



Contribution ID: 317

Type: **Session**

AI for Science - CODATA Convened Session

Wednesday 15 October 2025 09:00 (1h 30m)

This session will explore cutting edge developments in relation to AI for Science and, in particular, AI and Data. AI offers groundbreaking opportunities and poses numerous challenges regarding the reproducibility of science and the conduct of scientists. How do we prepare AI-ready data, build and use specialized AI models, while establishing ground truths and a new form of explainability? What response is necessary from science policy makers, leaders, and decision makers? The session will include a discussion of the essential scientific response to AI developments from various perspectives, including those of scientists and AI specialists, and of policy makers. The session will also feature the launch of ISC Policy Primers and the CODATA White Paper on the Role of Data in AI for Science. The presentations will be followed by structured discussion and audience involvement.

Chair: Simon Hodson, CODATA Executive Director

Speakers:

1) Mercè Crosas, Director of Computational Social Science, Barcelona Supercomputing Center and CODATA President: 'Becoming AI-ready in Science'.

Mercè will open the session and describe the landscape of AI developments in scientific research, highlighting the steps needed to use, modify, or develop AI models in science without compromising rigor and open science principles. The talk will include a discussion of the best practices and standards that need to be extended to help domain scientists and research/AI engineers prepare for the adoption of AI throughout the research lifecycle. Finally, a few examples will be presented of the initial attempts to apply AI-readiness principles to research in the social sciences and humanities at the Barcelona Supercomputing Center.

2) Vanessa McBride, Director of Science, International Science Council: Launch of ISC Policy Primers on 'Data for AI Science', 'Environmental Impact of AI' and 'Types of AI for Science'.

The International Science Council's Centre for Science Futures has been gathering inputs from across the globe to assess how different science systems are preparing to harness the opportunities and mitigate the risks of artificial intelligence (AI) to science systems. Across these country case studies, from Chile to China, some common themes began to emerge. Two of these themes: the environmental impact of AI, and data for AI for science, have been explored in more depth in a set of primers written for a non-specialist, and possibly decision-making, audience. A third primer on the Types of AI in Science aims to provide a snapshot of terminologies, methods and example applications of AI in science. These primers will be launched at International Data Week, and we hope they will prove a useful starting point in translation between the scientific research communities and the policy-making communities as they navigate a rapidly evolving landscape in science.

3) Christine Kirkpatrick, Director of Research Data Services, San Diego Supercomputer Center, and CODATA Secretary General, and Tyng-Ruey Chuang, Associate Research Fellow at the Institute of Information Science, Academia Sinica, Taiwan: CODATA White Paper 'The role of data in AI for science'.

Christine and Tyng-Ruey will discuss the CODATA Concept Paper on the role of data in AI for Science. The advance of science is a common enterprise. Scientific collaborations across national borders, research institutions, and academic disciplines are now a norm, facilitated in part by timely online dissemination of research publications and data. Data management is often a joint effort, as project teams work together in the production, curation, and reuse of research datasets. As research work is increasingly driven by data and computation, especially in the context of building prediction models from large datasets, there is an urgent need to articulate a principled approach to research data so as to sustain and advance science. The presentation will give a brief overview of the opportunities, challenges, and recommendations from the paper, as well as a perspective on redefining FAIR, for the AI age.

4) Slava Tykhonov, Head of AI and Interoperability, CODATA: 'Semantic Croissant and CDIF for Responsible AI'

This talk introduces Semantic Croissant, a knowledge-graph based extension of the CroissantML standard that links datasets with external resources such as ontologies and vocabularies, while also providing structure, resources, transformation rules and semantic context to enhance their usability for AI models. Built on schema.org, extended with Responsible AI fields, and enriched through linkage to multilingual controlled vocabularies and ontology alignment, it enables machine-readable discovery and validation of datasets. The Cross-Domain Interoperability Framework (CDIF) further extends Croissant ML with a variable cascade model that captures measurement units, attributes, properties, and inter-variable relationships. This approach ensures FAIR compliance and supports seamless integration across institutional boundaries, including the transformation of variables across disciplines and the documentation of scientific experiments.

Primary author: HODSON, Simon (CODATA)

Co-authors: KIRKPATRICK, Christine (San Diego Supercomputer Center / CODATA); CROSAS NAVARRO, Merce (BSC-CNS); CHUANG, Tyng-Ruey (Academia Sinica, Taiwan); MCBRIDE, Vanessa (ISC); TYKHONOV, Vyacheslav (CODATA)

Presenters: KIRKPATRICK, Christine (San Diego Supercomputer Center / CODATA); CROSAS NAVARRO, Merce (BSC-CNS); HODSON, Simon (CODATA); CHUANG, Tyng-Ruey (Academia Sinica, Taiwan); MCBRIDE, Vanessa (ISC); TYKHONOV, Vyacheslav (CODATA)

Track Classification: SciDataCon2025 Specific Themes: Rigorous, responsible and reproducible science in the era of FAIR data and AI