

Twenty-third meeting of the Executive Committee
of the Statistical Conference of the Americas of the ECLAC
28 and 29 August 2024



AI for official statistics

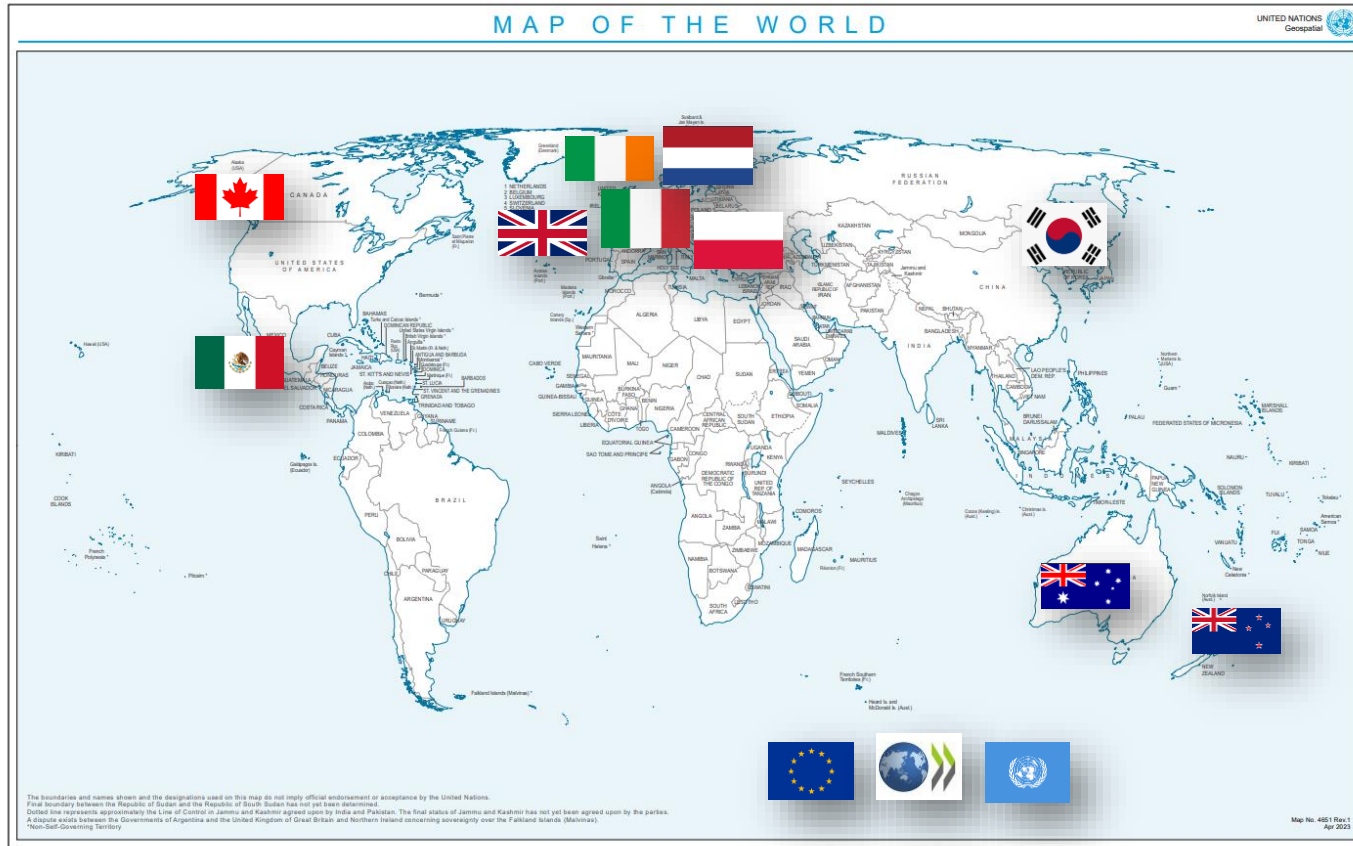
experiences and lessons learned based on HLG-MOS initiatives

InKyung Choi

Statistical Management and Modernisation Unit
United Nations Economic Commission for Europe (UNECE)

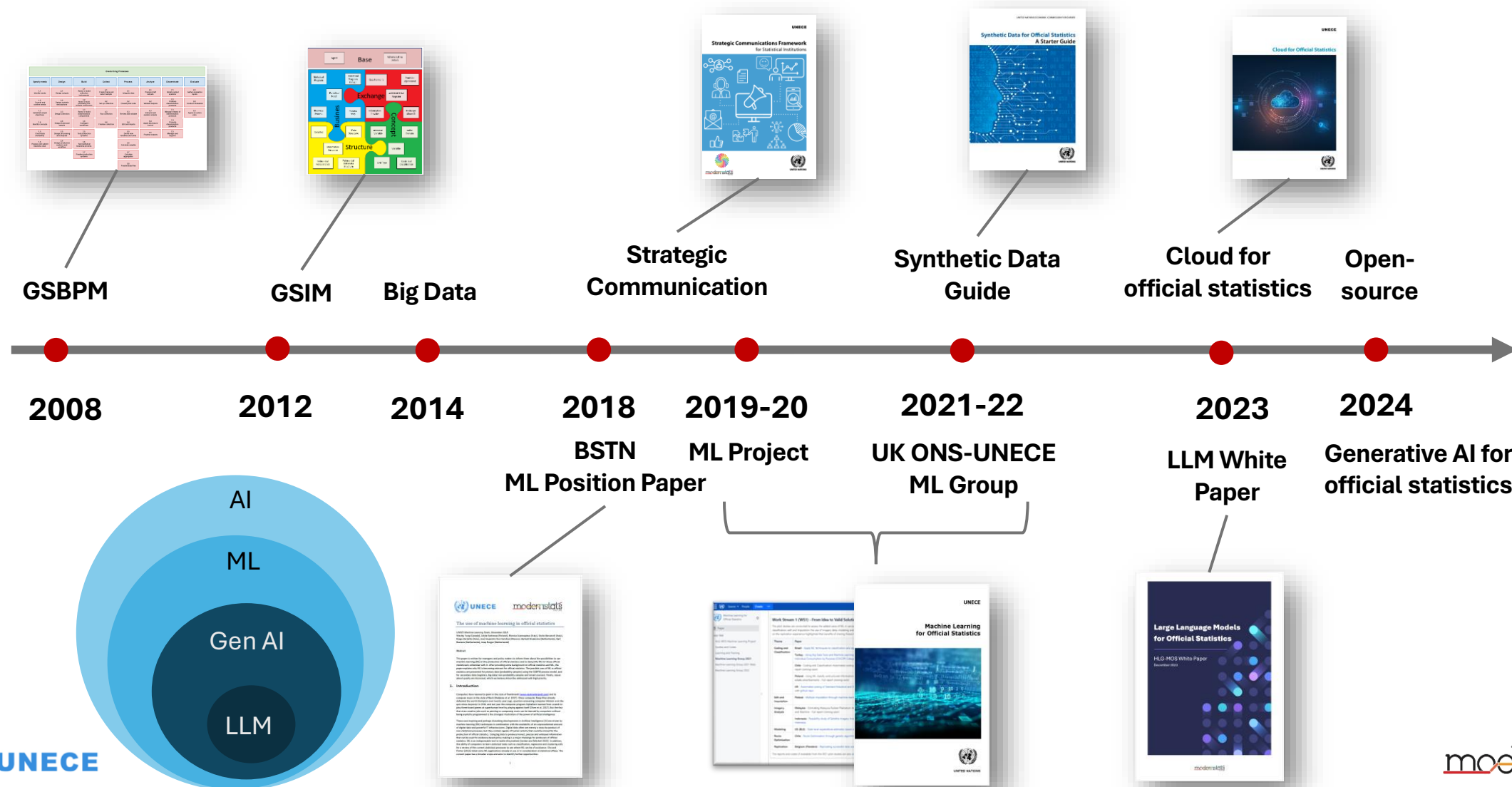
Disclaimer: The views and opinions expressed in this presentation are those of the author and do not necessarily reflect the official policy or position of UNECE, and it does not mean endorsement of any services or tools

HLG-MOS

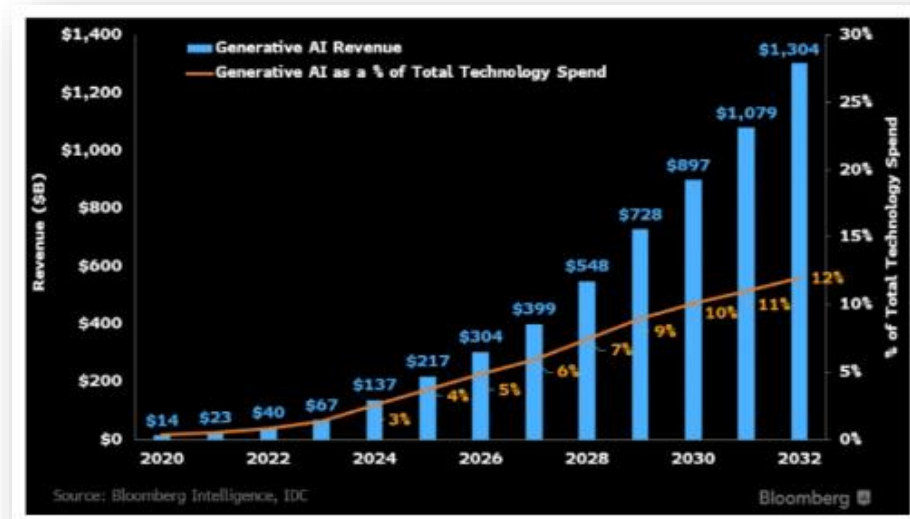


- **High-Level Group for the Modernisation of Official Statistics (HLG-MOS)** was established in 2010 by the Conference of European Statisticians (CES) **to drive and steer the modernisation for the official statistics**
- Consists of heads of 13 national statistics organisations (NSOs) and international organisations (Australia, Canada, Ireland, Italy, Mexico, Netherlands, New Zealand, Poland, Republic of Korea, UK, Eurostat, OECD and UNECE)
- Currently co-chaired by Netherlands and Poland

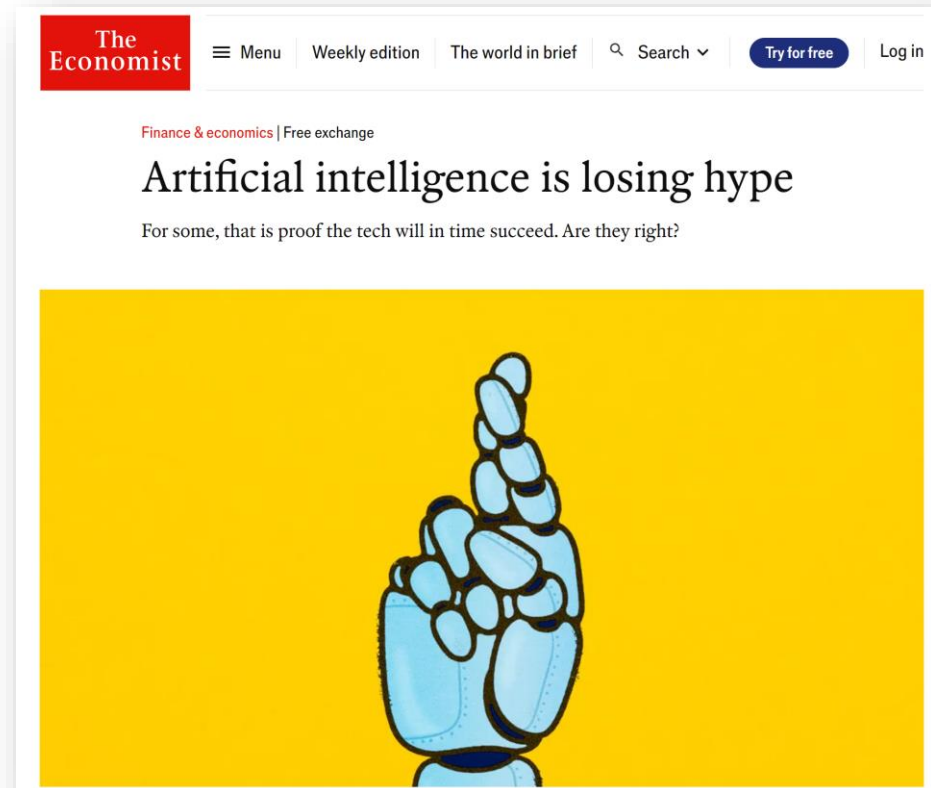
HLG-MOS and AI



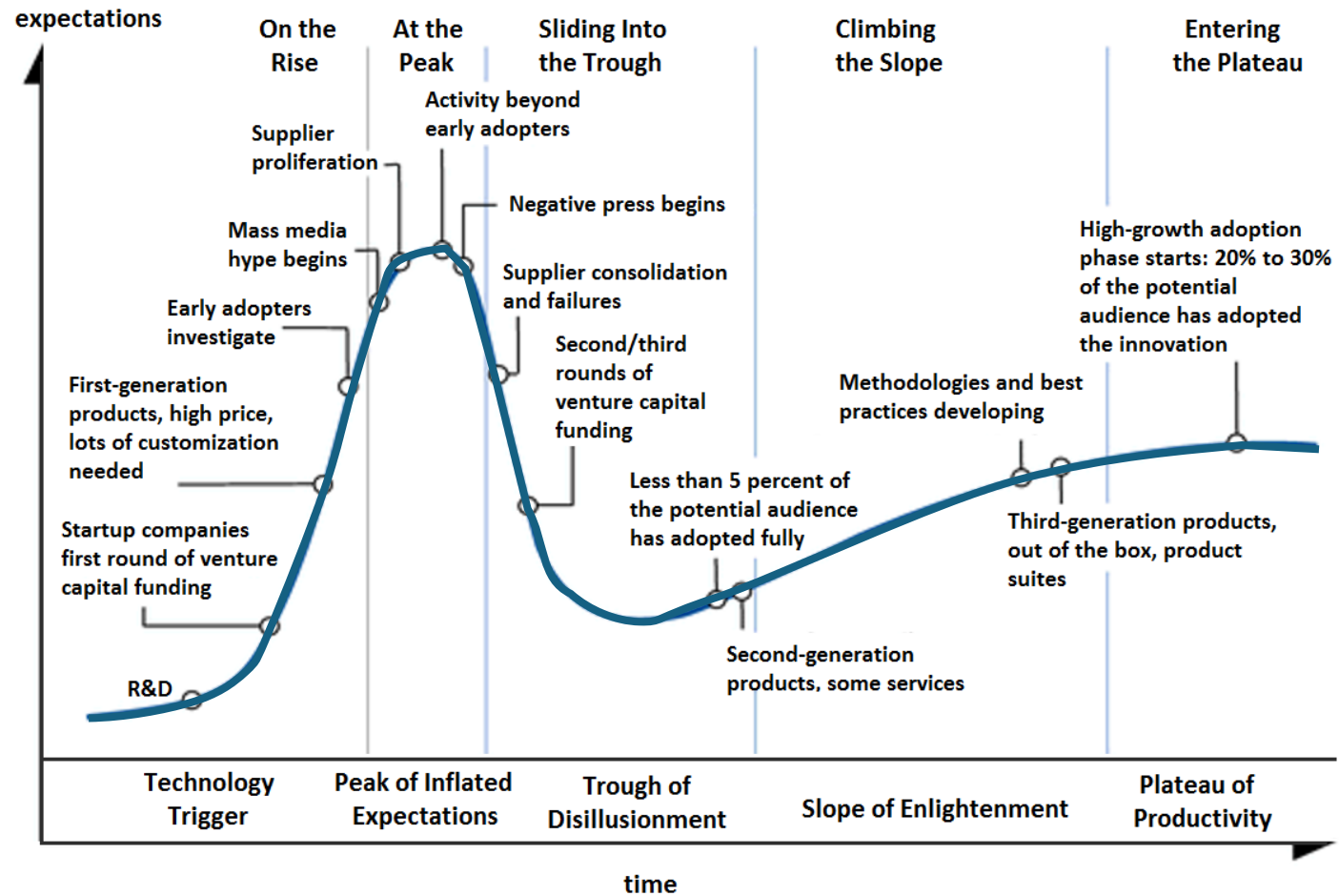
Rise of AI



Rise of AI

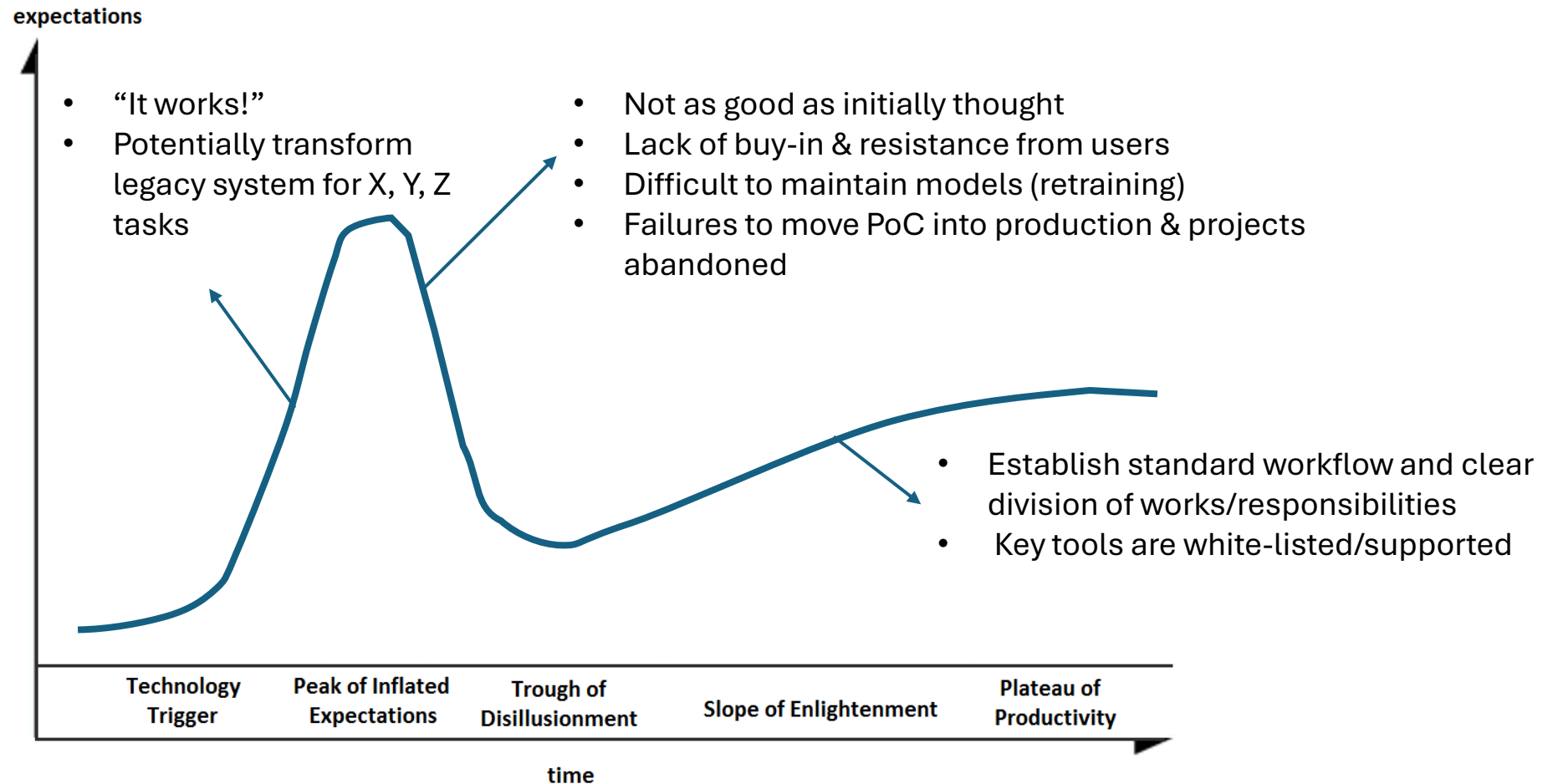


Rise of AI



Gartner general hype cycle for technology ([Wikipedia](#))

Rise of AI



AI for modernisation of NSOs

FIGURE 1
From the printing press to the global internet, technology has evolved, and human societies with it

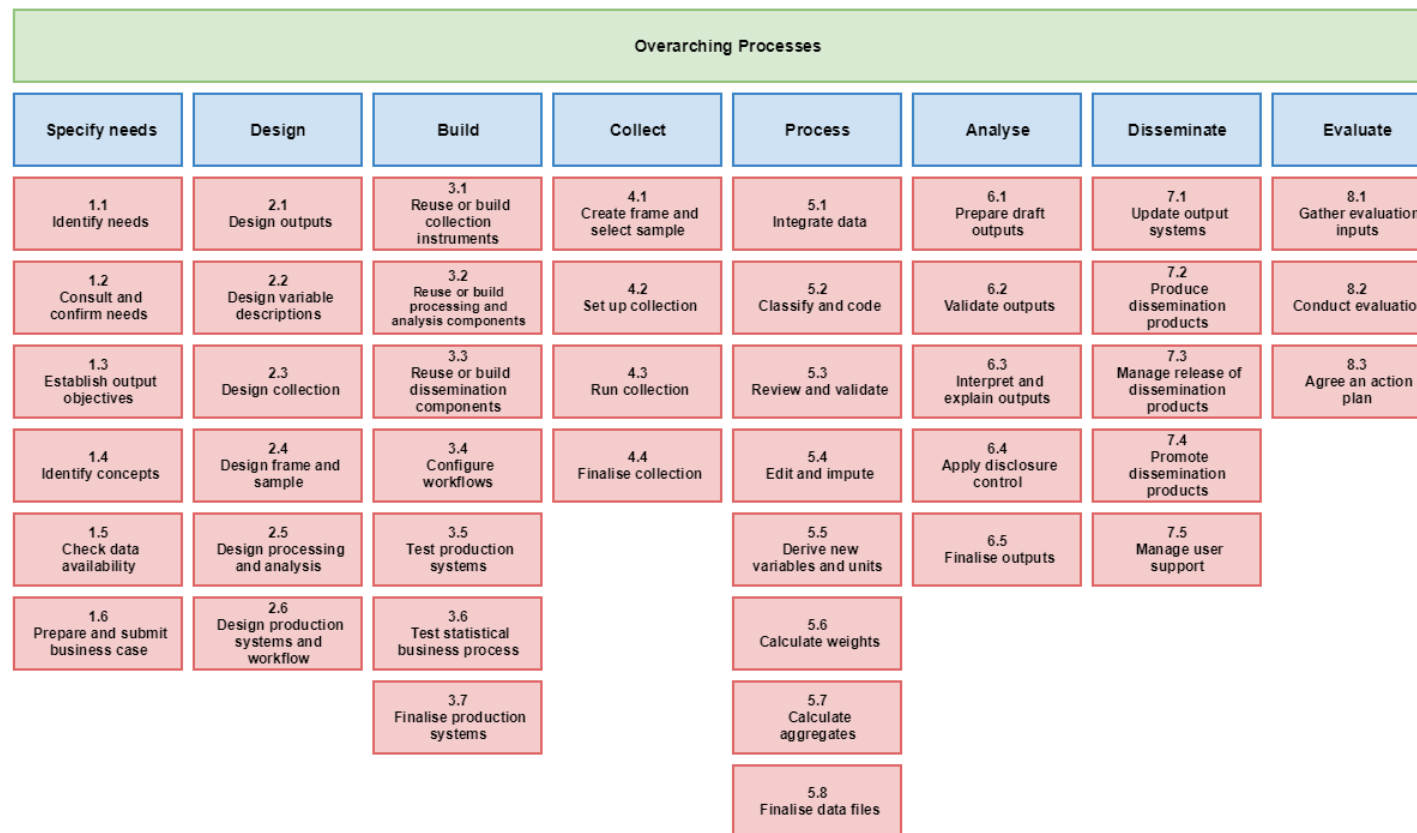


How technology has changed the way we work.

Image: Barclays

- Enhance productivity and efficiency
 - Improve service delivery and user experiences
 - Offer new products
- But not without challenges

Application areas



Generic Statistical Business Process Model (GSBPM)

Application areas



- Areas with manual and repetitive tasks can be automated with help of machine learning

Text classification

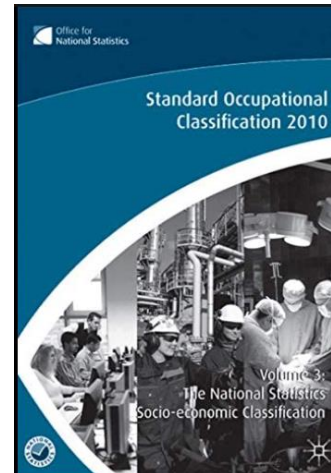


Survey response “I am a cook at a French restaurant”

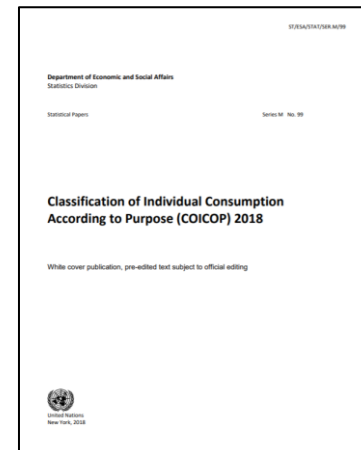


SOC “35-2014.00 Cooks, Restaurant”

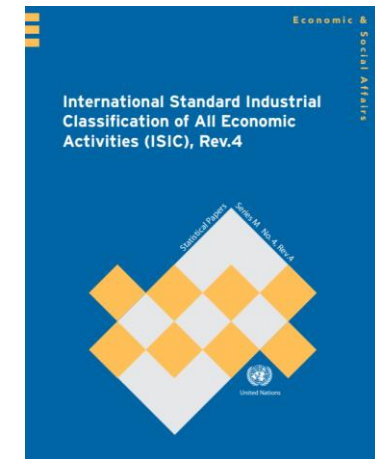
5.2
Classify and code



SOC for occupation



COICOP for product



ISIC for activity

Text classification

	produkt	kategoria
0	Lisner Herring fillets with chili and parsley ...	Other preserved or processed fish and seafood-...
1	Herring fillets in vegetable oil and ...	Other preserved or processed fish and seafood-...
2	Orasi - Soy drink with vanilla	Other milk products
3	Pate With Celery	Dried vegetables, other preserved or processed...
4	Hochland - Four Cheese cream cheese melted in ...	Cheese and curd
...
17094	Be Raw! - Energy bar	Confectionery products
17095	Bobovita - Apple dessert and sweet carrots	Baby food
17096	Hortex Vegetables for a pan with spinach 350 g	Frozen vegetables other than potatoes and other...
17097	Czajka Valley - NFC 100% pressed apple juice	Fruit and vegetable juices
17098	Hipp dessert. Strawberries and raspberries wit...	Baby food

17099 rows x 2 columns

Product descriptions in text

COICOP Code

Builds

ML
Model

Uses

	product description	predicted codes	probability	True codes
0	Lisner Herring fillets with chili and parsley ...	Other preserved or processed fish and seafood-...	0.971882	Other preserved or processed fish and seafood-...
1	Herring fillets in vegetable oil and ...	Other preserved or processed fish and seafood-...	0.933938	Other preserved or processed fish and seafood-...
2	Orasi - Soy drink with vanilla	Other milk products	0.901197	Other milk products
3	Pate With Celery	Other meat preparations	0.535415	Dried vegetables, other preserved or processed...
4	Hochland - Four Cheese cream cheese melted in ...	Cheese and curd	0.979830	Cheese and curd
...
2995	Edam cheese piece	Cheese and curd	0.930995	Cheese and curd
2996	Tea sir Williams royal taste yerba mate dame 1...	Tea	0.919722	Tea
2997	Beef tripe in broth	Other meat preparations	0.568275	Other meat preparations
2998	Lettuce - butterhead	Fresh or chilled vegetables other than potatoe...	0.760290	Fresh or chilled vegetables other than potatoe...
2999	Serenada Radamer cheese	Cheese and curd	0.934961	Cheese and curd

3000 rows x 4 columns

New set of product descriptions
in text

Predicted COICOP
Code

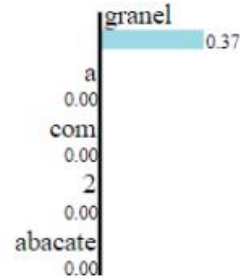
Text classification

Table 4. NAICS and NOC model metrics for the CCHS production pipeline. Overall Accuracy, F1, Precision, and Recall were calculated on the 'Interactive: QC Sample'. (A) Q3 Record number = 343. (B)

A	Measure
	Error Rate (%)
	Weighted Average F1-Score
	Weighted Average Precision
	Weighted Average Recall

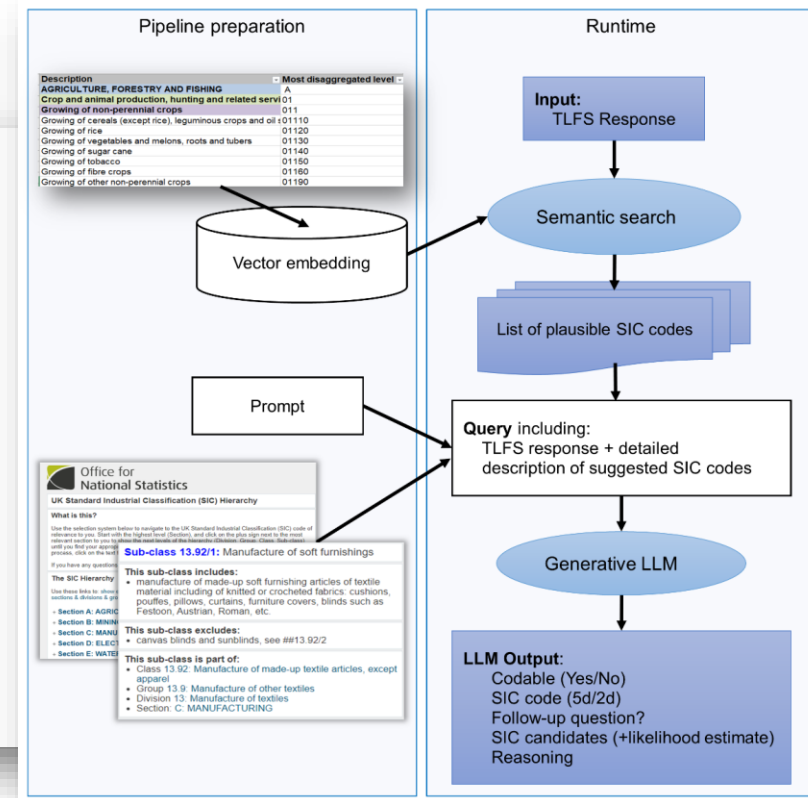
B	Metric
	Error Rate (%)
	Weighted Average F1-Score
	Weighted Average Precision
	Weighted Average Recall

(a)
 True: Abacate --> Pred: Goiaba | Prob: 0.05
 NOT Goiaba Goiaba



Text with highlighted words

abacate a **granel** com 2 unidades

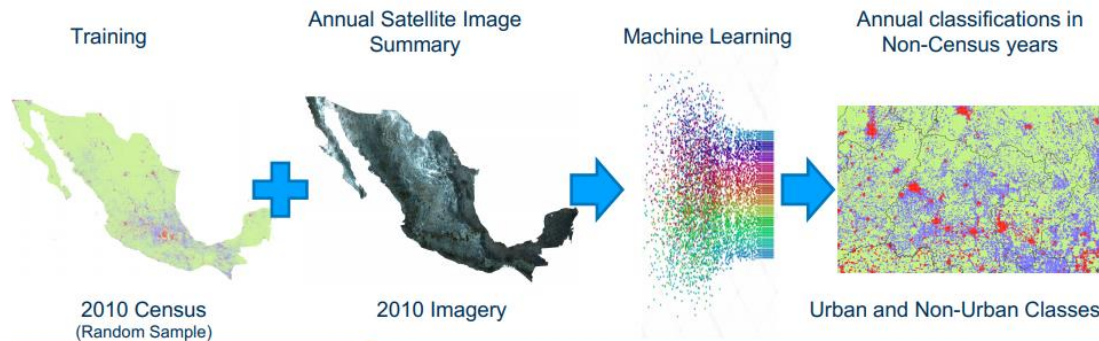


UK ONS ClassifAI

Imagery classification

Objective of this Imagery Pilot Project (Practical Application)

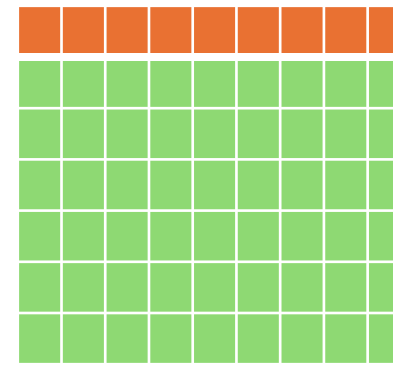
Expand the use of imagery data in the production of official statistics through the further development of knowledge and sharing of ML solutions and practices.



Imagery Pilot Project

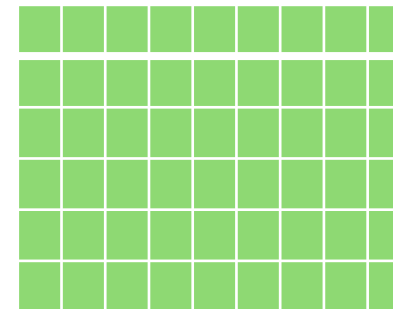
[INEGI \(ML Project\)](#)

1,975,719 (1km x 1km) grid cells



40,000 done by human

.....



based on this data
1,935,719 done by ML

Other examples in data processing

Record Matching Problem

Census file

Person	Company	Age	Narrative
John Smith	ACME Inc.	25	Car accident
Susan Carter	Tree Co.	74	Hit by tree
Hank Long	Big Box	34	Homicide

Source document file

Person	Company	Age	Narrative
Suzy E. Carter	Joe's	74	Hit by tree
Frank Garcia	Cola	74	Hit by tree
Jonathan Smith	A.C.	25	Car accident
Henry Long	BB F	34	Homicide

Use case – Turkish Statistical Institute

- Project 3 – Create a framework for government R&D survey

```

    graph LR
      Terms((Terms)) --- WebScraping((Web Scraping))
      URLList((URL List)) --- WebScraping
      WebScraping --- ListLinks((List of links containing the terms with their frequency))
  
```

Imputation

- Cache awareness and out-of-core computing
- Tree pruning using depth-first approach
- Parallelized tree building

Turkish Statistical Institute (ML Group)

Statistisk sentralbyrå
Statistics Norway

- Record matching
- Imputation
- Create survey framework
- Identify survey respondents to follow up

Statistics Norway (SDE Expert Meeting 2020)

Coding and IT development

The screenshot shows a web-based interface for a 'SAS to R code Assistant (Pilot)'. The interface is split into two panes. The left pane contains SAS code, and the right pane contains the corresponding R code. The SAS code defines a dataset 'TEMP' with columns for ID, NAME, SALARY, and DEPARTMENT, and calculates a commission based on salary. The R code uses 'data.frame' to create the same dataset and calculates the commission using a vectorized operation.

```

SAS Code:
DATA TEMP;
INPUT ID $ NAME $ SALARY DEPARTMENT $;
comm = SALARY*0.25;
LABEL ID = 'Employee ID' comm = 'COMMISSION';
DATALINES;
1 Rick 623.3 IT
2 Dan 515.2 Operations
3 Michelle 611 IT
4 Ryan 729 HR
5 Geary 843.25 Finance
6 Nina 578 IT
7 Simon 632.8 Operations
8 Guru 722.5 Finance
;
RUN;

R Code:
1 # Create a data frame with the given data
2 TEMP <- data.frame(ID = c("1", "2", "3", "4", "5", "6", "7", "8"),
3 NAME = c("Rick", "Dan", "Michelle", "Ryan", "Geary", "Nina", "Simon", "Guru"),
4 SALARY = c(623.3, 515.2, 611, 729, 843.25, 578, 632.8, 722.5),
5 DEPARTMENT = c("IT", "Operations", "IT", "HR", "Finance", "IT", "Operations", "Finance"),
6
7 # Calculate the commission
8 TEMP$comm <- TEMP$SALARY * 0.25
9
10 # Assign labels to the columns
11 names(TEMP)[names(TEMP) == "ID"] <- "Employee ID"
12 names(TEMP)[names(TEMP) == "comm"] <- "COMMISSION"

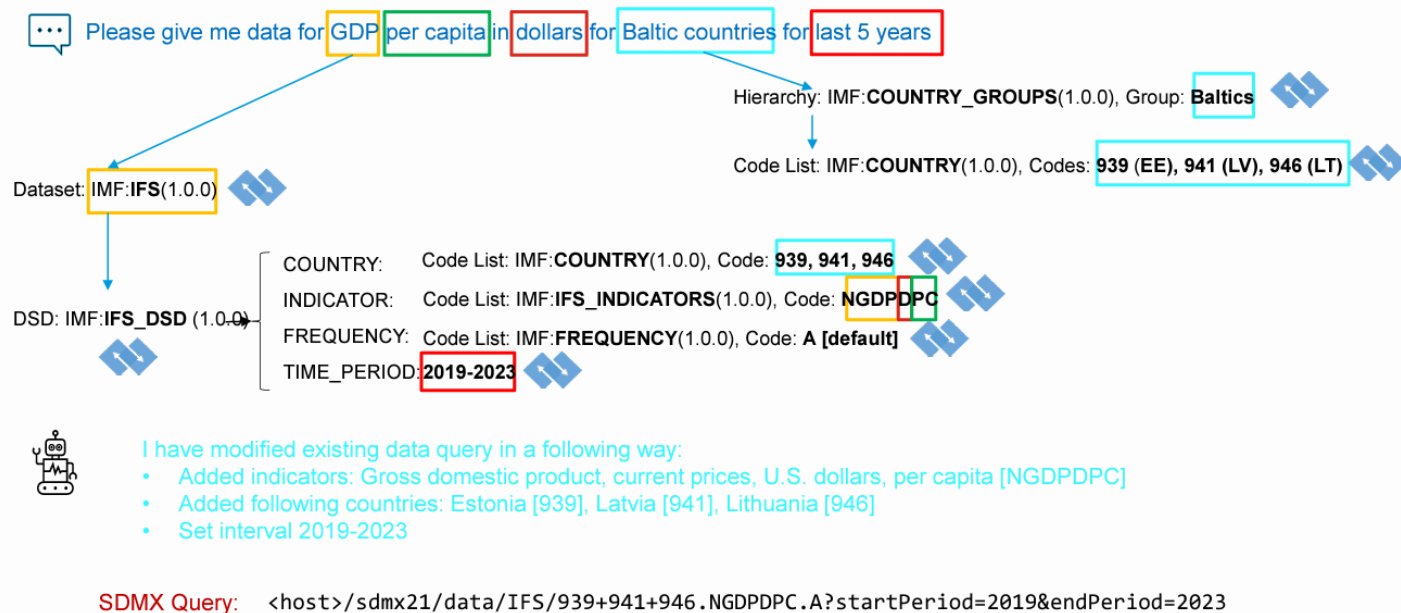
```

Central Statistics Office of Ireland (Generative AI Project)

- Use case that has impacts in many working areas in NSOs
- Particularly important these days as many of them are transitioning to open-source such as R and python
- Seen positive side-effect: standardisation and good coding practices

Data dissemination

Stat-GPT: How it works



IMF (CES 2024)

- Provision of statistics is fundamental role of NSOs
- Data platform is not easy to navigate
- Help users find data with natural language, improving user experience

Information search and text generation

The screenshot shows a web interface for document upload and AI text generation. On the left, there is a sidebar with the following sections:

- Upload your own documents**: Includes a URL input field with the value `https://unctad.org/system/files/official-document/rmt2023_en.pdf`, a file upload area with a "Browse files" button, and a "Process corpus" button.
- Model parameters**: Includes a "Which LLM" dropdown menu set to "mistral-docsgpt", a "Which corpus" dropdown menu set to "temporary_daniel_hopp", an "Advanced model parameters" dropdown menu, a "Reinitialize model" button, and a "Reset model's memory" button.

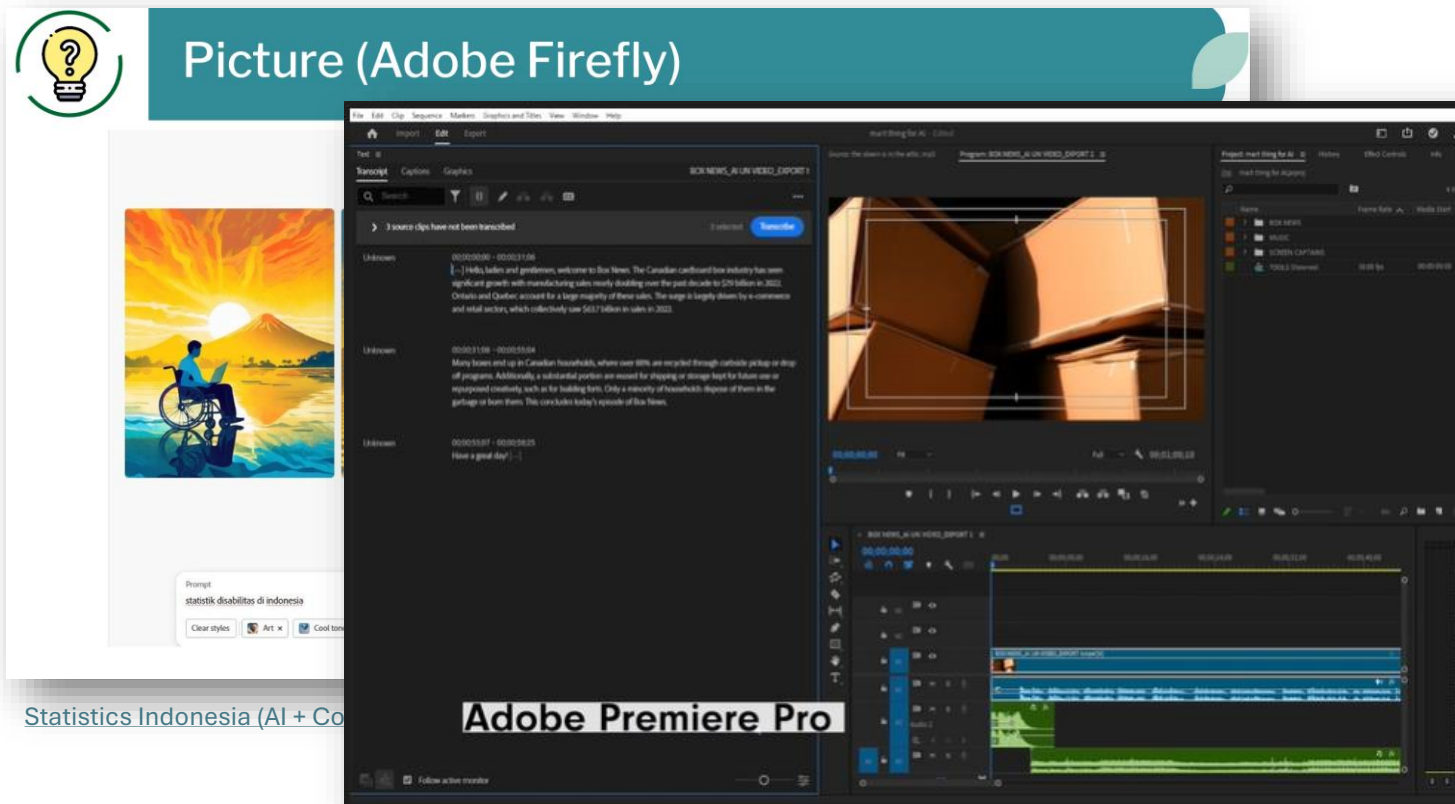
On the right, the main content area displays the following:

- A header: "Give 3 key takeaways from the 2023 Review of Maritime Transport 2024-04-19 10:55" (2024-04-19 10:56).
- A list of three key takeaways:
 1. The global economy remains fragile, with high inflation, geopolitical risks, and debt burdens constraining growth prospects. The outlook for global trade is pessimistic, with expected annual growth rates of 1.7% in 2023 and -0.6% for the second quarter of 2023.
 2. The shipping industry continues to face disruptions and uncertainty, with shifts in the geography of maritime trade, continued disruptions to shipping and ports, and high fuel costs impacting energy and food security. The outlook for shipping in 2023 is uncertain, with decarbonization efforts and fuel transition pathways presenting both opportunities and challenges for the industry.
 3. The Review highlights the importance of digitalization and electronic trading systems in facilitating maritime trade and transport, with the adoption of such systems expected to increase efficiency and reduce costs for stakeholders. However, the Review also notes the need for continued efforts to address cybersecurity risks and ensure the interoperability and compatibility of different systems.
- A "Sources:" section (2024-04-19 10:56).
- A header: "Elaborate on global trade growing at a slower pace" (2024-04-19 10:57).
- A paragraph of text: "According to the 2023 Review of Maritime Transport, global merchandise trade grew by only 2.7% in 2022, which represents a significant slowdown from the 9.4% rebound witnessed in 2021. This slowdown can be attributed to several factors, including high inflation, geopolitical risks, and debt burdens constraining economic growth, as well as disruptions to shipping and ports caused by the COVID-19 pandemic and the war in Ukraine. The outlook for global trade in 2023 is also pessimistic, with an expected annual growth rate of 1.7%, and trade growth is projected to..."
- A query input field: "Query 'mistral-docsgpt' contextualized on 'temporary_daniel_hopp' corpus" with a right-pointing arrow.

- Can also be used for internal information search & research

[UNCTAD \(CES2024\)](#)

Communication



Statistics Indonesia (AI + Co)

Adobe Premiere Pro

Statistics Canada (AI+Comm Task Team under HLG-MOS)

- Growing emphasis on tailored communication
- Image/audio/video generation
- Generation of communication materials in line with certain styles and tone

Challenges and some lessons learned

- Challenges about AI

Different methodologies

Black-box

Different programming languages

- ... compound by being **national statistics organisations**

Higher stake for error
Procurement limitation

Alignment with quality
standards

Experiment vs. production
Change management

Quality

United Nations National Quality Assurance Framework quality principles and supporting Fundamental Principles of Official Statistics

Quality principles	Fundamental Principles of Official Statistics									
	1	2	3	4	5	6	7	8	9	10
Level A: Managing the statistical system										
1: Coordinating the national statistical system										
2: Managing relationships with data users, data providers and other stakeholders	*								*	
3: Managing statistical standards										
Level B: Managing the institutional environment										
4: Assuring professional independence	○	*								
5: Assuring impartiality and objectivity	*	○	○	○	○					
6: Assuring transparency			*							
7: Assuring statistical confidentiality and data security										*
8: Assuring commitment to quality		*								
9: Assuring adequacy of resources	○									
Level C: Managing statistical processes										
10: Assuring methodological soundness		*				○				
11: Assuring cost-effectiveness									*	
12: Assuring appropriate statistical procedures		*				○				
13: Managing the respondent burden									*	
Level D: Managing statistical outputs										
14: Assuring relevance	*		○			○				
15: Assuring accuracy and reliability	*					○				
16: Assuring timeliness and punctuality	*					○				
17: Assuring accessibility and clarity	*		○							
18: Assuring coherence and comparability	*		○							○
19: Managing metadata			*							○

A quality framework for statistical algorithms

Article type: Research Article

Authors: Yung, Wesley^{a,*} | Tam, Siu-Ming^b | Buelens, Bart^c | Chipman, Hugh^d | Dumpert, Florian^e | Ascari, Gabriele^f | Rocci, Fabiana^g | Burger, Joep^h | Chol, Inkyungⁱ

Affiliations: [a] Statistics Canada | [b] National Institute of Applied Statistical Research, University of Wollongong, Wollongong, NSW, Australia | [c] Vlaamse Instelling voor Technologisch Onderzoek (VITO) | [d] Department of Mathematics and Statistics, Acadia University, Canada | [e] Federal Statistical Office of Germany | [f] Italian National Institute of Statistics | [g] Statistics Netherlands | [h] United Nations Economic Commission for Europe

Correspondence: [*] Corresponding author: Wesley Yung, Statistics Canada. Tel.: +1 613 404 2203; Fax: +1 613 951 1462; E-mail: Wesley.Yung@canada.ca.

Abstract: As national statistical offices (NSOs) modernize, interest in integrating machine learning (ML) into official statisticians' toolbox is growing. Two challenges to such an integration are the potential loss of transparency from using "black-boxes" and the need to develop a quality framework. In 2019, the High-Level Group for the Modernisation of Official Statistics (HLG-MOS) launched a project on machine learning with one of the objectives being to address these two challenges. One of the outputs of the HLG-MOS project is a Quality Framework for Statistical Algorithms (QF4SA). While many quality frameworks exist, they have been conceived with traditional methods in mind, and they tend to target statistical outputs. Currently, machine learning methods are being looked at for use in processes producing intermediate outputs, which lead to a final statistical output. Therefore, the QF4SA does not replace existing quality frameworks; it complements them. As the QF4SA targets intermediate outputs and not necessarily the final statistical output, it should be used in conjunction with existing quality frameworks to ensure that high-quality outputs are produced. This paper presents the QF4SA, as well as some recommendations for NSOs considering the use of machine learning in the production of official statistics.

Keywords: Machine learning, official statistics, explainability, reproducibility

DOI: 10.3233/SII-210875

Journal: Statistical Journal of the IAOS, vol. 38, no. 1, pp. 291-308, 2022

Published: 21 March 2022

- Accuracy
- Explainability
- Reproducibility
- Timeliness
- Cost-effectiveness

UN National Quality Assurance Framework

AI in production

Stakeholder buy-in

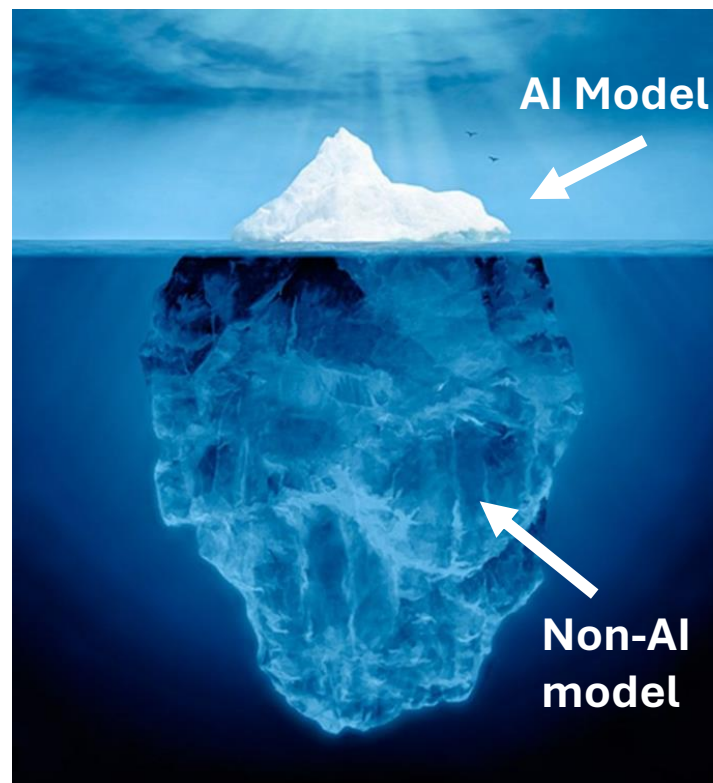
Model serving

Data security

Monitoring and re-training

.....

Data management



Developing a model

Ethics

Standardisation

Quality control

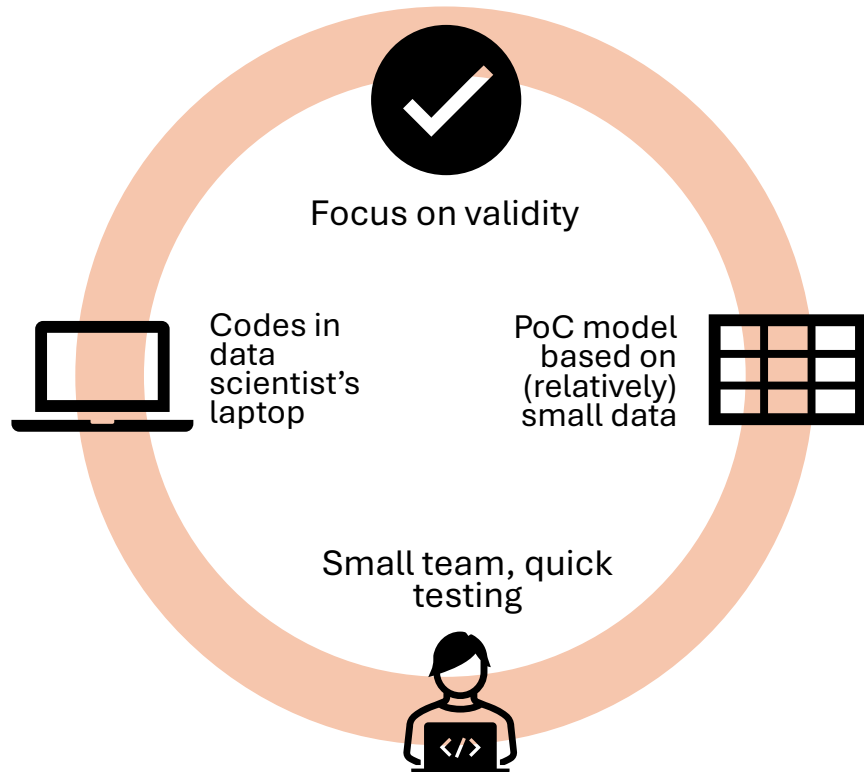
..... Versioning

Documentation

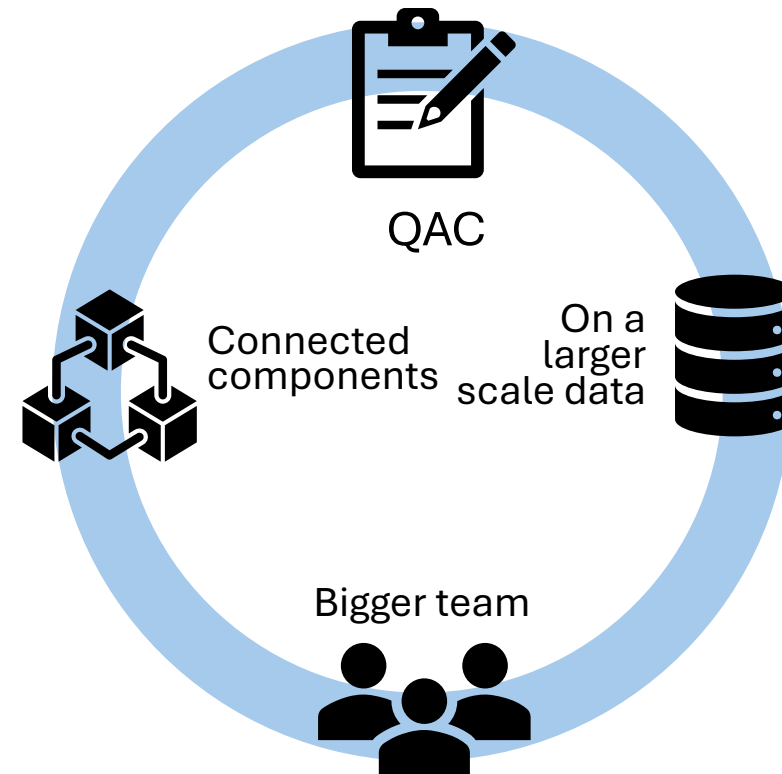
.....

AI in production

AI in experimental phase



AI in production



Skills

- People with AI skills are in high demand *everywhere*
- There are limitations for recruitments as public organisation



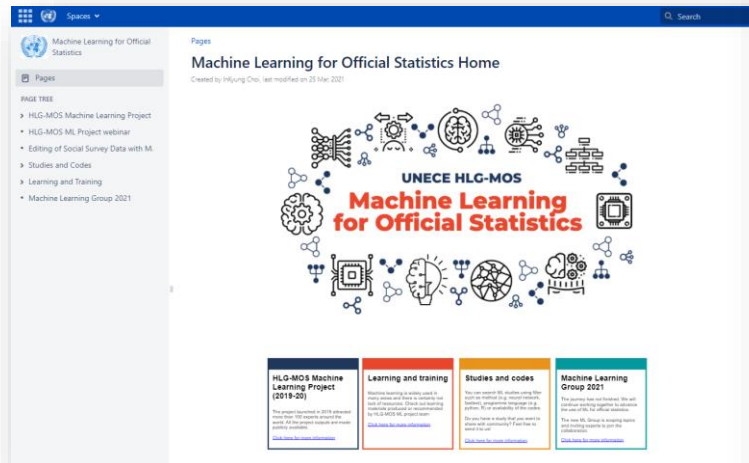
- Emphasis on what NSOs can offer (e.g., public good, data)
- Partnership (e.g., public & academia consortium)

AI for official statistics

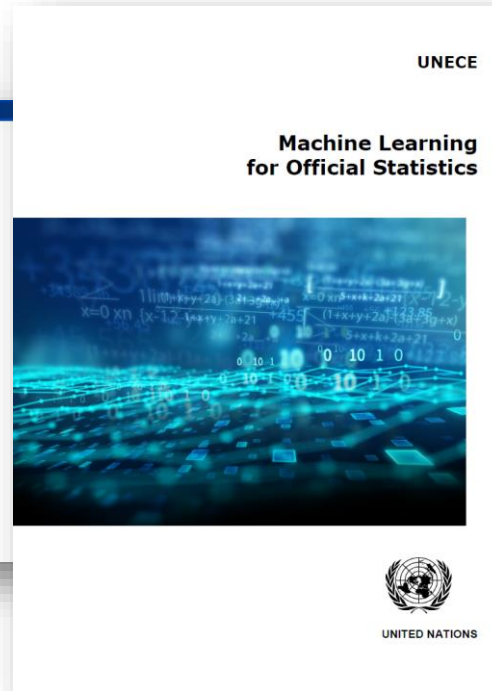
Some final remarks

- AI may not transform NSOs overnight, but can help modernise in various ways
- Concrete uses cases from NSOs
- Fully integrating AI is a long journey
- Sharing and collaboration is key to facilitating AI adoption

Resources



Statistics Wiki on ML



ML for Official Statistics

Generative AI use cases in official statistics

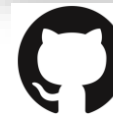
This page contains generative AI use cases in national and international statistical organizations presented at various relevant meetings. If you have any use case / research you wish to add to this list, please submit it via pull request to [the data file on github](#) or [this form](#).

List of use cases

See data file in JS

Use case type	Title	Resource type	Organisation	Date	AI models used	Approach	Tools	Code availability
1	Code and IT development SAS to R translation	Paper	CSO Ireland	2023-12	GPT-3.5			
2	Code and IT development SAS to R conversion: the use of LLM	Presentation	France Insee	2023-11	GPT-3.5			
3	Text generation for analysis Report Generation Using LLMs	Paper	Statistics Canada	2023-11	GPT-3.5			
4	Text generation for analysis AI in Statistics Norway	Presentation	Statistics Norway	2024-05	GPT-4	Prompt engineering	Chat GPT	N/A
5	Text generation for data processing (editing) Leveraging AI-assistants for better metadata and communication	Presentation	Bank of International Settlements	2024-03	GPT-3.5	Prompt engineering, RAG	OpenAI AI Assistant	N/A
6	Text generation for data processing (editing) Leveraging AI-assistants for better metadata and communication	Presentation	Bank of International Settlements	2024-03	GPT-3.5	Prompt engineering, RAG	OpenAI AI Assistant	N/A
7	Communication - textual materials Use of AI in the communication and dissemination of statistics	Presentation	Indonesia BIS	2024-03	GPT-3.5	Prompt engineering		N/A
8	Communication - images or videos materials Use of generative AI in ESTAT	Presentation	EUROSTAT	2024-04		Prompt engineering	DALL-E, ChatGPT, Adobe Firefly, Image Creator, Midjourney	N/A
9	Data dissemination and information search Free to play: UNCTAD's experience with developing its own open-source RAG LLM application	Paper	UNCTAD	2024-05	mistral	RAG	LlamaIndex, PostgreSQL, Streamlit	Yes

modernstats



[Github repo on generative AI use cases](#)

Thank you for your attention!