

## Official Statistics and Data Science in the Fourth Industrial Era

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

**Organiser:** Osuolale Peter Popoola

**Chair:** Ola Awad-Shakhshir

### **Participants:**

**Osuolale Peter Popoola:** The Fourth Industrial Era and Evolution of Data Science

**Ronald Jansen:** Application of Data Science in Official Statistics

**Lisa Grace Bersales:\*** Data Science: Capacity Building for Official Statistician

**Manana Dewage Diana Dilshanie Deepawansa:** Data Science and Official Statistics: Challenges and Opportunities

\* Work presentation not available or non-existent



# Official Statistics and Data Science in the Fourth Industrial Era

**Osuolale Peter Popoola, PhD**

Mathematics and Statistics Department. Adeseun Ogundoyin Polytechnic, Eruwa . Oyo State, Nigeria





# The Fourth Industrial Era and Evolution of Data Science.

**Osuolale Peter Popoola, PhD.**

**Mathematics and Statistics Department,  
Adeseun Ogundoyin Polytechnic, Eruwa,  
Oyo State. Nigeria**



## ❑ Table of Content

- ✓ Industrial Revolution: What Does it Mean?
- ✓ Industrial Era and Development of Statistics
- ✓ The 4-IE and Data Revolution
- ✓ Why Data Science?
- ✓ Data Science General Overview
- ✓ Conclusion





# What is Industrial Revolutions?

- ❖ IR Are:
- ❖ Wave of major innovations
- ❖ linked to each other and together bringing about a fundamental change in human society
- ❖ Wherein new technologies are developed and introduced
- ❖ Times of technological change
- ❖ Have a particular set of characteristics that are connected to, and contemporaneous with, broader social transformation.
- ❖ Changes that go beyond discreet technological capabilities and, shift entire systems of Humans interaction.
- ❖ Link to Evolutions and Transformation in Data and Statistics Production

## IE and Development of Data Production

- ❑ The 1-IE was the transition from human and animal labour technology into machinery
- ❑ It brought about chemical manufacturing and iron production processes
- ❑ It improved efficiency of water power, increasing use of steam engine, and development of machine tools.
- ❑ Statistics and Data Collection have also been evolving into more complex forms over a period of time or Era
- ❑ The earliest data collections took the form of census, Hand Counting, the use of Tallies
- ❑ Data Storage in 1-IE was difficult

## 2-IE and Data Production

- ❑ It builds on 1-IE.
- ❑ Brought about by expansion of electrical technology.
- ❑ Transformed by unprecedented urbanization and rapid territorial expansion.
- ❑ Time of great technological advancement.
- ❑ Rapid changes in communication and manufacturing technologies.
- ❑ In 2-IE, Sampling was discovered which gave rise to surveys, and Multi-stage survey.
- ❑ Data Storage evolve using Vacuum Tubes and Transistors technologies
- ❑ Data storage becomes easier and adding machine was discovered
- ❑ Leading to huge improvements in quality of life for people all over the world

## 3-IE and Data Production

- ❑ It came into being through combinations of science, technology and demand for products
- ❑ 3-IE is “The Digital Revolution Era”
- ❑ Move from mechanical and analogue electronic Technology to Digital Electronics
- ❑ It is based on energy transition and digital technologies, and the internet.
- ❑ Data becomes big, diskette, flash-Drive, CD-ROM, Micro-chips etc were discovered



## 4-IE and Data Production

- The 4-IE is building on the Third, the digital revolution that has been occurring since the middle of the last century.
- It is characterized by a fusion of technologies that is blurring the lines between the physical, digital, and biological spheres.
- reduce barriers between inventors and markets due to new technologies such as 3D printing for prototyping
- increasing trends in artificial intelligence, Robotic, Electric Car, Ultra-Fast Train ,Drive less cars etc

## The Fourth Industrial Era...

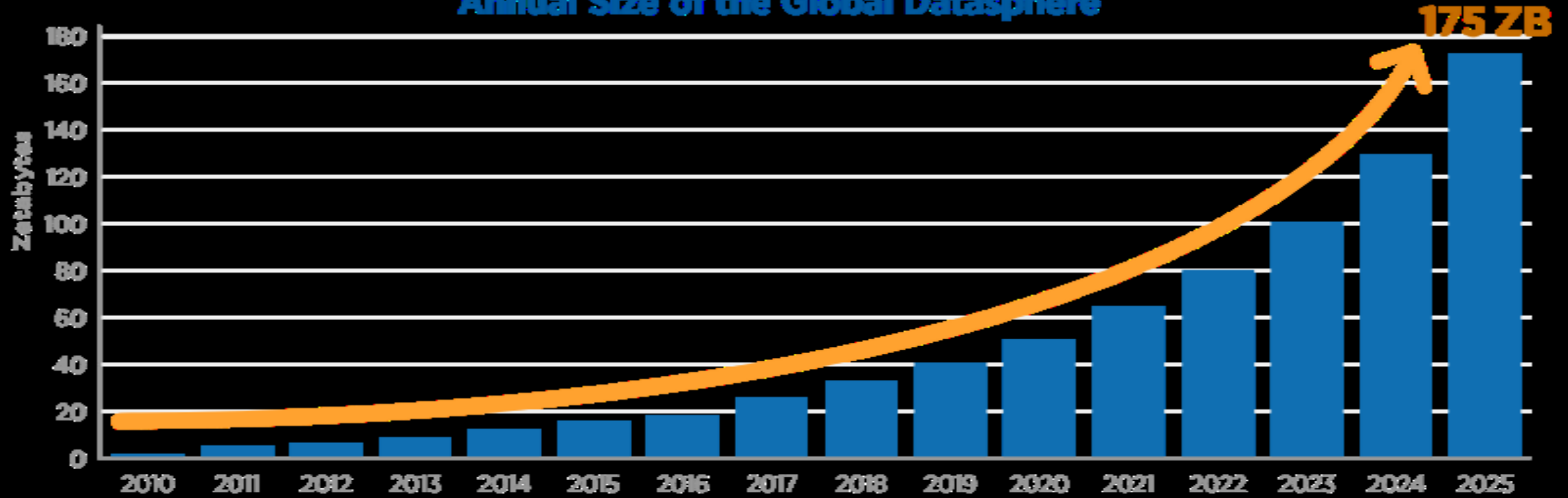
- It an exponential growth of several key technological fields' concepts, such as intelligent materials and block chain technology
- A name for the current trend of automation and data exchange in manufacturing technologies, including cyber-physical systems, the Internet of things, cloud computing and cognitive computing and creating the smart factory.
- A world where individuals move between digital domains and offline reality with the use of connected technology to enable and manage their lives

# 4-IE and Data Revolution....

- ❑ Data represents a post-industrial opportunity.
- ❑ Through the use of internet, social media, commercial transactions, digital images etc sweaty of Data is being created every seconds
- ❑ The world creates 2.5 exa-bytes [ $10^{18}$ ] of data every day
- ❑ 90% of the data in the world today were created in the last two years
- ❑ In 2022, the digital universe was estimated to consist of 44 zeta-bytes of data
- ❑ Predilected approximately 175 Zeta-bytes data would be created every 24 hours worldwide
- ❑ 4-IE is
- ❑ The Era of Big Data;
- ❑ Data Science; and
- ❑ AI, ML, Cloud Computing...

Figure 1 - Annual Size of the Global Datasphere

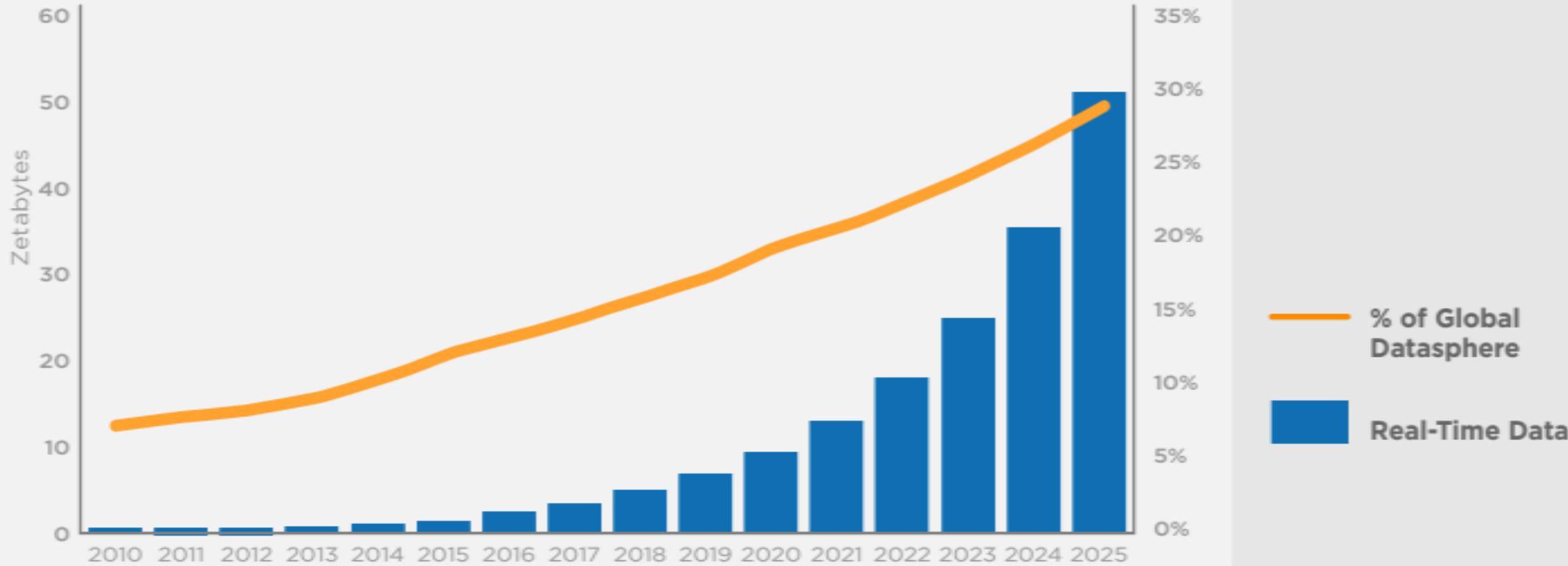
## Annual Size of the Global Datasphere



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Figure 8 - Real-Time Data

## How Much of Global Datasphere is Real-Time?

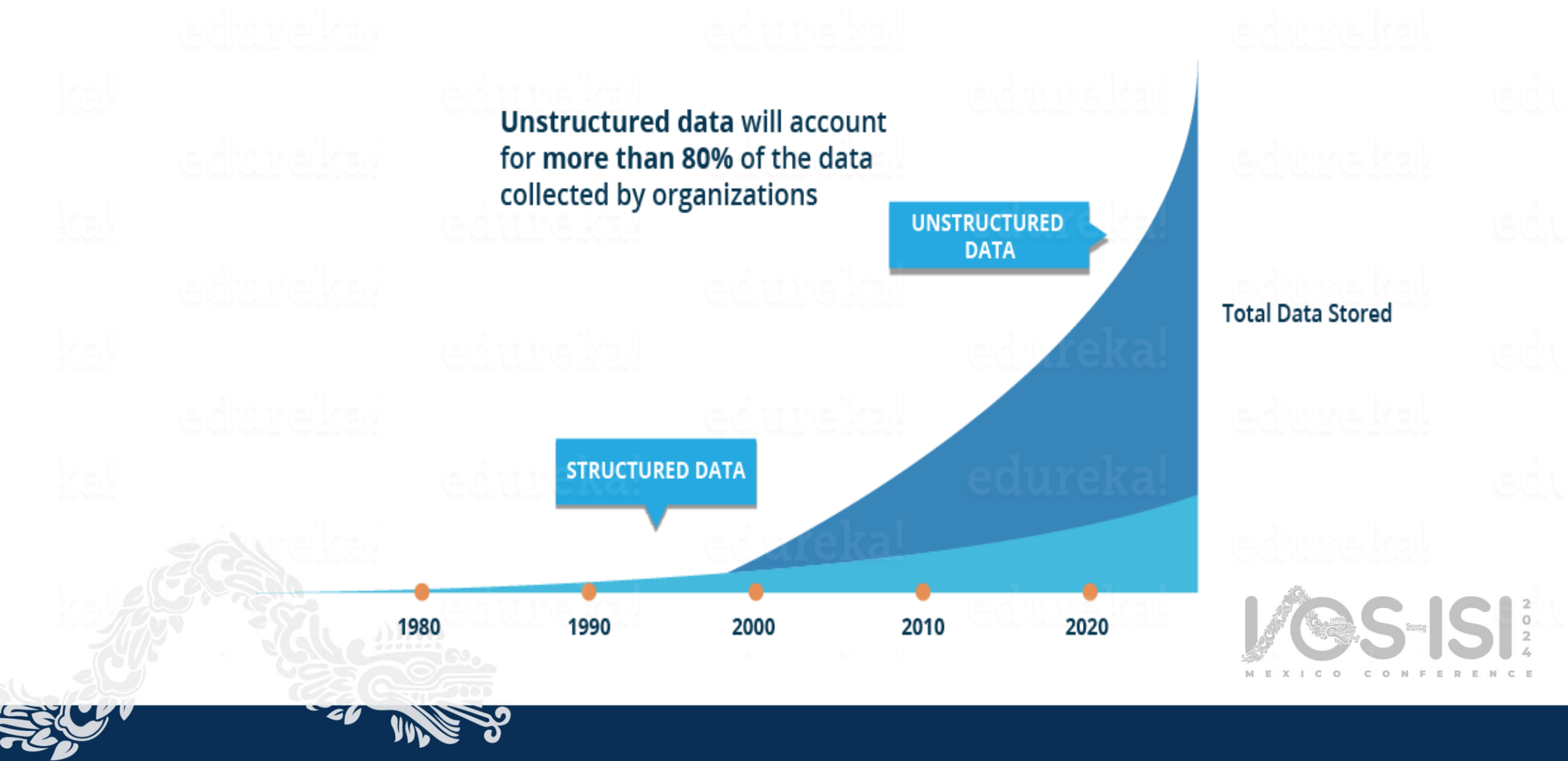


# Why Data Science?

- ❑ Data revolution is already here, and it will continue to grow more and more at an accelerated rate in the days and months and years to come.
- ❑ In this contemporary world, there is no shortage of data. Rather we have a plethora of data almost everywhere
- ❑ Before this Era, Official Statistics were obtained either from:
  - Regional sample survey;
  - Large scale census;
  - Vital statistics ;
  - or Administrative methods
- ❑ Which are:
  - mostly structured,
  - Small in size,
  - Analyzed by using simple statistical tools and statistical packages.



➤ The below shows more than 80 % of the unstructured data in the world as of 2020.



- ❑ In 4-IE, data are being generated every day in large quantity, every seconds through:
  - use of internet,
  - social media, E-transactions, digital images,
  - astronomical tracking, emails, security devices,
  - satellite activities, research laboratories, communications, transport, bank cards, weather indices etc.
- ❑ Unlike data in the traditional systems which was mostly structured, today most of the data is unstructured or semi-structured..
- ❑ As data continue to grow in size and complexity, new algorithms need to be developed so as to learn from eclectic data sources
- ❑ At the same time, computers have become far more powerful, networking is ubiquitous, and algorithms that can connect datasets to enable broader and deeper analyses than previously possible.

- ❑ What do we do with all of these data? How do we make it useful to us? What are its real-world applications?
- ❑ These questions are the domain of data science
  - The volume and variety of data have far outstripped the capacity of manual analysis.
- ❑ This huge volume, variety and complexity of data being generated in this era need
  - advanced analytical tools
  - algorithms for processing,
  - analyzing and drawing meaningful insights out of it.
- ❑ The limitation of conventional statistics to manage and analyze big data has inspired data analysts to venture into data science.

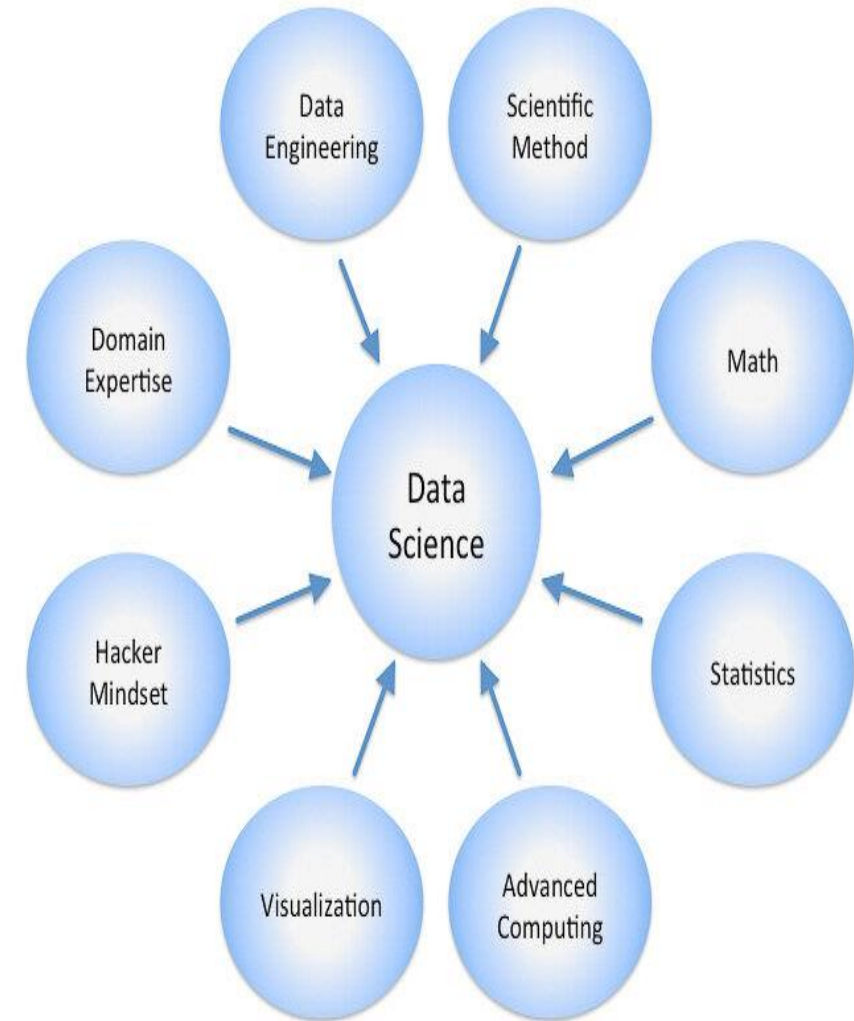
# Data Science General Overview

❑ Data Science is a combination of multiple disciplines that use:

- Statistics;
- Data analysis;
- Machine learning;
- Artificial Intelligence; and,
- Virtualization

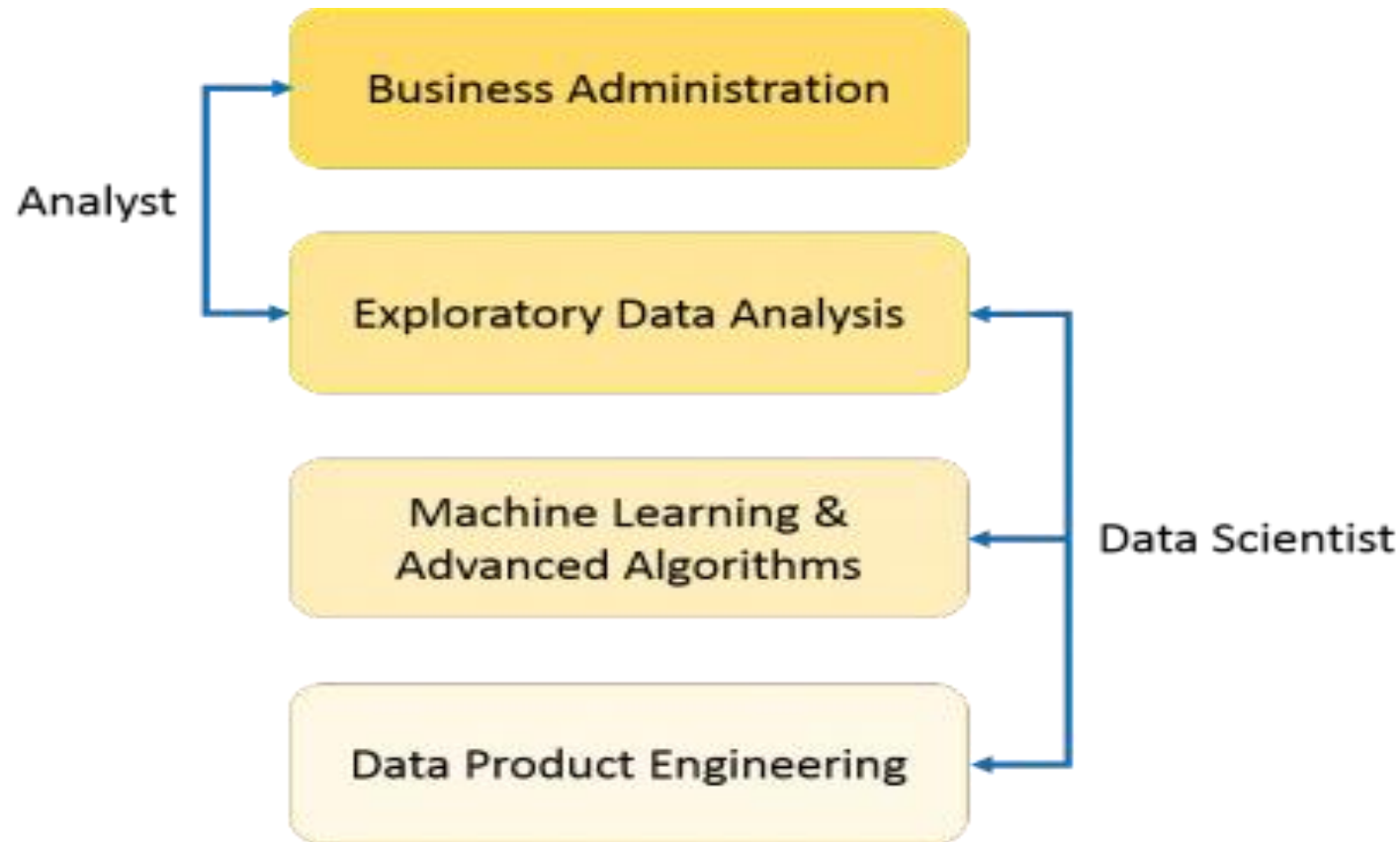
to analyze and extracting meaningful insights from the complex and large sets of data

❑ It is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the complex and large sets of data.



- ✓ Data Science is about data gathering, analysis, visualization and decision-making from unstructured or semi-structured data.
- ✓ Data Science is about finding patterns in data, through analysis, and make future predictions.
- ✓ By using Data Science, NSOs should able to make:
  - Better decisions (should we choose A or B)
  - Predictive analysis (what will happen next?)
  - Pattern discoveries (find pattern, or maybe hidden information in the data)
  - One purpose of Data Science is to structure data unstructured data, making it interpretable and easy to work with.

- ❑ Data Scientist not only does the exploratory analysis, but also uses various advanced machine learning algorithms to identify the occurrence of a particular event in the future



## ❑ How Does a Data Science Work?

A Data Science requires expertise in several backgrounds:



- Machine Learning
- Statistics
- Programming (Python or R)
- Mathematics
- Databases
- A Data Scientist must find patterns within the data. Before he/she can find the patterns, he/she must organize the data in a standard format.
- ❑ **Here is how a Data Science works:**
  - ✓ **Ask the right questions** - To understand the business problem.
  - ✓ **Explore and collect data** - From database, web logs, customer feedback, mobile network provider.

- ✓ **Extract the data** - Transform the data to a standardized format.
- ✓ **Clean the data** - Remove erroneous values from the data.
- ✓ **Find and replace missing values** - Check for missing values and replace them with a suitable value (e.g. an average value).
- ✓ **Normalize data** - Scale the values in a practical range (e.g. 140 cm is smaller than 1,8 m. However, the number 140 is larger than 1,8. - so scaling is important).
- ✓ **Analyze data, find patterns and make future predictions.**
- ✓ **Represent the result** - Present the result with useful insights in a way the "Government" can understand.

❑ The future of Official Statistics lies in incorporation of Data Science into the production of official data for National development



**Thank you**





# APPLICATION OF DATA SCIENCE IN OFFICIAL STATISTICS

- Gabriel Gamez & Ronald Jansen
- United Nations Statistics Division
- Thursday, 16 May 2024, 16:00 – 17:30




# OUTLINE

- Data Science and Official Statistics – Stefan Schweinfest & Ronald Jansen
- Data Science Leaders Network
- Future Perspective
  - *Data Science teams across government*
  - *Changes to Statistical Legislation*



# DATA SCIENCE AND OFFICIAL STATISTICS

***Paper in Harvard Data Science Review – Stefan Schweinfest & Ronald Jansen, Dec 2021***

- Organizing a complex national eco-system of data sources
  - Protecting privacy while accessing relevant data
  - New Job Profiles: Data scientist, Data engineer and Data analyst
  - Communication of uncertainty – Experimental Data
  - Fundamental Principles for the larger data landscape?
- 



# DATA SCIENCE LEADERS NETWORK



Address challenges and constraints of a culture of innovation within NSOs



Implement data science into the statistical production process



Build and strengthen effective data science partnerships nationally and internationally

Innovation culture

International collaboration

Production process change

Data science projects

# WHAT IS DATA SCIENCE IN OFFICIAL STATISTICS?

- **Automation** of the statistical production processes (increase efficiency and improve quality)
- **Supplementary indicators** produced for emerging issues to provide additional insights
- **Changing statistical production**






# EXAMPLES OF DATA SCIENCE IN OFFICIAL STATISTICS

- *Satellite data for Agriculture Statistics*
  - Machine Learning for Crop mapping and identification
- *Scanner data and webscraping for Price statistics*
  - Optimizing the web crawlers
  - Automated Classification of product descriptions
- *AIS vessel tracking data for Maritime Transport indicators*
  - Use of Polygons around harbors to estimate Port activity



# KEY FEATURES OF THE DSLN PLAYBOOK

- **Purpose:** Guide NSOs in integrating and adopting data science effectively.
- **Living Document:** Continuously updated to reflect new insights and technologies.
- **Balanced Approach:** Complements traditional statistics with new tools and methodologies.
- **Practical approach:** provide use cases and coding examples as part of the Playbook through GitLab format



# OVERALL STRUCTURE OF THE PLAYBOOK



**Section 1:** Leveraging basic tools of data science for immediate efficiency gains in NSO operations



**Section 2:** Generating additional insights in response to emerging needs



**Section 3:** Full transformation of official statistics through digitalization



**Section 4:** Cross-sectional themes

## FUTURE PERSPECTIVE

- *Data Science is here to stay – NSOs need to adapt (Data Science teams inside NSOs)*
- *New Job Profiles; New training – continuous learning*
- *Necessary collaboration with Private sector, Research institutes and Academia*
- *Changes to Statistical Legislation*
- *Communication of uncertainty – Experimental Data*





## FUTURE PERSPECTIVE

- ***Data Science teams across government, led by the NSO – How could this work?***
  - *For emerging issues (like COVID pandemic) – the NSOs can lead the Data Science task force with expertise and adherence to fundamental principles*
  - *Data Scientists of NSOs can advise other parts of Government with data quality assurance framework*



# FUTURE PERSPECTIVE

## **Changes to Statistical Legislation**

- *Statistics Legislation is part of a broader data regulatory framework*
- *Legislation can facilitate collaboration among many stakeholder communities in the society-wide data eco-system*
- *Legislation can enable access to privately held data*



# Official Statistics and Data Science in the Fourth industrial Era

**Dr Dilshanie Deepawansa**  
Department of Census and Statistics  
Sri Lanka





# Data Science and Official Statistics: Challenges and Opportunities

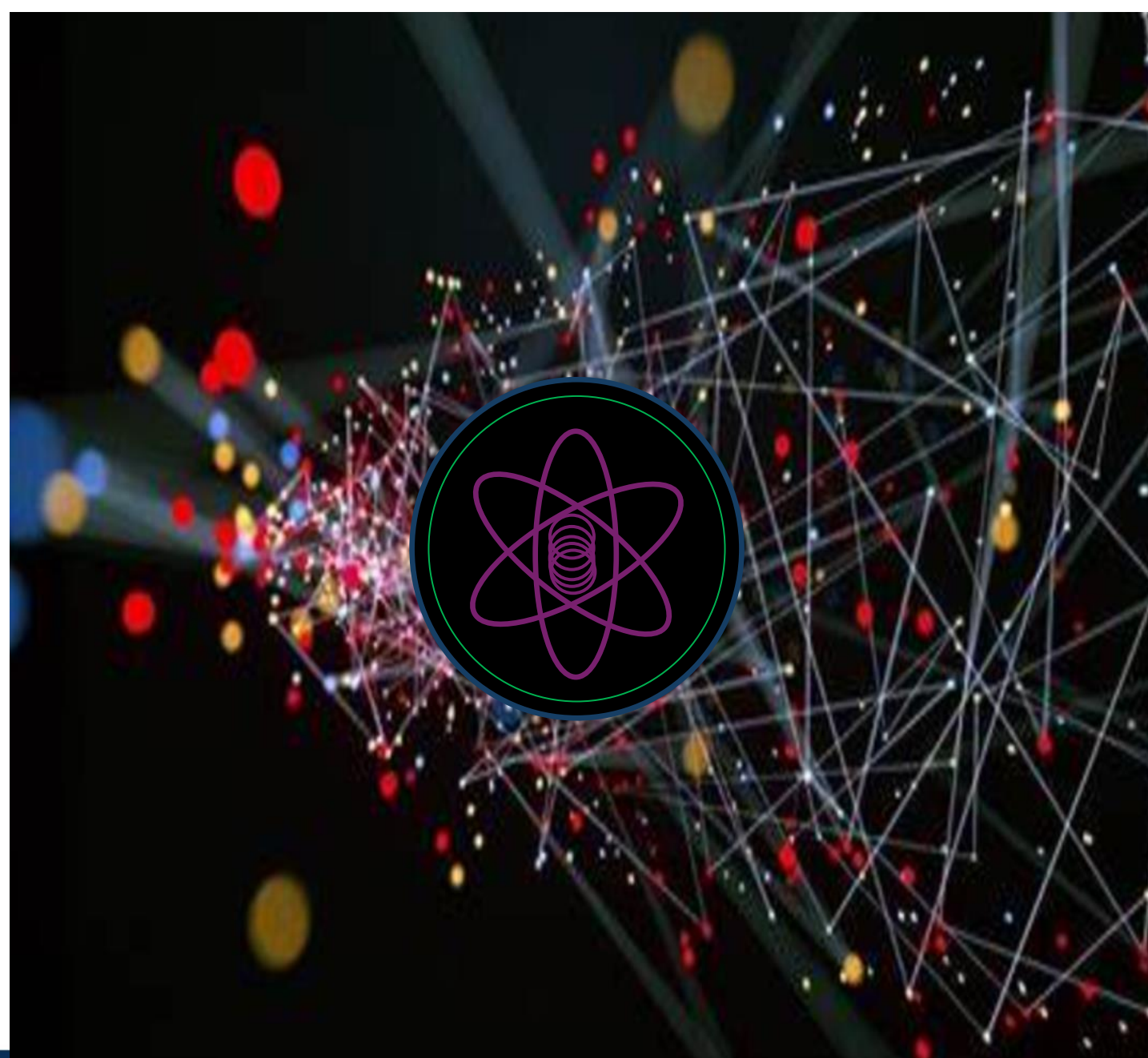


# Data Sáciense

Data science is an interdisciplinary academic field that uses :

- statistics
- scientific computing,
- scientific methods
- processes
- Algorithms and systems

to extract or extrapolate knowledge and insights from potentially noisy, structured, or unstructured data.



# Official Statistics

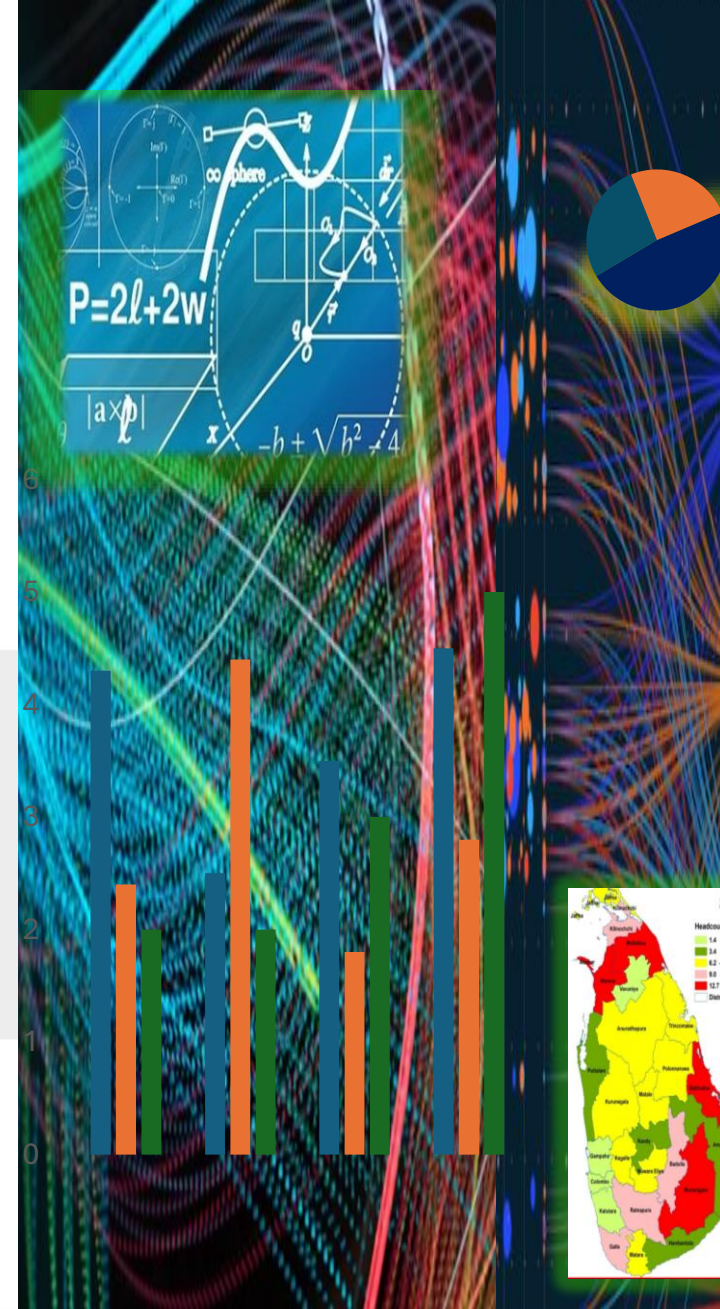
- Official statistics are statistics published by government agencies or other public bodies such as international organizations as a public good.
- They provide quantitative or qualitative information on all major areas of citizens' lives, such as economic and social development, living conditions, health, education, and the environment.
- Official statistics result from the collection and processing of data into statistical information.
- They are then disseminated to help users develop their knowledge about a particular topic or geographical area, make comparisons between countries or understand changes over time.
- Official statistics make information on economic and social development accessible to the public, allowing the impact of government policies to be assessed, thus improving accountability





# Data Science and Official Statistics

- Extracting meaningful patterns from data using machine learning (ML), artificial intelligence (AI), and visualization for Official statistics
- National statistical offices (NSOs) can benefit from these opportunities to improve data for effective decision-making for development policies
- Data Science methodologies offer a powerful set of tools and techniques for enhancing the production and analysis of official statistics, enabling statisticians to work with larger, more complex datasets and extract meaningful insights to inform policymaking and decision-making processes
- Traditional statistical methods remain indispensable for their interpretability and robustness in many statistical tasks, emerging data science techniques offer powerful tools for handling complex data





# Challenges

- **Data Quality:** Address issues related to data quality, accuracy, completeness, and consistency
- **Privacy and Ethics:** Discuss the ethical considerations and privacy concerns when dealing with large datasets, especially in the context of official statistics.
- **Technological Infrastructure:** Explore challenges related to the infrastructure required to handle large volumes of data and perform complex analyses.
- **Skills Gap:** Highlight the need for specialized skills in both data science and official statistics, and discuss potential strategies for addressing this gap.

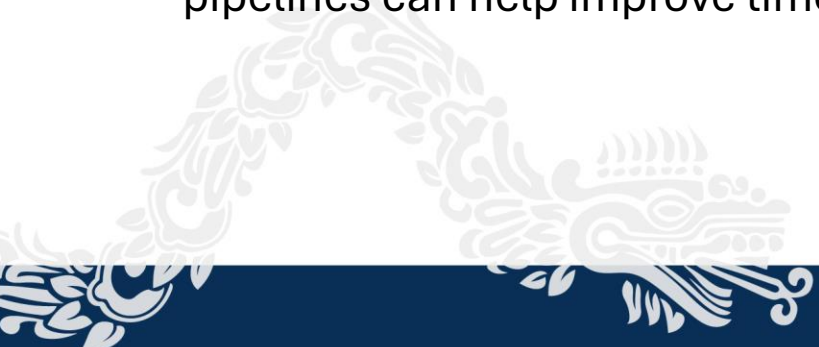
# Challenges to Data Quality in Official Statistics

**Data Accuracy** : In official statistics, inaccuracies can arise due to measurement errors, reporting errors, or sampling biases. Similarly, in data science, inaccuracies can be arising from data collection methods, data entry errors, or algorithmic biases.

**Data Completeness** : Addressing missing data requires careful imputation methods

**Data Consistency** : Inconsistencies can arise from variations in data collection methodologies, definitions, or classifications over time or across regions. Data harmonization efforts are necessary to reconcile discrepancies and ensure consistency.

**Data Timeliness**: Official statistics often need to be produced and disseminated in a timely manner to support decision-making processes. Leveraging real-time data sources and implementing efficient data processing pipelines can help improve timeliness.

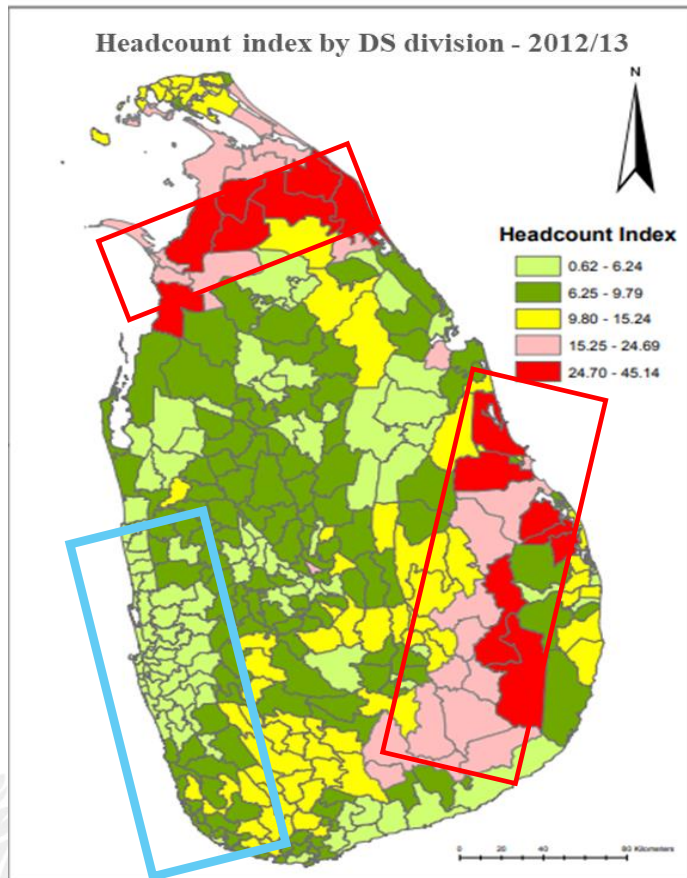


# Opportunities in Data Science

- **Enhanced Insights:** Show how data science techniques can provide deeper insights and uncover hidden patterns in official statistics.
- **Real-Time Data:** Discuss the potential for using real-time data sources (like social media, internet and electronics devices) to complement traditional survey methods and improve the timeliness of official statistics.
- **Predictive Analytics:** Explore how predictive analytics can be applied to official statistics to forecast trends and anticipate future changes.
- **Data Visualization:** Highlight the importance of data visualization techniques in communicating official statistics effectively to policymakers and the general public.

# Data Integration and poverty mapping: The Experience of Sri Lanka

## Official Statistics

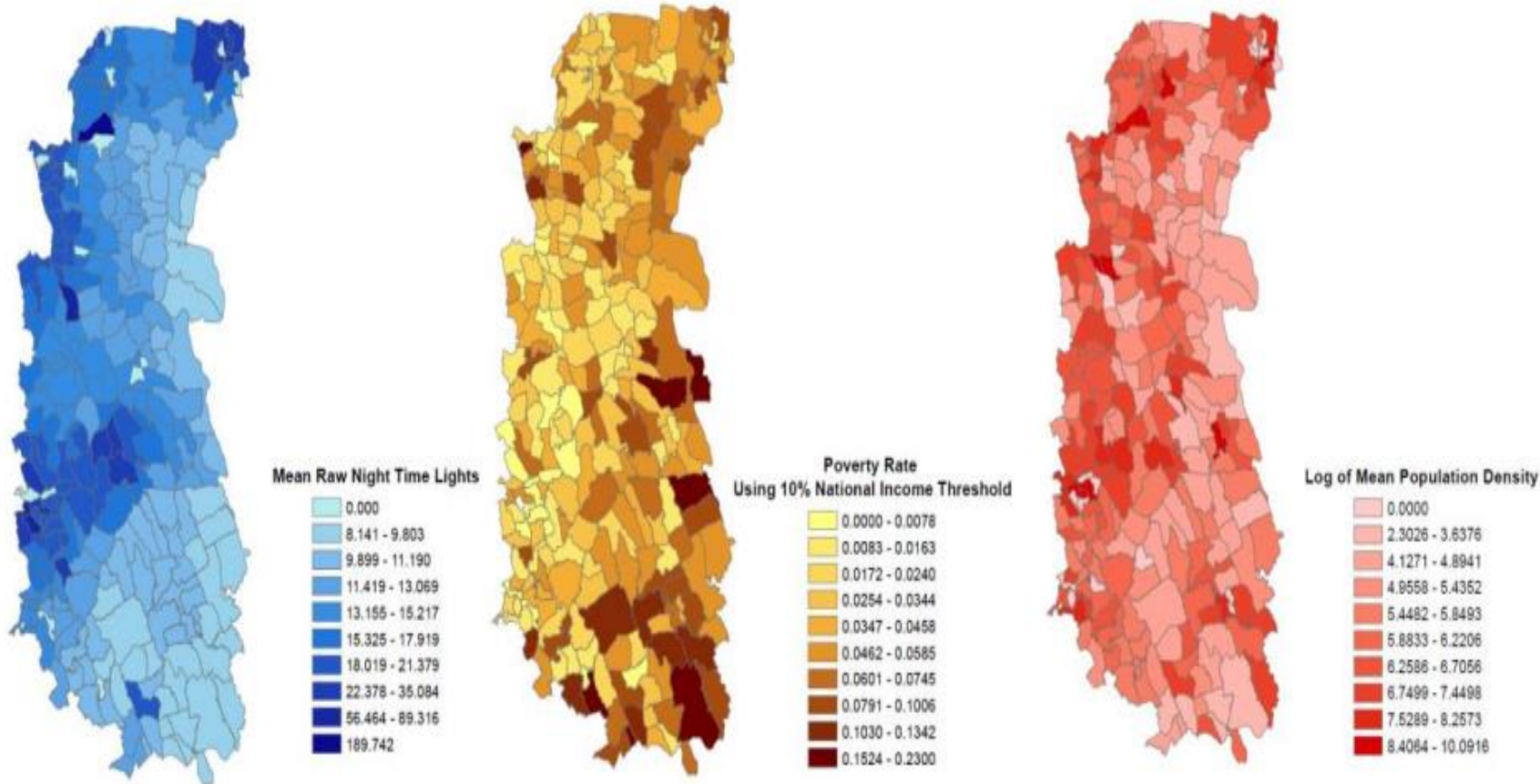


## Night-time lights



Satellite imagery and other innovative big data sources to measure poverty  
E.g.;  
using covariates from either high-resolution spatial features, night-time lights

## Poverty from Space: Using High Resolution Satellite Imagery for Estimating Economic Well-being



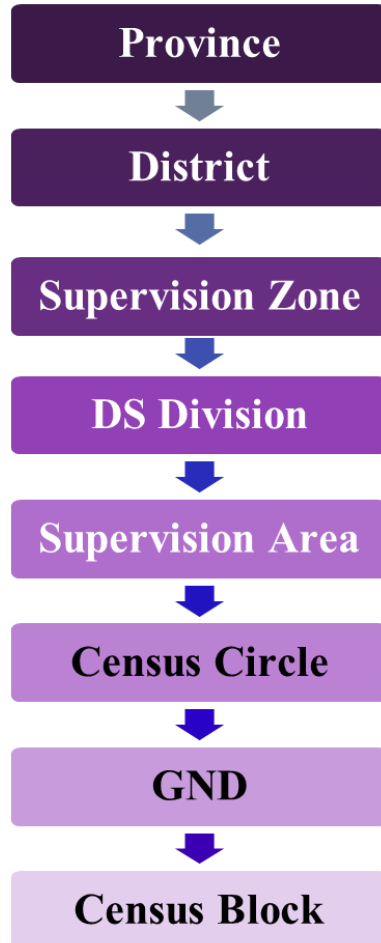
Engstrom, Hersh and Newhouse, (2017) have accomplished a study using high spatial resolution satellite images to accurately estimate poverty and economic well-being for the Divisional Secretariat of Seethawaka in Sri Lanka.



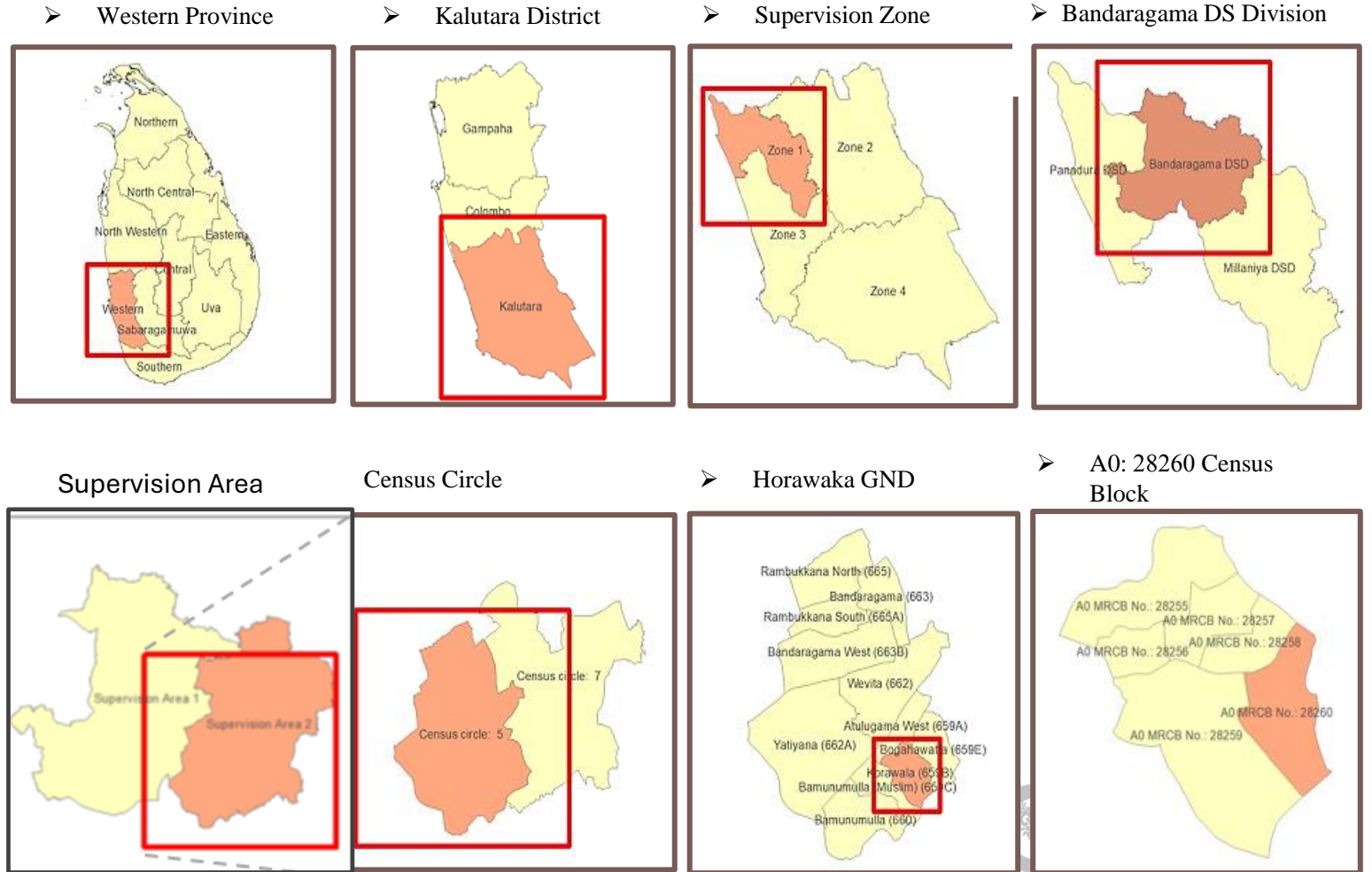
# Future Directions

## Census of Population and Housing 2024 – Listing Stages

Geographical administration structure of CPH

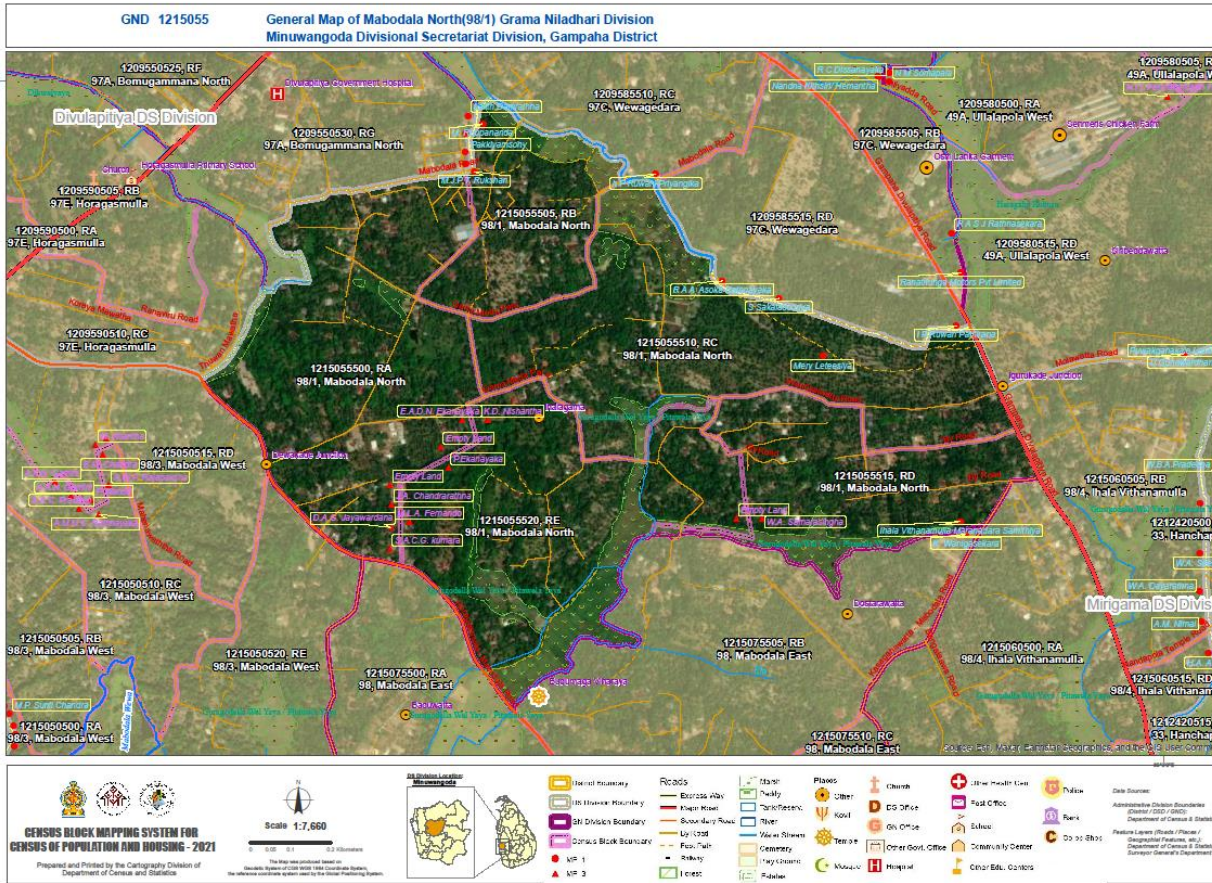


88,000

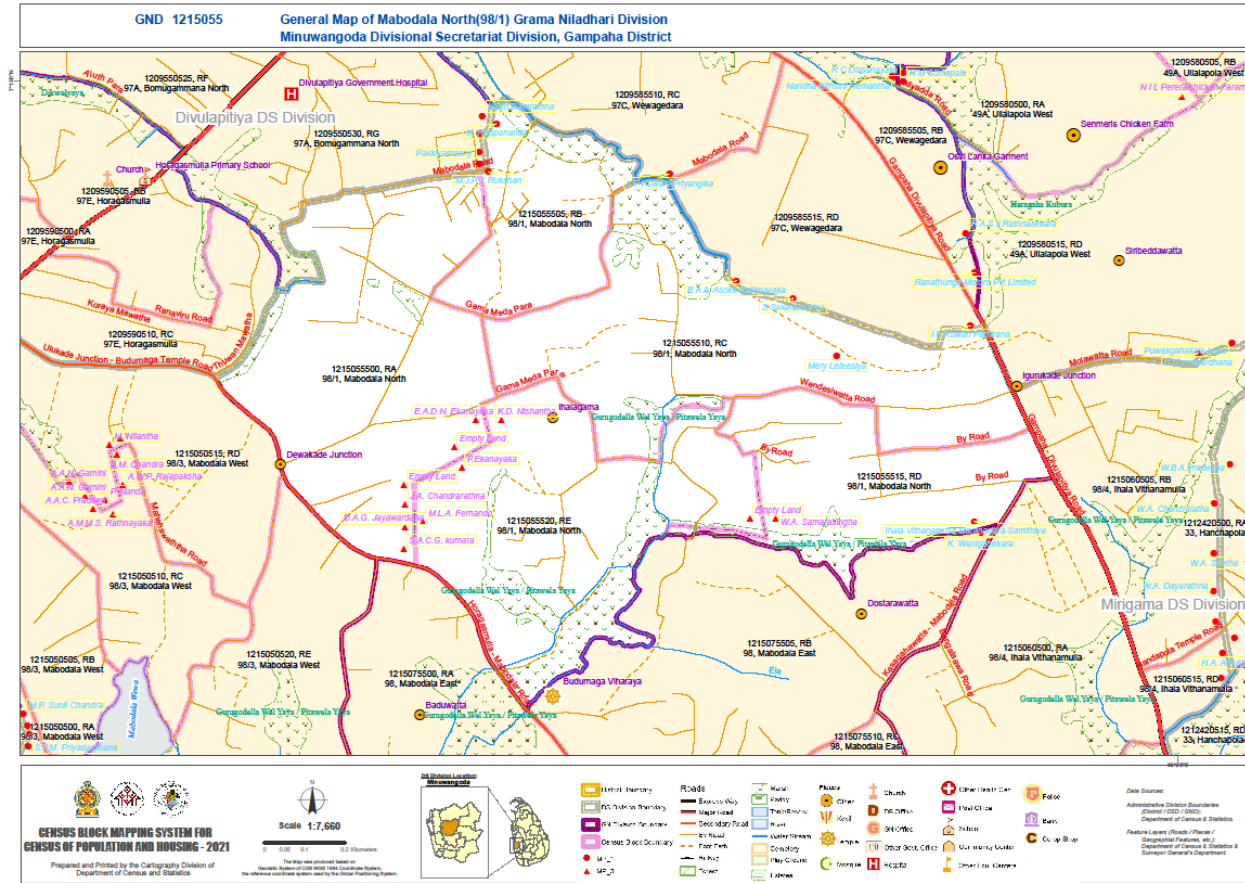


# Geo-codes

GN Division Map (with satellite images)



GN Division Map (without satellite images)

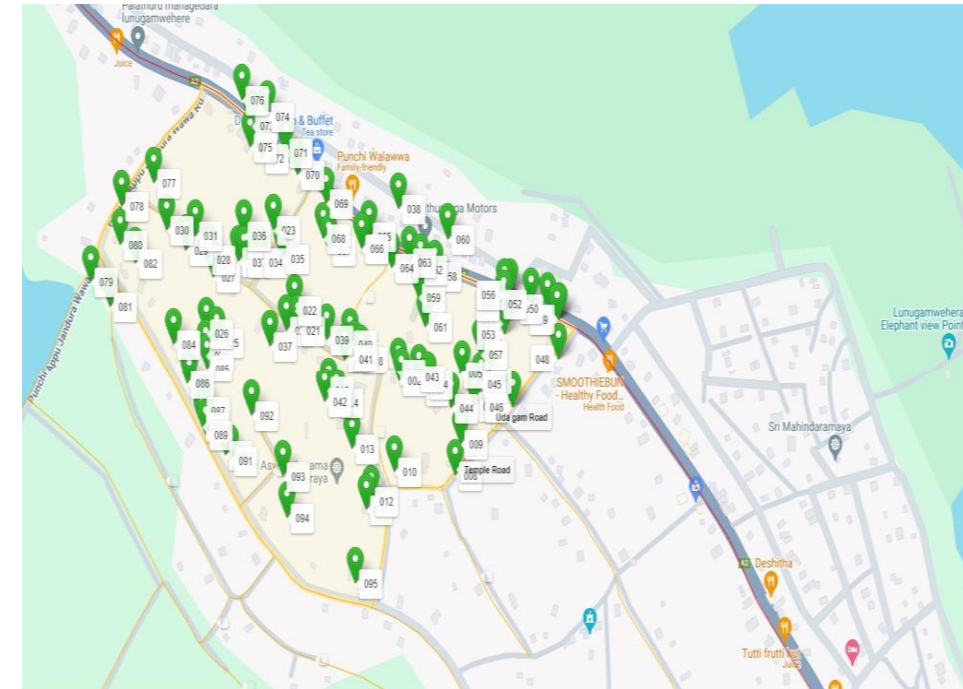
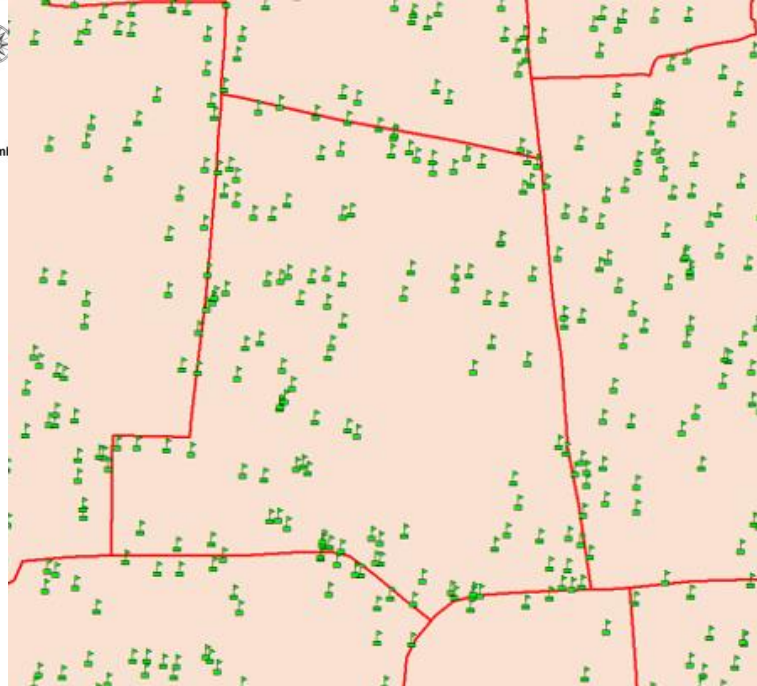
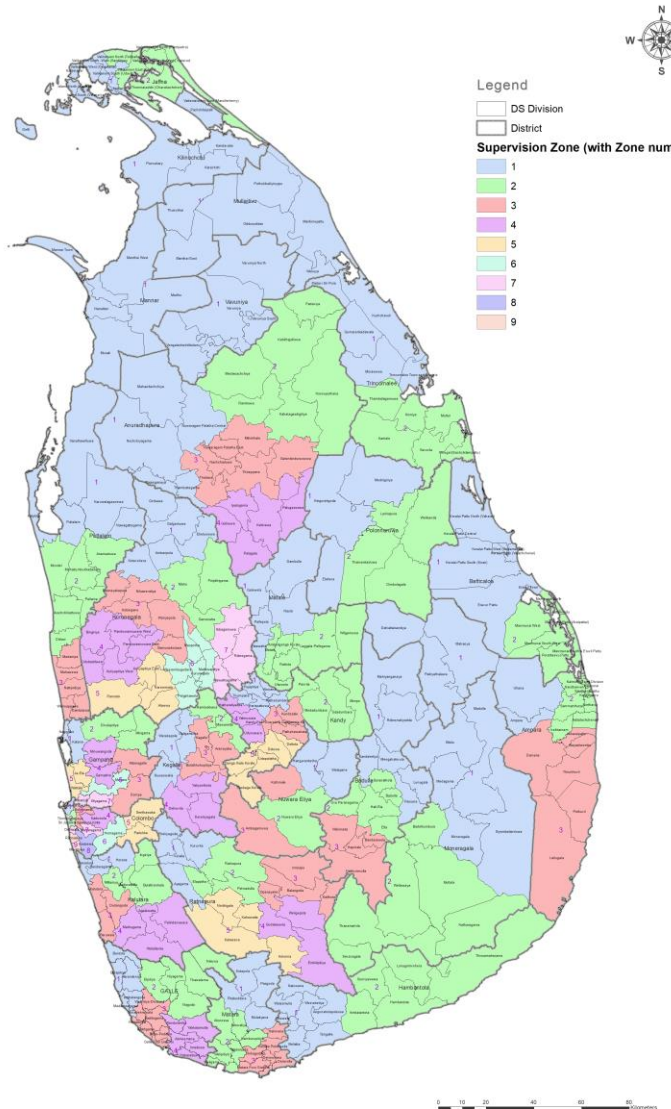




# Geo-codes

Number of building coordinates prepared at the census listing stage

Supervision Zone by District in Sri Lanka - CPH 2024



# Benefits of Preparing a separate Geo-special database for Sri Lanka

Further development of different boundary areas ( Ex: Educational zone, Police area, The medical officer of health (MOH) areas

- Adding new data fields to the database(ex: cultivated land, bear lands, forest...
- Data analysing , forecasting and decision making
- An efficient information gathering and public reporting system
- Creating Apps to develop the systems





# Conclusión

- Collaboration between data scientists and statisticians is essential for connecting the full potential of data science in official statistics.
- Data scientists and statisticians can address the challenges and capitalize on the opportunities presented by data science to produce high-quality, reliable, and actionable statistical information for evidence-based decision-making and policy formulation by;
  - Leveraging their complementary expertise
  - Fostering methodological innovation
  - Ensuring data quality and validation
  - Promoting interdisciplinary insights
  - Investing in capacity building

# Reference

Engstrom, R., Hersh, J., & Newhouse, D. (2022). Poverty from space: Using high resolution satellite imagery for estimating economic well-being. *The World Bank Economic Review*, 36(2), 382-412.

Gibson, J., Olivia, S., & Boe-Gibson, G. (2020). Night lights in economics: Sources and uses 1. *Journal of Economic Surveys*, 34(5), 955-980.

Ngestrini, R. (2019). Predicting poverty of a region from satellite imagery using CNNs (Master's thesis).

Engstrom, R., Hersh, J., & Newhouse, D. (2016). Poverty in HD: What does high resolution satellite imagery reveal about economic welfare. Available online: Pubdocs. worldbank.org/en/60741466181743796/Poverty-in-HD-draft-v2-75.pdf (accessed on 1 December 2016).

Afzal, M., Hersh, J., & Newhouse, D. (2015). Building a better model: Variable selection to predict poverty in Pakistan and Sri Lanka.





**Thank you**

