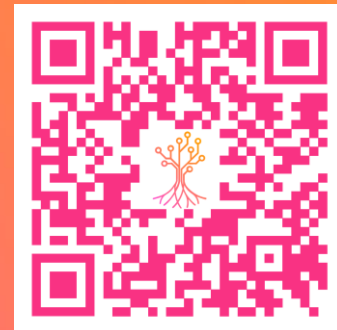


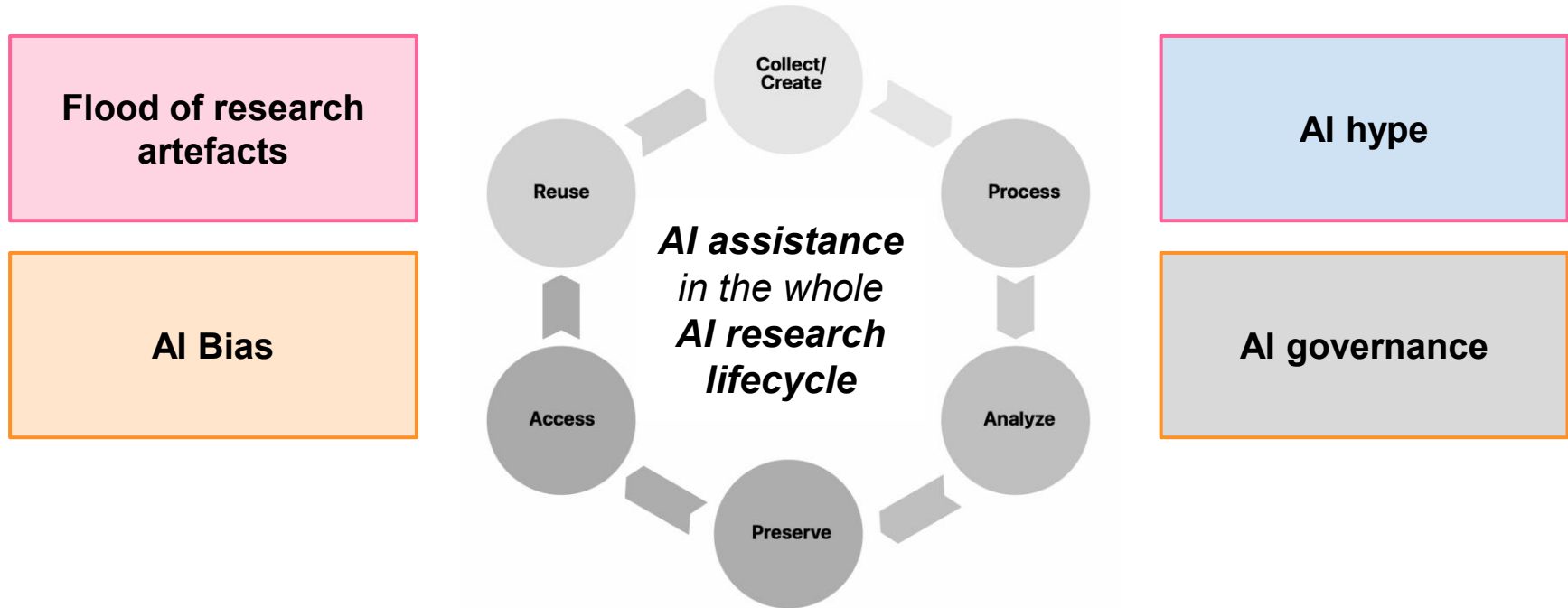
**On reproducibility
(and provenance) activities at NFDI4DS**

**Christine Hennig
Fraunhofer FOKUS**

19.05.2026



(Some) Challenges in Data Science and Artificial Intelligence



An AI task is a specific function for which an AI system was created.

We intensively use “shared (AI) tasks” to solve actual challenges.

- *Develop, establish and sustain a national **research data infrastructure***
- *Offer innovative **tools** and **services***
- *Make **all digital artefacts** available*
 - **Workflows**
 - **Models**
 - **Data**
 - **scripts/code**
 - **Articles**
- ***Interlink** all digital artefacts*
- *Foster **FAIR** (Findable, Accessible, Interoperable, Reusable or **Fully AI Ready**) and **open** research*
- *Increase **transparency** and **reproducibility***
- *Consider **ethical**, **legal** and **social** aspects*

1. Workflows

- A **data science workflow** defines the phases (or steps) in a data science project. Using a well-defined data science workflow is useful in that it provides a simple way to remind all data science team members of the work to be done to do a data science project.
- Typical **workflow schema**
 1. Problem Definition (*match the ML type*)
 2. **Data** Collection (*what type of data?*)
 3. Success Criteria (Initial **Evaluation**) (*which evaluation method?*)
 4. Features (*feature variables predict target variables*)
 5. Model Building (*choosing and training data, model tuning, model comparison, data splitting*)
 6. **Model** Evaluation (*use the metric from 3*)
 7. Experiment (*iterate previous steps*)
 8. Deployment (Optional) (*usable product or service*)

2. Models

- \ A **large language model** (LLM) is a neural network trained on a vast amount of text for **natural language processing** tasks, especially **language generation**

- \ **Model cards** are files that accompany the models and provide information:
 - the model
 - its intended uses & potential limitations, including biases and ethical considerations
 - the training params and experimental info
 - which datasets were used to train your model
 - the model's evaluation results

*Model cards are making visible the dataset **strength**, **weaknesses**, **biases** ...*

3. Data

Benchmark Dataset

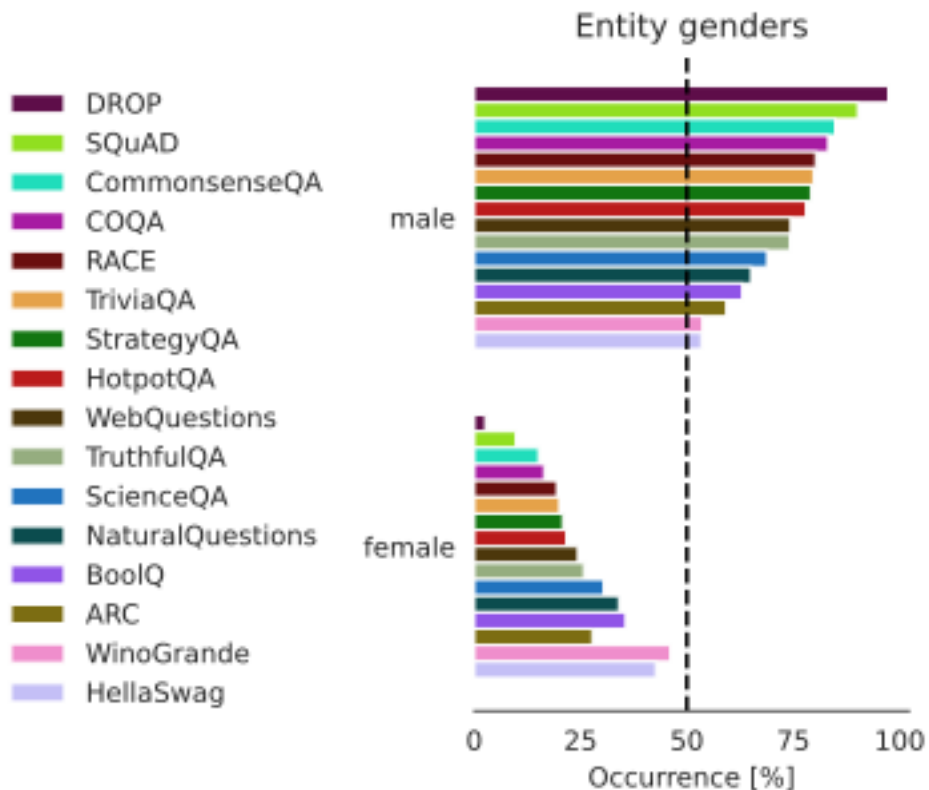
- \ is a **standardized**
- \ **high-quality**
- \ collection of data
- \ designed to **evaluate** the **performance** of **machine learning (ML) models**
- \ in a fair, **reproducible**, and objective manner.

A **benchmark dataset**

- \ serves as a public "measuring stick"
- \ for the research and development community

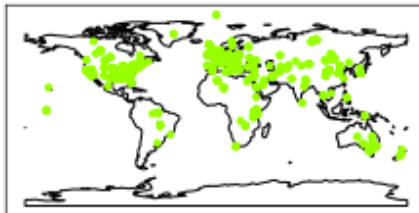
*If a benchmark dataset is biased → we **reproduce the bias***

AI benchmarking: Benchmark dataset bias: gender

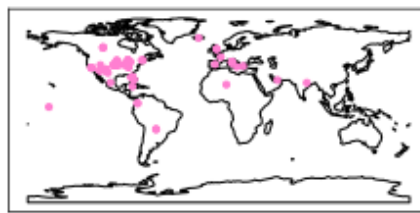


AI benchmarking: Benchmark dataset bias: **region**

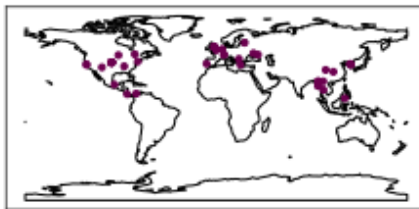
SQuAD



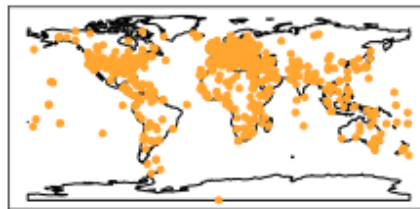
WinoGrande



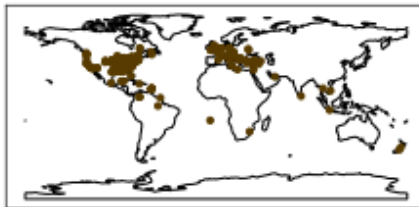
DROP



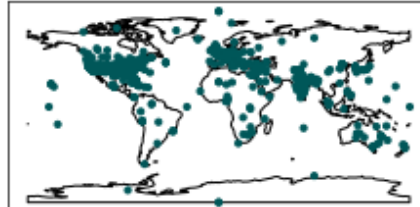
TriviaQA



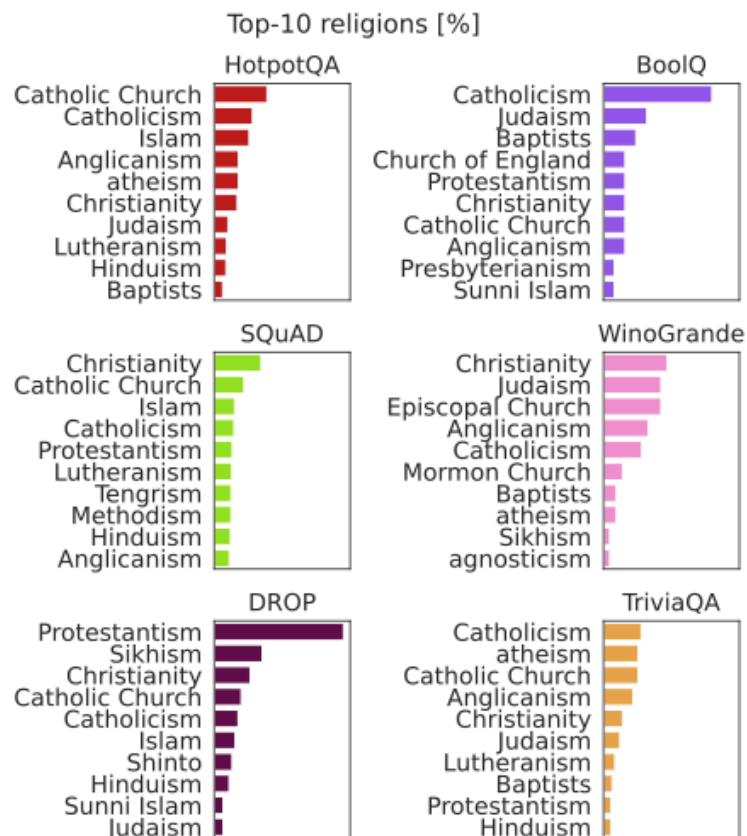
WebQuestions



NaturalQuestions



AI benchmarking: Benchmark dataset bias: religion



4. Code / Scripts

- Code **repository** with **version control** (e.g. git –widely adopted)
- **Describe** the code (e.g. cff)
 - **Readme** file
 - **CITATION.cff** files are plain text files with human- and machine-readable citation information for software (and datasets). Code developers can include them in their repositories to let others know how to correctly **cite their software**.
- **Document** the code
- Reproducible: Runnable code (e.g. containerization)
 - **Containerization** is the packaging of software code to create a single lightweight executable—called a container—that runs consistently on any infrastructure
- **Shared task:** *GitHub ReadMe to Knowledge Graph*

Numbers are not facts

When are 2 numbers the same?

- Rounding
- Error bars
- Storing numbers in a computer → resolution
- Calculate with numbers in a computer → precision
- *Measurement accuracy*
- *Measuring devices: calibration, environment, ...*

Algorithms that are not reproducible in itself

- **Non deterministic systems**

5. Publications / scholarly data

- **Metadata** enhancing of scholarly publications using **AI**
 - Shared tasks:
 - *Information Extraction from Scientific **Figures** in Materials Science*
 - *Scientific **Visual** Question Answering*
 - ***Software Mention Detection in Scholarly Publications***
 - *LLM-based **Subject** Tagging*
 - *Field of Research Classification*
 - ***Metadata** Extraction from Scholarly Documents*
 - *Tracking the **State-of-the-Art** in AI Scholarly Publications*

Putting it all together

- **Metadata** machine actionable
 - **RO-Crates** to package model + data + code
 - **FAIR Digital Objects** for findable & reusable training data, reproducible workflows
 - **Metadata Standards**
- (Research) **knowledge graphs**: represent (meta)data and operate on the data
 - Knowledge representation and reasoning
 - **Shared tasks**:
 - *Large Language Models for Ontology Learning*
 - *Knowledge Graph Question Answering*
 - *Hybrid Question Answering over Scholarly Knowledge Graphs and Text*
- **Neuro-Symbolic** methods: combine statistical, neural and symbolic methods

Why Reproducibility?

- \ A scientist **wants** to ... discover the **truth**
- \ A scientist **wants** to ... **understand** the world
- \ A scientist **needs** to ... **trust** (most of) the existing knowledge

Links and Literature

<https://www.nfdi4datascience.de/>

<https://www.datascience-pm.com/data-science-workflow/>

https://en.wikipedia.org/wiki/Large_language_model

<https://huggingface.co/docs/hub/model-cards>

<https://www.ultralytics.com/glossary/benchmark-dataset>

<https://www.ibm.com/think/topics/containerization>

<https://dev.to/dinakajoy/a-beginners-guide-to-the-data-science-workflow-4772>

<https://citation-file-format.github.io/>