BUILDING LOCAL LANGUAGES AI INFRASTRUCTURE TO ENABLE VERNACULAR CAPABLE DIGITAL SOLUTIONS

Dunstan Matekenya, PhD
Data Scientist
The WBG, Washington DC

ICTAM AGM
November 25, 2023

B.Ed (Maths)

UNIVERSITY OF MALAWI

2001-2006

Statistician — GIS-lead

NSO, Malawi

Government of Malawi

NATIONAL STATISTICAL OFFICE

2007-2008 — 2009-2013

- PhD research- Big Data
- Freelance Data Scientist
- Part-time ML Engineer- Startup

東京大学 THE UNIVERSITY OF TOKYO

2013-2016

Data Scientist, WBG

- Mobile phone data and development
- Machine learning and other ad hoc DS work
- Teaching DS with AIMS in Rwanda
- NLP and AI for local languages in Malawi

THE WORLD BANK

2017-Present

TALK OUTLINE

1. VERNACULAR CAPABILITY IN DIGITAL SOLUTIONS

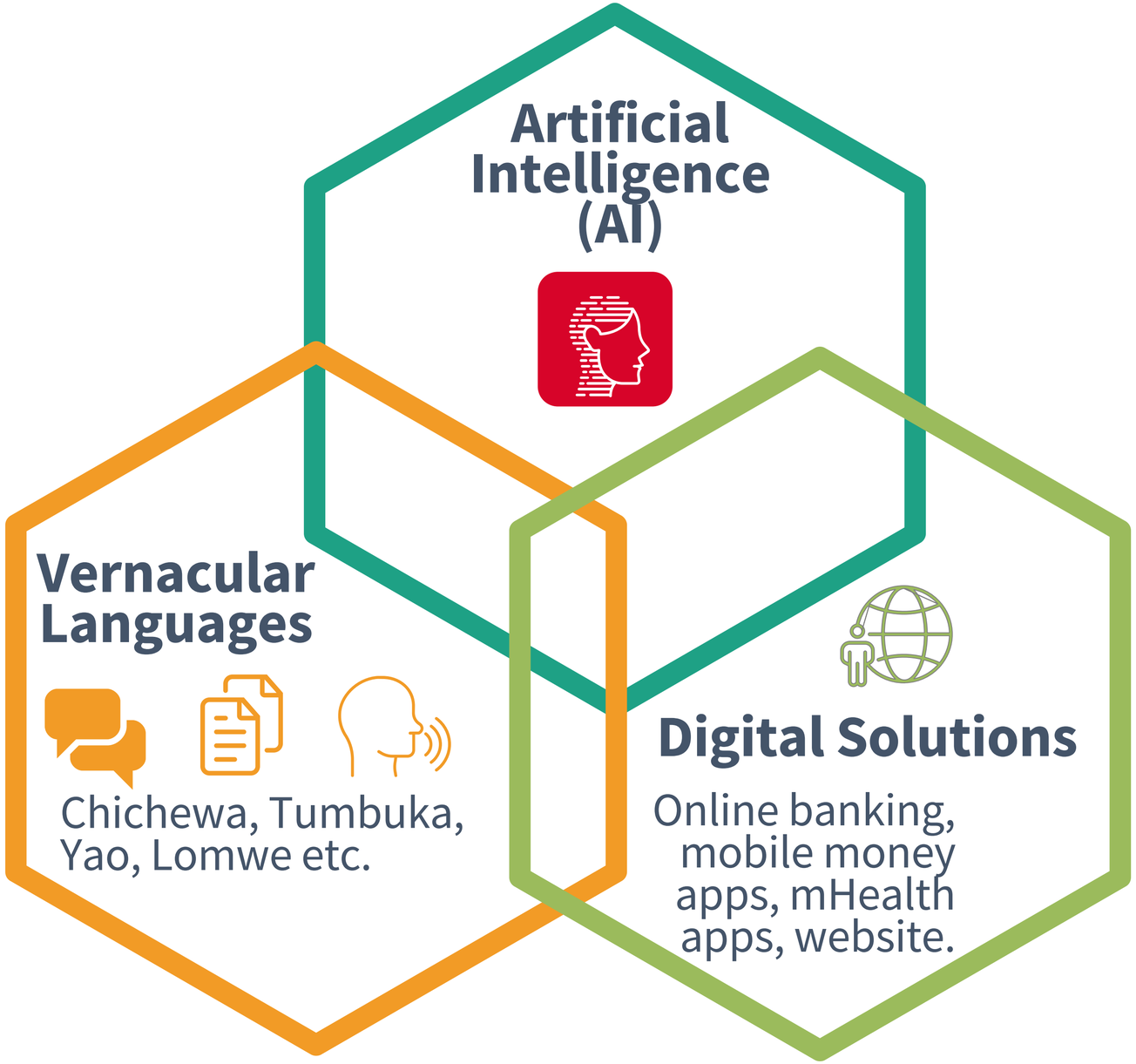
2. AI AND VERNACULAR LANGUAGES IN MALAWI

3. NLP FOR LOCAL LANGUAGES

4. ROADMAP TO BUILDING AI INFRASTRUCTURE

AI, VERNACULAR LANGUAGES AND DIGITAL SOLUTIONS

The Three Main Focus of this Talk



1. VERNACULAR CAPABILITY IN DIGITAL SOLUTIONS IN MALAWI

1. What do we mean by vernacular language capability?
2. Why is this important?
3. Problems associated with lack of vernacular language support?

WHEN DO WE SAY APPS ARE VERNACULAR CAPABLE?

Selected Characteristics of Digital Capable Solutions

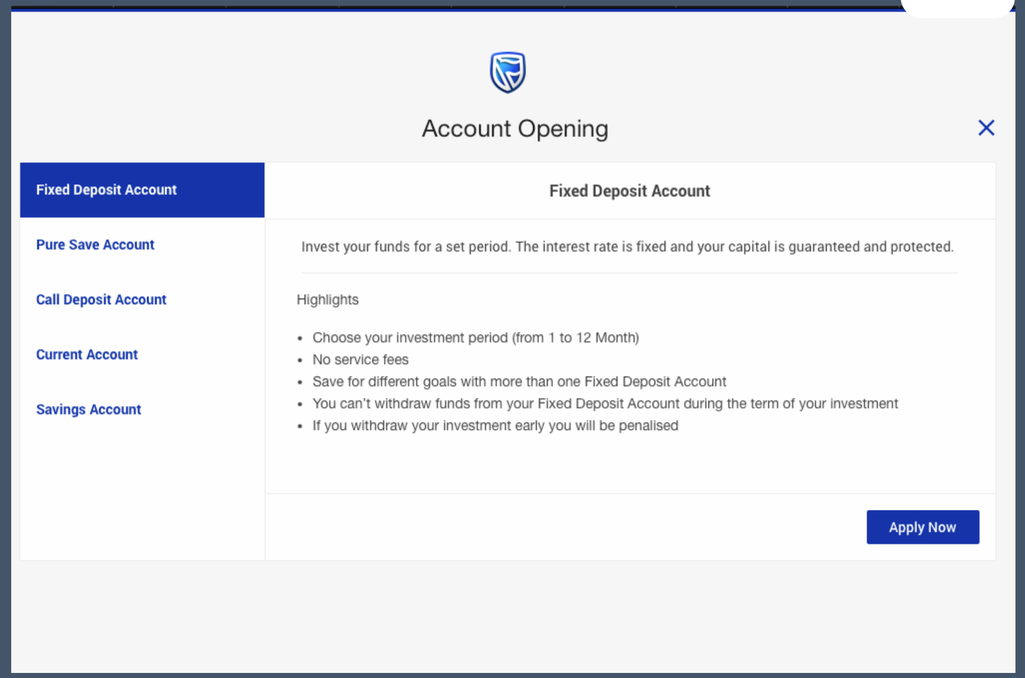
	Mobile Apps	Web apps	Devices
AUDIO	• Accept as input • Process, understand • Use for downstream tasks • Use for output (e.g., Siri, Alexa) • Display, show information	• Accept as input • Process, understand • Use for downstream tasks • Use for output • Display, show information	• Accept as input • Process, understand • Use for downstream tasks • Use for output • Display, show information
CHATS			
DOCS			
VIDEO			
IMAGE			

ACCESSING CONTENT, ONLINE SERVICES IN LOCAL LANGAUE?

Selected Characteristics of Digital Capable Solutions

Language options

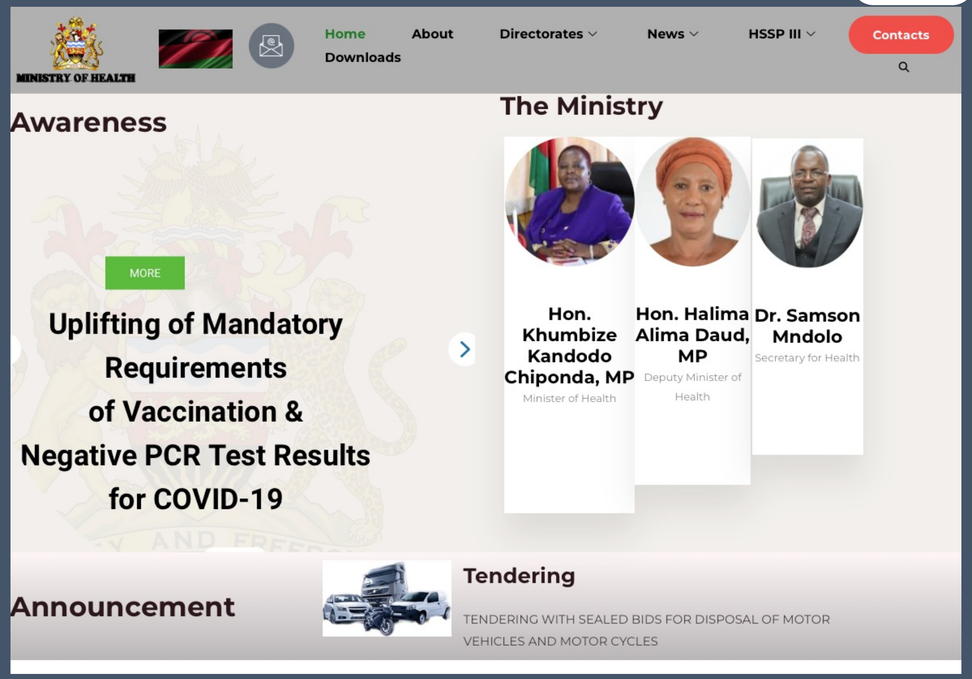
1. English
2. Chichewa
3. Tumbuka



A Mr. Phiri accessing an online banking website in CHICHEWA

Language options

1. English
2. Chichewa
3. Yao



A chemusa accessing Ministry of health website in YAO

VERNACULAR LANGUAGE CAPABILITIES CAN MEAN SEVERAL THINGS

Users can choose to use their own language when they visit a website for crucial services such as banking, agriculture trading

Users can use their own language to interact with apps (e.g., mHealth, banking, ecommerce) as well as control devices (e.g., mobile phone, computers, IoT devices and other smart devices)

When needed (e.g., due to illiteracy, sickness or mere preference) users can interact with devices using alternative inputs such as voice in their mother

Organizations have access to language technology tools (e.g., **MT translation** from English to Yao; **automatic transcription** of audio interviews in Tumbuka; **summarization** of documents in Chichewa, **sentiment analysis** of user comments in Tumbuka

Systems and applications can process local language content and integrate with other structured datasets .

VERNACULAR LANGUAGE CAPABILITIES

A summary with Technical Terminology

AUTOMATIC SPEECH RECOGNITION(ASR)

1. **Speech To Text (STT)**

2. **Text to Speech (STT)**

Direct translation from speech (e.g., Tumbuka speech to English text), wake words detection (e.g., for mHealth apps, detect key words from a voice note which indicate severity)

MACHINE TRANSLATION (MT)

1. **English to Chichewa, Tumbuka, Lomwe etc**
2. **Chichewa to Tumbuka, Lomwe**

Ability to provide accurate translations which can be used by organizations. Go further to topic specific translations as well

OTHER NLP & NLU

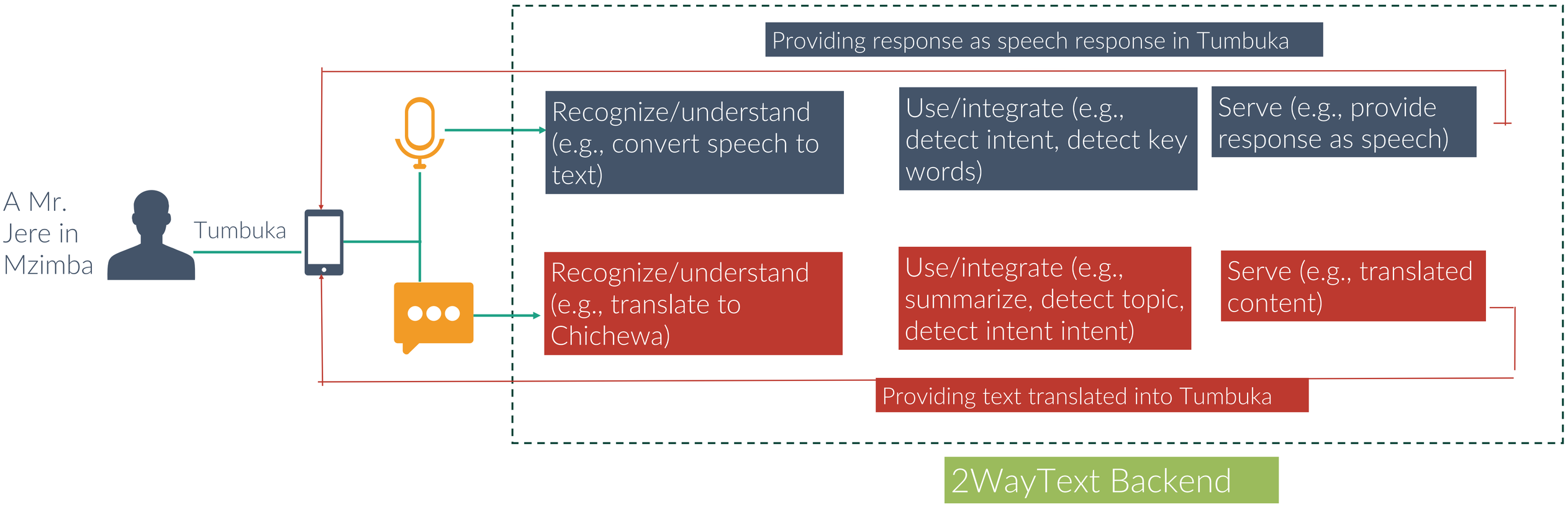
1. **Language generation (chatGPT style)**
2. **Text summarization, classification**

3. **Conversational systems**

Numerous other Natural Language (NLP) and Natural Language Understanding (NLU) capabilities which are useful

CONSIDER A HEALTH APP- 2WayText

The 2WayText app connects HIV/AIDs patients on ART with providers and enables communication such as reminders for taking drugs, motivations to keep going, appointment setup and reminders, patients messaging providers about their issues.



DOESN'T EVERY (MOST) MALAWIAN UNDERSTAND THESE TWO TOP LANGUAGES WE USE?

SKILL	ENGLISH	CHICHEWA
Read	✗	?
Write	✗	?
Speak	✗	?
Fully comprehend and comfortable	✗	?

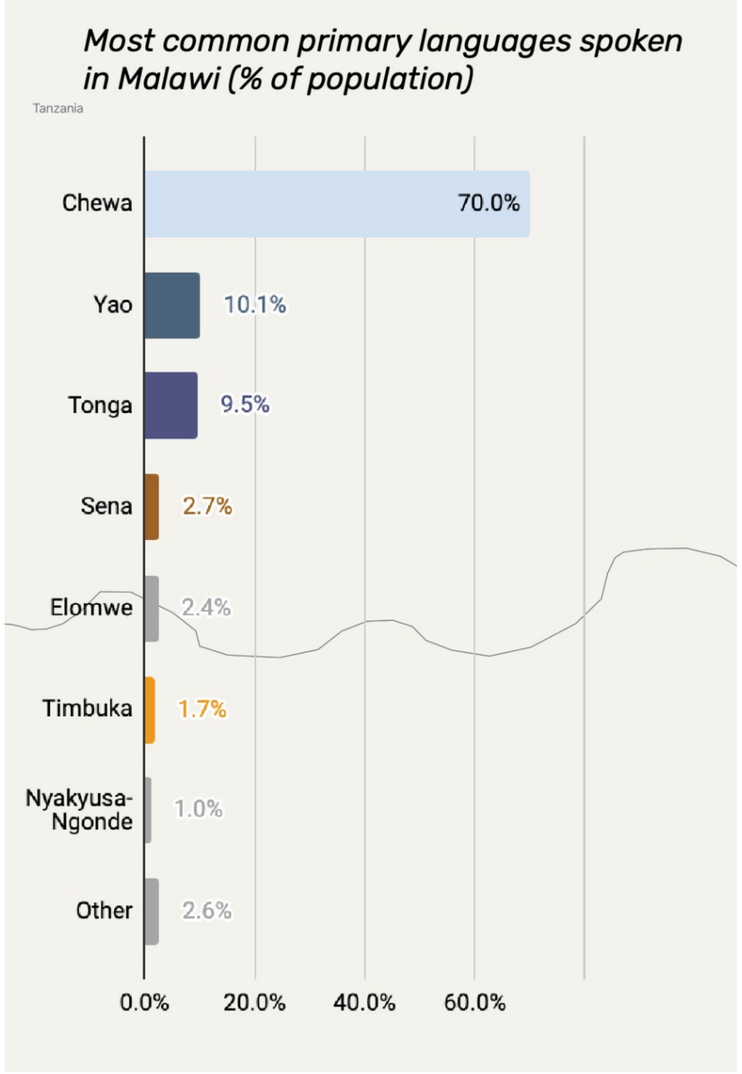
What’s your best guess of how many Malawians can speak English?

“Although English is the official language, the 2008 Census reports that only 26 percent of the population above the age of 14 is able to speak English. ”

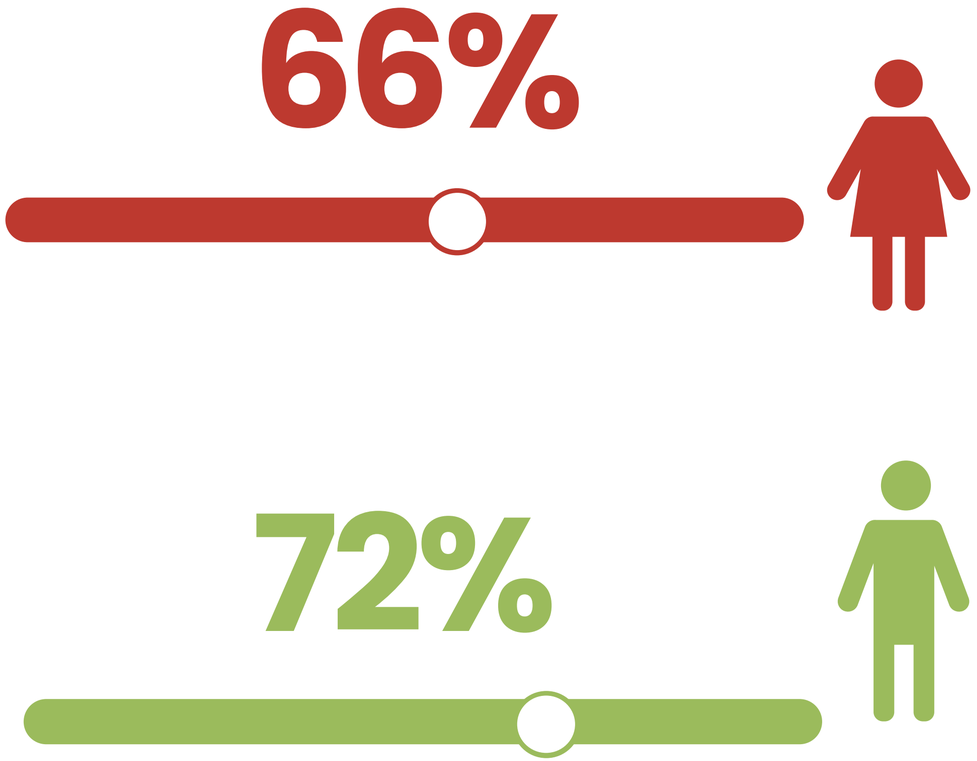
https://translatorswithoutborders.org/language-data-for-malawi

MORE LANGUAGE STATISTICS

Speakers by language



Literacy Rates



1. Source: https://translatorswithoutborders.org/language-data-for-malawi

IN SUMMARY, AS THE GUEST OF HONOR NOTED YESTERDAY, WHAT WE HAVE IS A HUGE LANGUAGE BARRIER, WHICH LEADS TO SERIOUS CHALLENGES IN TECHNOLOGY ADOPTION & SCALABILITY WIDENING THE DIGITAL DIVIDE FURTHER

LANGUAGE BARRIER CHALLENGES

Consequences of on Digital Solutions

Failure to Scale

If you consider mHealth apps as example, they cannot scale nationwide if there is always a need to have human read/listen/respond to messages .

Non-user Friendly Solutions

- People are more comfortable and willing to use technology when its in their vernacular.
- Due to failure to process and work with speech, users are forced to use text input even when they are incapable

Digital Solution Gaps

No digital solutions to allow people to do the following:

1. Translate from English to other local languages other than Chichewa (offered by Google and Bing)
2. Transcribe speech in local languages (e.g., from Yao, Tumbuka) to text
3. No Apps to automatically translate from speech in local langueges to speech or text in foreign languages

Non-smart solutions

Lack of smart/intelligent solutions which can incorporate text, audio, image data with local language content.

2. AI AND VERNACULAR LANGUAGES IN MALAWI

1 | Artificial Intelligence (AI) and how it Works

2 | AI vs. vernacular languages

3 | Available support for vernacular languages ?

ARTIFICIAL INTELLIGENCE(AI)

A simplified look at major areas

GENERATIVE AI

Creating new and original content (e.g., images, text) by learning from existing data patterns using LLMs

chatGPT

NLP & NLU

Large large models (LLMs), Machine translation (MT), summarization, conversation, question answering, .

Google Translate

SPEECH RECOGNITION

Convert speech to text, text to speech, classify audios, understand intent from speech .

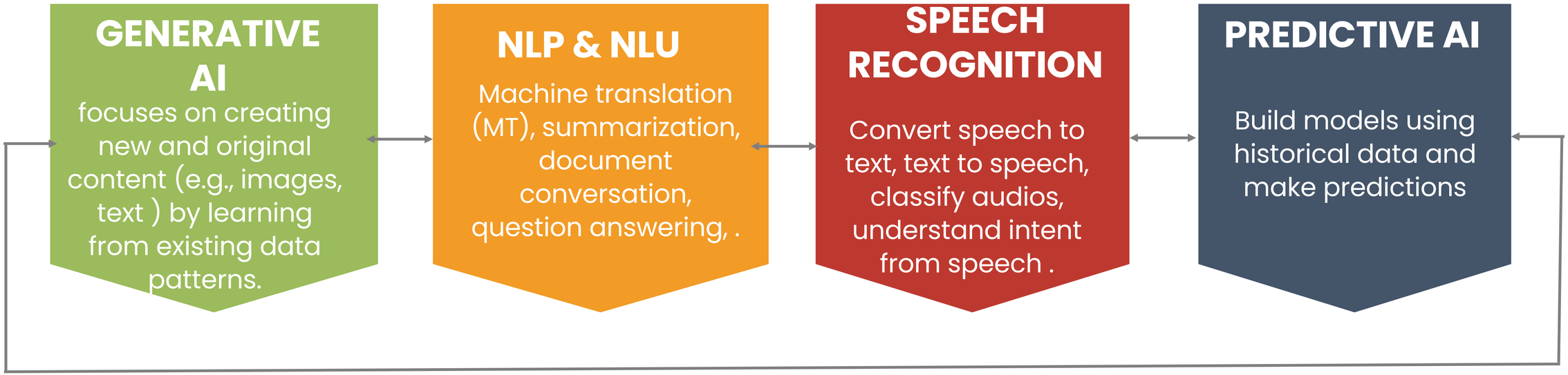
Siri, Alexa

PREDICTIVE AI

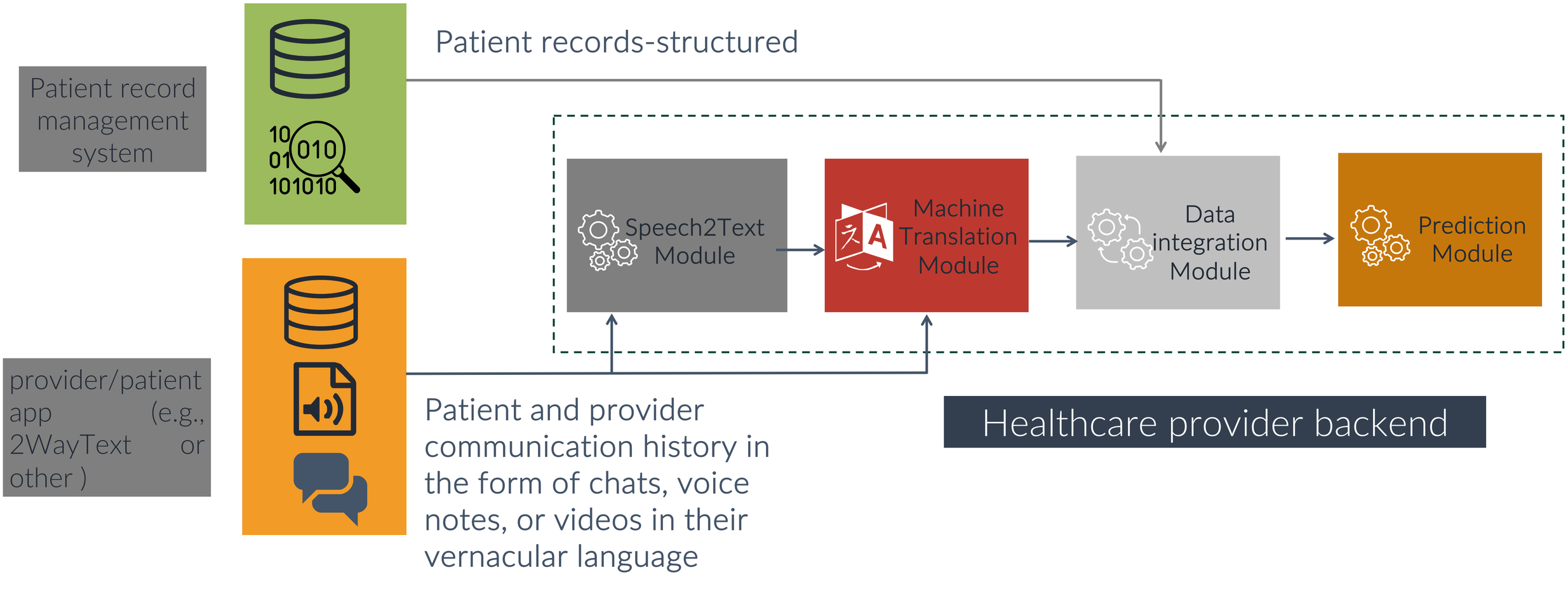
Build models using historical data and make predictions

A commercial bank using models to decide whether to give a loan

Although the branches of AI were presented in a hierarchical manner, in practice, they are deep connections across these sub-fields



Consider our example of a patient on ART treatment using an app- what if the provider wanted to predict likelihood of the patient abandoning the treatment?



Main Components of an AI Model

Data, Model and Documentation

DATA

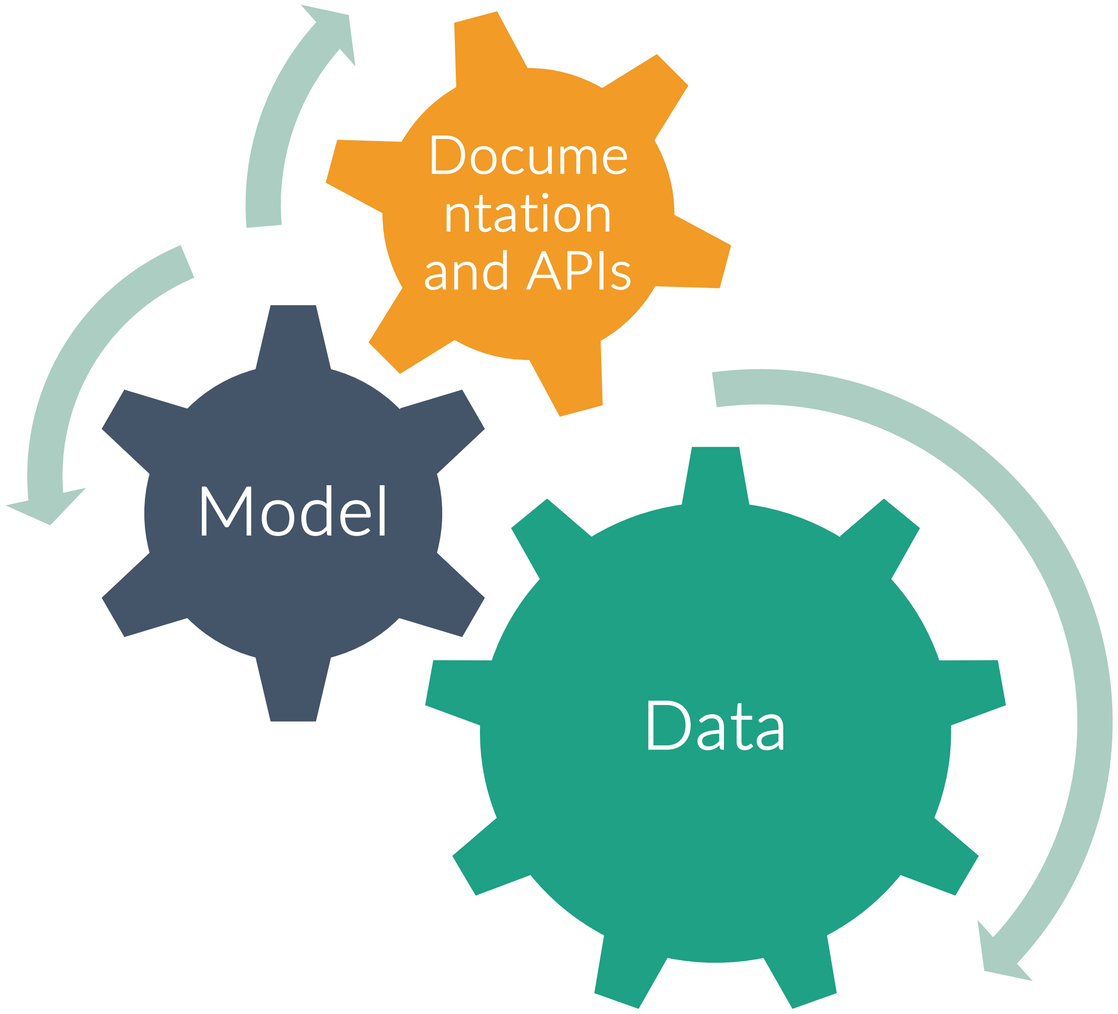
AI models learn from existing/historical data

MODEL

The mathematical formulation which defines how to learn from the data

DOCUMENTATION AND APIs

Provides information on how to use the model and model characteristics, metadata



HOW WOULD YOU BUILD A MODEL TO CONVERT TUMBUKA SPEECH INTO TEXT (AKA SPEECH2TEXT)

High Level Overview of the Main Steps

1

COLLECT DATA

Record people speaking in Tumbuka or gather existing recordings

2

ANNOTATE DATA

For each recording, provide a transcript matching the spoken words

3

TRAIN MODEL

Choose whether to train an ASR model from scratch or fine-tune existing models

4

EVALUATE MODEL

Select a metric (e.g., WER) to determine how well the model is doing

5

DOCUMENT AND USE MODEL

Use the model in apps to automatically transcribe Tumbuka speech

WHICH OF OUR VERNACULAR LANGUAGES HAVE AI SUPPORT?

Vernacular Language	Machine Translation (en->lan)	Speech Recognition	Generative capabilities	summarization, question answering etc.
Chichewa	☑	⚗	⚗	⚗
Yao	?	✗	✗	✗
Tumbuka	?	✗	✗	✗
Sena	✗	✗	✗	✗
Lomwe	✗	✗	✗	✗
Tonga	✗	✗	✗	✗

OUT OF THE 4 MAJOR LANGUAGES IN MALAWI, CHICHEWA IS THE ONLY ONE WITH SOME AI SUPPORT

Yao, Tumbuka, Sena and other vernacular languages lack basic language technology support such as translating from English to these languages

All Malawian languages fall within the category of low-resource languages in NLP literature

Low resource languages have negligible support because they have little digital content readily accessible online and the languages are also not commercially viable for big tech to invest in.

New multilingual methods for NLP are making progress in low resource language support more rapid.

3. CHICHEWA AI PROJECT

1 Goals and Objectives

2 Main Activities & Results

3 Project use cases

PROJECT GOALS

Contribute to building underlying AI infrastructure (data, models and documentation) to enable basic language technology support for all local languages in the country.

MASSIVE LOCAL LANGUAGE CONTENT

Text and speech data in local languages which can be used and re-used by researchers, commercial companies and other organizations to continuously improve AI models for language support

MODEL FINE-TUNING AND EXPERIMENTATION

Make available fine-tuned models for anyone to use; perform experimentation to benchmark model performance (e.g., which model is best for what language)

DOCUMENTATION

Document datasets, model APIs, performance tests, best practices to enable use and re-use of these assets

PROJECT TEAM

Technical Lead

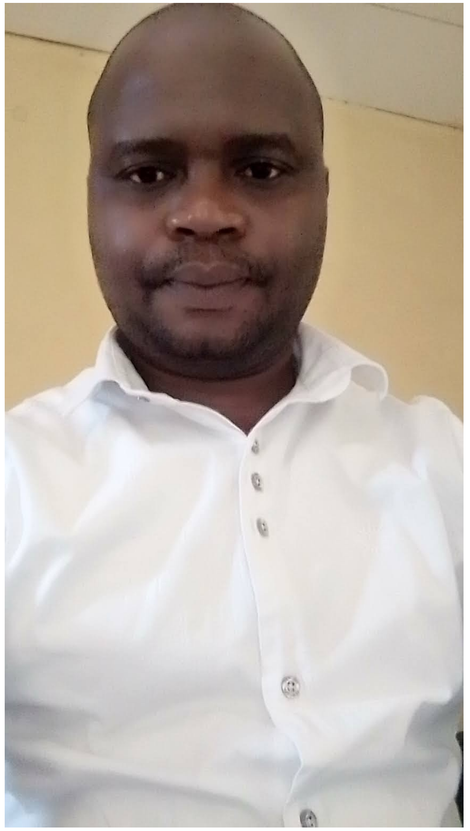
Dr. Dunstan Matekenya

NLP-Researcher-ASR

Willy, [PhD candidate]

NLP Researcher-MT

Stephen Kiilu, MSc



Data Annotation

Evance Mathewe,

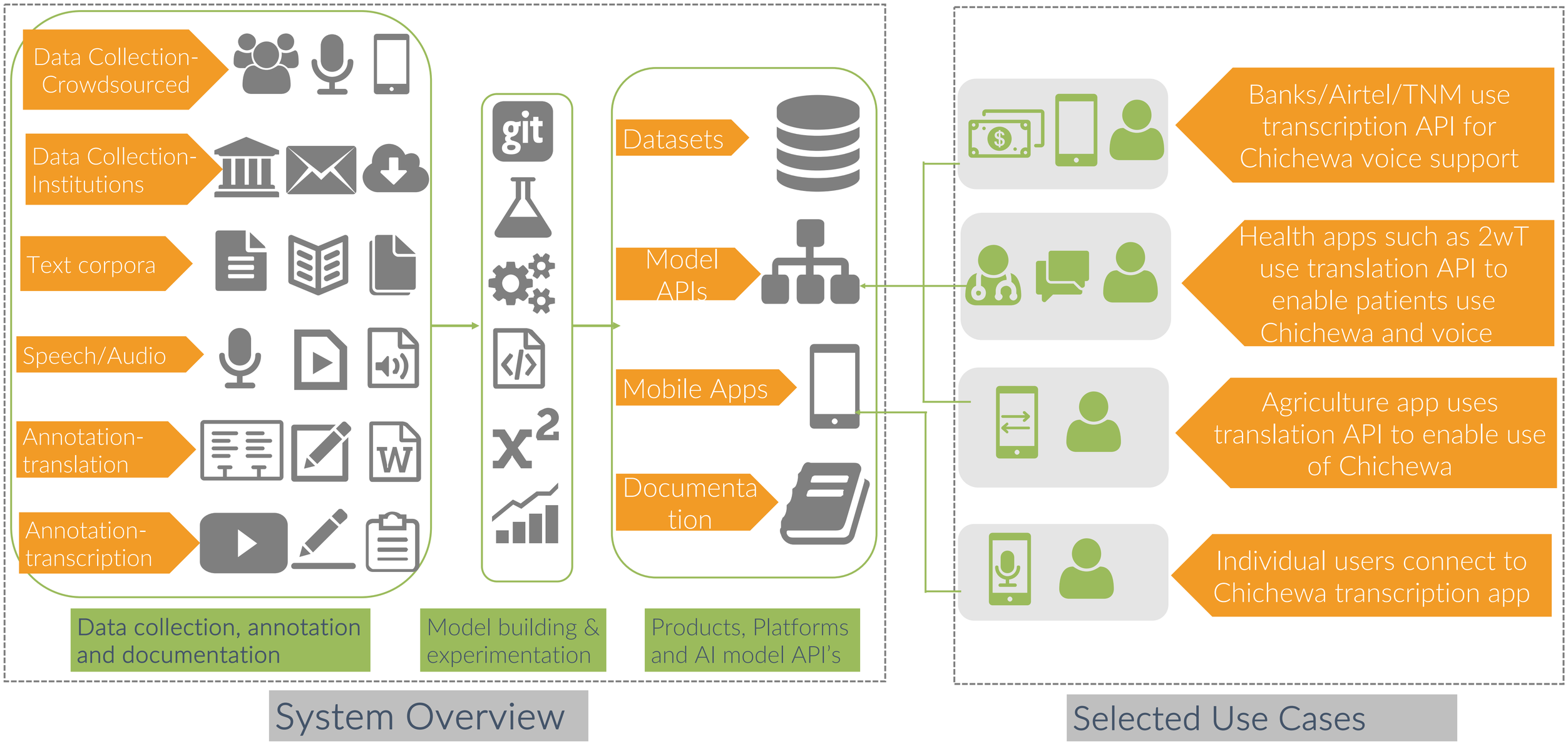
Data Annotation

Gloria Chirwa

“There is a lot of local language content in Malawi, we just need to gather from disparate sources (ministries, universities, private companies etc), digitize from analog sources, make content easily accessible online”

DUNSTAN

SYSTEM OVERVIEW FOR THE PROJECT



FOUR MAIN FOCUS AREAS FOR THIS PROJECT

Data, Models & Documentation

DATA COLLECTION AND ANNOTATION

Using low-cost approaches to collect, annotate and publish data for building LLMs

- Unlabeled text and speech data
- Labeled text and speech data

MODEL FINE-TUNING AND EXPERIMENTATION

Take base open source models and enhance their performance using different approaches

- Which base model is best?
- How much data is needed?

DOCUMENTATION

Enable others users (including commercial entities) find appropriate documentation about AI for local languages

- What's the best model for what language?
- Where to find data?

APPS AND DEMOS

Make a big impact with professional slides, charts, infographics and more.

- Speech2Text apps or WhatsApp Bot
- English to local language translation

DATA COLLECTION AND ANNOTATION

The Main Activities

1. TEXT CONTENT FROM INSTITUTIONS AND ORGS

Fiction, non-fiction books, articles, text books, communiques, speeches

2. TRANSLATED TEXT CONTENT FROM INSTITUTIONS

Translated government documents (e.g., Land reform law), other institutions/orgs

3. AUDIO RECORDINGS WITH/WITHOUT TRANSCRIPTS FROM INSTITUTIONS

FGDs, radio station recordings, court systems recordings etc.

4. TEXT/AUDIO CONTENT FROM SOCIAL MEDIA

Publicly shared voice notes on WhatsApp, Facebook, posts

5. CROWD-SOURCE BASED DATA COLLECTION

Collect speech data from diverse population

6. ANNOTATION AND CURATION

1. Translate from source (e.g., EN) to target (e.g., Yao)
2. Transcribe audio
3. Curate data for publishing

DATA COLLECTION AND ANNOTATION

What Has Been Achieved So Far

SPEECH DATASETS

1. A **70 hours** (about **20 transcribed**) Chichewa speech dataset published through Google NLP Hack Series
2. Ongoing collection and processing of 100+ hours of untranscribed speech (e.g., from social researchers (FGD's))

CHICHEWA TEXT CORPORA

1. About **5k Chichewa** Newspaper articles
2. About **25 Chichewa** books (fiction, non-fiction, text books)

PARALLEL MT DATASETS

1. **15k sentences English to Chichewa** parallel dataset
2. 1k English strings translated as part of on **Localization of Chichewa on Mozilla**
3. Ongoing gathering, processing and cleaning of already translated content (e.g., from NSO-questionnaires, Ministry of Lands etc)

MODEL FINE-TUNING AND EXPERIMENTATION

The Main Activities

1. FINE-TUNE & CUSTOMIZE LLMs & ASR MODELS

Enhance models to perform better on target languages

2. BENCHMARK MODELS & PLATFORMS

How well LLMs & other platforms perform on target languages for specific tasks

3. COMPARE LLMs, ASR MODELS & PLATFORMS

Out of numerous LLMs (NLLB, mBart), speech models (Whisper, MMS) & platforms (Google, Bing) which one works better

4. STRATEGIES FOR ENHANCING MODELS ON LOCAL LANGUAGES

Fine-tune (transfer learning), pivot-languages, use specific similar language, unlabeled data

5. HOW MUCH DATA IS REQUIRED

How many translated sentences (MT) and transcribed speech hours (ASR) lead to required performance

6. BUILD LLMs FROM SCRATCH

Can we build LLMs for local languages from scratch

MODEL FINE-TUNING AND EXPERIMENTATION

Results in Machine Translation(MT) Experiments

Experiment Objective: Which LLM/platform provides the best machine translation (MT) from English to Chichewa?

Performance Metric: BLEU score (out of 100%) which measures quality of MT model

Datasets used: Custom dataset from this project (Chichewa NLP) and other online datasets

LLM	Chichewa NLP		VV		FLORES-200		MAFAND-MT		Average	
	BLEU	ChrF	BLEU	ChrF	BLEU	ChrF	BLEU	ChrF	BLEU	ChrF
1 Google Translate	**17.1838**	**49.8466**	**21.2157**	**52.8351**	16.7542	50.0673	18.3846	50.9163		
2 MS Bing	15.0289	48.7437	19.5783	51.3756	**17.2489**	**50.2119**	17.2854	50.1104		
4 ChatGPT(3.5)	4.5473	30.8839	5.6283	30.4527	6.113	33.2838	5.0962	31.8735		
NLLB (600M)	12.8351	45.1251	15.8566	46.7084	14.7799	46.9726	14.8239	46.2687		
NLLB (1.3B)	13.662	46.4778	16.5077	47.7076	15.7197	47.28	15.9638	47.4885		
3 NLLB (3.3B)	14.0357	46.8637	17.2807	48.6876	15.7805	48.0068	15.6989	47.5194		

MODEL FINE-TUNING AND EXPERIMENTATION

First ASR Model for Chichewa

Try this WhatsApp Speech2Text AI bot to convert Chichewa voice note to text

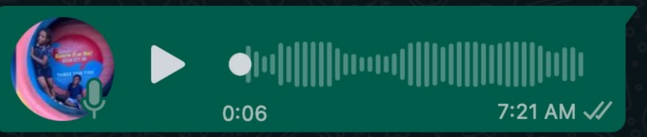
1. Access by scanning the QR code
2. Or Add +1 (564) 544-5403 to your WhatsApp
3. Send an audio or VN in Chichewa.
4. Examine the generated text

ZaZo AI
click here for contact info

ZaZo AI
Mwakhutitsidwa ndi zotsatilazi?
Ayi 👎 6:05 AM ✓

Chonde konzani kamasulidwe ka nkhani ndinakupatsani 6:05 AM

TODAY



You
0:06
Mwadzuka bwanji kumeneko? ndimafuna ndikufunsa kuti odwalaadzuka bwanji lero.
7:21 AM

ZaZo AI
Mwadzuka bwanji kumeneko? ndimafuna ndikufunsa kuti odwalaadzuka bwanji lero.
Mwakhutitsidwa ndi zotsatilazi? 7:21 AM

Eya 👍 Ayi 👎

ZaZo AI
Mwakhutitsidwa ndi zotsatilazi?
Ayi 👎 7:40 AM ✓

Chonde konzani kamasulidwe ka nkhani ndinakupatsani 7:40 AM

HOW ARE THE OUTPUTS OF THIS PROJECT BEING USED IN MALAWI?

CURRENT USE CASES

How The Outputs from this Project Are Being used

AGRICULTURE CHAT-BOT

1. Opportunity International (OI) working on AI based tool to assist extension workers ask Agriculture questions
2. The tools is using MT to convert input Chichewa questions into English and vice-versa
3. The goal is to also use speech input and provide answers as speech (VNs)

TRANSCRIPTION OF SOCIAL RESEARCH INTERVIEWS

1. Chitetezo is a longitudinal study of an adolescent-focused advocacy intervention designed to decrease the frequency of road traffic collisions
2. In talks to use the ASR model to automatically transcribe 300 FGD interviews

CURRENT USE CASES

Opportunity International (OI) Use Case

Broad Goals and Objectives

1. Enable Agriculture Extension workers access crucial agricultural information through using Chichewa and accessible apps (e.g., WhatsApp)
2. Ultimately, enable farmers access same information

Technical Details

1. Using Retrieval Augmented Generation (RAG) which enables use of external datasets to enhance LLMs
2. The Good Agriculture Practices (GAP) is used as the external document to provide authoritative information

ANTICIPATED USE CASES-THE MUUNI CASE

LOCAL COUNCIL	AGRICULTURE	BANKING & FINANCIAL SERVIVES	EDUCATION	HEALTH & ENVIRONMENT	ICT AND DIGITAL SERVICES
		REVENUE – Introduction of e-ticketing			
PHALOMBE DC	Electronic platform to link farmers to reliable off takers.	Introduction of E-ticketing mechanisms and e-licensing mechanisms	Integrating ICT in schools	Utilization of CMA app	
RUMPHI DC	Website and database for value chain actors and stakeholders	Development of user-friendly digital financial apps.	Platform for school activities and performance	Electronic Medical Records	Digital platforms that can use vernacular language.

3. ROADMAP TO ACCLERATED AI SUPPORT FOR LOCAL LANGUAGES

1 What Challenges are we Solving?

2 Suggested RoadMap

3 Call to Action

WHAT PROBLEMS ARE WE SOLVING?

01. LACK OF LOCAL LANGUAGE DIGITAL CONTENT

02. LACK OF DOCUMENTATION ON NLP FOR LOCAL LANGUAGES

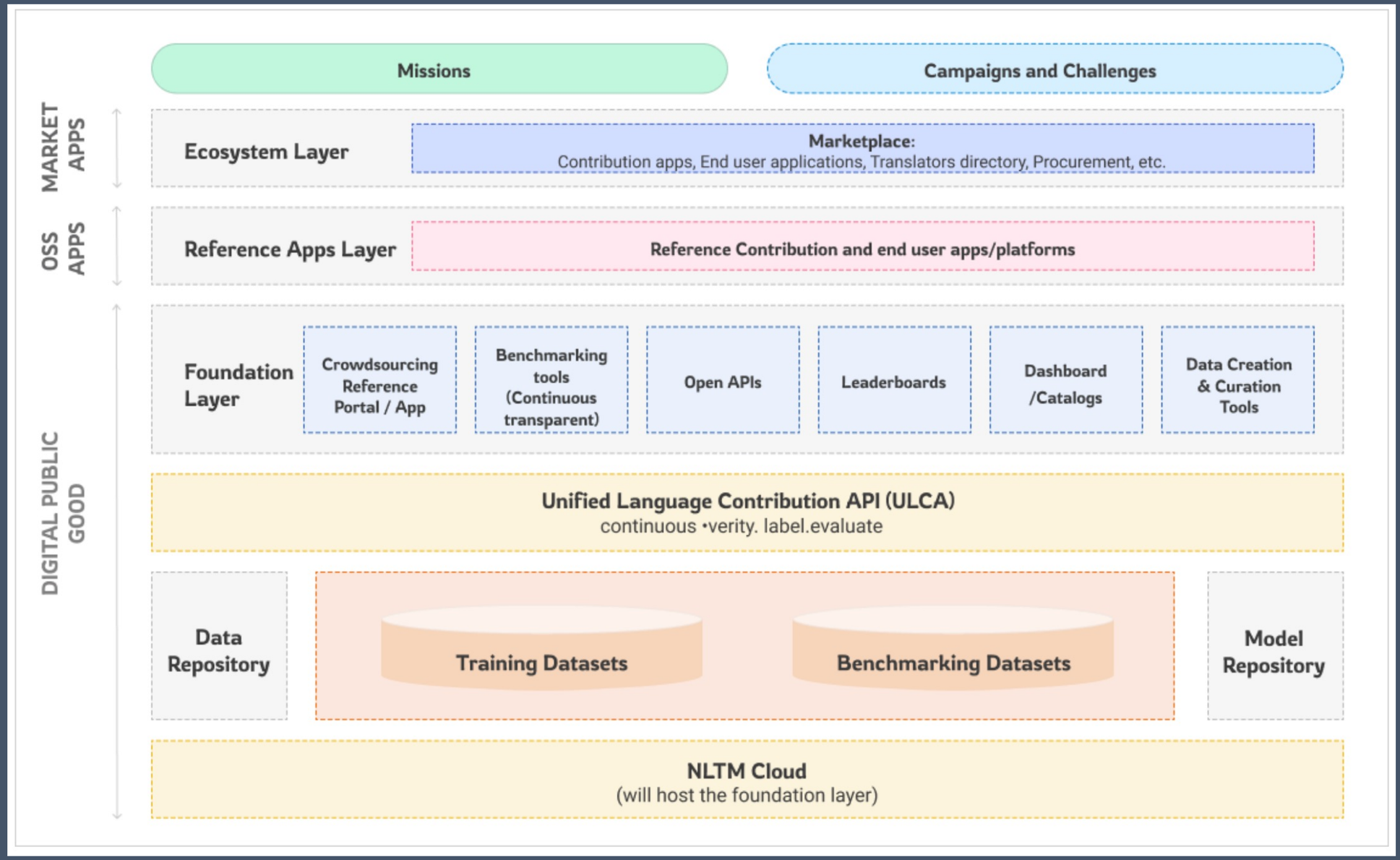
03. LACK OF LOCAL RESEARCH AND DEVELOPMENT

04. LACK OF GUIDELINES AND STANDARDS

Let’s create a National AI Infrastructure (NAII) to do the following:

1. Data repository. Coordinate local language content harvesting, sharing and access
2. Governance. Create and manage standards and best practices
3. Model zoo. Act as repository for customized LLMs and speech models
4. Documentation. Technical documentation on research in NLP

Potential Architecture for The National AI Infrastructure (NAII) Computing Platform



Source: https://bhashini.gov.in/about-bhashini

HOW CAN YOU HELP?

1. Sharing local language content from your organization (e.g., translated documents)
2. Contribute to research in in NLP
3. Contribute to data creation (e.g., share voice notes, translate content etc)

THANK YOU!

ANY QUESTIONS?

Email:
dmatekenya@gmail.com