



# AI in Africa: The state and needs of the ecosystem

Diagnostic and solution set for data

March 2024



Sida



IDRC · CRDI



AI4D  
AFRICA

**G:ENESIS**  
25 YEARS OF UNLOCKING VALUE

DATA

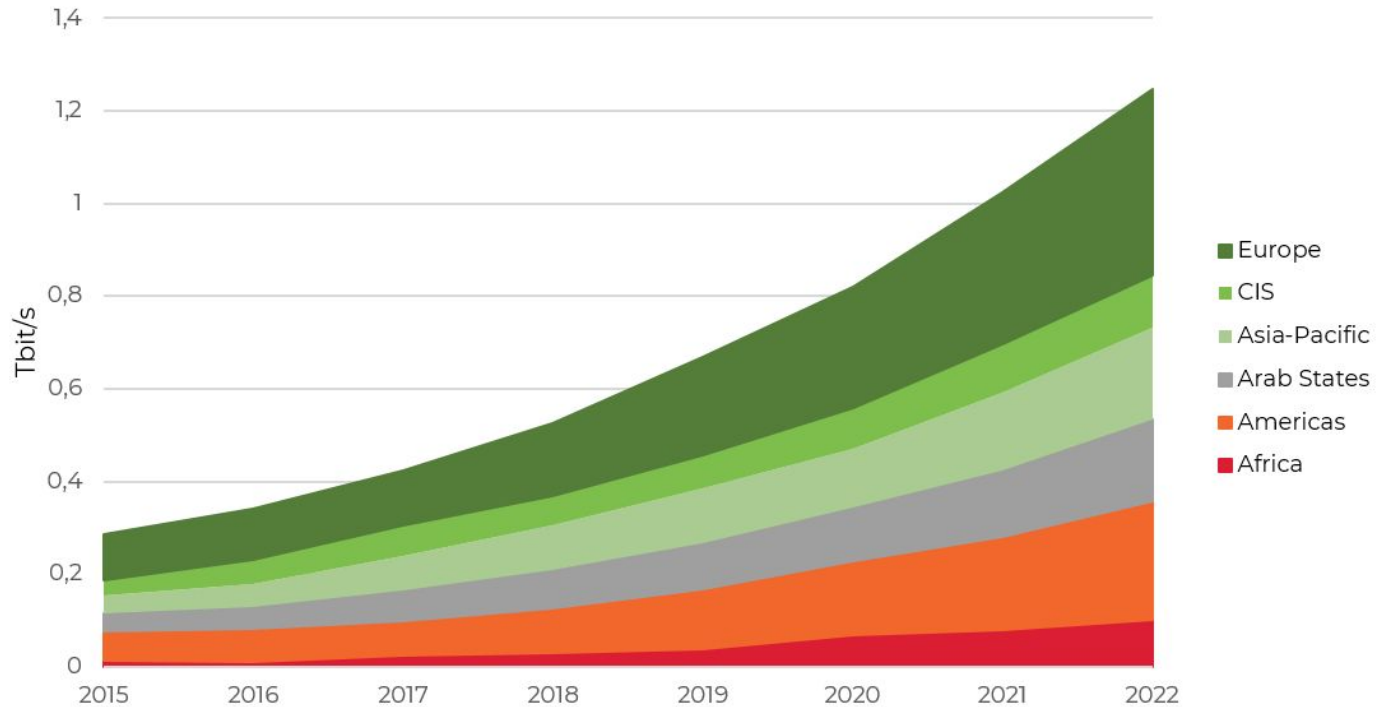


## WHAT IS COVERED IN THIS PACK

1. Data challenges across the continent
2. Channels for plugging the data gaps on the continent
3. Stylised case studies with associated lessons learnt and high level statistics on:
  - a. Developing tangible use cases to scale sector-related datasets
  - b. Using public data to achieve a quick win
  - c. Mapping non-standard smallholder farms to improve decision making
4. Quantifying the gap in data and associated investment ask
5. Suggested areas for intervention

# In a world of increasing data generation, Africa is a laggard

Africa lags behind in the number of citizens connected to the internet, and by proxy able to generate data.



International bandwidth usage (Tbit/s) per million internet users as a proxy for volume of data created per capita in different regions over time.

# One of the largest gaps in the availability of data is in language

'Brute force' alone is unlikely to achieve parity in data generation in African languages compared to Anglophone, given the extent of the data imbalance.

	Language (Nov '23)	% Share of Internet content in local language
Global Benchmarks	English	52.60%
	Hindi	0.1%
Top African Languages	Afrikaans	0.003%
	Twi	0.00195%
	Swahili	0.00135%
	Malagasy	0.00022%
	Bambara	0.00025%
	Venda	0.000115%
	Hausa	0.00011%
	Average with other African languages*	0.000999%
Sum of African Languages	0.01999%	

English  
53%

African  
0.02%

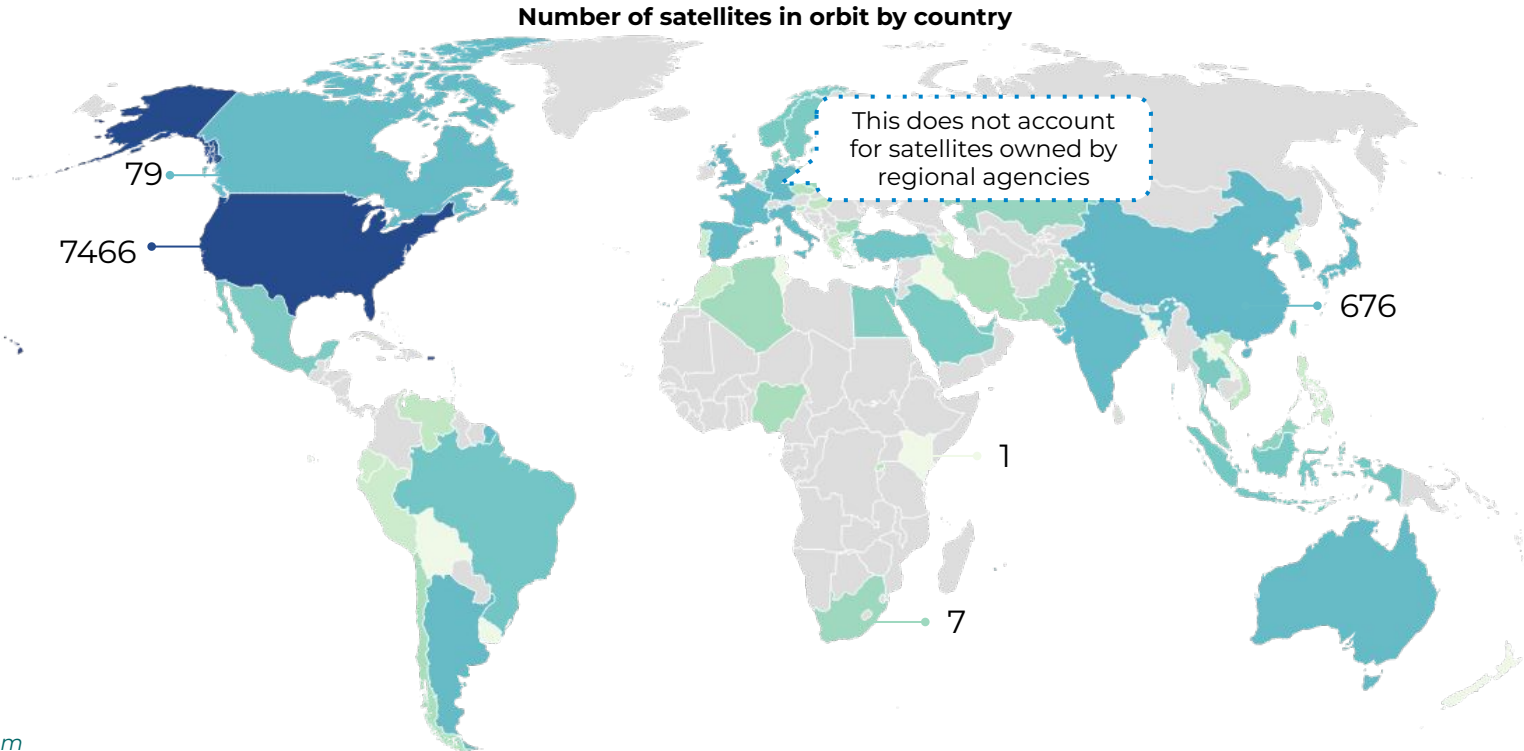
2,650 x more content in English than African Languages

Lacuna Fund and Masakhane are key interventions, but only scratch the surface

\* Afrikaans, Twi, Swahili, Bambara, Malagasy, Hausa, Venda, Haitian, Haitian Creole, Igbo, Luba-Katanga, Ndonga, Rundi, Tokelau, Tswana, Akan, Chichewa, Chewa, Nyanja, Fulah, Ganda, Masai

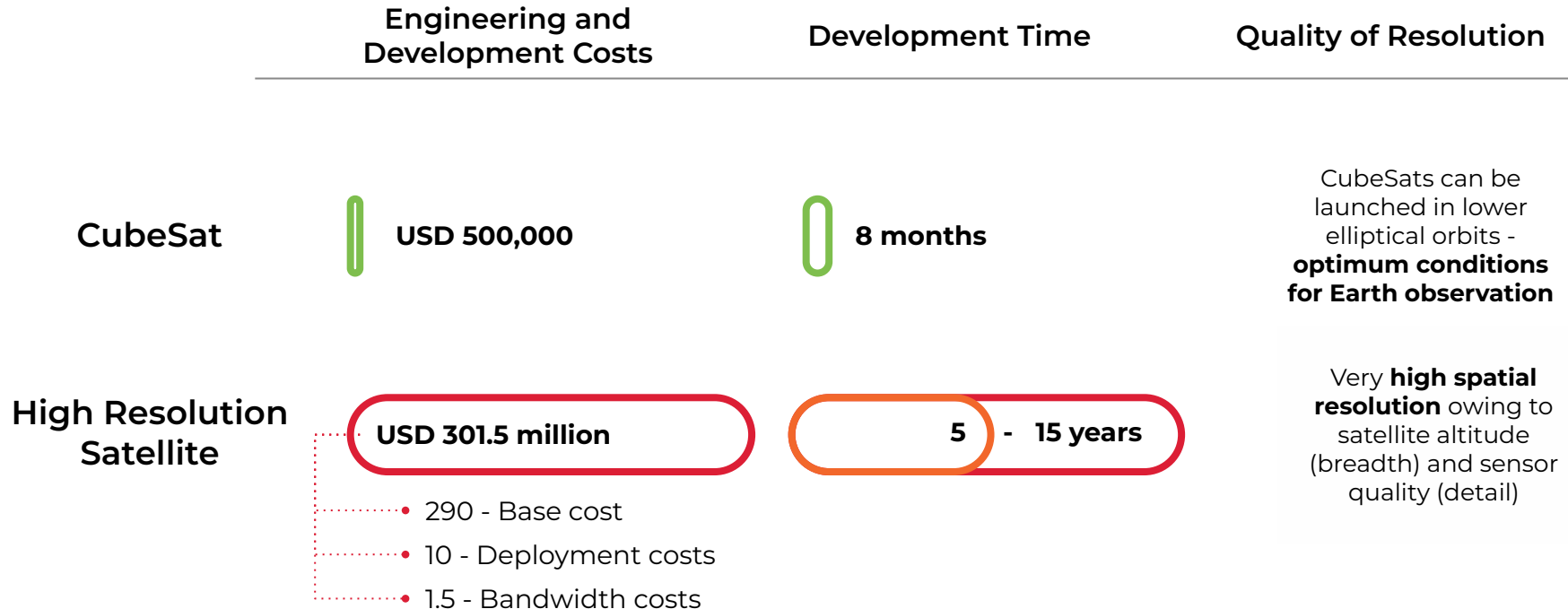
# Majority of African countries are without a satellite in orbit

High spatial resolution remote sensing enables real-time decision-making. However, the cost of the sophisticated sensors required for such quality imagery limits their accessibility, particularly in Africa.



# Before CubeSats, earth observation data required a compromise in data - either occasional high spatial resolution, or frequent low resolution

A CubeSat is a miniaturized satellite that is used to collect data and enable surveillance and monitoring and communications. It is a **low-cost, lightweight option with a standard design** that provides for high spatial resolution data collection.



# Data gaps result from constraints in the underlying environment

These constraints impact the development or scaling of business or government sector-related use cases; the quality or completeness of datasets; and the availability of data to build solutions in vernacular.

1

## CONNECTIVITY

The rate of internet penetration determines the rate of natural data generation

### Internet penetration

African average: 43.2%

Global average: 67.9%

2

## DIGITISATION

The maturity of country's digital economy determines the extent to which data can be digitised

### Digital development level

South Africa (49%) Kenya (37%) Nigeria (32%)  
84% (Denmark - highest)

3

## DIGITAL SKILLS

Digital literacy and proficiency in engaging with the digital economy, is essential to bridge the digital divide

### Individuals using the internet

Africa: 46%

Upper-middle-income: 71%

4

## INNOVATION & GLOBAL CONNECTIVITY

Digital trade allows for the the collective flow and creation of data; stimulating innovation and additional use cases

### Digital Services Trade Restrictiveness Index

South Africa (0.34)

UK (0.06) Canada (0.0) France (0.12)

5

## POLICY & REGULATION

Inter-country interoperability of information systems together with appropriate data protection regulation supports accessibility of data

### G5 Benchmark (for digital regulation)<sup>1</sup> (max: 100)

Africa region: 39.96

Europe region: 67.60

6

## BROAD & NARROW ICT INFRASTRUCTURE

To engage and reap the benefits of AI, reliable and affordable internet infrastructure (broad) is required to support the operation of devices reliant on bandwidth (narrow)

### Number of satellites in orbit

South Africa (7) Nigeria (5) Kenya (1)

UK (656) France (94) Canada (79)

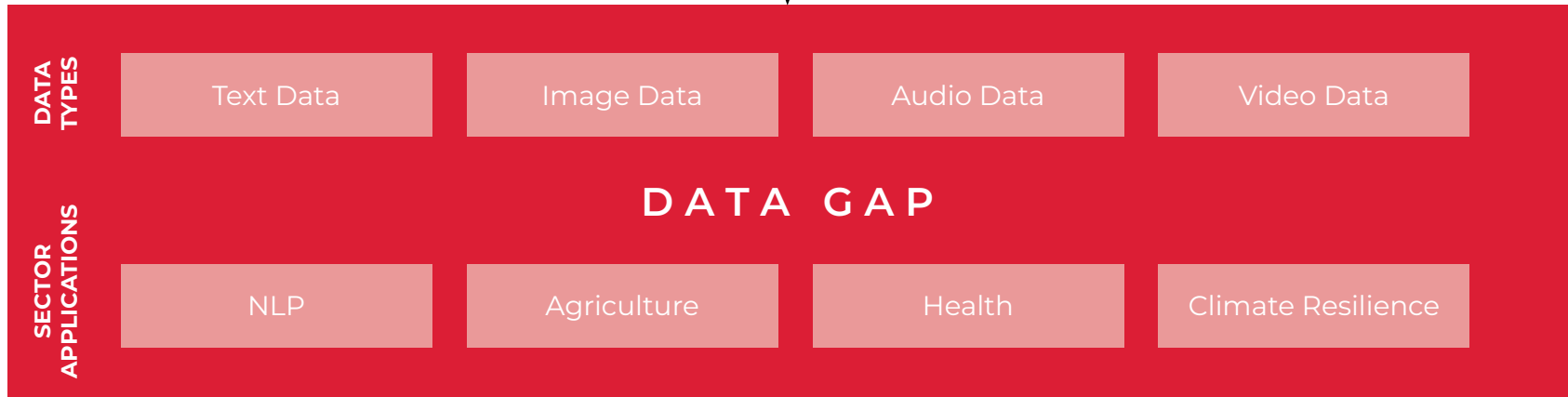
<sup>1</sup> Measures the state of collaborative digital regulation



# Two main channels exist for plugging the data gaps on the continent

---

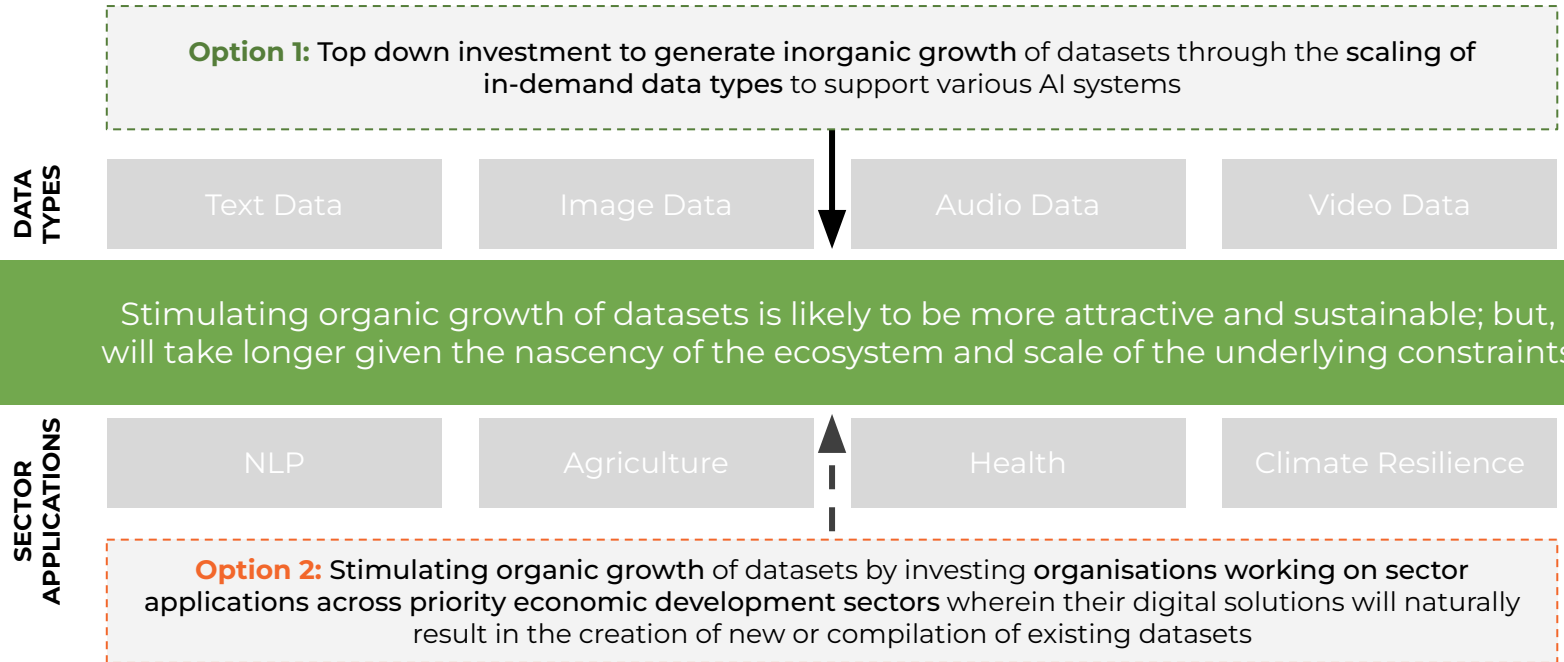
**Option 1:** Top down investment (inorganic growth) to generate data in in-demand data types and sectors to support various AI systems



**Option 2:** Bottom up investments (organic growth) by investing in organisations working on applications and in environments across priority development areas wherein their digital solutions will naturally result in the creation of new or compilation of existing datasets

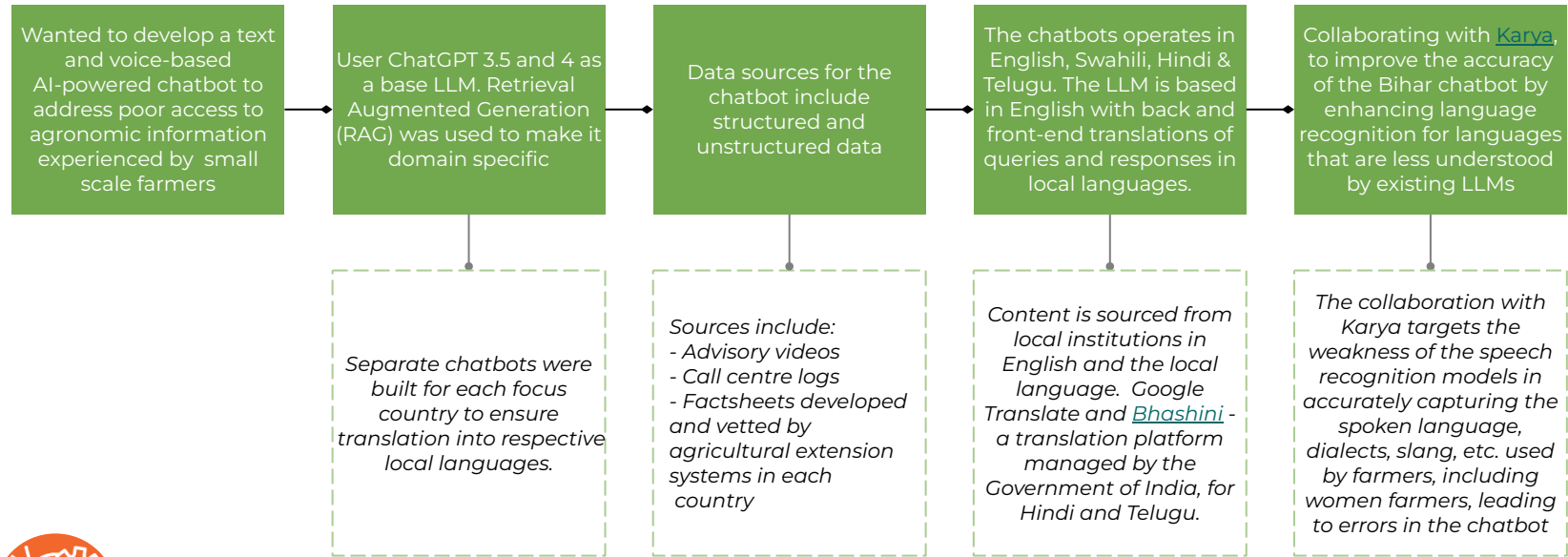
# Two main channels exist for plugging the data gaps on the continent

---



# Digital Green: Developing tangible use cases to scale sector-related datasets

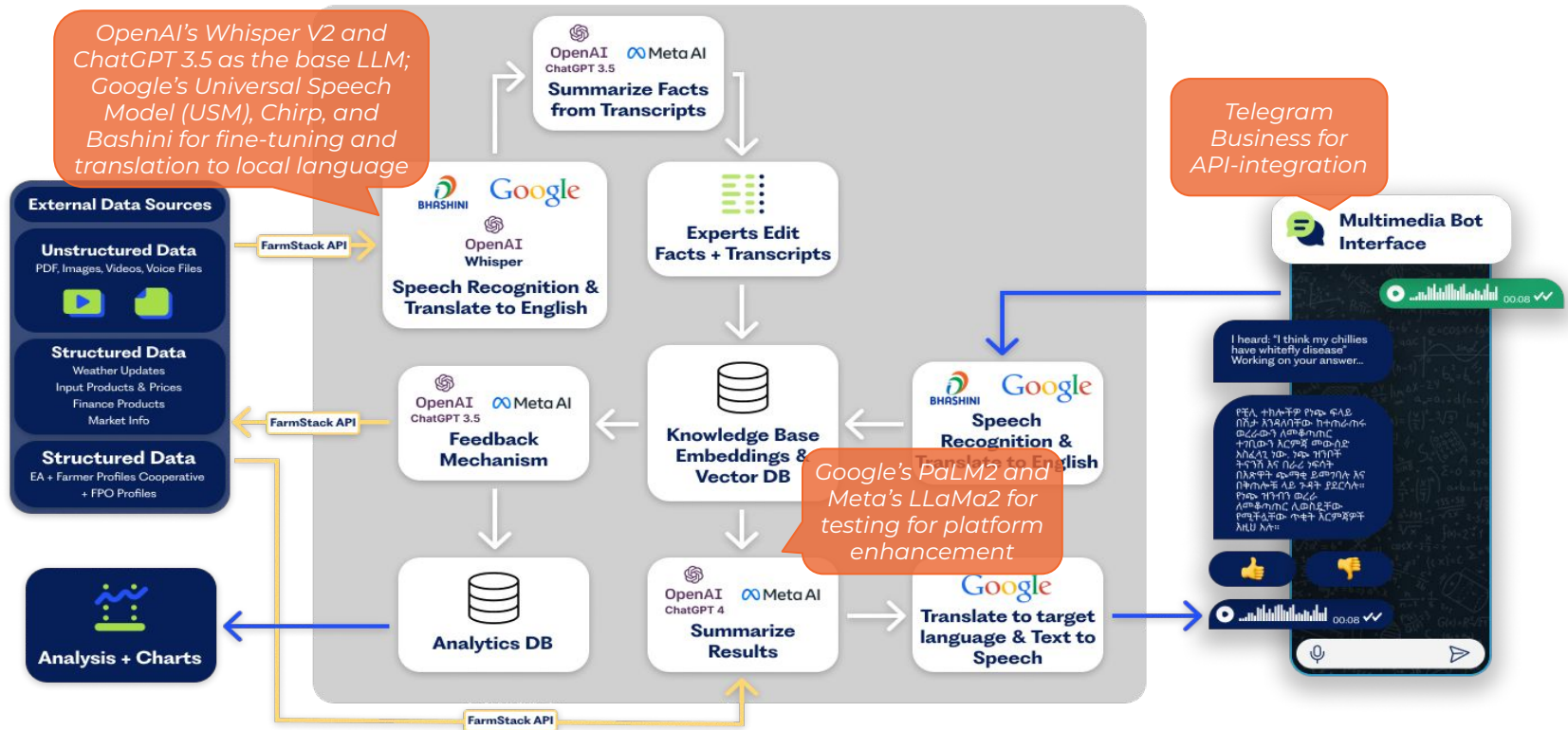
Digital Green's Farmer.CHAT solution leveraged multiple data types, technology services providers and open source platforms to build an AI-powered chatbot that supports agricultural extension workers.



The **successful business use case** of Farmer.CHAT has resulted in significant demand from public sector partners. Digital Green has raised **USD 30 million** to support the development and rollout of similar AI-powered agronomic chatbots for agricultural ministries in India, Kenya and Ethiopia - with the aim of reaching over 220 million farmers in each country.

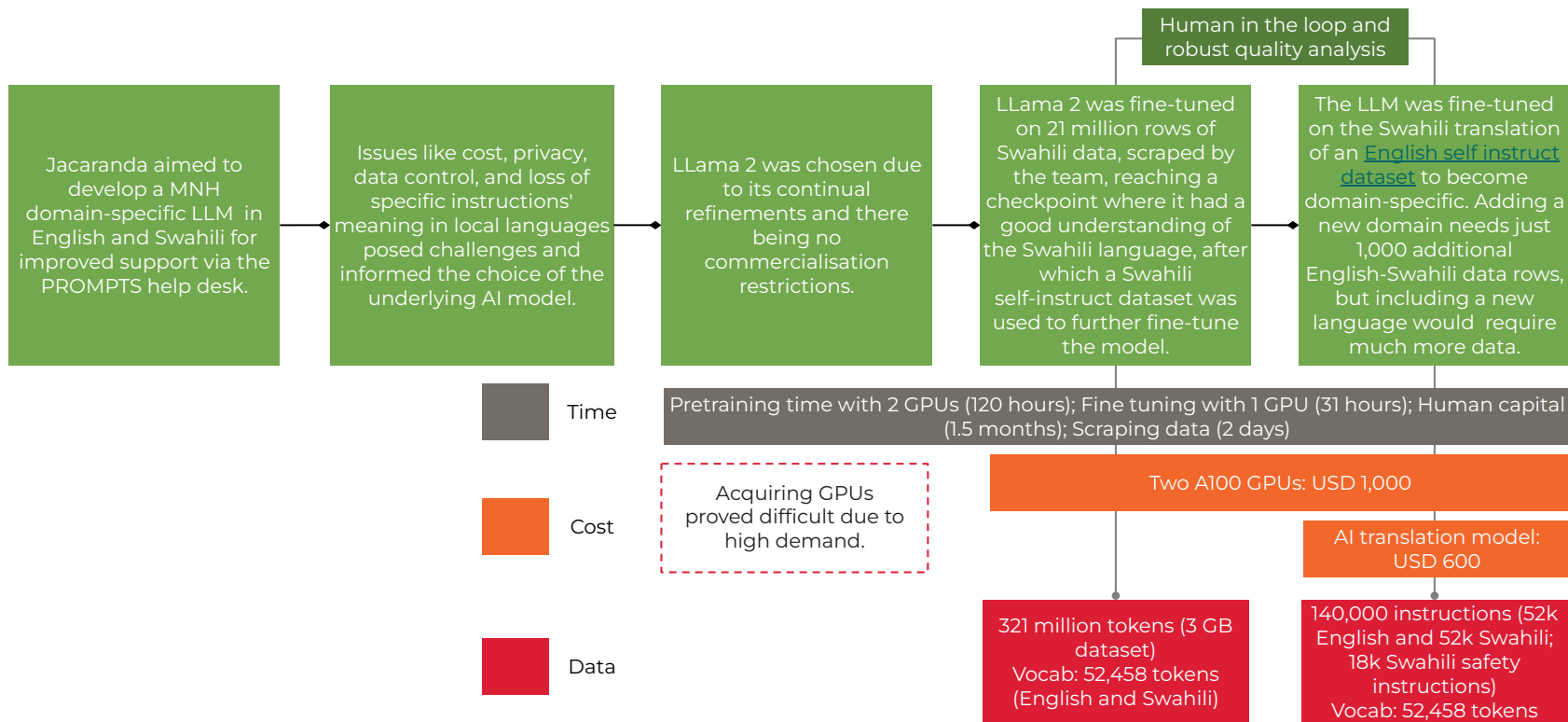
# The success of Farmer.CHAT is contingent on important partnerships with Technology Service Providers

Digital Green's Farmer.CHAT solution leveraged multiple data types, technology services providers and open source platforms to build an AI-powered chatbot that supports agricultural extension workers. The diagram below demonstrates the complexity of the solution, and the numerous areas where Big Tech can contribute



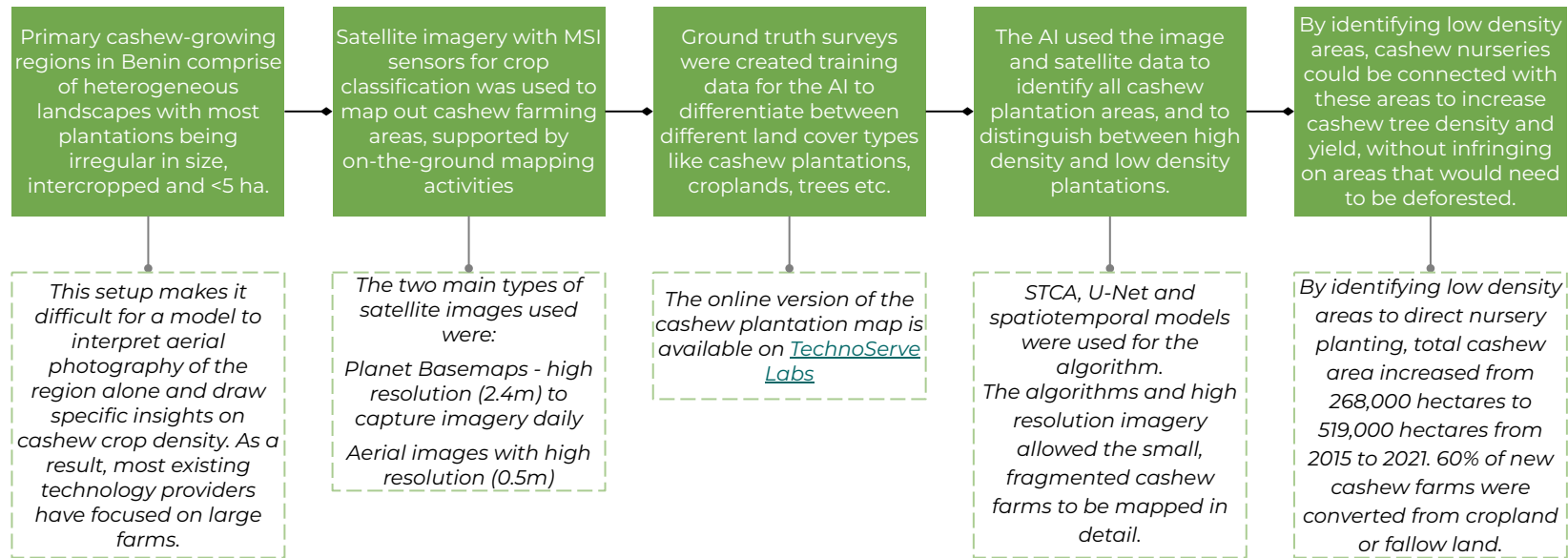
# Jacaranda Health: using public data to achieve a quick win

Jacaranda Health developed UlizaLlama to enhance accuracy and context-specificity of its automated responses on the maternal and newborn health (MNH) digital platform, PROMPTS. The project showcases the feasibility of scaling similar applications and contributes to broadening the African NLP ecosystem.



# TechnoServe Labs: mapping non-standard smallholder farms to improve decision making

Accurate crop mapping is vital for securing livelihoods, yet current models struggle to identify nonstandard smallholder farms predominant in Africa. According to Technoserve, **USD 1.5 million, coupled with 5-7 software developers**, would be the **investment required per crop type**.



Access to GIS and satellite imagery will become increasingly important in the agriculture and climate resilience sectors, particularly as regulations aim to protect deforestation efforts present new challenges for smallholder farmers to map out and certify their farming areas to ensure compliance with export standards.

# Quantifying the gap in data and associated investment ask

The investment gap for data is estimated using funding amounts provided to grantees in priority sectors by Lacuna Fund. Funding provided by Lacuna aims to address the gap in the lack of openly accessible datasets in African languages.

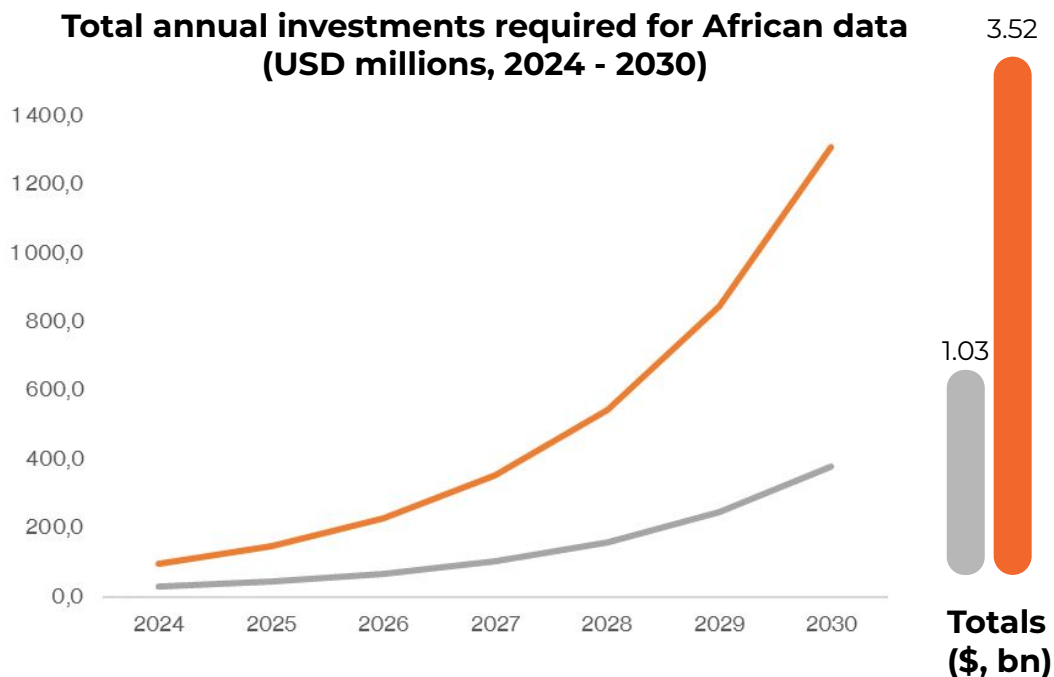
Inputs: Lacuna Fund Data			Range calculated using Lacuna Fund data	Future Value of Lower Bound Total to 2030	Multiplied to account for demand outside of Lacuna Fund Grantees	Annual CubeSat launch per region	Total Investment required by 2030
Illustrative Example: Agriculture sector							
Total Amount Funded (2020 - 2022)	2.2 million	Lower Bound	Annual shortlisted - annual funded	Grown from 2024 to 2030 using average growth rate of number of AI startups	3.6 as the calculated multiplier	USD 500,000 per CubeSat x Four regions	Total
Ave. annual	0.73 million						
Total Amount Shortlisted (2020 - 2022)	13.5 million	Upper Bound	Annual requested - Annual funded	Grown from 2024 to 2030 using average growth rate of number of AI startups	3.6 as the calculated multiplier	USD 500,000 per CubeSat x Four regions	Total
Ave. annual	4.5 million						
Total Amount Requested (2020 - 2022)	36 million						
Ave. annual	12 million						

Source: Lacuna Fund\*  
(\*Compiled in 2023 for Grantee Cohorts in 2020-2022)

See Methodology [here](#)

# Quantifying the gap in data and associated investment ask

Quantifying the data gap across the continent for all sectors is a complex task - it can be sized in orders of magnitude of error. The estimated investment required ranges from just over **\$1 billion to \$3.5 billion**. Based on Lacuna Fund data. Agricultural and Climate data make the bulk (64%) of necessary investments.



Effective partnerships, efficient scalability planning, and community engagement can help reduce the overall cost and investment needed for successful AI implementation in Africa.

Ensuring effective data sharing is paramount to ensuring that investments are maximised.



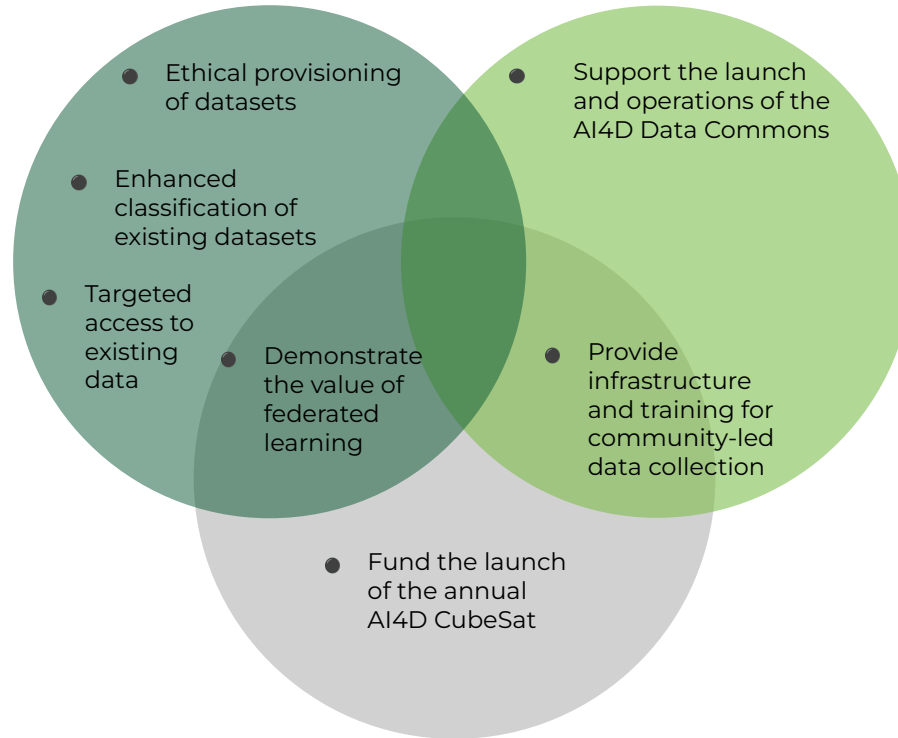


# WHERE TO INTERVENE

# Data Quick Wins

---


**Scale up** methods of unlocking safe access to new sources of data



**Scale out** necessary infrastructure to support AI4D Data Commons and AI Language Hub

**Empower** the organic creation of data

# Support the launch and operations of the AI4D Data Commons



Data sharing and integration can have a profound impact on private and public sector applications. However, AI use cases that demonstrate the effective integration of data are limited due to data scarcity and limited access to essential resources like GPUs or licensed datasets.

The launch of the AI4D Data Commons can provide a central point for data collection and sharing. The AI4D Data Commons would be a cloud-based platform that enables community members to submit, analyse, store, and share data; assisting with matching research requests with available datasets.

Big Tech can invest in a number of ways. Firstly, Google Cloud could provide free storage services. Secondly, through the provision of translation engines such as Google's Chirp to expand the usability of LLMs into different African languages. Thirdly, by allocating more training hours to fine tune existing models like USM (Google AI) and MMS (META) using African language data.

# Enhancing existing data collection processes



Limited access to data at a community-level can hinder accurate representation in datasets and stifle the development of community-level representative datasets.

Community-Led Monitoring (CLM) is a participatory approach to collecting, analysing and using data, and involves community members collecting data from their community across healthcare, education, clean water and other essential infrastructure priorities.

Internet Service Providers like MTN or Safaricom can provide CLM leaders with ICT infrastructure (mobile devices, tablets, data bundles) and digital literacy onboarding to allow these community members to capture data digitally. Players like Opera or Google can provide access to online data collection tools such as survey management platforms to execute data collection processes. This data can be stored on the AI4D Data Commons, prompting further research.



# Unlocking safe access to existing and new sources of data



The scarcity of high-quality and diverse text data in African languages hinders the development of effective African AI and NLP models.

Enhanced data classification, as offered by Google's Dataset Search and potentially Amazon's AWS Data Exchange, can categorise and index datasets as 'relevant to Africa' making them discoverable through keyword searches. Players like Google and Uber can also offer existing support services like help desks, and FAQs in key African languages, generating valuable, machine readable data in African languages.

ISPs like Vodafone, Telefonica, Airtel, and others can enhance local language NLP data by ethically making anonymised call center data open source to the AI4D Data Commons. Organisations like GSMA can also advocate for its members to provide similar anonymised data available.

# Demonstrating the value of federated learning



Federated learning is a privacy preserving approach to developing AI models where the underlying data for training is not shared. The AI4D network is regionally distributed and well positioned to experiment with and demonstrate the value of federated learning.

Big Tech firms such as MNOs can contribute to a federated learning Proof of Concept (PoC). They can provide secure data access and infrastructure for select AI4D innovation lab members to experiment with and develop federated learning models in areas such as credit extension, customer support, or others.

These PoC's will demonstrate the opportunity of federated learning to participant firms, positively impact Africans should the PoC be well designed ,and provide the AI4D network with an opportunity to 'specialise' in an area that has impact and that they are well positioned to lead on.

# Addressing barriers preventing access to quality data



Limited access to high-resolution GIS or satellite imagery due to funding constraints can lead to incomplete datasets, especially in sectors like agriculture. Smallholder farms' undocumented perimeters inhibit the feasibility of building data-driven solutions to their challenges, leading to reliance on drone technology for data collection, which is expensive and lacks continuous data access.

The launch of four annual AI4D CubeSats, costing around \$500,000 and offering faster deployment and high-resolution imagery, could enable continuous and up-to-date monitoring of expansive regions, supporting accurate national planning decisions for sectors including agriculture, water and sanitation. Further, overlaying this data with additional information including weather forecasts can allow for more proactive climate change policy. The data collected from the CubeSat can be accessed via the AI4D Data Commons, and stored similarly.

An aerial, top-down view of a road intersection. The image is dark and semi-transparent, serving as a background. In the center, the word "APPENDIX" is written in white, uppercase, sans-serif font. The background shows a multi-lane road with several cars, a parking lot with many cars, and some buildings. The overall scene is a complex urban or suburban environment.

# APPENDIX



# Methodology

---

## Slide 16: Quantifying the gap in data and associated investment ask (1/2)

- Total amount funded, shortlisted and requested is provided by Lacuna Fund for 2020-2022 grantees (see table below, average annual investment by sector)
- Lower band is calculated as *Total Amount Funded* subtracted from *Total Amount Shortlisted* (\$7.6 million in 2024)
- Upper band is calculated as *Total Amount Funded* subtracted from *Total Amount Requested* (\$26.17 in 2024)
- These bands are then multiplied by the average annual growth rate of AI startups (55%) and the calculated 3.6 multiplier.
- This is summed with the annual investments in a CubeSat per region.

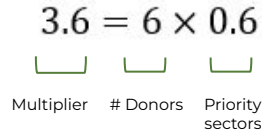
	Agri	NLP	Equity and Health	Climate	<b>Total (millions)</b>
Funded	0,73	0,70	0,57	1,17	<b>3,17</b>
Shortlisted	4,50	2,20	1,17	2,87	<b>10,73</b>
Requested	12,00	7,67	3,00	6,67	<b>29,33</b>

# Methodology

---

## Slide 16: Quantifying the gap in data and associated investment ask (2/2)

- This multiplier figure is calculated as the number of additional key donor organisations<sup>1</sup> providing grant funding for datasets multiplied by the weighted average of the number of priority sectors<sup>2</sup>

$$3.6 = 6 \times 0.6$$


Multiplier   # Donors   Priority sectors

1. This includes: [BMFG Grant Challenge for “Catalyzing Equitable AI Use”](#); The [Rockefeller Foundation](#); [Science Granting Councils Initiative in sub-Saharan Africa](#); [UNICEF](#); [GIZ](#); and the [Wellcome Trust](#)

2. These sectors include: NLP; Agriculture; Health and Equity; Climate Resilience; Education and Financial Inclusion