Empowering Africa's Languages in AI and Machine Learning

Prof Vukosi Marivate (ABSA UP Chair of DS)
CTO, Lelapa AI
vukosi.marivate@cs.up.ac.za

September 2023

The pirates in the yacht club - midjourney

LELAPA AI

CC BY SA

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Affiliations

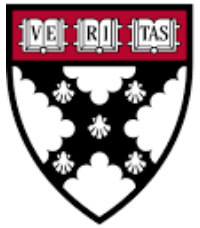
DEEP LEARNING
INDABA

LELAPA AI

Lacuna Fund
Our voice on data

Masakhane

A grassroots NLP community for Africa, by Africans



Harvard
Business
School

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter
www.up.ac.za

Declaring My Assumptions

VUKOSI MARIVATE
Chair of data science

ABIODUN MODUPE
Senior Member

SEANI RANANGA
Senior Member

MICHELLE TERBLANCHE
Student : PhD

BAPHUMELELE MASIKISIKI
Student : PhD CS

MPHO MOKOATLE
Student : PhD CS

HERKULAAS COMBRINK
Student : PhD CS

ROCKERFELLER
Student : PhD AIMS

TEMITOPE KEKERE
Student : PhD IT

MHLASAKULU LEKA MVUBU
Student : PhD

PITSO KHOBOKO
Student :CS MSc

AMICA DE JAGER
Student :CS MSc

KATHLEEN SIMINYU
Student :CS MSc

MACK MAKGATHO
Student :CS MSc

THAPELO SINDANE
Student :CS MSc

TSHOLOFELO GOMBA
Student :CS MSc

MIEHLEKETO MATHEBULA
Student :CS MSc

NELSON PHORA
Student :CS MIT

SHAUN JOHNSON
Student :CS MIT

YOLANDA NKALASHE
Student :CS MIT

DERWIN NGOMANE
Student :CS MIT

MATIMBA SHINGANGE
Student :CS MIT

MOHLATLEGO NAKENG
Student :CS MIT

NDAMULELO NEMAKHAVHANI
Student :CS MIT

FISKANI BANDA
Student :CS MIT

MATOME LEDWABA
Student :CS MIT

ROZINA MYOYA
Student :CS MIT

MARC GAGIANO
Student :CS MIT

PENELOPE MATLOGA
Student :CS MIT

GODFREY TSHEHLO
Student :CS Honours

MULWELI MUKWEVHO
Student :CS Honours

MADUVHA MALANGE
Student :CS Honours

DAVID WALKER
Student :CS Honours

AMUKELANI CHAUKE
Student :CS Honours

ALEX MOLOKOMME
Student :CS Honours

LESIBA LEDWABA
Student :CS Honours

NHLAKANIPHO NGWEYA
Student :CS Honours

MATOME MAKGOPA
Student :CS Honours

RICHARD LASTRUCCI
Student : Research Assistant

Data Science for Social Impact

Others not pictured
Alumni

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter
www.up.ac.za

Africa and the ongoing AI revolution - An opinionated view

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

What is Artificial Intelligence?

Classical definition

An ***Agent*** (machine)

Exists in some ***Environment***

Takes ***Action***

To achieve

some ***Goal(s)***

Artificial Intelligence: A Modern Approach by Peter Norvig and Stuart J. Russell

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

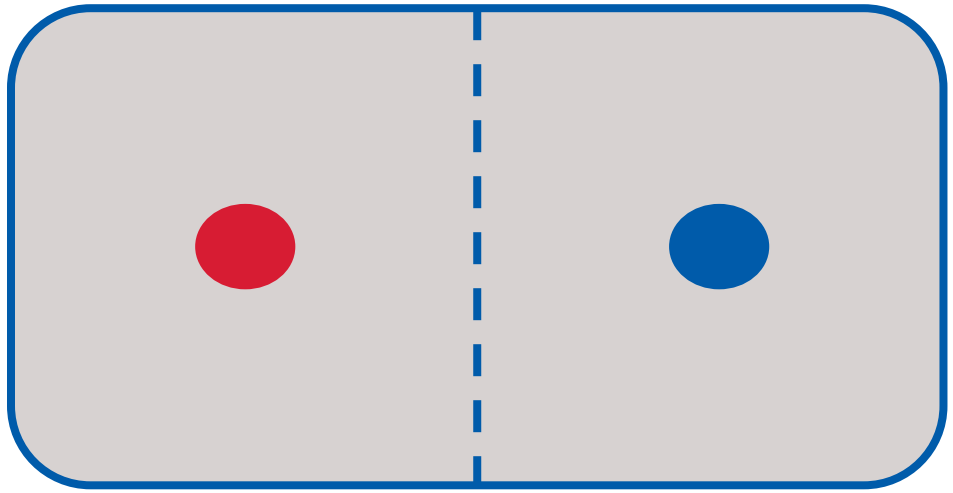
Make today matter

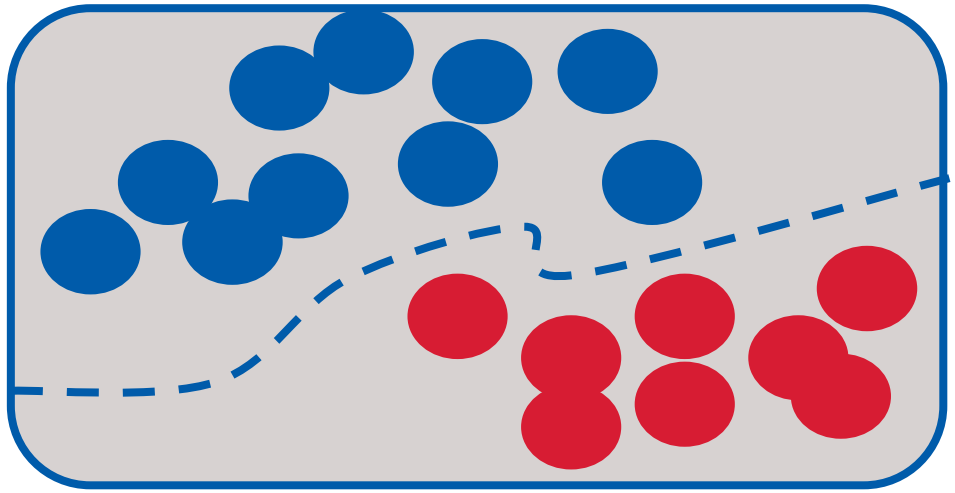
www.up.ac.za

What is Machine Learning?

Subset of **AI** that deals with

A machine ***learning*** patterns from data





UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

IS THIS A
CAT or DOG?

CAT DOG

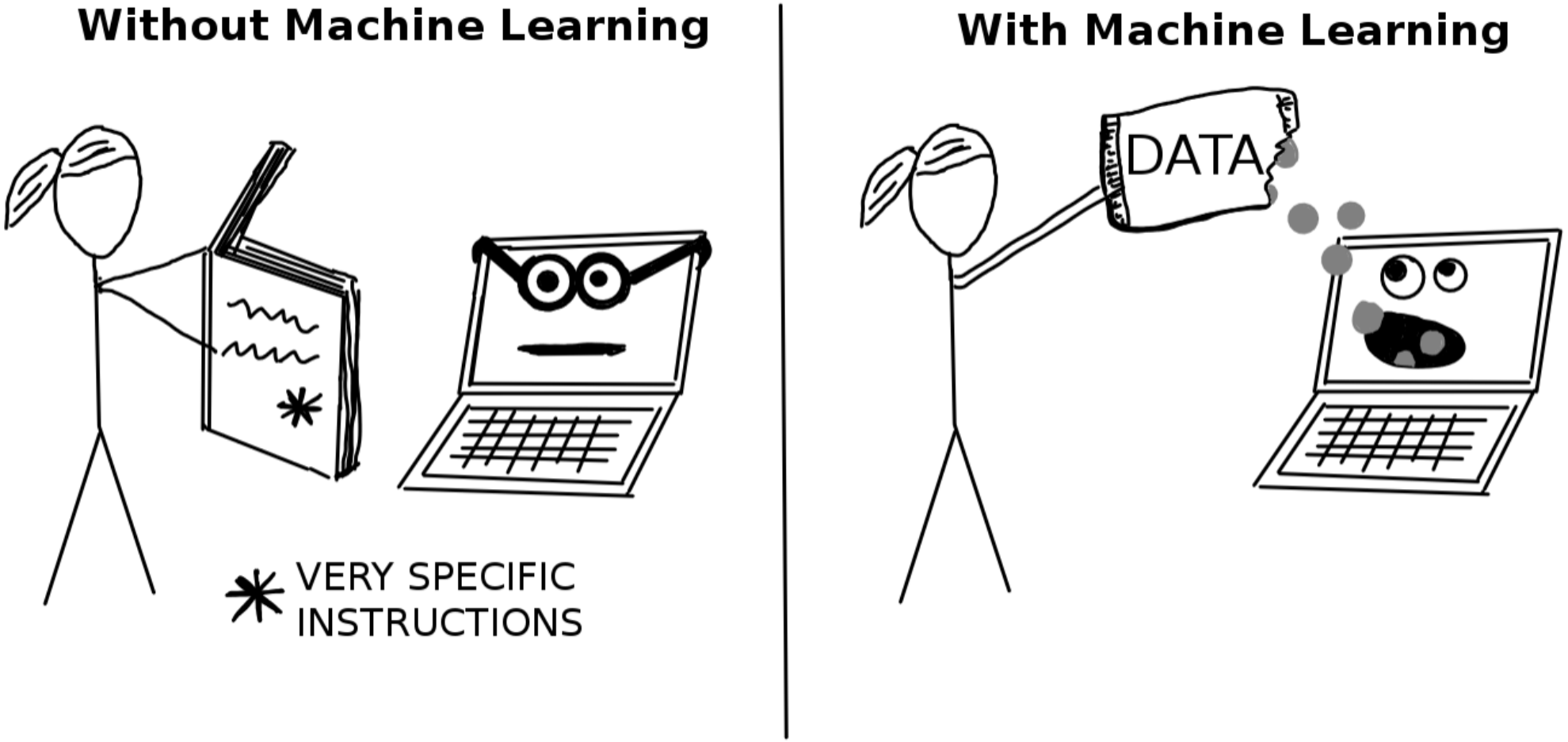
OUTPUT LAYER

ACTIVATED NEURONS

INPUT LAYER

DEEP neural NETWORK

Why ML?



Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.
Christoph Molnar

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Make today matter
www.up.ac.za

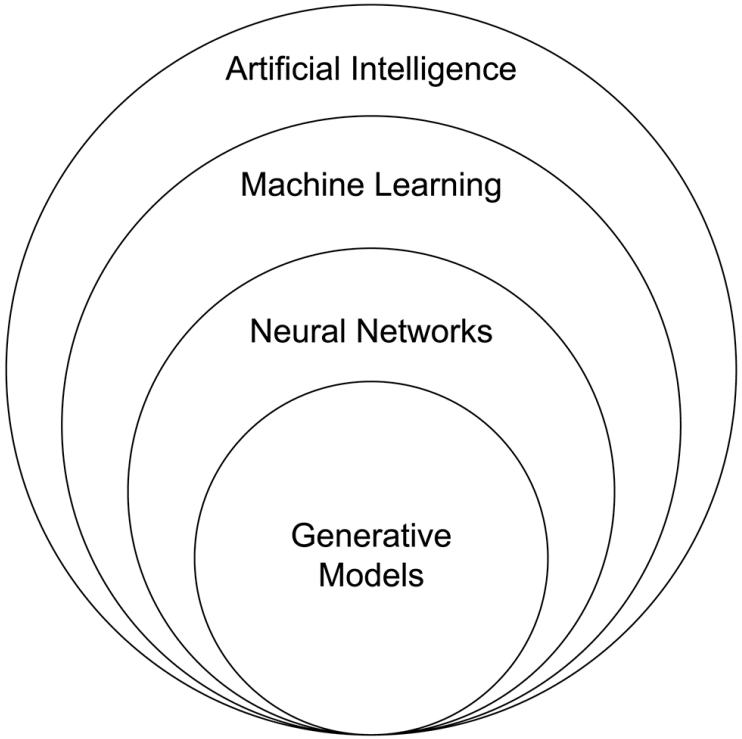
What is old is new again

This is not the first time we are going through an AI/ML hype cycle.

We are now in the age of Generative AI

It won't be the last.

How do we learn from the past to also get a grip on the future?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter
www.up.ac.za

There will be change

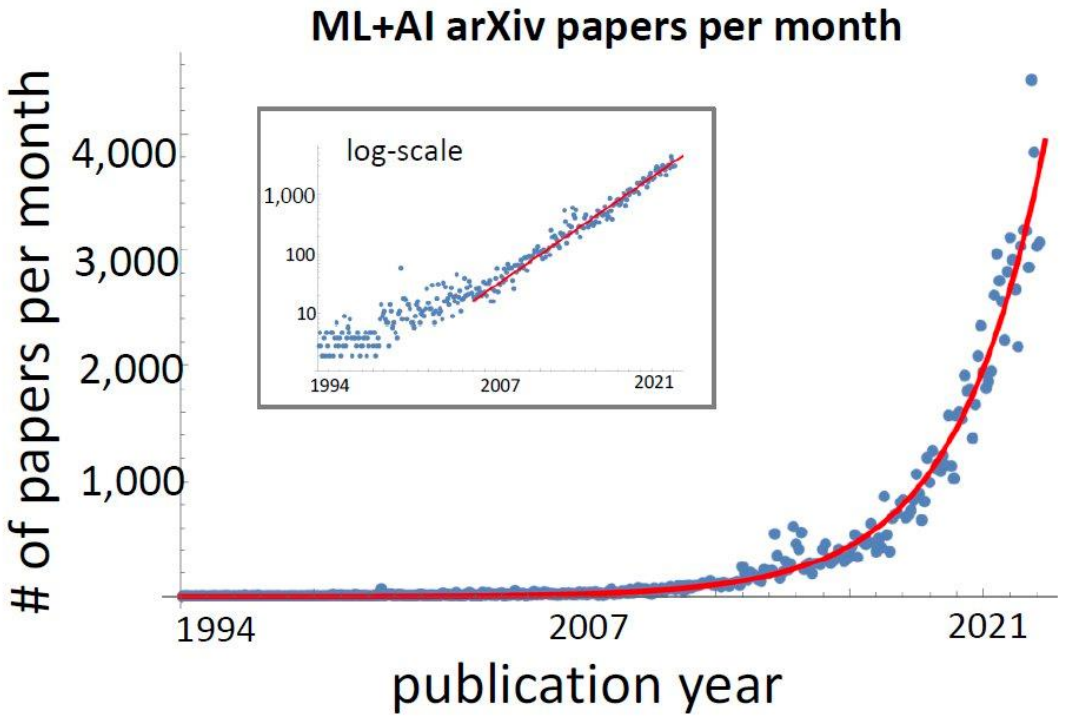
Many ideas that will be floating around.

Most will wither away, some will float to the top.

Predicting the Future of AI with AI:
High-Quality link prediction in an exponentially growing knowledge network

Mario Krenn,[1,*] Lorenzo Buffoni,[2] Bruno Coutinho,[2] Sagi Eppel,[3] Jacob Gates Foster,[4]
Andrew Gritsevskiy,[3,5,6] Harlin Lee,[4] Yichao Lu,[7] João P. Moutinho,[2] Nima Sanjabi,[8] Rishi Sonthalia,[4]
Ngoc Mai Tran,[9] Francisco Valente,[10] Yangxinyu Xie,[11] Rose Yu,[12] and Michael Kopp[6]

[1] *Max Planck Institute for the Science of Light (MPL), Erlangen, Germany.*
[2] *Instituto de Telecomunicações, Lisbon, Portugal.*
[3] *University of Toronto, Canada.*
[4] *University of California Los Angeles, USA.*
[5] *Cavendish Laboratories, Cavendish, Vermont, USA.*
[6] *Institute of Advanced Research in Artificial Intelligence (IARAI), Vienna, Austria.*
[7] *Layer 6 AI, Toronto, Canada.*
[8] *Independent Researcher, Barcelona, Spain.*
[9] *University of Texas at Austin, USA.*
[10] *Independent Researcher, Leiria, Portugal.*
[11] *University of Pennsylvania, USA.*
[12] *University of California, San Diego, USA.*

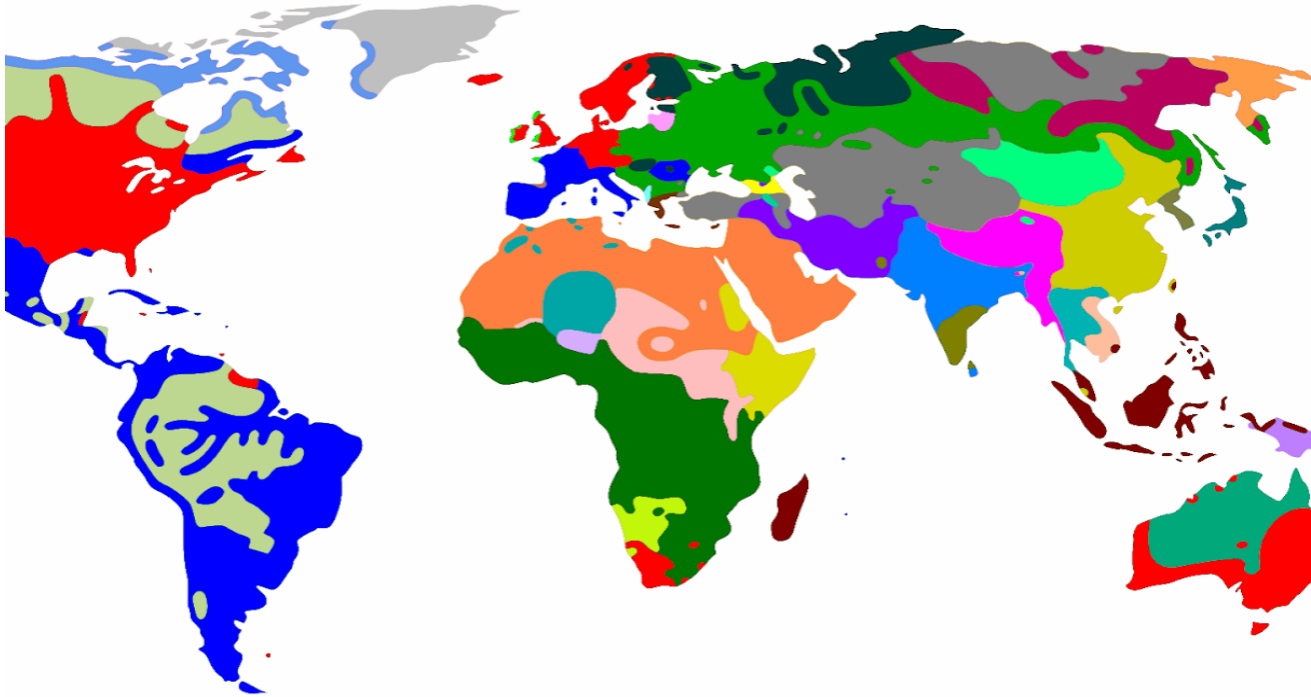


UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter
www.up.ac.za

Defining the challenges to Low Resource Languages

- Low availability of resources (Data, Tools, etc.)
- Discoverability
- Reproducibility
- Focus
- Benchmarks
- Scale and Complexity



Repartition map of the languages over the world (version blank of key) [Wikimedia:User:Industrius]

A Focus on Neural Machine Translation for African Languages

Laura Martinus
Explore / Johannesburg, South Africa
laura@explore-ai.net

Jade Abbott
Retro Rabbit / Johannesburg, South Africa
ja@retrorabbit.co.za

A Focus on Neural Machine Translation for African Languages

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter
www.up.ac.za

How we can all contribute to Language Development, lessons from the Grassroots

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Building on Grassroots AI

Lowering barriers to entry

Multidisciplinary

Inclusive

Address Systemic Issues

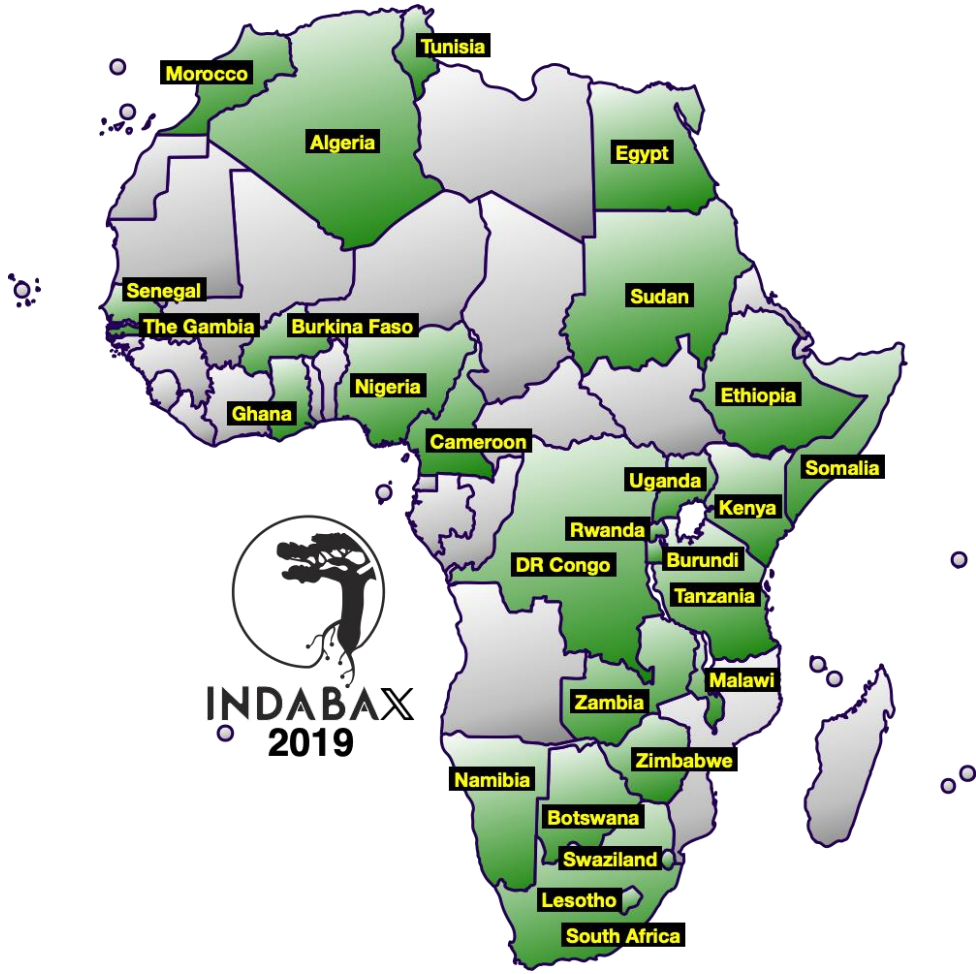
Collaborative

Independent / Multi-institutional

Holistic

Community Focused

The Deep Learning Indaba



There was an Idea #SautiYetu

Strengthening African Machine Learning

Vision

Africans becoming

- critical contributors,
- owners,
- and shapers

of the coming advances in artificial intelligence and machine learning.

https://deeplearningindaba.com/

The Talking Drums of Africa [2017, South Africa]

330 Attending [700 Applicants]

Masakhane [South Africa, 2018]

We Build Together

650 Attending [~1500 Applicants]

Sauti Yetu [Kenya, 2019]

Our Voice

650 Attending [~1800 Applicants]

Masirouna [Tunisia, 2022]

Our Destiny

400 Attending

Yɛbɛtumi (We Can) [Ghana, 2024]

Our Destiny

800 Attending, 3000+ Applications

36 African Countries

800+ attendees

45% women in machine learning

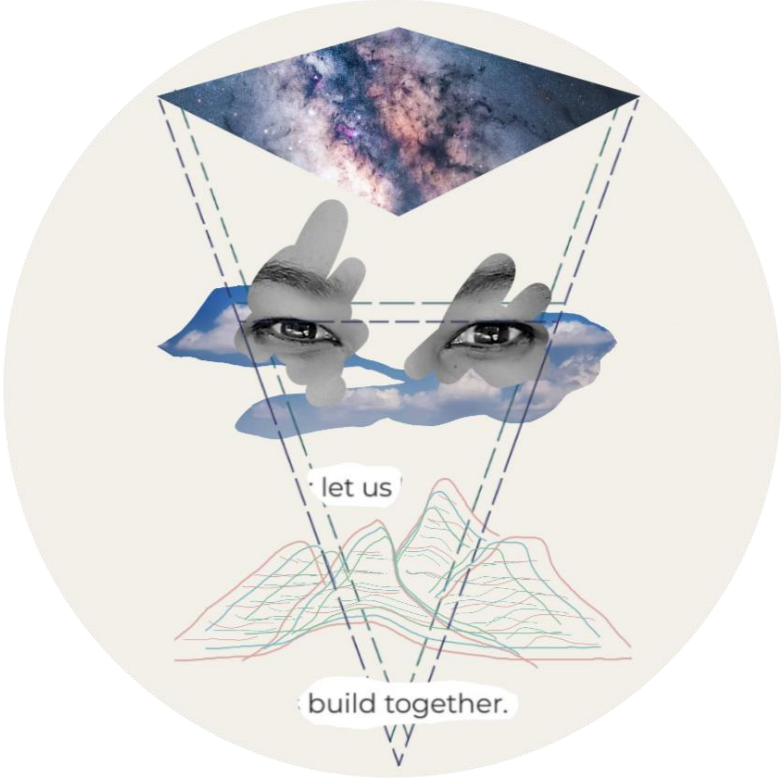
200+ travel grants

DEEP LEARNING INDABA

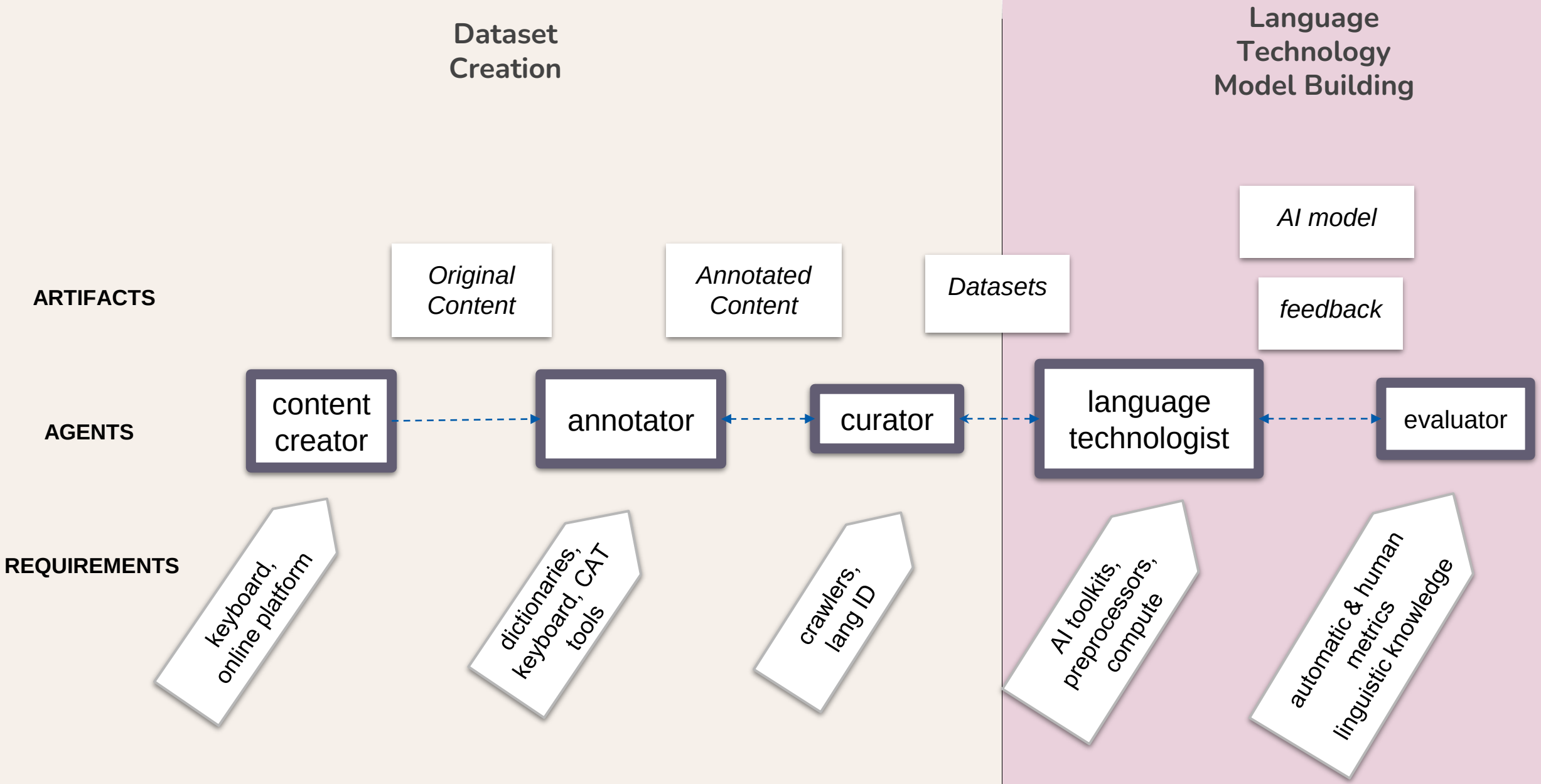
Masakhane is a **grassroots AI organisation** whose mission is to **strengthen and spur NLP research in African languages, for Africans, by Africans.**

Masakhane roughly translates to “We build together” in isiZulu.

Our goal is for Africans to shape and own these technological advances towards human dignity, well-being and equity, through inclusive community building, open participatory research and multidisciplinarity.



https://www.masakhane.io/



https://www.masakhane.io/

Preparing the Vuk'uzenzele and ZA-gov-multilingual South African multilingual corpora

Richard Lastrucci[1], Isheanesu Dzingirai[1], Jenalea Rajab[2], Andani Madodonga[1], Matimba Shingange[1], Daniel Njini[1], Vukosi Marivate[1,3]

[1]Department of Computer Science, University of Pretoria

[2]School of Computer Science and Applied Mathematics, University of the Witwatersrand

[3]Lelapa AI

richard.lastrucci@tuks.co.za, ishe.dzingirai@gmail.com, jenalea.rajab@gmail.com, andanim412@gmail.com, mrosslyns@gmail.com, vukosi.marivate@cs.up.ac.za

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Goals of the Project

Liberate and prepare textual multilingual data from Government Communication Information System [GCIS] of the South African government.

Extract and align sentence pairs for all South African languages within the data.

Automate as much of the process as possible.

Provide ***benchmarks for Machine Translation*** using this data.

Release the cleaned and machine readable data openly for other researchers to use.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

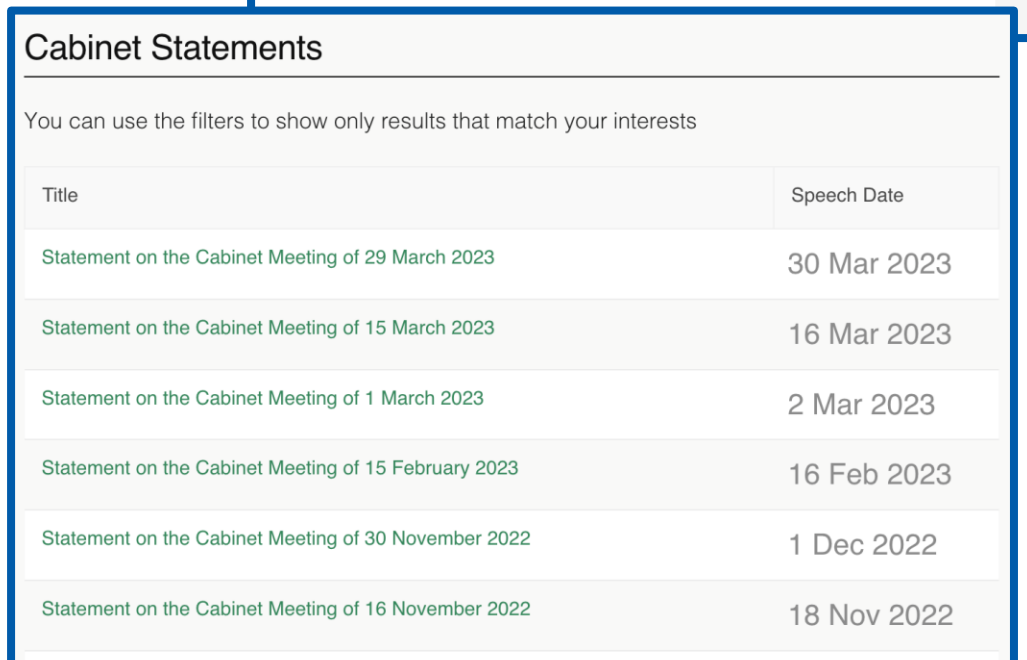
Cabinet Statements

What is this source?

- Statement of Cabinet Meetings of the South African government.
- A few weeks after the English statement is published, **translated statements are made available**.

Need

- **Extract statements** and make them easily accessible for NLP.
- Automate the process with **monitoring of website changes** and grow the dataset.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Make today matter
www.up.ac.za

Vuk’uzenzele Gov Newspaper

What is this source?

- “Vuk’uzenzele is a free Government Newspaper, committed to making a difference in the lives of South Africans”
- Published in 2 editions per month. Available physically and digitally [PDF} in all 11 South African languages.

Need

- **Extract text from PDFS.** Clean and make them easily accessible for NLP.
- Make the **extraction process repeatable** with humans in the loop.

https://www.vukuzenzele.gov.za/

Vuk’uzenzel

Yi humelerisiwile hi: Vufambisi bya Mfumo bya Vuhlanganisi na Mahungu (GCIS) English/Xitsonga Dzivamisoko 2

Ku endla swimakiwa swa mi na swo tisasekisa hi mpepo

Sihle Manda

Namhla Dotwana wa Namhla Collection u endla xiave eka swimakiwa swa misisi na swo tisasekisa hi ku tirhisa impepho.

Health • Rural Development • Employment • Safety & Security • Education

Vuk’uzenzele

Produced by: Government Communication & Information System (GCIS) April 2023 Edition 1

Producing hair and beauty products with *impepho*

Sihle Manda

Consumed by boredom while stuck in her flat in Cape Town at the height of the COVID-19 one evening, Namhla Dotwana had a light bulb moment that gave birth to a blossoming start-up.

Dotwana, 31, founder of Namhla Collection, recounted her journey to *Vuk’uzenzele*.

Namhla Collection is an innovative organic skincare range business based in Egotyibeni village in Tsolo, in the Eastern Cape.

The start-up journey began when Dotwana remelted an old cinnamon-scented candle deformed by harsh sunlight on her window sill.

When the media relations practitioner’s contract with the City of Cape Town’s social development unit lapsed at the end of 2020,

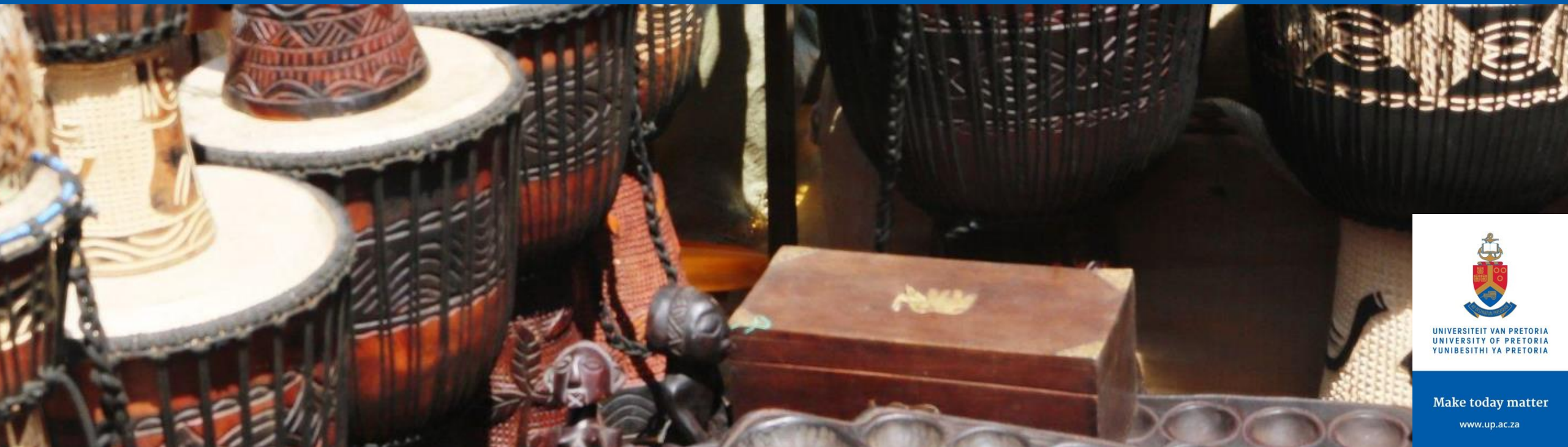
Namhla Dotwana of Namhla Collection making her mark in hair and beauty products with the use of sage (*impepho*).

Cont. on page 2

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter
www.up.ac.za

Sentence Alignment and MT Benchmarks



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Data Characteristics for Aligned Translations

Table 5: Characteristics of the translation data for the Vuk'uzenzele (Vuk.) datasets

Translation Direction	Size #sents (#src / #trg tokens)
nbl→eng	136 (3.4k / 3.9k)
nso→eng	1715 (53.7k / 41.6k)
ssw→eng	1588 (29.9k / 37.6k)
sot→eng	260 (9.9k / 7.5k)
tsn→eng	1366 (49.8k / 31.7k)
tso→eng	1998 (58.9k/ 46.6k)
ven→eng	230 (9.1k / 7k)
xho→eng	1338 (25.8k / 31.5k)
zul→eng	1874 (34.1k / 43k)

Table 6: Characteristics of the translation data for the ZA-gov-multilingual (Gov.) datasets

Translation Direction	Size #sents (#src / #trg tokens)
nbl→eng	3513 (63.9k / 107k)
nso→eng	14742 (460.9k / 375k)
ssw→eng	15139 (291k/ 377.8k)
sot→eng	4995 (145.9k / 153.5k)
tsn→eng	14068 (493.1k / 362.2k)
tso→eng	15393 (466.4k / 381.2k)
ven→eng	3404 (68.2k / 96.6k)
xho→eng	15853 (318.2k / 389.5k)
zul→eng	15503 (327.5k / 384.1k)

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter
www.up.ac.za

MT Results

The highest BLEU scores are distributed across the ZA-gov-multilingual and Autshumato NMT models

ZA-gov-multilingual models achieving a higher score for **Setswana, Xitsonga and isiZulu**

Highest benchmark result for **Xitsonga** across all datasets, demonstrating the effectiveness of transfer learning for new low-resource language datasets.

Our contributions:

- Extend the benchmark translation resources (in the government data domain) to **isiNdebele, isiXhosa, siSwati and Tshivenda**
- Broaden the translation direction **beyond English** as the source language.

Table 4: BLEU scores for Massively Multilingual Transfer on xxx-eng translations using the Vuk'uzenzele (Vuk.), ZA-gov-multilingual (Gov.) and subsets of the available Autshumato datasets (Aut.)

Translation Direction	BLEU		
	Vuk.	Gov.	Aut.
nbl→eng	7.33	8.04	**12.24**
nso→eng	9.29	**26.50**	-
ssw→eng	4.80	**28.72**	-
sot→eng	4.55	10.21	**14.83**
tsn→eng	2.80	**29.68**	28.04
tso→eng	13.86	**35.40**	32.10
ven→eng	2.32	9.68	**17.24**
xho→eng	6.05	26.81	-
zul→eng	9.97	**30.03**	25.90

• Source languages which were not including in the original M2M100 pre-training are highlighted

It is noted that variations in the Autsumato subset selections will yield different results and an in-depth analysis is left for future work.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter
www.up.ac.za

SABC Tindzaba
@SABCTindzaba

Badlali beBafana Bafana Ronwen Williams na Percy Tau banelitsemba lekutsi lelicembu liyowucudzelana ku AFCON e Ivory Coast hhayi nje kutsi liyowungeta sibalo.

#SABCNews
#SABCTindzaba

Hosted inference API

Text2Text Generation

Badlali beBafana Bafana Ronwen Williams na Percy Tau banelitsemba lekutsi lelicembu liyowucudzelana ku AFCON e Ivory Coast hhayi nje kutsi liyowungeta sibalo.

Compute ⌘+Enter

0.0

Computation time on Intel Xeon 3rd Gen Scalable cpu: 3.522 s

Bafana Bafana Ronwen Williams and Percy Tau are confident that the team is competing to AFCON on Ivory Coast and not only raise awards.

Make today matter
www.up.ac.za

Daily News BOTSWANA
THE PAPER THAT INFORMS AND EDUCATES

DOWNLOA
DAILYNEWS PDF

Breaking News

Like
Mo, 11 sep. 30° Tu, 12 sep. 24° We, 13 sep. 23° Th, 14 sep. 27° Fr, 15

Our NEW facebook page name & Look DailyNewsbw

Home / News / Dikgang / Ba lebisitswe molato wa petelelo le go kgothosa

BA LEBISITSWE MOLATO WA PETELELO LE GO KGOTHOSA

10 Sep 2023

Borre ba le babedi ba ba tlholegang kwa lefatsheng la Zimbabwe ba itshupile fa pele ga lekgotla ka Labone mabapi le molato wa go tsaya ka dikgoka le go thubetsa mme yo o dingwaga tse di masome a matlhano le bobedi (52) wa kgotla ya Kanamo kwa Mahalapye.

Borre bao, wa dingwaga tse di masome a mararo le bobedi (32) le wa tse di masome a mararo le borataro (36) mo bekeng e e fetileng ba ne ba tsenelela mme yo mo ntlong ba mo amoga megala ya letheka e le mebedi ba bo ba mmetelela.

Mookamela Mapodisi a Mahalapye Central Supt. Boitshepho Mudongo o kaile fa ka lesego ba kgonne go ba tshwara mme megala yotlhe ka bobedi e bonwe.

A re bobedi jo bo amanngwa le dikgetsi tse dingwe tsa go kgothosa di le pedi ka ba fitlhetswe ka dilwana tse di neng di begilwe fa di

Hosted inference API

Token Classification

Borre ba le babedi ba ba tlholegang kwa lefatsheng la Zimbabwe ba itshupile fa pele ga lekgotla ka Labone mabapi le molato wa go tsaya ka dikgoka le go thubetsa mme yo o dingwaga tse di masome a matlhano le bobedi (52) wa kgotla ya Kanamo kwa Mahalapye.

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

Borre ba le babedi ba ba tlholegang kwa lefatsheng la Zimbabwe LOC ba itshupile fa pele ga lekgotla ka Labone DATE mabapi le molato wa go tsaya ka dikgoka le go thubetsa mme yo o dingwaga tse di masome a matlhano DATE le bobedi DATE (52) wa kgotla ya LOC Kana PER mo kwa LOC Mahalapye LOC.

JSON Output

Maximize

Hosted inference API

Token Classification

Borre ba le babedi ba ba tlholegang kwa lefatsheng la Zimbabwe ba itshupile fa pele ga lekgotla ka Labone mabapi le molato wa go tsaya ka dikgoka le go thubetsa mme yo o dingwaga tse di masome a matlhano le bobedi (52) wa kgotla ya Kanamo kwa Mahalapye.

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.042 s

Borre NOUN ba le DET babedi NUM ba DET ba PRON tlholegang VERB kwa DET lefatsheng PROPN la DET Zimbabwe PROPN ba PRON
itshupile VERB fa PUNCT pele ADV ga ADP lekgotla NOUN ka ADP Labone mabapi NOUN le AUX molato NOUN wa DET go tsaya VERB ka ADP
dikgoka NOUN le CCONJ go thubetsa VERB mme ADV yo DET o dingwaga NUM tse DET di masome NUM a PROPN matlhano NUM le CCONJ
bobedi NUM (PUNCT 52 NUM) PUNCT wa DET kgotla NOUN ya DET Kana PROPN mo kwa DET Mahalapye PROPN . PUNCT

JSON Output

Maximize

A few shot learning approach for a multilingual recommendation system for Agro-Information Question Answering

F Banda (MITC Big Data Science), J Nakatumba-Nabende, M Axmed, V Marivate

Data Science for Social Impact

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

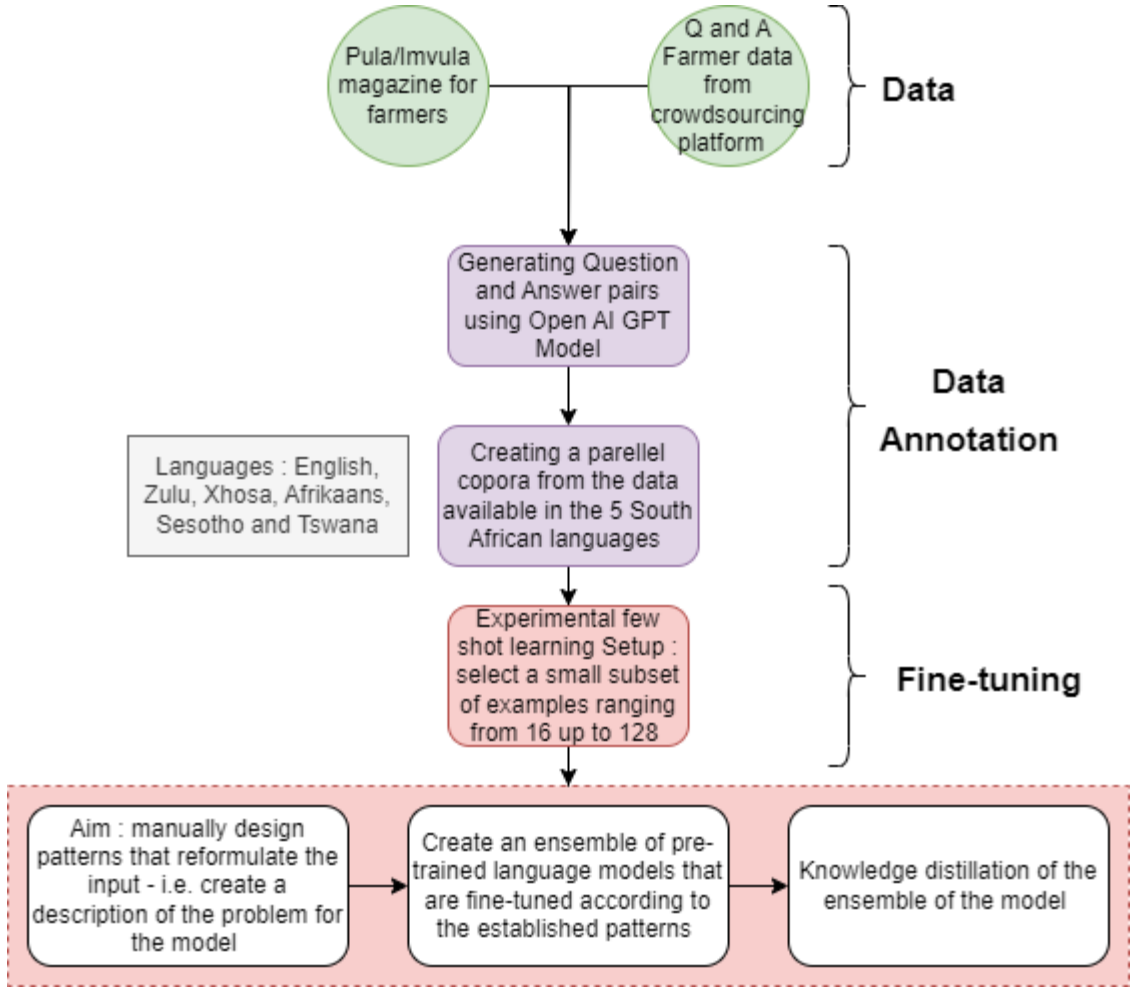
Make today matter

www.up.ac.za

A few shot learning approach for a multilingual recommendation system for Agro-Information Question Answering

- In many Sub-Saharan African households, smallholder farming is at the centre of their income and food security and face many challenges without any resources
- To bridge this gap , building low IT maintenance support systems , can provide support and reach a vast majority of smallholder farmers in need.

This led to the following research question: Can the main concept of Pattern exploiting training (PET), which is a few shot fine-tuning method limited to text classification, be adapted for closed domain multilingual extractive question answering ?



Data Science for Social Impact
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Make today matter
www.up.ac.za

Closing it off

A Call to Action

Get in touch with us.

We are looking for partners for language data (text, audio video)

Setting up communities of practice, training for young people with these skills.

Increasing R&D Investment

Use Cases

Prof. Vukosi Marivate
vukosi.marivate@cs.up.ac.za
https://dsfsi.github.io
@vukosi
@DSFSI_Research

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter
www.up.ac.za

Collaborations

Mail&Guardian
AFRICA'S BEST READ

Titimuleni Nthavela

DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION

MEDIA MONITORING AFRICA

AIMS

CSIR
Touching lives through innovation

Dedan Kimathi University of Technology

DEEP LEARNING INDABA

UVS
UNIVERSITE VIRTUELLE DU SENEGAL
foo nekk foofu la

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter
www.up.ac.za

DSFSI Support - Thank You

Data Science for Social Impact

Grants/Gifts/Fellowships

moz://a

Google Cloud

HARVARD UNIVERSITY
CENTER FOR AFRICAN STUDIES

absa

IDRC · CRDI
International Development Research Centre
Centre de recherches pour le développement international

Canada

DST-NRF CENTRE OF EXCELLENCE IN MATHEMATICAL & STATISTICAL SCIENCES
MaSS

FUTURE PROFESSORS PROGRAMME (PHASE 01)

CREATING THE NEXT GENERATION OF SOUTH AFRICAN PROFESSORS

The Future Professors Programme (Phase 01) is a flagship programme of the Department of Higher Education and Training. It is a national, collaborative initiative, aimed at developing the qualities of academic excellence and leadership in university scholarship in order to contribute to the development of a future South African professoriate.

Meta
facebook

TensorFlow

aws

Google Research

science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

NRF National Research Foundation

saMRC advancinglife

mastercard foundation

NVIDIA

JPMorgan Chase & Co.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter
www.up.ac.za

Thank you

Questions

Prof. Vukosi Marivate
vukosi.marivate@cs.up.ac.za
https://dsfsi.github.io
@vukosi
@DSFSI_Research

Keep in touch
Join our research group newsletter
https://tinyletter.com/datascience-up/

Made with ♡ in Tshwane

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za