Data and AI for science: Key considerations

CENTRE FOR SCIENCE FUTURES

International Science Council

To cite this report:

International Science Council (September 2025). *Data and AI for Science.*
DOI: 10.24948/2025.11

Authors: Natalia Norori, Denisse Albornoz and Vanessa McBride
Reviewers: Mohamed Farahat, Gloria Guerrero, Simon Hodson
Project coordination: Dureen Samandar Eweis, Vanessa McBride
Project chair: David Castle

Funding acknowledgement: This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Ottawa, Canada. The views expressed herein do not necessarily represent those of IDRC or its Board of Governors.

Design: Mr Clinton

Cover photo: imaginima

About the International Science Council

The ISC is an international non-governmental organization with a unique global membership that brings together 250 international scientific unions and associations, national and regional scientific organizations including science academies, research councils, regional scientific organizations, international federations and societies, and academies of young scientists and associations.

The ISC works at the global level to catalyse change by convening scientific expertise, advice and influence on issues of major importance to both science and society.

Contents

Summary

- The efficient and reliable use of artificial intelligence (AI) in science depends on well-curated machine-readable data.
- Algorithms that enable contextually aware AI approaches hold potential for scientific discovery and specialist knowledge generation. They depend on AI-ready scientific data.
- High-quality data is important for the ground-truthing of AI in science.
- AI tools are becoming useful aids in data stewardship of (AI-ready) scientific data.
- Biases and structural inequalities in datasets and access to infrastructure risk amplifying disparities in scientific capacity and outcomes.

Recommendations

- Convergence to existing data frameworks and standards, for example, FAIR-R and Croissant, should be used by scientists and data stewards.
- Data governance structures should go beyond technical standards to promote equity, access to compute resources, and capacity-building.
- Investment in data infrastructure and skills development is a prerequisite for efficient and competitive use of AI in science.
- Recognition of data stewardship careers in science, and incentives to encourage these skills, is a cornerstone implementation pathway of the above investment.

About this paper

This paper provides an overview of the technical, ethical and environmental factors to consider when preparing scientific data for artificial intelligence (AI), and how these factors align with the 'Open Science' movement. The information presented is relevant to researchers, data practitioners, scientific bodies and policy-makers for science.

The first section introduces the foundational concepts and discusses the advantages and challenges of making scientific data AI-ready. The second section examines the key considerations for data readiness for AI, and conversely, AI to curate data. We build on data standards while discussing AI-specific considerations such as machine-readability and bias mitigation, while highlighting ethical and environmental considerations around data readiness for AI in science. The third section discusses data readiness within the framework of Open Science, presents two case studies that illustrate how Open Science practices can support AI-readiness for scientific research.

It uses the following concepts:

- **Artificial intelligence**: In this paper, we use the Organisation for Economic Co-operation and Development (OECD) definition for AI, meaning an AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment (OECD, 2006).
- **Data and AI for science:** The use of data-driven AI for scientific research and discovery. The considerations discussed in this paper apply to the two types of data that are part of the scientific process:
  - **Scientific data**: The original observations, measurements and records collected and analysed during scientific research. These may be in their raw format or higher-level products.
  - **Scientific knowledge data**: The outputs and findings that emerge from analysing scientific data.
- **Metadata:** The considerations discussed in this paper apply to the use of data and the metadata that accompanies it. The OECD defines metadata as data that defines and describes other data (OECD, 2007).

The document focuses on examining the use of data and AI in the context of scientific research and knowledge production. This encompasses **AI as an instrument of the scientific method**, including its use in hypothesis formulation, data analysis and interpretation of results and **AI-enabled research outputs**, where AI contributes directly to the generation of new scientific knowledge.

The paper is part of a series of three primers that explore various technical dimensions of AI and its impact on science. The other primers are "Types of AI in science" and "Environmental considerations of AI for science".

Introduction

Data has long been central to scientific discovery, being the foundation for testing hypotheses, validating results and generating new insights. The volume of data produced as part of scientific research is currently at its peak, and this trend is expected to continue in the years to come (Clissa et al., 2023; Samborska, 2025). In addition to the rise in data volume, emerging technologies, such as wearable devices, have transformed the different formats and structures in which data can be collected (Clark et al., 2024).

The rapid growth of data has opened a wide range of possibilities for scientific research. Before the rise of AI, the adoption of data standards and open collaboration practices gained traction as the world began to recognize the many ways in which data could support and accelerate science. A key example is the Human Genome Project, which demonstrated the power of global collaboration and data sharing in mapping the entire human genome (Bentley, 2000).

In the age of AI, scientific data is being analysed at unprecedented speed across domains (Gao and Wang, 2024). For instance, in radiology, AI algorithms are trained on diagnostic imaging data to help identify abnormalities and tumours in X-rays, CT scans and MRI (Najjar, 2023). AI has also played an important role in anticipating and managing health emergencies, such as the COVID-19 pandemic (Islam et al., 2020). In environmental science, AI is being used for climate modelling and prediction, as well as to support biodiversity and conservation efforts, and to identify patterns of deforestation. In physics, AI is used to analyse collision events and identify anomalous features that may indicate the presence of new particles (Berman et al., 2023). AI is also being used as an agent to accelerate scientific discovery by suggesting alternate hypotheses, executing workflows to test them, and/or making inferences (see, for example, Google's AI co-scientist and GeoGPT).

Despite its widespread adoption, **data readiness for AI**, which refers to preparing scientific data and ensuring it is properly collected, cleaned and structured for AI use, remains a challenge. For example, the data preparation process requires a significant amount of time and workforce capacity. Data scientists report spending up to 80 percent of their time cleaning and preparing their data (Directorate-General for Research and Innovation and PwC EU Services, 2018). A lack of systems designed to evaluate data readiness at scale, data access and ownership concerns, and gaps in digital skills further contribute to this challenge.

Although the importance of **data readiness for AI** is widely recognized, there is limited evidence of how it is being assessed and what key factors researchers need to consider when making their data AI-ready.

Section 1: Foundational concepts

The quality of an AI model is inherently tied to the quality of the data used to train it. For AI to meaningfully contribute to scientific research, the data behind it must be carefully prepared and optimized for AI applications (Verhulst et al., 2025).

This concept, known as **data readiness for AI,** involves ensuring that datasets are properly collected, cleaned and structured before being used to train models. The Bridge2AI project defines data readiness as 'a set of characteristics of a dataset and its associated metadata that permit reliable, ethical analysis by AI methods within defined use cases and operational limits, with sufficient metadata to support reliable, appropriate post-model explainability analysis' (Clark et al., 2024).

Although high-quality data is critical for AI, data preparation is often viewed as a tedious aspect of model development, requiring a significant amount of time (Sambasivan et al., 2021). In response to this challenge, several tools have been developed to assess data readiness for AI (Hiniduma et al., 2025). However, it remains unclear to what extent these tools are being utilized and whether the criteria they propose are being met in practice.

There are several benefits to centring data readiness in AI for science. AI-ready data and metadata can enhance the trustworthiness of AI, improve model performance and minimize the risks associated with data bias (Hiniduma et al., 2024; Verhulst et al., 2025). High-quality data and metadata also minimize the need for extensive preprocessing and supports faster, resource-efficient model training, which can significantly reduce costs and data preparation time (Sorscher et al., 2022). Additionally, AI-ready data can contribute to transparency and reproducibility. Accurate data and metadata labelling, alongside information on data provenance, help researchers gain a clear understanding of the data's characteristics, allowing them to validate their research findings and build upon prior work (Kidwai-Khan et al., 2024). Beyond reproducibility, the process of preparing AI-ready data may naturally foster interdisciplinary collaboration, as it often requires input from experts across domains (Sambasivan et al., 2021).

However, despite the global use of AI in scientific research, not all scientific domains and geographies have the same levels of data readiness (Oxford Insights, 2025). Domains with well-established data practices, such as medicine and chemistry, are often overrepresented in AI-driven research (Bianchini et al., 2024). While these fields typically have more training data and resources to develop AI, this overrepresentation does not necessarily indicate greater AI-readiness. The inconsistent adoption of metadata schema and the availability of semantic resources are barriers to the effective use of AI in some areas of science. Similarly, the lack of funds and limited compute and storage infrastructure and gaps in training also limit the use of AI in science (Van Noorden and Perkel, 2023).

To add, researchers working in resource-constrained settings face additional challenges in preparing data for AI. Studies have found that although AI practitioners across settings have access to similar models, there are marked differences in data quality and access to

computing power in East and West African countries and India when compared to the United States (Sambasivan et al., 2021). Other challenges include a lack of digital data, limited data availability and volume, reduced internet and electrical connectivity, unclear data protection regulations and a gap in AI skills (Open Data Charter, 2025).

This reflects broader inequalities in AI. For instance, the vast majority of AI models are developed in the Global North, with data that does not accurately reflect the Global South. Despite this, many AI models trained on Global North data are generalized to Global South contexts, although it is well known that this can lead to poor performance and real-world consequences (Hagerty and Rubinov, 2019).

There is ongoing work to address these gaps. One example is the Data Science Without Borders project, led by the African Population and Health Research Centre, which aims to improve data systems in Africa by enhancing the data collection, management and analysis capabilities across institutions in Ethiopia, Senegal and Cameroon. Another example is the Data Science-Innovation Africa (DS-I Africa) Open Data Science Platform (eLwazi ODSP), an open data platform for depositing, sharing and accessing data and deploying tools on computer environments suited to the African context.

Lack of data readiness can lead to the development of AI algorithms that produce biased outputs (Hiniduma et al., 2024). The consequences of such biases are especially serious in high-stakes domains where AI is increasingly used to support critical decision-making (Sambasivan et al., 2021). This issue is even more pronounced in Global South settings, where people face a higher risk of the discriminatory outcomes of AI when compared to those in the US and Western European countries (Hagerty and Rubinov, 2019). Given these challenges, and to ensure the effective integration of AI in scientific research for the public good, it is essential to consider the different factors required for the shift towards AI-ready data.

Section 2: Key considerations for AI-ready scientific data

Data need to meet a set of characteristics before they can be used responsibly in AI for science. This includes the application of data standards, as well as specific characteristics relevant to AI (Verhulst et al., 2025). In addition, the ethical, environmental, and structural implications of the use of data in AI need to be evaluated throughout the data lifecycle. In this section, we discuss some of the frameworks available to support researchers in interrogating and improving the quality of their data. Some of these were developed before the emergence of AI in scientific research, while others have been modified to meet its specific needs.

At the same time, a suite of AI tools is emerging to facilitate data and metadata preparation (sometimes referred to as FAIRification – see FAIR principles in the next section). These include the use of large language models to verify metadata standards (Sundaram et al., 2024), cloud-based research data ecosystems (The European Open Science Cloud research data commons, 2025) and agentic AI tools to curate and publish data (SENSCIENCE, n.d.).

2.1 Technical considerations

2.1.1 GENERAL DATA STANDARDS

The importance of data standards in scientific research is well-documented and has gained traction over the last decade (Wilkinson et al., 2016). While concerted efforts to improve management, sharing and use of scientific data have taken place within specific disciplines for some time, the introduction of the FAIR principles provided a catalyst for additional initiatives across domains, institutions and regions.

FAIR PRINCIPLES

One of the most influential shifts towards higher-quality metadata and semantics has been the introduction of the FAIR Guiding Principles for scientific data management and stewardship: **F**indability, **A**ccessibility, **I**nteroperability and **R**eusability. These guiding principles emphasize the importance of well-structured metadata to enhance data discoverability. The FAIR principles, if implemented, can help improve data usability. They are a component of Open Science practice – discussed in Section 4 – and could also serve to foster collaboration throughout the research ecosystem.

According to the FAIR principles, metadata and data should be *findable* by both humans and machines. They also need to be *accessible*, with clear guidance on the authorization steps required. In addition, the metadata and semantics should enable the integration of datasets and their *interoperation* with the applications and workflows used for analysis, storage and processing. Finally, both metadata and data need to be reusable and well-described to enable *replication* (Wilkinson et al., 2016).

2.1.2 AI-SPECIFIC DATA STANDARDS

AI-specific data standards are beginning to emerge in practice. This section discusses some general considerations as well as a specific framework.

AI requires machine-readable data, i.e. structured in a way that allows computers to process it directly, without requiring extensive reformatting. Therefore, in some cases, the process of making scientific data machine-readable may involve tedious extraction, reformatting and cleaning. Scientific data, which includes the original observations collected and analysed during research, is not always made available, and when it is made available, it is not often in machine-readable formats. While machine-readability has become particularly relevant in the context of AI-ready data, its importance was established well before the current emphasis on AI.

Scientific knowledge data, which includes the outputs and findings that emerge from analysing scientific data, is made machine-readable typically only after publication. To facilitate integration with existing knowledge, nanopublication approaches like *Reborn*, propose shifting towards producing machine-readable scientific knowledge at the pre-publication stage by utilizing infrastructures built on knowledge graphs (Stocker et al., 2025).

While the FAIR principles provide a solid foundation for data, they do not address AI-specific needs. To bridge this gap, the FAIR-R conceptual framework extends the original FAIR principles by incorporating AI-readiness. It emphasizes that datasets should also be structured to meet the quality requirements of AI applications, such as providing labelled data for supervised learning or ensuring comprehensive and representative coverage for unsupervised learning (Verhulst et al., 2025).

CROISSANT

Croissant is a metadata schema for AI-readiness by MLCommons that was developed to enhance the discoverability, portability and interoperability of machine learning datasets. Before Croissant, there was no standardized format to organize the metadata behind machine learning datasets, making the process of finding machine learning data tedious and time-consuming. Croissant can describe most types of data used in machine learning workflows and lets users add semantic descriptions and machine learning-specific information. It allows the data to be discoverable through search engines and loaded into different machine learning platforms without the need for reformatting (Akhtar et al., 2024).

2.2 Data quality, volume and bias

2.2.1 DESCRIPTION

While data standards provide a foundation for data management in AI for science, they are only one aspect of the process. Data quality is of key importance.

High-quality data, metadata and semantics ensure that models are trained on inputs that accurately reflect the complexity and diversity of real-world conditions, reducing the risk of biased or inaccurate outputs. Poor data quality can compromise model performance and the trustworthiness of AI-driven decisions (Balahur et al., 2022). Data quality refers to, among

other characteristics, accuracy, completeness, consistency, uniqueness and fitness for purpose. These considerations are essential to ensure that AI models are trained on data that will translate to real-world settings and use cases (Open Data Institute, 2023). Lower quality data, with relevant metadata, can still prove useful in answering specific scientific questions, but not others. An element of judgement, supported by metadata descriptors, is still required to assess whether certain data is appropriate and relevant for a given research question.

Completeness ensures that all relevant variables are captured and fully described, enabling models to learn from the full context. **Accuracy**, on the other hand, measures the degree to which the data aligns with the information it represents, and seeks to quantify the probable error (Hiniduma et al., 2024). Data-centric approaches can help improve the quality of AI data and complement model development. Unlike methods that focus on increasing data volume to boost model performance, data-centric AI focuses on improving the quality of the training data (Bhatt et al., 2024). Including machine-readable expressions of estimated data quality in the metadata would be an important step to achieve this. This can reduce computational demands without compromising model performance and aligns with sustainable AI principles by lowering the carbon footprint associated with compute-intensive training (Sorscher et al., 2022). **Uniqueness** ensures that no data points are duplicated, as such duplication will affect the results. **Consistency** explores the extent to which different measurements of the same quantity in or across datasets agree and is one of the components in determining how precisely a result might be estimated from a given set of data.

DATA VOLUME

The volume of available training data has grown exponentially. Since 2010, it has doubled every 9–10 months (Samborska, 2025). Adequate data volume is a key requirement for AI-readiness, as it ensures models can learn effectively from the training data. However, increasing data volume does not necessarily translate to improved output quality (Sorscher et al., 2022). If the data volume is too small, the models may perform well during the training phase but struggle when applied to different datasets: This is defined as **underfitting**. On the other hand, extremely large datasets can drive up costs, increase environmental impact, and may not lead to meaningful performance gains: this is defined as **overfitting** (Hiniduma et al., 2024).

HETEROGENEITY AND CLASS IMBALANCE

Heterogeneity refers to the diversity that exists within data, including differences in data sources, formats, categories, processes and represented groups (Liu and Cui, 2025). Working with heterogeneous data increases the time and effort required for data cleaning and preparation before model development. It also poses challenges in the deployment of AI as it makes learning patterns more complex and can impact model accuracy (Kakko, 2025).

Class imbalance arises when the different categories or classes within a dataset are unevenly distributed, resulting in some being more prevalent than others. Models trained on imbalanced data tend to be biased towards the more common classes, which can lead to poor predictions for the classes that make up the minority of the dataset. This happens because the model has not seen enough samples from these minority groups to learn how to recognize them accurately. Class imbalance can lead to misclassification and has a large

impact in high-stakes domains, particularly affecting the ability of AI models to predict infrequent events such as rare diseases and the identification of protected species (Hiniduma et al., 2024).

DATA PERISHABILITY

As data ages and the conditions under which it was created evolve, it can become less relevant. Understanding the data 'shelf life' helps inform decisions about how long data should be stored and the timeline for retraining models. Limiting data storage could help reduce the environmental impact of AI, as holding large volumes of data consumes significant energy (Wu et al., 2021). On the other hand, not all data is created equal. Some data, such as natural phenomena data, represent unique observations that cannot be re-collected if lost. In such cases, preserving data is critical as its loss may hinder future research, and the priority should be to guarantee that the stored data is accessible and usable in the long term (Borgman, 2017).

DATA LOCALITY

Data locality is the principle of moving computing closer to where data is stored, rather than transferring data to a central computing location. This represents a shift from traditional approaches, where data was typically moved to centralized servers for processing. Data locality helps reduce the time and environmental cost associated with transferring large datasets by minimizing data movement. This approach can also enhance performance, as local data access is faster and more efficient than retrieving data from remote locations (Usman et al., 2022). Data locality is relevant for frugal AI models that pursue high performance with minimal resources, including data, computing power, and energy or edge AI applications that bring AI computation closer to where data is generated.

2.2.2 DATA OPTIMIZATION TECHNIQUES

Techniques such as **synthetic data, data augmentation** and **data pruning** can help achieve an optimal data sample size and reduce the risk associated with heterogeneous data and class imbalances.

Synthetic data involves generating artificial data that replicates the characteristics of real-world data. It is a useful technique for increasing the volume of training data required for model development. It can be used to mitigate the risk of class imbalance when training models in scenarios where real-world data is scarce or difficult to obtain (Liu et al., 2024). Synthetic data generation techniques can be used to create synthetic versions of sensitive datasets, allowing researchers to develop models with confidence that their results can be applied to the real datasets (Howe et al., 2017). This approach is widely applied in the health sciences, where, for example, synthetically generated AI images are being used to train skin cancer diagnostic models (Tai et al., 2024).

Data augmentation can also be used to increase AI data volume. It involves creating modified copies of existing data points within a training dataset and can help improve the model diversity by introducing variations of the training data (Tabbakh et al., 2024). Data augmentation can be considered in scenarios where collecting large amounts of labelled data is challenging. For example, in Indonesia, data augmentation is used to enhance underwater

images, allowing AI models to recognize different coral types and reef conditions (Andono et al., 2024).

Conversely**, pruning** can help reduce the sample sizes of datasets to avoid overfitting. It involves removing duplicated, low-quality or less informative data that is unlikely to impact the overall performance of the model. Data pruning enables AI models to focus their learning on the best available data, cutting down costs and improving the sustainability of AI models (Sorscher et al., 2022; Tabbakh et al., 2024). One risk of data pruning is that it can worsen class imbalance, which may lead to classification bias. To address this, a new pruning approach called *DRoP*, helps select how many samples of a class to take, ensuring that the pruning process retains a balanced class distribution (Vysogorets et al., 2025).

Oversampling and **undersampling** are two opposite data resampling techniques that can help reduce class imbalance and mitigate bias. They can help generate new data for classes that are underrepresented or remove data for overrepresented classes within a dataset (Mondal, 2023).

Data readiness for AI: Assessment tools

The Data Nutrition Project: This project enables researchers to create metadata labels that accurately describe the key classes and variables within a dataset (Holland et al., 2018).

AI Data Readiness Inspector (AIDRIN): A framework for evaluating data readiness for AI qualitatively and quantitatively. It produces reports that can support researchers in identifying flaws in data (Hiniduma et al., 2025).

The Earth Science Information Partners (ESIP) Data Readiness Checklist: This checklist was developed by the ESIP Data Readiness Cluster to assess the level of AI-readiness for Earth science applications (ESIP Data Readiness Cluster, 2022).

The Data Provenance Initiative: A volunteer collective of researchers that audits the licensing and attribution of datasets used in AI. The results are shared in their Explorer tool (Longpre et al., 2023).

Bridge2AI: A framework to assess and improve data readiness for AI in Biomedical data (Clark et al., 2024).

FAIR-R: A framework that incorporates readiness for AI into the FAIR principles (Verhulst et al., 2025).

UNESCO Readiness Assessment Methodology: An AI-readiness assessment tool designed to help countries evaluate their AI-readiness status and understand how prepared they are to apply AI ethically and responsibly (Global AI Ethics and Governance Observatory, n.d.).

2.3 Ethical considerations

As AI becomes increasingly integrated into science, it raises important ethical questions that extend beyond model performance. This section touches on key ethical considerations related to data handling and AI system deployment that need to be accounted for when working with data and AI in scientific contexts. It discusses bias awareness, the role of human oversight, data provenance, sovereignty, consent and transparency.

BIAS AWARENESS

Biased AI models make decisions that favour certain groups more than others in unfair ways (Ntoutsi et al., 2020). Different sources of bias can enter AI models at all stages of the development process.

Data-driven bias occurs when the training data used in AI models for science is not representative of the population. The lack of availability of scientific data can introduce bias, as researchers often have to rely on limited publicly available datasets to develop AI models. In addition, when algorithms are trained on skewed data, they tend to reinforce patterns and assumptions that reflect the dominant groups present in the data. This is particularly relevant in scientific studies, in which the majority of scientific data originates in the Global North. For instance, while genomic data is being used to accelerate precision medicine, over 80 percent of participants in genomic studies are of European ancestry (Norori et al., 2021). Therefore, the usability of AI algorithms trained on Global North data and then applied to the Global South has been questioned, as the Global South may be at higher risk of the harms of AI and less likely to benefit from its advantages (Hagerty and Rubinov, 2019). AI models trained on skewed data are more prone to produce inaccurate results that undermine the validity of research outputs (Royal Society, n.d.).

Gender and racial bias hinder data readiness for AI. A study that evaluated three major commercial gender classifications found that darker-skinned females are the most misclassified group with error rates of up to 34 percent. In comparison, the error rate for light-skin males was 0.8 percent (Buolamwini and Gebru, 2018). A different analysis of 133 biased AI systems found that 44 percent present gender bias, and 26 percent show both gender and racial bias (Smith and Rustagi, 2021). Gender and ethnicity disaggregated data should be collected, analysed and reported to identify gaps and assess potential underrepresentation or overrepresentation.

Human bias refers to the systematic and often unconscious tendencies that lead individuals to favour or disfavour certain people, groups or ideas, often in a way that is subjective or unfair. Human bias can also infiltrate AI models, where it may be replicated or even amplified, leading to real-world harm. Both human and data-driven biases are often challenging to detect (Norori et al., 2021). It is important to acknowledge that these biases exist and apply techniques to identify and reduce them during the data collection and preparation stages. These techniques include resampling to address class imbalance, generating synthetic data, applying standardization techniques to reduce heterogeneity, and improving the quality of data labels. Involving interdisciplinary experts in the data preparation process can also help identify bias at the early stages of model development (Slesinger et al., 2024).

HUMAN OVERSIGHT

Maintaining appropriate human supervision over AI systems can help reduce the risks to safety and fundamental rights that may occur when AI systems are used, especially in high-stake domains. The principle of human oversight acknowledges that while AI systems can process vast amounts of data and identify complex patterns, human judgement remains essential for contextual understanding, ethical reasoning and accountability. Therefore, AI systems need to be designed with features that enable humans to understand its judgement and decisions, monitor its performance and intervene when necessary (EU Artificial Intelligence Act, n.d.).

DATA PROVENANCE, SOVEREIGNTY AND CONSENT

Understanding the **provenance** and history of the data, including who created it, who owns it and what transformation it has gone through, helps when evaluating AI models. It also supports reproducibility, as researchers can understand the context in which the data was created and assess its reliability (Hiniduma et al., 2024). Data provenance is an important component of metadata and plays a role in helping assess the quality of datasets.

In AI for science, datasets are frequently pulled from public repositories or scraped from online sources. This raises concerns about data **sovereignty,** as there is a lack of clarity about who owns the data and whether the humans represented in the datasets provided their informed **consent** for their information to be collected or to be used in a specific manner. This is particularly relevant when collecting data from communities affected by extractive research practices, such as Indigenous Peoples. The **CARE** Principles for Indigenous data governance emphasize **C**ollective benefit, **A**uthority to control, **R**esponsibility and **E**thics, and encourage stronger control and oversight over data generated from Indigenous Peoples. The aim is to ensure that Indigenous data facilitates the collective benefit of Indigenous Peoples (Carroll et al., 2022). For example, the Local Contexts initiative was designed to support Indigenous Peoples with tools to gain control over how their data is collected, accessed and shared.

TRANSPARENCY

While data **transparency** is widely recognized as essential to maintain scientific rigour, there is an ongoing debate about the level of transparency required for AI. On one hand, disclosing details about training data and deployment processes allows researchers and the public to interrogate AI models and identify potential sources of bias. On the other hand, there are concerns that transparency might reveal sensitive information and create security risks (Luna, 2024). AI models are often developed and operate in black-box environments. This can create a barrier to research reproducibility, making it difficult for researchers to verify and replicate scientific discoveries (Hardings et al., 2023; Royal Society n.d.; von Eschenbach, 2021).

Due to the large-scale nature of the datasets used for AI, it has become increasingly difficult to assess if the training data unintentionally contains **sensitive** information. This raises concerns around privacy and security. Data classification tools can be used to anonymize sensitive information at the data preparation stage and minimize these risks (Feretzakis and Verykios, 2024). Synthetic data can also be used to reduce privacy risks (Howe et al., 2017).

The integration of AI in science involves broader societal issues not discussed in this section; including power asymmetries between data providers and researchers, systemic injustices that may be perpetuated through algorithmic decision-making, cultural sensitivity in cross-cultural research contexts, and extractive data practices. While these critical topics fall outside of the scope of this paper, they do need to be addressed to prevent unintended consequences and ensure applications of AI in science do not exacerbate social inequities.

2.4 Environmental considerations

The growth in data volume has raised new concerns about the environmental footprint of AI. The increase in sample size is facilitated by the rise in computing power and results in an increased environmental footprint of AI. It is estimated that by 2050, the data industry will be responsible for more carbon emissions than the automotive, aviation and energy sectors combined (Massey and Moriniere, 2023).

Data **augmentation** is often used to reduce the risk of bias and enhance model performance. However, since it involves growing the data volume, it can increase the storage and training demands, requiring more computational resources. On the other hand, when done well, data augmentation can improve the efficiency of AI while reducing energy consumption (Tabbakh et al., 2024). Similarly, data **pruning** can assist in cutting down storage volume and training time, lowering the overall footprint of AI models (Sorscher et al., 2022). Understanding data **perishability** can also help reduce the resource requirement for data storage, as data is only stored for the amount of time it is needed (Tabbakh et al., 2024).

To mitigate the environmental impact of AI in science, it is important to consider the direct and indirect effects of the data lifecycle. This awareness can help inform decisions and develop strategies to reduce the environmental footprint.

Section 3: Context-specific use of AI

The advent of large language models has enabled the deployment of tools that accelerate the production of scientific knowledge from existing knowledge and new data. While large language models are usually trained on vast amounts of language data, they can be tuned to respond to very specific questions by augmenting their broad training data with subject-specific training data.

Two approaches are gaining traction:

- **Retrieval Augmented Generation** allows large language models to incorporate new information (usually scientific knowledge data) without an arduous model retraining process, and in general with better results (Belcic, 2024). This type of finetuning is realized in science in tools such as Elsevier's *ScopusAI*, where the large language models are augmented by the corpus of scientific literature.
- **Model Context Protocol** is an open standard for connecting AI systems with data sources. While this facilitates the integration of local or specialized scientific data with multiple AI-agents, it also means that data privacy and security can be maintained. The protocol provides logging, which is key to supporting AI validation and reproducibility of scientific results. Model Context Protocol was publicly released by Anthropic in late 2024, and scientific use cases are still emerging.

Section 4: Data readiness for AI within an Open Science framework

Open Science is the movement to make scientific research, data and their dissemination available to any member of an inquiring society, from professionals to citizens (ORION Open Science, 2017). Open scientific knowledge (of which data is part) is a cornerstone of the 2021 UNESCO Recommendation on Open Science (UNESCO, 2021). The adoption of Open Science practices has helped accelerate scientific discovery. Examples include the Human Genome Project, which publicly released genomic data and led to major breakthroughs in medicine and genetics, the Sloan Digital Sky Survey, which made high-quality astronomical data freely available and transformed research in astrophysics, and the Copernicus Data Space ecosystem, which provides access to geospatial data from the Copernicus Sentinel missions (Bentley, 2000; York et al., 2000).

Open Science promotes a broad cultural shift and encourages reproducibility, collaboration, access to and inclusion in scientific research. The FAIR approach to data is part of the Open Science landscape and can help improve scientific research practices. These approaches strive to make data more transparent, accessible and usable for the scientists and the general public (Umbach, 2024).

The current challenge to Open Science posed by AI is in transparency, explainability and reproducibility. In many cases, the inherent black-box nature of deep learning algorithms can obscure biases, an understanding of root causes of the results, and they can hinder reproducibility – a central principle of science. Care must be taken in using data and algorithms in a way that builds, rather than undermines, trust in the scientific process (von Eschenbach, 2021; Lawrence and Montgomery, 2024).

Open Science fosters the development of open infrastructures and tools such as open-source software and open data repositories. It has, in many ways, paved the way for today's advancements in AI and computing by making software and data accessible and providing a foundation for others to build upon. For example, AlphaFold, which made groundbreaking contributions to predicting protein structures from amino acid sequencing, was made possible, in part, by the Protein Data Bank and the broader genomic revolution it helped initiate, which laid the groundwork for data sharing, infrastructure and collaborative research practices in genetics (Burley et al., 2017).

CASE STUDY: AlphaFold 1

Proteins are complex molecules composed of sequences of amino acids. They are a vital component of life and have been studied for decades. However, due to their complexity, the way in which proteins fold in nature is still not fully understood. It has been calculated that it would take longer than the age of the universe for a protein to fold in every conformation available to it. This challenge, known as the protein folding problem, has been a long-standing issue in biology (Finkelstein, 2018). Before AlphaFold, we could only predict the 3D structures of 17 percent of the proteins present in the human body. AlphaFold has substantially increased that number to 98 percent.

AlphaFold serves as an example of how open data has facilitated scientific discovery. The program relied on large, high-quality datasets from the Protein Data Bank, an open-source dataset that aligns with the FAIR principles, making it suitable for training deep learning models.

AlphaFold not only leveraged the power of open data, but it is also a project that directly contributes to the Open Science movement. DeepMind, the company behind AlphaFold, made the source code for AlphaFold freely available and launched an open dataset that includes the structure of 350,000 proteins, with more to be added in the near future (Jumper et al., 2021).

Open data sharing, as embodied by the FAIR principles, is a key component of the Open Science movement. Open data can be easily accessed and reused, facilitating research reproducibility and efficiency. However, making open data accessible relies on physical storage infrastructures and a skilled workforce to curate the data, apply metadata and manage the associated semantic resources. Strengthening these skills within the scientific community will help to realize the gains described above.

Open Science provides a framework for data for AI for science, but care should be taken so that open data is not employed for perverse or nefarious purposes. The **CARE** principles were developed to assert Indigenous data sovereignty and the need for data gathering and access to be bounded by ethical considerations and respect. CARE highlights that data collected from or about Indigenous Peoples should also be governed and shared on Indigenous Peoples' terms (Carroll et al., 2022). The ethical principles of CARE have wide applicability. The FAIR principles are explicitly focused on technical issues: how to describe data and metadata to make them machine-actionable. This includes specifying the conditions under which protected data might be accessed. CARE provides a complementary framework to FAIR towards responsible data sharing (Wong et al., 2024). If applied in conjunction, they can support the principles of Open Science and help mitigate safety risks.

CASE STUDY: PrevisIA

PrevisIA is a risk prediction model developed by the Institute for the Amazon's People and Environment. It assesses the risk of deforestation in the Amazon region and currently predicts that 8,959 km² of habitat in Brazil's Amazon, including Indigenous lands and conservation units, are at risk of deforestation. The tool leverages a wide range of data sources, including geospatial data, socio-economic data, urban infrastructure and historical records of deforestation.

PrevisIA serves as a case study of a partially open publication model, where the results of the model and methodology are made openly available, but the training data remains closed due to privacy concerns. PrevisIA was trained with data from Brazil and developed by Brazilian researchers who understood the context of the region and its specific needs. It demonstrates how local AI can be trained and developed in its intended environment, addressing issues relevant to the region.

PrevisIA was created through a collaboration between Imazon, Fundación Fundo Vale and Microsoft. The model generates real-time alerts, enabling the government and conservationists to develop preventive measures. This makes PrevisIA a prime example of a cross-sector partnership that involves government, the public sector, industry and academia.

While PrevisIA has helped improve efficiency by narrowing the focus to the regions at highest risk of deforestation, the environmental impact of PrevisIA has not been publicly disclosed (Alves et al., 2024).

Conclusion

AI is transforming the way science is conducted and will continue to impact both the generation of new knowledge and the application of research results. Data is the driving force behind AI, and as the use of AI increases, the volume of data is also expected to grow. For these results to benefit society, the training data must be AI-ready.

Data readiness for AI encompasses several key factors that must be considered at all stages of the data lifecycle, from data generation to the publication of research outputs. Data readiness for AI is not yet consistently evaluated within the production process of scientific knowledge, although various frameworks exist. In addition, data readiness levels for AI differ across domains, geographies and settings. Global collaboration is required to reduce the risks and maximize the benefits of data intensive AI approaches.

AI itself is becoming a contributor, not only to new scientific results built on data, but also to curating scientific data sets and designing the workflows to utilize them for science.

Open data is a core pillar of Open Science, and choices regarding data openness for use in AI should be supported by technical, social and environmental factors.

References

Akhtar, M. et al. 2024. Croissant: A Metadata Format for ML-Ready Datasets. https://doi.org/10.48550/ARXIV.2403.19546.

Alves, A.M. et al. 2024. Deforestation Control at Rural Properties in Pará: A Strategy Using the PrevisIA Risk Model. https://imazon.org.br/publicacoes/deforestation-control-at-rural-propertiens-in-para-a-strategy-using-the-previsia-risk-model/.

Andono, P.N. et al. 2024. Refining Parameter Values in Convolutional Neural Networks Using Weight Modification Particle Swarm Optimization for Enhanced Coral Reef Condition Classification, *International Journal of Intelligent Engineering and Systems*, 17(6), pp. 779–790. https://doi.org/10.22266/ijies2024.1231.59.

Balahur, A. et al. 2022. *Data quality requirements for inclusive, non-biased and trustworthy AI: putting science into standards.* LU: Publications Office (European Commission. Joint Research Centre). https://data.europa.eu/doi/10.2760/365479 (Accessed 14 August 2025).

Belcic, I. 2024. What is RAG (retrieval augmented generation)?, *IBM*, 21 October. https://www.ibm.com/think/topics/retrieval-augmented-generation (Accessed 7 August 2025).

Bentley, D.R. 2000. The Human Genome Project—An Overview, *Medicinal Research Reviews*, 20(3), pp. 189–196. https://doi.org/10.1002/(SICI)1098-1128(200005)20:3<189::AID-MED2>3.0.CO;2-%23.

Berman, G. et al. 2023. The Use of Artificial Intelligence in Science, Technology, Engineering, and Medicine, *The Royal Society* [Preprint]. https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-ai-taxonomy-report.pdf.

Bhatt, N. et al. 2024. A Data-Centric Approach to improve performance of deep learning models, *Scientific Reports*, 14(1), p. 22329. https://doi.org/10.1038/s41598-024-73643-x.

Bianchini, S. et al. 2024. 'Drivers and Barriers of AI Adoption and Use in Scientific Research'. *arXiv*. https://doi.org/10.48550/ARXIV.2312.09843.

Borgman, C.L. 2017. *Big Data, Little Data, No Data: Scholarship in the Networked World*. The MIT Press. https://mitpress.mit.edu/9780262529914/big-data-little-data-no-data/.

Buolamwini, J. and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, *Proceedings of Machine Learning Research* [Preprint]. https://proceedings.mlr.press/v81/buolamwini18a.html.

Burley, S.K. et al. 2017. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive, in A. Wlodawer, Z. Dauter, and M. Jaskolski (eds) *Protein Crystallography*. New York, NY: Springer New York (Methods in Molecular Biology), pp. 627–641. https://doi.org/10.1007/978-1-4939-7000-1_26.

Carroll, S.R. et al. 2022. Using Indigenous Standards to Implement the CARE Principles: Setting Expectations through Tribal Research Codes, *Frontiers in Genetics*, 13, p. 823309. https://doi.org/10.3389/fgene.2022.823309.

Clark, T. et al. 2024. AI-readiness for Biomedical Data: Bridge2AI Recommendations. *Bioinformatics*. https://doi.org/10.1101/2024.10.23.619844.

Clissa, L. et al. 2023. How big is Big Data? A comprehensive survey of data production, storage, and streaming in science and industry, *Frontiers in Big Data*, 6, p. 1271639. https://doi.org/10.3389/fdata.2023.1271639.

Copernicus Data Space Ecosystem. n.d. *Homepage.* https://dataspace.copernicus.eu/.

ESIP Data Readiness Cluster. 2022. Checklist to Examine AI-readiness for Open Environmental Datasets. ESIP, p. 179221 Bytes. https://doi.org/10.6084/M9.FIGSHARE.19983722.V1.

EU Artificial Intelligence Act. n.d. *Article 14: Human Oversight*, *EU Artificial Intelligence Act*. https://artificialintelligenceact.eu/article/14/ (Accessed 11 July 2025).

European Commission. Directorate General for Research and Innovation. and PwC EU Services. (2018) *Cost-benefit analysis for FAIR research data: cost of not having FAIR research data.* LU: Publications Office. https://data.europa.eu/doi/10.2777/02999 (Accessed 14 August 2025).

FARR. n.d. *Resources*, *FAIR in ML, AI Readiness, & Reproducibility Research Coordination Network*. https://www.farr-rcn.org/resources (Accessed 11 July 2025).

Feretzakis, G. and Verykios, V.S. 2024. Trustworthy AI: Securing Sensitive Data in Large Language Models, *AI*, 5(4), pp. 2773–2800. https://doi.org/10.3390/ai5040134.

Finkelstein, A.V. 2018. 50+ Years of Protein Folding, *Biochemistry (Moscow)*, 83(S1), pp. S3–S18. https://doi.org/10.1134/S000629791814002X.

Gao, J. and Wang, D. 2024. Quantifying the use and potential benefits of artificial intelligence in scientific research, *Nature Human Behaviour*, 8(12), pp. 2281–2292. https://doi.org/10.1038/s41562-024-02020-5.

Global AI Ethics and Governance Observatory. n.d. *Readiness Assessment Methodology*, *UNESCO*. https://www.unesco.org/ethics-ai/en/ram (Accessed 11 July 2025).

Gottweis, J. et al. (2025) Towards an AI co-scientist. *arXiv*. https://doi.org/10.48550/ARXIV.2502.18864.

Hagerty, A. and Rubinov, I. 2019. Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence. *arXiv*. https://doi.org/10.48550/ARXIV.1907.07892.

Hardinges, J. et al. 2023. We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models, *Harvard Data Science Review* [Preprint], (Special Issue 5). https://doi.org/10.1162/99608f92.a50ec6e6.

Hiniduma, K. et al. 2024. AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI, in *Proceedings of the 36th International Conference on Scientific and Statistical Database Management*. *SSDBM 2024: 36th International Conference on Scientific and Statistical Database Management*, Rennes, France: ACM, pp. 1–12. https://doi.org/10.1145/3676288.3676296.

Hiniduma, K., Byna, S. and Bez, J.L. 2025. Data Readiness for AI: A 360-Degree Survey, *ACM Computing Surveys*, 57(9), pp. 1–39. https://doi.org/10.1145/3722214.

Holland, S. et al. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *arXiv*. https://doi.org/10.48550/ARXIV.1805.03677.

Howe, B. et al. 2017. Synthetic Data for Social Good. *arXiv*. https://doi.org/10.48550/ARXIV.1710.08874.

Islam, M.N. et al. 2020. A Systematic Review on the Use of AI and ML for Fighting the COVID-19 Pandemic, *IEEE Transactions on Artificial Intelligence*, 1(3), pp. 258–270. https://doi.org/10.1109/TAI.2021.3062771.

Jumper, J. et al. 2021. Highly accurate protein structure prediction with AlphaFold, *Nature*, 596(7873), pp. 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kakko, A. 2025. *Data Heterogeneity in AI: A Deep Dive into the Challenges and Solutions*, *Alphanome.AI*. https://www.alphanome.ai/post/data-heterogeneity-in-ai-a-deep-dive-into-the-challenges-and-solutions.

Kidwai-Khan, F. et al. 2024. A roadmap to artificial intelligence (AI): Methods for designing and building AI ready data to promote fairness, *Journal of Biomedical Informatics*, 154, p. 104654. https://doi.org/10.1016/j.jbi.2024.104654.

Lawrence, N.D. and Montgomery, J. 2024. Accelerating AI for science: open data science for science, *Royal Society Open Science*, 11(8), p. 231130. https://doi.org/10.1098/rsos.231130.

Liu, J. and Cui, P. 2025. Data Heterogeneity Modeling for Trustworthy Machine Learning, in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*. *KDD '25: The 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Toronto ON Canada: ACM, pp. 6086–6095. https://doi.org/10.1145/3711896.3736560.

Liu, R. et al. 2024. 'Best Practices and Lessons Learned on Synthetic Data'. *arXiv*. https://doi.org/10.48550/ARXIV.2404.07503.

Longpre, S. et al. 2023. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. *arXiv*. https://doi.org/10.48550/ARXIV.2310.16787.

Luna, A. 2024. The Open or Closed AI Dilemma, *Bipartisan Policy Center*, 2 May. https://bipartisanpolicy.org/blog/the-open-or-closed-ai-dilemma/ (Accessed 9 June 2025).

Massey, J. and Moriniere, S. 2023. Why we need to be responsible about data and the environment, *Open Data Institute*, 7 September. https://theodi.org/news-and-events/blog/why-we-need-to-be-responsible-about-data-and-the-environment/.

Mondal, D. 2023. Imbalanced data classification: Oversampling and Undersampling, *Medium*, 6 February. https://medium.com/@debspeaks/imbalanced-data-classification-oversampling-and-undersampling-297ba21fbd7c.

Najjar, R. 2023. Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging, *Diagnostics*, 13(17), p. 2760. https://doi.org/10.3390/diagnostics13172760.

Norori, N. et al. 2021. Addressing bias in big data and AI for health care: A call for open science, *Patterns*, 2(10), p. 100347. https://doi.org/10.1016/j.patter.2021.100347.

Ntoutsi, E. et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey, *WIREs Data Mining and Knowledge Discovery*, 10(3), p. e1356. https://doi.org/10.1002/widm.1356.

OECD. 2006. *What Is AI? Can You Make a Clear Distinction between AI and Non-AI Systems?*, *Organisation for Economic Co-operation and Development*. https://oecd.ai/en/wonk/definition (Accessed 11 July 2025).

OECD. 2007. *Data and Metadata Reporting and Presentation Handbook*. Paris: Organisation for Economic Co-operation and Development. https://www.oecd.org/en/publications/data-and-metadata-reporting-and-presentation-handbook_9789264030336-en.html.

Open Data Charter. 2025. Open Data Charter, "Data and AI Development in the Global South: Ethical Practices, Regulations, and Challenges in India, Sri Lanka, and Latin America", *Open Data Charter*, 7 April. https://medium.com/opendatacharter/data-and-ai-development-in-the-global-south-76205d3421d1.

Open Data Institute. 2023. *Data-Centric AI*. https://theodi.org/insights/projects/data-centric-ai/.

ORION Open Science. 2017. What is Open Science? *ORION Open Science*, 27 September. https://www.orion-openscience.eu/resources/open-science.

Oxford Insights. 2025. *Government AI Readiness Index 2024*. https://oxfordinsights.com/ai-readiness/ai-readiness-index/.

Royal Society. n.d. *Science in the age of AI*. https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/.

Sambasivan, N. et al. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI, in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21: CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, pp. 1–15. https://doi.org/10.1145/3411764.3445518.

Samborska, V. 2025. *Scaling up: how increasing inputs has made artificial intelligence more capable*, *Our World in Data*. https://ourworldindata.org/scaling-up-ai.

Schmoldt, A. et al. 1975. Digitoxin metabolism by rat liver microsomes, *Biochemical Pharmacology*, 24(17), pp. 1639–1641. https://pubmed.ncbi.nlm.nih.gov/10/.

SENSCIENCE. n.d. *Home*. https://www.senscience.ai.

Slesinger, I. et al. 2024. Training in Co-Creation as a Methodological Approach to Improve AI Fairness, *Societies*, 14(12), p. 259. https://doi.org/10.3390/soc14120259.

Smith, G. and Rustagi, I. 2021. When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity, *Stanford Social Innovation Review*, 31 March. https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity (Accessed 11 July 2025).

Sorscher, B. et al. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. *arXiv*. https://doi.org/10.48550/ARXIV.2206.14486.

Stocker, M. et al. 2025. Rethinking the production and publication of machine-readable expressions of research findings, *Scientific Data*, 12(1), p. 677. https://doi.org/10.1038/s41597-025-04905-0.

Sundaram, S.S. et al. 2024. Use of a Structured Knowledge Base Enhances Metadata Curation by Large Language Models. *arXiv*. https://doi.org/10.48550/ARXIV.2404.05893.

Tabbakh, A. et al. 2024. Towards sustainable AI: a comprehensive framework for Green AI, *Discover Sustainability*, 5(1), p. 408. https://doi.org/10.1007/s43621-024-00641-4.

Tai, C.A. et al. 2024. Cancer-Net SCa-Synth: An Open Access Synthetically Generated 2D Skin Lesion Dataset for Skin Cancer Classification. *arXiv*. https://doi.org/10.48550/arXiv.2411.05269.

The European Open Science Cloud research data commons. n.d. *Services for inter- and cross-disciplinary data discovery, access, sharing and reuse in the EOSC Federation*, *Horizon Europe*. https://doi.org/10.3030/101188179.

Umbach, G. 2024. Open Science and the impact of Open Access, Open Data, and FAIR publishing principles on data-driven academic research: Towards ever more transparent, accessible, and reproducible academic output?, *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, 40(1), pp. 59–70. https://doi.org/10.3233/SJI-240021.

UNESCO 2021. *UNESCO Recommendation on Open Science*. UNESCO. https://doi.org/10.54677/MNMH8546.

Usman, S. et al. 2022. Data Locality in High Performance Computing, Big Data, and Converged Systems: An Analysis of the Cutting Edge and a Future System Architecture, *Electronics*, 12(1), p. 53. https://doi.org/10.3390/electronics12010053.

Van Noorden, R. and Perkel, J.M. 2023. AI and science: what 1,600 researchers think, *Nature*, 621(7980), pp. 672–675. https://doi.org/10.1038/d41586-023-02980-0.

Verhulst, S. et al. 2025. Moving Toward the FAIR-R principles: Advancing AI-Ready Data. SSRN. https://doi.org/10.2139/ssrn.5164337.

Von Eschenbach, W.J. 2021. Transparency and the Black Box Problem: Why We Do Not Trust AI, *Philosophy & Technology*, 34(4), pp. 1607–1622. https://doi.org/10.1007/s13347-021-00477-0.

Vysogorets, A. et al. 2025. DRoP: Distributionally Robust Data Pruning. *arXiv*. https://doi.org/10.48550/ARXIV.2404.05579.

Wilkinson, M.D. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data*, 3(1), p. 160018. https://doi.org/10.1038/sdata.2016.18.

Wong, M.R. et al. 2024. Uplifting FAIR and CARE across Earth and Environmental Science (E&ES) Data. A Discussion Paper to inform the Data Targeted Discussion of the National Digital Research Infrastructure Strategy. *Australian Research Data Commons.* https://doi.org/10.5281/ZENODO.14241825.

York, D.G. et al. 2000. The Sloan Digital Sky Survey: Technical Summary, *The Astronomical Journal*, 120(3), pp. 1579–1587. https://doi.org/10.1086/301513.

Appendix 1: Glossary

For other concepts related to data in science, see CODATA's terminology for Research Data Management.

AI An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.

Source: OECD. 2006. *What Is AI? Can You Make a Clear Distinction between AI and Non-AI Systems?, Organisation for Economic Co-operation and Development*. https://oecd.ai/en/wonk/definition (Accessed 11 July 2025).

Metadata Data that defines and describes other data with purposes that may include finding data, recording its provenance, describing its structure for facilitating analysis.

Source: OECD. 2007. *Data and Metadata Reporting and Presentation Handbook*. Paris: Organisation for Economic Co-operation and Development. https://www.oecd.org/en/publications/data-and-metadata-reporting-and-presentation-handbook_9789264030336-en.html.

Data readiness for AI Data readiness for AI involves ensuring that datasets are properly collected, cleaned and structured before being used to train models.

Source: Clark, T. et al. 2024. AI-readiness for Biomedical Data: Bridge2AI Recommendations. *Bioinformatics*. https://doi.org/10.1101/2024.10.23.619844.

Data lifecycle The data lifecycle encompasses the different stages that data goes through, from data generation to data interpretation.

Source: Wing, J. M. 2019. The Data Life Cycle. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.e26845b4 (Accessed 12 July 2025).

High-quality data High-quality data is accurate, complete, verifiable and contains the data needed to answer a specific question or set of questions.

Source: National Center for Advancing Translational Sciences. *Data Quality.* https://toolkit.ncats.nih.gov/glossary/data-quality/ (Accessed 12 July 2025).

Scientific data Data generated by experiments and observations as part of the scientific process. The definition encompasses raw data, processed data and higher-level data products across and includes data in any format, for example, numerical, text, images, sound, etc. For this work, we assume that data is in a digitized, machine-readable format.

Open Science Open Science is the movement to make scientific research, data and their dissemination available to any member of an inquiring society, from professionals to citizens.

Source: ORION Open Science. 2017. What is Open Science?, *ORION Open Science*, 27 September. https://www.orion-openscience.eu/resources/open-science.

Knowledge graphs A representation of entities and the relationships between them, preserving meaning.

Large language model An AI model that has been trained on a large amount of text for the purpose of language processing tasks, often of a generative nature.

Data augmentation A suite of techniques that can be used to increase the size or enhance the quality of a dataset for training.

Source: Shorten, C. and Khoshgoftaar, T.M. 2019. A survey on Image Data Augmentation for Deep Learning, *Journal of Big Data*, 6(1), p. 60. https://doi.org/10.1186/s40537-019-0197-0.

Work with the ISC to advance science as a global public good.

About the International Science Council

The International Science Council (ISC) works at the global level to catalyse change by convening scientific expertise, advice and influence on issues of major importance to both science and society.

The ISC is a non-governmental organization with a unique global membership that brings together more than 250 international scientific unions and associations, national and regional scientific organizations including academies and research councils, international federations and societies, and young academies and associations.

Connect with us at:

council.science

secretariat@council.science

International Science Council

5 rue Auguste Vacquerie

75116 Paris, France

bsky.app/profile/sciencecouncil.bsky.social

threads.net/@council.science

facebook.com/InternationalScience

https://www.linkedin.com/company/international-science-council

instagram.com/council.science