An Overview of the DARPA Data Driven Discovery of Models (D3M) Program*

Richard Lippmann, William Campbell, and Joseph Campbell
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420
{lippmann, wcampbell, jpc}@ll.mit.edu

Abstract

A new DARPA program called Data Driven Discovery of Models (D3M) aims to develop automated model discovery systems that can be used by researchers with specific subject matter expertise to create empirical models of real, complex processes. Two major goals of this program are to allow experts to create empirical models without the need for data scientists and to increase the productivity of data scientists via automation. Automated model discovery systems developed will be tested on real-world problems that progressively get harder during the course of the program. Toward the end of the program, problems will be both unsolved and underspecified in terms of data and desired outcomes. The program will emphasize creating and leveraging open source technology and architecture. Our presentation reviews the goals and structure of this program which will begin early in 2017. Although the deadline for submitting proposals has past, we welcome suggestions concerning challenge tasks, evaluations, or new open-source data sets to be included for system development and evaluation that would supplement data currently being curated from many sources.

1 Introduction Background

Understanding the complex and increasingly data-intensive world around us relies on the construction of robust empirical models of real, complex systems that enable decision makers to predict behaviors and answer "what-if" questions. Empirical models lie at the heart of much science (e.g., quantitative physics, material science, chemistry, biology/medicine, etc.). Basic research often aims to develop these models, which engineers and scientists can then use to develop new technologies (e.g., to fabricate new semiconductors or develop new sensors).

Today, construction of complex empirical models is largely a manual process requiring access to data scientists who perform many tasks including: 1) collaborate with subject matter experts to define a suitable modeling problem, 2) curate, select and annotate appropriate data, 3) transform, cleanse and structure data, 4) extract features from data, 5) model the data, and 6) visualize and explain the modeled outcomes. For any given empirical modeling problem, a team of subject matter experts and data scientists is typically required with expertise spanning all six of these areas to build a custom solution. There are currently not enough data scientists to investigate emerging data sources that could speed scientific discovery, improve human-computer interfaces, improve United States Government (USG) logistics and workforce management, and perform other important tasks.

*This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 Program Overview

An overview of the D3M program is available in [1]. It aims to develop automated model discovery systems that enable users with subject matter expertise but no data science background to create empirical models of real, complex processes. This capability will enable subject matter experts to create empirical models without the need for data scientists, and will increase the productivity of data scientists via automation. The automated model discovery systems developed by the D3M Program will be tested on real-world problems that will progressively get harder during the course of the program. Toward the end of the program, D3M will target problems that are both unsolved and underspecified in terms of data and instances of outcomes available for modeling.

Research will focus on three areas:

1. Create and provide a library of selectable primitives that serve as basic building blocks for complex modeline pipelines.
2. Automatically compose complex models from primitive building blocks. Select and compose building block primitives into complex modeling pipelines based on user-specified data and outcome(s) of interest.
3. Develop human-model interaction that supports creation of models by subject matter experts. A method and interface will be developed to facilitate human-model interaction that enables formal definition of modeling problems and curation of automatically constructed models by users who are not data scientists.

The program will be divided into two phases, of 24 months each, with evaluations structured around real world modeling problems that become increasingly more difficult as the program progresses. During the first phase of the program, performers will develop the capability to build models for a class of empirical science problems where complete data (but possibly including distractor variables) is given a priori (e.g., social and bio-science problems from repositories like dataverse, Kaggle-style business intelligence/machine learning problems, machine learning problems from OpenML). Each problem supplied during this phase will have prior expert-generated solutions. We will measure each system's ability to recover expert-generated solutions as part of annual performer evaluations. During the second phase, the program will work on problems that are both underspecified and unsolved (e.g., team formation in massive multi-player games, riot/disease outbreak prediction, political instability prediction, models of genetic factors for disease, etc.).

This program differs from past automated machine learning challenges such as [2] in that preprocessing including data cleaning and feature extraction must be automated, datasets are more complex and include many types of raw data, human-model interaction must be automated, and a wide variety of datasets will be included from many application areas.

This talk will provide an overview of the D3M program and goals. We will solicit suggestions concerning challenge tasks, evaluations, and new open-source datasets for system development and evaluation that supplement data currently being curated from sources including OpenML, Kaggle, Dataverse, Zenodo, and many sites with text, speech, image, video, medical, and time series challenges.

References

[1] Defense Advanced Research Projects Agency. Data-Driven Discovery of Models (D3M), June 2016. URL `https://www.fbo.gov/index?s=opportunity&mode=form&id=06049eb4ae38a68b7b8c54b94d9c7979&tab=core&_cview=1`.

[2] I. Guyon, I. Chaabane, H. J. Escalante, S. Escalera, D. Jajetic, J. R. Lloyd, N. Macía, B. Ray, L. Romaszko, M. Sebag, A. Statnikov, S. Treguer, and E. Viegas. A brief review of the ChaLearn AutoML challenge. In *Proc. of AutoML 2016@ICML*, 2016. URL `https://docs.google.com/a/chalearn.org/viewer?a=v&pid=sites&srcid=Y2hhbGVhcm4ub3JnfGF1dG9tbHxneDoyYThjZjhhNzRjMzI3MTg4`.