AFRL-RI-RS-TR-2021-076

DATA DRIVEN DISCOVERY OF MODELS

RAYTHEON BBN TECHNOLOGIES

APRIL 2021

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2021-076 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
MICHAEL J. MANNO
Work Unit Manager

/ S /
SCOTT D. PATRICK
Deputy Chief,
Intelligence Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
APRIL 2021	FINAL TECHNICAL REPORT	SEP 2017 – JUL 2020

4. TITLE AND SUBTITLE

DATA DRIVEN DISCOVERY OF MODELS

5a. CONTRACT NUMBER
FA8750-17-C-0176

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
62702E

6. AUTHOR(S)

Prasanna Muthukumar

5d. PROJECT NUMBER
D3ME

5e. TASK NUMBER
00

5f. WORK UNIT NUMBER
08

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Raytheon BBN Technologies
10 Moulton Street
Cambridge, MA 02138

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIED
525 Brooks Road
Rome NY, 13441-4505

DARPA/I2O
675 North Randolph Street
Arlington, VA 22203-2114

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2021-076

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

D3M 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. BBN focused on one aspect of the automation problem: Developing selectable primitives for time-series pattern discovery. BBN developed prototype primitives that are discoverable by other D3M performers who aim to automatically compose primitives into a model building pipe-line.

15. SUBJECT TERMS

Automated model discovery, primitives, time series, pattern discovery, model building pipeline

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**MICHAEL J. MANNO**
U	U	U	UU	11	19b. TELEPHONE NUMBER (*Include area code*) **N/A**

TABLE OF CONTENTS

1.0 SUMMARY

D3M 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 expert 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.

BBN focused on two aspects of the automation problem: (1) Developing selectable primitives for time-series pattern discovery; and, (2) Developing a method to compose complex models by converting primitives or primitive pipelines into Neural Nets (NN) to enable joint tuning and optimization.

BBN developed prototype primitives that are discoverable by other D3M performers (Technical Area 2) who aim to automatically compose primitives into a model building pipeline. Performance metrics include (1) the ability to identify meaningful units in time series data; and, (2) the ability to be automatically discovered by TA-2 performers' systems.

2.0 INTRODUCTION

The D3M program aims to provide a Subject Matter Expert (SME) the means to run powerful algorithms and extract results without incurring the expense of hiring a data analyst. Achieving this task requires a diverse set of primitives as well as the ability to compose these primitives into rapid pipelines.

Time-series problems are a particularly hard problem for SMEs because of the variable lengths of each data sample. These problems also tend to have longer-term and sparse patterns that are uncommon in other forms of data. Dealing with such problems requires specialized primitives carefully designed with time-series data in mind. BBN has used its decades-long expertise in time-series problems to contribute useful primitives that are extremely effective in this area.

When constructing pipelines by composing primitives, it is essential that the resulting pipelines be extremely fast. Not only does this enable faster prototyping and testing, but it also allows the production systems to unlock capabilities that they might not otherwise possess. With this view, BBN has demonstrated the ability of end-to-end neural networks to effectively mimic complex pipelines with several primitives. Neural networks run on specialized hardware; this hardware has been advancing at a rate that far outpaces the growth rate of the modern general-purpose processor. Therefore, this idea is likely to reap vast benefits in the future with progressively increasing returns as hardware gets faster.

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

For Technical Area 1 (TA1), a diverse primitive set that handles a variety of data types is necessary for successful automatic discovery of data driven models. Our proposal focused on time series datasets and involves tokenization of time series data and structure discovery, which provide powerful tools for building models on diverse sets of time series data, such as salary data over time, audio, video, handwriting, traffic, finance etc. Primitives needed for speech recognition have been included as part of this primitive set. These by themselves are very useful for supervised time series tokenization. Very few ML primitives focus on time series especially those that vary in duration. Capturing time-series patterns will greatly enrich the set of problems solvable by auto-ML. A full list of primitives is available in the Results and Discussion section.

For Technical Area 2 (TA2), we demonstrated the possibility of converting the best pipeline models discovered by the general approach to neural networks (NN), which will enable joint model fine-tuning. Pipelines constructed by the various TA2 processes tend to be complex, unwieldy, and slow despite the fact that they show excellent performance. These pipelines can be sped up by converting these pipelines into stand-alone end-to-end neural networks that can make use of specialized hardware. This conversion is achieved through student-teacher training where a neural network is trained on the outputs predicted by the original pipeline. BBN demonstrated this conversion successfully on a random forest pipeline. This achievement is particularly significant as random forests are well-known for being an incredibly strong baseline when the machine learning algorithm is not heavily tuned. More details on this technique are available in the paper that is referenced at the end of this report.

4.0 RESULTS AND DISCUSSION

- **Completed**: Delivered a wide variety of primitives, compatible with the D3M core package, that address various aspects of time series problems
- **Completed:** Delivered two time-series datasets, FMA and Språkbanken.
- **Completed:** Demonstrated potential of neural network conversion techniques to speed up existing ML pipelines

TA1:

Delivered ML primitives:

- MLP classifier
- Cluster curve fitting
- I-vector extractor
- Segment curve fitter
- Sequence to Bag of Tokens
- Mel Frequency Cepstral Coefficients
- TF-IDF transformer
- Uniform segmenter

Delivered glue primitives:

- Audio reader
- Channel average
- CSV reader
- Signal dither
- Targets reader
- Signal framer

Delivered two datasets:

- Free Music Archive (FMA)
- Språkbanken

TA2:

- Demonstrated potential of converting a random forest pipeline to an end-to-end neural network
- Published paper on this approach

BBN specific TA2 work put on hold per DARPA instructions in October 2019.

5.0 CONCLUSIONS

BBN has contributed to the TA1 and TA2 parts of the D3M program. On TA1, we provided a diverse set of ML primitives that address a variety of time-series problems. On TA2, we demonstrated the ability of a neural network to mimic and significantly speed-up computation of machine learning pipelines.

APPENDIX A – PUBLICATIONS AND PRESENTATIONS

List the dates, times, title, event and speakers of any presentations made under this effort and the title author and publication information for any publication made under this effort.

Sung, ML., Silovsky, J., Siu,MH., Gish,H., Pittapally, C., "Neural Network Conversion of Machine Learning Pipelines", ICML AutoML Workshop (2018)

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

(CDS) Cross Domain System or Solution

(BBN) Bolt, Beranek, and Newman

(SME) Subject Matter Expert

(D3M) Data-Driven Discovery of Models

(MLP) Multi-Layer Perceptron

(NN) Neural Network

(CSV) Comma Separated Value

(TF-IDF) Term Frequency Inverse Document Frequency

(TA) Technical Area

(ML) Machine Learning