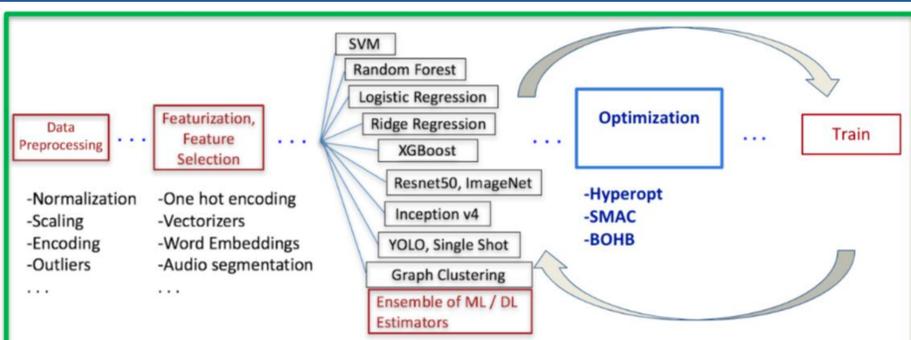


Automated Machine Learning (AutoML) as a Service for the Earth Sciences

Technical Lead: Brian Wilson

Co-Is: Alice Yepremyan, Diego Martinez, Sami Sahnoune, Edwin Goh, Sujen Shah, Kai Pak, Santiago Lombeyda, Chris Mattmann, and Wayne Burke
Jet Propulsion Laboratory / California Institute of Technology



Assembling an Optimal ML/DL Pipeline

Schematic of AutoML Pipeline Search: preprocessing, featurization/embeddings, feature selection, train an estimator, hyperparameter tuning, and rank by accuracy.

Abstract

As part of the **DARPA D3M program**, JPL is curating a library of ML/DL “primitives” (algorithms) with sufficient metadata and hyperparameter tuning hints to enable auto-assembly (in Python) of pipeline steps. These steps include preprocessing, feature extraction & selection, tuning an ensemble of models, ranking models using a metric, etc. The library contains 90+ classic ML algorithms from scikit-learn, pre-trained deep learning (DL) nets from Keras & PyTorch, and a set of advanced primitives from the D3M performer teams. JPL’s MARVIN tools provide an environment to annotate, discover, install, compose, and execute ML/DL primitives and pipelines. Pipelines and metadata are specified in a declarative manner using a community-defined JSON schema and taxonomy. MARVIN automates the creation of Docker containers containing the primitives and software dependencies, which are executed on a Kubernetes cluster either on premise or at any Cloud vendor supporting Kubernetes. D3M is designed to solve 15+ problem types:

- Classification, Regression, Clustering
- Image classification, object recognition
- Graph clustering/matching, Recommendations, Links
- Audio segmentation, video processing
- Time-series forecasting etc.

Exploring the library of ML algorithms, datasets/problems, pipelines

MARVIN enables an Automated ML environment similar to an “app store” in which a new “discoverable” ML/DL capability can be added by authoring a simple Python class satisfying the method interface, with tuning hints and a bit of metadata from the taxonomy. Currently, MARVIN contains 600+ datasets/problems, 330+ primitives, and 5 Million+ Pipeline Runs.

MARVIN

An Open Machine Learning Corp. Primitive Annotation and Execution Framework

Explore our metalearning database:

Datasets - 5165 results

Problems - 3939 results

Primitives - 2877 results

Pipeline Runs - 5362052 results

Other resources:

Docker Registry

Metalearning Information

The screenshots show the MARVIN web interface. The top section displays search results for 'Datasets', 'Problems', and 'Primitives'. The 'Datasets' section shows a search for 'World development indicators: Life expectancy prediction dataset'. The 'Problems' section shows a search for 'timeseries classification problem'. The 'Primitives' section shows a search for 'sklearn.neural_network.multilayer_perceptron.MLPClassifier'. The interface includes search bars, filters, and pagination controls.

5M+ Pipeline Runs

- Ranked by score
- Cmp pipelines
- Cmp key primitives
- Cmp hyperparameters
- Group by problem type & team
- Leaderboards

Future Work

- Inject Earth science remote sensing problems -- “phenomena recognition”, anomaly detection and time-series forecasting problems -- into the D3M program.
- Soliciting datasets and problems.
- Enable Meta-learning across 10M+ solution pipelines to be used for future model selection.
- GUI’s for domain experts: einblick.ai, Harvard Two Ravens, AutoML from Jupyter Notebooks

D3M Ecosystem and Community Outreach

datadrivendiscovery.org

