Machine Learning Development Environment for Single-cell Sequencing Data Analyses

Dong Xu

Department of Electrical Engineering and Computer Science

C.S. Bond Life Sciences Center

University of Missouri

Developers: ***Lei Jiang****, Yuexu Jiang, Clement Essien, Juexin Wang*

Collaborator: Qin Ma, Ohio State University

NIH: R35-GM126985

Challenges for Method Development

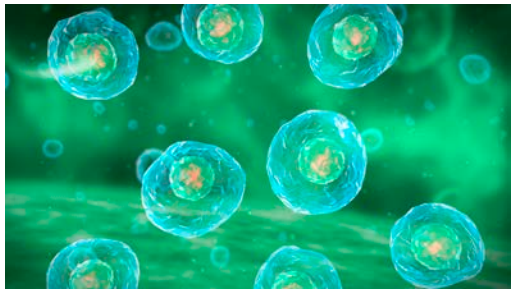
Technological complexity | biological knowledge | data quality | data formatting

Single-cell data are complex and inherently unstandardized.

Most data require preprocessing, such as quality control and normalization before any meaningful analysis can be conducted.

Problem formulations typically require domain knowledge about single-cell technologies and underlying biology.

It is hard to improve a certain method (e.g., just clustering algorithm) without mastering an entire analysis pipeline.



University of Missouri

Ecosystem for Single-Cell Data Analyses

One platform

Workflow engine

Streamlined
Standardized
Automated
Reproducible

Data engine

AI-ready
Large-scale
Benchmarks
Findable

UI engine

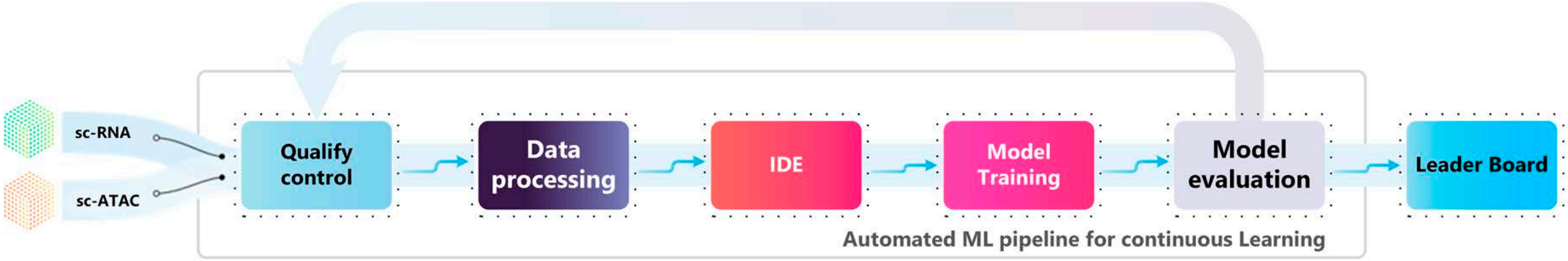
Accessible
Customizable
User-friendly
Visualized

Orchestration engine

Scalable
Available
Sustainable
Flexible

Workflow Engine

Automated ML pipeline for continuous learning



Public Datasets
- GEO
- SRA
- CancerSEA
- 10X Genomics
- Simulation
- ...

Benchmarks
- Raw counts

Feature store
- Normalized data
- Corrected data
- Summarized data

Template Libraries

GitHub

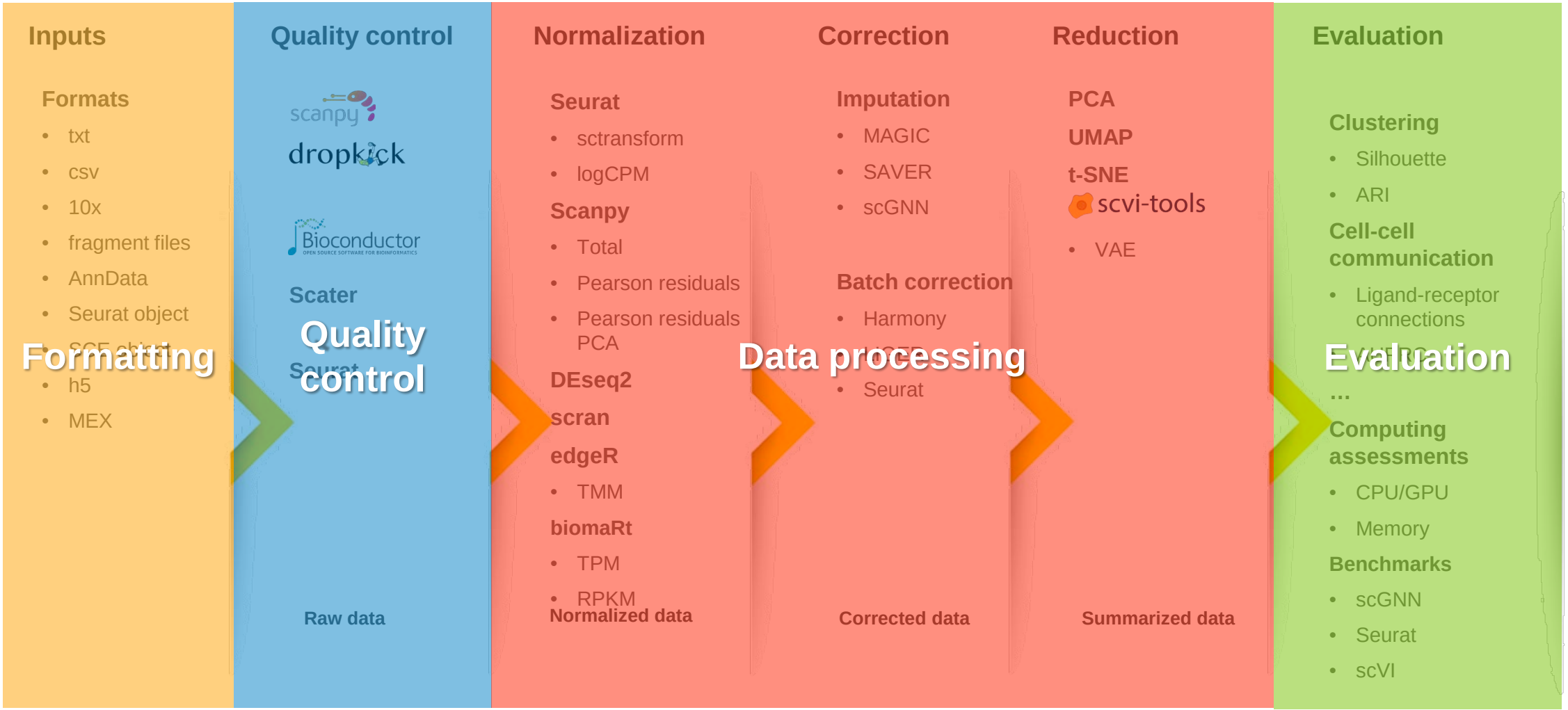
Environments

YML

Performance Assessments
- Performance metrics
- Computing assessments

Ranking
- Marker genes
- Clustering
- Trajectory inference
- Integration
- Cell-cell communication
- ...

Workflow Engine



Data Engine

Large-scale AI-ready benchmarks

Datasets

Sources

- GEO
- SRA
- CancerSEA
- 10X Genomics
- Simulation
- …

Species

- Human
- Mouse
- Virus
- …

Tissues

- Brain
- Liver
- COVID
- …

Stage of processing

Raw data

Normalized data

Corrected data

Summarized data

Tasks

- Differential expression
- Marker genes
- Genes over condition
- Genes over time

- Differential expression
- Marker genes
- Genes over condition
- Genes over time
- Cell-cell communication
- Gene regulatory relations

- Visual comparison of data
- Trajectory inference
- Imputation
- Multi-omic data integration

- Visualization
- Trajectory inference
- Clustering
- KNN graph inference
- Cell type identification

University of Missouri

UI Engine

Large-scale AI-ready benchmarks

Web

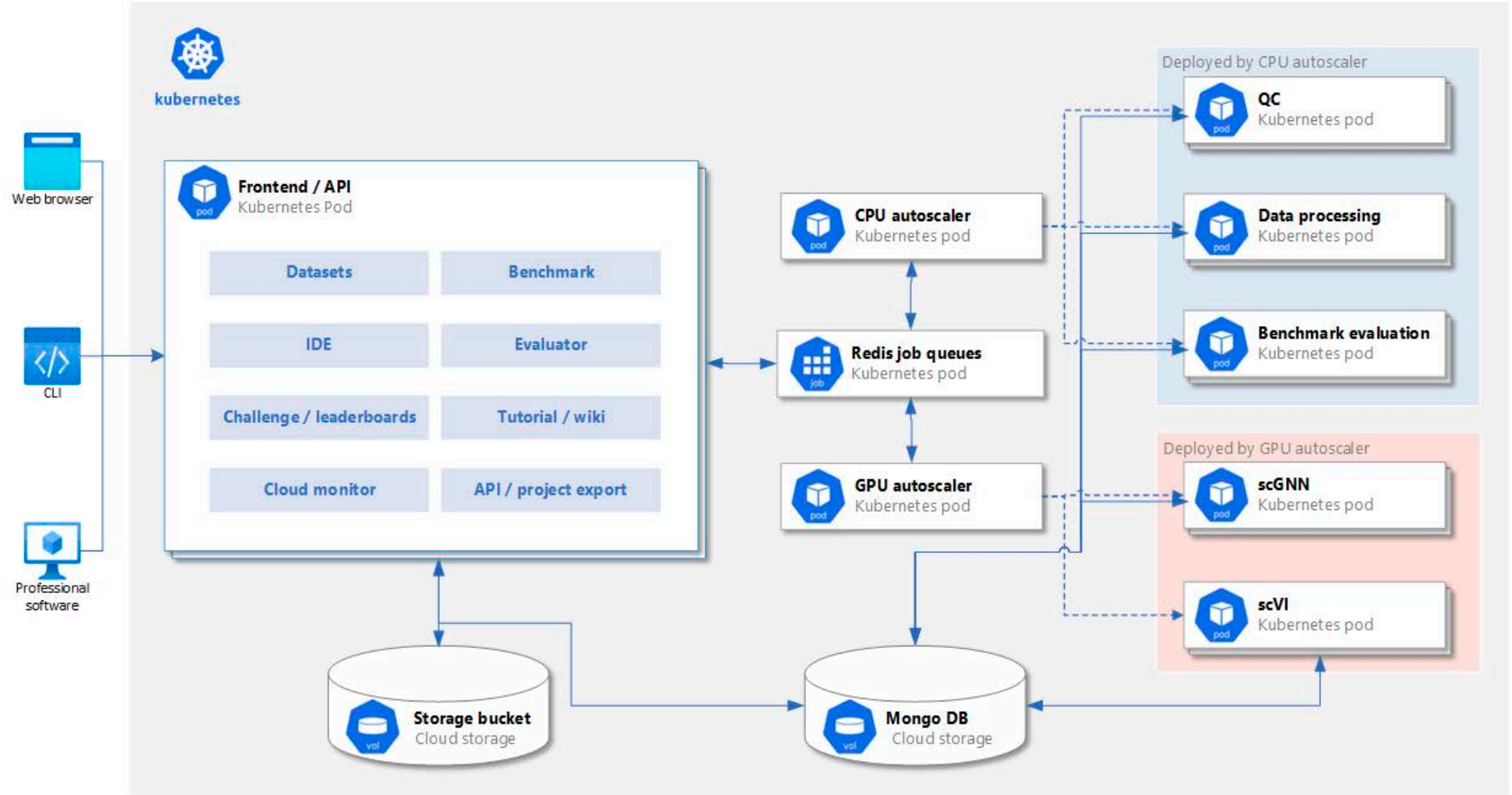
- Analysis
- Method development
- Education
- Documentation
- Visualization

CLI

- IDE
- Libraries

Professional software

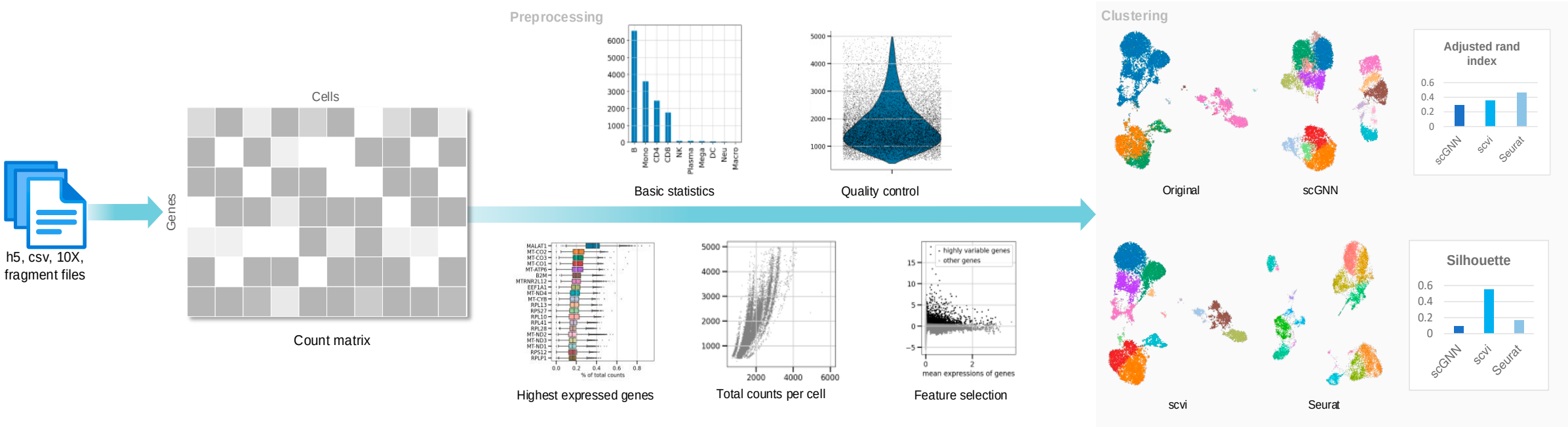
- Plugin
- API



University of Missouri

UI Engine

Single-cell sequencing data analyses



University of Missouri

Orchestration Engine

Configuration-based | code as architecture

Configuration-based development

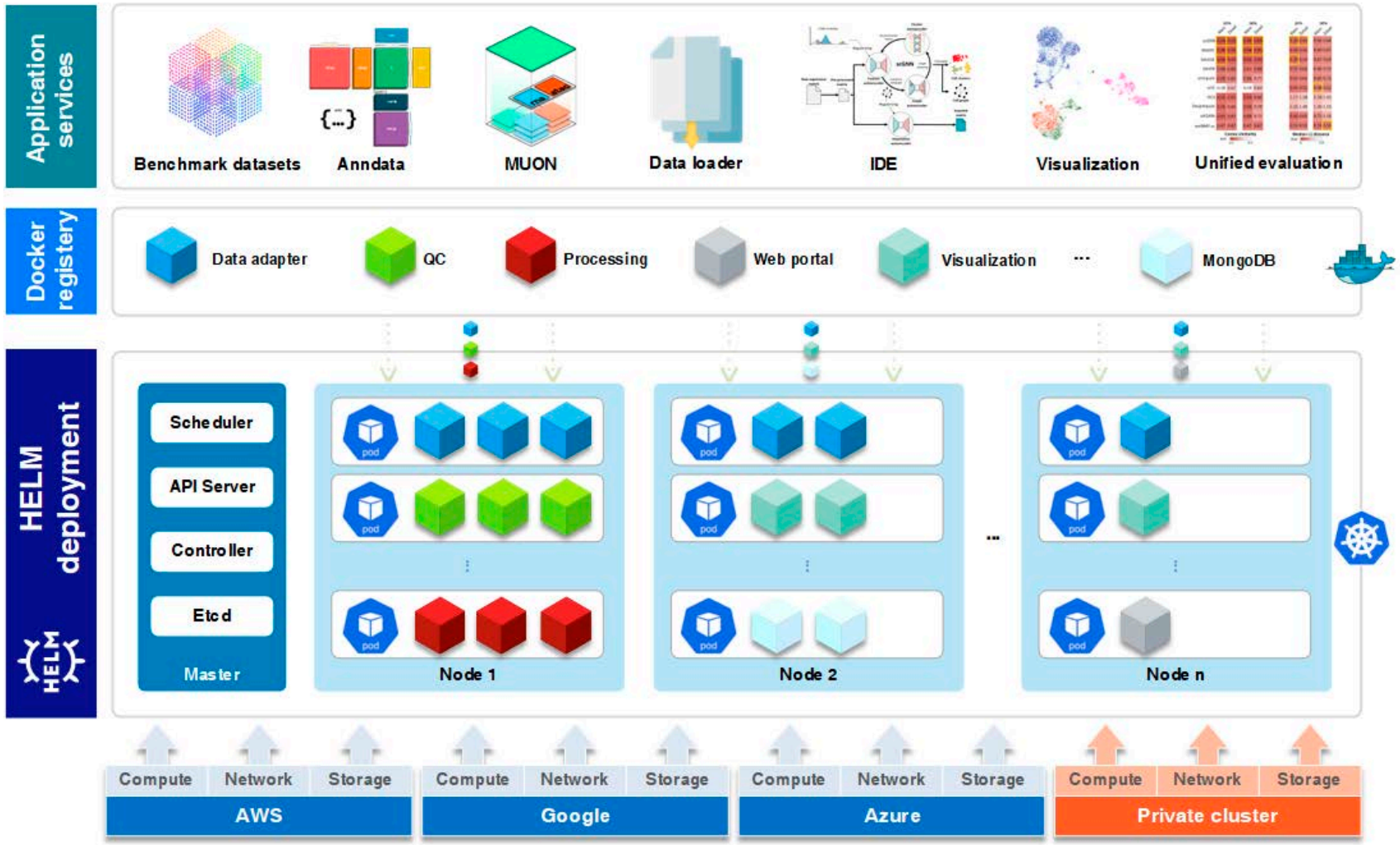
- Snakemake config
- Dockerfile
- Docker-compose
- Helm template

Code as architecture

- Container
- Kubernetes
- Cluster
- Cloud

Management

- Version control
- Update
- Monitoring
- Backup



University of Missouri

Summary

- **Automated end-to-end single-cell analyses ML pipelines** are developed to simplify and standardize single-cell data formatting, quality control, loading, model development, and model evaluation.
- The platform can significantly lower the method development barrier in single-cell data analyses for machine-learning researchers.
- **Future work**: open the portal for community data submission and analysis
- **Publications**
  - Qin Ma and Dong Xu. Deep learning shapes single-cell data analysis. Nature Reviews Molecular Cell Biology, 23:303–304, 2022. (https://www.nature.com/articles/s41580-022-00466-x)
  - Anjun Ma, Juexin Wang, Dong Xu, Qin Ma. Deep learning analysis of single-cell data in empowering clinical implementation. Clinical and Translational Medicine, 12(7): e950, 2022. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9299757/)
  - Lei Jiang, Yuexu Jiang, Cankun Wang, Clement Essien, Juexin Wang, Anjun Ma, Qin Ma, Dong Xu. Machine learning development environment for single-cell sequencing data analyses. Poster, ISMB, 2022. (https://iscb.junolive.co/ismb2022/library/search/ismb2022_poster_503)