LabCAS: An integrated data-intensive environment for machine learning, archiving, processing, AI-based analyzing, and sharing data

Daniel Crichton[1], David Liu[1], Asitang Mishra[1], Heather Kincaid[1], Sean Kelly[1], Ashish Mahabal, PhD[2], Alphan Altinok, PhD[1], Chris Amos, PhD[3], Kristen Anton[4], Maureen Ryan[4], Christos Patriotis, PhD[5], Sudhir Srivastava, PhD, MPH[5]

[1]NASA Net Propulsion Laboratory, California Institute of Technology [2]California Institute of Technology [3]Baylor College of Medicine [4]University of North Carolina, Chapel Hill [5]National Cancer Institute

LabCAS ("Laboratory Catalog and Archive Service") is a web-enabled environment that provides a comprehensive suite of services for managing scientific data sets captured in biomedical research through its full lifecycle, supporting both pre-publication restricted access and post-publication open access. It is especially suited for capturing data for training machine learning (ML) algorithms. The LabCAS architecture is composed of a front-end web portal, where users can login to browse, inspect, visualize and download data; and a back-end software stack that exposes a rich set of data and metadata APIs for programmatic client access.

The LabCAS front-end allows for external tool integration to provide visualization and other analytical tools to be plugged-in. Currently in LabCAS, image files such as DICOM, SVS, SCN, and other formats can be visualized in an integrated viewer (the Digital Slide Archive or DSA) providing seamless visualization and interactivity. One-click downloading enables easy integration into ML and AI pipelines.

Additionally, LabCAS can be used to execute data intensive processing pipelines as structured workflows with scalable computation on the cloud. It also provides customizable user input, automatic publication of output with controlled access and allows for repeatable and reproducible results. As part of the scalable processing pipelines, LabCAS also offers an expanding array of ML models, complemented by recent foundation models such as SAM (Segment Anything). These models are seamlessly executable within the LabCAS user interface, compatible with a diverse range of available datasets. Furthermore, they are conveniently accessible externally through a REST API, thereby facilitating their application to novel datasets.

This poster will present a general overview of LabCAS, its application into ML and AI, the open-source image viewers, and examples of existing machine learning algorithms plugged into a processing pipeline as a structured workflow.