INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

» Artificial Intelligence and Official Statistics

the need for a new approach

Departamento de Metodologia e Sistemas de Informação
Serviço de Metodologia

pedro.campos@ine.pt

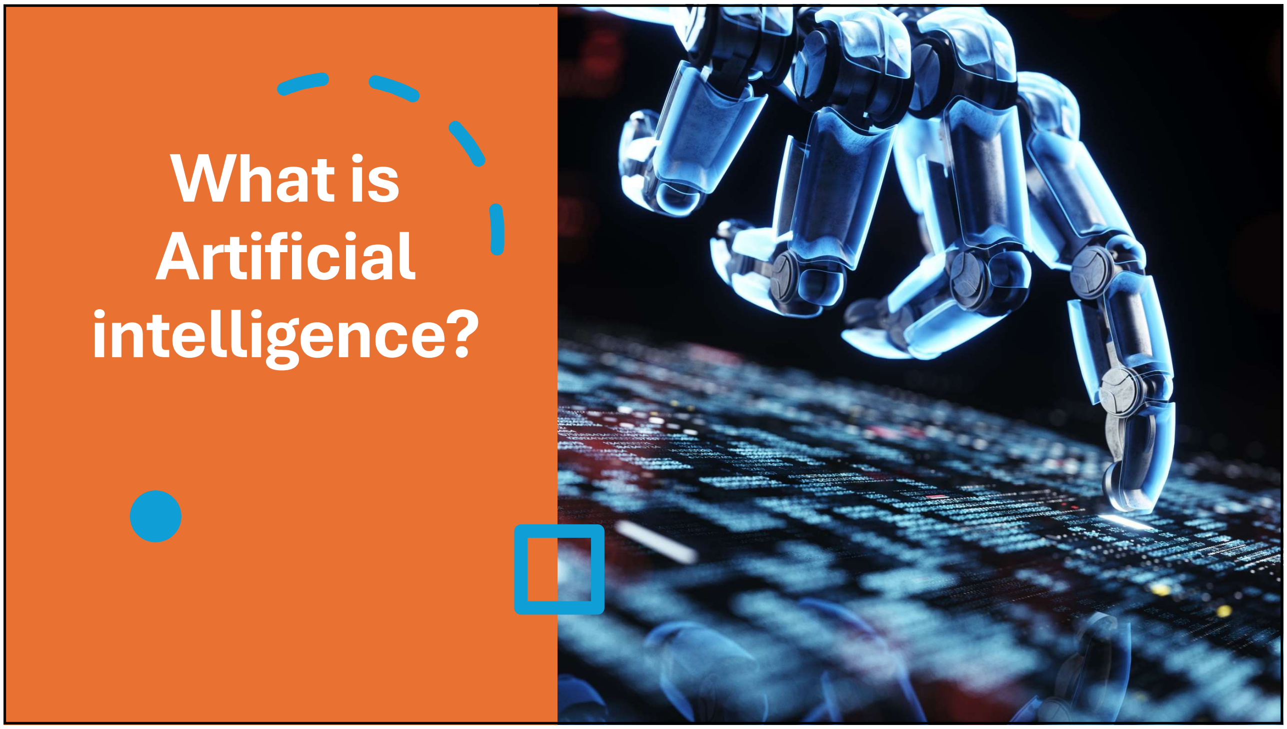
Pedro Campos

1

Contents

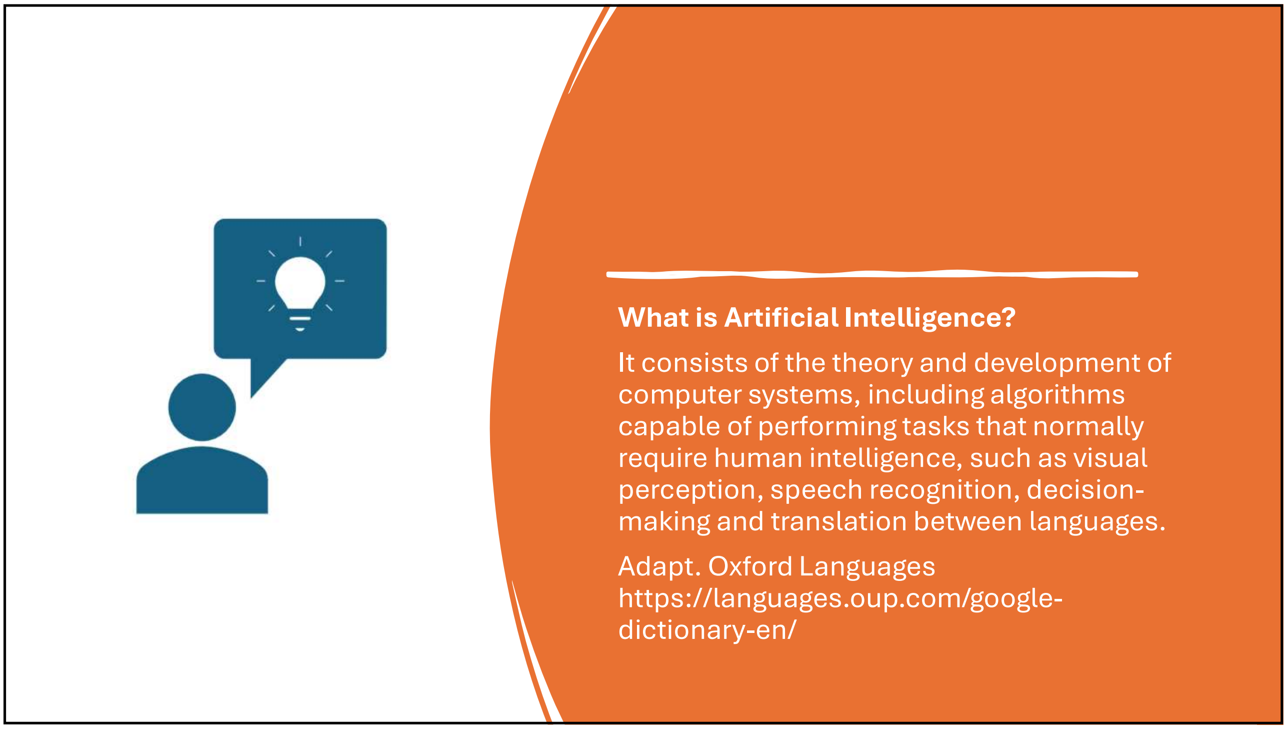
- Official Statistics and Artificial Intelligence
- Machine Learning in Official Statistics tasks
- LLM - Large Language Models

2



What is Artificial intelligence?

3

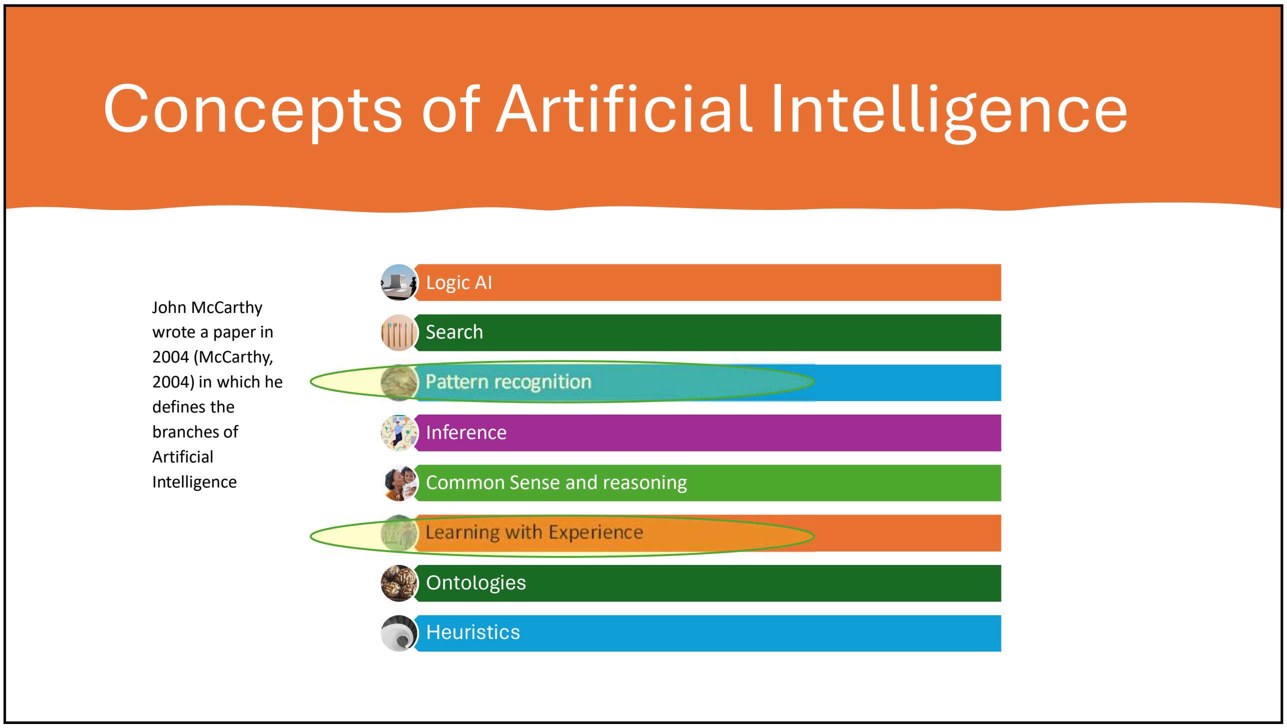


What is Artificial Intelligence?

It consists of the theory and development of computer systems, including algorithms capable of performing tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making and translation between languages.

Adapt. Oxford Languages https://languages.oup.com/google-dictionary-en/

4

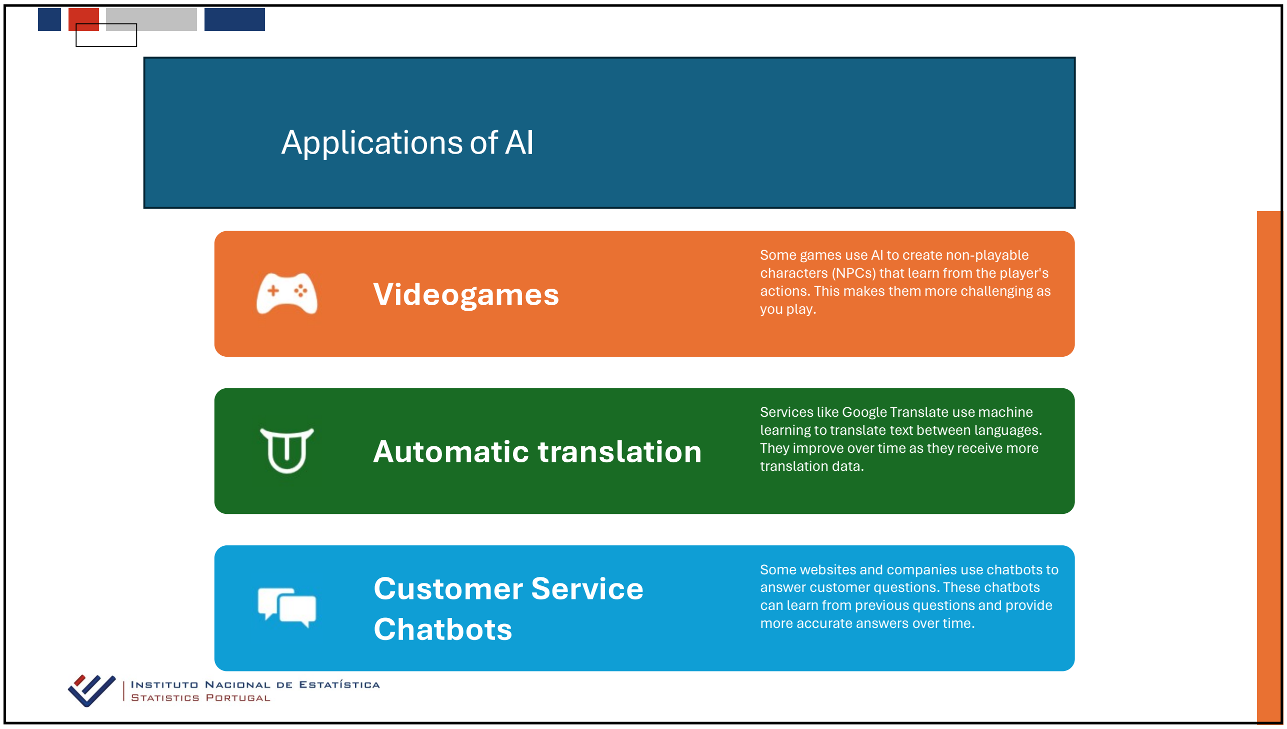


Concepts of Artificial Intelligence

John McCarthy wrote a paper in 2004 (McCarthy, 2004) in which he defines the branches of Artificial Intelligence

- Logic AI
- Search
- Pattern recognition
- Inference
- Common Sense and reasoning
- Learning with Experience
- Ontologies
- Heuristics

5



Applications of AI

Videogames

Some games use AI to create non-playable characters (NPCs) that learn from the player's actions. This makes them more challenging as you play.

Automatic translation

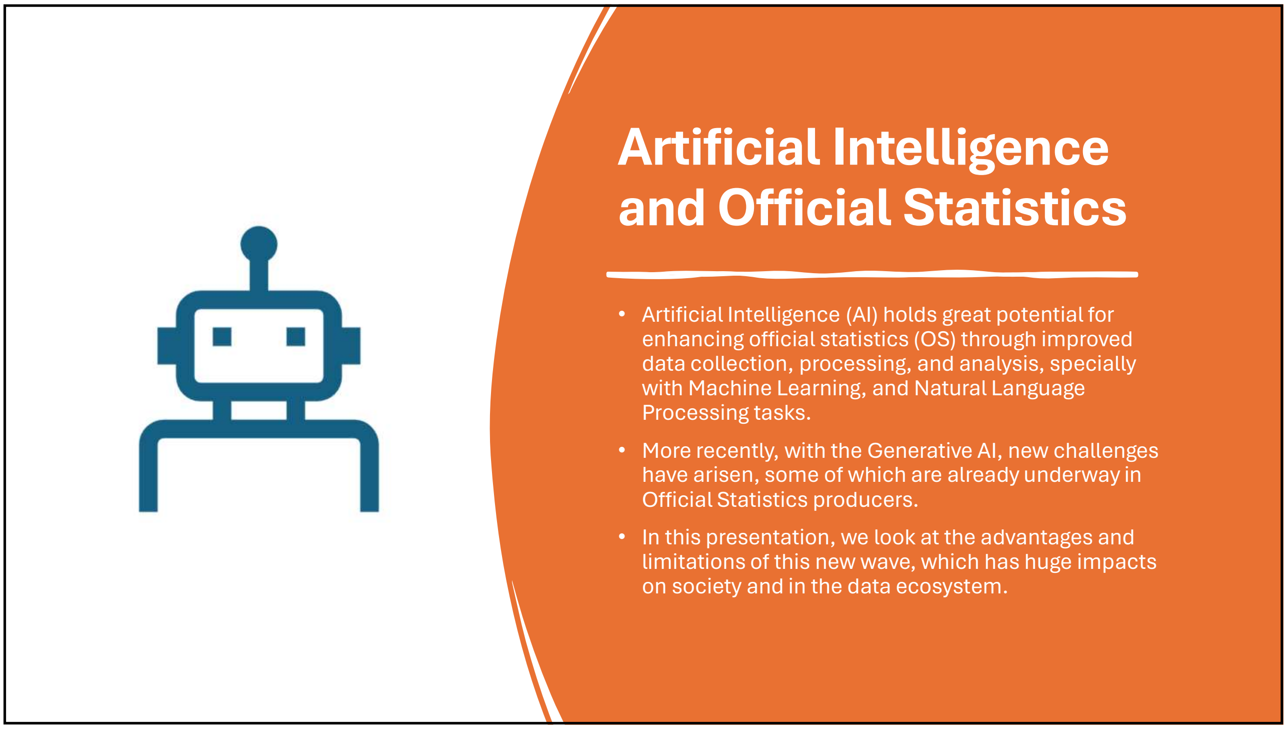
Services like Google Translate use machine learning to translate text between languages. They improve over time as they receive more translation data.

Customer Service Chatbots

Some websites and companies use chatbots to answer customer questions. These chatbots can learn from previous questions and provide more accurate answers over time.

Instituto Nacional de Estatística
Statistics Portugal

6



Artificial Intelligence and Official Statistics

- Artificial Intelligence (AI) holds great potential for enhancing official statistics (OS) through improved data collection, processing, and analysis, specially with Machine Learning, and Natural Language Processing tasks.
- More recently, with the Generative AI, new challenges have arisen, some of which are already underway in Official Statistics producers.
- In this presentation, we look at the advantages and limitations of this new wave, which has huge impacts on society and in the data ecosystem.

7

Artificial Intelligence and Official Statistics

Automated Survey Questionaires

- AI-powered chatbots or virtual assistants can be used to conduct surveys and collect information from respondents.
- Natural language processing (NLP) algorithms can help understand and process unstructured data from open-ended survey questions.
- Voice and image recognition can play an important role

8

Machine Learning in Official Statistics tasks

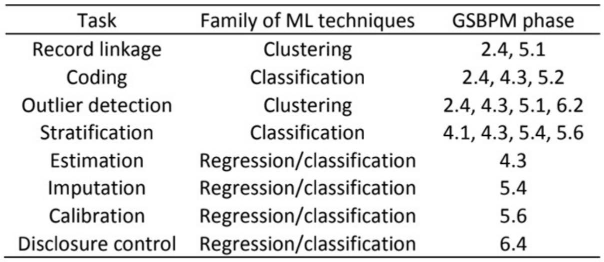
- **Why is machine learning becoming relevant to official statistics?**
- **Yung et. al (2018) survey the potential of Machine Learning in Official Statistics**

- **The NSIs are currently facing a number of challenges, which is causing them to reflect on their data sources:**
- **Low response rates to primary data collection efforts**
- **Access to secondary data from alternative sources (administrative and other data)**

9

Machine Learning in Official Statistics tasks

Primary Data



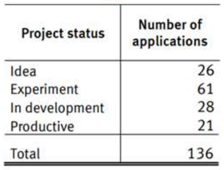
Task	Family of ML techniques	GSBPM phase
Record linkage	Clustering	2.4, 5.1
Coding	Classification	2.4, 4.3, 5.2
Outlier detection	Clustering	2.4, 4.3, 5.1, 6.2
Stratification	Classification	4.1, 4.3, 5.4, 5.6
Estimation	Regression/classification	4.3
Imputation	Regression/classification	5.4
Calibration	Regression/classification	5.6
Disclosure control	Regression/classification	6.4

Yung et. al (2018)

10

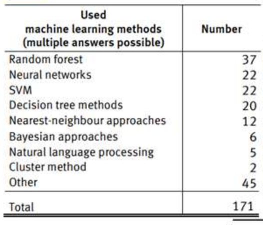
Machine Learning in Official Statistics tasks

Why?

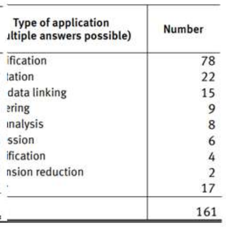


Project status	Number of applications
Idea	26
Experiment	61
In development	28
Productive	21
Total	136

How?



Used machine learning methods (multiple answers possible)	Number
Random forest	37
Neural networks	22
SVM	22
Decision tree methods	20
Nearest-neighbour approaches	12
Bayesian approaches	6
Natural language processing	5
Cluster method	2
Other	45
Total	171



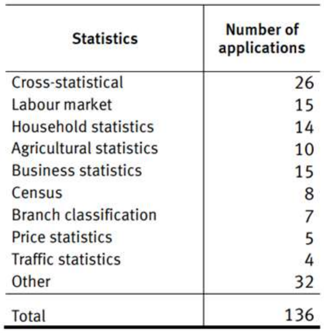
Type of application ıltiple answers possible)	Number
ification	78
ıation	22
data linking	15
ering	9
ınalysis	8
ssion	6
ification	4
nsion reduction	2
	17
	161

(Beck, Dumpert , Feuerhake, 2018)

11

Machine Learning in Official Statistics tasks

Where?



Statistics	Number of applications
Cross-statistical	26
Labour market	15
Household statistics	14
Agricultural statistics	10
Business statistics	15
Census	8
Branch classification	7
Price statistics	5
Traffic statistics	4
Other	32
Total	136

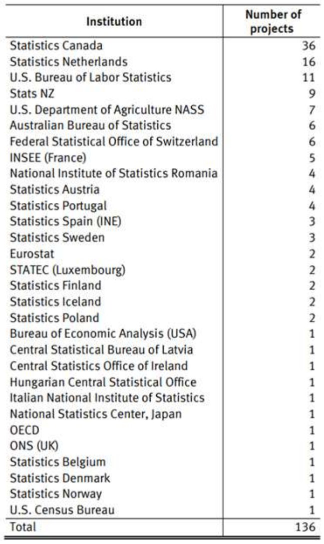
(Beck, Dumpert , Feuerhake, 2018)

12

Machine Learning in Official Statistics tasks

(Beck, Dumpert , Feuerhake, 2018)

Who?



Institution	Number of projects
Statistics Canada	36
Statistics Netherlands	16
U.S. Bureau of Labor Statistics	11
Stats NZ	9
U.S. Department of Agriculture NASS	7
Australian Bureau of Statistics	6
Federal Statistical Office of Switzerland	6
INSEE (France)	5
National Institute of Statistics Romania	4
Statistics Austria	4
Statistics Portugal	4
Statistics Spain (INE)	3
Statistics Sweden	3
Eurostat	2
STATEC (Luxembourg)	2
Statistics Finland	2
Statistics Iceland	2
Statistics Poland	2
Bureau of Economic Analysis (USA)	1
Central Statistical Bureau of Latvia	1
Central Statistics Office of Ireland	1
Hungarian Central Statistical Office	1
Italian National Institute of Statistics	1
National Statistics Center, Japan	1
OECD	1
ONS (UK)	1
Statistics Belgium	1
Statistics Denmark	1
Statistics Norway	1
U.S. Census Bureau	1
Total	136

13

Machine Learning in Official Statistics tasks

In Portugal

- **Data Cleaning, Anomaly Detection** (Santos and Campos, 2022) - correction and imputation of DMR (Monthly SS Income Declarations) values based on past months, using Support Vector Machine.
- **Privacy Preserving / Statistical Disclosure Control**
- **Forecasting** - Demographic Projections

Instituto Nacional de Estatística
Statistics Portugal

14

Machine Learning in Official Statistics tasks

In Portugal

- **Protection of confidentiality -** use of decision trees to generate synthetic values in PUFs (Public Use Files)
- **OJA - Online Job Advertisements** (Web Intelligence Hub) - detect online job offers https://www.cedefop.europa.eu/en/tools/skills-intelligence/trend-focus/skills-online-job-advertisements?year=2022&country=EU27_2020#1

Instituto Nacional de Estatística
Statistics Portugal

15

LLM – Large Language Models

User Experience Generative AI

LLMs (Large Language Models) are a class of artificial intelligence capable of understanding, interpreting and generating texts, based on extensive training on vast data sets.

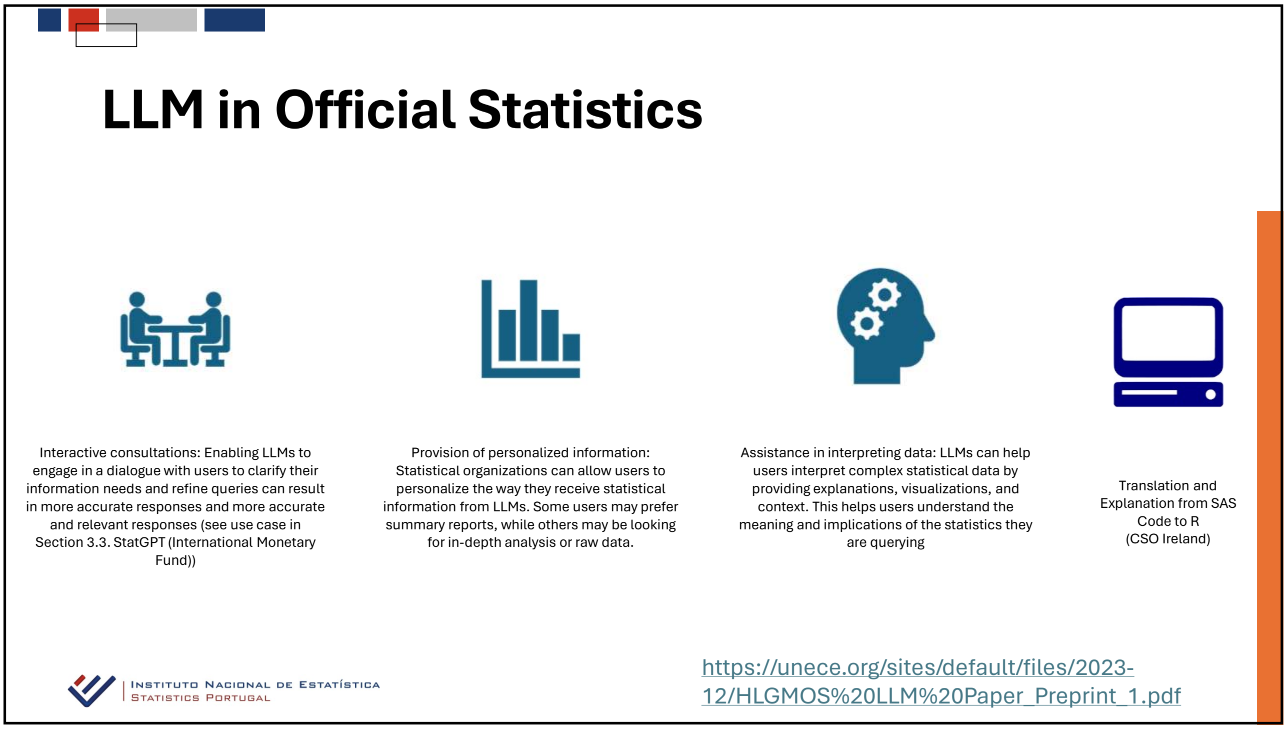
LLMs are capable of understanding and generating texts at a level indistinguishable from human beings.

Large Language Models for Official Statistics

HLG-MOS White Paper
December 2023

https://unece.org/sites/default/files/2023-12/HLGMOS%20LLM%20Paper_Preprint_1.pdf

16



LLM in Official Statistics

Interactive consultations: Enabling LLMs to engage in a dialogue with users to clarify their information needs and refine queries can result in more accurate responses and more accurate and relevant responses (see use case in Section 3.3. StatGPT (International Monetary Fund))

Provision of personalized information: Statistical organizations can allow users to personalize the way they receive statistical information from LLMs. Some users may prefer summary reports, while others may be looking for in-depth analysis or raw data.

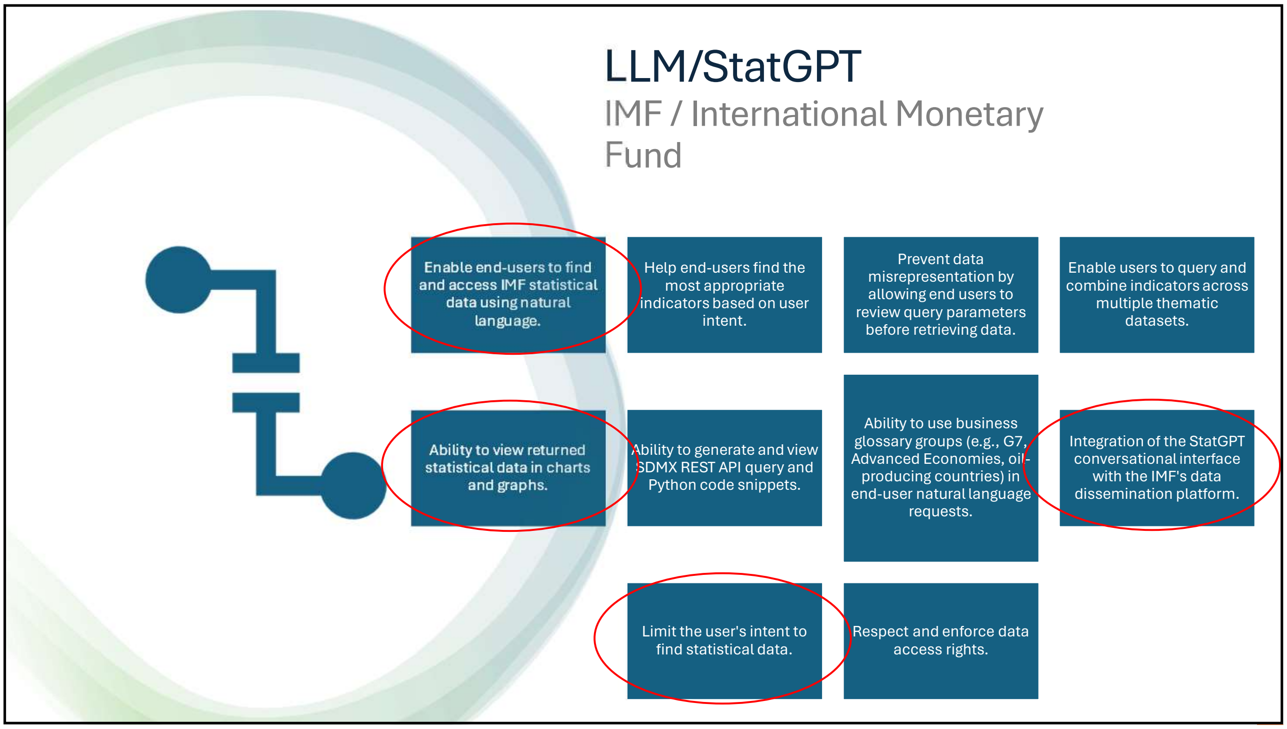
Assistance in interpreting data: LLMs can help users interpret complex statistical data by providing explanations, visualizations, and context. This helps users understand the meaning and implications of the statistics they are querying

Translation and Explanation from SAS Code to R (CSO Ireland)

Instituto Nacional de Estatística
Statistics Portugal

https://unece.org/sites/default/files/2023-12/HLGMOS%20LLM%20Paper_Preprint_1.pdf

17



LLM/StatGPT

IMF / International Monetary Fund

- Enable end-users to find and access IMF statistical data using natural language.
- Help end-users find the most appropriate indicators based on user intent.
- Prevent data misrepresentation by allowing end users to review query parameters before retrieving data.
- Enable users to query and combine indicators across multiple thematic datasets.
- Ability to view returned statistical data in charts and graphs.
- Ability to generate and view SDMX REST API query and Python code snippets.
- Ability to use business glossary groups (e.g., G7, Advanced Economies, oil-producing countries) in end-user natural language requests.
- Integration of the StatGPT conversational interface with the IMF's data dissemination platform.
- Limit the user's intent to find statistical data.
- Respect and enforce data access rights.

18

LLM in Official Statistics (risks)

- Issues of timeliness and quality of the data produced, based on the time and source of the datasets used by the LLMs.
- The problems of timeliness and accuracy may not always be obvious to the "average" user of LLMs, nor may it be obvious that LLMs cannot currently produce relevant and up-to-date statistics.

https://unece.org/sites/default/files/2023-12/HLGMOS%20LLM%20Paper_Preprint_1.pdf

References

Beck, M., Dumpert, F., Feuerhake, J., (2018), Machine Learning in Official Statistics, (n.p.), disponível em: https://arxiv.org/abs/1812.10422

McCarthy, J., 2004, What is Artificial Intelligence (n.p.), disponível em: https://www-formal.stanford.edu/jmc/whatisai.pdf

ModernStats, Large Language Models for Official Statistics, HLG-MOS White Paper, December 2023, disponível em: https://unece.org/sites/default/files/2023-12/HLGMOS%20LLM%20Paper_Preprint_1.pdf

UNECE Machine Learning Team, November 2018, Wesley Yung (Canada), Jukka Karkimaa (Finland), Monica Scannapieco (Italy), Giulio Barcarolli (Italy),
Diego Zardetto (Italy), José Alejandro Ruiz Sanchez (Mexico), Barteld Braaksma (Netherlands), Bart Buelens (Netherlands), Joep Burger (Netherlands), The Use of Machine Learning in Official Statistics

UNECE, Machine Learning for Official Statistics, (2021), disponível em: https://unece.org/sites/default/files/2022-09/ECECESSTAT20216.pdf