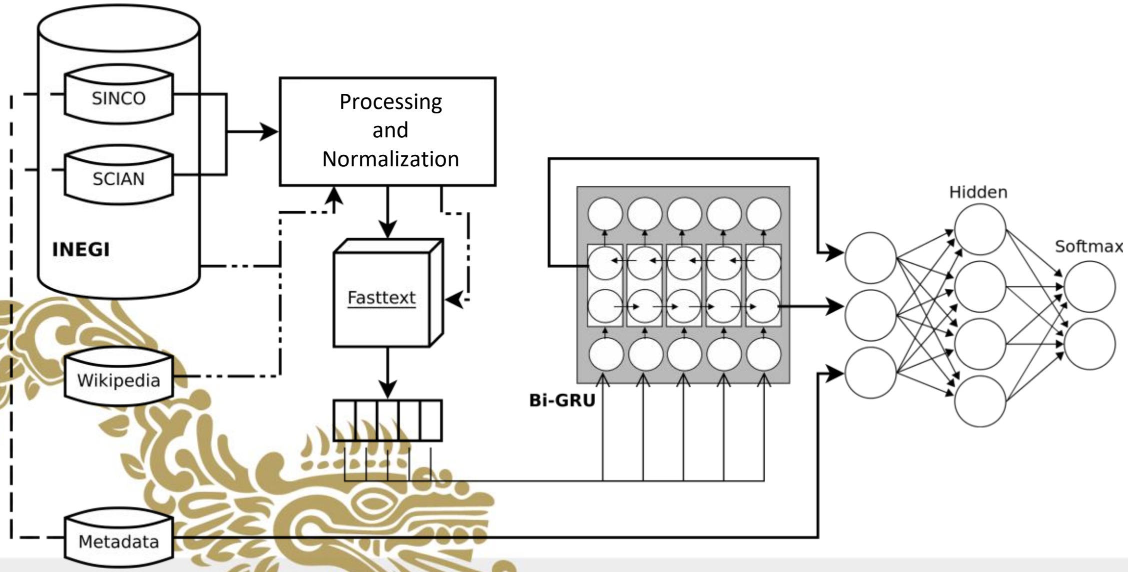


*Circle and choropleth map color is proportional to the average payroll
**Circle size is proportional to the population of each municipality

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.

Research projects

Item Reclasification:

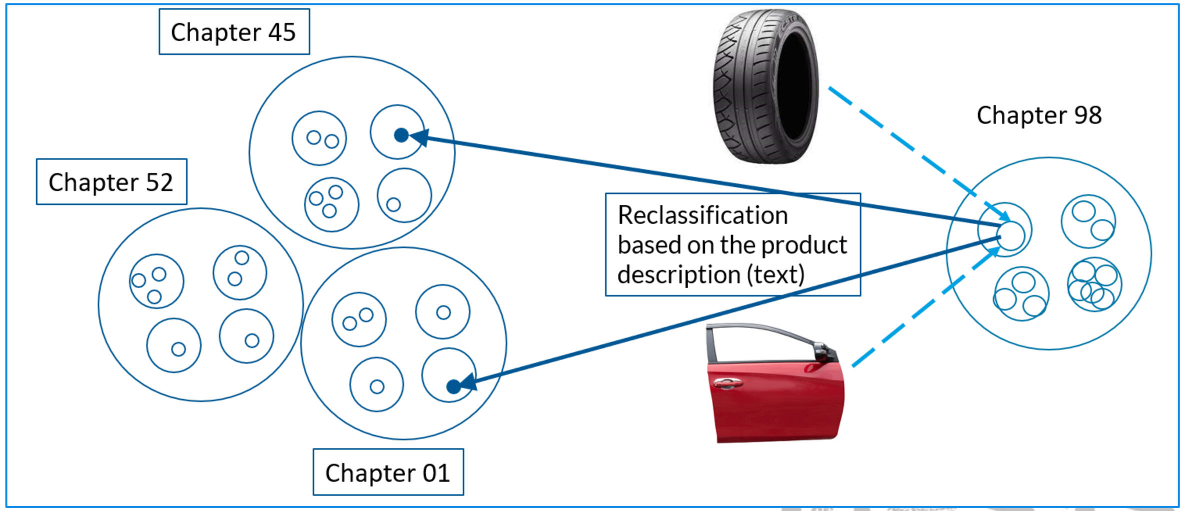
Import and Export General Law, Chapter 98.

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.



Next Future

Generation of New Information Products

- **Utilizing Alternative Data Sources**: AI can leverage diverse and non-traditional data sources such as satellite imagery, social media feeds, and Internet of Things (IoT) sensors to generate new types of statistical data and insights. This expands the scope and depth of information products available.
- **Enhanced Analytical Capabilities**: Through advanced algorithms and machine learning, AI can analyze complex datasets to produce nuanced geospatial analyses, environmental studies, and other specialized information products that were previously unattainable.

Next Future

Increased Efficiency in Production Processes

- **Automation of Data Collection and Processing**: AI automates labor-intensive and time-consuming tasks involved in data collection and processing, such as data cleaning, sorting, and analysis, significantly speeding up these processes.
- **Improving Accuracy and Reducing Errors**: AI's ability to learn from data and continuously improve its algorithms minimizes errors in data handling, leading to more accurate outputs.
- **Cost Reduction**: By automating routine tasks, AI helps reduce operational costs associated with human resources and allows staff to focus on more strategic, value-added activities.

Next Future

Improved Access to Information for Users

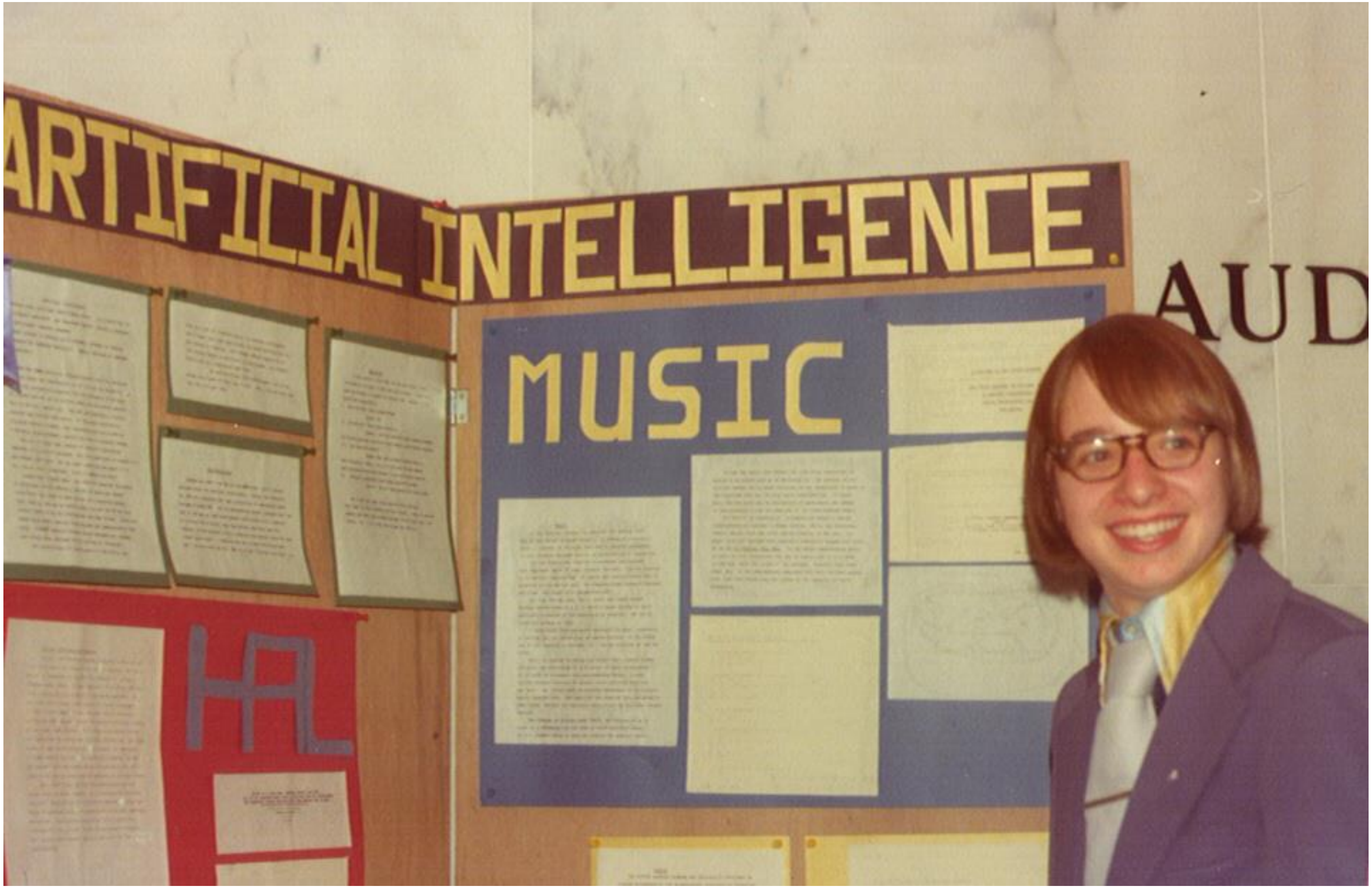
1. **Deployment of Large Language Models (LLMs)**: NSOs can use LLMs to interpret and respond to user inquiries, making data more accessible and understandable to a wider audience. This includes generating automatic responses, summaries, and even detailed reports based on user queries.
2. **Personalized Data Interaction**: AI can tailor the presentation and delivery of data to meet the specific needs of different user groups, ranging from policymakers to researchers and the general public.
3. **Enhanced Data Dissemination Platforms**: AI technologies can be used to develop more interactive and user-friendly data platforms that provide real-time access to updated statistical information, enhancing transparency and user engagement.

Thank you

Developing AI-Ready Public Data

Kenneth Haase

U.S. Census Bureau


ARTIFICIAL INTELLIGENCE.
MUSIC
HAL
AUD

Plan

- **AI & Census Data: Opportunity & Risk (and an RFI)**
- **3-Slide History of AI**
- **Hybrid Solutions**

Opportunity

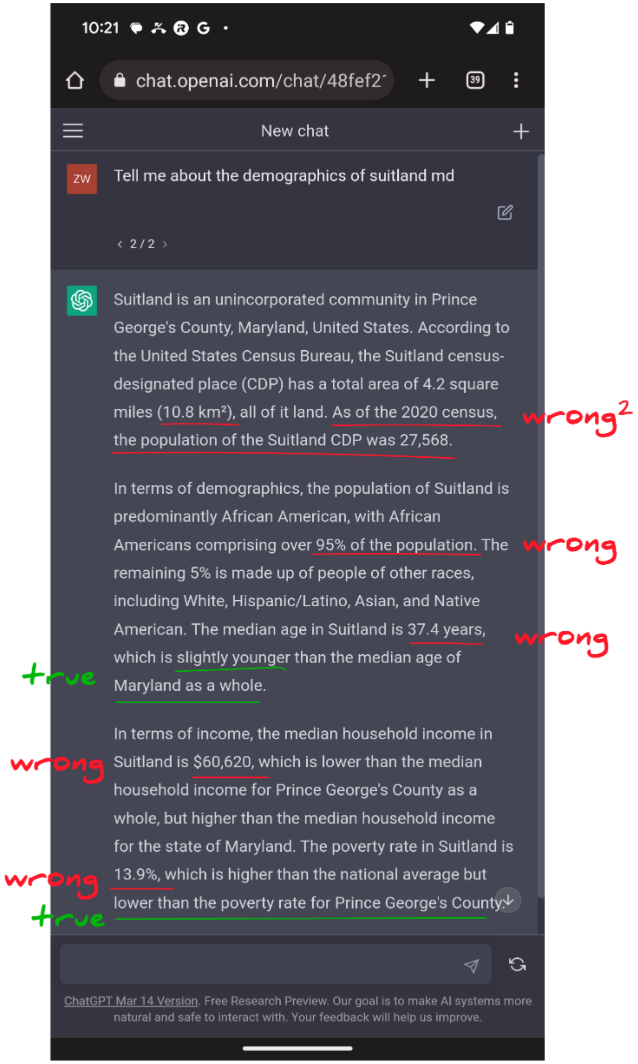
AI systems can enable dramatic democratization of access to data where intelligent intermediaries help non-expert users understand, analyze, and apply public data to local challenges

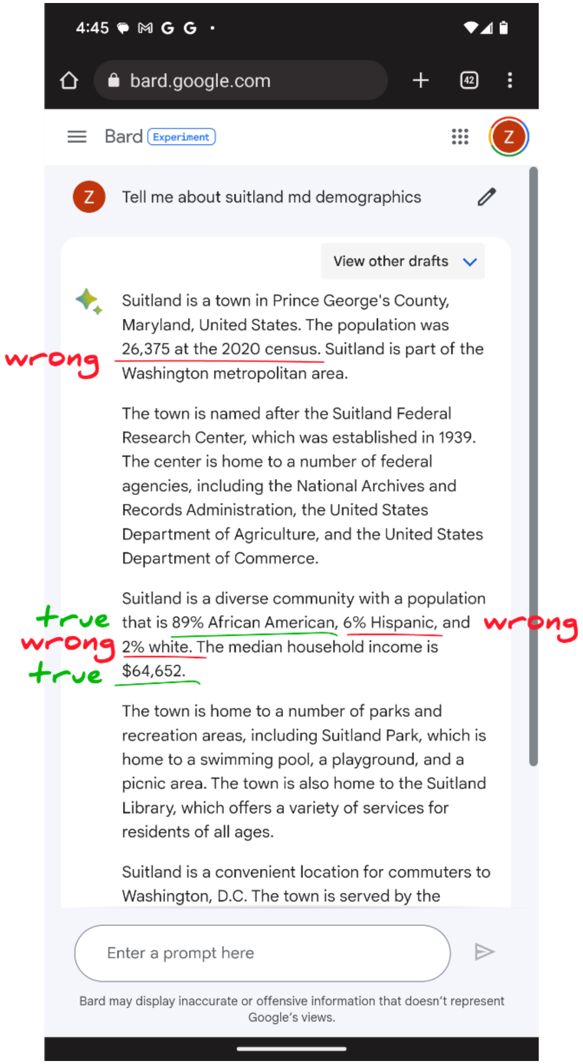
Risk

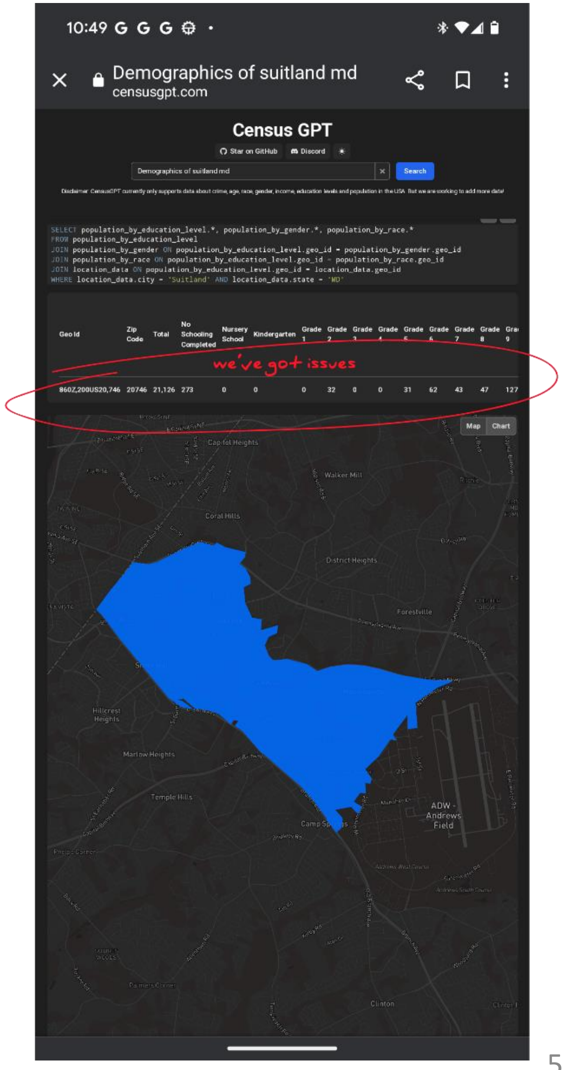
AI systems lead to misunderstanding and misuse of data with the appearance of competence and authority

Sometimes wrong, never in doubt

ChatGPT, Bard, and CensusGPT (2023)







Commerce Working Group & AI-ready Data RFI

Commerce Data Governance Board created the **AI and Open Government Data Assets Working Group** to explore improving DOC's data publishing standards and making our data assets AI-ready (i.e., machine-understandable)

Launched Request for Information (RFI) on April 17th, 2024 to inform DOC's journey to AI-Readiness.

RFI will close on July 17th, 2024.

Home » News » Blog

Request for Information: AI-Ready Open Government Data Assets

April 17, 2024

Artificial Intelligence

A 3-Slide History of AI

1961: Slagle’s SAINT

- SAINT (Symbolic Automatic INTegrator) was able to solve freshman calculus problems
- Based on this success and others like it, many AI researchers believed that human-level intelligence wouldn’t be far away (they were wrong)
- Calculus was ”easy” for computers because it was
  - teachable: humans could learn it formally
  - modular: small pieces, easily assembled
  - polysemic: many ways to reach identical results

1960-1990: Extended Infancy

- Many things that are hard for people are easy for computers
  - ➢ calculus, spectral analysis, large numerical problems, etc.
- Many things that are easy for people are hard for computers
  - ➢ recognizing speech, understanding stories, telling cats from dogs
- Smart researchers struggled for years to get computers to do what humans learn to do by the time they were five years old
- They learned a lot about cognition and reasoning but some of the simplest human behaviors confounded their machines

1990-2020: Accelerating Training

Meanwhile:

- computer power grew exponentially (Moore's Law)
- digital data also grew enormously as humanity moved "online"

This enabled **VERY** large numerical models to be created (trained) based on massive amounts of data

The easy-for-humans hard-for-machines problems began to fall before these new models whose internal complexities (millions/billions of parameters) matched or surpassed those problems' complexities.

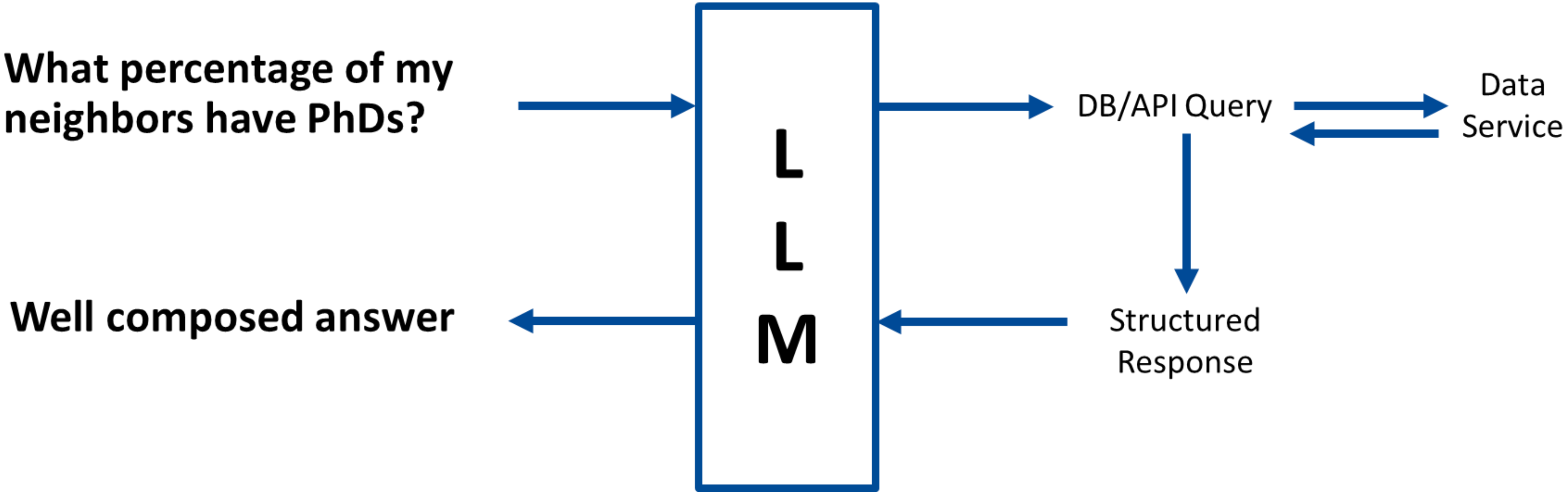
We Are Here

Generative AI Systems – powered by very large numeric models – are fluent in answer questions, generating explanations, and performing a range of creative tasks

BUT, they struggle with hallucinations, biases, factual errors, and fabrications

Which is not surprising, because the training of LLMs is about *generalization*, *fluency*, and *flexibility*, not **facts** of **precision**.

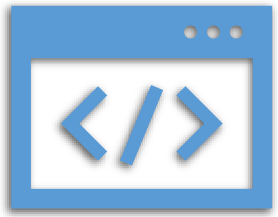
Early Solution: Hybrid Architectures (RAG)



The Challenge

- To both interpret questions and explain answers, public resources used by AI systems must expose the structure and meaning of the underlying data.
- The categories and connections of variables and results must be sufficient to the detail and complexity of human languages and purposes
- This kind of machine-*understandable* data is needed to guide the use, combination, and explanation of data in “responsible” ways

AI-readiness requires a shift in how government thinks about data publishing standards



From machine readable

- Can be automatically processed by a computer
- Common formats (.csv, JSON, HTML)

To machine understandable

- Structured data
- Standardized schemas with enriched metadata
- Semantics expressed in a knowledge graph

Making It Work

Co-Evolution

- AI systems must evolve to engage with the rules and reasoning which underly the data they interpret
- Our publishing processes and standards must evolve to expose the knowledge and reasoning behind the data and statistics we deliver.

What We Will Need

- Knowledge graphs for variable level metadata, linking human terms to data elements (linguistic meaning and variation) and data elements to one another (dependency, correlation, covariance, etc).
- Embracing standardized ontologies such as schema.org or NIEM but extending those ontologies in principled ways
- Adopting data formats which allow for rich metadata as well as generating metadata “sidecars” for more traditional formats such as CSV or SAS

How we will get it

- Gathering internal and external written documentation for data products and artifacts and mining them for terminology to use in annotation, harmonization, and linking
- Releasing some documentation as raw text for the training of AI models
- Harmonizing and linking internal ontologies and vocabularies using knowledge graphs grounded in those standardized ontologies
- Analyze the data itself for regularities which reflect implicit meaning

Discussion: The Role of Artificial Intelligence in the Production of Official Statistics—Now and in the Future

Linda J. Young
USDA National Agricultural Statistics Service (NASS)

May 15, 2024

Thank You, Speakers

- Elio Atenógenes Villaseñor Garcia
  - INEGI
- Kenneth Haase
  - U.S. Census Bureau

The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

Traditional AI

Traditional AI Has Been Here for a Long Time

- Machine Learning and Text-based Analyses
- Models trained to recognize patterns in existing data
- Can be used to obtain estimates or forecasts
- Can make decisions based on predefined rules
- Focused on AUTOMATION

Widely Used Work and Personal Lives

- Robots fill customer orders or complete tasks on the assembly line
- Track inventory
- Grab-and-go stores
- Facial recognition
- Weather forecasts
- Traffic signals
- Google searches
- Getting directions to a destination
- Traveling by air
- Home security systems

Traditional AI: Long Used for Official Statistics

- Web scraping
  - Identify farms not currently on the NASS list of all U.S. farms
  - Collect data for use in a report
- Predict the probability that a producer will respond to a survey
- Forecast the yield of major crops: corn, soybeans, cotton, wheat
- Assess the value of publications by determining how often they are used in newspapers, magazines, journal articles, etc.
- Identify most common topics tweeted about NASS
- Identify the comments that producers provide that may improve customer service

Traditional AI: Increasing Need and Opportunities for Applications in Official Statistics

- Challenges with survey-based production of official statistics
  - List frame coverage decreasing
  - Increasing demand for statistics on finer temporal and spatial scales
  - Response burden increasing
  - Response rates declining
- Alternative (non-survey) data increasingly available
  - Challenges with data integration
- Increased computer power
  - Cloud computing

Traditional AI: Newer Applications in Official Statistics

- Monitor and assess urban growth
- Identify agricultural areas
- Characterize vulnerability at the residential block level
- Publishing monthly statistics of private consumption
- Assigning codes based on responses
- Record linkage

Generative AI

Characteristics of Generative AI Use

- Focused on ALMAGMATION
- Models trained on massive amounts of data
- Learning patterns and generating content
- Creation of new data
- Creative capabilities

Generative Pre-trained Transformer (GPT)

- Rely upon high performance computing environment
- Developed and refined over several decades
  - New/improved models
    - Large Language Models
    - Convolutional Neural Networks (CNNs)
    - Generative Adversarial Networks (GANs)
  - Numerous applications
- GPTs
  - November, 2022: OpenAI ChatGPT
  - February, 2023: Microsoft Copilot
  - December 6, 2023: Google Gemini

Generative AI in Our Daily Lives

- Optimize calendar, suggest meeting times
- Draft or sort e-mails
- Summarize articles, books or papers, highlighting key points or generating questions to consider
- Take dictated notes, suggest improvements, generate summaries or brainstorm ideas
- Elicit actionable ideas
- Customer service chatbots
- Drug discovery, medical imaging, and genomic analysis

Generative AI: Opportunities for Applications in Official Statistics

- Development of code
  - Describe problem and specify programming language
  - Convert old code to another programming language
  - Continuously improve algorithms
- Customizing survey questionnaires
- Answer questions about survey questionnaires
- Provide easier access to data
  - Chatbots can be created making access easier
  - Interactive and user-friendly data platforms

Generative AI: Challenges for Applications in Official Statistics

- Creating machine understandable data
- Biases and errors in generated content
- Preserving confidentiality of individual information
- Impact on the workforce
- Lack of transparency

Questions for Our Speakers

- What AI application within your NSO have you found most exciting so far?
- Without regard to how challenging it may be, what AI application would you most want to move forward within you NSO?
- What is the largest challenge to the implementation of new AI applications within your NSO?

IAOS-ISI 2024 MEXICO CONFERENCE

Thank you!

Linda.J.Young@usda.gov