Sprint on Artificial Intelligence and Data Science for Economic Statistics

Webinar 2, 12 December 2024, 07:00–10:00 AM (GMT-4)

Session 1:
Welcome and opening remarks

The event began with remarks from Stefan Schweinfest, Director of the UN Statistics Division, and Osama Rahman, Chair of the UN Data Science Leaders Network. They outlined the webinar's objectives, emphasizing AI's role in modernizing statistical systems and empowering NSOs with state-of-the-art tools and methodologies.

Session 2:
Introduction and overview of AI and data science for official Statistics

Presentation 1: Generative AI for Official Statistics

InKyung Choi from UNECE delivered a foundational presentation on generative AI, a subset of artificial intelligence that creates new data and content. She discussed AI's broad potential to enhance statistical organizations' capabilities, including data expansion, operational modernization, and improved engagement with citizens. Specific examples included generative AI's role in text classification (e.g., COICOP and NACE systems), coding assistance, chatbot-based data collection, and dissemination tools for communication and user support.

The presentation underscored critical risks associated with generative AI, such as confidentiality breaches, accuracy issues, misuse, ethical concerns, and copyright challenges. To address these risks, it is crucial to implement technical safeguards, like open-source, self-hosted models, and governance strategies with human oversight.

Presentation 2: Statistics Canada’s Roadmap to AI Adoption

Wesley Yung presented Statistics Canada’s systematic approach to adopting generative AI technologies, focusing on maintaining organizational relevance and improving services to citizens. Their strategy prioritizes integrating AI responsibly across their operations, underpinned by governance structures, workforce upskilling, and enabling infrastructure.

Highlights included piloting use cases such as LLM-driven census reference searches, chatbot tools for interviewers, and AI-augmented workflows. The AI governance framework emphasized ethical guidelines, technology reviews, and legal compliance. StatCan's modular playbooks and tailored workforce training were pivotal to AI democratization, with a phased roadmap ensuring scalability and risk mitigation.

The presentation also showcased high-impact use cases, including the automation of code generation, NLP-based classification, and Microsoft Co-Pilot deployment for productivity enhancement. StatCan's strategic governance model ensures that AI aligns with organizational priorities while balancing risks.

Session 3: Generative AI and Data Science Initiatives in Statistical Production

Presentation 1: Guidelines for Preparing Open Data for Generative AI

Victoria Houed outlined the U.S. Department of Commerce's vision for democratizing access to public data using generative AI. She emphasized developing guidelines to make open data AI-ready, ensuring its reliability, transparency, and effective utility for model training. The department prioritizes rich metadata, improved data navigation systems, and licensing clarity to enable trustworthy AI applications.

Commerce’s generative AI-ready data pilots aim to enhance accessibility, monitor model performance, and innovate continuously, balancing openness with security and integrity.

Presentation 2: ClassifAI for classifying free text to a statistical classification

Mat Weldon from ONS, UK, introduced ClassifAI, a framework leveraging advanced LLMs for text classification tasks within NSOs. By using a Retrieval-Augmented Generation (RAG) approach, the system efficiently assigns free text data from labor surveys to industry

classifications (SIC). Early results showed notable accuracy improvements, with an API-ready framework designed for scalability, security, and interoperability.

Challenges addressed included model uncertainty, mitigated through calibrated confidence scores and conformal prediction techniques. ClassifAI's principles—security, efficiency, quality, and sustainability—ensure the system's readiness for large-scale operational deployment.

Presentation 3: Caliper Deployment at the UN Global Platform

Carola Fabi from FAO introduced Caliper, a comprehensive toolset for managing statistical classifications. Caliper supports the entire lifecycle of classifications, emphasizing metadata as a public good. The solution, based on open-source technologies like VocBench and ShowVoc, ensures FAIR (Findable, Accessible, Interoperable, Reusable) principles through RDF and SPARQL standards.

By offering modular tools and fostering partnerships (e.g., UNSD, FAO, and BC3), Caliper enables seamless classification updates, querying capabilities, and cost-effective infrastructure sharing across stakeholders.

Presentation 4: Leveraging Open-Source Language Models

Alejandro Pimentel from INEGI shared their experience automating employment and economic activity codification using AI models such as BERT and FastText. The AI-driven approach significantly reduced manual workloads, improved accuracy, and highlighted the importance of high-quality training data. Future plans include advancing with LLMs and building a robust ground truth database.

Presentation 5: Generative AI and official statistics: the project of the UNECE High-level Group for the Modernization of Official Statistics

Vytas Vaiciulis from CSO Ireland and Olivier Sirello from BIS presented an extensive collaborative project on the integration of generative AI in official statistics, involving 13 statistical offices, 3 international organizations, and other stakeholders. The project aims to explore real-world applications, prioritize expectations, and address the challenges,

limitations, and risks associated with generative AI. The ultimate outcome will be a comprehensive report titled *"Generative AI in Official Statistics."*

The project began with an international survey conducted under the Conference of European Statisticians to collect and classify use cases, resulting in the creation of a UNECE repository. The framework focuses on two interlinked components: technical implementation and governance, ensuring balanced organizational capability while preparing for future challenges.

The work is organized in 6 chapters:

- Ch. 1: Building organizational capability
- Ch. 2: Using and Implementing Generative AI
- Ch. 3: Governing and Managing Generative AI
- Ch. 4: Mitigating and monitoring risks
- Ch. 5: Sharing data, tools and knowledge
- Ch. 6: Preparing for the future

Session 4:
Generative AI for Dissemination and Interpretation of Statistics

Presentation 1: Compilers Hub: A Global Solution for Statistical Collaboration

Michael Stanger from IMF presented the Compilers Hub project, which will develop a user-friendly digital platform designed to serve the global statistical community by fostering knowledge sharing, collaboration, and co-development among stakeholders, including the IMF, international organizations, and government officials.

The Hub is organized around three core experiences:

1. Resources: A centralized digital library offering manuals, training materials, tools, templates, and conference links, all tagged for easy search and navigation.
2. Community: A space for knowledge sharing through forums, discussions, and technical Q&A.
3. Collaboration: Facilitating co-development projects and skill-matching opportunities among users.

Access to the Hub is tiered into four levels—Visitors, Members, Administrators, and Owners—each with varying permissions to view, contribute, or manage content. A notable feature is the 'Talk2manuals' Bot, an AI-powered chatbot designed to simplify navigation of

extensive manuals by providing user-friendly, interactive access to information while maintaining accuracy through resource rail-guarding.

Presentation 2: IntelliStatCan: AI Chatbot for Statistical Information

Milana Karaganis from Statistics Canada presented IntelliStatCan, a generative AI chatbot designed to enhance access to Statistics Canada's 18,000+ publications. Built with a RAG architecture, IntelliStatCan integrates Azure OpenAI services to deliver accurate responses with relevant citations. The MVP demonstrated early successes but also highlighted challenges in search orchestration and response quality.

Next steps involve hybrid search optimization, GPT-4 integration, and custom workflows to refine user experience and expand accessibility for Canadian audiences.

Presentation 3: Automatic Document Generation

Alejandro Pimentel and Elio Villaseñor from INEGI, Mexico, showcased their application for generating "industry insights" based on official statistics using LLMs. The approach involves defining key indicators (e.g., GDP, inflation), retrieving data via internal APIs, and applying fine-tuned LLMs for interpretation. Challenges like arithmetic errors and inconsistent language outputs were mitigated through prompt engineering and meta-prompt strategies.

INEGI's AI-powered insights offer a scalable solution for communicating complex statistical data in user-friendly formats, combining interactive visualizations with automated narratives.

Presentation 4: Governance and Responsible AI

Christos Sarakinos from Statistics Canada provided their experience in the design of responsible AI governance frameworks. He highlighted the principles of AI ethics, transparency, and accountability, aligning them with the Fundamental Principles of Official Statistics. Statistics Canada's Office of Responsible AI (ORAI) ensures robust governance through audits, playbooks, AI reviews, and red-teaming exercises.

The session emphasized the importance of balancing innovation with risk mitigation and preparing statistical systems for emerging AI challenges.