

# An expert system for semi-automated classification of product and industry descriptions

E.L. Cano<sup>1,2</sup>, A.Deveza<sup>2</sup>, X. Gong<sup>2</sup>, D. Boko<sup>2</sup>, E. Keeble<sup>2</sup>, P. Debanes<sup>2</sup>

<sup>1</sup>Data Science Laboratory, Universidad Rey Juan Carlos

<sup>2</sup>United Nations Economic Commission for Africa (UNECA)

**Presenter: Emilio L. Cano**

Associate Professor at Rey Juan Carlos University, Madrid, Spain;  
former Consultant of UNECA



# The AfCIOT Project

# AfCIOT Project background

- Joint project with [UNECA](#), [WTO](#), and [OECD](#)
- International Trade, Environmental and Employment Indicators
- International and multidisciplinary team
- **Industry and product classification**, gap estimation, policy simulation and visualization
- Key result: A **shiny app** for reporting, visualization and policy simulation
- Several important challenges for the **multinational** scope



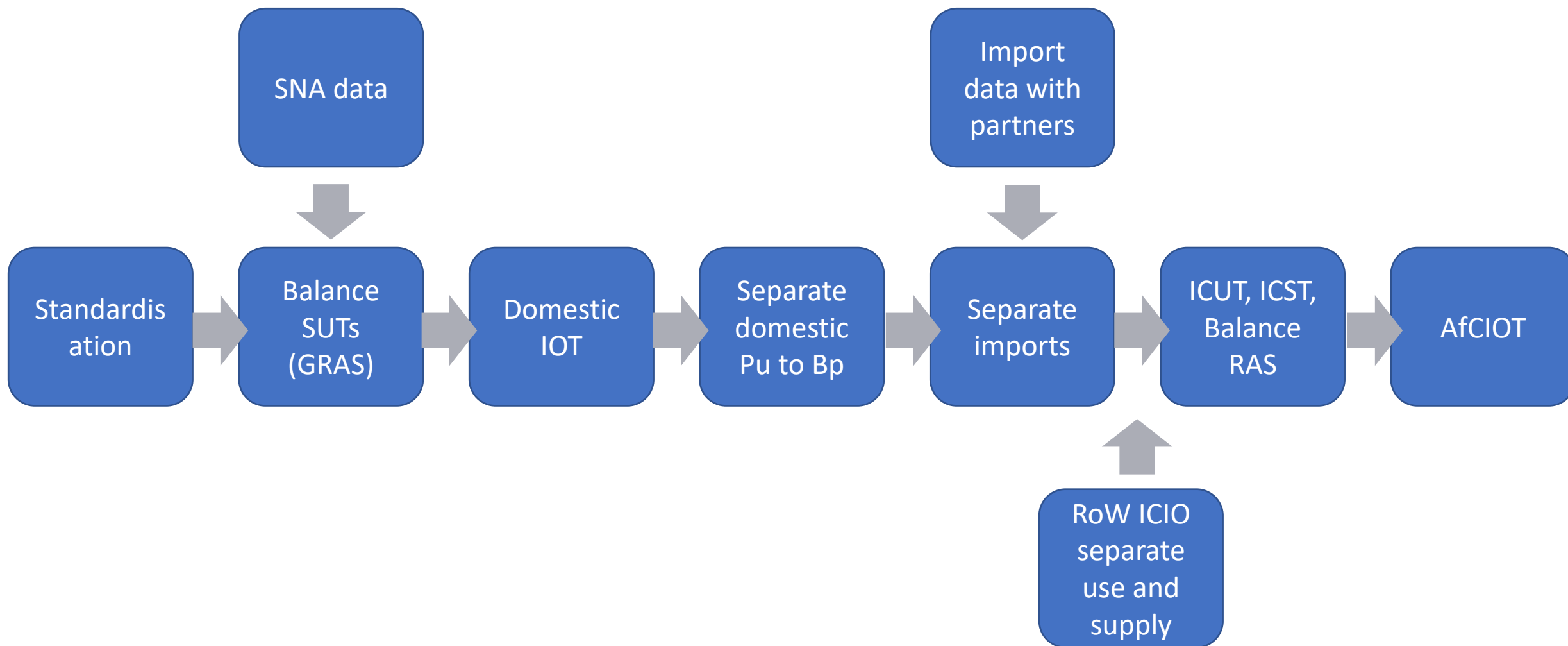
# African Continental Input-Output Table (AfCIOT)

- From National Accounts to TiVA indicators
- Via Supply and Use Tables (SUT)
- Each country using different classifications
- Different languages
- Transformations needed to standardize into 45 economic activities of TiVA



Image from [Gerd Altmann](#) at [Pixabay](#)

# AfCIOT Methodology



# Visualization dashboard

Overview

Home

Country profile

Model

Analysis

Indicators

Summary

Visualisation

Insights

Policy simulations

Metadata

Indicators

Dimensions

Industries

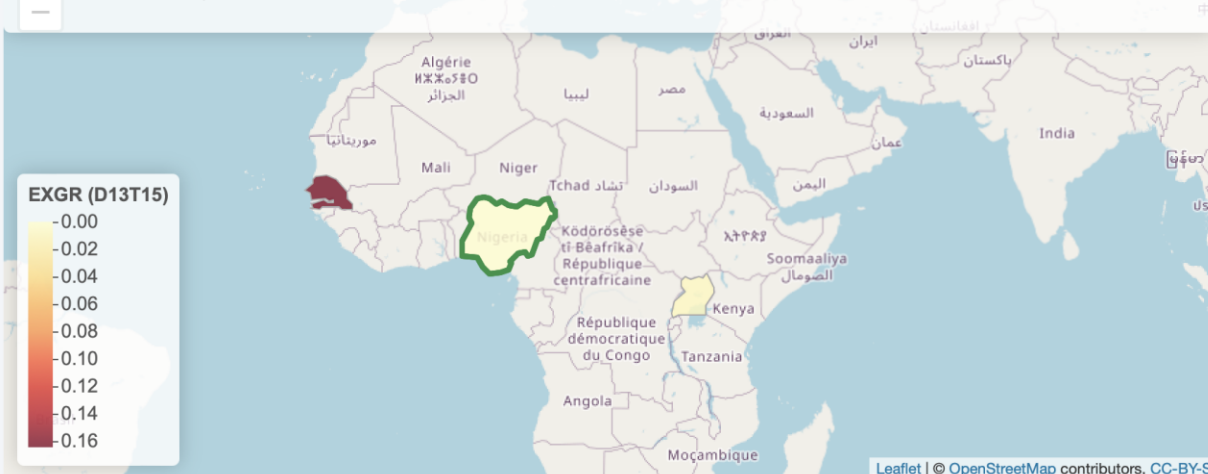
TIVA Environment Employment

Map Time Series Ranking Matrix

Selections instructions

Insights

Gross exports, (USD) 2018 by Import country; Export country: **Nigeria**; Export Products producing industry: **Textiles, textile products, leather and footwear**;



Selections

Explanations

Indicator

EXGR

Year

2018

Export country

NGA

Export Products producing industry

D13T15

Import country

Pick one or more countries

# The classification task

# The problem

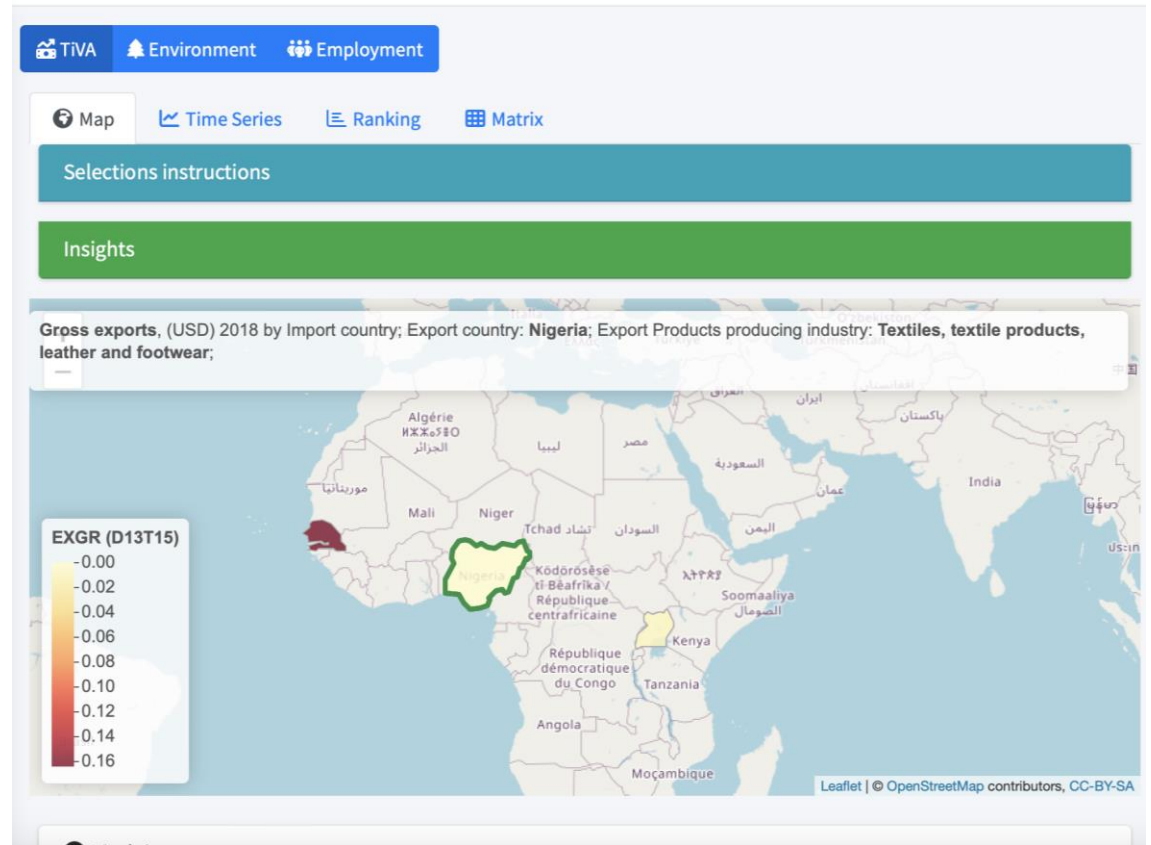
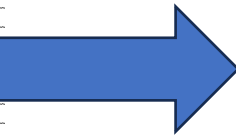
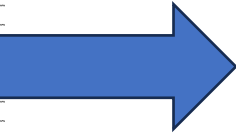
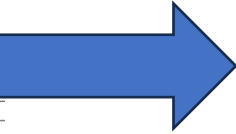
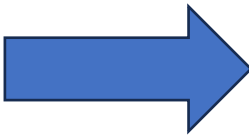
Match national products and industries descriptions with standard CPA 2.1 and ISIC 4 codes at two digits level





# Initial input vs. final result

A	B		C	D	
code			1	2	
			AGRICULTURE VIVRIERE	AGRICULTURE D'EXPORTATION	SYLURFOSE
1	PRODUITS DE L'AGRICULTURE				
2	PRODUITS AGRICOLES DES				
3	PROD. SYLVICULTURE EXP	Código dos produtos	Descrição dos produtos		Oferta total a preços de consumidor
4	PRODUITS DE L'ELEVAGE E				
5	PRODUITS DE LA PECHE				
6	PRODUITS D'EXTRACTION				
7	INDUSTRIES AGROALIMEN				
8	AUTRES INDUSTRIES MANU	01.01	Milho, trigo e outros cereais		75.634
9	PRODUCTION D'ELECTRICI	01.02	Mandioca		17.392
10	TRAVAUX DE CONSTRUCTI	01.03	Feijão		92.049
11	COMMERCE, SERVICES DE	01.04	Batata rena		23.627
12	TRANSPORTS ET ACTIVITE	01.05	Outros legumes, raízes e tubérculos		167.594
13	SERVICES D'INTERMEDIAT	01.06	Banana		17.412
14	AUTRES SERVICES MARCHA	01.07	Outros produtos agrícolas		29.451
15	SERVICES D'ADMINISTRAT	01.08	Gado bovino, outros gados e animais vivos		6.077
16	EDUCATION	01.09	Aves domésticas		16.266
17	SANTE ET ACTION SOCIALE	01.10	Leite, ovos e outros produtos de origem animal		11.050
18	ACTIVITES A CARACTERE CO	01.11	Carvão vegetal		26.623
19	ACTIVITES DES MENAGES E	01.12	Lenha e outros produtos da silvicultura		8.093
20	SIFIM	02.01	Peixe vivo, fresco e refrigerado		169.079
21	CORRECTION TERRITORIAL	02.02	Outros produtos da pesca		17.941
999	PRODUITS EN ATTENTES	03.01	Petróleo bruto e gás associado		4.568.457
		03.02	Serviços petrolíferos		147.949
		04.01	Diamantes		94.371
		04.02	Minerais metálicos e outros minerais não metálicos		24.091
		05.01	Carnes e cortes comestíveis de gados bovinos, de outros		61.198
		05.02	Carnes e cortes comestíveis de aves, frescas, refrigeradas		48.185
		05.03	Conservas e preparados de carne		25.929
		05.04	Preparados e conservas de peixes, outros produtos da		49.272
		06.01	Óleos e gorduras animais e vegetais		40.156
		07.01	Produtos lácteos		30.063
		08.01	Farinha e fuba de milho		139.363
		08.02	Farinha e fuba de mandioca		30.100
		08.03	Farinha de trigo		24.581



# Technological options

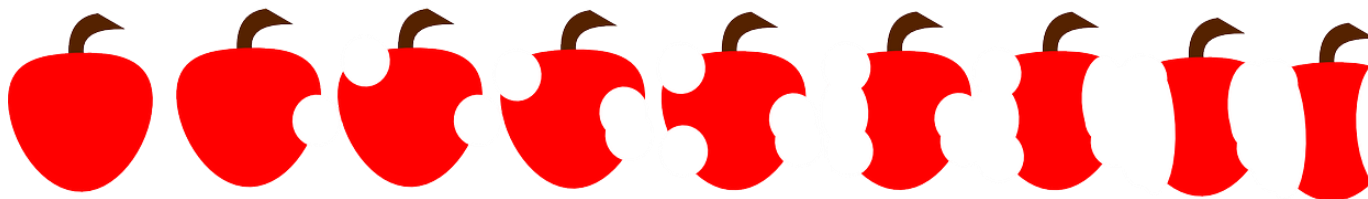
- Machine Learning
  - Worked well for 1 digit classification
  - Lack of labeled samples
  - No one-to-one matching
  - Needs large amount of data
- LLMs (chatGPT and friends)
  - Not enough mature (IMHO)
  - Ethical concerns
- Expert assignment
  - Subjective
  - Very slow



Image from [Jan Alexander](#) at [Pixabay](#)

# Chosen method: stepwise algorithm

0. Standards consolidation with their relationships
1. Text mining data curation
2. Exact matches
3. Less (lemmatized) string distance
4. Number of word matches
5. Expert review and final code allocation



# AfCIOT data

## Input classifications

- African countries
- Heterogenous levels of detail
- Different languages
- Initial set of countries, but further to come

country	language	# Categories
Guinea	fr	19
Kenya	en	154
Mali	fr	38
Mauritania	fr	26
Mauritius	en	59
Mozambique	pt	175
Rwanda	en	80

# Standards data

## Myriad of standards

- Target: CPA 2.1 & ISIC 4
- Target level: 2 digits
- Existing relationships
  - Other standards
  - Different versions
  - Different sources

### Correspondence tables

CPC Correspondence Tables					
FROM / TO	CPC prov	CPC Ver. 1.0	CPC Ver. 1.1	CPC Ver. 2	CPC Ver. 2.1
CPC prov	-			-	-
CPC Ver. 1.0		-		-	-
CPC Ver. 1.1			-		-
CPC Ver. 2	-	-		-	
CPC Ver. 2.1	-	-	-		-
ISIC Rev. 3			-	-	-
ISIC Rev. 3.1	-	-		-	-
ISIC Rev. 4	-	-	-		
BEC Rev. 5	-	-	-	-	
SITC Rev. 4	-	-	-		-

# Relationships consolidation+lemmatization

code	name	language	standard	cpa21
A	agriculture, forestry and fishing	en	isic4	
01	crop and animal production, hunting and related service activities	en	isic4	
011	growing of non-perennial crops	en	isic4	
0111	growing of cereals (except rice), leguminous crops and oil seeds	en	isic4	01.11.11
0111	growing of cereals (except rice), leguminous crops and oil seeds	en	isic4	01.11.11
0111	growing of cereals (except rice), leguminous crops and oil seeds	en	isic4	01.11.12
0111	growing of cereals (except rice), leguminous crops and oil seeds	en	isic4	01.11.12
0111	growing of cereals (except rice), leguminous crops and oil seeds	en	isic4	01.11.20
0111	growing of cereals (except rice), leguminous crops and oil seeds	en	isic4	01.11.20
0111	growing of cereals (except rice), leguminous crops and oil seeds	en	isic4	01.11.31

To classify a new name:

1. Look for exact matches
2. Lemmatize new name
3. Look for partial matches
4. Get target code

code	name	language	standard	tokens	lemmas
A	agriculture, forestry and fishing	en	isic4	agriculture fishing forestry	agriculture fishing forestry
01	crop and animal production, hunting and related service activities	en	isic4	activities animal crop hunting production related service	activity animal crop hunt production relate service
011	growing of non-perennial crops	en	isic4	crops growing nonperennial	crop grow nonperennial
0111	growing of cereals (except rice), leguminous crops and oil seeds	en	isic4	cereals crops growing leguminous oil rice seeds	cereal crop grow leguminous oil rice seed
0112	growing of rice	en	isic4	growing rice	grow rice
0113	growing of vegetables and melons, roots and tubers	en	isic4	growing melons roots tubers vegetables	grow melon root tuber vegetable
0114	growing of sugar cane	en	isic4	cane growing sugar	cane grow sugar
0115	growing of tobacco	en	isic4	growing tobacco	grow tobacco
0116	growing of fibre crops	en	isic4	crops fibre growing	crop fibre grow
0119	growing of other non-perennial crops	en	isic4	crops growing nonperennial	crop grow nonperennial

Implementation

# R Statistical Software

## R

- Open source
- Statistical software and programming language
- ML, TM, and any modelling method



## Shiny apps

- Framework for building reactive web applications with just R regular code





# Scripts for building the standards data frames

The screenshot shows the RStudio interface with the following elements:

- Editor:** Contains R code for loading a CSV file and writing an RDS file. The code is as follows:

```
10 library(tidyr)
11
12 source("R/fun_correspondence.R")
13
14 ## - CLASSIFICATIONS ----
15 ### 1. CPC 2.1 (en) ----
16 cpc21_en <- structure(
17   read_csv("data/raw_data/standards/CPC_Ver_2_1_english_structure.txt",
18     locale = locale(encoding = "ISO-8859-1"),
19     col_names = c("code", "name"),
20     col_types = "cc",
21     skip = 1),
22   language = "en",
23   standard = "cpc21"
24 )
25 str_std <- "cpc21_en"
26 write_rds(eval(parse(text = str_std)),
27   paste0("data/clean_data/standards/", str_std, ".rds"))
28
29 ### 2. CPA 2.1 (multilanguage) ----
```
- Search Bar:** Shows a search for "##" with options for "Next", "Prev", "All", and "Replace".
- Options:** Includes checkboxes for "In selection", "Match case", "Whole word", "Regex", and "Wrap" (checked).
- Table of Contents (Sidebar):**
  - CLASSIFICATIONS**
    1. CPC 2.1 (en)
    2. CPA 2.1 (multilang...
    3. CPC 2 (en)
    4. CPC 1.1 (en, fr)
    5. CPC 1.0
    6. CPC prov
    7. ICIO
    8. ISIC 4
    9. ISIC 3.1
    10. ISIC 3
    11. HS5
  - CONVERSION TABLES**
    1. CPA 2.1 to CPC 2.1
    2. CPA 2.1 to CPC 2
    3. CPA 2.1 to CPC 1.1
    4. CPA 2.1 to CPC 1.0
    5. CPA 2.1 to CPC prov
    6. CPA 2.1 to ICIO
    7. CPA 2.1 to ISIC 4
    8. CPA 2.1 to ISIC 3.1
    9. ISIC 3 to CPC 1.0
    10. CPA 2.1 to HS5
  - Bind and join data
  - Save clean data
- Status Bar:** Shows "3:1 (Top Level)" and "R Script".

# Core function

```
source("R/matching_algorithm.R")
pnew <- read_rds("data/clean_data/pnew.rds") |>
  filter(country == "Guinea") |> slice(1:3)

std_corresp <- read_rds("data/clean_data/cpa21_corresp.rds")
lstandards <- read_rds("data/model_data/lstandards.rds")

test <- match_names(dfnew = pnew,
  newcountry = "Guinea",
  target = "cpa21",
  std_corresp = std_corresp,
  lstandards = lstandards,
  model_path = "data/model_data",
  fpath = "R",
  save = FALSE)
```

dfnew: any arbitrary data frame with codes and names

# Shiny app

## Check industry and product matching

Existing files

Upload an Excel file

IndustryMatch.xlsx

Country

Language

Target standard

Target digits

### Matches found

Show  entries

Search:

	quality	code	name	target_code	matched_name	matched_code	match_std	target_digits_code	step
1		1	agriculture forestry and fishing	01;02;03	agriculture forestry and fishing	A	isic4	A	1
2		10	wholesale and retail trade repair of motor vehicles and motorcycles	45;46;47	wholesale and retail trade repair of motor vehicles and motorcycles	G	isic4	G	1
3		11	transportation and storage	49;50;51;52;53	transportation and storage	H	isic4	H	1
4		12	accommodation and food service activities	55;56	accommodation and food service activities	I	isic4	I	1

[https://lcano.com/share/tmp/screencast\\_classification\\_app\\_V1.mp4](https://lcano.com/share/tmp/screencast_classification_app_V1.mp4)

# Discussion

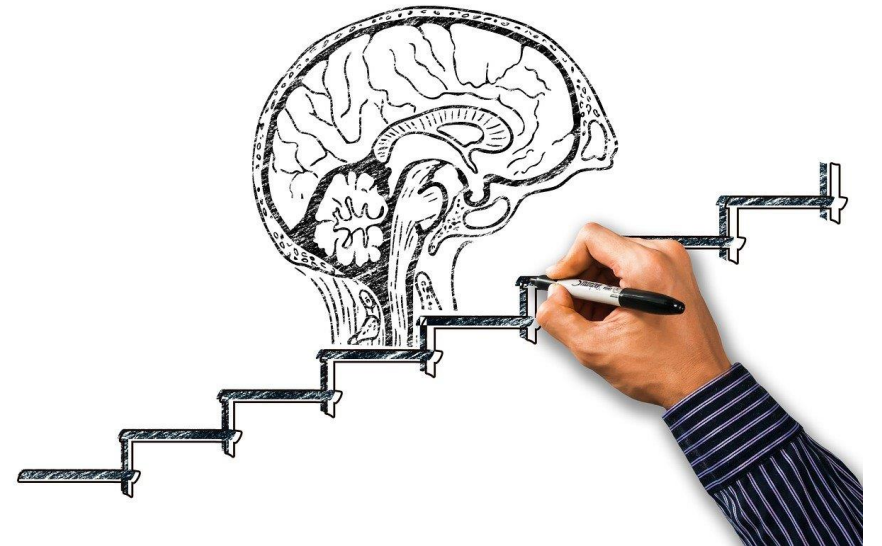
# Conclusions

- Standardization of categories appears in the initial steps of the indicators building pipeline
- Wrong classifications may lead to wrong indicator values, hence in wrong information for decision making (garbage in-garbage out)
- Automatic classification and matching is a major challenge in economic statistics
- Hybrid models are useful in many situations and provide robustness



# Room for improvement

- Add more standards and their relationships
- Include models that learn from the experts' input
- Explore how to include LLMs for difficult matches
- Add rule-based controls to avoid clear missclassifications and manage one-to-many and many-to-many relationships
- Try further Data Science techniques, e.g., unsupervised classification or Bayesian models.



# An expert system for semi-automated classification of product and industry descriptions

Emilio L. Cano

Rey Juan Carlos University, Madrid, Spain

# Thanks!

[emilio.lopez@urjc.es](mailto:emilio.lopez@urjc.es)



# An expert system for semi-automated classification of product and industry descriptions

E.L. Cano<sup>1,2</sup>, A.Deveza<sup>2</sup>, X. Gong<sup>2</sup>, D. Boko<sup>2</sup>, E. Keeble<sup>2</sup>, P. Debanes<sup>2</sup>

<sup>1</sup>Data Science Laboratory, Universidad Rey Juan Carlos

<sup>2</sup>United Nations Economic Commission for Africa (UNECA)

**Presenter: Emilio L. Cano**, Associate Professor at Rey Juan Carlos University, Madrid, Spain; former Consultant of UNECA

# Thanks!

[emilio.lopez@urjc.es](mailto:emilio.lopez@urjc.es)