Librarian or Researcher

Source

I have some scanned documents on my laptop, some materials from the web and collection materials from SDR. They are all in different formats.

Data selection

My datasets are really large and, for this project, I need to select a subset of the data.

Data storage and processing

I need a place to build my corpus and do some pre-processing of the data.

Access to GPUs local and cloud

I need to bring the data together with machine learning models, python libraries to build, run and test models in a collaborative space.

Web server

I need to quickly and easily publish a web app as a verification/labelling tool and/or to show visualizations, documentation, etc.

Preservation and reuse

The final outputs of this project need to be available for download and citable.

I want to be able to use the same data set and models again for the next project.

User-facing

Google Drive

globus

REDIVIS

jupyter

Google Cloud

OpenAI

aws

Streamlit

SDR

Data Catalog

AI Service infrastructure

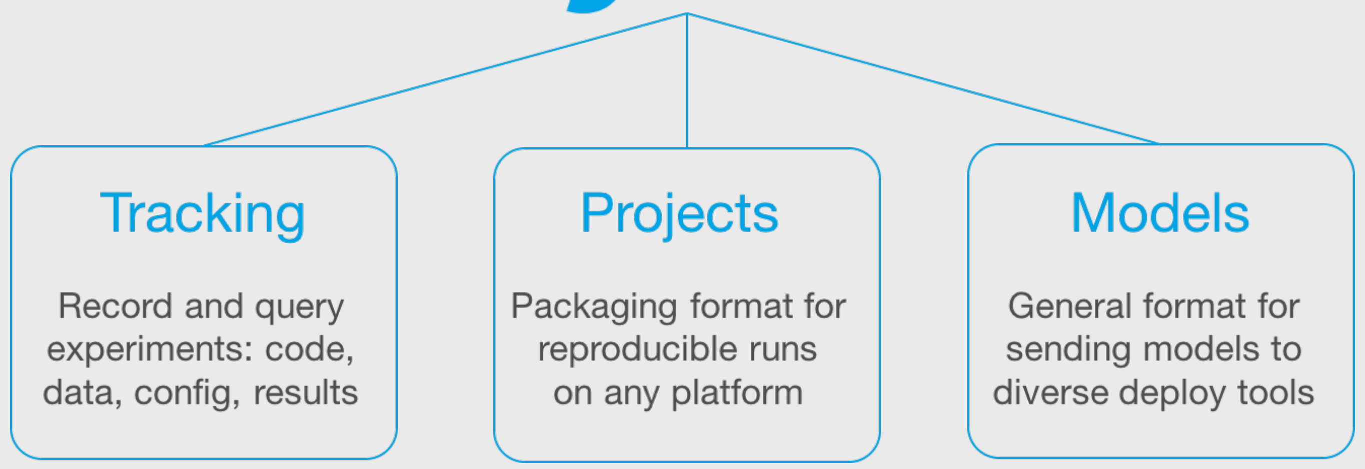
iiif

Apache Airflow

Data catalog and models management

mlflow

INTAKE



Tracking

Record and query experiments: code, data, config, results

Projects

Packaging format for reproducible runs on any platform

Models

General format for sending models to diverse deploy tools

?

Solr