LIRNEasia
Pro-poor. Pro-market.

Applied Data Science Research for Social Good

Industrial session at UOK

2 October 2023

At the end of the session you,

- **Have a better understanding of applied Data Science Research and how it differs from traditional academic research & the private sector**
- **Are seriously considering AI 4 Social Good as an area you want to work in**
- Know what LIRNEasia is and can explain the work we do to a friend
- Find our work as inspiring as we do and are seriously considering joining us
- Have enough information to decide LIRNEasia is a good fit for you and feel comfortable enough to reach out to us

Structure of the Session

- Part 1 – About LIRNEasia and our work
- Part 2 – A Case study – how we do applied research
- Part 3 – Broader QnA

About LIRNEasia

LIRNEasia is a non-profit, multi-disciplinary research organization that aims to bring about policy change and develop solutions through research to improve the lives of people in the Asia and Pacific using knowledge, information and technology. Our **experts come from diverse backgrounds such as economics, law, computer science, and social sciences, enabling us to approach problems in a multi-disciplinary manner**.

Our research covers a wide range of topics including **ICT policy and regulation, infrastructure development, health, social safety nets, disability, and disaster risk reduction**. We work closely with governments, private sector organizations, and civil society groups to ensure our research has real-world impact and contributes to positive change in the region.

https://lirneasia.net/category/themes/

LIRNEasia – Senior Staff

Rohan Samarajiva
Chairman, Board of Directors

Helani Galpaya
Chief Executive Officer

Nilusha Kapugama
Chief Operating Officer

Ayesha Zainudeen
Senior Research Manager – Gender specialist

Tharaka Amarasinghe
Project Manager & Statistician – Survey specialist

Gayani Hurulle
Senior Research Manager - Economist

Isuru Samaratunga
Research Manager – Qualitative specialist

Chiranthi Rajapakshe
Research Manager – Legal specialist

A glimpse of “multi-disciplinary”

- **Social Safety Nets-** Using data and technology to understand poverty and improve the efficiency, reach, and effectiveness of social safety nets. We aim to inform/influence policy design and implementation, ensuring that the limited resources reach the most vulnerable groups. *(surveys, KIIs, FGDs, desk research, web scraping, ML, policy analysis)*

- **Misinformation experiments –** Understanding which interventions will make people less susceptible to harmful effects of misinformation through randomized controlled trials *(Surveys, RCTs, desk research)*

- **Data governance –** A comparative analysis of countries in the Asian Region with respect to their policies, regulations & strategies around data governance (Policy & legal analysis, KIIs)

LIRNEasia
Pro-poor. Pro-market.

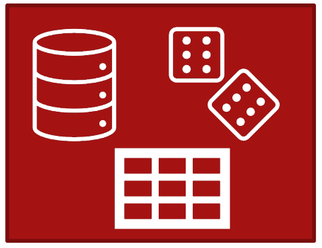
The DAP Team

Data, Algorithms, and Policy

The DAP Team - Introduction

Apply data science and machine learning to inform policy & do social good

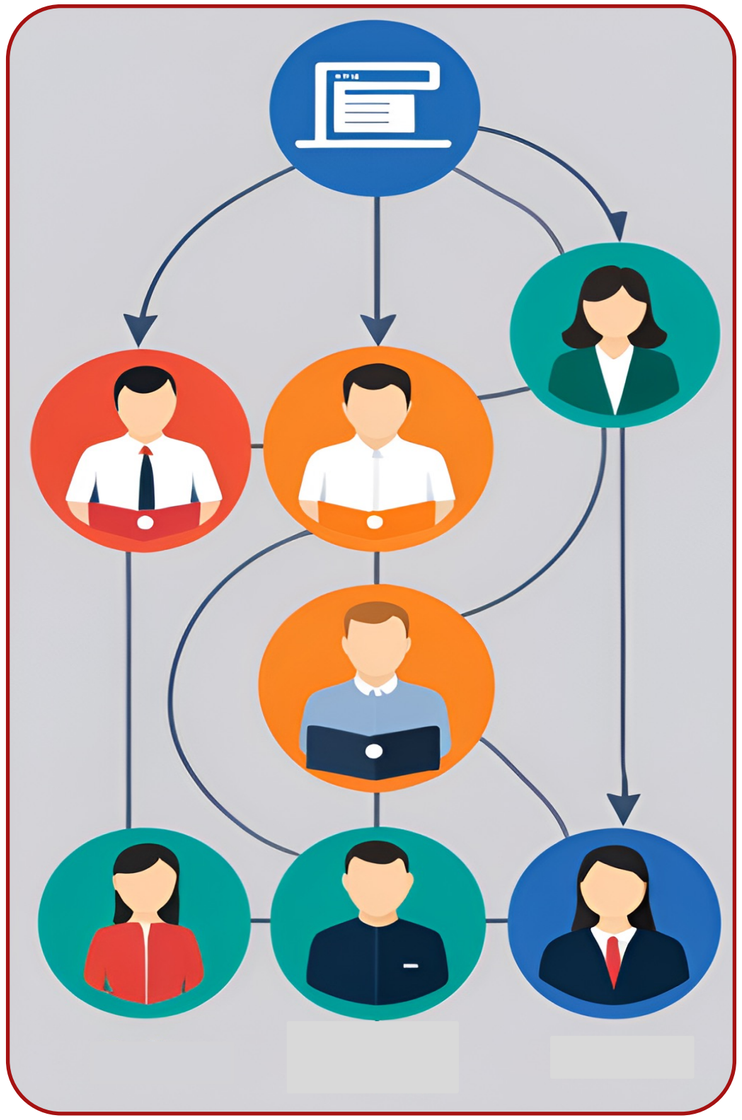
Big Data For Development → **Language & NLP** → **Applying ML to the real-world**



Data

Algorithms

Policy



Our team – DAP Team (the CORE team)

Merl Chandana
Research Manager & Team Lead (Data, Algorithms, & Policy)

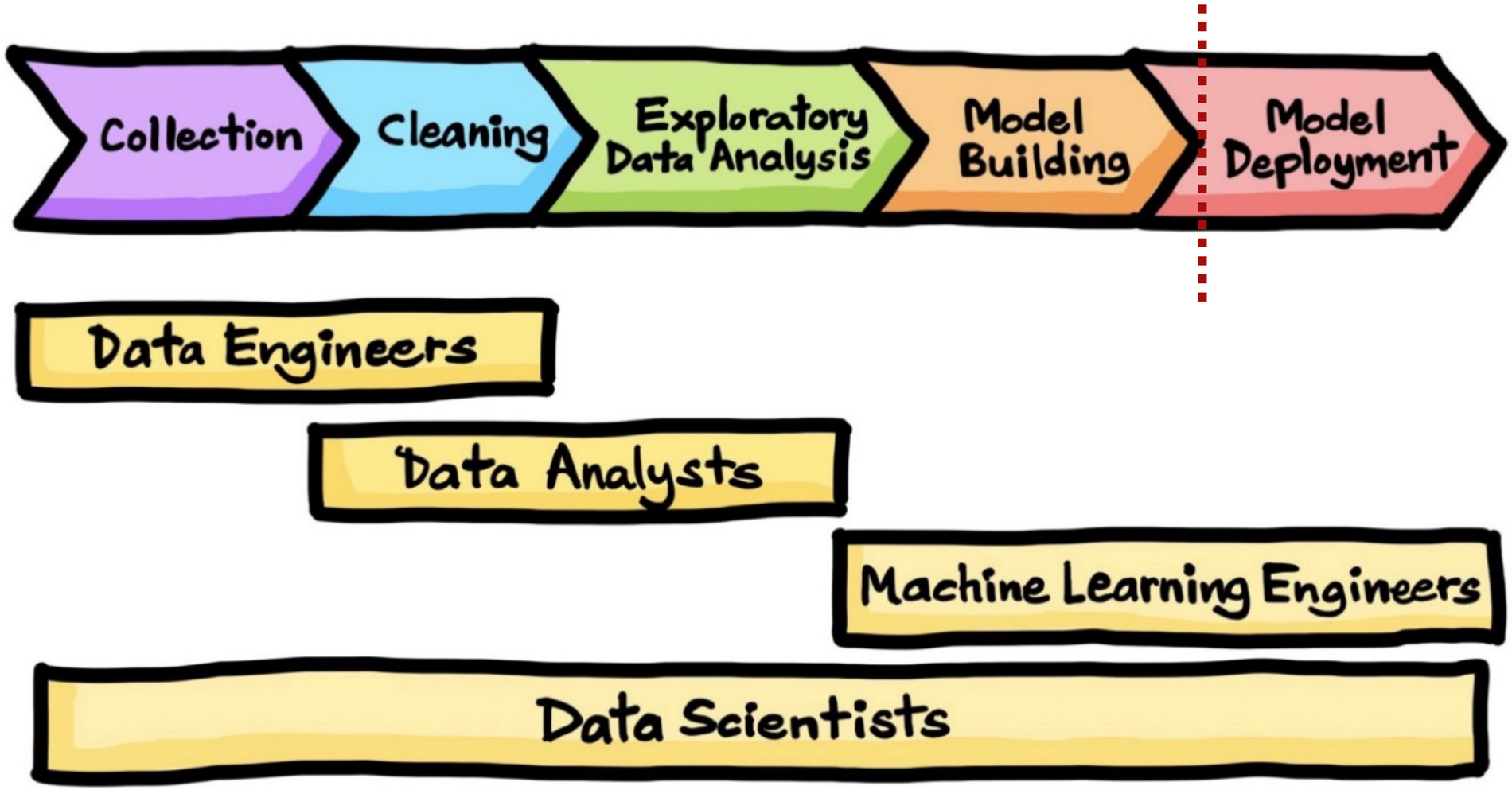
Kasun Amarasinghe
Research Fellow & Consultant

Chanuka Algama
Junior Researcher

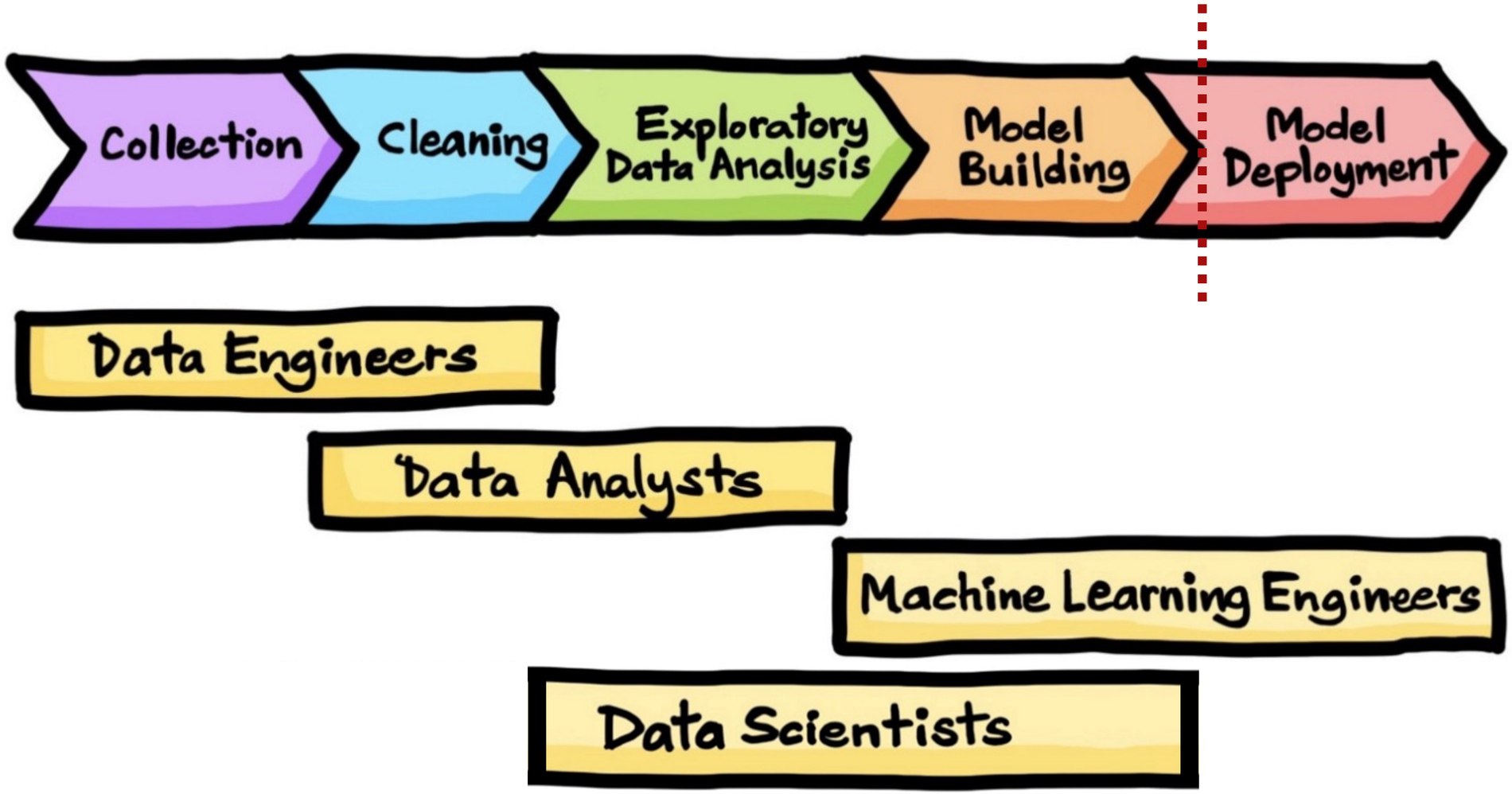
The Process of Applied Data Science

1. Does this look right?
2. What is wrong with it?
3. What is missing?
4. How would you improve this?

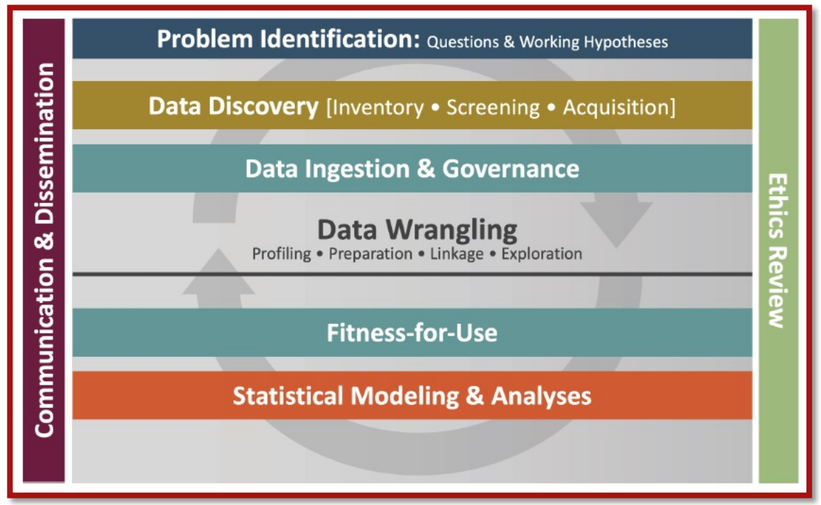
Roles Within the DAP Team - I



Roles within the DAP Team - II



Data Science Project Ethics Checklist



Problem Identification

Establish the ethical basis for undertaking the project as well as the project requirements of both the protection of research participant and the equitable allocation of all potential project benefits and risks.

- ✓ What are the expected benefits of the project to the "public good," and do they outweigh potential risks to certain populations?
- ✓ Are there implicit assumptions and biases in the framing of the project regarding the studied communities and how will they be addressed?
- ✓ What type of Institutional Review Board approval process is needed? Has the team reviewed the protocol?

Data Discovery, Inventory, Screening, & Acquisition

Consider potential biases that may be introduced through the choice of datasets and variables.

- ✓ Do the data include disproportionate coverage of the different communities of study?
- ✓ Do data have adequate geographic coverage?
- ✓ Have checks and balances been established to identify and address implicit biases in the data?

Data Ingestion and Governance

Put in place data platforms and processes to ensure data transfer, storage, and database development adheres to data governance agreements and best practices for data quality assurance.

- ✓ Have team members reviewed standard operating procedures (SOPs) and data management plans?
- ✓ Do additional procedures need to be defined for this project?

Fitness-for-Use Assessment

Critically assess the overall utility of the results in achieving the predicted benefits of the study, to be transparent about potential limitations of the study, and to ensure that unintended biases haven't been introduced as a result of data choice and model refinement.

- ✓ What are the limitations of the results? Are the results useful given the purpose of the study?
- ✓ Do the statistical results support the potential benefits of the study previously stated?
- ✓ Do the statistical results support the mitigation of the potential risks of the study previously stated?
- ✓ Have any data been deemed unusable that require revisiting the question of potential biases being introduced through the choice of datasets and variables?

Microsoft’s Principles of Responsible AI

Defining what’s important

Microsoft’s six principles to guide AI development and use.

Fairness

Ai systems should treat all people fairly

Reliability and Safety

Ai systems should perform reliably and safely

Privacy and Security

Ai systems should be secure and respect privacy

Inclusiveness

Ai systems should empower everyone and engage people.

Transperancy

Ai systems should be understandable

Accountability

People should be accountible for AI systems

Google’s AI Principles

AI SHOULD:

1. Be socially benificial
2. Avoid creating or reinforcing unfair bias
3. Be built and tested for safety
4. Be accountable to people
5. Incorporate privacy design principles
6. Uphold high standards of scientific excellence
7. Be made available for uses that accord with these principles

- primary purpose and use
- Nature and uniqueness
- Scale
- Nature of google involvement

APPLICATIONS WE WILL NOT PURSUE:

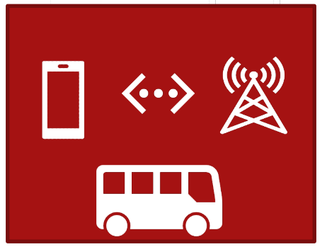
1. Likely to cause overall harm
2. Principle purpose to direct injury
3. Surveiliance violating internationally accepted norms
4. Purpose contravenes international law and human rights

LIRNEasia
Pro-poor. Pro-market.

Our Work

Past & Present

A few snapshots from the past

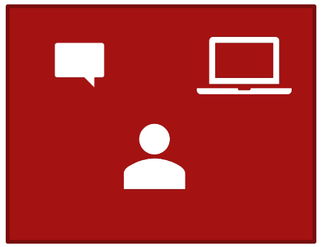


Understanding Human Mobility

Determining the need for roads and bridges. Evaluating their effects upon completion
Understanding the spread of infectious diseases
Studying patterns of internal migration

Policy analysis at the intersection of data, AI and ethics

Comparative study of policies, regulations, and applications of AI in Singapore and India
Developing a discussion paper on the elements of a potential AI policy for Sri Lanka



Online job portal data for labor market analysis

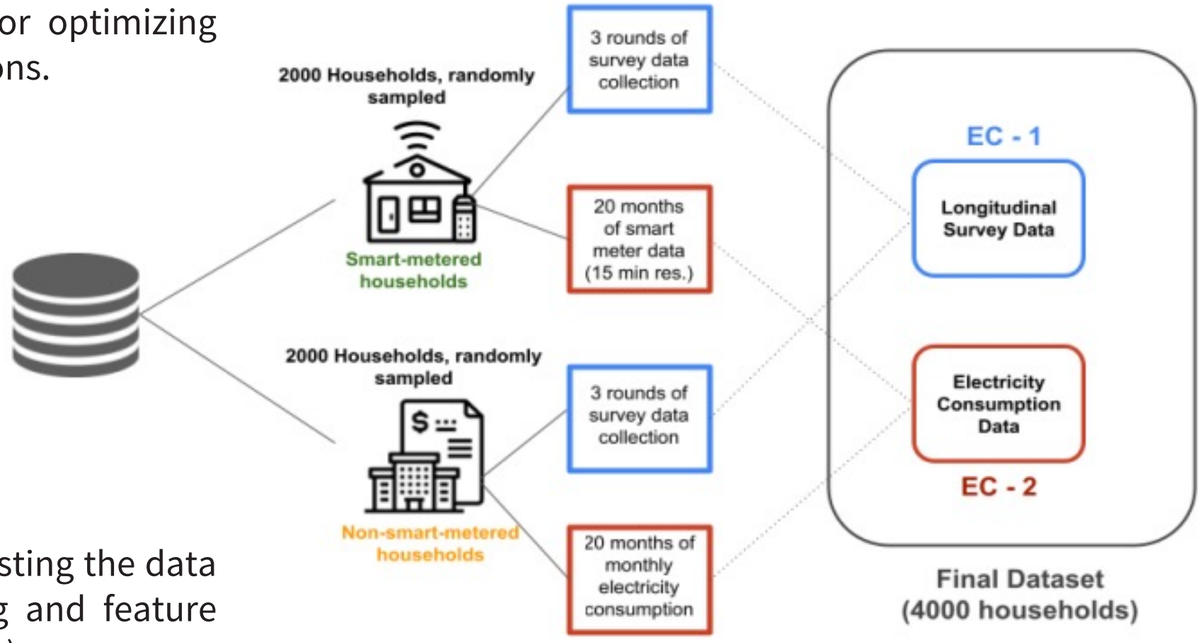
Scoping study on job portals of 12 countries in the region
Analyzing online job advertisements to understand the demand for skills

Ongoing project 1: Building machine learning datasets on electricity consumption

Problem/Context: The project aims to address the need for a comprehensive dataset for machine learning research related to household electricity consumption. Understanding the factors that drive electricity consumption is essential for optimizing energy usage, promoting energy efficiency, and making informed policy decisions.

Data Sources:

- Electricity Consumption Data: The primary data source is the electricity consumption data obtained from electricity providers.
- Longitudinal Survey Data: This survey will consist of three rounds of data collection, capturing various demographic and behavioral factors that influence electricity consumption.



Methods:

Sampling, Data collection, Data integration, Data processing and cleaning, Hosting the data for public access, Machine learning analysis(Predictive modeling, clustering and feature engineering to uncover patterns and insights related to electricity consumption).

Expected Outputs: A comprehensive two-part dataset

Ongoing project 2: Urban boundary mapping: Towards a Better Understanding of Sri Lankan Cities Using Satellite Imagery

Problem/Context: The research project addresses the critical need for accurate and up-to-date information on urbanization in Sri Lanka. Urbanization is a complex process with significant implications for urban planning, infrastructure development, and resource allocation.

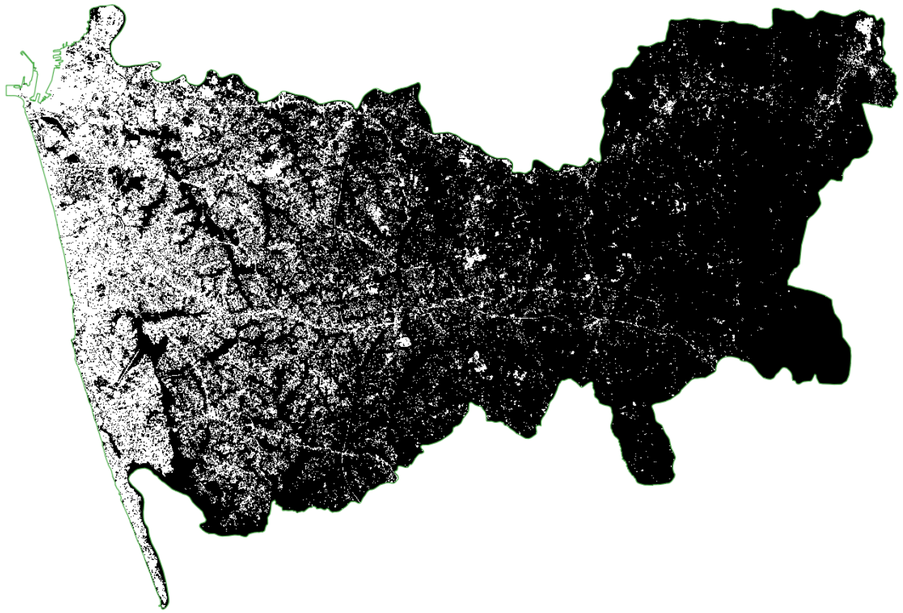
Data Sources:

- Satellite Imagery: The primary data source is high-resolution satellite imagery of Sri Lanka. This imagery captures spatial and spectral information about the landscape, including urban and non-urban areas.
- Remote Sensing Data: The study involves processing and analyzing large volumes of remote sensing data, which may include various spectral bands, resolution levels, and temporal information.

Methods:

Data collection, Image processing, Machine learning (to differentiate urban areas from non-urban areas within satellite imagery), Spectral analysis, Spatial analysis, validation

Expected Outputs: Urbanization maps, urbanization trends, Machine learning models, policy recommendations



10-meter built-up area map of the Colombo district for the year 2020. White denotes built-up; black denotes non-built-up

Ongoing project 3: Understanding spatial disparity of poverty: Leveraging Mobile Call Records and Remote Sensing Data

Problem/Context: How public and private data sources commonly available for low- and middle-income countries can provide novel insight into the spatial distribution of poverty.

Data Sources:

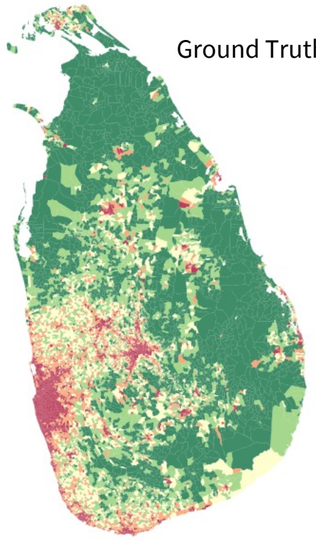
- Call detail records: Phone usage patterns, location/ mobility of the user, social network, handset type

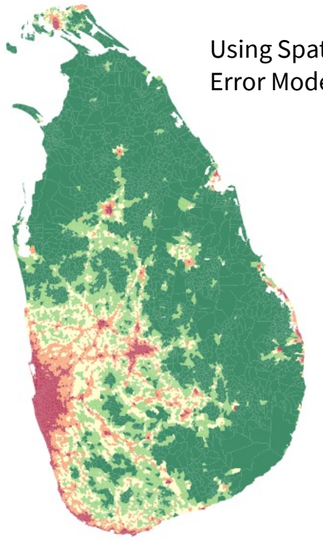
- Remote Sensing Data: Provide features such as Accessibility, Population, Climate, Night-time lights, Elevation, Vegetation, Distance to roads and waterways, Urban/rural, Protected area, landcover, and demographic features.

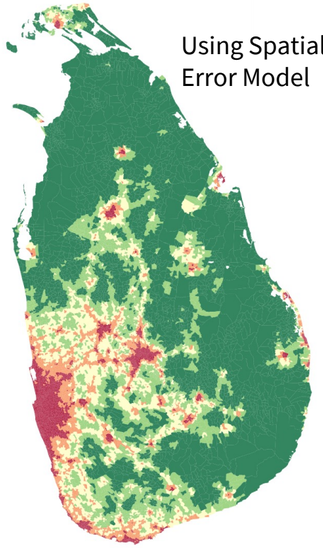
Methods:

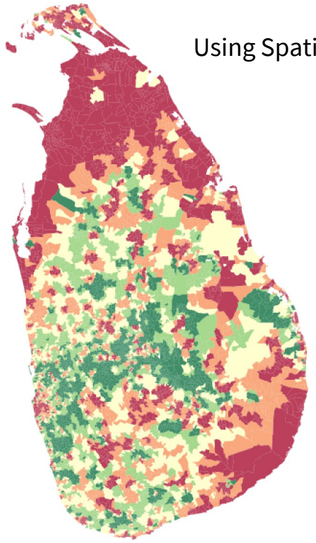
Data collection, Data integration, Machine learning (using Spatial error models, spatial lag error models, spatial durbin error models), Spectral analysis, Spatial analysis, validation

Expected Outputs: Poverty maps, insights into poverty disparities, Machine learning models, policy recommendations









Ongoing project 4: Global Index on Responsible AI

In this context, the term **“responsible AI”** is used here to refer to the **development, use, and governance of AI** in ways that **fully uphold human rights and democratic values** throughout the AI lifecycle and value chain (development, deployment, and maintenance).

A week-in-life at LIRNEasia

And the required skills

Research

Research Admin

Journal Clubs & Colloquiums

Biz Dev

A few of our past folks

Nisansa De Silva
Senior Lecturer - University of Moratuwa
PhD, University of Oregon, NLP

Lasantha Fernando
Research Fellow - LIRNEasia
PhD, University of Waterloo (reading), Distributed Systems

Danaja Maldeniya
Research Fellow - LIRNEasia
PhD, University of Michigan
Computational Social Science

Yudhanjaya Wijeratne
Research Fellow - LIRNEasia
Award winning author (Nebula-nominated and Gratiaen-winning), Self-taught

LIRNEasia
Pro-poor. Pro-market.

For more information
www.lirneasia.net

CDR Dataset Description

Table 1: Sample observations from a CDR dataset.

Call Direction	Subscriber ID	Other ID	Tower ID	Timestamp	Duration
Incoming	A	B	A024	2012/11/10 06:35:37	20 seconds
Outgoing	B	A	A375	2012/11/10 06:35:37	20 seconds
Outgoing	C	D	A129	2013/05/29 20:07:55	35 seconds
Incoming	E	F	A754	2013/03/22 00:03:22	153 seconds

Notes

- The **subscriber ID** and **other ID** are pseudonymized to protect the subscriber's privacy.
- If both parties of a call belong to the same operator, there will be two **reciprocal** records to represent the call. Eg: rows 1 and 2 in the table above.
- A subscriber will preferentially connect to the **nearest** tower.
- The latitude and longitude corresponding to each **tower ID** is known.
- The **timestamp** is accurate to a second.
- The **duration** is accurate to a second.