



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

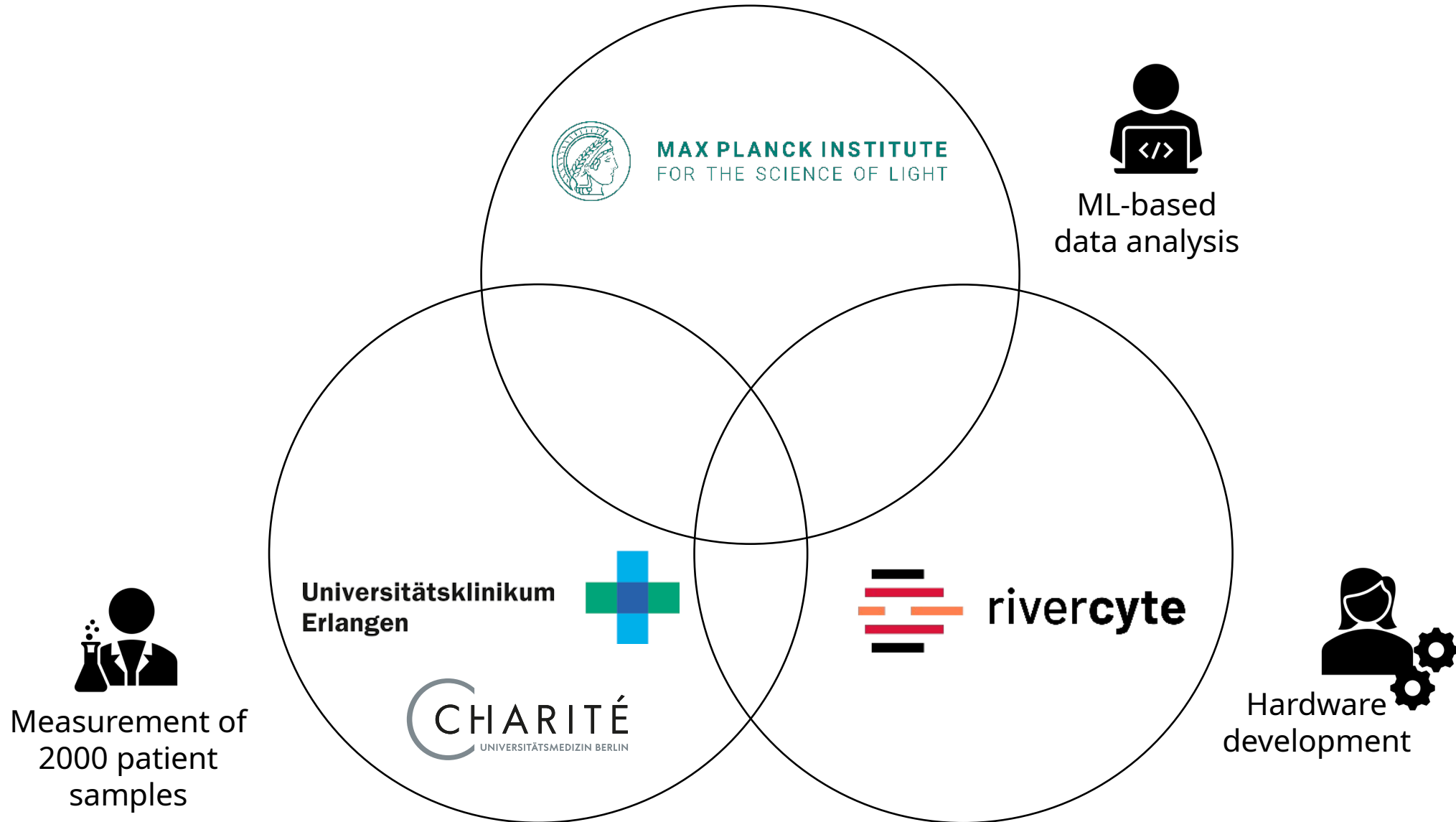


CONFLICTS OF INTEREST DECLARATION

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.

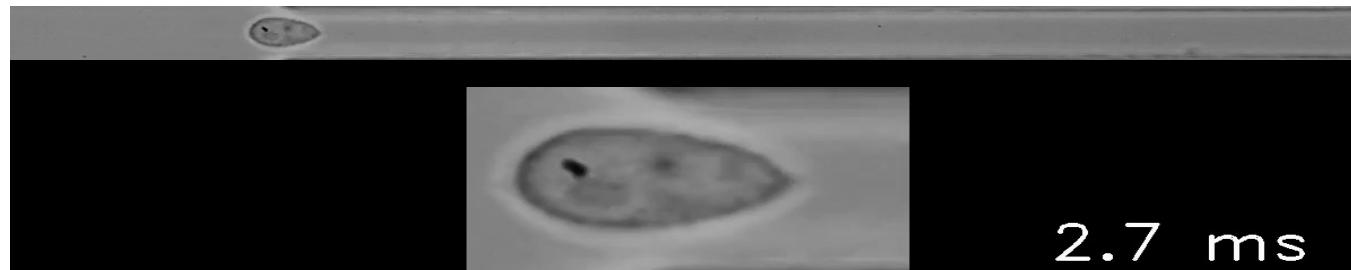
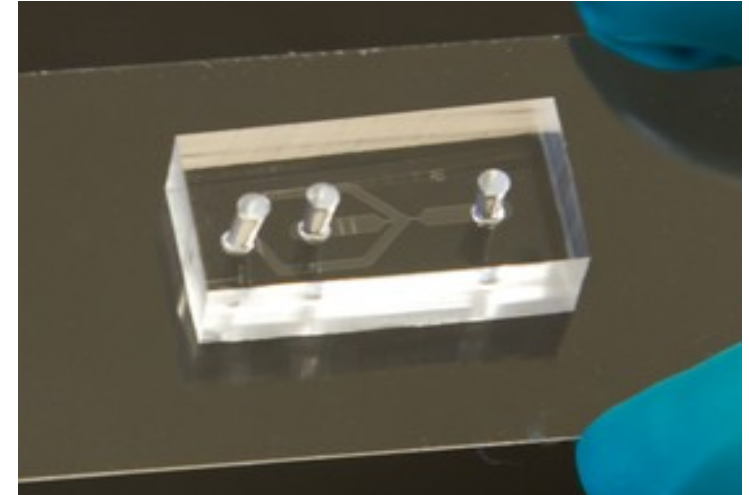
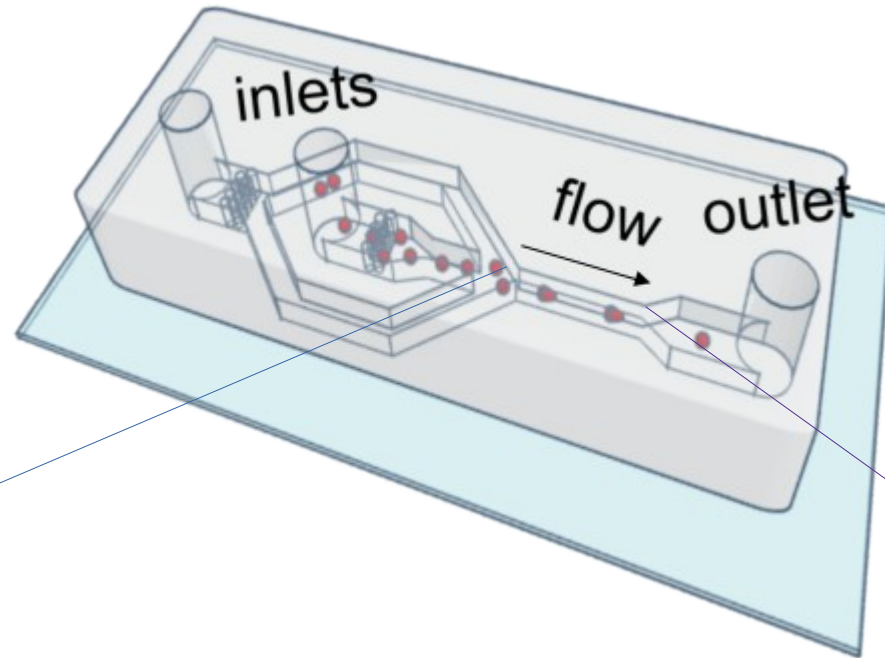


THE BIG PICTURE





DATA ACQUISITION



- 10 min measurement
- 1 000 000 cell images from 1 μ L 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 languages.

The screenshot displays the HDFView 3.3.0 application window. The main interface is divided into several sections:

- File Explorer:** Shows a tree view of the file structure. The selected file is `SO2-export_4_p1_mc_syto13_1.rtdc`, which contains an `events` sub-directory. Under `events`, there are several datasets: `area_cvx`, `area_msd`, `area_ratio`, `area_um`, `aspect`, `bright_avg`, `bright_sd`, `circ`, and `contour`.
- Object Attribute Info:** A panel on the right showing details for the selected object. It indicates that the `Attribute Creation Order` is `Creation Order NOT Tracked` and that the `Number of attributes = 0`. There are buttons for `Add Attribute` and `Delete Attribute`.
- Table View:** A window titled `aspect at /events/ [...]` displays a table of data. The table has columns for index, value, and another value. The data is as follows:

| 0 | 4.0 | |
|---|--------------------|--|
| 1 | 3.909090757369995 | |
| 2 | 3.615384578704834 | |
| 3 | 3.8181817531585693 | |
| 4 | 4.0 | |
| 5 | 3.909090757369995 | |
- Image View:** A window titled `image at /events/ [SO2-ex...]` displays a grayscale image of two elongated, teardrop-shaped objects. The image is shown with a color scale on the right, ranging from `0.00E0` to `9.00E1`. The image is currently showing a slice at index `2969` of a total of `2975` slices.



DATA STORAGE SOLUTION: DCOR

Motivation:

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

Implementation (MPCDF):

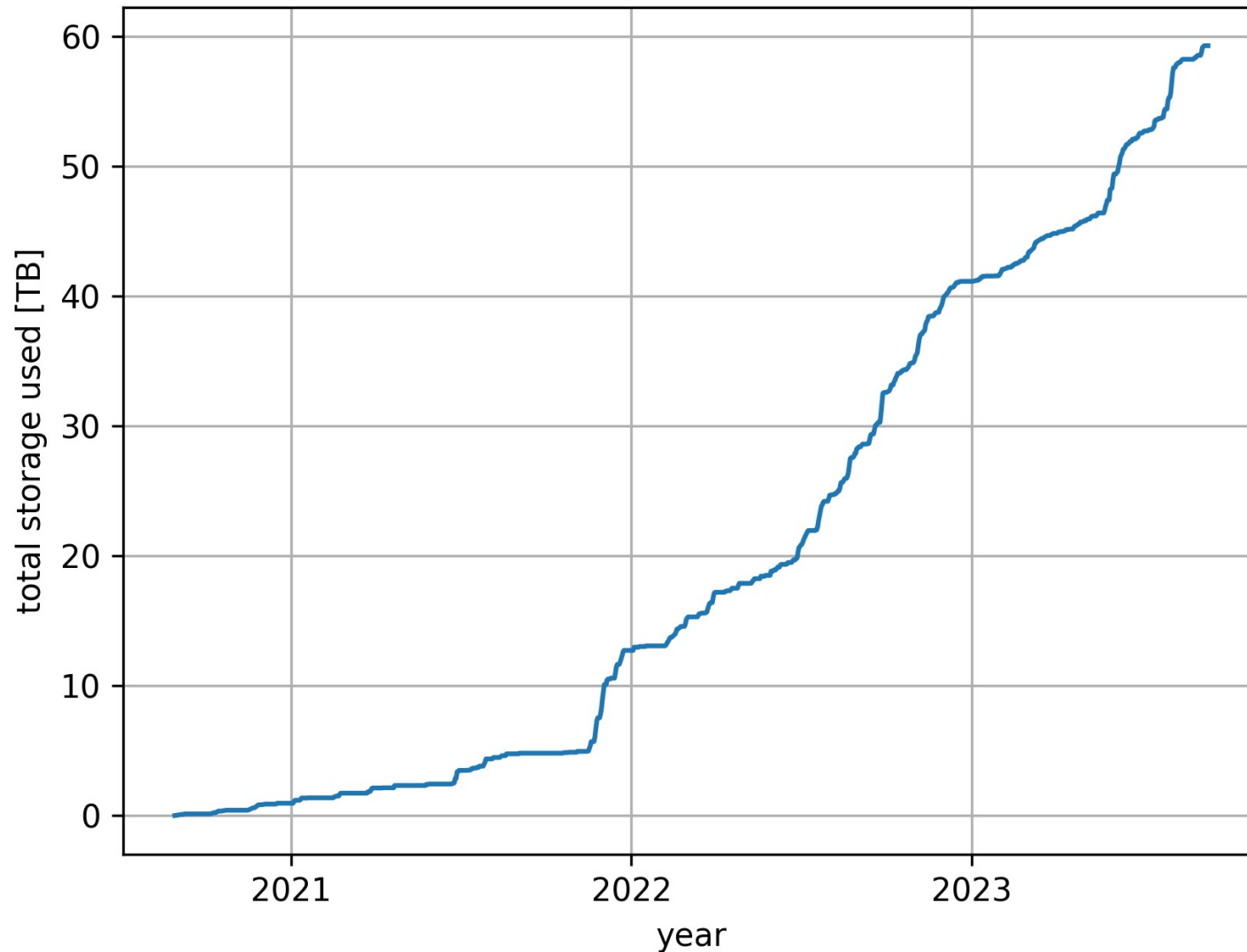
- Ubuntu 20.04
- CKAN with self-written plugins (authorization, visualization, data API, ...)
<https://github.com/DCOR-dev>
- S3 object storage (HPC-Cloud)
- S3 archived at MPCDF

The screenshot shows a web browser window with the address bar displaying `https://dcor.mpl.mpg.de`. The page has a dark blue header with the DCOR logo and navigation links for "Log in" and "Register". The main content area features a "Welcome to DCOR" message, a search bar with the placeholder text "E.g. environment", and a section for "Popular tags". The background of the page is decorated with a pattern of small, light blue icons.



DATA STORAGE

DCOR data storage

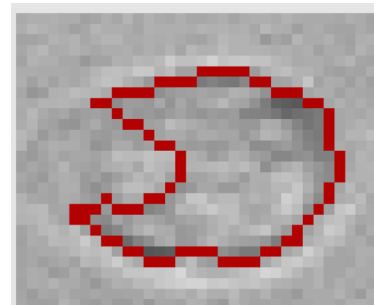
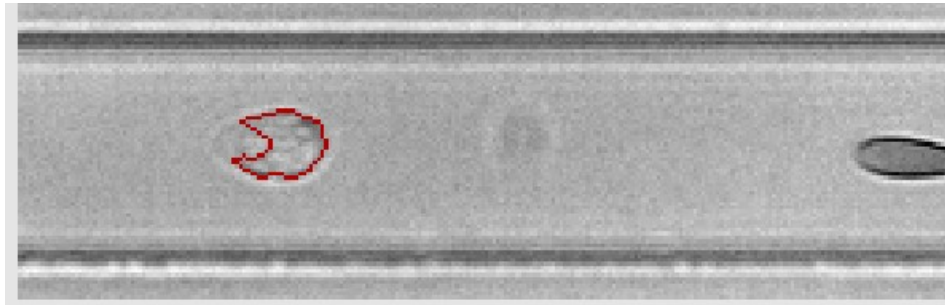


- Data stored on S3 (HPC-Cloud / OpenStack)
- pre-signed URLs for upload, download, or **partial access**
- DCOR-Aid GUI software for data management

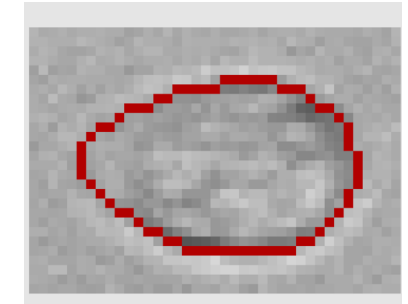
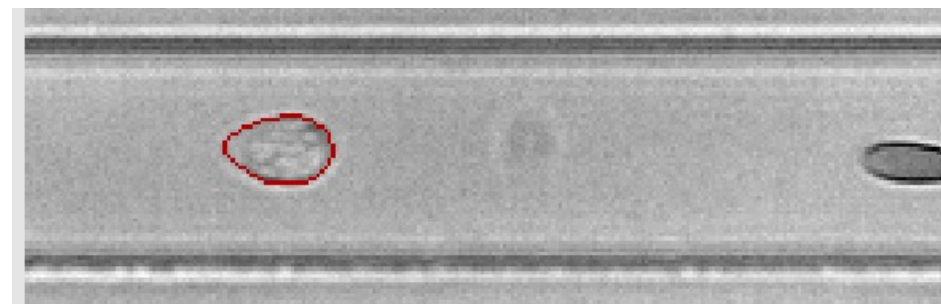
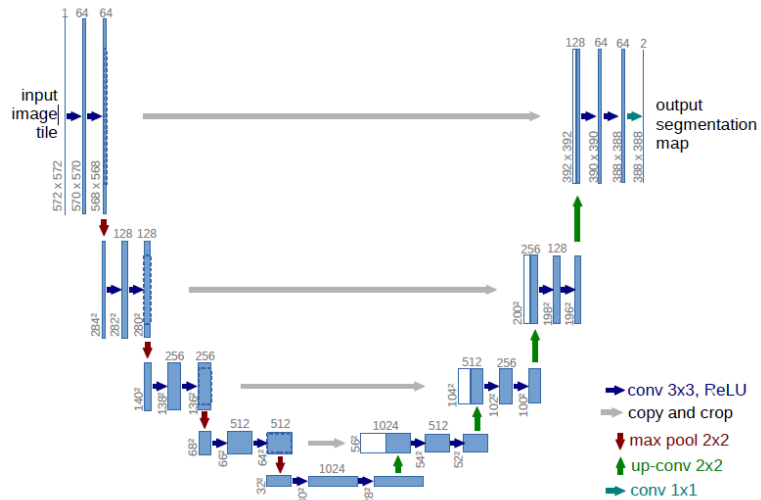


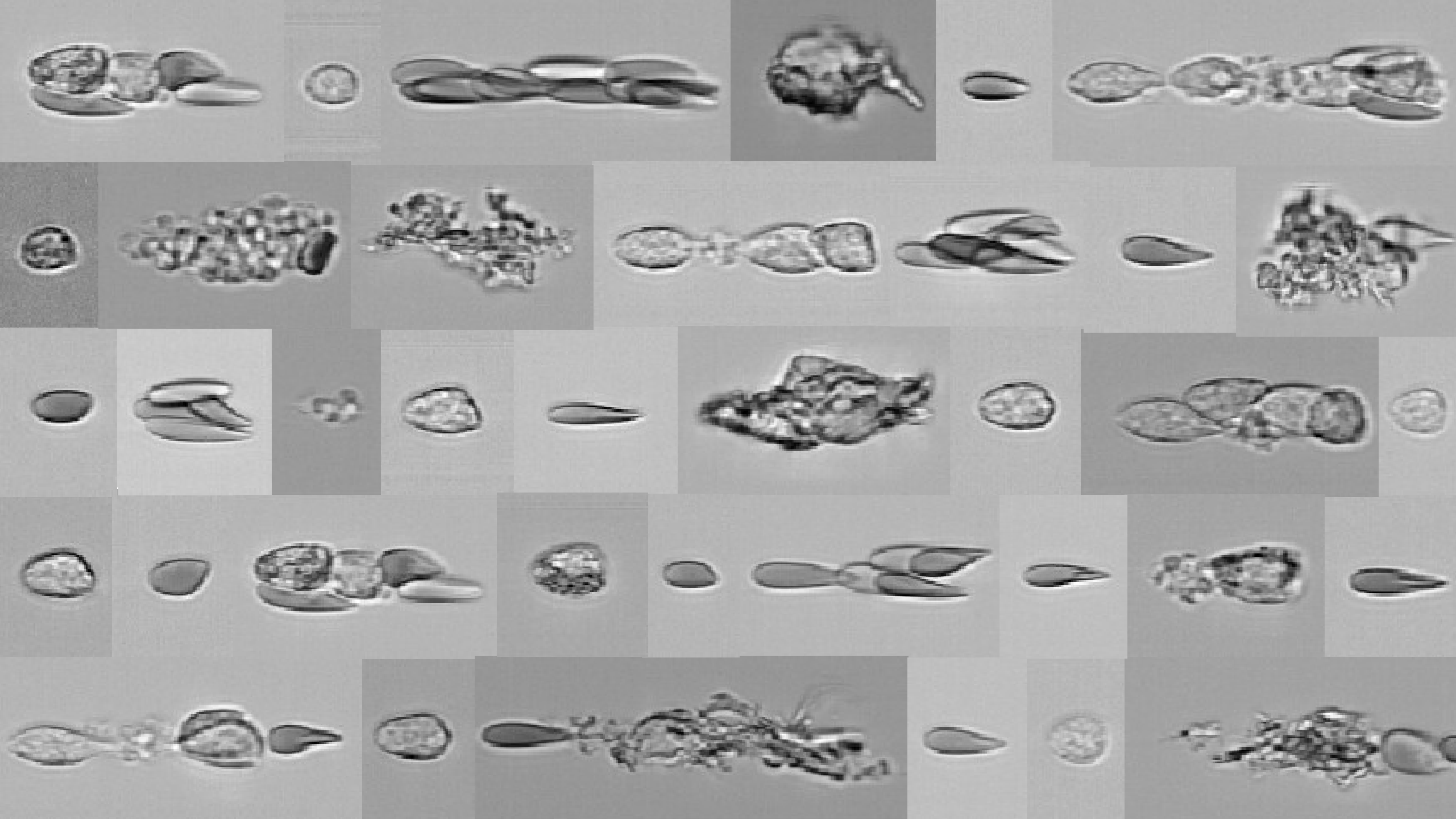
DATA POST-PROCESSING

Threshold-based segmentation



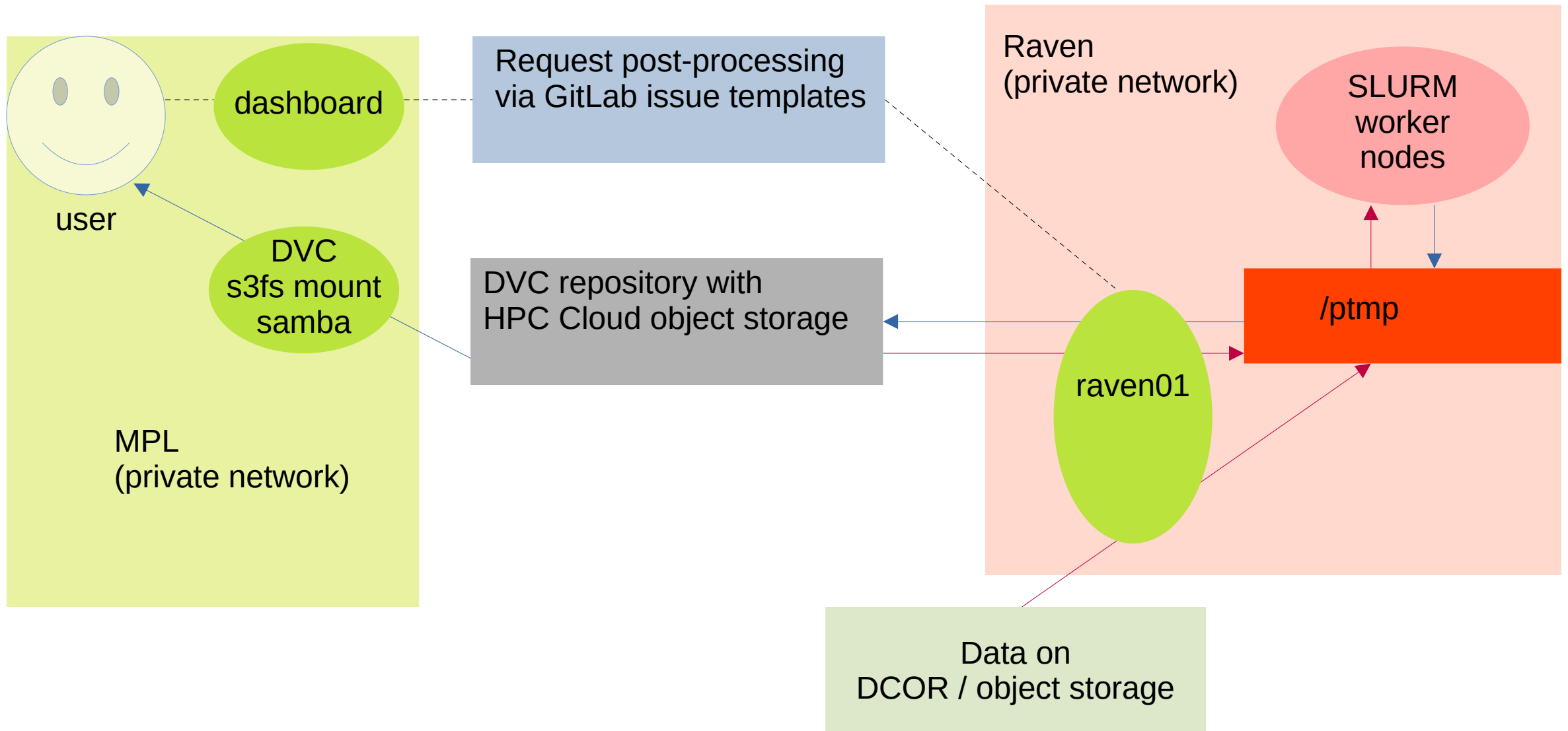
Machine-learning (UNET) segmentation





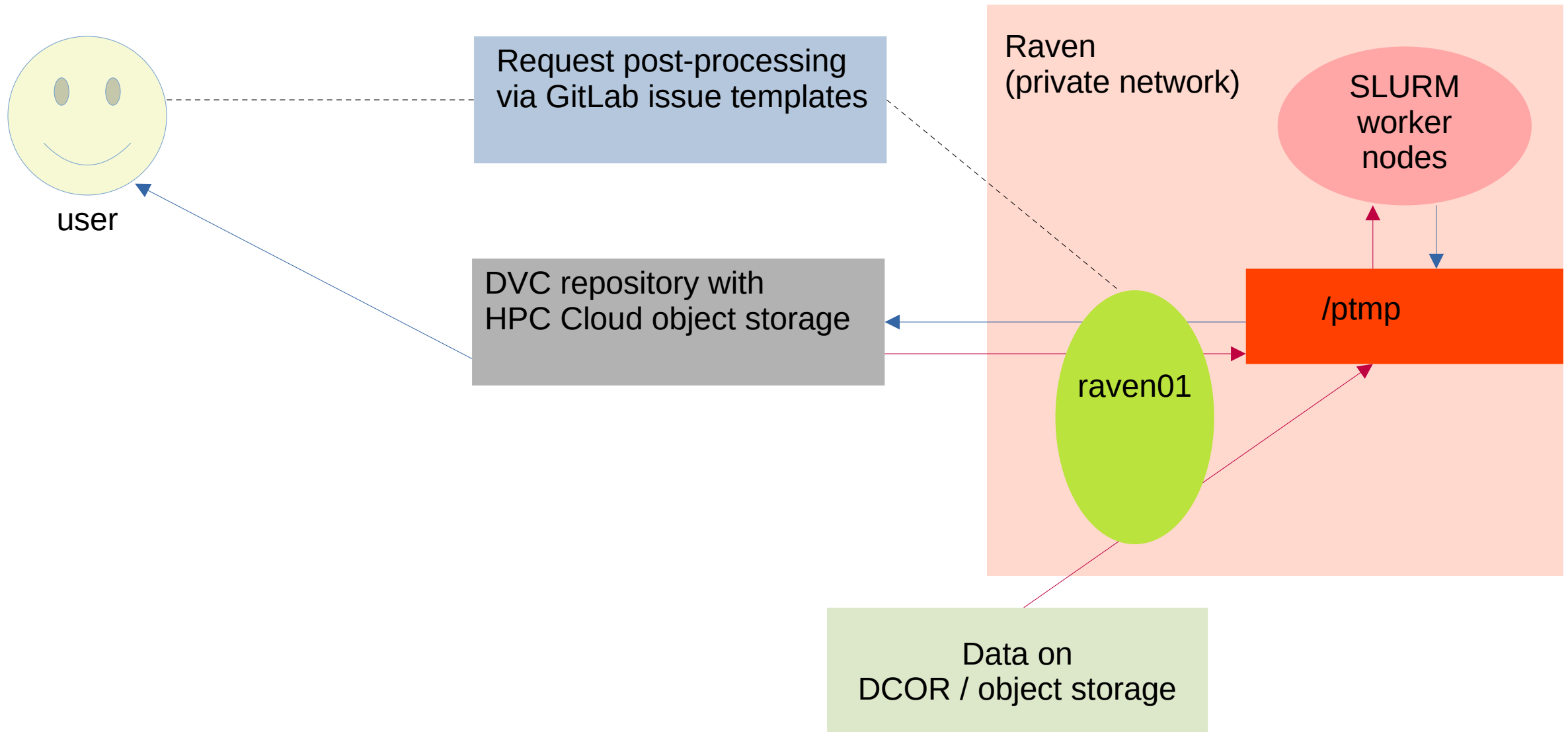


AUTOMATED DATA POST-PROCESSING AT MPCDF





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



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