ITU, WHO & WIPO Global Initiative on AI for Health (GI-AI4H): Purpose, governance, working & topic groups

Focused deep-dive:

Data standards for health AI: Benchmarking, metadata and federated data discovery

Overview

ITU/WHO/WIPO Global Initiative on AI for Health (GI-AI4H)

Simão Campos
ITU

AI for Good Webinar | 27 March 2026

What is the GI-AI4H?

Joint UN initiative led by ITU, WHO and WIPO to harness AI for the benefit of health.

Global coordination platform for standards, governance, knowledge sharing and implementation support.

Builds on the legacy and deliverables of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

Strategic pillars

Enable

- Development of international standards, norms, guidance and evidence-based technical products.

Facilitate

- Convening stakeholders and operating enabling mechanisms (e.g. Open Code Infrastructure).

Implement

- Supporting country-level scaling, capacity building and sustainable implementation models.

Governance structure

Steering Committee
Senior representatives from ITU, WHO and WIPO.

Joint Secretariat
Operational coordination across the three organizations.

Working Groups, Topic Groups & Facilitation Groups

Partners and donors
Supporting activities and implementation.

How the work is organized

Working Groups (WGs)
Cross-cutting themes such as ethics, regulation, data, evaluation and intellectual property.

Topic Groups (TGs)
Specific health domains (e.g. traditional medicine, maternal health).

Facilitation groups
Stakeholder networks supporting knowledge exchange and implementation.

Current Working & Topic Groups (snapshot)

Working Groups

- WG – Regulatory considerations
- WG – Ethics and governance
- WG – Data
- WG – Evaluation
- WG – Intellectual property & innovation

Topic Groups (tentative)

- Traditional medicine √
- Maternal & reproductive health
- Point-of-care / primary health care
- Oral health (currently on hold)

WG-DATA: scope and rationale

Addresses fragmentation and high transaction costs in health-AI data usage.

Develops a Data and Model Exchange Protocol (DMXP) to support discoverability and access.

Builds on FG-AI4H benchmarking experience and modern data standards (e.g., Croissant).

WG-DATA objectives

G1: Dataset metadata standardization and basic matchmaking.

G2: Searchable, ontology-linked index of health datasets.

G3: Secure transaction protocols
(authentication, licensing, access control).

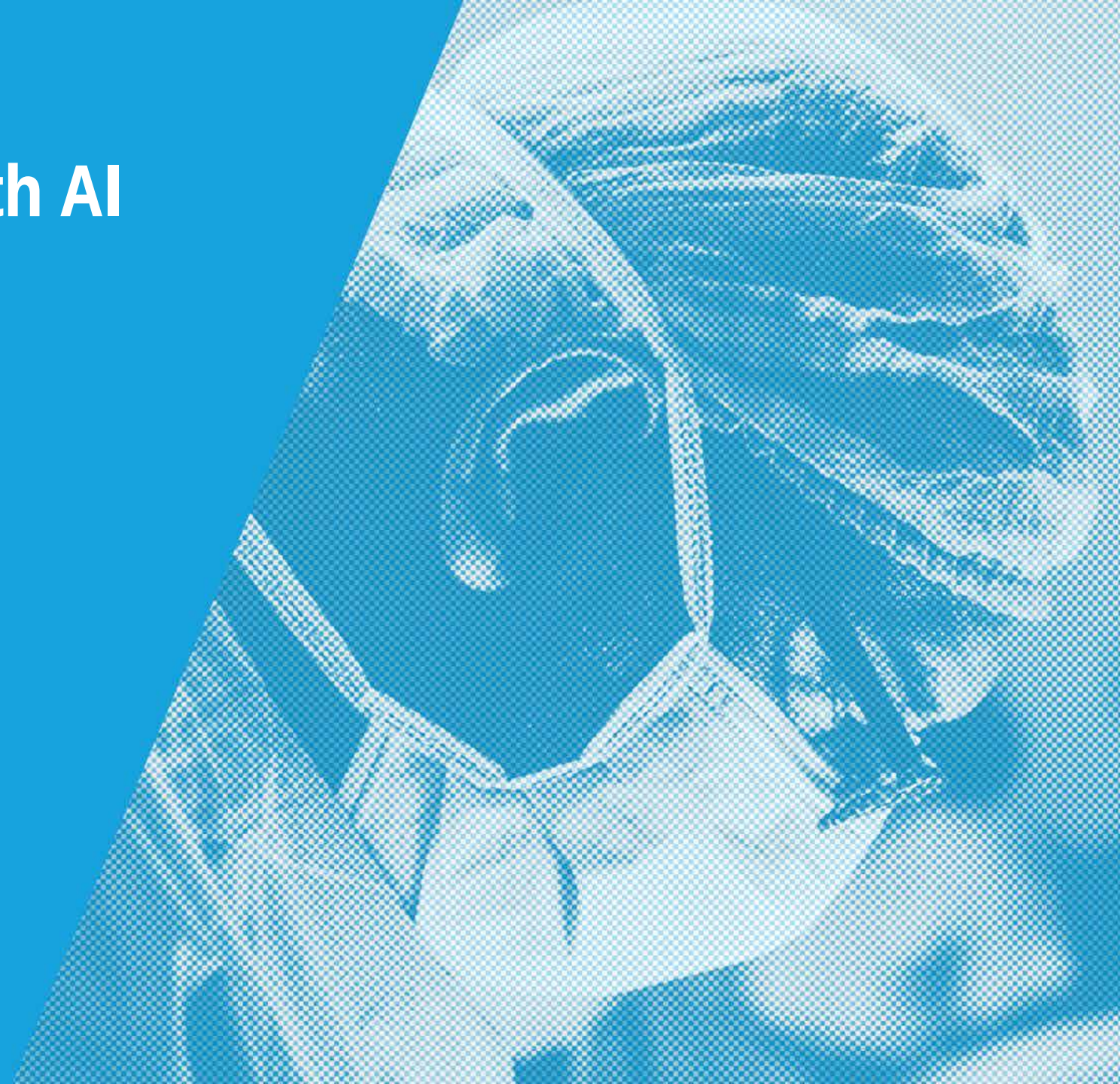
G4: Federated testing environments and real-world pilots.

Data Standards for Health AI

Benchmarking, Metadata and Federated Data Discovery

Marc Lecoultre
Open Code Infrastructure | GI-AI4H
ITU-WHO-WIPO Global Initiative on AI for Health

AI for Good Webinar | 27 March 2026



The Data Challenge in Health AI

- Health data is siloed across countries, institutions, and systems
- No consistent way to describe, discover, or compare datasets
- Every dataset has its own format, documentation, and access model
- Regulatory requirements vary by jurisdiction

Impact on Health AI

- AI developers cannot find suitable training and benchmarking data
- Regulators cannot verify what data was used to build a model
- Research is duplicated — datasets are not discoverable
- Benchmarking is inconsistent — models tested on incomparable data
- Health equity gaps widen when underrepresented data stays invisible

We need a standardized, federated approach to health dataset discovery

The Open Code Infrastructure: A Public Good

The AI4H Assessment Platform is a free, open-access platform built as a public good, providing a secure, end-to-end framework for developing and rigorously assessing health AI algorithms globally.

Open Source

Transparent, auditable,

community-driven

Universal

Across borders,

stakeholders, disciplines

Health-Specific

Privacy, ethics, consent,

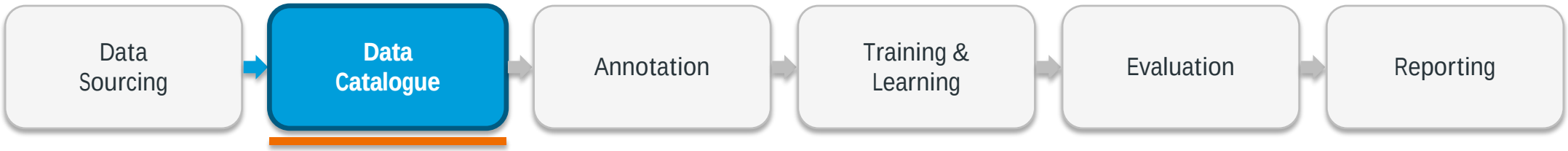
regulatory alignment

- 10 module drivers, 40+ contributors from 5 continents
- Engineers, academics, medical doctors, regulators
- Integrates data privacy, ethics, governance, and secure metadata management

Data Catalogue

Discovering Health Datasets
for AI Benchmarking

End-to-End Platform: Focus on Data Catalogue



The Data Catalogue is the backbone of health AI benchmarking

- Standardized dataset descriptions using Croissant metadata
- Browse and search datasets by modality, disease, population
- Machine-readable metadata enables automated quality checks
- Federated architecture: data stays at source, metadata travels
- Supports regulatory traceability and audit requirements
- Interoperable with Kaggle, Hugging Face, OpenML, Google Dataset Search

Why Data Standards Matter for Health AI

Without Standards

- ✕ Dataset described in a README — "100 chest X-rays, contact PI for access"
- ✕ No consistent vocabulary across institutions or countries
- ✕ Manual effort to understand format, structure, and licensing
- ✕ Cannot compare datasets used in different benchmarks
- ✕ Regulators have no audit trail for data provenance
- ✕ Discovery relies on personal networks and word-of-mouth

With Croissant

- ✓ Machine-readable JSON-LD metadata indexed by search engines
- ✓ Shared vocabulary linked to biomedical ontologies (SNOMED, ICD, LOINC)
- ✓ Self-describing structure: fields, types, splits, and ML semantics
- ✓ Datasets are comparable across challenges and institutions
- ✓ Full provenance: collection method, annotation, consent, governance
- ✓ Global discovery via federated catalogue and standard APIs

Croissant: The Metadata Standard for ML Datasets

Community built and open standard by MLCommons • Built on schema.org • JSON-LD format

1. **Metadata** — Dataset description, license, creators, responsible AI attributes
2. **Resources** — Pointers to data files, URLs, checksums for integrity
3. **Structure** — Records, fields, data types — how raw data maps to tables
4. **ML Semantics** — Train/test/validation splits, intended tasks and use cases

Adopted by: Google Dataset Search • Kaggle • Hugging Face • OpenML • CKAN • TensorFlow Datasets

WG "BioCroissant" - Extending Croissant for Health & Life Sciences

The life sciences extension of Croissant (v1.1) — domain-specific vocabulary for biomedical datasets

Biomedical Ontologies

Links to SNOMED CT, ICD, LOINC, MeSH, UMLS

Clinical Context

Imaging modality, anatomical region, pathology, cohort

Patient Metadata

Demographics, consent status, cohort characteristics

Regulatory Attributes

IRB approval, data governance, privacy level, jurisdiction

Provenance

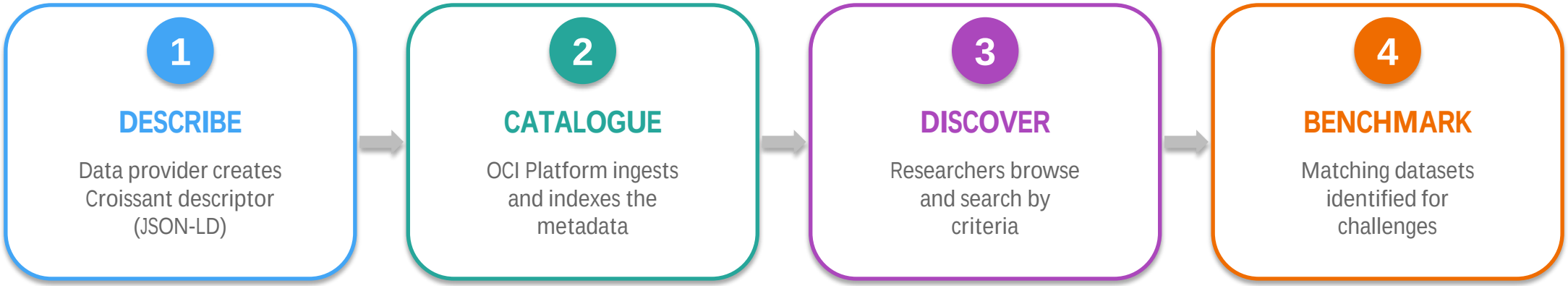
Collection protocol, annotation methodology, ground truth

Cross-border Comparison

Shared vocabulary enables global benchmarking

A chest X-ray dataset from Kenya and one from Germany can be described, compared, and discovered using the same standardized vocabulary

How It Works: Dataset Cataloguing with Croissant

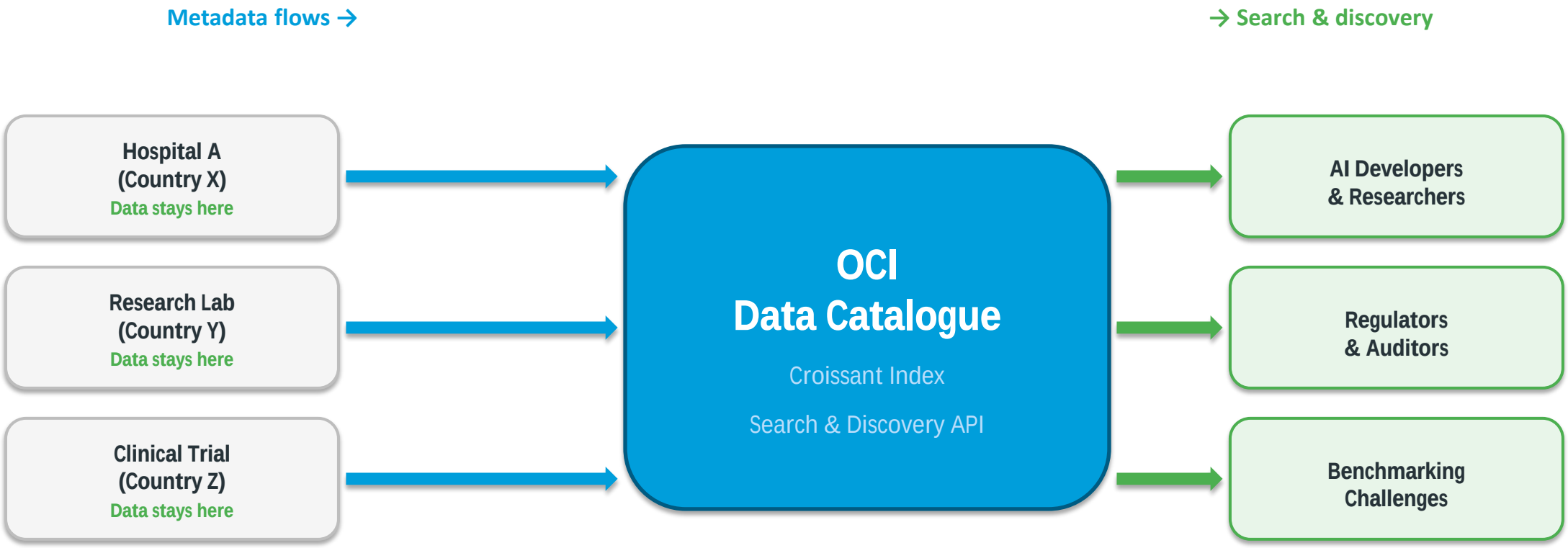


Key Principle: Data stays at source — only metadata travels

Federated catalogue respects data sovereignty while enabling global discovery

Browse by: disease area • imaging modality • population • geography • consent status • annotation quality

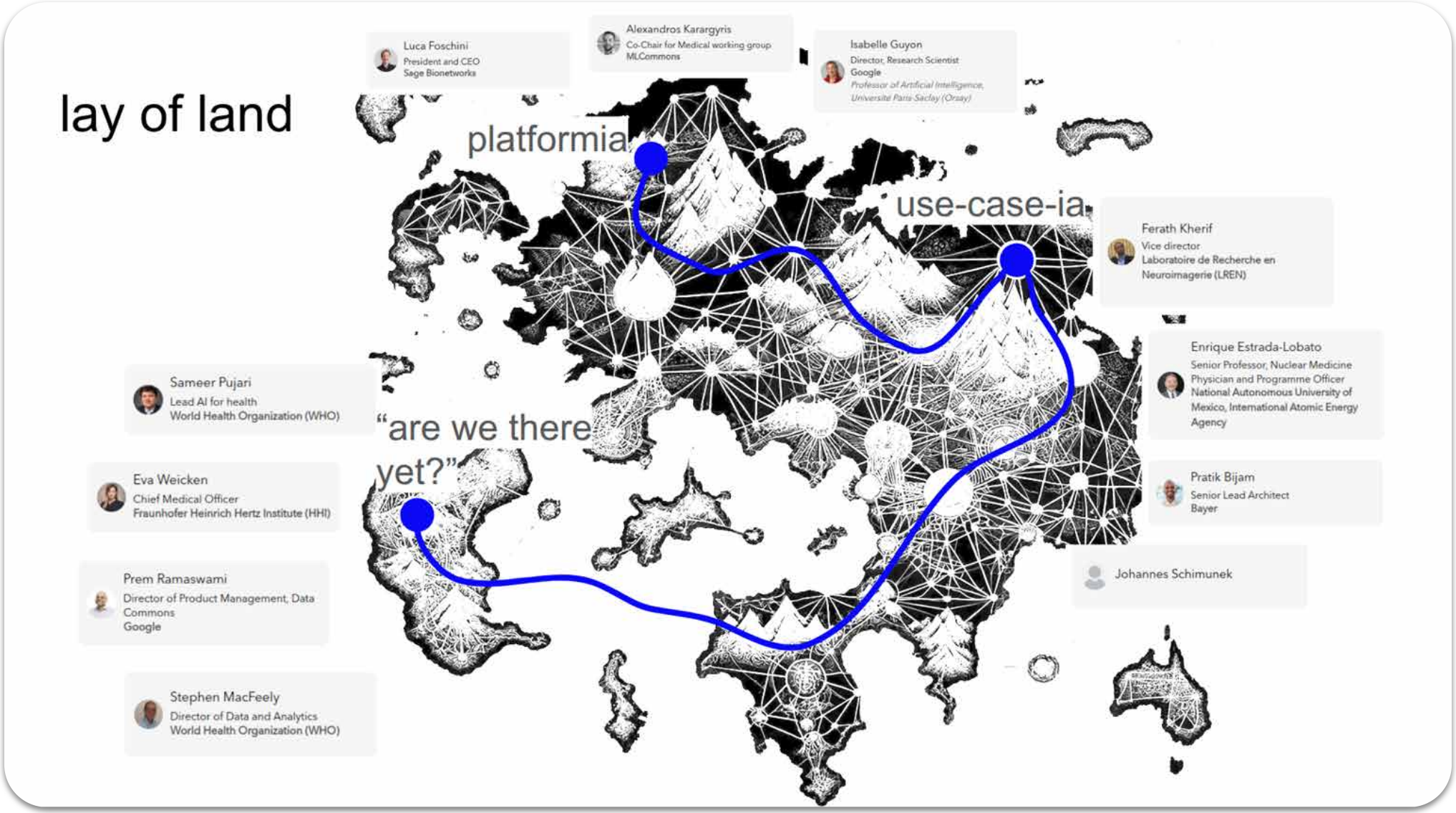
Federated Data Discovery Architecture



Data Multiplexers

Infrastructure for Open Data Marketplaces and Exchanges

Luis Oala
ITU-WHO-WIPO Global Initiative on AI for Health
Brickroad



Throwback: Geneva, 31 May 2024

DATA-CENTRIC CONCEPTS OVER TIME

In 1990 and 1992, Wilson, Garris and Wilkinson from NIST release data that will later become known as MNIST. To date, the papers have <150 citations

Early works by Cohn, MacKay, Schmidthuber and others in active learning pave the way for data-centric concepts

First mentions of "data-centric" in computer science papers such as Breitbart et al. (1993)

Seminal data sheets paper by Gebru et al. (2018)

Landmark ML datasets such as ImageNet and MNIST are reevaluated by its creators (Yang et al., 2019; Yadav and Bottou, 2019)

Andrew Ng and others launch datacentricai.org

NeurIPS D&B track launches for first time

First DMLR workshop at ICML to unify community

EleutherAI, LAION, and other groups compile and release new generations of popular ML datasets

Data cascades paper by Sambavisan et al. (2021)

Data-centric AI workshop at NeurIPS 2021

Nurturing the DMLR ecosystem to build the infrastructure and community for next gen public datasets

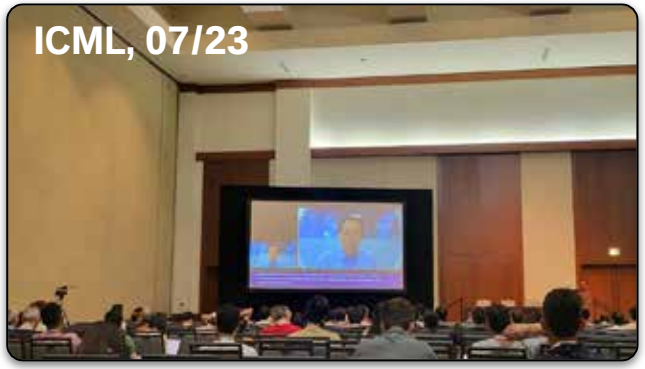
DMLR Journal launches in the JMLR family

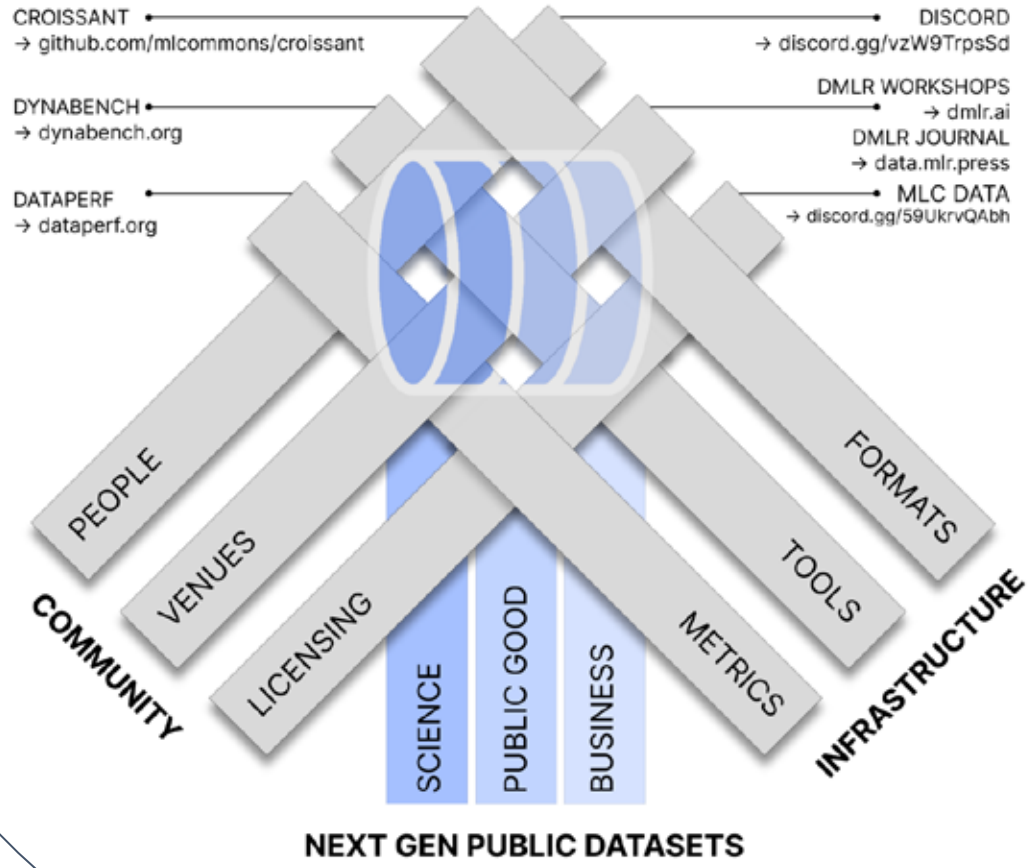
1990s and 2000s
Systems and active learning

2010s
Data scrutiny

Early 2020s
Data-centric ML

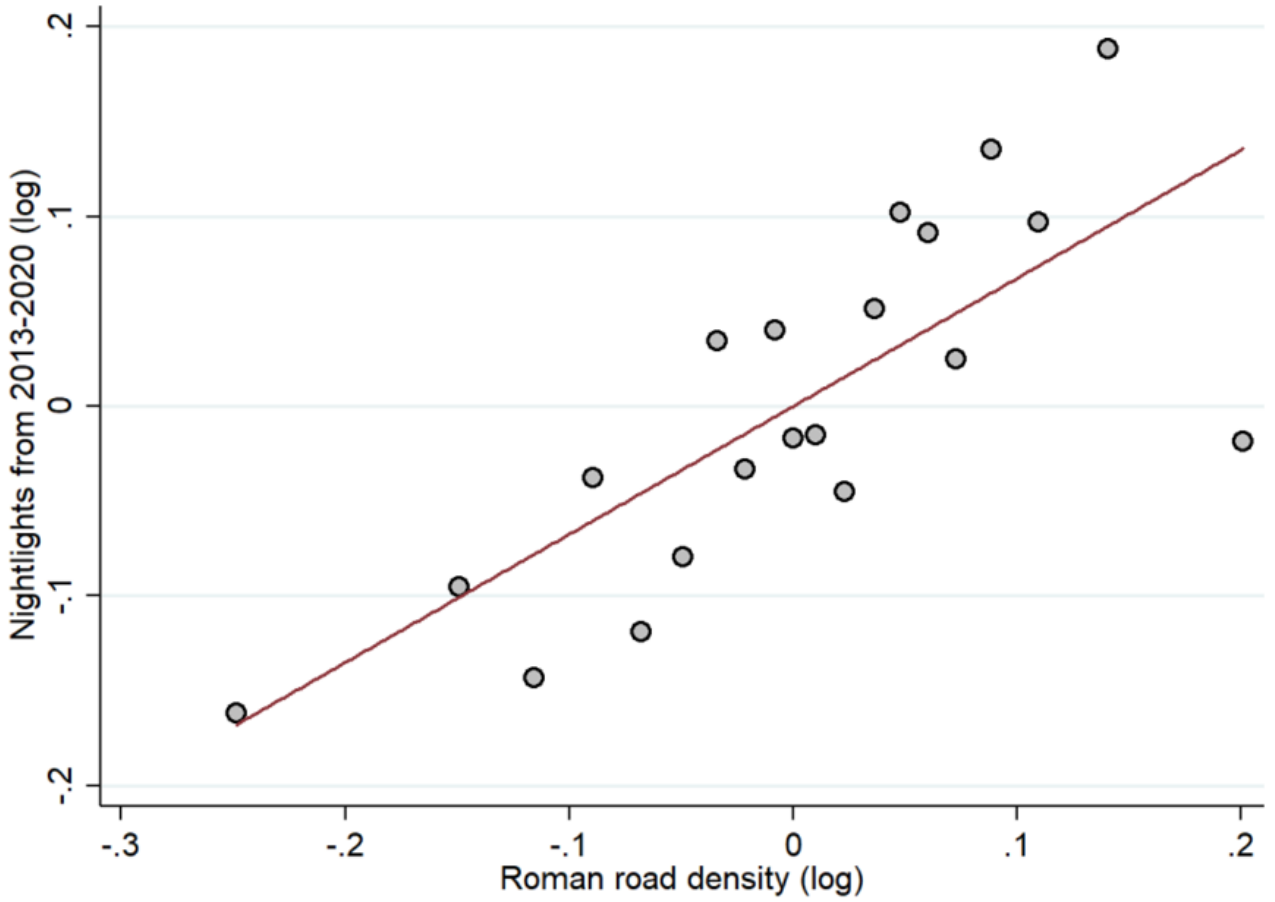
Future
The DMLR ecosystem





Oala, Luis, et al. "DMLR: Data-centric Machine Learning Research-Past, Present and Future." Journal of Data-centric Machine Learning Research (2024).

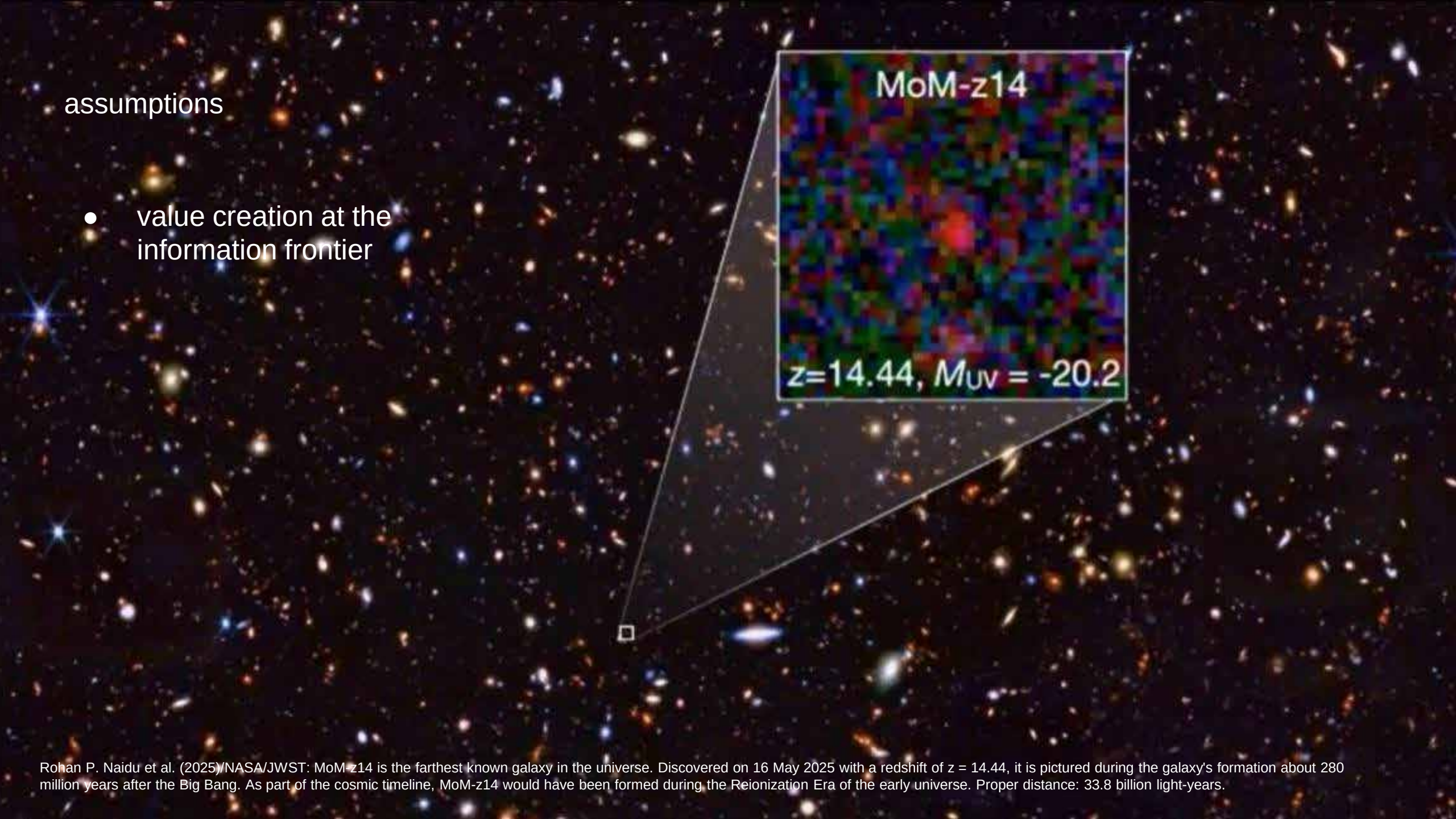
Assumptions



Dalgaard, Carl-Johan, et al. "Roman roads to prosperity: Persistence and non-persistence of public infrastructure." Journal of Comparative Economics 50.4 (2022): 896-916.

assumptions

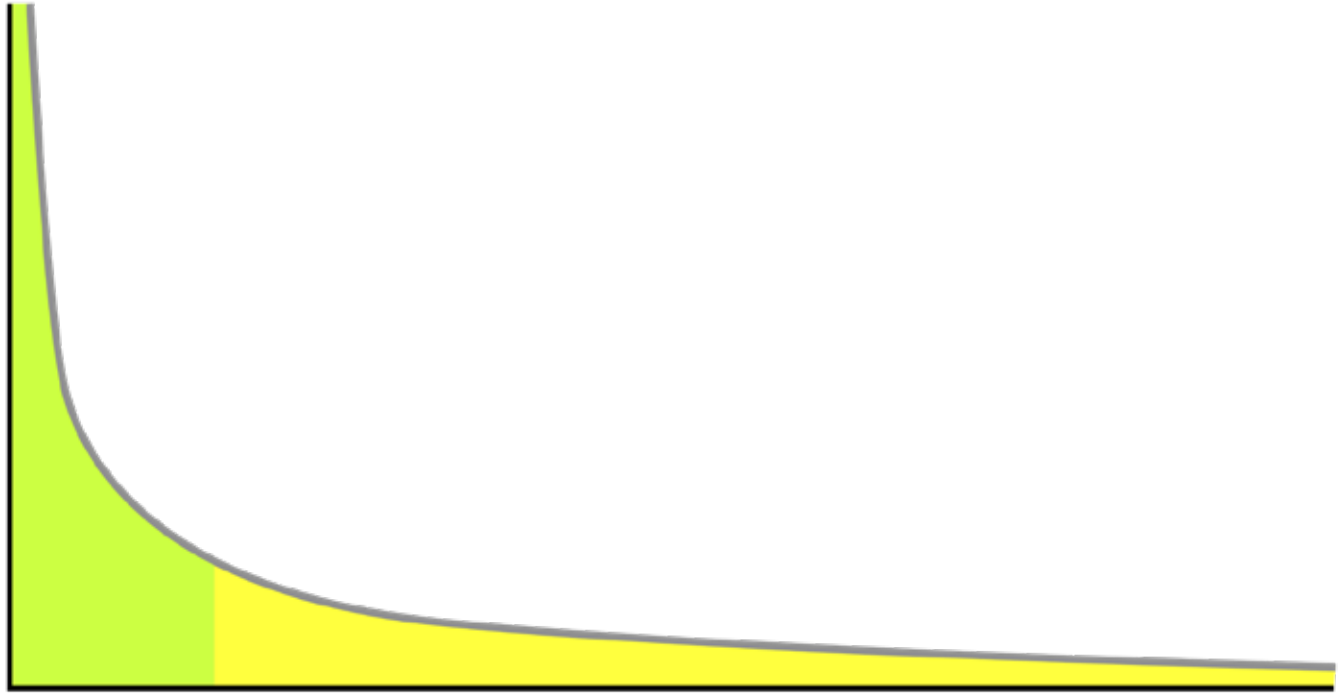
- value creation at the information frontier



Rohan P. Naidu et al. (2025)/NASA/JWST: MoM-z14 is the farthest known galaxy in the universe. Discovered on 16 May 2025 with a redshift of z = 14.44, it is pictured during the galaxy's formation about 280 million years after the Big Bang. As part of the cosmic timeline, MoM-z14 would have been formed during the Reionization Era of the early universe. Proper distance: 33.8 billion light-years.

Assumptions

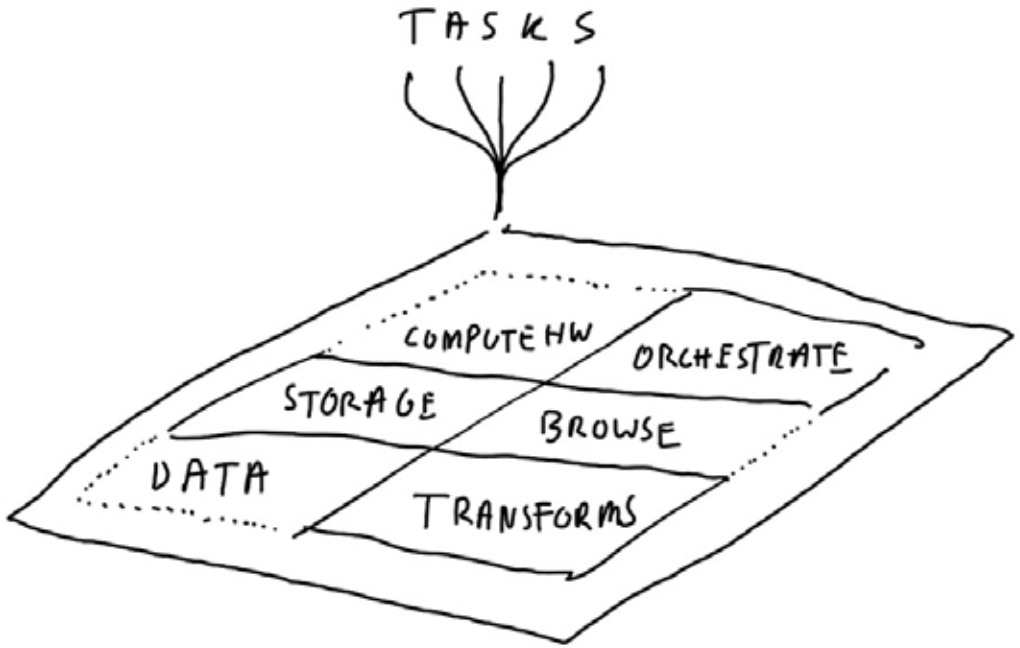
- value creation at the information frontier
- long tail of tasks

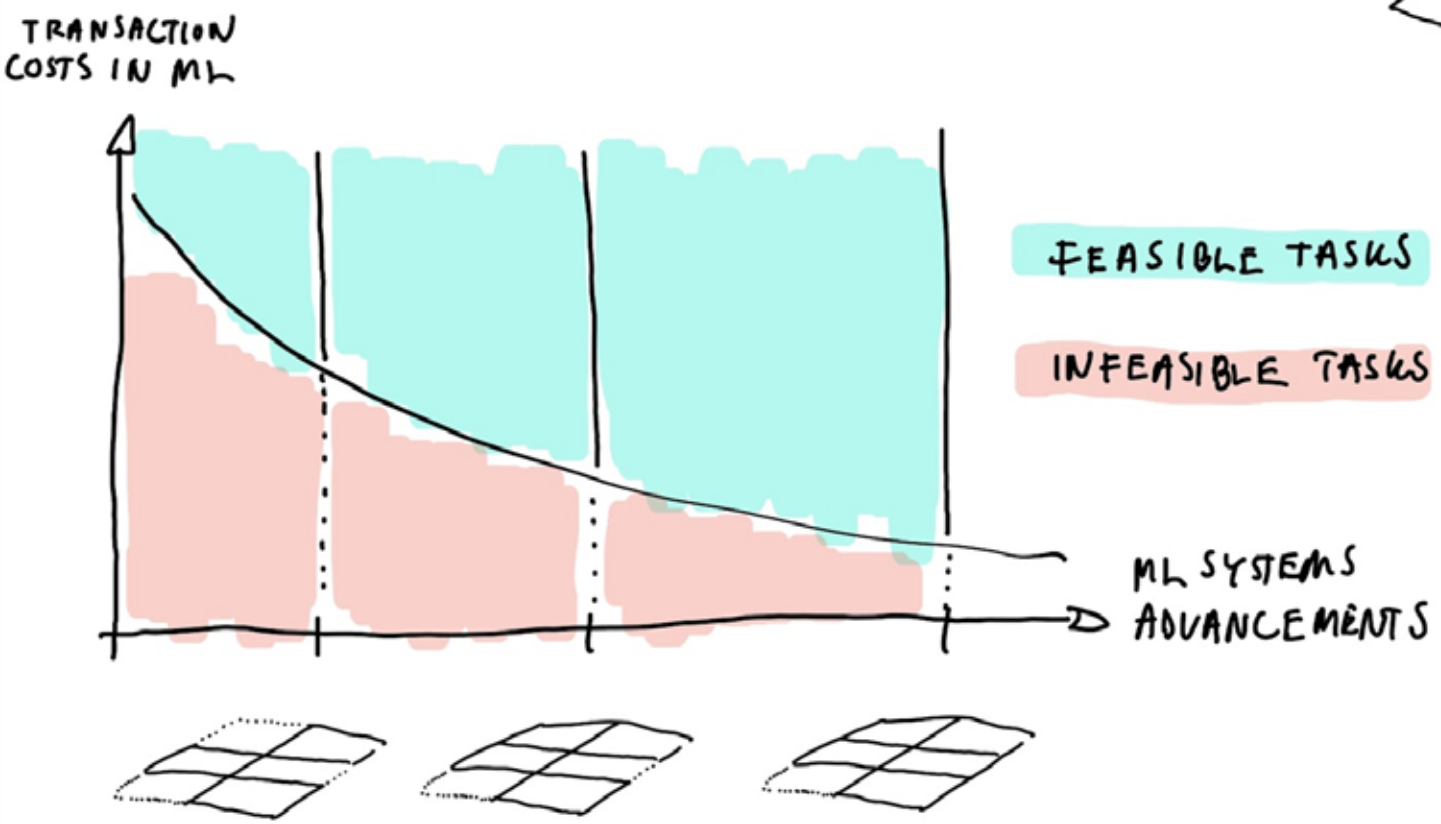


Picture by Hay Kranen

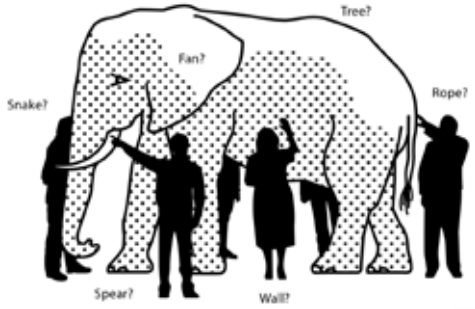
Goals

Increase the space of feasible tasks



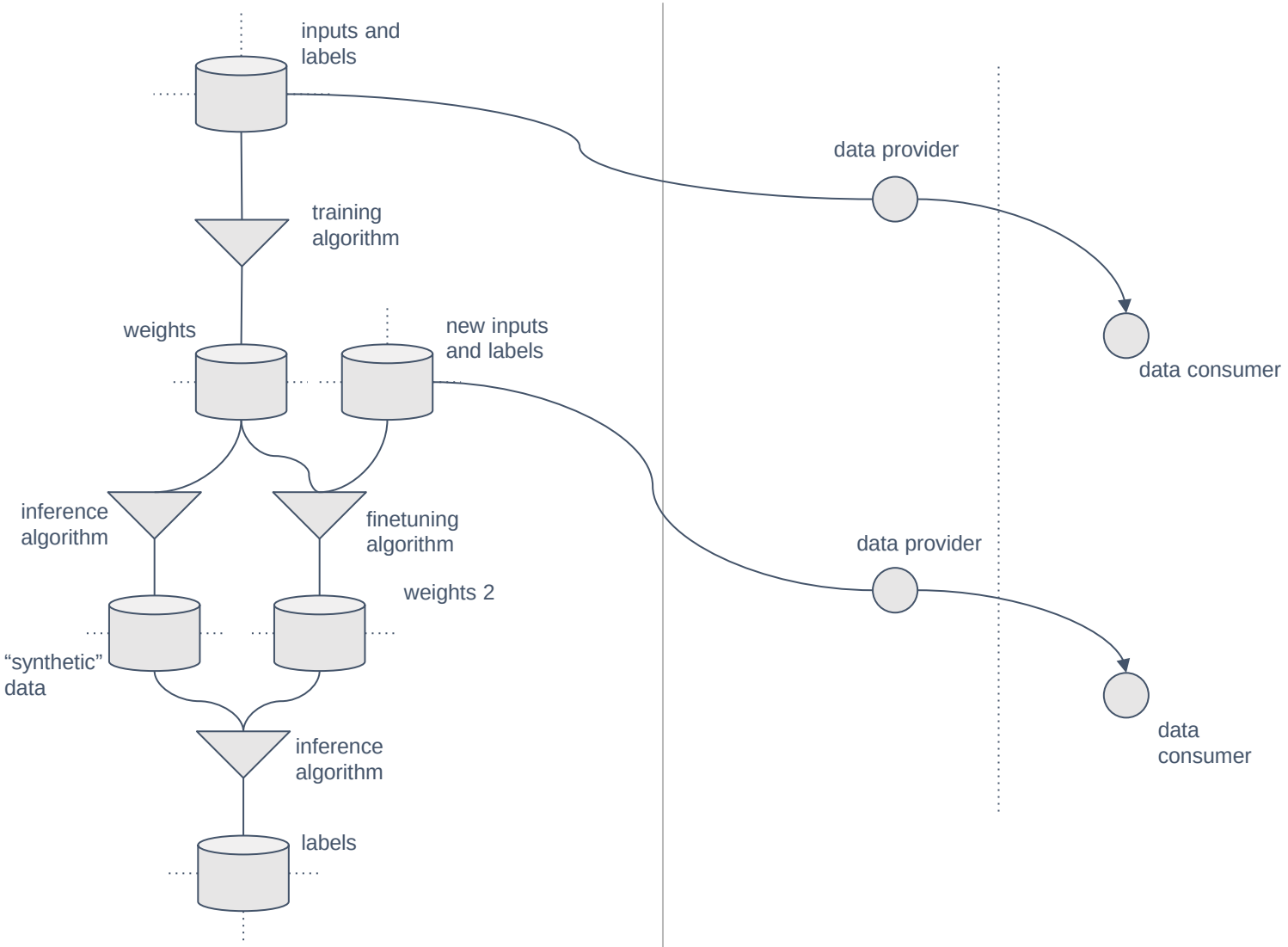


Challenges



https://blogs.sas.com/content/datamanagement/2016/09/27/sas-odpi-interoperability-hadoop/

Data value chain



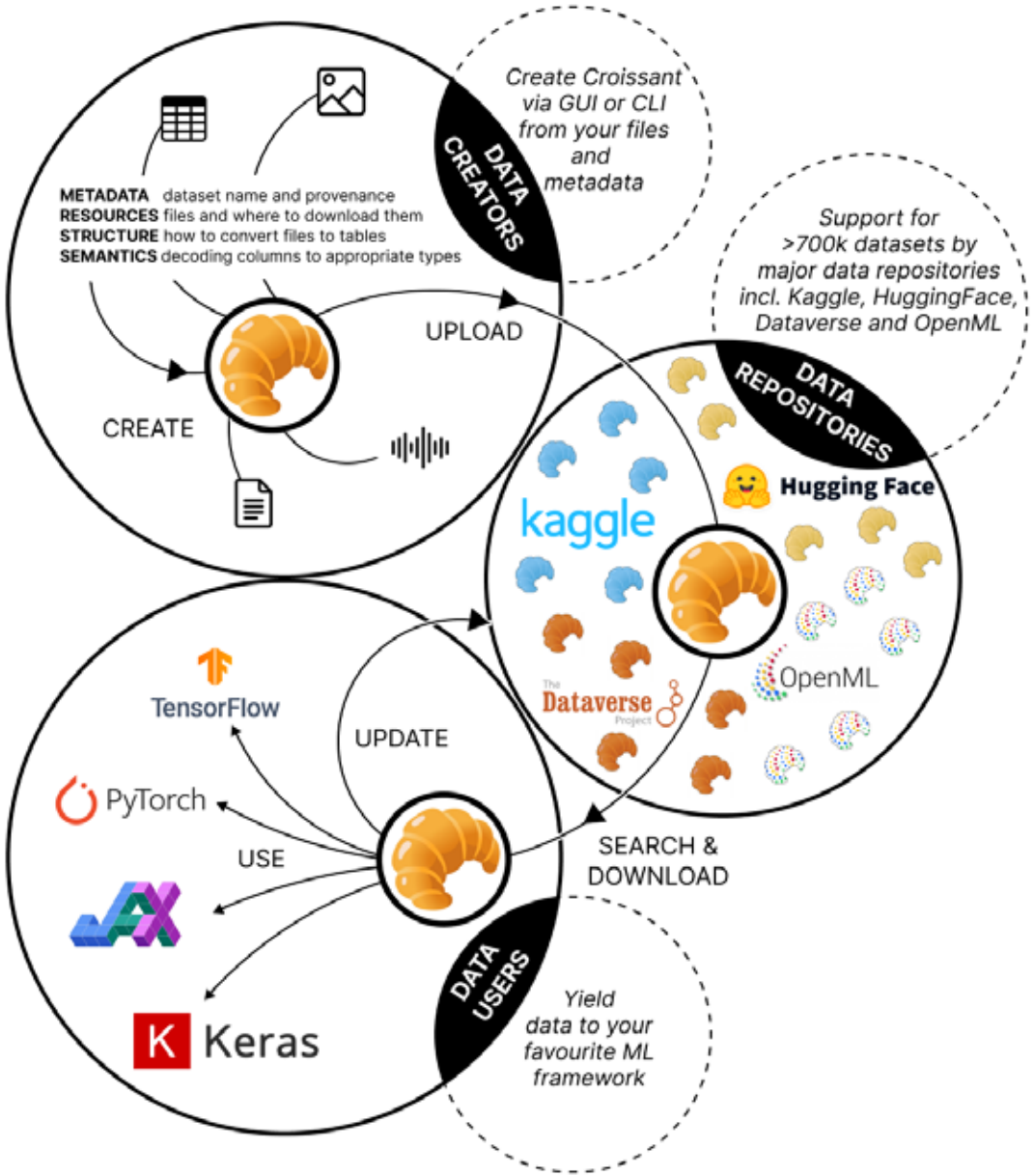
Transaction graph

Actor spectrum

- data consumer
  - frontier labs
  - agent wrappers
  - gen pop agents
  - gen pop eng.
- data provider
  - “mom and pop”
  - waking giants
  - web 2 brokers
  - scraping pros

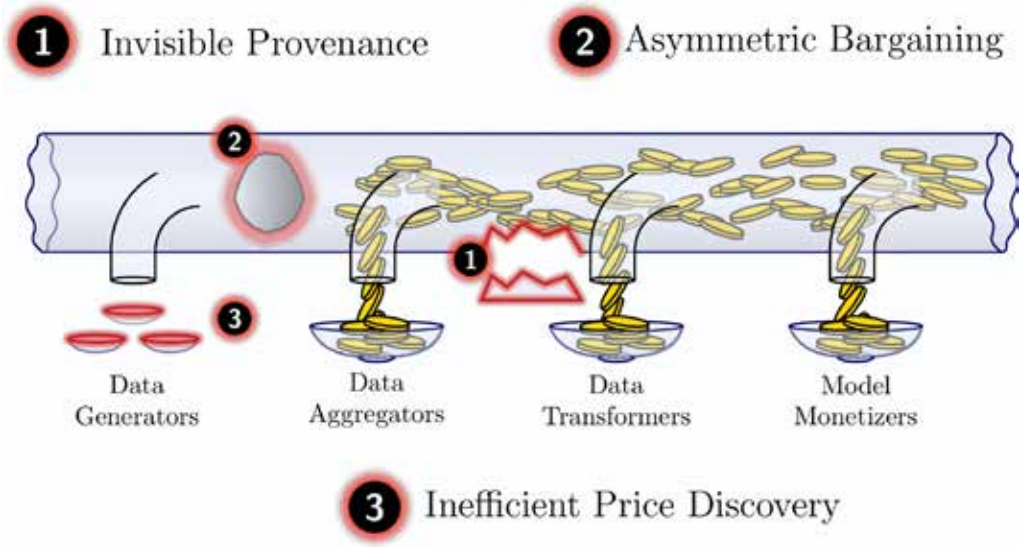
Recent steps

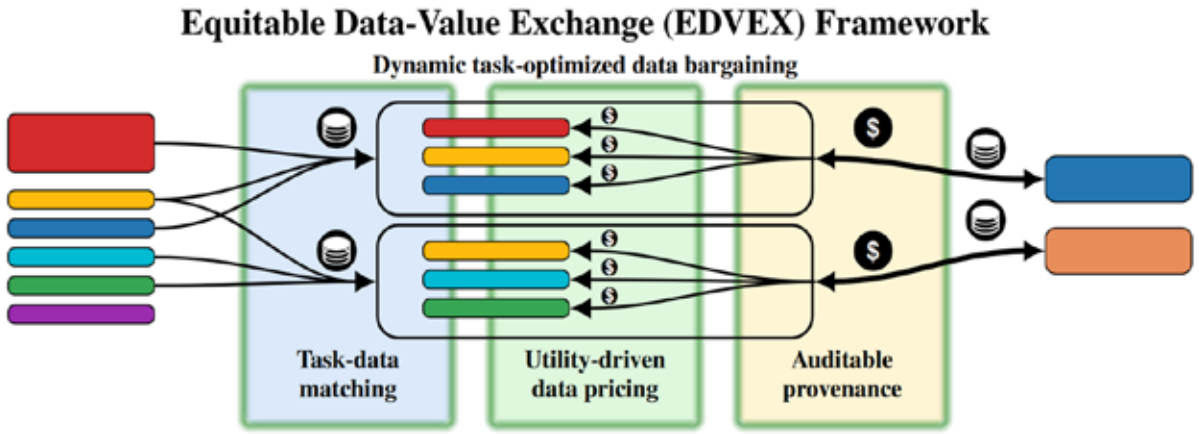
Croissant

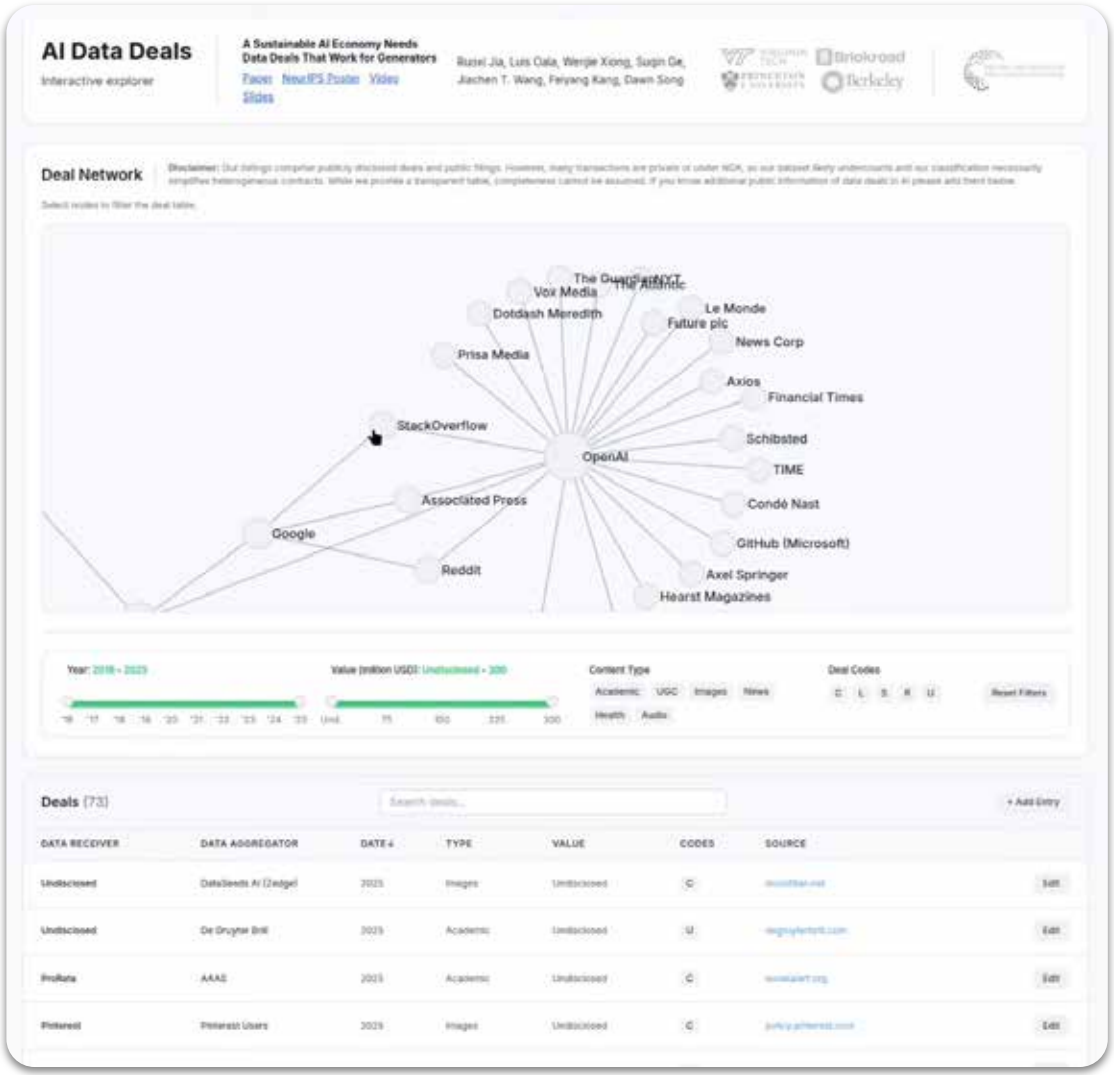


Akhtar, Mubashara, et al. "Croissant: A metadata format for ml-ready datasets." Advances in Neural Information Processing Systems 37 (2024): 82133-82148.

Data value chains

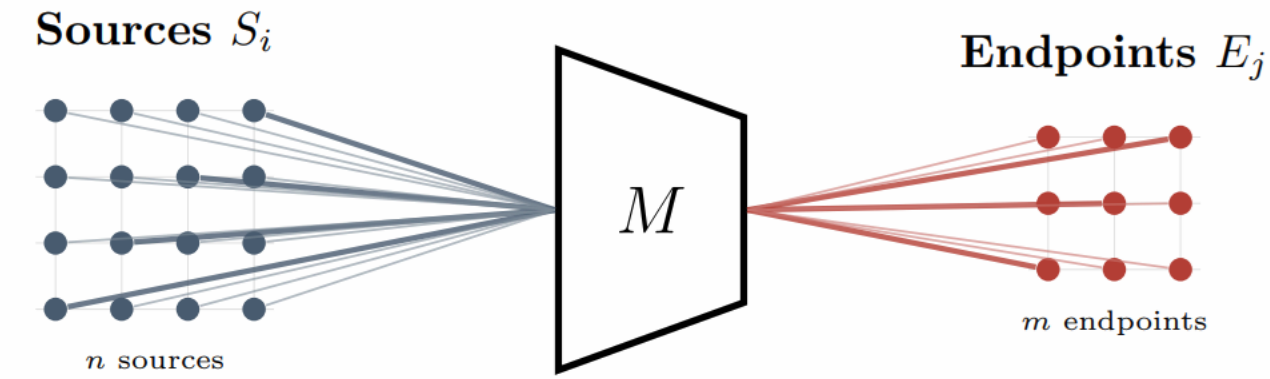


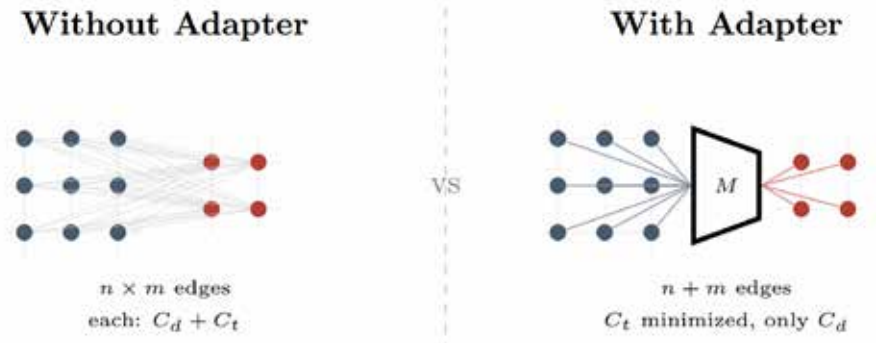




Jia, Ruoxi, et al. "A Sustainable AI Economy Needs Data Deals That Work for Generators." The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track (202

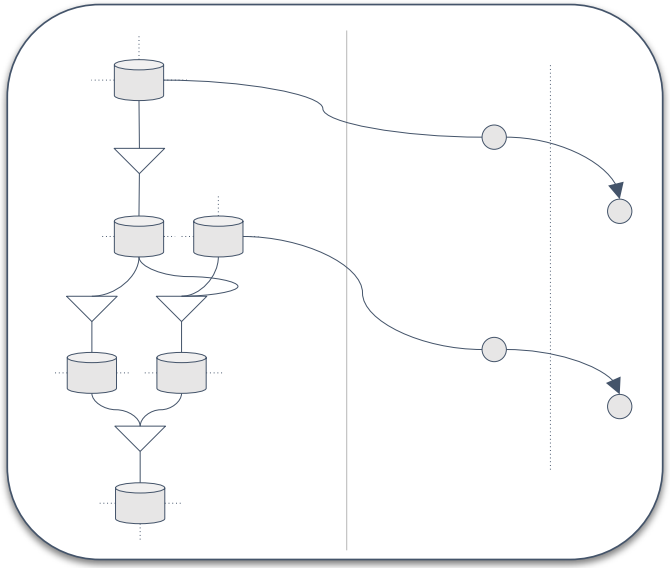
Frontier: Closing the loop via multiplexers

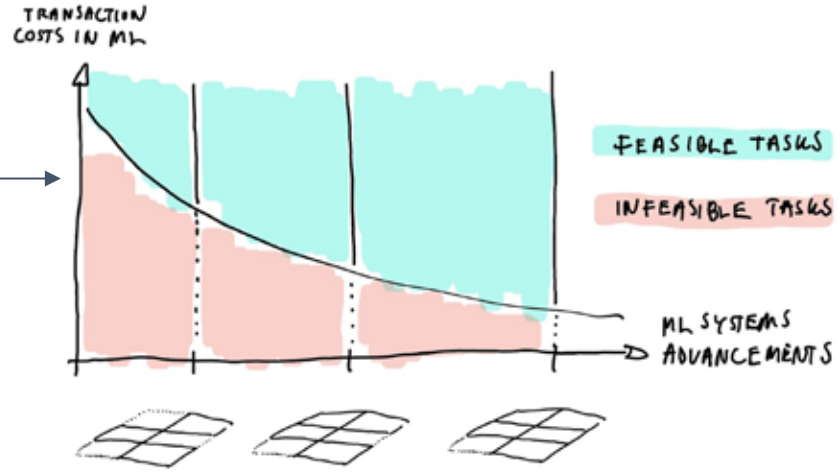




The data multiplexer M acts as a universal adapter between sources S_i and endpoints E_j, routing flows that maximize utility U while minimizing total cost $C = C_d + C_t$. Edge thickness indicates utility; color indicates direction (blue: source→M, red: M→endpoint).

Our north star: A road network for data without congestion





luis@brickroadapp.com
https://github.com/mlcommons/croissant

What's next

Expanding the federated catalogue
Onboarding partner institutions and data sources across GI-AI4H benchmarking challenges

More targeted dataset discovery
Using UMLS vocabulary and AI tools to enable detailed search across the catalogue

Tooling for Croissant adoption
Automated metadata generation, validation, and visual editing for data providers

Integration with evaluation pipeline
Connecting catalogue directly to benchmarking challenges for seamless dataset selection

Scaling the data mesh
Decentralized data ownership with cross-functional domain teams managing data as products

➞ **Now let me hand over to our expert from "BioCroissant" working group** for a deep dive into the standard, its specification, tooling, and how you can start using it.

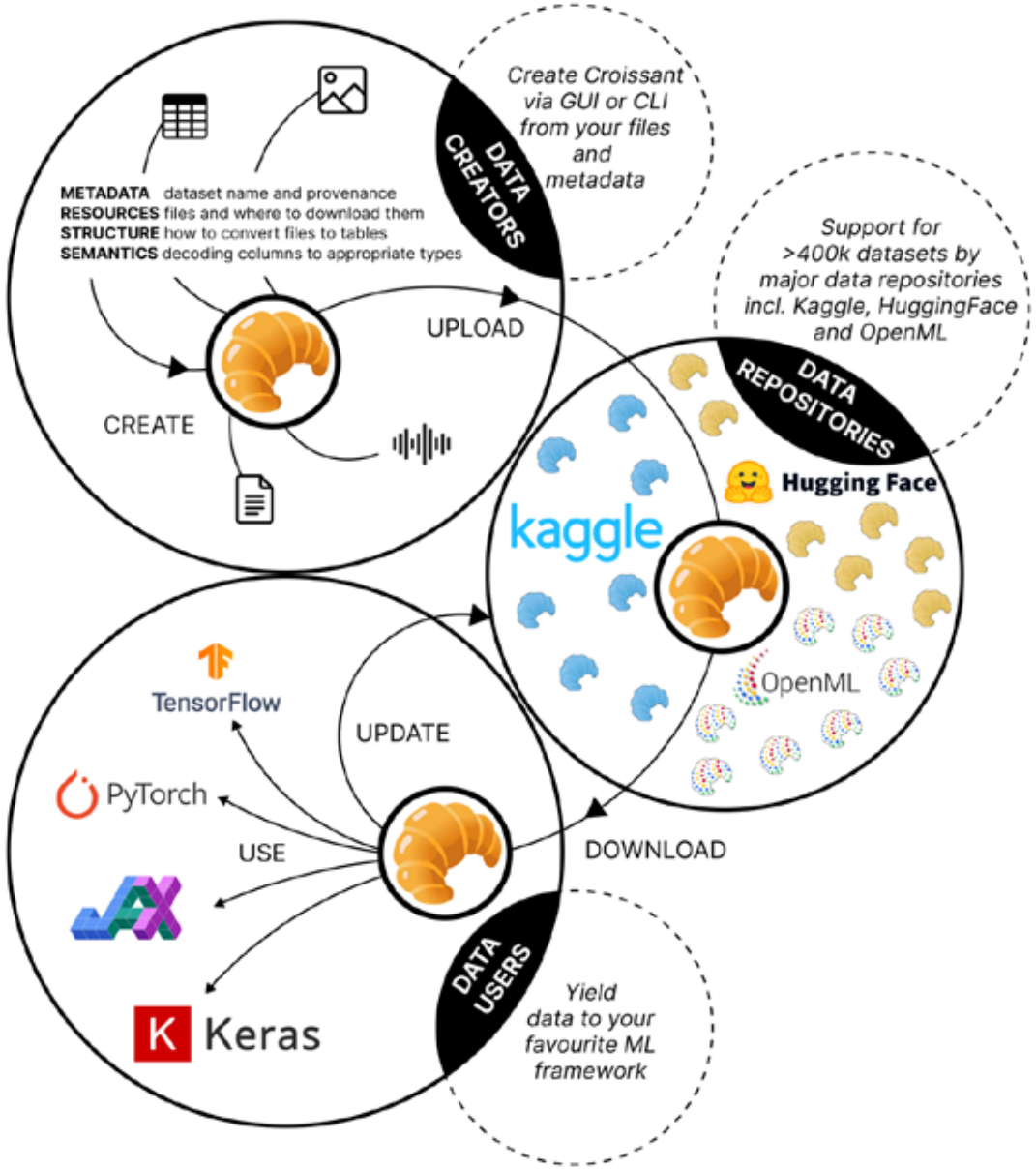
Croissant

A Metadata Format for ML-Ready Datasets

Raeesa Yousaf
Helmholtz Center Munich
German Diabetes Research Centre

Data is Source Code for AI

- **AI quality scales with data quality - but health data lives in silos**
  - Manual assembly does not scale
- **Croissant makes datasets machine-readable without changing the data**
  - One metadata layer across storage locations
  - Compatible for Agentic tools
- **It describes everything AI needs to know about data**
  - Access, structure, licensing, provenance, annotation, and intended use



Croissant on the inside


```
{
  "@type": "sc:Dataset",
  "name": "minimal_example_with_recommended_fields",
  "description": "This is a minimal example, including the required and the
recommended fields.",
  "license": "https://creativecommons.org/licenses/by/4.0/",
  "url": "https://example.com/dataset/recipes/minimal-recommended",
  "distribution": [
    {
      "@type": "cr:FileObject",
      "@id": "minimal.csv",
      "name": "minimal.csv",
      "contentUrl": "data/minimal.csv",
      "encodingFormat": "text/csv",
      "sha256":
"48a7c257f3c90b2a3e529ddd2cca8f4f1bd8e49ed244ef53927649504ac55354"
    }
  ],
  "recordSet": [
    {
      "@type": "cr:RecordSet",
      "name": "examples",
      "description": "Records extracted from the example table, with their
schema.",
      "field": [
        {
          "@type": "cr:Field",
          "name": "name",
          "description": "The first column contains the name.",
          "dataType": "sc:Text",
          "references": {
            "fileObject": { "@id": "minimal.csv" },
            "extract": {
              "column": "name"
            }
          }
        },
        {
          "@type": "cr:Field",
          "name": "age",
          "description": "The second column contains the age.",
          "dataType": "sc:Integer",
          "references": {
            "fileObject": { "@id": "minimal.csv" },
            "extract": {
              "column": "age"
            }
          }
        }
      ]
    }
  ]
```

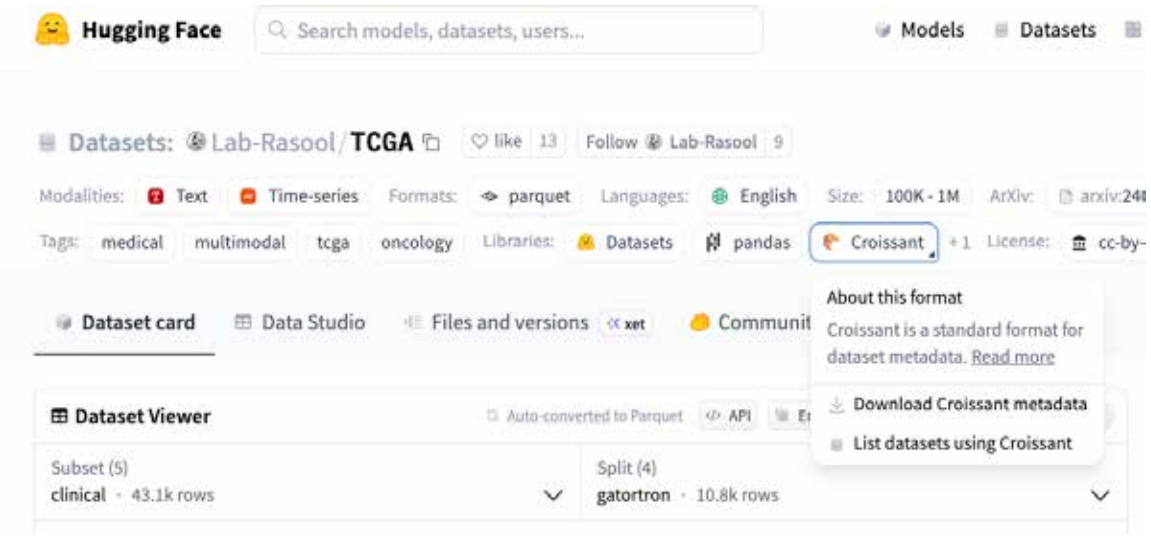

Metadata

Structure

Resources

Semantics

How to use it?




```
pip install mlcroissant
```

```python
# 1. Point to a local or remote Croissant file
import mlcroissant as mlc
url = "https://huggingface.co/api/datasets/fashion_mnist/croissant"
# 2. Inspect metadata
print(mlc.Dataset(url).metadata.to_json())
# 3. Use Croissant dataset in your ML workload
import tensorflow_datasets as tfds
builder = tfds.core.dataset_builders.CroissantBuilder(
    jsonld=url,
    record_set_ids=["record_set_fashion_mnist"],
    file_format='array_record',
)
builder.download_and_prepare()
# 4. Split for training/testing
train, test = builder.as_data_source(split=['default[:80%]', 'default[80%:]'])
```

Croissant is a great tool and very easy to use.
But is there something missing to be used in the medical domain?

Why BioCroissant WG is needed – what AI for Health is missing

In analogy to the European Healthcare Data Space (EHDS):

Governance above; technical execution below

EHDS Governance Layer

Legal framework · purpose limitation · access rights

HealthDCAT-AP

High level · Not auto-loadable

Dataset Discovery & Approval

Catalogues · portals · HDABs · access requests

HealthDCAT-AP

Findable · Not executable

Technical Dataset Specification

Features · splits · schemas · provenance · task definitions

BioCroissant

Machine-readable · Auto-loadable

Secure Processing & Analysis

SPEs · TREs · federated ML · AI model training

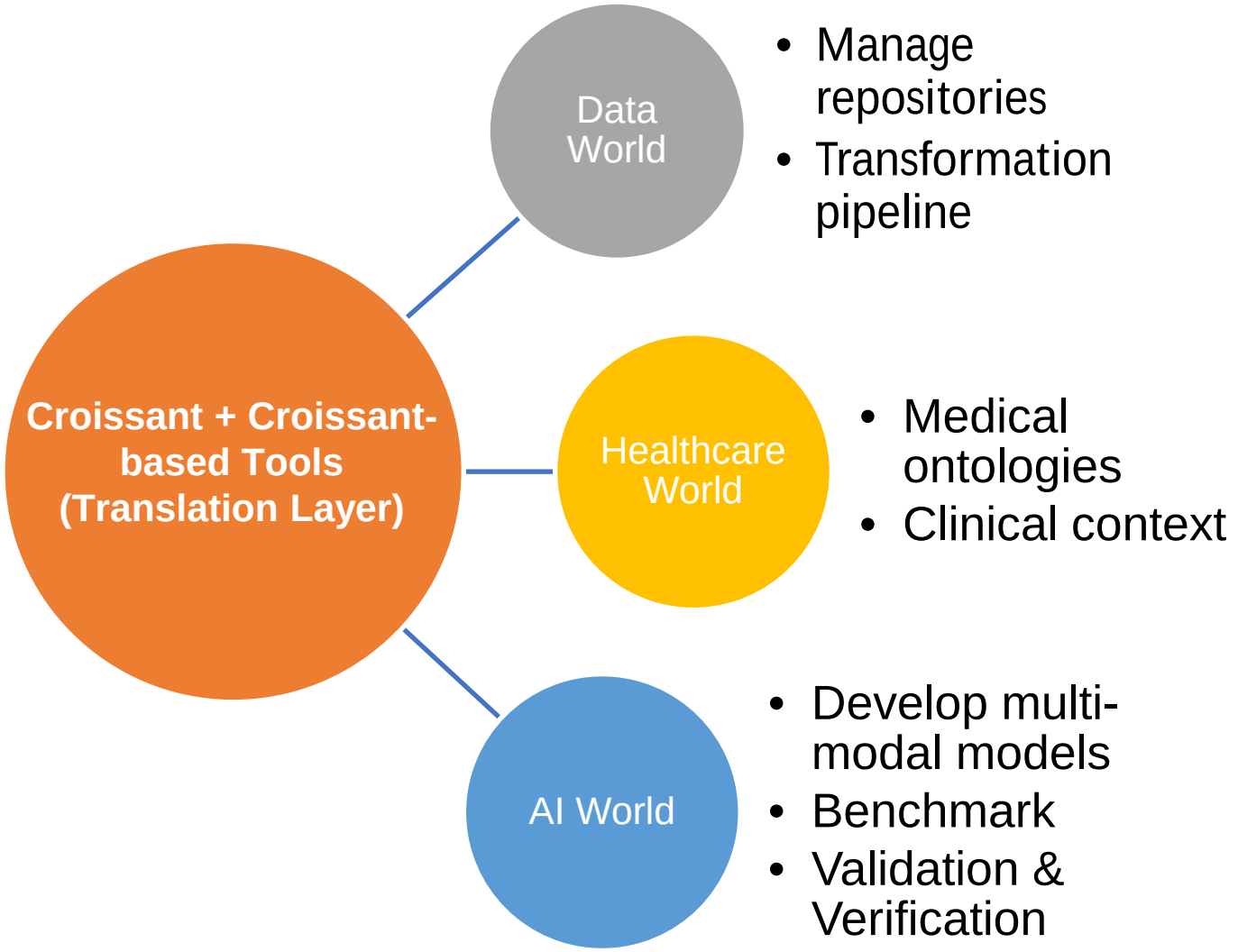
Execution layer

Requires EHDS-compliant access grant

- Non-normative & voluntary
- Downstream of EHDS governance
- Enables auto-loading of datasets
- EHDS-compatible technical profile

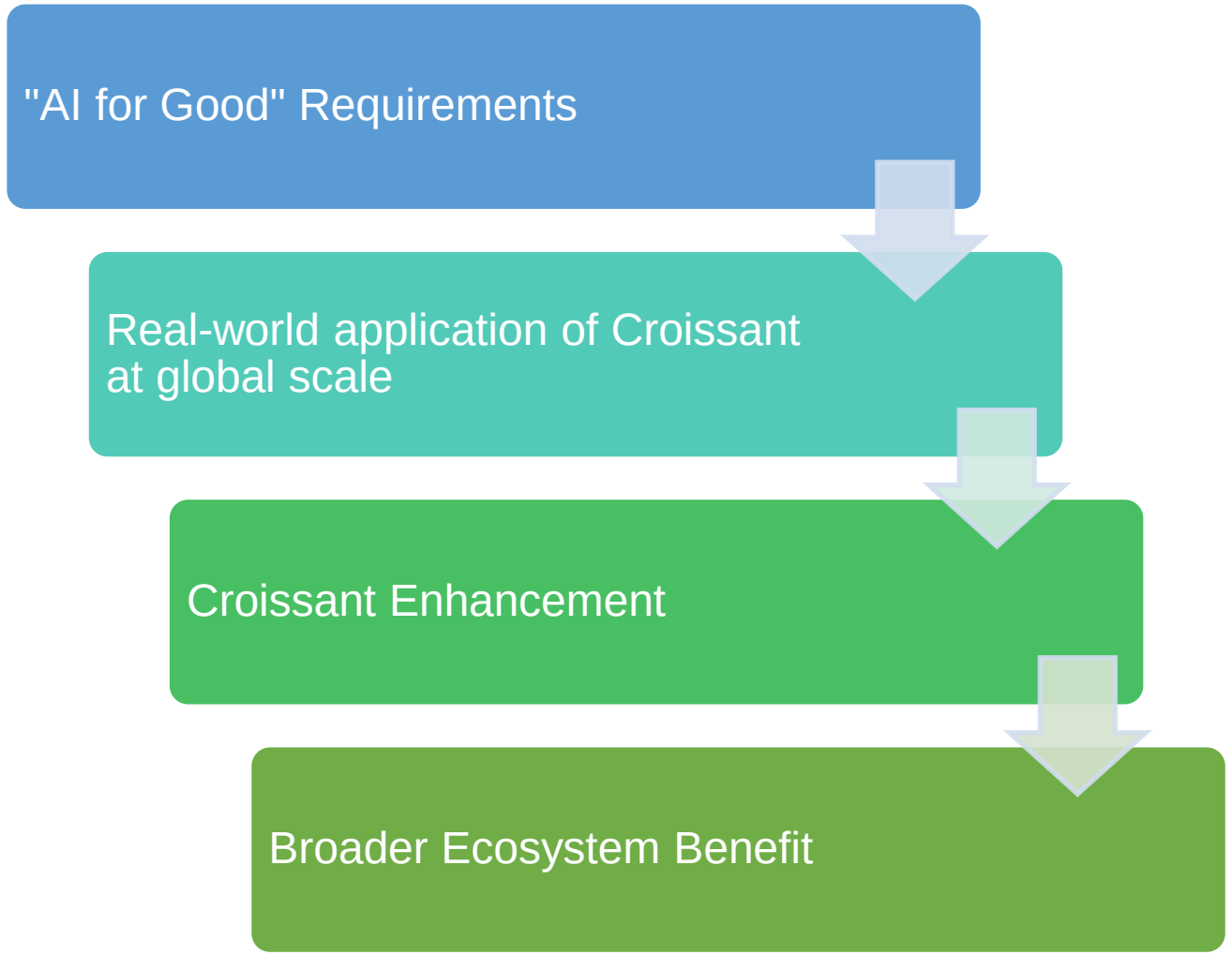
Goal: Bridging Three Communities with one data language

- Smart **metadata wrapping** across modalities, formats, and systems
- **Complexity stays under the hood.**
  Tooling handles structure, provenance, licensing, and semantics
- **Experts stay in their lane.**
  Each community works in familiar tools and concepts.



From "AI for Health" to Croissant (and Back)

- "AI for Health" brings **real-world requirements** to Croissant
- Global benchmarking turns **theory into practice**
- **Feedback** flows back into Croissant and **improves** it
- Since Croissant is domain-agnostic this "AI for Health" project can indirectly **accelerate standards maturity** for the entire AI field



Call to action: Join the Movement!

Integrate Croissant support into data repositories, analysis tools, and ML frameworks to expand ecosystem

Tool Developers

Reference Croissant standards in data governance frameworks to promote interoperability and FAIR compliance

Policy Makers

Researchers

Contribute to BioCroissant working group, share implementation experiences, and help refine health-specific extensions

https://mlcommons.org/working-groups/data/croissant/

Data Providers

Adopt Croissant for health datasets to improve discoverability and enable responsible reuse across platforms

Thank you for your attention!