MAX PLANCK INSTITUTE
FOR THE SCIENCE OF LIGHT

DCOR: A DATA REPOSITORY INTEGRATING POSTPROCESSING, BASED ON CKAN+S3+SLURM AT MPCDF

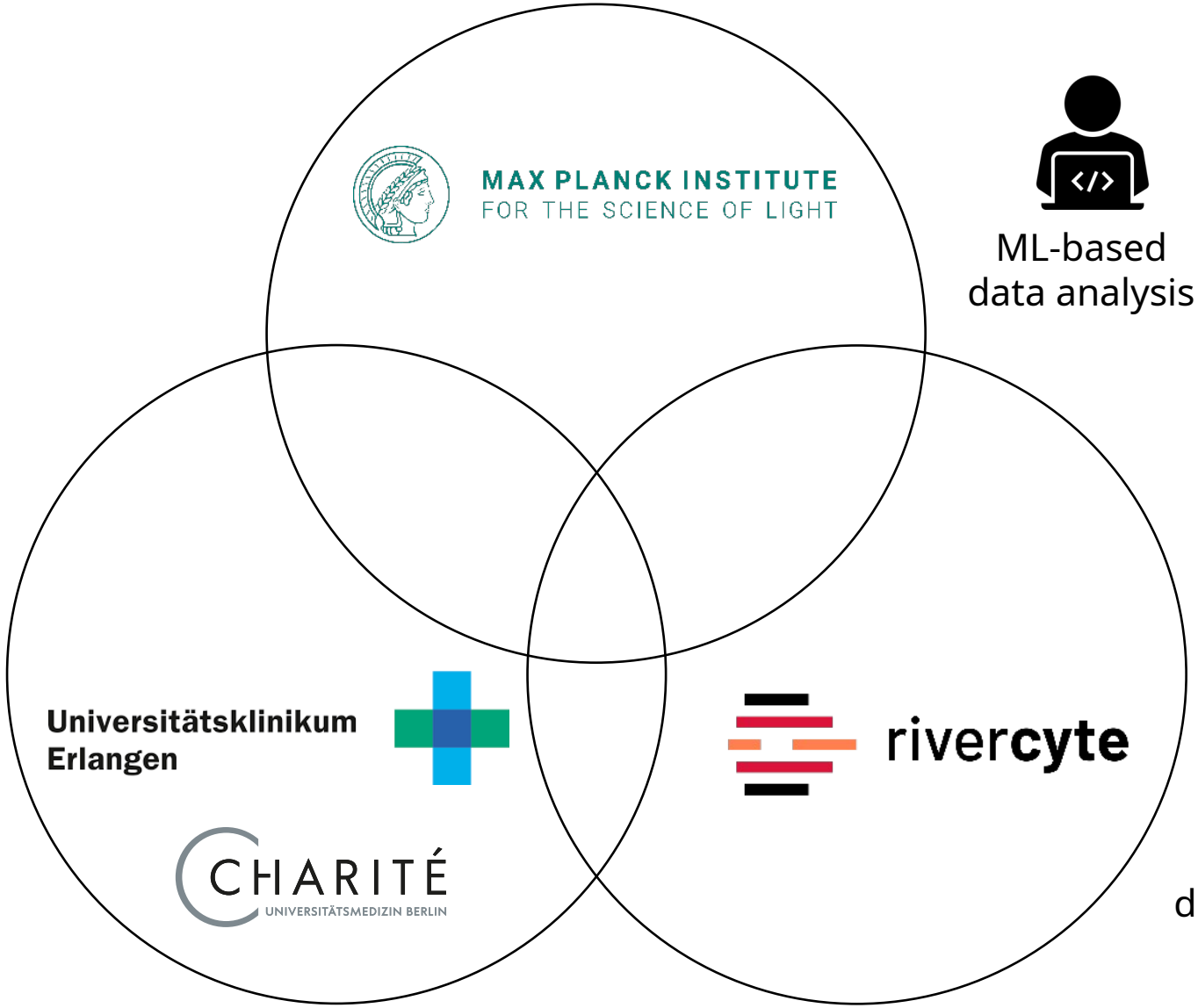
HPC Cloud Workshop
Max Planck Computing & Data Facility, Garching

Paul Müller
2024-09-11
Max Planck Institute for the Science of Light, Erlangen

CONFLICTS OF INTEREST DECLARATION

Paul Müller is Co-Founder of Rivercyte GmbH, a company that markets a deformability cytometry device with the aim of establishing a new medical blood test.

THE BIG PICTURE



ML-based
data analysis

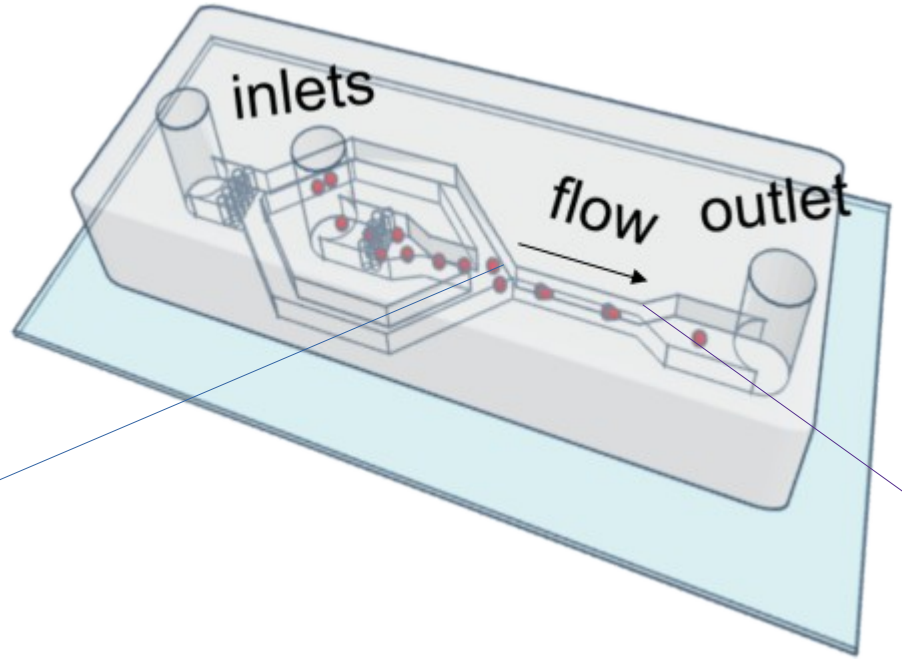
Measurement of
2000 patient
samples

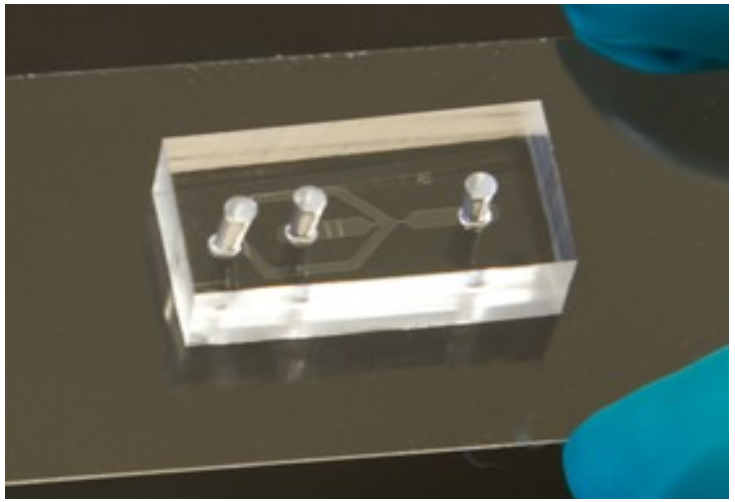
Hardware
development

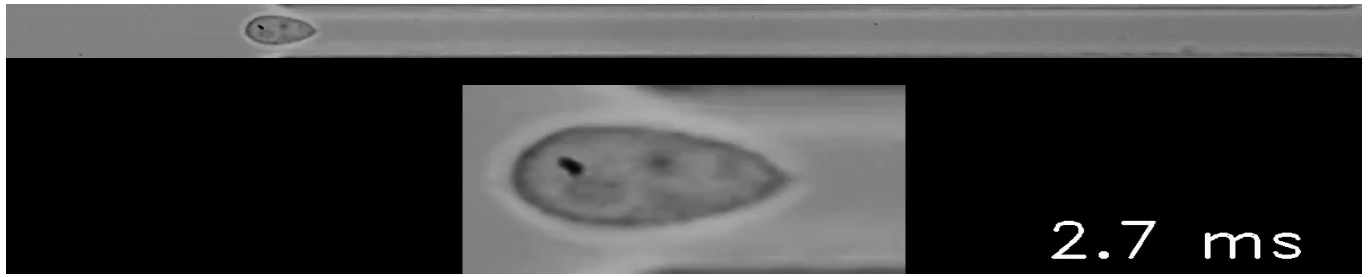
bayern innovativ

Bayerisches Staatsministerium für
Wirtschaft, Landesentwicklung und Energie

DATA ACQUISITION







- 10 min measurement
- 1 000 000 cell images from 1µL of blood
- label-free
- single-cell image-based analysis

Otto O, et al. Nature Methods (2015)
Rosendahl P, et al. Nature Methods (2018)

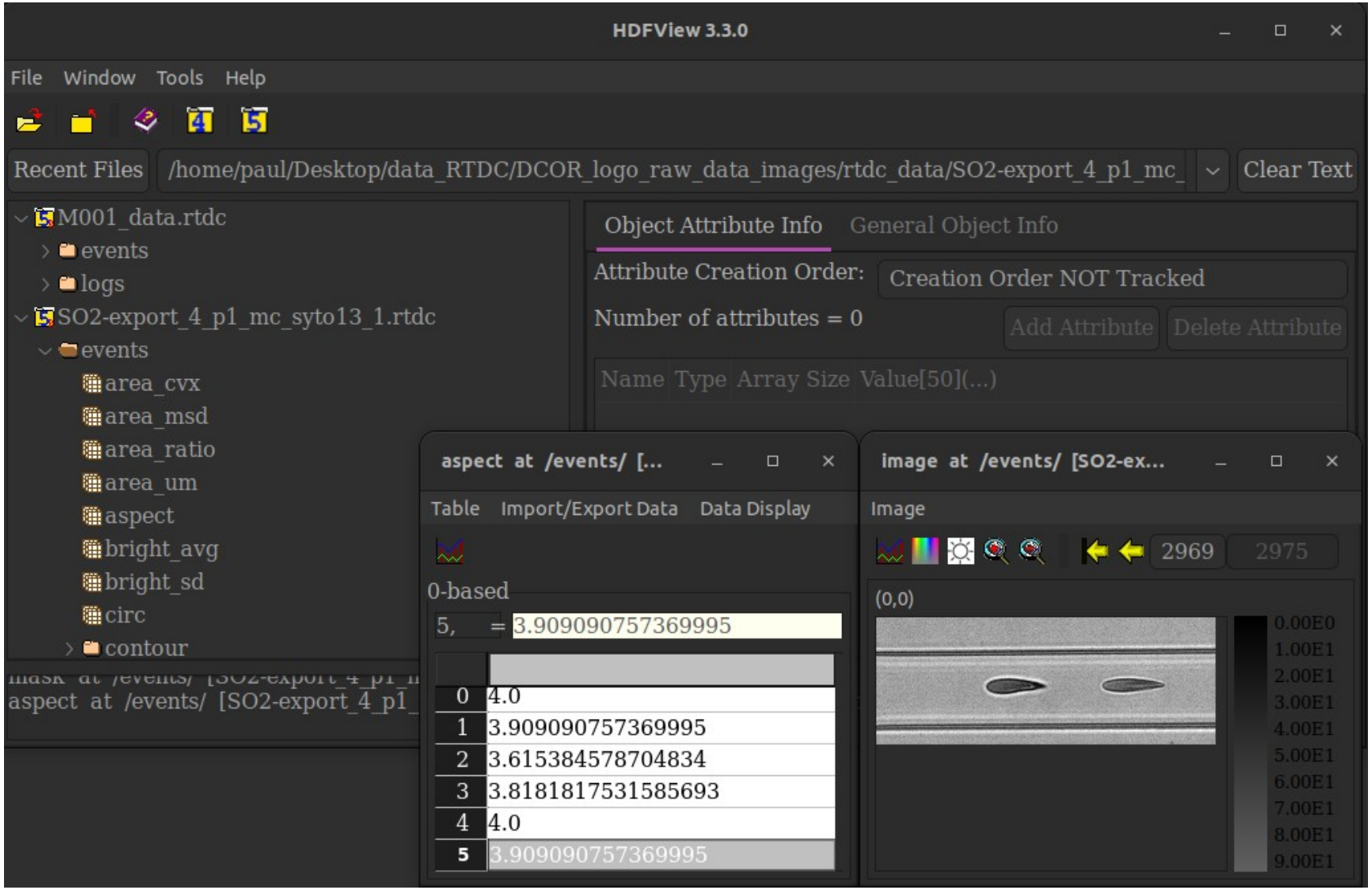
DEFORMABILITY CYTOMETRY DATASETS

HDF5 file format
~20GB per file

Data types
- Image data
- Scalar data
- Meta data
- Logs

Features
- Zstd compression
- Fletcher32 checksums
- Dataset slicing
- Custom chunk sizes

Integrates well with all major programming lanugages.



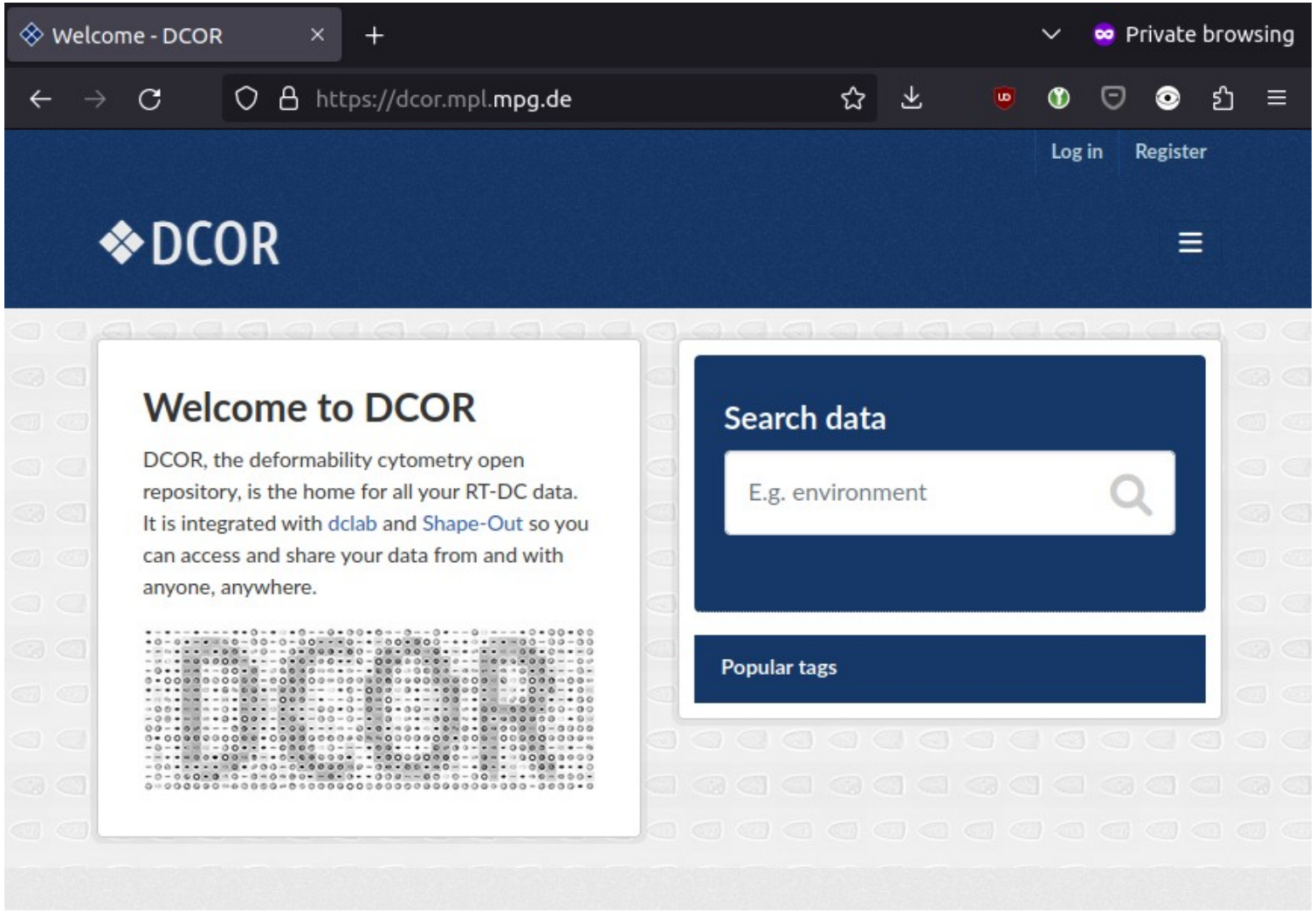
DATA STORAGE SOLUTION: DCOR

Motivation:

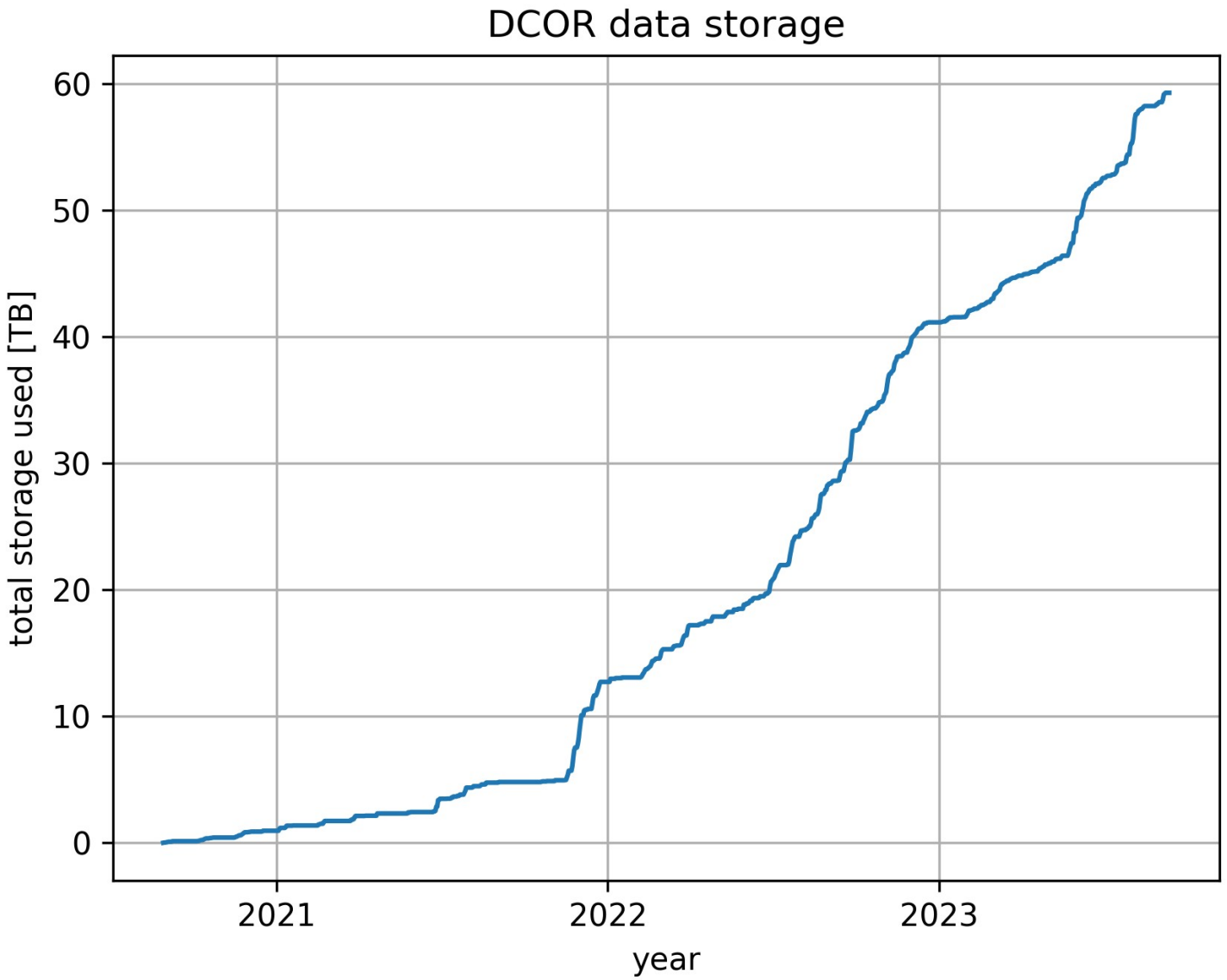
- data accessible from everywhere
- data upload from anywhere
- share with collaborators
- Immutable, citable datasets
- data access (slicing)

Implementation (MPCDF):

- Ubuntu 20.04
- CKAN with self-written plugins (authorization, visualization, data API, …) https://github.com/DCOR-dev
- S3 object storage (HPC-Cloud)
- S3 archived at MPCDF



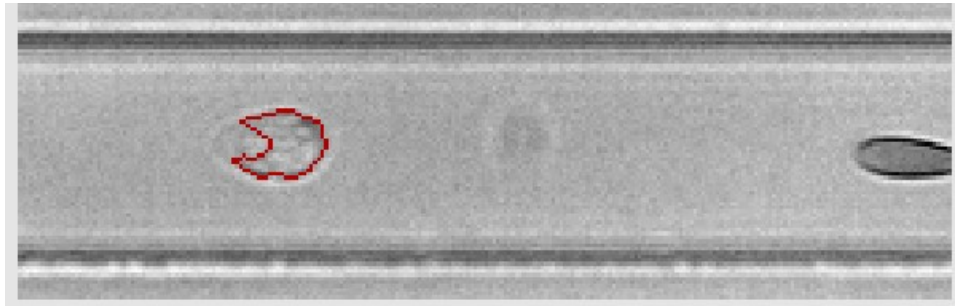
DATA STORAGE

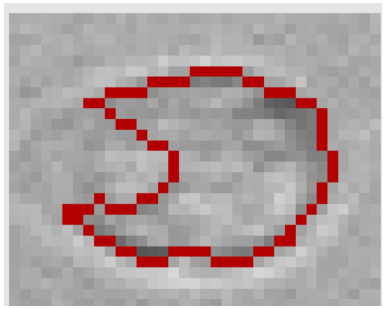


- Data stored on S3 (HPC-Cloud / OpenStack)
- pre-signed URLs for upload, download, or **partial access**
- DCOR-Aid GUI software for data management

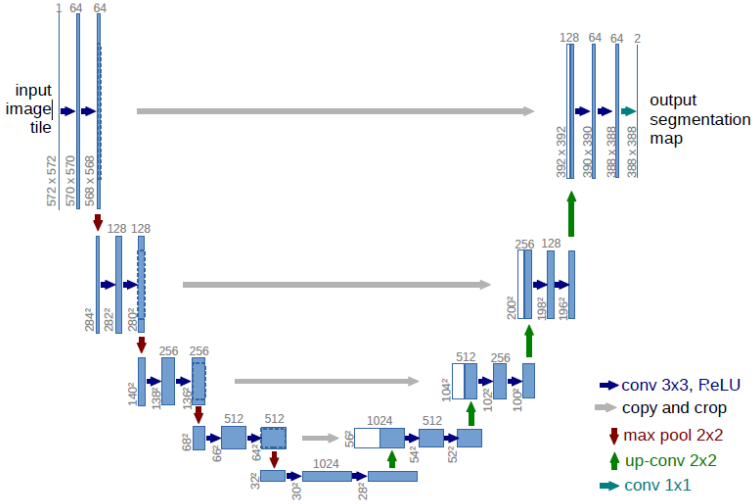
DATA POST-PROCESSING

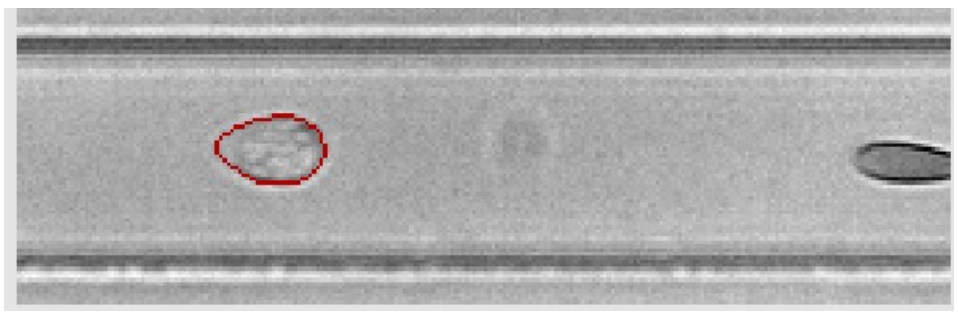
Threshold-based segmentation

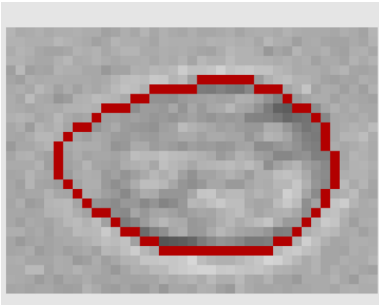


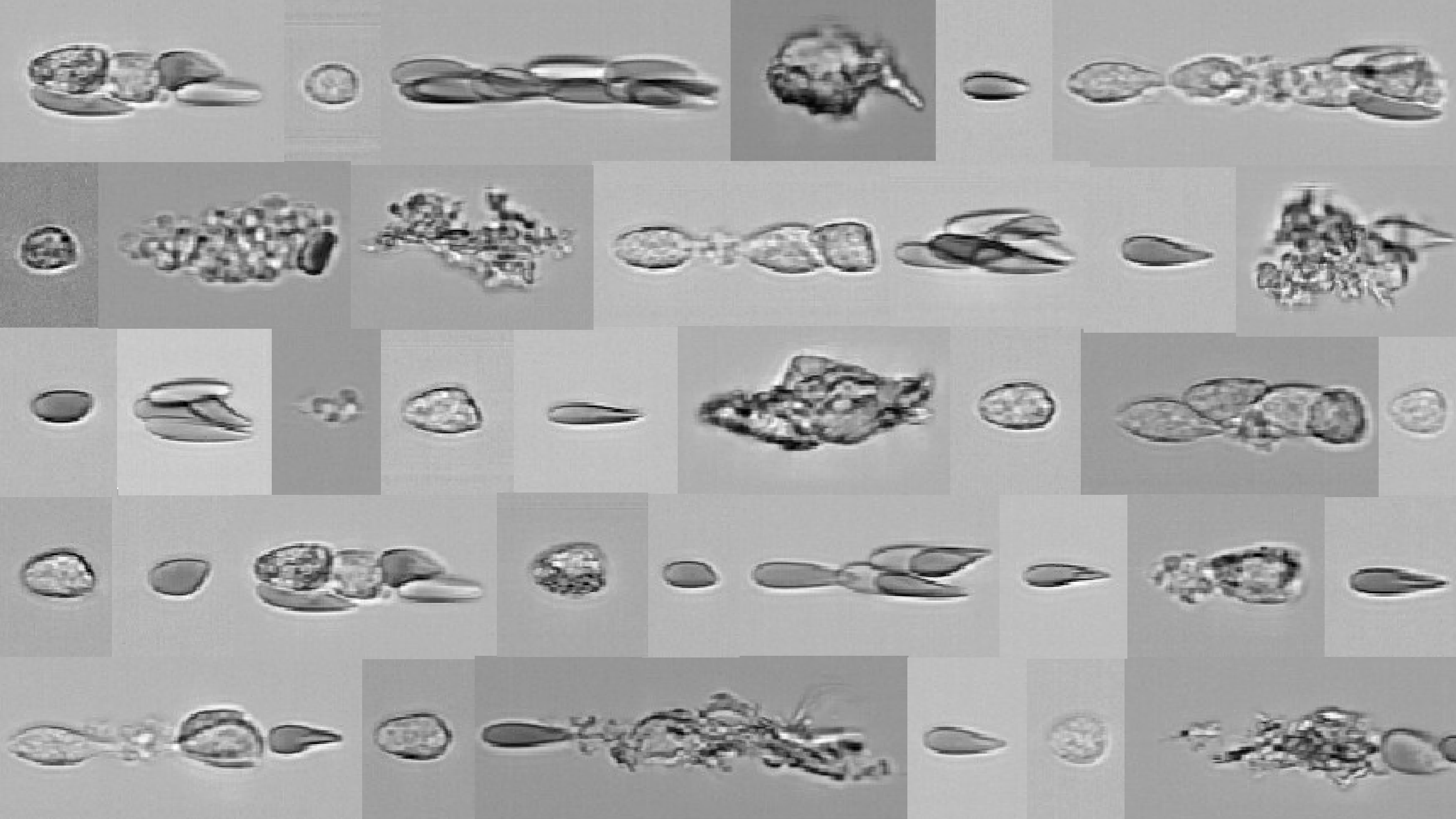


Machine-learning (UNET) segmentation

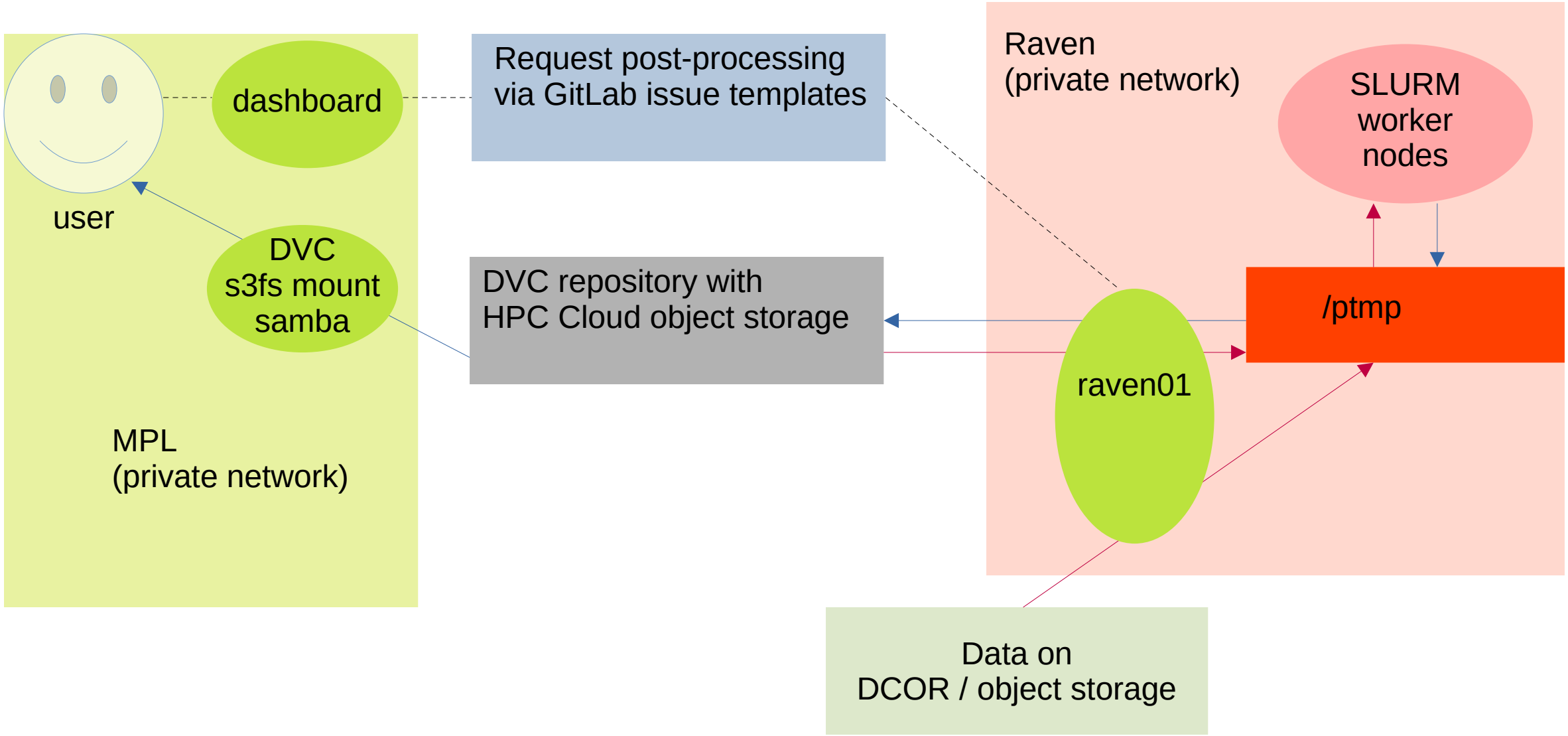




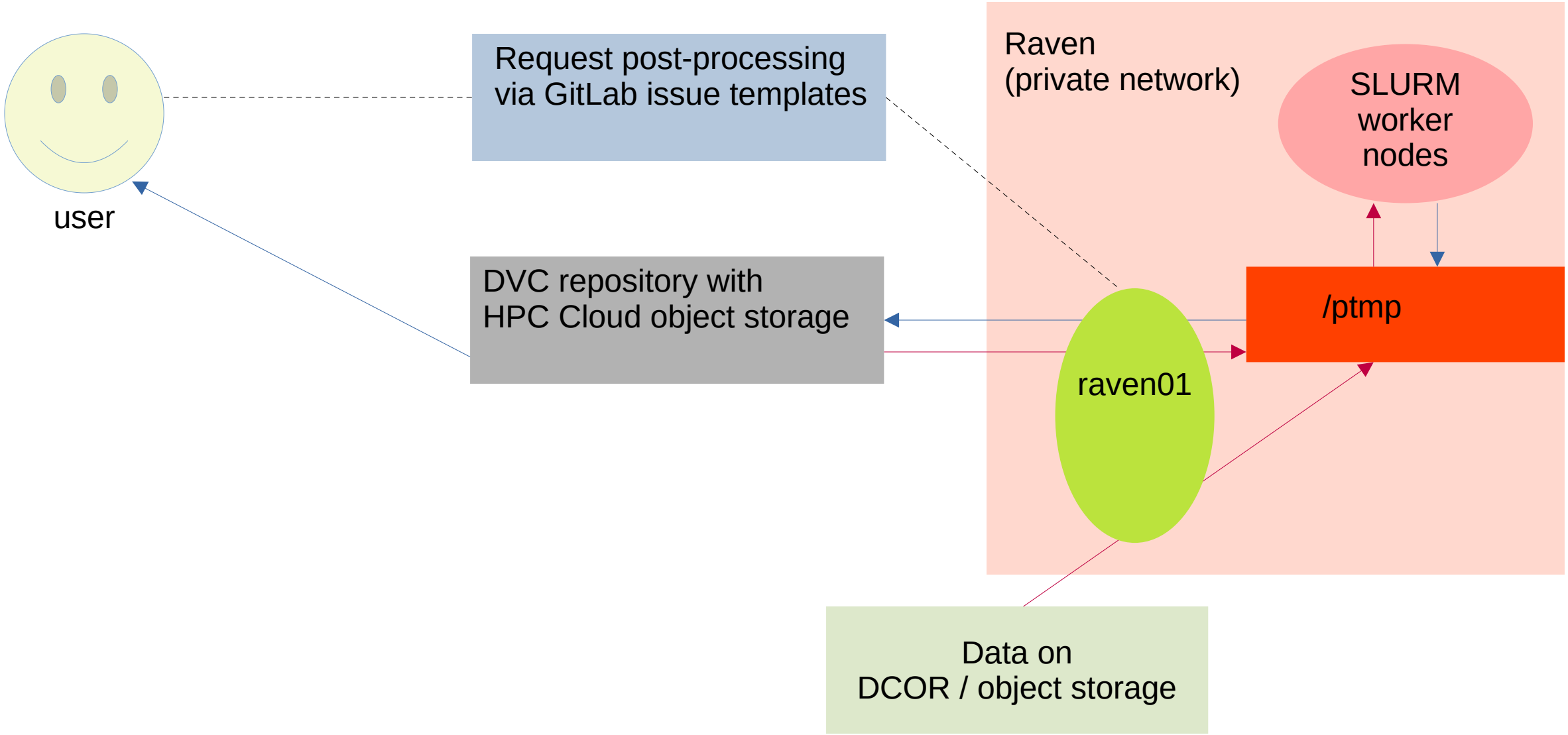




AUTOMATED DATA POST-PROCESSING AT MPCDF



AUTOMATED DATA POST-PROCESSING AT MPCDF



Thank You!

Guck Division
Benedikt Hartmann
Eoghan O'Connell
Felix Reichel
Jochen Guck
Lena Schnörer
Marta Urbanska
Martin Kräter
Maximilian Schlögel
Nadia Sbaa
Parth Patel
Raghava Alajangi
Salvatore Girardo
Sara Kaliman
Shada Abuhattum

paul.mueller@mpl.mpg.de

MPCDF
MAX PLANCK
COMPUTING & DATA FACILITY

Brian Standley
Florian Kaiser
Frank Berghaus
John Kennedy
Lorenz Huedepohl
Mykola Petrov
Nicolas Fabas
Raphael Ritz
Thomas Zastrow
Wolfgang Ryll