LA-UR-24-27102

Approved for public release; distribution is unlimited.

Title:	Enabling ML workflows using Data Science Infrastructure
Author(s):	Pathak, Aishani Jayant Banesh, Divya Pulido, Jesus Jr Ahrens, James Paul
Intended for:	Los Alamos National Laboratory (LANL) Student Symposium, 2024-07-30/2024-07-31 (Los Alamos, New Mexico, United States)
Issued:	2024-07-12 (Draft)

Los Alamos
NATIONAL LABORATORY

Enabling ML workflows using Data Science Infrastructure

Aishani Pathak[1,3], Divya Banesh[1], Jesus Pulido[1], James Ahrens[2] | (1) CCS-3, (2) NSEC, LANL; (3) Arizona State University

Motivation

Subject matter experts across LANL regularly run a multitude of simulations and experiments. In many instances, the results of such runs are not stored in an easily accessible and reusable manner. The Data Science Infrastructure (DSI) project aims to make data more readily available for scientists by enacting FAIR data principles while maintaining user and group restrictions. This is primarily achieved by extracting metadata (automated, user-defined, and resource-dependent) from scientific data and generating ensembles to manage suites of simulation or experimental results.

This metadata extraction provides a unique opportunity to develop machine learning (ML) capabilities as an additional component of the DSI framework. In this work, we present new capabilities that allow us to generalize and integrate ML functionality into the DSI framework for a multitude of ML-related tasks. Specifically, we provide capabilities to train, predict and evaluate different ML models and neural network results using metadata extracted with DSI and their associated scientific data.

The Data Science Infrastructure Project

The Data Science Infrastructure project aims to enable FAIR data principles for LANL data to make it more **f**indable**, a**ccessible**, i**nteroperable, and **r**eusable while maintaining LANL user and group permission restrictions. The DSI framework provides a standard way to store and access data and associated metadata to support evolving scientific goals. The framework provides methods to enable data sharing and collaboration across group members, and ensure data retention when scientists change projects or retire.

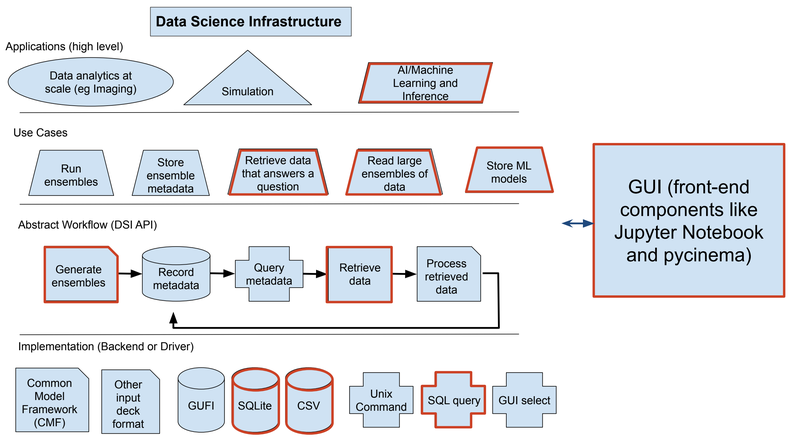


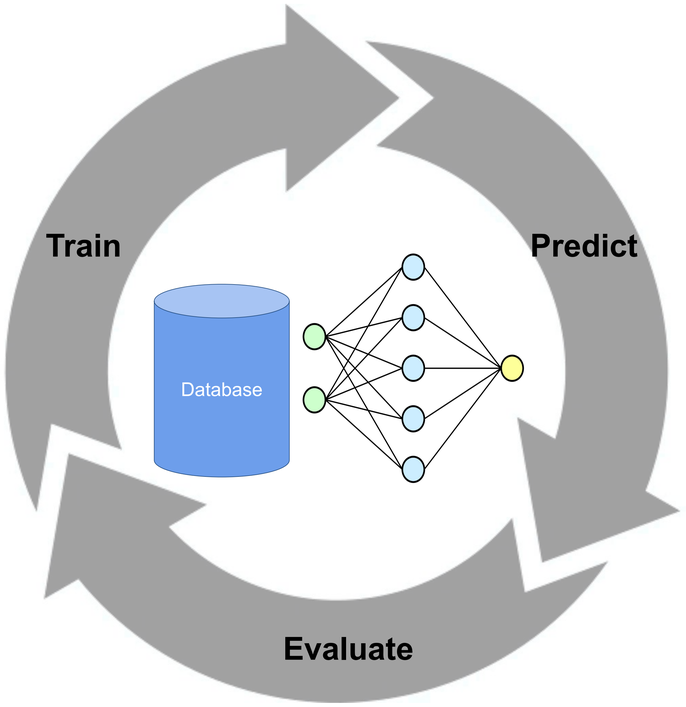
Figure 1. DSI Infrastructure - Multiple levels of the infrastructure allow us to link backend components to frontend components to analyze our results. Highlighted components include contributions of this project to the larger DSI goal

Machine Learning Workflows

As part of the DSI project, for any set of data, we store associated metadata that captures input/output parameters and associated values, details of compute resources used to generate the data, and any other information the scientist might consider valuable. Using this data, we can facilitate the training, prediction and evaluation of neural networks.

- Train: We train the model by utilizing data sourced from DSI backend databases. We also include the ability to update training using newly predicted data assuming they meet standards as defined by our evaluative process.
- Predict: Prediction capabilities allow the model to generate results based on new data inputs or test inputs, leveraging the model's learnt weights and insights
- Evaluate: Qualitative and quantitative evaluative metrics allow the user to determine the quality of their neural network predictions. We explore the following metrics:
  Quantitative: Inception Score, Frechet Inception Distance, SHAP values
  Qualitative: Subject matter expert evaluation through user studies
  Using front-end frameworks such as Jupyter Notebook or pycinema[1], we can visualize these metrics to evaluate additional training and prediction needs.

Metric	Usage	Ideal Value
Inception Score	Image Quality	Lower = Better
FID	Image Quality	Higher = Better



Results and Evaluation

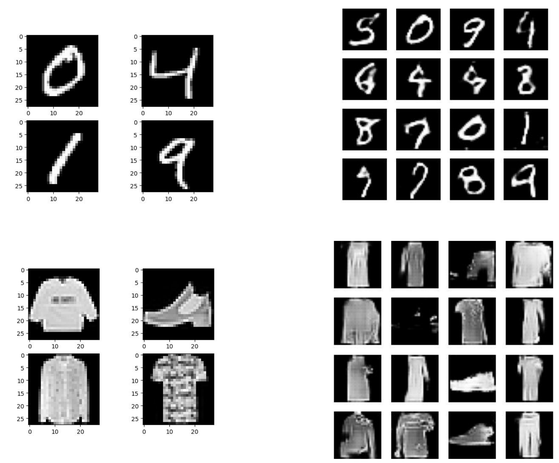
Workflows to support GANs

Our first DSI-based ML workflow supports a generalization of the Generative Adversarial Network (GAN). In this network, two neural networks, a generator and a discriminator contest with one another to form a zero-sum game.

- Train: We access data stored in back-end databases such as SQLite or CSV files through the DSI framework. Users can query for images and associated values from the database to train the network.
- Predict: Given a trained GAN network, we also provide functionality to predict new images for particular parameter combinations. These images and parametric values can be saved into the database for further analysis.
- Evaluate: Metrics such as the inception score and Frechet inception distance help determine if the results meet the scientist's standards for new data or if the model must be further tweaked for more accurate prediction.

Results

We test our DSI workflow capabilities on two sample datasets and associated networks: the MNIST and the Fashion MNIST data. Each sample dataset provide 60,000 training images that we use to train our Deep Convolutional GAN neural network. The MNIST-based GAN network was trained on 50 epochs and the Fashion MNIST-based GAN network was trained on 100 epochs. Predictions for both networks are shown below:



Citation: Cinemascience. "Cinemascience/Pycinema: Cinema Engine Toolkit." *GitHub*, github.com/cinemascience/pycinema. Accessed 10 July 2024.

Los Alamos NATIONAL LABORATORY
NNSA National Nuclear Security Administration
Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.
LA-UR# number goes here