

Сессия 5

ИИ для изучения опыта официальной статистики и извлеченных уроков на основе инициатив ГВУ-МОС

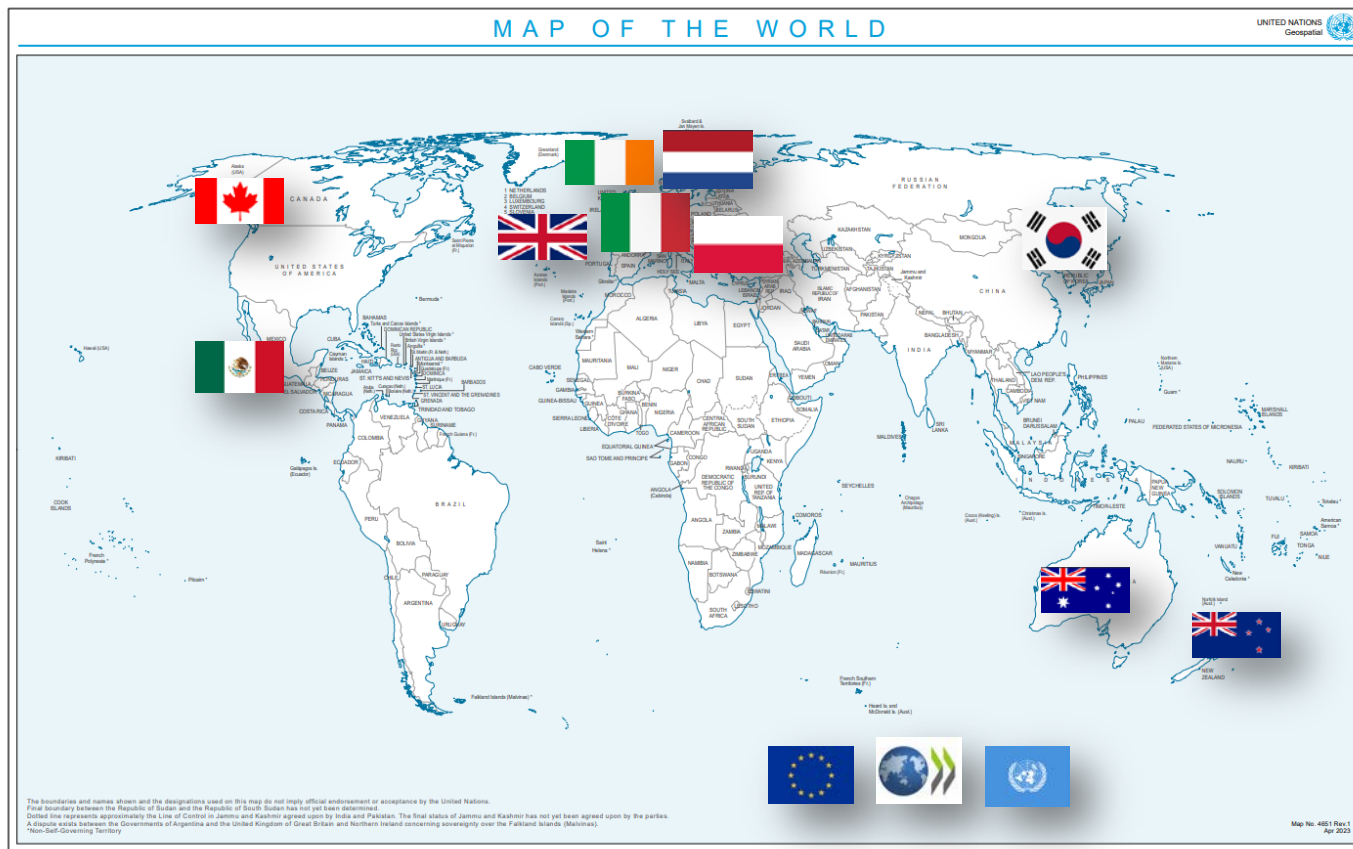
Тиина Луиге, ИнКьюнг Чой, ЕЭК ООН

Форум СНГ: новые технологии и источники данных в статистике
Ташкент, Узбекистан, 2-4 октября 2024



UNESCO

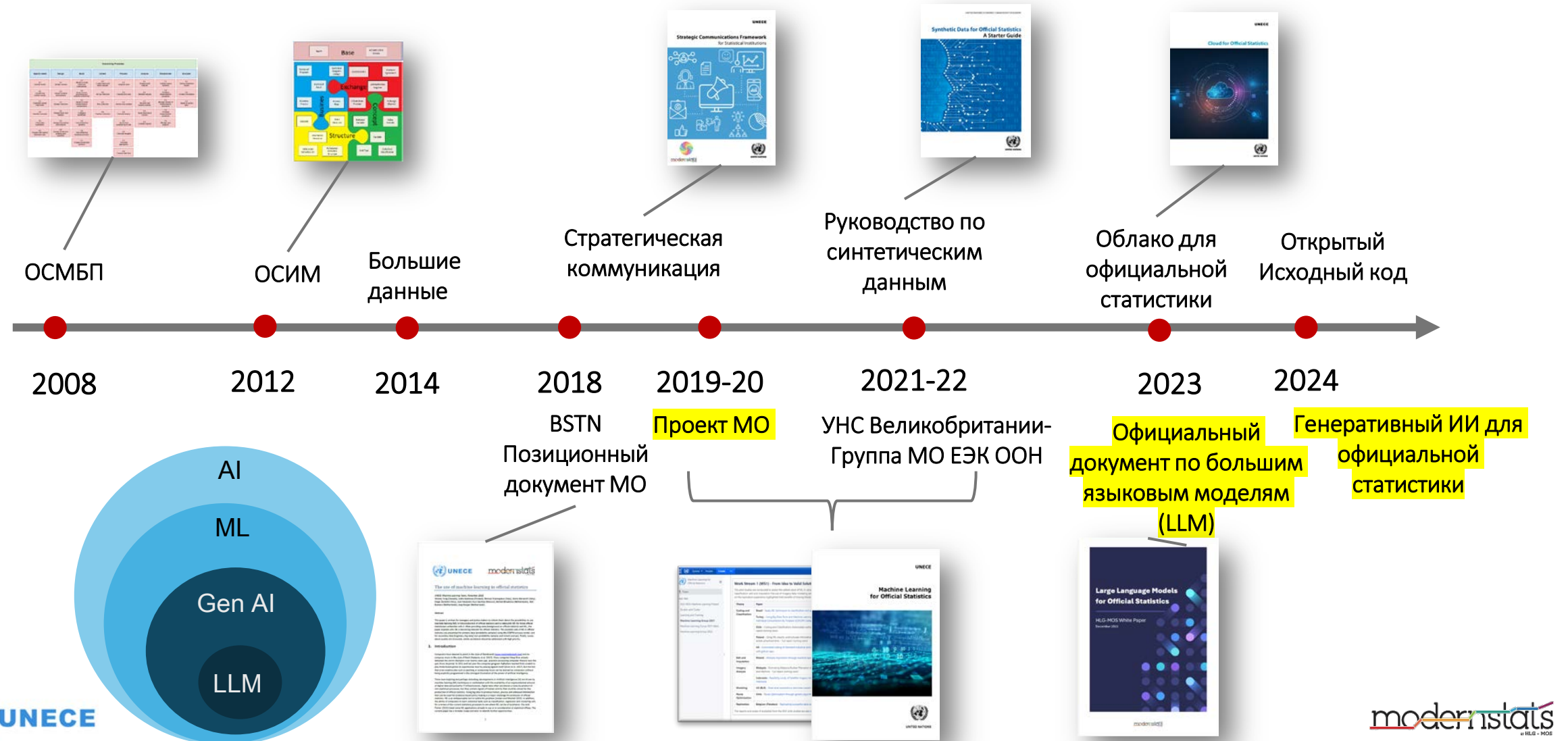
ГВУ-МОС



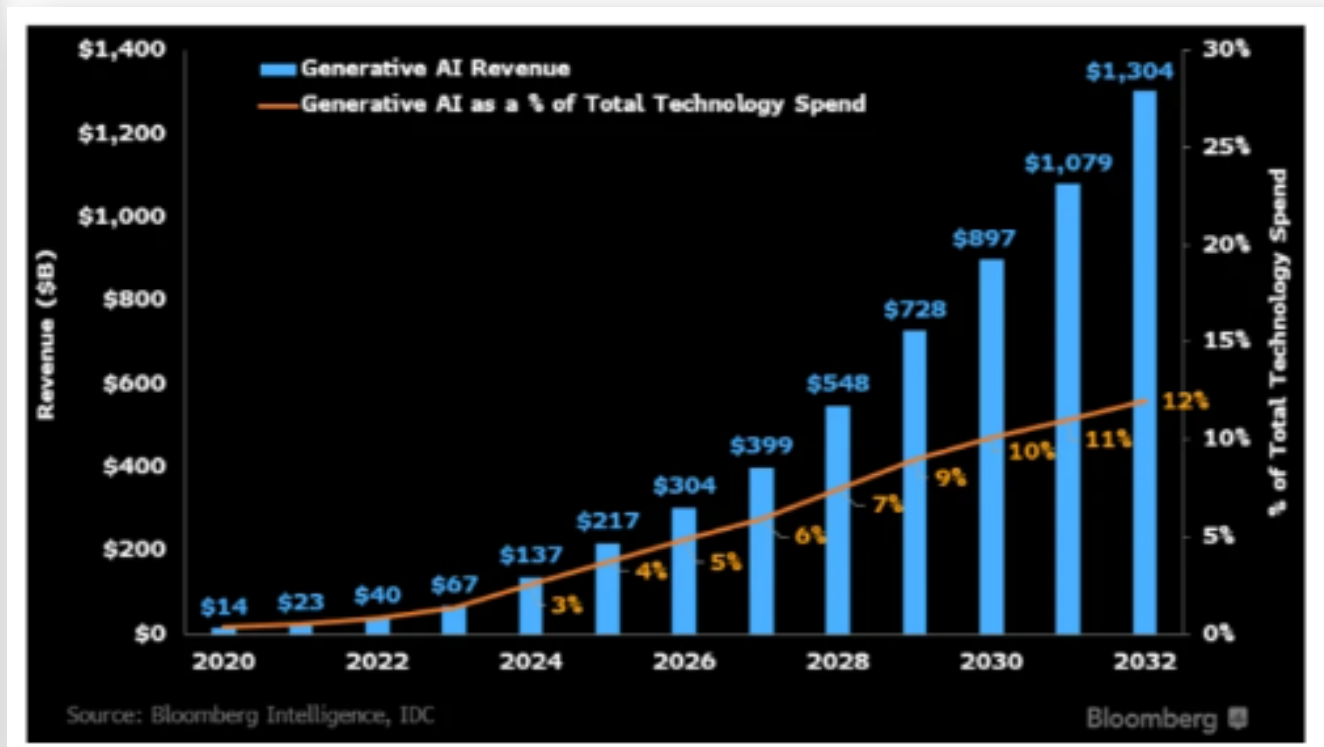
- Группа высокого уровня по модернизации официальной статистики (ГВУ-МОС) была создана в 2010 году Конференцией европейских статистиков (КЕС) для проведения модернизации официальной статистики
- Состоит из руководителей 13 национальных статистических служб (НСС) и международных организаций (Австралия, Канада, Ирландия, Италия, Мексика, Нидерланды, Новая Зеландия, Польша, Республика Корея, Великобритания, Евростат, ОЭСР и ЕЭК ООН).
- В настоящее время сопредседателями являются Нидерланды и Польша

ГВУ-МОС и ИИ

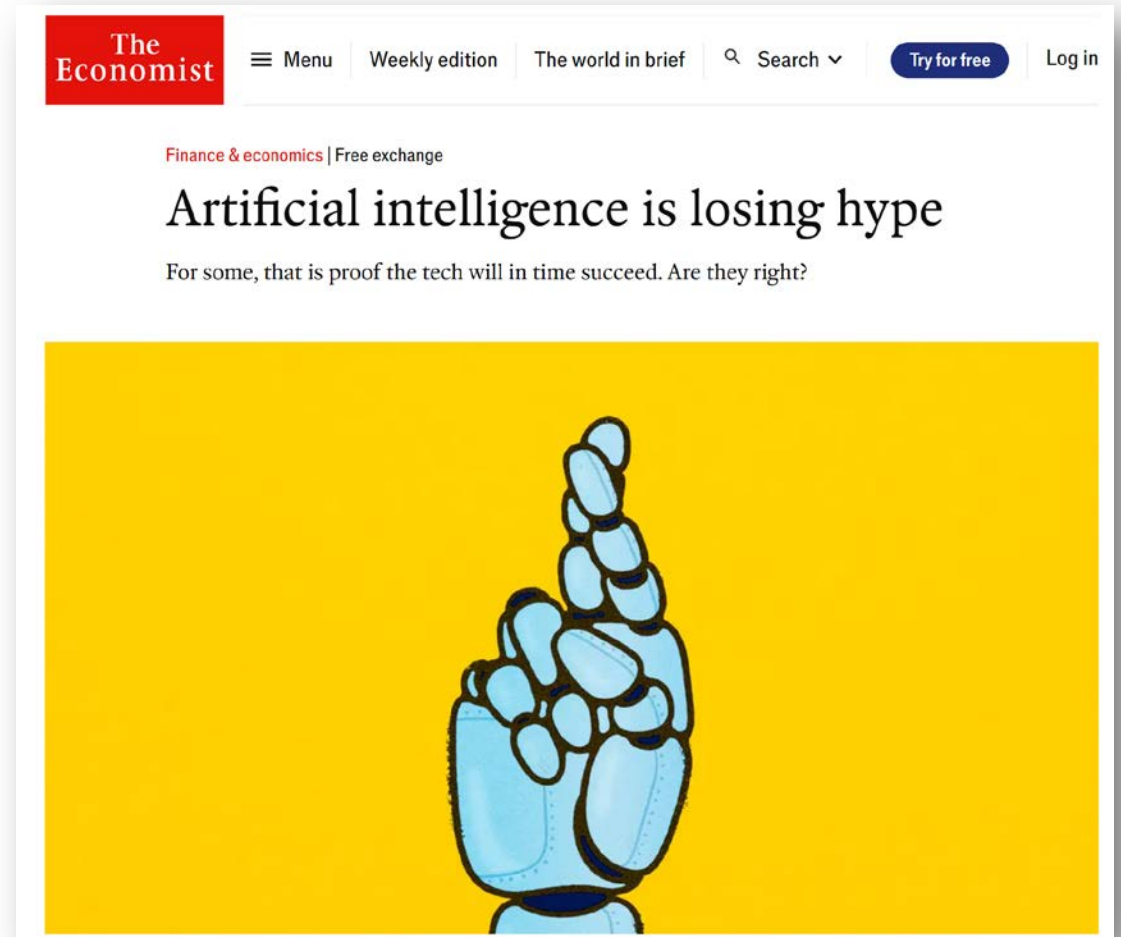
ОСМБП (GSBPM) - Общая статистическая модель бизнес-процесса
 ОСИМ (GSIM) - Общая статистическая информационная модель
 МО (ML) – машинное обучение



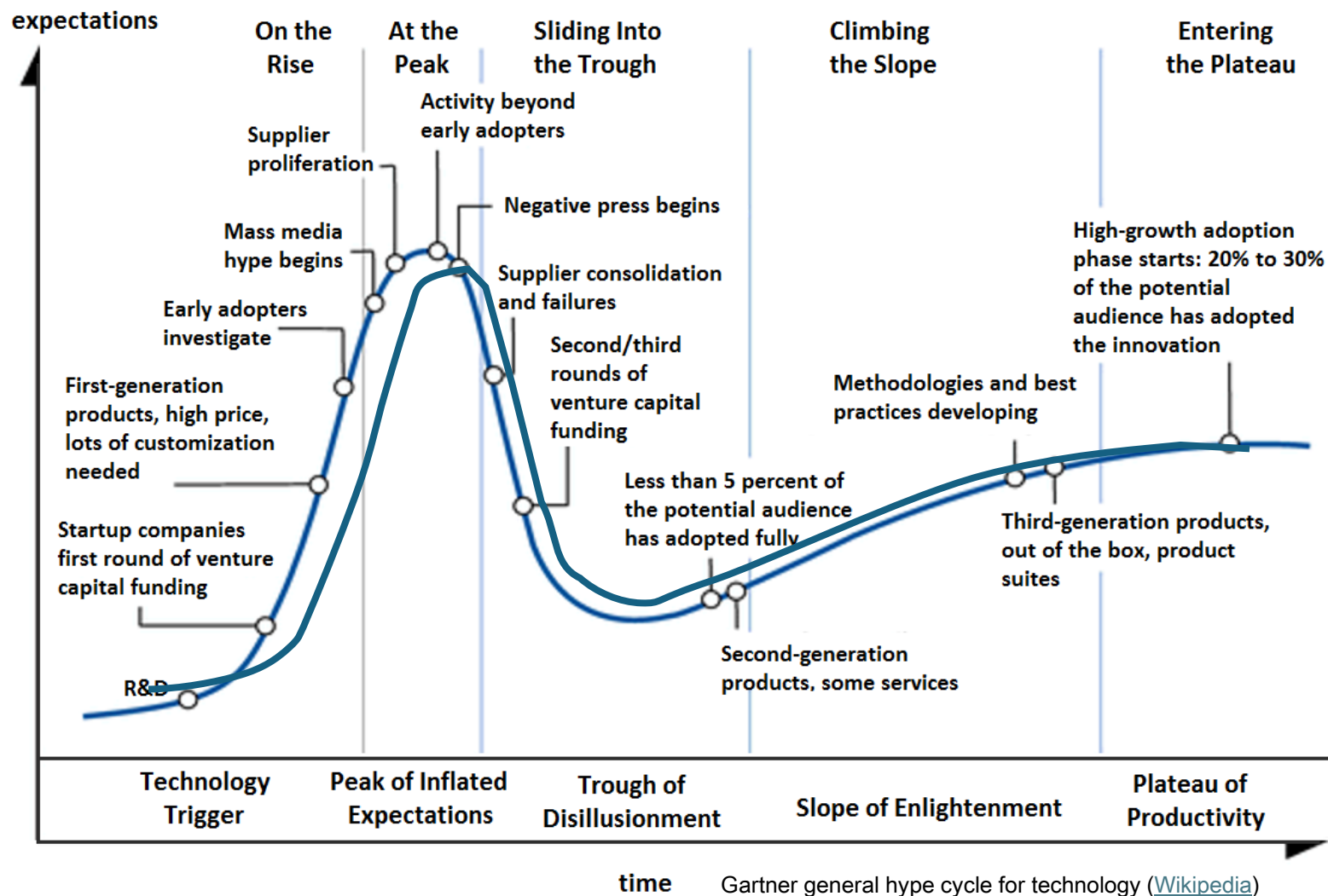
Расцвет ИИ



Расцвет ИИ



Расцвет ИИ



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

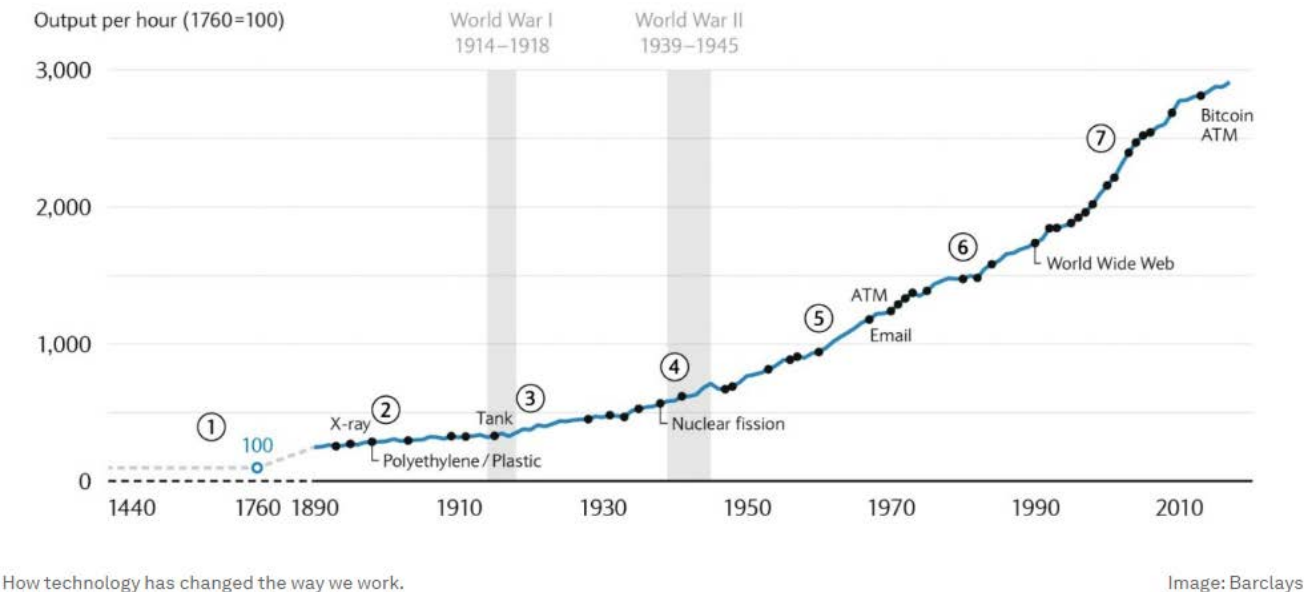
Расцвет ИИ



ИИ для модернизации НСС

FIGURE 1

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



- Повышение производительности и результативности
- Улучшение качества предоставляемых услуг и взаимодействия с пользователями
- Предложение новых продуктов

Но не без проблем

Области применения



- Области с ручными и повторяющимися задачами могут быть автоматизированы с помощью МО

Общая статистическая модель бизнес-процессов (GSBPM)

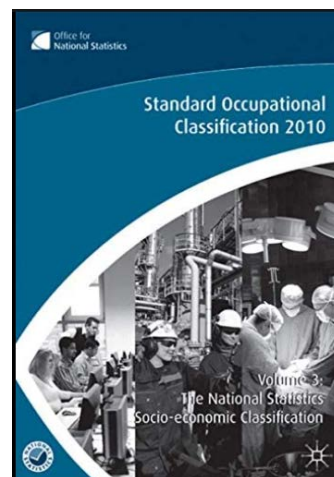
Классификация текста



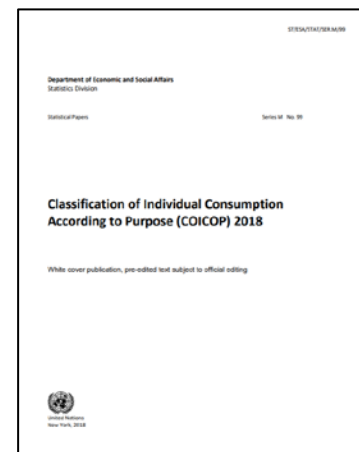
Ответ на опрос “Я работаю поваром во французском ресторане”



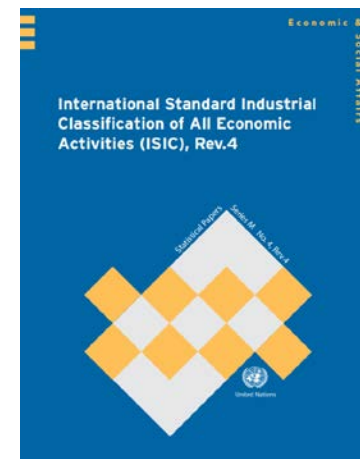
SOC “35-2014.00 Повара, Ресторан”



Система стандартной профессиональной классификации (SOC) по категориям занятости



Классификация индивидуального потребления по целям (КИПЦ).



Международная стандартная отраслевая классификация (МСОК)

Классификация текста

	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 100g	Confectionery products
17095	Bobovita - Apple dessert and sweet carrots	Baby food
17096	Hortex Vegetables for a pan with spinach 450 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

Текстовое описание продукта

Код COICOP

Настройки

Модель
МО

Пользователь
и

	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 potatoes...	0.760290	Fresh or chilled vegetables other than potatoes...
2999	Serenada Radamer cheese	Cheese and curd	0.934961	Cheese and curd

3000 rows x 4 columns

Новый набор текстовых
описаний товаров

Прогнозируемый
код COICOP

Классификация текста

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)

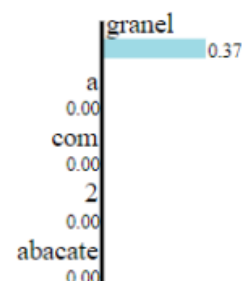
A	Measure
	Error Rate (%)
	Weighted Average F1-Score
	Weighted Average Precision
B	Metric
	Error Rate (%)
	Weighted Average F1-Score
	Weighted Average Precision
	Weighted Average Recall

(a)

True: Abacate --> Pred: Goiaba | Prob: 0.05

NOT Goiaba

Goiaba

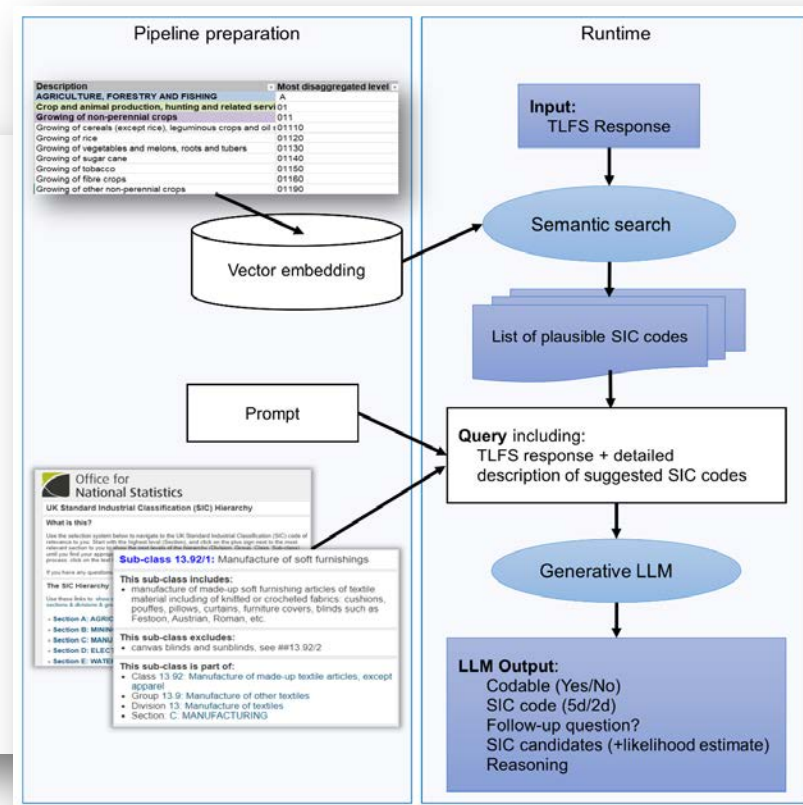


Text with highlighted words

abacate a **granel** com 2 unidades

[Statistics Canada \(ML Project\)](#)

[Brazilian Institute of Geography and Statistics \(ML Group\)](#)



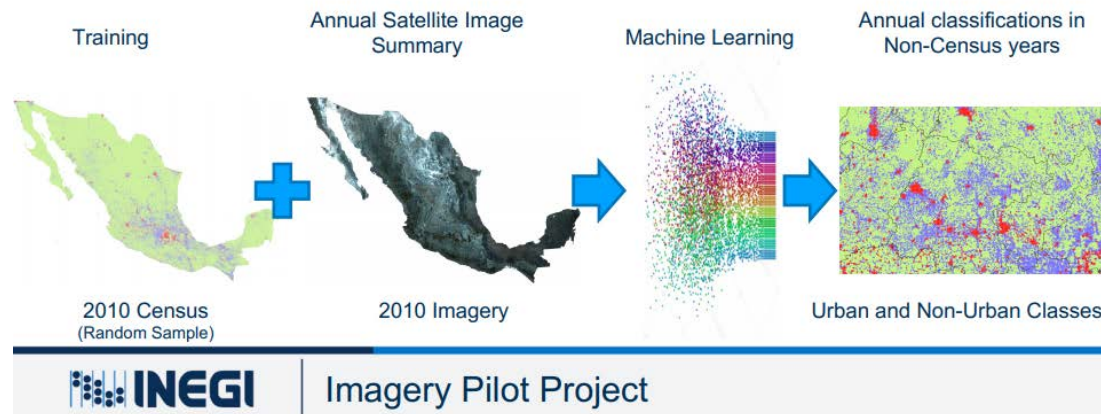
[UK ONS ClassifAI](#)



Классификация изображений

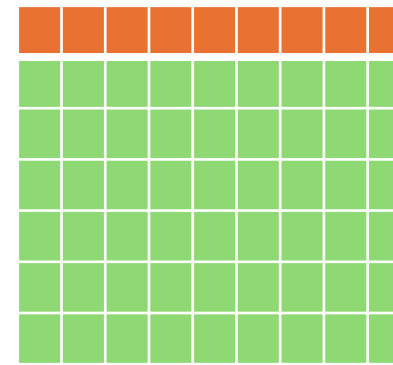
Objective of this Imagery Pilot Project (Practical Application)

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.



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

1 975 719 (1 км x 1 км) ячеек

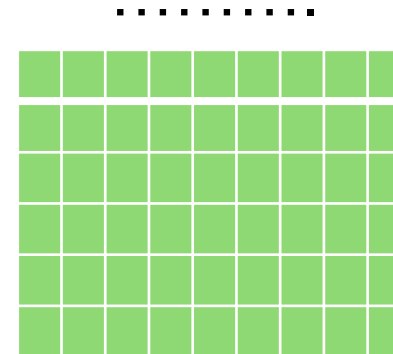


40,000 созданные человеком

Основываясь на этих данных

1,935,719

созданные при помощи МО



Другие примеры

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
Suzy E. Carter	Joe
Frank Garcia	Cola
Jonathan Smith	A.C.
Henry Long	BB



Use case – Turkish Statistical Institute

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



Turkish Statistical Institute (ML Group)

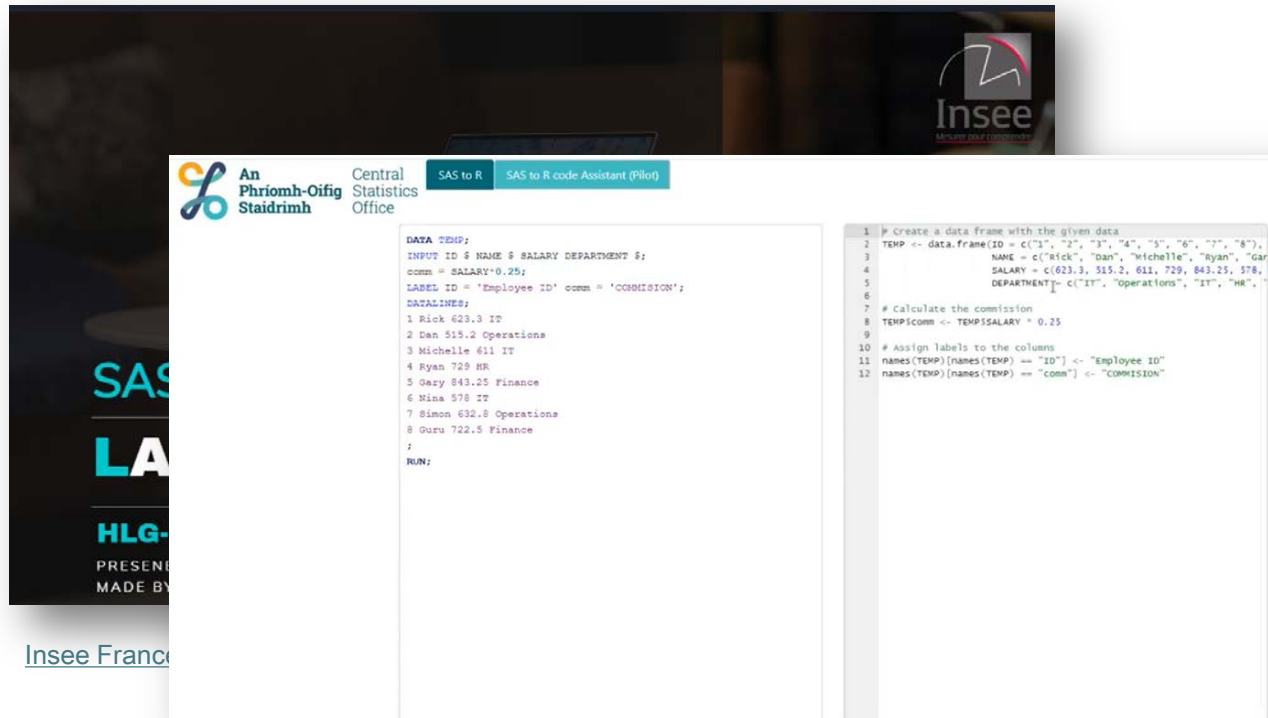
Statistisk sentralbyrå
Statistics Norway

Source: Analytics Vidhya

Statistics Norway (SDE Expert Meeting 2020)

- Запись соответствия
- Рассчеты
- Создание структуры опроса
- Определение респондентов опроса с дальнейшим мониторингом

Кодирование и развитие ИТ

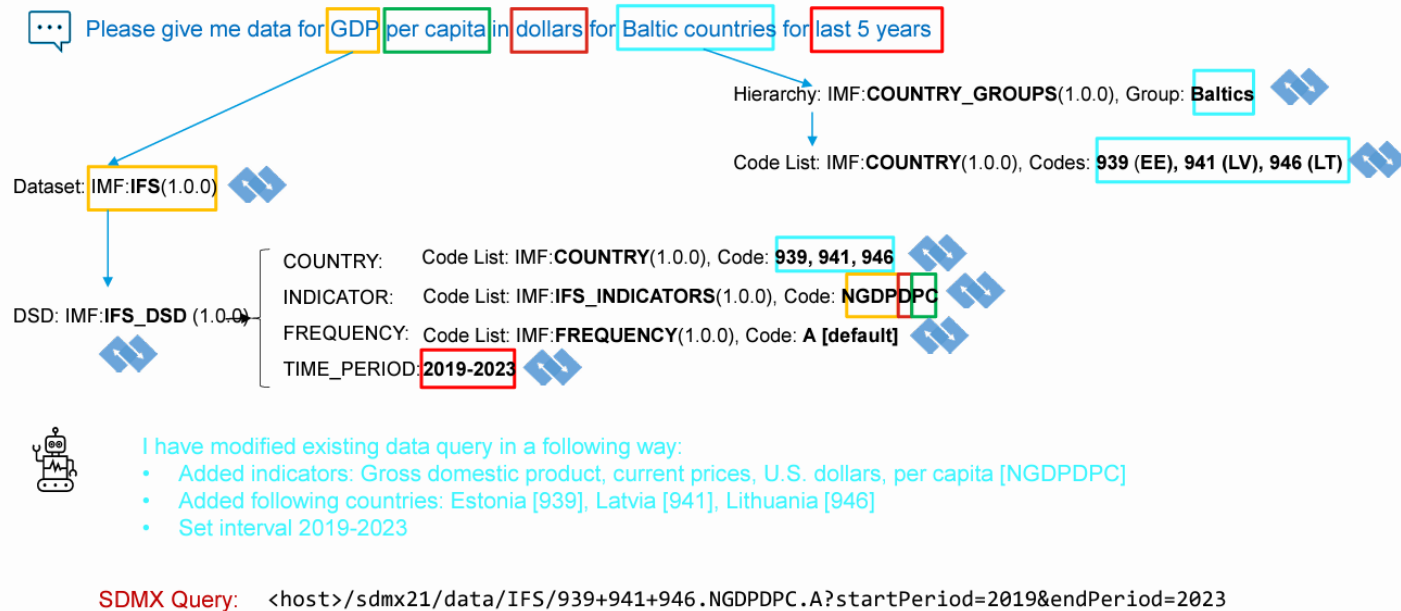


Central Statistics Office of Ireland (Generative AI Project)

- Вариант использования, который влияет на многие рабочие области в НСС
- Особенно актуально, так как многие из них переходят на ПО с открытым исходным кодом, такие как R и Python
- Положительный «побочный эффект»: стандартизация и эффективные методы кодирования

Распространение данных

Stat-GPT: How it works



IMF (CES 2024)

- Предоставление статистики является фундаментальной ролью НСС
- Платформа данных не проста в навигации
- Помощь пользователям находить данные, используя естественный язык и улучшая пользовательский опыт

Поиск информации и генерация текста

The screenshot displays a web interface for document processing. On the left, a sidebar contains sections for 'Upload your own documents' and 'Model parameters'. The 'Upload your own documents' section 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. The 'Model parameters' section features dropdown menus for 'Which LLM' (set to 'mistral-docsgpt') and 'Which corpus' (set to 'temporary_daniel_hopp'), along with an 'Advanced model parameters' dropdown, a 'Reinitialize model' button, and a 'Reset model's memory' button. The main content area on the right shows a list of generated text snippets. The first snippet is titled 'Give 3 key takeaways from the 2023 Review of Maritime Transport 2024-04-19 10:55' and contains three numbered points. The second snippet is titled 'Elaborate on global trade growing at a slower pace' and contains a paragraph of text. The third snippet is titled 'Sources:' and contains a paragraph of text. At the bottom of the main content area, there is a query input field with the text 'Query 'mistral-docsgpt' contextualized on 'temporary_daniel_hopp' corpus' and a search button.

Upload your own documents ⓘ

URLs ⓘ

https://unctad.org/system/files/official-document/rmt2023_en.pdf

Upload your own documents ⓘ

Drag and drop file here
Limit 200MB per file • ZIP, DOCX, DOC, TXT, PDF, CSV

Browse files

Process corpus

Model parameters ⓘ

Which LLM ⓘ

mistral-docsgpt

Which corpus ⓘ

temporary_daniel_hopp

Advanced model parameters

Reinitialize model

Reset model's memory

Give 3 key takeaways from the 2023 Review of Maritime Transport 2024-04-19 10:55
2024-04-19 10:56

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.

Sources: ⓘ
2024-04-19 10:56

Elaborate on global trade growing at a slower pace
2024-04-19 10:57

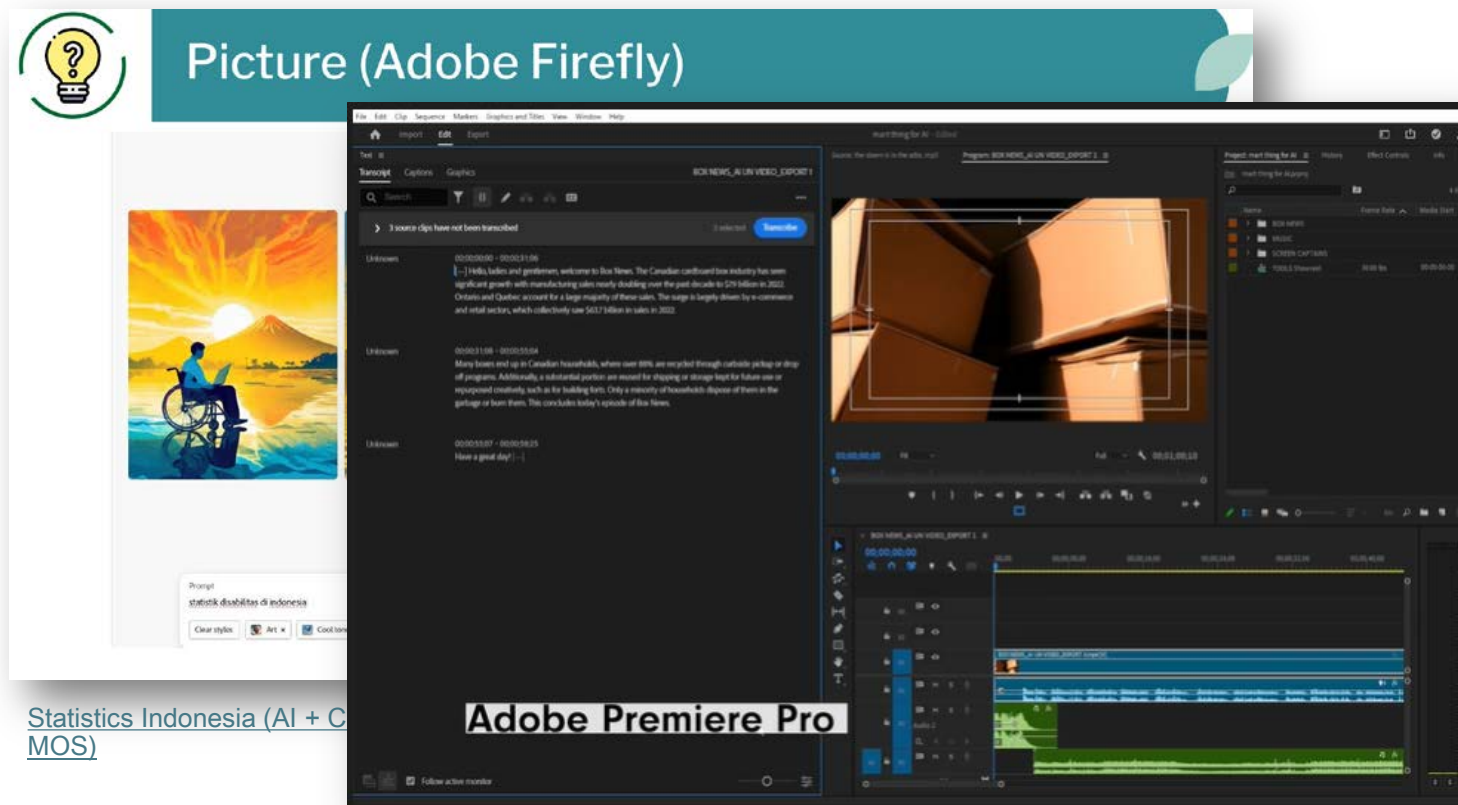
According to the 2023 Review of Maritime Transport, global merchandise trade grew by only 2.7% in 2022, which represents a significant slowdown from the 9.4% rebound witnessed in 2021. This slowdown can be attributed to several factors, including high inflation, geopolitical risks, and debt burdens constraining 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

Query 'mistral-docsgpt' contextualized on 'temporary_daniel_hopp' corpus

- Может также использоваться для внутреннего поиска и исследования информации

[UNCTAD \(CES2024\)](#)

Коммуникация



- Растущий акцент на индивидуальной коммуникации
- Генерация изображений/аудио/видео
- Генерация коммуникационных материалов в соответствии с определенными стилями и тоном

Вызовы и некоторые усвоенные уроки

- Вызовы, связанные с ИИ

Различные методологии
Черный ящик
Различные языки
программирования

- ... осложняется тем, что речь идет о национальных статистических службах

Более высокая цена
ошибки
Ограничение закупок

Соответствие
стандартам качества

Эксперимент vs. производство
Управление изменениями

Качество

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			*						○	

UN National Quality Assurance Framework

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 | Ascarl, Gabriele^f | Rocci, Fabiana^g | Burger, Joep^h | Choi, Inkyung^h

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/SJI-210875

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

Published: 21 March 2022

Quality Framework for Statistical Algorithm (2022)

- Точность
- Возможность объяснения
- Воспроизводимость
- Своевременность
- Экономическая эффективность



ИИ в производстве

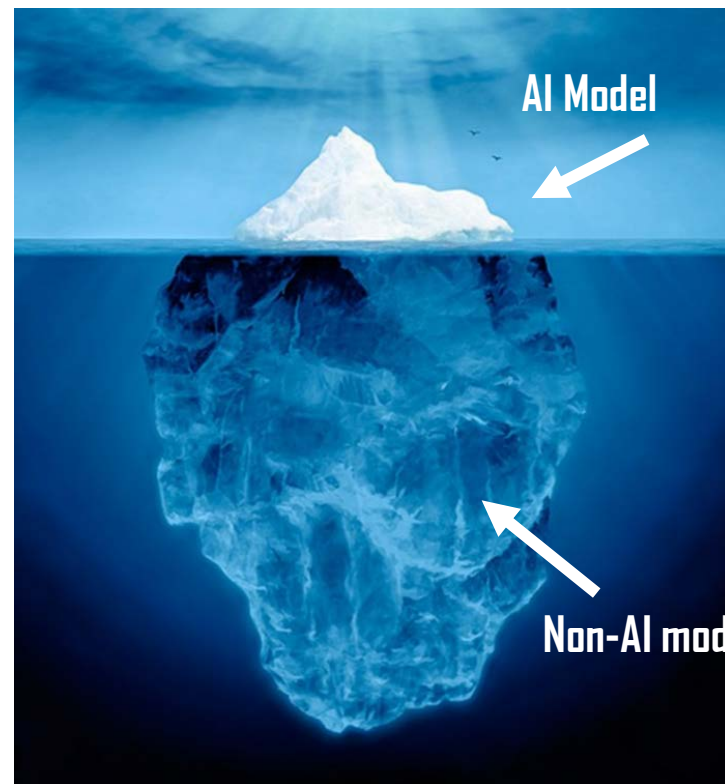
Участие заинтересованных сторон

Модель обслуживания

Безопасность данных

Мониторинг и переподготовка

Управление данными



Создание модели

Этика

Стандартизация

Контроль качества

..... Управление версиями

Документирование

.....



ИИ в производстве

ИИ в экспериментальной фазе



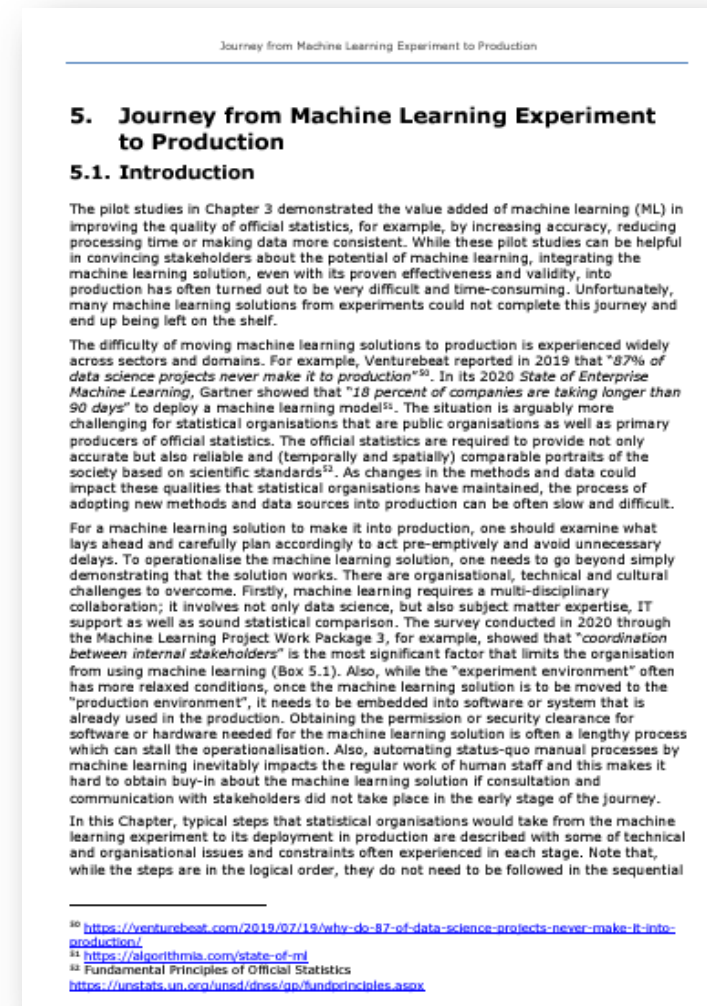
ИИ в производстве



ИИ в производстве



HLG-MOS ADSaMM Group task team on ML editing



ML for Official Statistics

Навыки

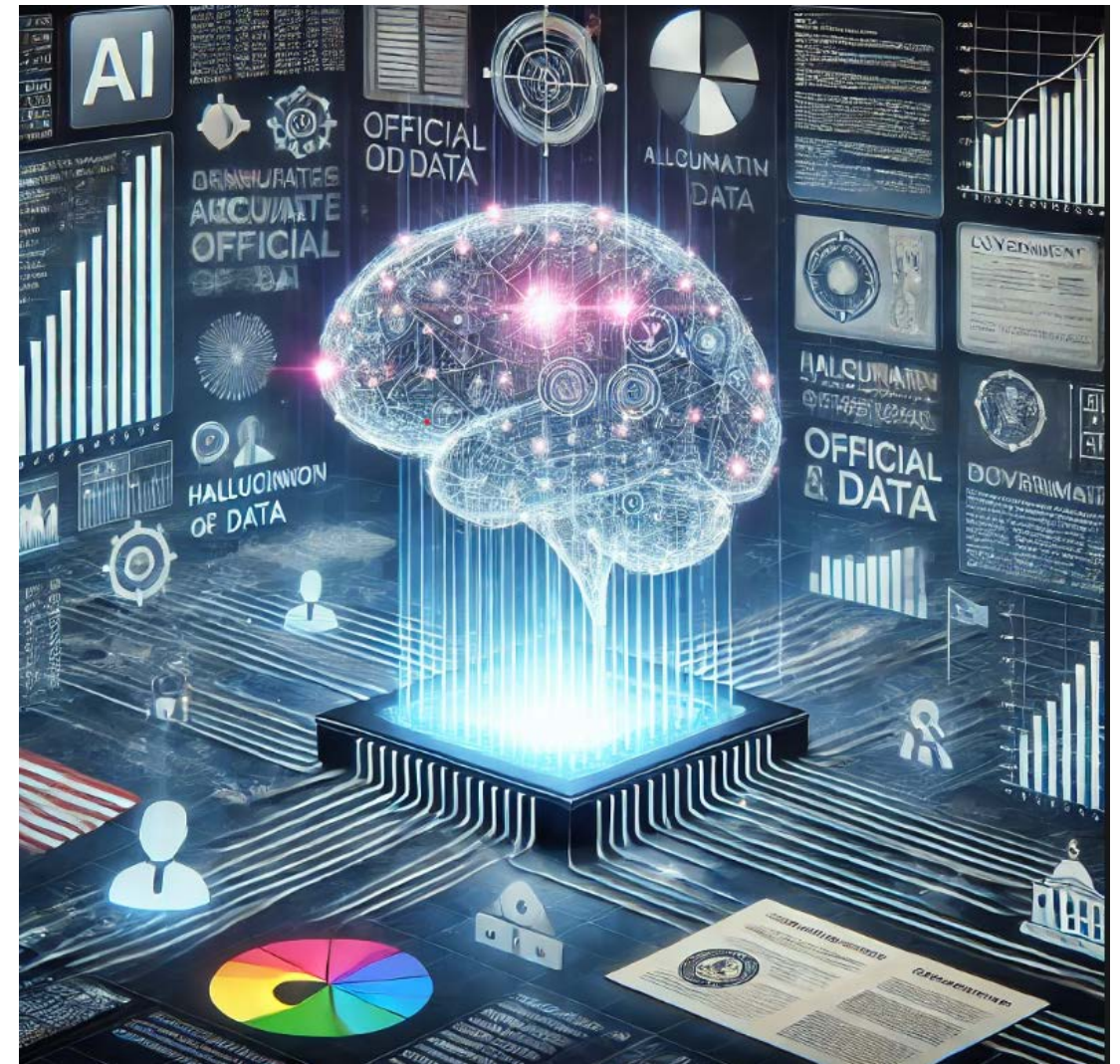
- People with AI skills are in high demand *everywhere*
 - There are limitations for recruitments as public organization
 - Люди с навыками работы с ИИ востребованы везде
 - Существуют ограничения для набора в государственные организации
- Повышение квалификации
 - Акцент на том, что могут предложить НСС (например, общественное благо, данные)
 - Партнерство (например, консорциум общественности и академических кругов)



Использование ИИ для официальной статистики

Проблема нахождения официальных статистических данных в ИИ

- Для ИИ любые данные считаются одинаково надежными
- Улучшение: теперь указывается источник



ИИ для официальной статистики

Some final remarks

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

Некоторые заключительные замечания

- ИИ не может преобразовать НСС в одночасье, но может помочь модернизировать различными способами
- Конкретные примеры использования в НСС
- Полная интеграция ИИ — это долгий путь
- Обмен и сотрудничество — важный фактор на пути к принятию ИИ

ИСТОЧНИКИ



[Statistics Wiki on ML](#)

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 [the form](#).

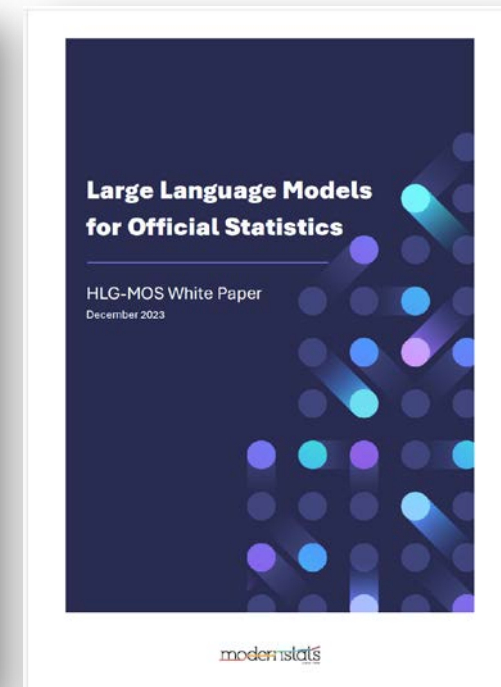
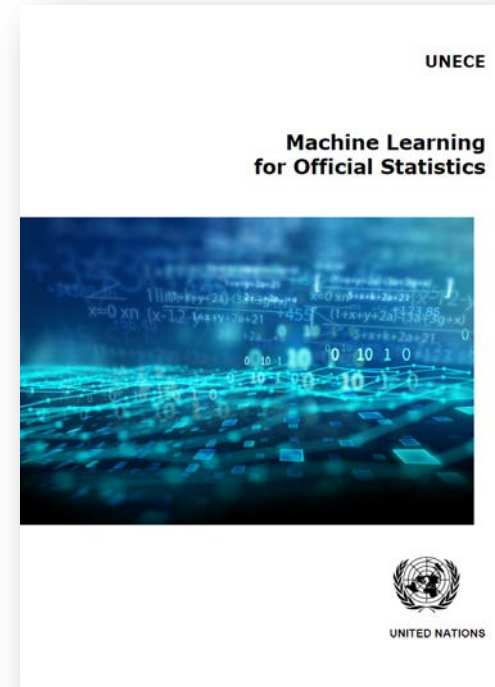
List of use cases

See data file in JS

User case type	Title	Resource type	Organization	Date	AI models used	Approach	Tools	Code availability
1. Code and IT development	SQL to R translation	Paper	CSO Ireland	2023-12	GPT-3.5			
2. Code and IT development	SQL to R conversion using LLM	Presentation	France Insee	2023-11	GPT-3.5			
3. Text generation for analysis	Statistical Generation using LLMs	Paper	Statistics Canada	2023-11	GPT-3.5			
4. Text generation for analysis	AI in Statistical Research	Presentation	Statistics Norway	2024-05	GPT-4	Prompt engineering	Chat GPT	N/A
5. Text generation for data processing (writing)	Language LLM assistants for better production 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 (writing)	Language AI assistants for better production 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 communication of EMGOs	Presentation	Indonesia BPS	2024-02	GPT-3.5	Prompt engineering		N/A
8. Communication - images or videos materials	Use of generative AI in ESTAT	Presentation	UNCTAD	2024-04		Prompt engineering	DALL-E, ChatGPT, Adobe Firefly, Image Creation, Midjourney	N/A
9. Data dissemination and information search	E-stat: AI assistant for ESTAT	Paper	UNCTAD	2024-05	meta	RAG	OpenAI, ChatGPT, Bing, Google, etc.	Yes



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



[ML for Official Statistics \(2021\)](#) [LLMs for Official Statistics \(2023\)](#)



[CES paper on compilation of AI initiatives under HLG-MOS \(April 2024\)](#)

c
tina.iuige@un.org