ML Commons

What is MLCommons?
What is Medical Accuracy WG?
What is MedPerf?

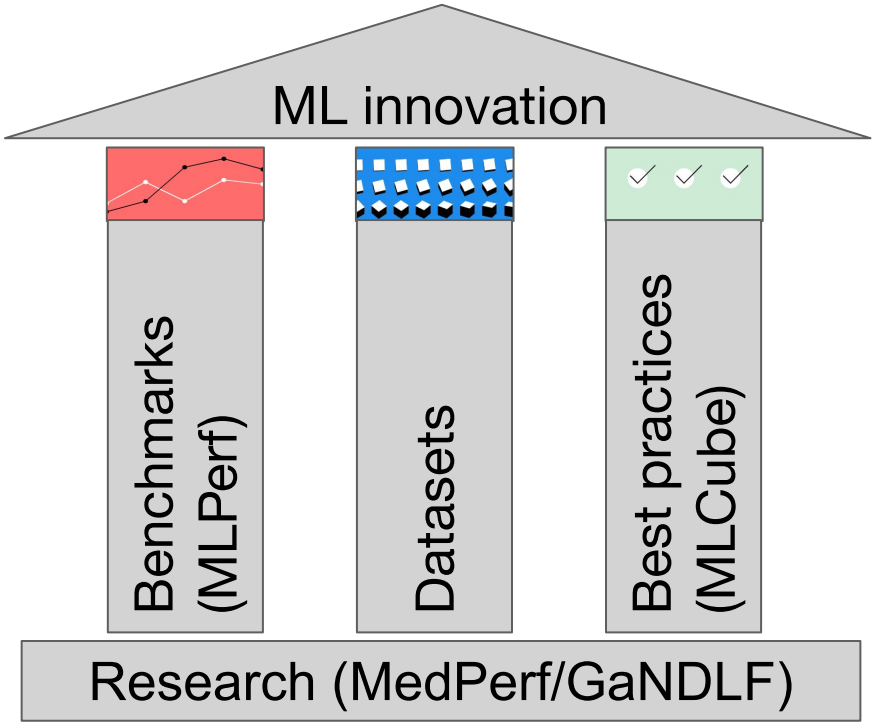
01-20-2023

Alexandros Karargyris, PhD
MLCommons/IHU Strasbourg

Some of us at Medical Accuracy WG

- Alex Karargyris: Research scientist at IHU Strasbourg
- Renato Umeton: Associate Director of Artificial Intelligence Operations at Dana Farber
- Micah Sheller: Research scientist at Intel
- Spyros Bakas: Assistant Professor at UPenn
- Sarthak Pati: Senior Software Developer
- Johnu George: Staff engineer at Nutanix
- Alejandro Aristizabal: Machine Learning engineer at Factored
- Hasan Kassem: Machine Learning scientist at IHU Strasbourg
- Peter Mattson: Research Scientist at Google

Mission: Better ML for Everyone



MLCommons is a global community

Founding Members

Academics from educational institutions including:

Harvard University
Indiana University
Polytechnique Montreal
Peng Cheng Laboratory
Stanford University
University of California, Berkeley
University of Toronto
University of Tübingen
University of York, United Kingdom
Yonsei University

MLPerf
For ML Benchmarks

MLPerf breadth: µWatts to MegaWatts

Evolution over time

Scale	2018	2019	2020	2021
Training - HPC				
Training				
Inference - Datacenter				
Inference - Edge				
Inference - Mobile				
Inference - Tiny (IoT)				
Storage				‘21?

Improving technical maturity

New training/inference benchmarks

- Recommendation: DLRM + 1TB dataset
- Medical imaging: 3D U-NET
- Speech-to-text: RNN-T
- NLP: BERT + wikipedia

Standardized methodology for Training

- Optimizer definitions
- Hyperparameter definitions
- Reference Convergence Points (RCP)

Added power measurement to Inference

Launched Mobile App

Tiny launched in June 2021

MLCommons Research

Algorithmic Research Working Group

- Benchmarks for algorithms to improve efficiency: better accuracy/compute

Medical Accuracy Working Group

- Federated *evaluation* across distributed data: research ~= clinical practice

Scientific Research Working Group

- Better datasets and software for science

DataPerf Working Group

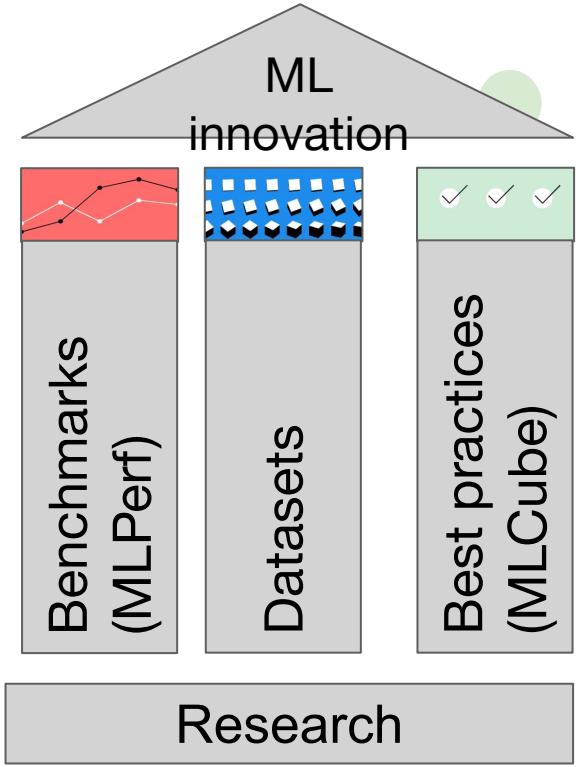
- Research & develop benchmarks for datasets

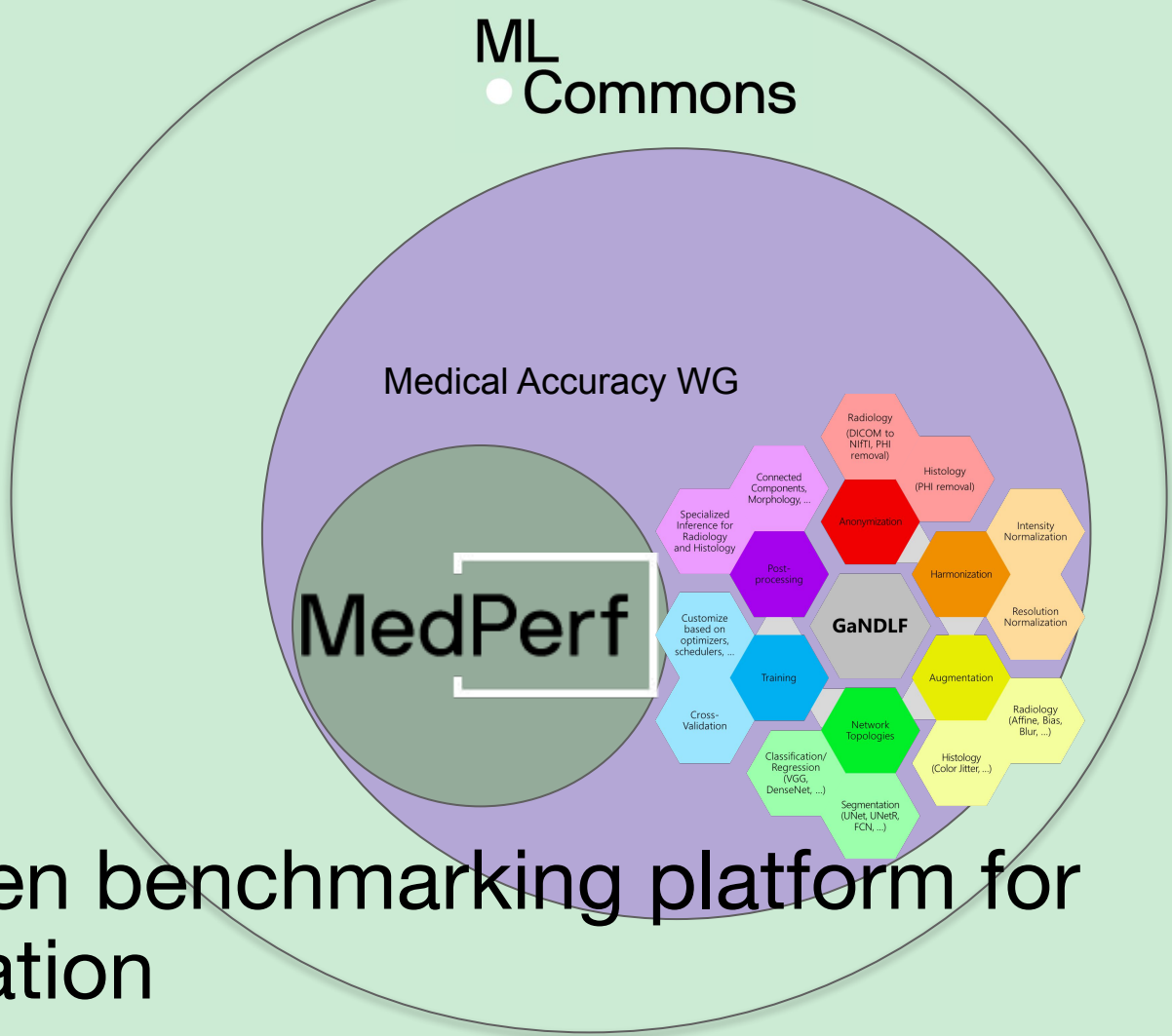
DynaBench

- Research platform for dynamic data collection and benchmarking

Storage Research Working Group

- Benchmarks to characterize performance of storage systems that support machine learning workloads



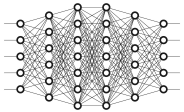


MedPerf: An open benchmarking platform for federated evaluation

ML
•C

Creating clinical impact through medical ML benchmarking

ML Research



Long Research Translation

Clinical World

Benchmarking can:

- Incentivize stakeholders: ML developers, Clinicians, Medical Data Owners, Decision Makers
- Help streamline research translation through best practices
- Create an open community with groups focused on improving key components of research translation

Grand Aims

3 areas of focus to bridge gap between research development and clinical efficacy:

- Model Evaluation
  - Focus on multi-institutional data
  - Evaluate model performance under multiple scenarios

- Benchmarking
  - Establish rigorous benchmarks with multi-disciplinary experts: health orgs, model developers, clinicians, patient advocacy groups, regulatory orgs
  - Clinical impactful and scalable
  - Support benchmarks by fostering an ecosystem of organizations

- Research
  - Utilize methods that measure out-of-distribution in new (target) data and expected change in model accuracy

Current Stage

- Technical Infrastructure Development
- Feasibility studies with UPenn, Univ. Strasbourg, and Dana Farber
- MedPerf infrastructure for the Federal Tumor Segmentation Challenge at MICCAI 2022
- Continuous evaluation for neuro-oncology in 2023

White paper is available at mlcommons.org/medperf and medperf.org

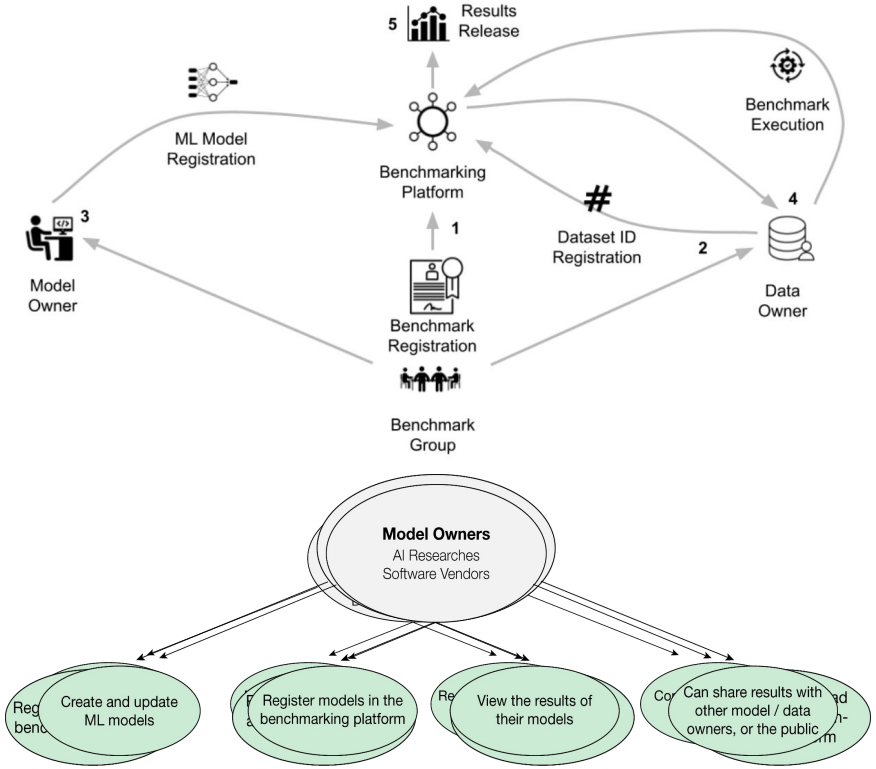
Members from 18 companies, 13 universities, 5 hospitals and 10 countries

MedPerf technical description

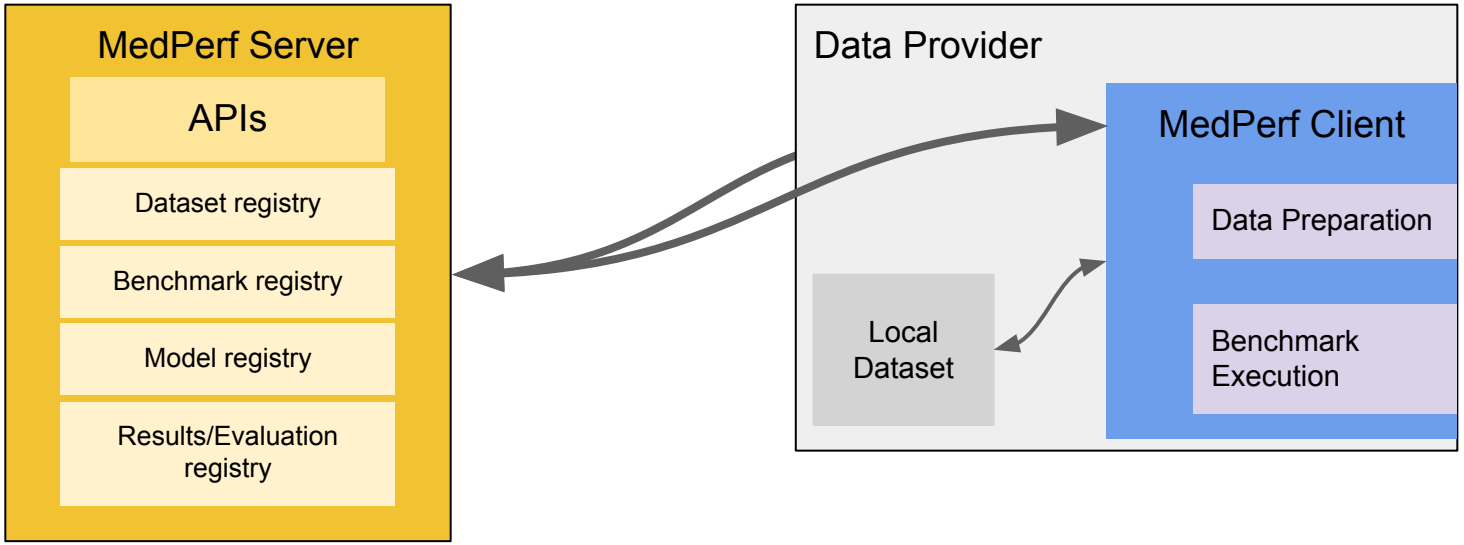
MedPerf

ML
C

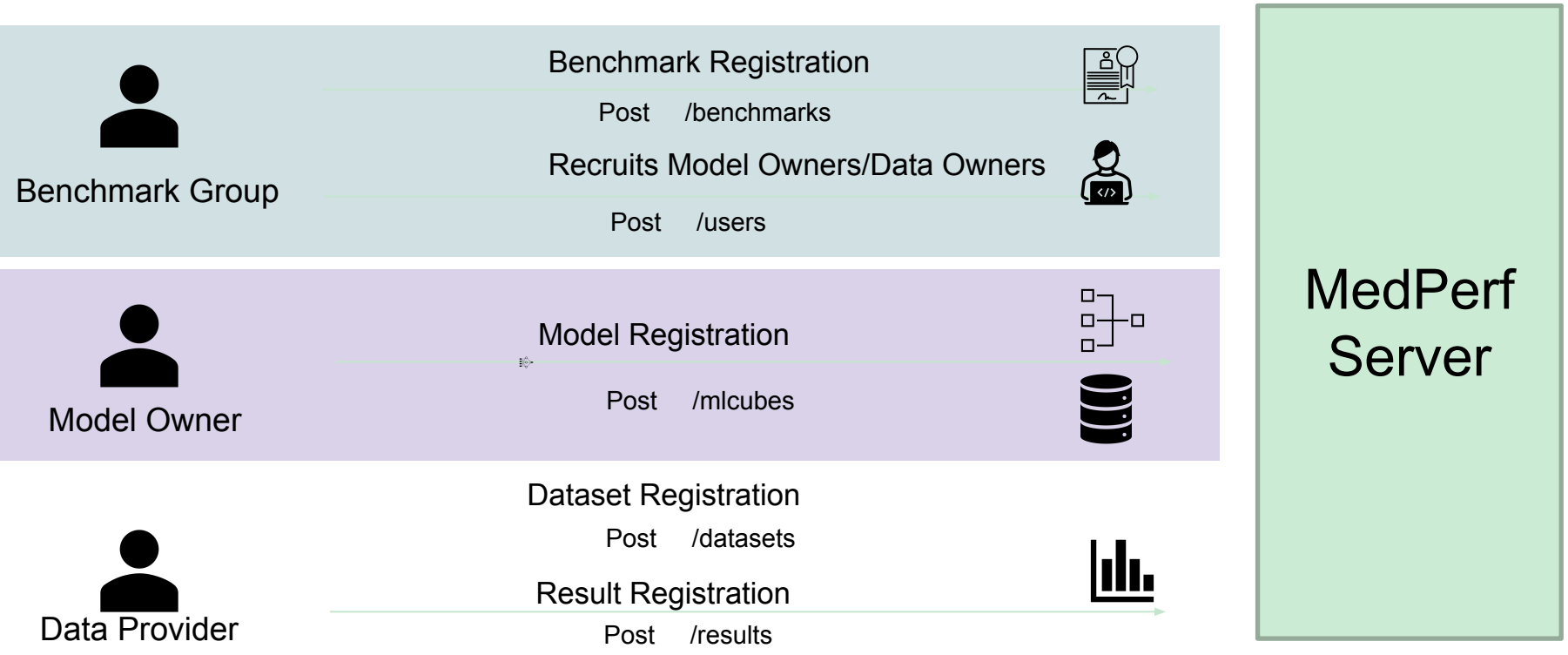
Design Philosophy



MedPerf Design



MedPerf Server Workflow



MedPerf CLI

- **Data Preparation Flow**
- Benchmark Execution Flow

Data Preparation Flow


```
$ medperf prepare -b benchmark -d data/ -l data/labels.csv
```

Prepare a raw dataset for federated evaluation.

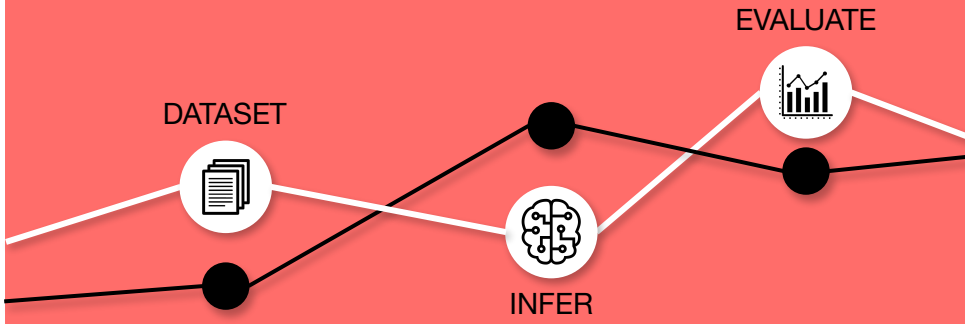
1. PREPARE: Multiple formats to single standard format.
2. SANITY CHECK: Ensure data is clean and in the standard format.
3. STATISTICS: Compute statistics from the dataset.

The prepared dataset stays in the Data Owner's machine. Statistics are uploaded to the server under Data Owner's approval

MedPerf CLI

- Data Preparation Flow
- **Benchmark Execution Flow**

Benchmark Execution Flow




```
$ medperf execute -b benchmark -d data_uid -m model_uid
```

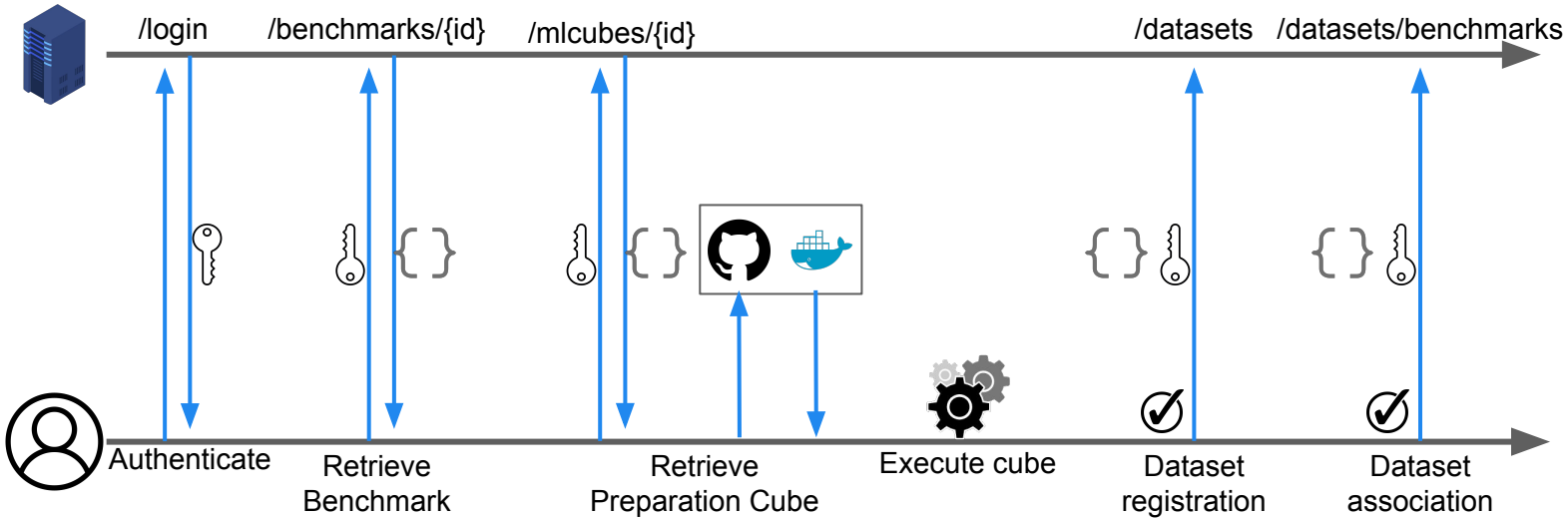
Evaluate a model's performance on a prepared dataset.

1. **INFER:** Obtain predictions from the specified model.
2. **EVALUATE:** Calculate model performance on certain metrics.

Model performance metrics are uploaded to the server under Data Owner's approval.

Data Preparation Flow

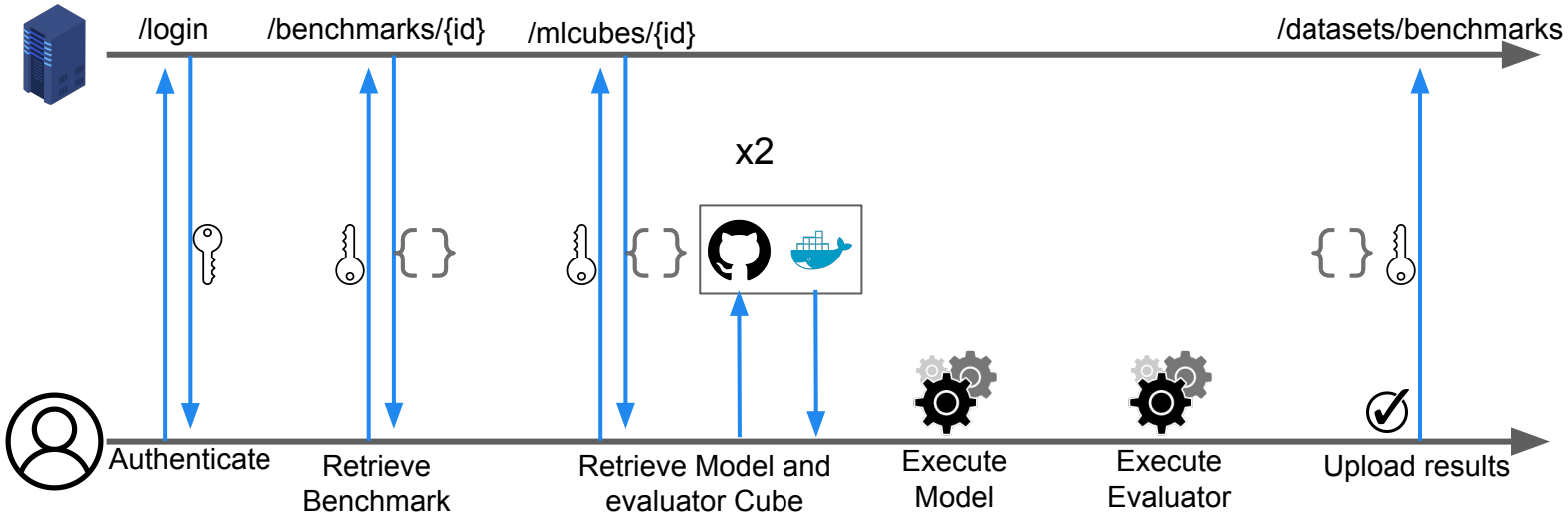
```
$ medperf prepare -b benchmark -d data/ -l data/labels.csv
```



Glossary: Auth Token | JSON Payload | Local Process Execution | User Confirmation

Benchmark Execution Flow

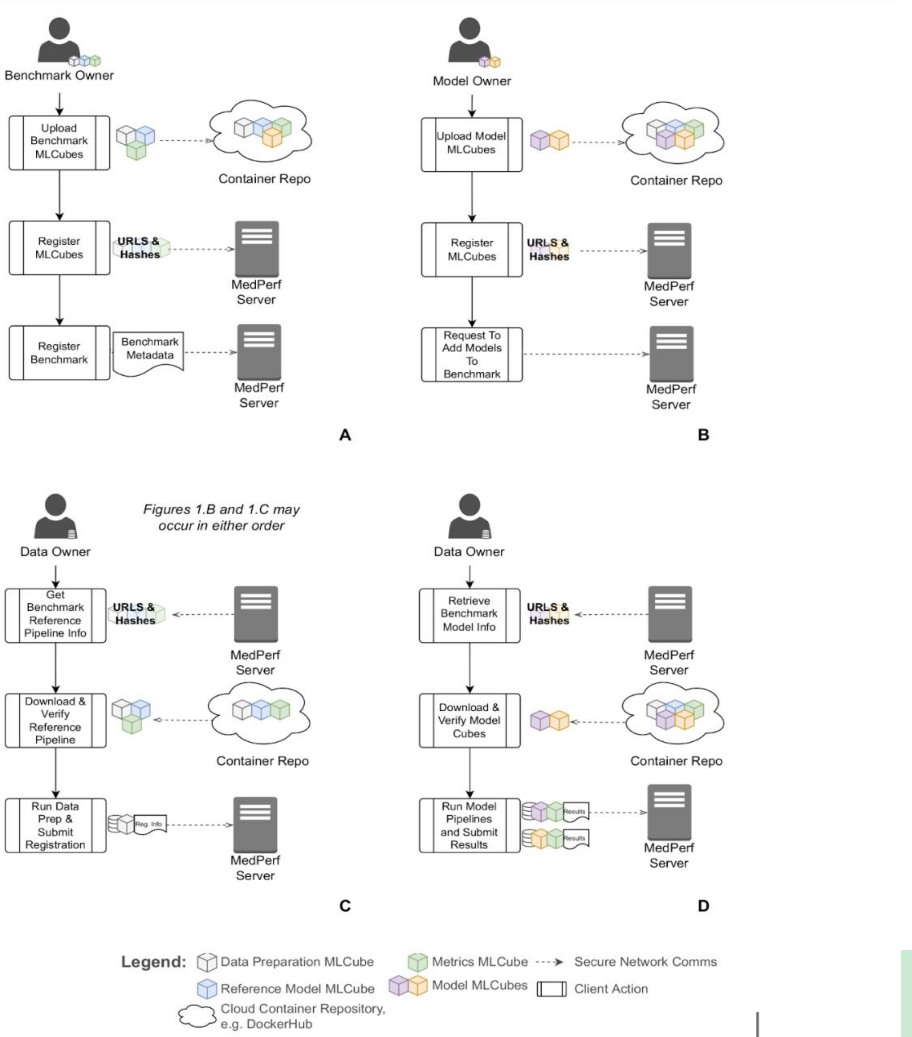
```
$ medperf execute -b benchmark -d data_uid -m model_uid
```



Glossary: Auth Token | JSON Payload | Local Process Execution | User Confirmation

High Level Workflow Diagram

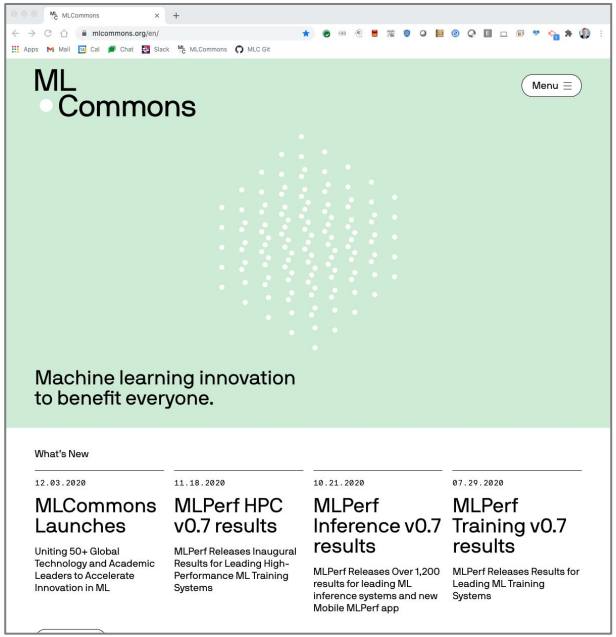
Please visit medperf.org or mlcommons.org/medperf For more information



How can I get involved?

We welcome people who want to make ML better.

- Join our mailing list
- Attend community events
- Become a member (free for academics)
- Participate in working groups
- Create or join benchmark at medperf.org



Join us at **mlcommons.org**

Visit medperf.org

Email us at **medperf-hello@mlcommons.org**