Synchronizing the data science workflow with data management at scale

David R Stirling, Sébastien Besson, Emil Rozbicki
Glencoe Software Inc., Seattle, WA, USA

Background

- The vast quantities of data generated by scientists necessitate efficient management and analysis solutions to aid discovery and understanding of complex biological processes.
- OMERO is a powerful image data management platform developed by the Open Microscopy Environment (OME), specifically designed for handling life sciences data, including experimental metadata and analytics.
- Python has become a popular choice for data manipulation, analysis, and visualization due to its extensive catalogue of third-party libraries, many of which are tailored to the scientific community.
- This poster presents the omero2pandas package, designed to integrate with popular Python data science tools and enable analysis workflows which utilize tabular data stored on an OMERO server.

OMERO Plus

Enterprise image database for scientific images and associated metadata. Supports more than 150 bio-image formats and together with OME-NGFF provides first truly cloud native image data management solution. Tabular data is stored and accessed using the OMERO.tables ecosystem.

OMERO-Segmentation Connector

A powerful tool that integrates with popular deep learning models such as StarDist, Cellpose, and custom ONNX models to analyze brightfield and multiplexed digital pathology images. The flexible connector supports various computing environments, including local compute, High-Performance Computing (HPC), and cloud-based compute, ensuring scalability and adaptability to various research and clinical workflows.

PathViewer

A versatile digital pathology viewer designed for both pathologists and researchers, offering support for brightfield and multiplexed fluorescent data. The user-friendly platform enables efficient analysis and visualization of complex pathological samples, catering to the needs of modern research and clinical applications.

Aim

Bridge the gap between OMERO and data science with a seamless connection to Python data science tools.

omero2pandas

omero2pandas is an open-source Python library designed to streamline data retrieval and storage by converting OMERO.tables to Pandas DataFrames and vice versa. This will assist in analysing data stored on OMERO with the scientific Python stack.

Key features:

- Load OMERO.tables to DataFrame remotely

```
df = omero2pandas.read_table(file_id=402)
```

- Download the table for local use

```
omero2pandas.download_table(
    "/path/to/output.csv", file_id=2, chunk_size=1000)
```

- Upload a results table to OMERO

```
ann_id = omero2pandas.upload_table(
    my_data_frame, "Table Name", 142, "Image")
```

- Retrieve a list of columns from a remote table.

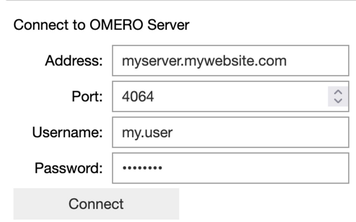
```
columns = omero2pandas.get_table_columns(
    annotation_id=142)
```

- Read specific rows and/or columns

```
my_dataframe = omero2pandas.read_table(
    file_id=10, column_names=['object', 'intensity'],
    rows=range(0, 100, 10))
```

- Server connection management, featuring automatic Jupyter environment detection and login widget

```
connector = omero2pandas.connect_to_omero(server="myserver.mywebsite.com")
```



- Connector also supports auth tokens generated with omero-user-token package.

Use Case: Data

Unsupervised clustering of Multiplexed Digital Pathology Data

Multiplexed digital pathology allows for the simultaneous visualization and analysis of multiple biomarkers within a single tissue sample, enabling a more comprehensive understanding of disease processes and patient-specific treatment strategies.

Goal:

Automatically group similar data patterns, such as cell types or tissue structures, without prior knowledge or labeled data, enabling the discovery of hidden relationships and structures within the tissue samples.

Data:

- Tonsil tissue sample: 13 x 9 mm at 20x magnification
- Fluorescent imaging of 19 biomarkers with RareCyte's Orion Platform
- Image file size: 50 GB. Stored on OMERO Plus server.

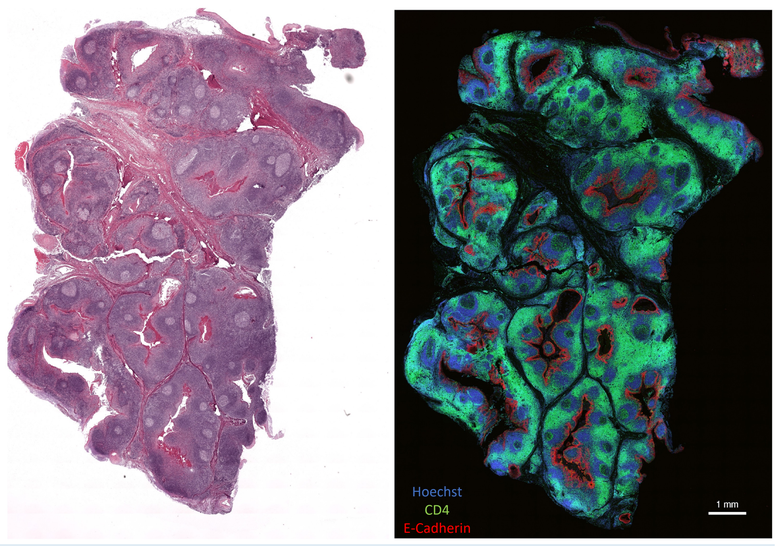


Figure 1: Left: vH&E whole slide image of a tonsil tissue sample, Right: Mutiplexed fluorescent imaging of the same sample. Sample was stained with 19 biomarkers and imaged with RareCyte's Orion Platform. 3 selected channels are shown for clarity.

- Image was segmented using Glencoe's OMERO-Segmentation Connector. The process identified 1.2 million objects and calculated more than 300 features describing each object.

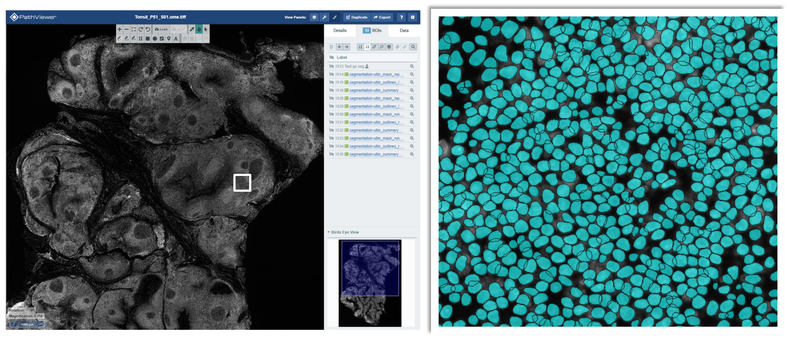


Figure 2: Left: Original image showing nuclear stain (Hoechst). Right: White rectangle region in the left panel showing segmentation results from the StarDist segmenter executed with Glencoe's OMERO-Segmentation Connector.

- Results stored online as an OMERO.table associated with the parent image.

	object	polygon_area	polygon_perimeter	polygon_longest_axis	polygon_convexity	polygon_circ_compact
0	1	41.077892	24.635889	8.168496	0.946690	0.783852
1	2	462.501984	77.326464	26.921430	0.997796	0.812508
2	3	395.988050	71.441465	24.117911	0.994416	0.866788
3	4	438.167280	75.089452	25.066534	0.995221	0.887895
4	5	647.472070	100.521910	40.676245	0.981223	0.498252
...	...	...	...	...	...	...
1201944	1201945	424.909953	74.126120	25.447321	0.994178	0.835455
1201945	1201946	408.398495	73.169765	25.063975	0.995308	0.827741
1201946	1201947	444.333152	76.978537	28.470617	0.994205	0.697951
1201947	1201948	280.224372	59.988894	19.621862	0.993733	0.926692
1201948	1201949	585.777897	101.598922	42.182618	0.946069	0.419156

1201949 rows × 306 columns

Table 1: Table showing the features calculated by OMERO-Segmentation Connector for the objects segmented with the StarDist model. Each row represents one object.

omero-user-token

Python package that creates long running user tokens for use with the OMERO API under non-interactive, headless conditions.

Create token:

```
omero_user_token set -s where.is.omero -u your.username --time_to_idle 0
```

Use the token:

```
omero_user_token get
```

```
from omero_user_token import getter
token = getter()
```

Use Case: Analysis

Unsupervised clustering of Multiplexed Digital Pathology Data

Unsupervised clustering pipeline:

- Load OMERO.table to Pandas DataFrame using omero2pandas
- Pre-process the data by replacing missing values and normalizing all features
- Remove strongly correlated features using Spearman correlation
- Run dimensionality reduction using UMAP
- Build an unsupervised clustering model (Birch) and classify the cells
- Sample representative cells and visualise by retrieving image data using the OMERO API.
- Save the result to OMERO.tables for visualization with PathViewer and further downstream workflows

Results:

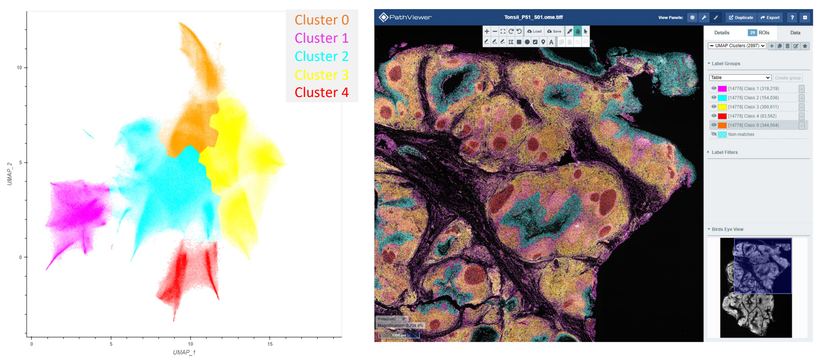


Figure 3: UMAP clustering of objects (left), identifying 5 primary clusters. (Right) original image with objects coloured according to cluster identity, revealing relative localization of objects from different clusters.

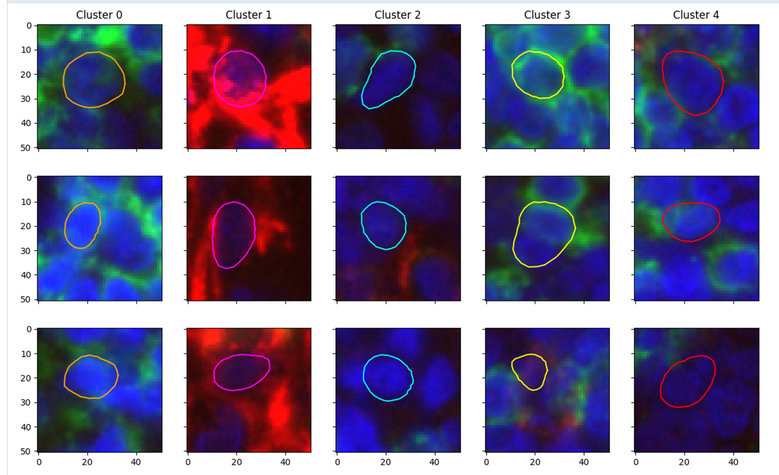


Figure 4: Representative cells sampled from the clusters identified by the UMAP. Blue=Hoechst, Green=CD4, Red=E-Cadherin. Individual cluster staining is likely to also differ in additional channels which are not displayed in these previews.

Jupyter Notebook available at:

www.github.com/glencoesoftware/webinar-notebooks

Conclusions

- The omero2pandas package seamlessly integrates the OMERO data management platform with popular Python data science tools, including Jupyter.
- The omero2pandas API provides a simple interface for manipulating tabular data without needing knowledge of the full omero-py API.
- Dedicated data management for analytical results in a multiuser collaborative environment fosters improved sharing and accessibility of data among researchers.
- Integrated applications, such as PathViewer, enable remote visualization of complex data, supporting diverse research and clinical workflows.

Future work:

- Enhance the backend performance of OMERO.tables to improve scalability and efficiency, especially when handling large datasets.
- Add support for OMERO.tables queries, enabling researchers to perform more advanced data filtering when requesting data with omero2pandas.

omero2pandas | omero-user-token | OMERO Plus