



AFRL-RI-RS-TR-2021-209

DARPA DATA-DRIVEN DISCOVERY OF MODELS - A SUMMARY OF PARENTHETIC'S WORK AND ACTIVITIES

PARENTHETIC, LLC

DECEMBER 2021

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

**AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE**

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 was cleared for public release by the 88th ABW, Wright-Patterson AFB Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (<http://www.dtic.mil>).

AFRL-RI-RS-TR-2021-209 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /

PETER A. JEDRYSIK
Work Unit Manager

/ S /

JULIE BRICHACEK
Chief, Information 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

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE DECEMBER 2021	2. REPORT TYPE FINAL TECHNICAL REPORT	3. DATES COVERED <table style="width: 100%; border: none;"> <tr> <td style="width: 50%; border: none;">START DATE SEPTEMBER 2019</td> <td style="width: 50%; border: none;">END DATE MAY 2021</td> </tr> </table>		START DATE SEPTEMBER 2019	END DATE MAY 2021	
START DATE SEPTEMBER 2019	END DATE MAY 2021					
4. TITLE AND SUBTITLE DARPA DATA-DRIVEN DISCOVERY OF MODELS - A SUMMARY OF PARENTHETIC'S WORK AND ACTIVITIES						
5a. CONTRACT NUMBER FA8750-19-C-1522	5b. GRANT NUMBER N/A	5c. PROGRAM ELEMENT NUMBER 62702E				
5d. PROJECT NUMBER D3MP	5e. TASK NUMBER S0	5f. WORK UNIT NUMBER 13				
6. AUTHOR(S) Michael Liang, Reva Schwartz, and Catherine Chapman						
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Parenthetic, LLC 6023 Woodmont Rd. Alexandria VA 22307			8. PERFORMING ORGANIZATION REPORT NUMBER			
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory/RISB 525 Brooks Road Rome NY 13441-4505		10. SPONSOR/MONITOR'S ACRONYM(S) RI	11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-RI-RS-TR-2021-209			
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited. PA# AFRL-2021-3670 Date Cleared: 20 Oct 2021						
13. SUPPLEMENTARY NOTES						
14. ABSTRACT On the DARPA Data-Driven Discovery of Models (D3M) program, Parenthetic evaluated the ability of various software systems which aimed to empower users with little to no data science background in creating empirical models of real, complex processes. Parenthetic developed an evaluation framework in which the systems were assessed across a number of key areas such as Explainability and Meaningfulness to ensure that users were not only able to create their own data-driven models, but were able to understand their models' outputs as well. In the final phases of the program, Parenthetic facilitated transition activities including information exchanges and technology demonstrations with several United States Government agencies						
15. SUBJECT TERMS Machine Learning; Artificial Intelligence; Software Evaluation; Models; Data Science; User Testing; Analytics; Automation; Program Evaluation						
16. SECURITY CLASSIFICATION OF: <table style="width: 100%; border: none;"> <tr> <td style="width: 33%; border: none;">a. REPORT U</td> <td style="width: 33%; border: none;">b. ABSTRACT U</td> <td style="width: 33%; border: none;">c. THIS PAGE U</td> </tr> </table>			a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	17. LIMITATION OF ABSTRACT SAR
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U				
18. NUMBER OF PAGES 155						
19a. NAME OF RESPONSIBLE PERSON PETER A. JEDRYSIK			19b. PHONE NUMBER (Include area code) N/A			

TABLE OF CONTENTS

Section	Page
LIST OF FIGURES	ii
LIST OF TABLES	ii
1.0 SUMMARY	1
2.0 INTRODUCTION	3
3.0 METHODS, ASSUMPTIONS, AND PROCEDURES	4
3.1 Evaluations	5
3.1.1 Fall 2019 Evaluation	5
3.1.2 Summer 2020 Evaluation	7
3.2 Transition Partners Engagements and Other Activities	10
3.2.1 Transition Partners and Stakeholders	12
3.2.2 Transition Partner Use Cases	12
3.2.3 Glossary of Machine Learning Terms	15
3.2.4 Transition Partner Experiment Dry-Run	20
4.0 RESULTS AND DISCUSSION	22
4.1 Fall 2019 Evaluation Results	22
4.1.1 Fall 2019 Evaluation Outcomes	23
4.2 Summer 2020 Evaluation Results	24
4.3 Summary of Evaluation Findings	26
4.4 Transition Partner Experiment Dry-Run Results	28
5.0 CONCLUSIONS	29
6.0 APPENDIX	30
6.1 Fall 2019 Bi-Annual Evaluation Brief	31
6.2 Summer 2020 Bi-Annual Evaluation Brief	110
7.0 LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS	150

LIST OF FIGURES

Figure	Page
Figure 1. Transition Partner Engagement Process	10
Figure 2. Transition Experiment Dry-Run Flow	20

LIST OF TABLES

Table	Page
Table 1. Program Events and Activities	4
Table 2. Summary of TA3 Evaluations	5
Table 3. Protocol Key Concepts	6
Table 4. Workflow Stages for Summer 2020 Evaluation	8
Table 5. Rubric for Summer 2020 Evaluation	9
Table 6. D3M Transition Partners and Stakeholders	12
Table 7. TA3 System by Use Case	14
Table 8. Glossary of Machine Learning Terms	15
Table 9. TA3 System Performance Across Stages	24
Table 10. Summary of General Evaluation Findings	26

1.0 SUMMARY

Parenthetic works on research and development projects in both the public and private sectors focused on new methods or technology for information and communication. For many clients, we support user testing and evaluation of software through test engagements. Our goal for these engagements is to inform and push the development of capabilities that will bring high fidelity to addressing national security interests. These activities enable us to translate end user requirements to technical teams and, similarly, translate technology capabilities and features to targeted end users. We also maintain—and continue to grow—a database of “users” from across public, private, and academic institutions that are available for these test engagements. This community spans many domains and levels of expertise, but our users have one commonality - interest in new technology that optimizes human-machine interaction for their domain. Additionally, we develop operationally-relevant challenge problem sets and evaluation metrics that are rooted in real-world data and circumstances.

The goal of DARPA’s Data-Driven Discovery of Models (D3M) program is 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. For D3M, Parenthetic assisted the program evaluation team by focusing exclusively on assessing the performance of the systems in supporting subject matter expert’s (SME) creation of data-driven models, and understanding model output. For machine learning (ML) technology to provide domain experts with utility and enable sensemaking, tools must easily connect the dots and allow experts to answer questions using their contextual knowledge and expertise.

As of March 2021, the Technical Area 3 (TA3) component of the D3M program—tasked to develop user interfaces to make optimal use of user effort and domain knowledge to derive the best model possible for a given problem—consisted of the following three systems and performer teams:

- **System #1:** System #1 will enable SMEs with no data science training to dynamically derive, understand and apply data-driven models in answering mission-critical questions. This approach augments human intellect with data discovery, enrichment, analytic model recommendation, and automated visualization to discover observable relationships in data, empowering intelligence action.
- **System #2:** System #2 provides a fundamentally re-thought data analytics stack: domain experts without programming experience or background in statistics can now explore and mine data through an intuitive user interface, and effortlessly build, analyze, and evaluate machine learning pipelines.

- **System #3:** System #3 is a data-driven web application designed to bring researchers to insights fast. It facilitates intuitive machine learning, model discovery, and data exploration for researchers. As System #3's intelligent back-end automatically seeks interesting relationships in the data and builds models to predict outcomes, researchers impart substantive knowledge about their data and own research questions to guide the automated generation of AI assistance for data analysis in an interactive paradigm we call *human-guided machine learning*.

For D3M, Parenthetic specifically:

- Met regularly with targeted transition partner end users to develop evaluation challenge problems, identify relevant datasets and users, and work towards SME adoption of developed technology
- Recruited SMEs and data science users for TA3 system testing and evaluation timelines
- Developed targeted user training
- Created and updated evaluation protocols
- Facilitated information exchanges between users and technology developers
- Ensured TA3 testing compliance with institutional review board (IRB) requirements
- Analyzed TA3 testing results and delivered user-focused conclusions and recommendations for the evaluation team
- Provided guidance on AI explainability and interpretability for non-computer scientist communities

Parenthetic's role consisted of two primary responsibilities. The first was to test and evaluate software developed by D3M TA3 performer teams to meet and exceed DARPA requirements and user expectations throughout the majority of the program. This included the development of pre-testing tasks, protocol development and refinement, completion of IRB requirements, user recruitment and testing tasks, and effective coordination of the entire end-to-end evaluation process.

The second responsibility was developed after a February 2020 meeting with DARPA D3M program leadership and led to Parenthetic pivoting to a sole focus on transition activities such as coordinating discussions with domain experts and carrying out technology demonstrations with various U.S. Government (USG) agencies and organizations. These activities were aimed at facilitating the adoption and/or continued funding of the technologies beyond the life of the D3M program which was scheduled to end in mid-2021.

In addition to these two major responsibilities, Parenthetic also conducted supplemental work to better connect TA3 performer teams with transition partner users, including:

- Coordinating exchanges and discussions with domain experts to better understand their experience with current tools, as well as their expectations and needs for newly developed modeling software
- Assisting performer teams in optimizing their training materials and methods to more effectively onboard new users, especially those that are unfamiliar with data modeling tools
- Identifying datasets and domain use cases that could serve as exemplar challenge problems for performer teams to use in benchmarking their systems

These activities helped to guide performer teams and improve D3M technology readiness for transition partner use cases. This work also enabled performer teams to showcase their systems' capabilities that are of particular interest to various USG organizations, increasing the likelihood of successful transition.

2.0 INTRODUCTION

This report summarizes Parenthetic's work on the DARPA D3M program from September 2019 to March 2021. This final phase of the D3M program focused primarily on continuous system testing and evaluation, and transition partner engagements. The former elements will be discussed at a higher level of detail, as the bi-annual evaluation reports attached as appendices covered these activities more in-depth. The full reports for these evaluations can be found in the Appendix of this report. The transition efforts, which began in October 2020, will be covered in greater detail.

Note that since the D3M systems are still in the process of transitioning with various USG agencies, the names of the systems, transition partners, and transition partner datasets will be appropriately anonymized (e.g., System #1, Transition Partner #1, etc.) to avoid violating existing memorandums of understanding (MOUs).

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

This section describes the various events and activities that Parenthetic either supported or led from September 2019 onward, as shown in Table 1 below.

Table 1: Program Events and Activities

Event	Date	Objective
System Evaluations		
TA3 Evaluation #2¹²	Fall 2019	Test and evaluate four TA3 systems
TA3 Evaluation #3³	Summer 2020	Test and evaluate three TA3 systems
Transition Partner Engagement Events		
Transition Partner #1 Demo	Oct. 8, 2020	Demonstrate System #2 to Transition Partner #1
Transition Partner #2 Demos	Nov. 20, 2020	Demonstrate all D3M systems to Transition Partner #2
Transition Partner #2 Demos	Feb. 5, 2021	Demonstrate D3M systems to a second team within Transition Partner #2
Transition Partner #2 Demos	Feb. 11, 2021	Demonstrate D3M systems to a third team within Transition Partner #2
Winter 2021 Workshop Demos	Feb. 25-26, 2021	Demonstrate D3M systems to various transition partners and USG stakeholders
Experiment Dry-Run	Mar. 9, 2021	Dry-run of System #2 experiment in preparation of transition partner experiment in the near future

¹ TA3 Evaluation #1 occurred during Summer 2018 so will not be included in this report

² Full report found in the Appendix

³ Full report found in the Appendix

3.1 Evaluations

Parenthetic's D3M program support from September 2019 onward focused on the design, implementation, and completion of two evaluations to assess TA3 systems' capabilities to meet DARPA's requirements and transition partner user expectations. A high-level summary of these evaluations is shown in Table 2.

Table 2: Summary of TA3 Evaluations

Evaluation	User Pool	Performer Teams/ Systems Evaluated	Datasets Used⁴
Fall 2019	41 users @ 2 systems each	System #1 System #2 System #3 System #4	ACLED Hate Crimes TERRA Dataset #1 Dataset #2
Summer 2020	9 users @ 1-4 systems each (D3M government team personnel)	System #1 System #2 System #3	Cancer Visa Applications TERRA Dataset #1

The subsequent sections detail the key methods that were utilized in order to ensure the successful implementation of each evaluation.

3.1.1 Fall 2019 Evaluation

Evaluators for the Fall 2019 evaluation were recruited from external sources via the Parenthetic user database and the user sign-up hosted on Parenthetic's website. The evaluation required significant preparation and software readiness assessments to ensure that the TA3 systems were prepared for testing. The preparation work included:

- Identification and procurement of multiple open-source datasets that could be used for evaluation
- Dry-run tests to ensure that TA3 systems integrated successfully with their respective Technical Area 2 (TA2; focused on developing techniques for automatically selecting model primitives and for composing selected primitives into complex modeling pipelines based on user-specified data and outcome(s) of interest) system pairings
- Online user training sessions and webinars

⁴ Dataset descriptions are provided within the appendix

For this evaluation, a protocol was designed to measure TA3 systems on key concepts of interest to DARPA and the evaluation team. Parenthetic developed several indices to validate each of these key concepts, (both forward- and reverse-coded) and randomized the order of their presentation to users. Pilot results indicated that these scales were highly reliable, as shown in Table 3.

Table 3: Protocol Key Concepts

Factor	Description	Best-fitting protocol question	Reliability of all items in this factor, post-analysis (Cronbach's α)
Enabled Exploration/Discovery	Systems enabled users to discover new insights	The system made it easy to find new sources of relevant data	0.99
Meaningfulness	Systems provided output that users found meaningful, made sense, and they could explain. System answered "why"	I CANNOT explain what these results mean	0.97
Explainability	Systems did not provide explainable output. System answered "how"	The system did NOT allow me to see how it made predictions	0.94
Right Amount of Information	Systems gave users the right amount of information	The system gave me an overwhelming amount of information	0.89

All TA3 systems underwent a system suitability assessment prior to user assignment. For this evaluation, the process entailed the following steps:

Step 1: TA3 performer teams performed preliminary testing using seed datasets and submitted their results. Systems passing this stage moved on to Step 2.

Step 2: Data Machines Corporation, a member of the evaluation team, carried out smoke testing of systems using sequestered datasets. Systems passing this stage moved on to Step 3. Systems that functioned on some datasets but not

others were still able to move forward for evaluation on the subset of datasets that they passed.

Step 3: This consisted of two separate but simultaneous sub-steps:

a) **The DARPA program systems engineering and technical assistance (SETA) lead** carried out a detailed system review on factors such as accuracy of datamart datasets, reset buttons causing issues on the backend, and appropriateness of model explanations relative to the dataset.

b) **Parenthetic** conducted a review from the user perspective and ensured there were no major stumbling blocks to a user successfully completing the system test. At the end of this stage, recommendations were made to DARPA on which TA2/3 system combination for each sequestered dataset should be presented to users.

Step 4: Systems making it to this stage were deemed suitable for presentation to users. TA3 software prototypes that made it past the “user-readiness” stage were tested by users while following a test protocol.

3.1.2 Summer 2020 Evaluation

The Summer 2020 evaluation was unique in that users were drawn from the program's evaluation team per DARPA's request due to time and resource constraints. Users were selected from the DARPA, NASA Jet Propulsion Lab (JPL), Data Machines, and Parenthetic teams – yielding a user base with a wide range of modeling and machine learning (ML) experience.

Similar to previous evaluations, the Summer 2020 event was focused on the importance of the four key concepts mentioned previously (Enabled Exploration, Meaningfulness, Explainability, and Right Amount of Information). However, the protocol was refined to create a more practical and faster workflow, meant to represent the typical analytical process of USG domain experts. The workflows were developed based on structured interviews with USG analysts and domain experts that took place in the lead up to the evaluation. Each TA3 system was evaluated on its ability to successfully address each stage enumerated in Table 4.

Table 4: Workflow Stages for Summer 2020 Evaluation

Workflow Stage	Key Component
Load	Enables user to easily load new datasets
	Uploads datasets from local computer
	Provides additional options to upload datasets (e.g., Internet, Kaggle, etc.)
Explore	Enables user to easily explore the data
	Views single variable relationships
	Views multivariable relationships
	Explores correlations between features
	Identifies outliers in the data
	Displays an appropriate amount of information
Prepare	Enables user to easily prepare data for model construction
	Allows for addition/deletion of features
	Supports feature engineering using Datamart or other related services
	Allows user to address outliers or other data issues prior to constructing models
Model	Enables user to easily construct models
	Allows user to define parameters for model construction (e.g., processing time, number of models, etc.)
	Allows user to stop, restart or reset model construction
Explain	Explains model feature importance
	Explains the relationships between features and target variables
	Allows user to compare models prior to model selection
	Recommends the best model based on user preferences
Deploy	Allows models to be easily applied to new data
	Provides user with clear results in deployed settings

Using the rubric shown in Table 5, systems were scored by nine users who assessed each system on its ability to support the various stages.

Table 5: Rubric for Summer 2020 Evaluation

Score	Threshold
0	Not Supported (i.e., capability does not exist or user is unable to find/use the functionality)
1	Insufficiently Supported (i.e., capability exists but it not useful or effective)
2	Minimally Supported (i.e., capability is supported but needs improvement)
3	Adequately Supported (i.e., capability is useful and effective)
4	Exceptionally Supported (i.e., capability is unique, highly useful, and effective)

General findings from these two evaluations are discussed in more detail in the Results and Discussion section.

3.2 Transition Partner Engagements and Other Activities

The process for transition partner engagements is shown in Figure 1. The activities were originally planned to take place in-person, however, due to Covid-19 restrictions on travel and in-person meetings, all engagements were held virtually through video conferencing software such as Zoom and Microsoft Teams.

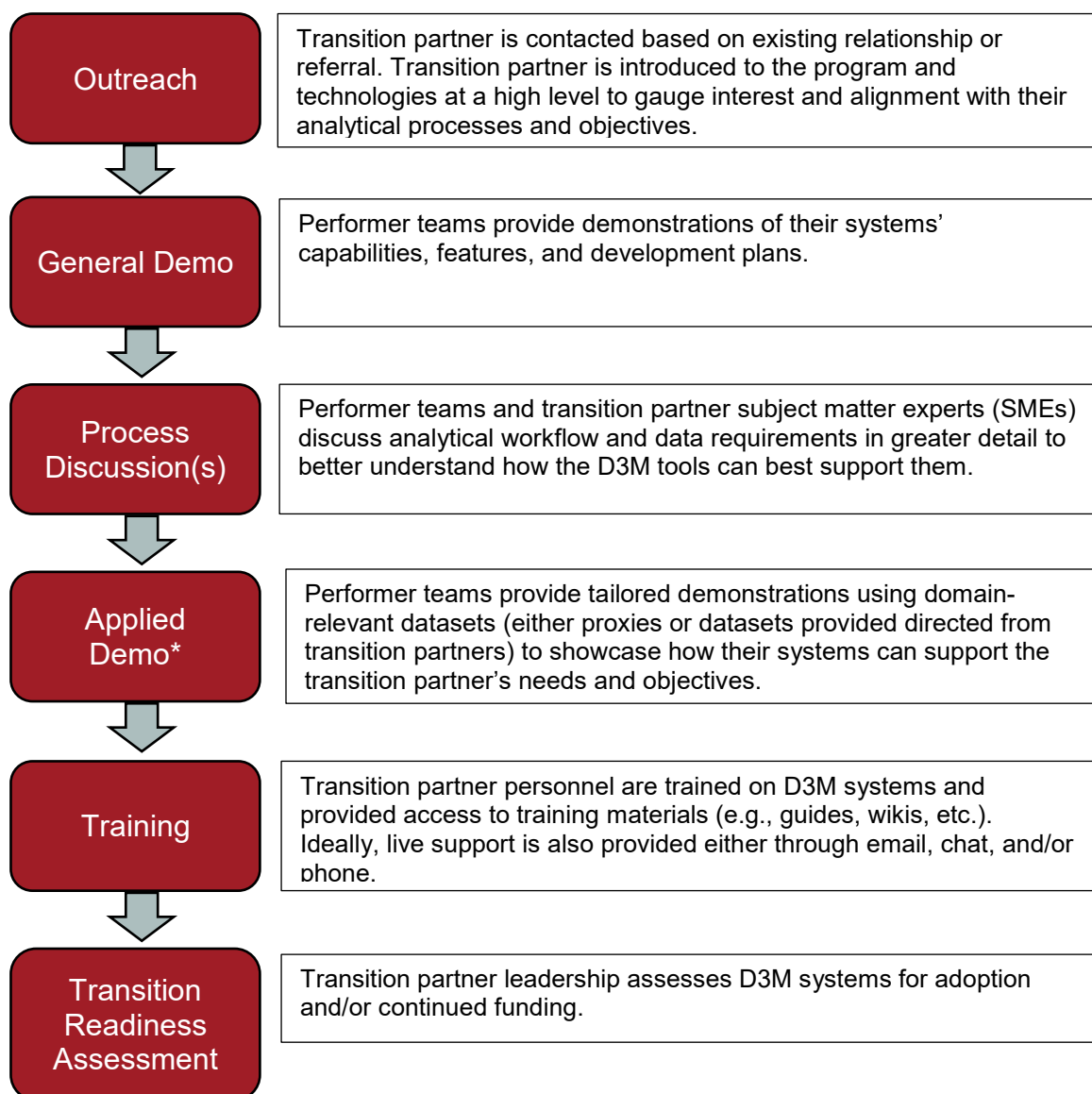


Figure 1: Transition Partner Engagement Process

*Note As of the writing of this report, transition partners have reached the "Applied Demo" stage with the latest round of demonstrations held on February 25-26, 2021.

The D3M program has been actively engaged with a number of transition partners throughout the duration of the program. For the latter phases of the program, transition partner engagements became the central focus. Although it was difficult to carry out these tasks due to COVID-19 restrictions, system demonstrations provided valuable insights. For example, the virtual setting allowed for more attendees to observe the demonstrations considering logistics such as travel were not necessary. Most importantly though, it allowed for direct discussions between performer teams and transition partners, giving the performer teams insight into how USG analysts typically work, what they need, and which system capabilities are currently missing. While it would have seemed more beneficial to allow users hands-on experience with the systems underperformer team guidance, the virtual conditions did not hinder transition activities as Parenthetic remained flexible, adjusting our approach to minimize disruption and meet the needs of the program.

3.2.1 Transition Partners and Stakeholders

The automated systems developed in the D3M program are of great interest and relevance to a number of USG agencies, especially those who need data science capabilities. As systems grew to focus on empowering analysts with modeling and ML capabilities, the program increased interaction with these USG stakeholders to varying degrees and engaged more extensively with specific primary transition partners. In addition to these transition partners, a number of additional USG agencies and organizations attended system demonstrations and discussions either as prospective transition partners, SMEs, or interested parties.

3.2.2 Transition Partner Use Cases

As the program engaged with transition partner stakeholders throughout the USG, Parthenetic developed a list of example use cases that aligned with the partners' internal datasets. The objective of this task was to identify domains and challenges that mirrored the workflows and problems that USG domain experts experience in their daily work. Additionally, these use cases helped to drive technology demonstrations, allowing performer teams to showcase system flexibility across a variety of domains and problem types. This, in turn, helped to increase system appeal to prospective transition partners. A number of these example use cases were used for system demonstrations after transition partners verified their suitability during iterative discussions, as shown in Table 6.

Table 6: List of Transition Partner Use Cases

Domain	Use Case Description	Problem Type(s)
Aerospace	Improve the identification and tracking of space debris and predict the likelihood of future risks to spacecraft and satellites	Remote Sensing
Agriculture	Examine patterns of locust infestation to model indicators related to food security	Remote Sensing
Climate	Forecast future climate conditions and trends (e.g., severe droughts, floods, etc.)	Time-Series Forecasting
Environmental	Predicting wildfire risk or damage levels to minimize loss of life and property.	Classification Regression
Immigration	Analyze an individual's information to aid the asylum or visa screening process	Classification
Immigration	Analyze large datasets of visa applications to identify patterns to streamline analysts' screening process	Classification

Table 6: List of Transition Partner Use Cases

Domain	Use Case Description	Problem Type(s)
Infrastructure	Analyze patterns of power usage to identify zones at greatest risk for outages	Classification
Security	Extract information related to events involving piracy.	Classification
Security	Gain insights into conflicted-affected countries by forecasting subnational political violence	Time-series Forecasting
Security	Identify vessels to aid monitoring efforts of maritime activity (e.g., port traffic, ship routes, etc.)	Remote Sensing
Security	Enhance crowd analysis including detecting crowd flow changes, anticipating crowd movement, etc.	Remote Sensing
Security	Automate analysis of satellite imagery to identify elements of interest to military and intelligence partners (e.g., buildings, vehicles, road networks, etc.)	Remote Sensing
Security	Better counter violent extremism by identifying at-risk individuals and the networks they are a part of (e.g., threat vs. no threat)	Classification
Law Enforcement	Leverage crime data to more accurately predict law enforcement deployment efforts by identifying areas at risk for high crime activity (e.g., low, medium, high)	Classification
Military	Weather monitoring and forecasting to aid ground and air operations	Time-series Forecasting
Military	Aid mapping and terrain analysis to aid military base and/or port siting or closing	Classification
Public Health	Forecast population and demographic shifts to better understand future needs of the population (e.g., healthcare, food supply, foreign aid, etc.)	Time-Series Forecasting
Public Health	Improve modeling of the spread of infectious diseases to help plan future interventions by identifying at-risk populations	Classification
Public Health	Build predictive models related to integrated Food Security Phase Classification (IPC) to address food security issues	Classification

 = Used for system demonstrations

Based on the Fall 2019 evaluation findings, D3M program leadership assigned each remaining TA3 system a specific real-world use case and problem

type on which to focus their efforts. These assignments were selected to meet system strengths and weaknesses from the Fall 2019 evaluation results, as well as the expertise and background of each performer team. Likewise, each use case had direct relevance with specific transition partner. Table 7 lists the specific use cases assigned to each TA3 team:

Table 7: TA3 System by Use Case

TA3 System	Use Case(s)
System #1	Model indicators related to food security due to locust infestation using multi-spectral satellite imagery.
System #2	Build and explain models focused on piracy events to explore patterns and predict piracy activity. Build predictive models of Integrated Food Security Phase Classification to improve decision-making around food security.
System #3	Forecast subnational political violence across various countries.

3.2.3 Glossary of Machine Learning Terms

To help educate domain expert users of D3M systems with limited data science and ML expertise, Parenthetic created a glossary of terms to serve as a training aid. The glossary was developed based on evaluation findings of user knowledge of these topics, as well as Parenthetic's continued exploration of each TA3 system. This glossary can serve as a more practical user resource than requiring performer teams to embed definitions, or other layman terms, in their systems and training materials. This glossary can be found below in Table 8.

Table 8: Glossary of Machine Learning Terms

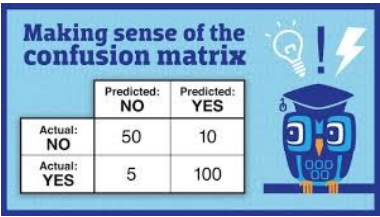
Term	Definition	Example (if appropriate)
Bias	The degree to which a reference value deviates from the truth. Bias can cause skew in data collection or interpretation. (Example: sample bias is when the data used to train your model does not accurately represent the environment that the model will operate in.) Bias is always a concern since automated machine learning can only be built on available data.	A model used to classify pictures of dogs into breeds was based on cat training data.
Categorical data	Features (e.g., column variables) that have a set number of possible values.	"Hair color" is either black, blonde, brown, grey, or red.
Classifier	An algorithm that categorizes and labels certain features within a dataset.	A model that classifies pictures of dogs into specific breeds.
Clustering	Automatically grouping data points based on shared attributes.	A business clusters its customers in groups to design a different marketing strategy for each (e.g., clustering by age, gender, etc.)
Confusion matrix	A table that summarizes how successful a classification model's predictions were by tracking predicted vs. actual values.	Example of a confusion matrix:  Image source: Data School. "Making sense of the confusion matrix." Online video clip. YouTube. YouTube, 25 March 2014, https://www.youtube.com/watch?v=8Oog7TXHvFY

Table 8: Glossary of Machine Learning Terms

Term	Definition	Example (if appropriate)
Continuous data	Data that has an infinite range of possible values.	A person's weight
Data augmentation	Improving a model by expanding the scope of the data (e.g., adding newly collected data).	Improving a weather forecasting model that was built with 2017 weather data by adding 2018 data.
Decision tree	A flow-chart-like model consisting of branching statements where each branch is determined by a test on a feature (i.e., Yes or No).	Example of a decision tree: <pre> graph TD Weather[Weather?] -- Sun --> Time[Time?] Weather -- Cloud --> Hungry[Hungry?] Weather -- Rain --> No[No] Time -- "> 30 mins" --> No2[No] Time -- "< 30 mins" --> Yes1[Yes] Hungry -- Yes --> No3[No] Hungry -- No --> Yes2[Yes] </pre> Image source: Hoare, Jake. "What is a Decision Tree?" Displayr.com. 2 August, 2018.
Discrete data	A feature that is measured in a fixed amount.	The number of daily admissions to a hospital
Example	A row in a dataset that contains one or more features.	Weight, height, name, and age are features of an example of a person.
Explainability	The degree to which a model's predictions can be readily explained. The most common approach used to ML understanding is analyzing model features by looking at feature importance and feature correlations.	We will use the "Husky vs Wolf example" to illustrate the importance of model interpretation. The classifier task was to identify if a wolf was in a picture or not. It falsely misclassified Siberian Husky as a wolf. Researchers were able to identify what areas of the pictures were important for the model. It turned out that if the picture contains snow it is classified as a wolf as almost all of the wolf pictures used as training data had snow in the background. The algorithm was using the background of the picture and totally ignoring animal characteristics. The model should look at the animal's eyes instead. Thanks to this discovery it was possible to fix the model and extend the training examples to prevent the reasoning snow = wolf.

Table 8: Glossary of Machine Learning Terms

Term	Definition	Example (if appropriate)
F1 Score	A statistical measure that rates the performance of a machine learning model. F1 scores range from 0 to 1 where 1 indicates the highest performance and 0 indicates the lowest performance.	N/A
Feature	The observable characteristic of interest in the data that you are seeking to model (e.g., column variables)	The various features of a person include their weight, height, name, age, etc.
Ground truth	The correct answer or reality. Often used as the gold standard for comparing a machine learning model's output to gauge performance.	Imagine a model that is designed to determine whether or not a person is pregnant. It ends up classifying a man as pregnant. Ground truth tells us that this is not possible and that model needs to be revised.
Histogram	A data visualization that shows the distribution of data for a given feature (or set of features).	A bar graph that shows the distribution of ages across a class of students.
Independent variable	The variable you change to test the effects on the dependent variable.	In a study to determine whether how long a student sleeps affects test scores; the independent variable is the amount of sleep. As the amount of sleep changes, you want to observe how test scores, the dependent variable, are affected.
Interpretability	The degree to which a model is easy for a human to comprehend.	Depending on the specific output, weather forecasts are more interpretable than credit scores
Machine learning (ML)	The application of artificial intelligence that provides systems the ability to learn and improve from experience (e.g., new data) without being explicitly programmed	Email services use artificial intelligence to filter incoming emails. Users can train their spam filters by marking emails as “spam”.
Model	A mathematical representation of a real-world process/system that is created by training a machine learning system using examples.	A model is used to identify and classify images of vehicles into categories (e.g., plane, boat, car, etc.).
Outlier	Data points that are distant from others.	The 6 ft 5 in tall student in a class where the average height of the other students is 5 ft 5 in.

Table 8: Glossary of Machine Learning Terms

Term	Definition	Example (if appropriate)
Overfitting	A form of error that occurs when a ML model corresponds too closely to data presented during training time and fails to generalize to data presented at testing time.	Imagine studying for an exam by memorizing previous versions of the test only to realize that come test day, you are confronted with a brand-new set of problems that you can't apply your studying to.
Precision	The ratio of correct positive predictions (see Prediction) to the total predicted positives.	Tells us what proportion of patients diagnosed as having cancer actually had cancer.
Prediction (added)	In machine learning, prediction refers to the output that a trained model generates when applied to new data. This differs from the general definition of prediction which refers to the forecast of a future event.	A machine learning algorithm is trained to find patterns in historical customer data. It is then applied to a new set of customer data to predict the likelihood of certain customers churning in the next 30 days.
Random forest	An approach to finding the decision tree that best fits the training data by creating many random decision trees and determining the average one.	See "decision tree"
Recall	The ratio of correct positive predictions to the total positive examples.	Tells us what proportion of patients that actually had cancer were diagnosed as having cancer.
Regression	In machine learning, used to predict the outcome of an event based on the relationship between variables in a dataset.	Predicting the price of a house based on its size, number of bedrooms, number of bathrooms, yard size, etc.
Target variable	The quantity that is being estimated by a machine learning model using other variables/features.	In developing a machine learning model to estimate students' test scores based on other factors like sleep, nutrition, exercise, etc., the test score is considered the target variable.

Table 8: Glossary of Machine Learning Terms

Term	Definition	Example (if appropriate)
Testing data	After a model is trained with training data, the model is then applied to testing data to measure its performance.	An ML system is fed training data that includes various features of animals, classified into species. The system trains itself on this accurate data to learn the patterns related to how animals are assigned to which species. It is then fed data with data on animals without assigned species, the testing data, and attempts to assign species based on what it learned from the training data to see how it performs.
Time series analysis	A use case in machine learning which is focused on the analysis of data collected at different points in time.	Forecasting 2020's hurricane season based on previous years' data.
Training set/data	The subset of the dataset used to train a model. Systems analyze the data and learn patterns that feed into the model which can be applied to other datasets.	An ML system is fed training data that includes various features of animals, classified into species. The system trains itself on this accurate data to learn the patterns related to how animals are assigned to which species.
Tuning	The process of adjusting the parameters of a machine learning model to maximize its performance without overfitting or increasing variability	Imagine these as dials that allow you to adjust the parameters of a model.

3.2.4 Transition Partner Experiment Dry-Run

A critical component of the transition process is allowing transition partners and stakeholders to witness firsthand how a system could be introduced, adopted, and utilized within their organization. On March 9, 2021, Parenthetic and System #2's performer team carried out a dry-run test of their system for a first-time user, with a specific focus on the onboarding process. The user's main task was to evaluate discrete components of the onboarding experience and the challenges it may present to future users. The general flow of the dry-run is shown below in Figure 2.

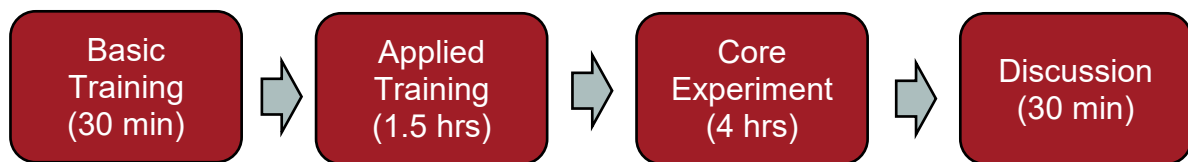


Figure 2: Experiment Dry-Run Flow

The system dry-run consisted of the following elements:

1. **User Background:** The test user was a senior Parenthetic analyst with an extensive background in intelligence analysis and general knowledge of statistical modeling. This user's understanding of analytical methods made them an ideal fit, aligning with System #2's transition partners who primarily reside within DoD organizations.
2. **Basic Training:** The performer team introduced the system and trained the user on basic system functionalities such as the user interface, menu system, and general navigation.
3. **Applied Training:** Under the supervision of the performer team, the user applied their learning as they interacted with the system using a predetermined dataset. Specifically, the user tested the following components of the system's onboarding experience:
 - Data upload and system navigation
 - Data visualization (histograms, scatterplots, line charts, pivot charts, etc.)
 - Statistical modeling and computation
 - Data extraction and pattern building
 - Building and interpretation of predictive models
 - Applying predictive models to new data
 - Scenario simulation and "what-if" analyses

- Real-time collaboration with the performer team (in this case, a member of the System #2 performer team)
 - Creating dashboards
- 4. Core Experiment:** The user was given ample time for independent problem exploration using an open source dataset consisting of customer churn rates for a banking institution. Although this domain was outside of the user's expertise, it presented a straightforward challenge problem with easily understandable variables and context. The user was encouraged to explore and work on their own, but had access to the performer team via video conferencing and remote screen sharing for troubleshooting assistance. This portion of the experiment was also video-recorded, enabling the performer team and DARPA program management to review the user's experience at a later time, or made available to transition partner stakeholders to view at their convenience.
- 5. Discussion:** The user presented their findings to the performer team, allowing the team to understand the effectiveness of the training material and how the user interacted with the system to solve the specific tasks (e.g., create a histogram, identify variable relationships, etc.). The user also completed a post-test survey about their experience, which the performer team could compare to previous test results.

This dry-run produced a number of useful results for System #2's performer team to update the systems as they prepare for future experiments with transition partners. These results are discussed in further detail in the Results and Discussion section.

4.0 RESULTS AND DISCUSSION

This section presents a detailed review of lessons learned from the activities during the specified time period.

4.1 Fall 2019 Evaluation Results

The Fall 2019 evaluation sought to assess to what degree non-data scientist domain experts were able to effectively utilize D3M TA3 systems. Results from the Fall 2019 evaluation are shown below:

- **System Functionality**
 - System #2 and System #3 were the only systems able to host users across all five evaluation datasets. Other systems were either unable to work with specific datasets (System #1 only worked with one of the five datasets, System #4 worked with four of the five datasets), or users were left waiting a long time for systems to load data augmentation, exploration, or modeling tasks.
 - While systems undergo suitability testing before being put in front of users, system stability is often fragile. The capabilities that users required were often either not available or not functioning.
- **Data Problem Exploration**
 - Overall, users sought to explore their data with the visualization and exploration features. Three of the four systems provided users the ability to explore their data and problems of interest. System #4 was unable to support the SME user cohort in this endeavor.
- **System Utility and Explainability**
 - In general, users perceived AI as very powerful, but, when using it, they often felt that their expectations were not satisfied. Often, they were unwilling to adapt to the process/tool.
 - Systems were focused on communicating *how* a model works instead of *why it generated the results that it did*. Users across all systems reported feeling unable to assess model usability. These models, therefore, lacked *explainability* - which is a key system requirement.

Based on these findings, the following general recommendations were made for each TA3 system:

System #1: With only 4 users testing System #1, we are unable to draw any specific conclusions. However, the team should investigate data ingest issues that caused System #1 to only work with the ACLED dataset (a tabular dataset and classification problem).

System #2: System #2 received mostly positive results. We recommend system improvement to enable SMEs to derive meaningful results from system output. This issue can be addressed with improved training and a user interface (UI) that is sufficiently flexible to support different user profiles. For example, some users wanted to see features that are currently hidden on the UI and wanted to hide the more advanced features that are currently visible. Other users wanted to see the opposite.

System #3: System #3 received positive ratings overall and the highest ratings from both SME and data scientist users. However, System #3's users often reported that they were unable to finish due to long load times, errors, and other functionality issues. These users also specifically sought additional education or information about the system and the ability to manipulate data. This system requires improvements to the data ingest process. We also recommend this team pivot to the more complex tasks of improved system utility and model explainability.

System #4: System #4 received mostly negative ratings and requires significant improvements to meet the goals of the program. Overall, System #4 underperformed other systems in model discovery. Furthermore, SMEs using System #4 especially reported that they were overwhelmed by information when compared to data scientists. The team could benefit from more frequent user feedback, and finding ways to meet the needs of non-data scientist users.

4.1.1 Fall 2019 Evaluation Outcomes

The findings of the Fall 2019 evaluations resulted in program leadership making a number of key decisions, including:

1. The down-selection of the System #4. Based on the generally negative experience of users, it was decided that this performer team's time and

resources would be better spent focusing solely on further developing their TA2 system.

2. Program leadership decided that it was more feasible to continue developing specialized systems that could do specific tasks well, rather than create less sophisticated systems that attempted to meet all of DARPA's original expectations. To that end, leadership assigned each TA3 system a specialization, focusing on specific types of problems and use cases, based on system strengths and weaknesses.

4.2 Summer 2020 Evaluation Results

While the Fall 2019 evaluation participants were mostly domain experts with little experience in data science, the Summer 2020 evaluation included users that were also experienced data scientists and modelers. However, similar to the Fall 2019 results, TA3 systems tested during the Summer 2020 evaluation were not able to fully satisfy the needs of these data science users. The results of this evaluation were pivotal in helping each performer team better understand key areas of improvement by prioritizing development plans as they prepared for transition activities. For example, these insights allowed performer teams to design their demonstrations so that they could highlight strengths, while also proactively addressing areas that need further development. A crucial element of software development is understanding not only the areas that need improvement, but also having a vision of how to remedy it. Table 9 shows a summary view of TA3 system performance across the various workflow stages.

Table 9: TA3 System Performance Across Stages (Overall Avg Scores*)

Stage	System #1	System #2	System #3
Load	2.3	3.8	2.3
Explore	2.0	2.9	2.6
Prepare	2.1	2.5	2.0
Model	1.9	2.8	1.8
Explain	1.3	2.2	1.9
Deploy	1.9	2.3	1.1
Overall	1.9	2.7	1.9

*Scoring system: 0=Not Supported, 1=Insufficiently Supported, 2=Minimally Supported, 3=Adequately Supported, 4=Exceptionally Supported. Color coding is to aid score differentiation where green represents higher scores, red represents lower scores, and orange represents mid-range scores.

An overview of system findings includes:

- **System Messaging:** TA3 systems must improve how they communicate with users. This can be done by displaying accurate progress bars, run-

time timers and other system prompts. Without this information, the systems place users in ambiguous situations where they are unsure if a system is processing a request or experiencing a freeze or bug. This has the effect of causing users to be unsure of what to do – for example, some users restarted their browsers thinking the systems had stalled completely.

- **User Guidance:** Users appreciate the level of freedom that the systems allowed them, but at times struggled with how to best approach their analysis given the wide array of options and functions presented in each system. Systems need to find a balance between user freedom and system guidance, especially when considering the potential diversity of data science expertise among users. This requirement for level of guidance spans the entire user experience spectrum - from helping users best prepare their data for modeling to understanding model outputs.
- **Explainability:** Every system overwhelmed users to some degree with information overload. Often systems placed the burden of interpreting large amounts of information on the user rather than helping users to connect the dots, leading to insights they would not have discovered otherwise – one of the more appealing functions these systems are supposed to feature. Requiring users to sift through various outputs and results is counterproductive and antithetical to the ultimate role ML systems are supposed to play: helping those with minimal data science knowledge construct and interpret models without a data scientist by their side.
- **General User Experience:** Each system should consider a number of adjustments to enhance the overall user experience, such as:
 - Adding tooltips – For example, providing more detailed data over charts or buttons to more clearly indicate their functionality.
 - Reducing redundancy – For example, allowing users to conduct actions in batches rather than requiring repetitious and tedious actions.
 - Automating processes – For example, automatically identifying outliers based on user-defined parameters or removing variables with no data.

4.3 Summary of Evaluation Findings

Table 10 summarizes findings from the two program evaluations, Parenthetic's recommendations, and end-of-program status for each system in meeting the recommendations.

Table 10: Summary of General Evaluation Findings

Finding	Parenthetic Recommendation/Outcome	Status
Systems struggled with generating meaningful output that is easily interpretable for users.	A wide variety of system-specific changes to improve users' ability to better understand system outputs.	Systems continue to improve explainability. For example, System #2 has widely implemented "wizards" that help simplify the model building process, prompting users with a series of basic questions to adjust their model parameters. As performer teams continue to interact with transition partners, explainability should improve.
Evaluations showed that each TA3 system had strengths across different problem types.	Based on evaluation results, program leadership assigned each TA3 system to specialize in a problem type. In doing so, the program was able to leverage the strengths of each system and the expertise of their respective performer team.	The following specializations were assigned after the Fall 2019 evaluation: • System #1: Remote-sensing • System #2: Predictive models • System #3: Time-series forecasting
Leveraging operational use cases that mirror real-world challenges was critical to accurately evaluating the systems.	Conduct research in coordination with transition partners to identify relevant challenge problems and datasets that serve as appropriate proxies.	Parenthetic researched and identified various use cases for consideration such as the H1-B visa application dataset. This is an ongoing effort with current transition partners.

Table 10: Summary of General Evaluation Findings

Finding	Parenthetic Recommendation/Outcome	Status
Training materials provided during the evaluations were confusing to follow.	Redesign training materials to better meet training objectives and user needs.	Parenthetic vetted and revised each system's training materials. This should be an ongoing effort as performer teams continue to iterate their systems with new features
User backgrounds need to be well understood.	Hold discussions with transition partner personnel to understand their roles, needs, and objectives. This information is also critical for curating training curriculum and materials.	Parenthetic coordinated several system demonstrations and discussions with domain experts and transition partner SMEs to better understand users.

More specific conclusions helped us reconfigure and streamline our processes for future evaluations:

- **User Engagement:** Identifying appropriate problem sets as early as possible helps to ease the process of user engagement between sign-up and task initiation. Also, more iterative small-scale testing requires less users, mitigating the impact of user turnover.
- **User Cognitive Load:** Parenthetic identified ways to lessen the cognitive load placed on users with regards to evaluation protocols, training, and interaction with the systems. Given the scale of previous evaluations, many users were overwhelmed by the amount of content presented to them, particularly with the training materials and the systems themselves. A more iterative testing approach would have enabled users to focus more on discrete features and functions, helping to alleviate the cognitive load associated with the full array of operability across multiple systems. As noted earlier, Parenthetic also actively worked closely with each performer team to enhance training materials, and redesign documentation to be more digestible for intended users, and simplified test protocols based on empirical findings from the Fall 2019 evaluation.

Implementing this more iterative user testing process was not possible as the program had progressed into the transition phase. This phase focused primarily on establishing transition partner needs in terms of technology adoption rather than more significant development.

4.4 Transition Partner Experiment Dry-Run Results

As mentioned previously, it is vital for transition efforts that prospective transition partners understand the logistics of transitioning new technology to their organization. The purpose of the dry-run held on March 9, 2021, was to ensure that the System #2 performer team, as well as the other performer teams, are able to enact an onboarding process that is intuitive and informative, leveraging everything they have learned from previous evaluation events. The purpose of such an experiment is to demonstrate that beyond the proven utility of a system, it is also feasible and practical to incorporate it into an organization's existing analytical ecosystem. Findings from the dry-run experiment include:

Areas of Success

- Training content was relevant, and the user was able to easily call on it when applying learned concepts and functions to new datasets.
- System #2's remote real-time collaboration function proved to be a critical training aid, allowing the performer team to seamlessly step in and assist the user when necessary.
- Despite having no experience with churn analysis or the training dataset's domain (finance), the user was able to complete the bulk of the requested analyses, showcasing System #2's ability to simplify the analytical and model-building process for new users, even in domains that are novel for them.

Areas of Improvement

- A brief training session or documentation for understanding data science fundamentals (e.g., Data Science 101) would be beneficial, particularly for users with no data science expertise. An example of this is a glossary of common machine learning and data science concepts that a user can easily reference, as provided by Parenthetic.
- A firm understanding of user background(s) prior to the experiment is essential to better curate the training. For example, this would help with identifying appropriate datasets as well as prioritizing specific training topics.
- Browser requirements should be explicit to ensure user machines are prepared to access systems.
- A small number of technical issues occurred, mostly related to server connectivity. The performer team will address these for future experiments.
- More breaks during the core experiment would be beneficial to avoid overburdening the user.

Overall, the experiment dry-run was considered a success as it will help System #2's performer team refine their onboarding process. The general learnings can be applied to the other D3M systems that may have transition partner engagements in the future.

5.0 CONCLUSIONS

Over the course of 19 months (September 2019 – March 2021), Parenthetic focused on furthering the key objectives of DARPA leadership for the D3M program which revolved around preparing the TA3 systems and their performer teams for transition. Parenthetic was flexible with program needs, pivoting when necessary, to ensure the program's goals were met by leveraging Parenthetic's expertise in understanding the more practical needs of technology users. Evaluations were conducted consisted of thorough and rigorous methodologies and resulted in systems that could produce more interpretable findings, giving TA3 performer teams a clear path for further developing their software in ways that brought them closer to DARPA's requirements. Upon the end of Parenthetic's contract, each of the remaining three performer teams had at least one USG agency or organization that was interested in continuing transition activities. Based on this, Parenthetic was able to meet and exceed its core role and responsibilities as a member of the D3M evaluation team.

6.0 APPENDIX

This section is comprised of the following technical reports:

- 2019 Bi-Annual Evaluation Brief – a comprehensive report covering the Fall 2019 evaluation
- 2020 Bi-Annual Evaluation Brief – a comprehensive report covering the Summer 2020 evaluation

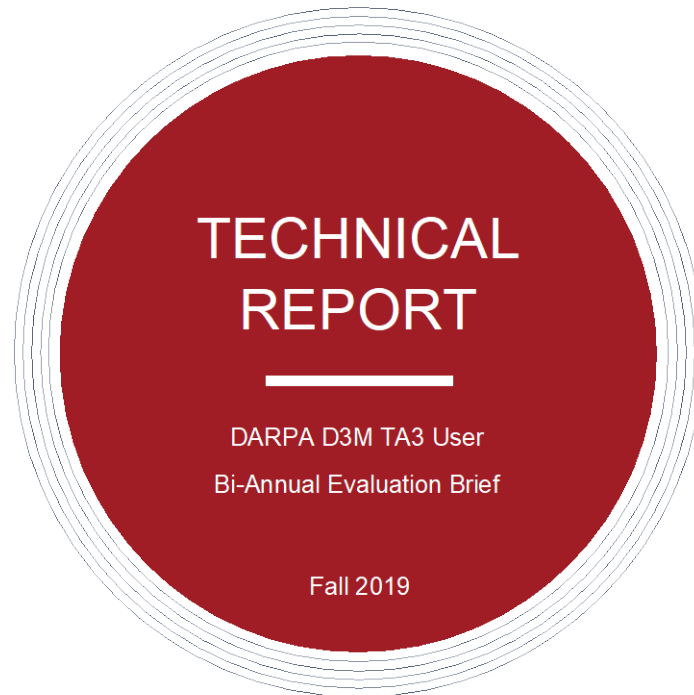
Note that these reports have also had any mentions of personnel, system names, transition partners, etc. anonymized to abide by MOUs that may still be in place.

6.1 Fall 2019 Bi-Annual Evaluation Brief

Alexandria, VA
703.406.8730
www.parenthetic.io

Contract No.: FA8750-19-C-1522
Performing Organization: Parenthetic
Deliverable: 0001AB Technical Report – Bi-Annual Evaluation Brief

PARENTHETIC



Prepared for:
DARPA Program Manager

Submission date:
06 January 2020

Submitted by:
Parenthetic
Catherine Tejeda
Reva Schwartz
Leann Down

Partnered with:
The George Washington University
David Broniatowski, PhD
Lydia Gleaves

*Unless otherwise stated, graphical output based on protocol results shows number of **protocol responses** (n=) on the y axis and **user rating** on the x axis. The average **user rating scale** runs from -3 to +3 where 0 means "neither agree nor disagree", +3 means "strongly agree", -3 means "strongly disagree".)

TABLE OF CONTENTS

Executive Summary	3 - 10
Background	12 - 13
Evaluation Elements	14 - 21
Evaluation Outcomes	22 - 30
Lessons Learned	31
Future Considerations	31
Appendices	32 - 77
Appendix A	32
Appendix B	33 - 35
Appendix C	36 - 40
Appendix D	41 - 43
Appendix E	44 - 48
Appendix F	45 - 78

EXECUTIVE SUMMARY

During the D3M fall 2019 user evaluation, a total of 41 subject matter expert (SME) and data scientist users assessed TA3 performer built systems (two systems were exercised by each user) for functionality and usability.¹ For this evaluation we captured user responses via an evaluation protocol that each user followed. We examined these responses using a reliable psychometric instrument based on principal component analysis (PCA; a widely used dimensionality reduction technique) – the results of which constitute the bulk of this report. We also collected the following user input: 1) user selected models generated by the TA3 system,² 2) post-test questionnaire responses, 3) comments provided by selected users during online feedback discussion sessions, 4) questionnaire responses from users who did not complete their tasks. A detailed review of results can be found in the body of this report.

Four (4) TA3 systems were evaluated using datasets from the following five (5) domains:

1. Armed conflict (Armed Conflict Location and Event Data (ACLED) data)
2. Plant science (ARPA-E/TERRA data)
3. Dataset #1 (A transition partner dataset)
4. Hate and violent extremism (Hate Crimes data)
5. Dataset #2 (A transition partner dataset)

Table 1: Number of user interactions by system and dataset.

Note that System #1 and System #4 did not work with all of the five datasets. The dataset cells also list the associated data and problem type.

System	ACLED - geospatial + tabular - multi-class classification	TERRA -tabular -time series forecasting	Dataset #1 ³ -tabular -binary classification	Hate Crimes -tabular -regression	Dataset #2 ⁴ -tabular -binary classification	Total Number of Interactions
System #1	4	-	-	-	-	4
System #2	4	13	1	3	2	23
System #3	3	13	1	6	1	24
System #4	4	15	-	2	1	22

¹ Parenthetic is responsible for the socio-technical evaluation of TA3 systems as performed by recruited users. Other government evaluation team members provide engineering and technical guidance across the TAs for the program.

² Collected and provided by Data Machines Corporation.

³ The low number of users for Dataset #1 is due to few SMEs in the field.

⁴ The low number of users for Dataset #2 is due to user non-completion. These results are discussed in the "User Feedback" section.

EXECUTIVE SUMMARY

Key Findings

The primary focus of TA3 systems is to support subject matter experts (SMEs) *without data science expertise* in automatically building empirical models and understanding their output. Therefore, we developed metrics to index: a) systems' abilities to enable model discovery and exploration; b) meaningfulness of system output; c) systems' abilities to deliver the right amount of information to users, and d) explainability of system output.

Compared to data scientists, we found that SMEs were:

- 1) Less able to use TA3 systems for model discovery
- 2) Less able to derive meaningful interpretations from TA3 system output
- 3) More likely to be overwhelmed by information provided by TA3 systems

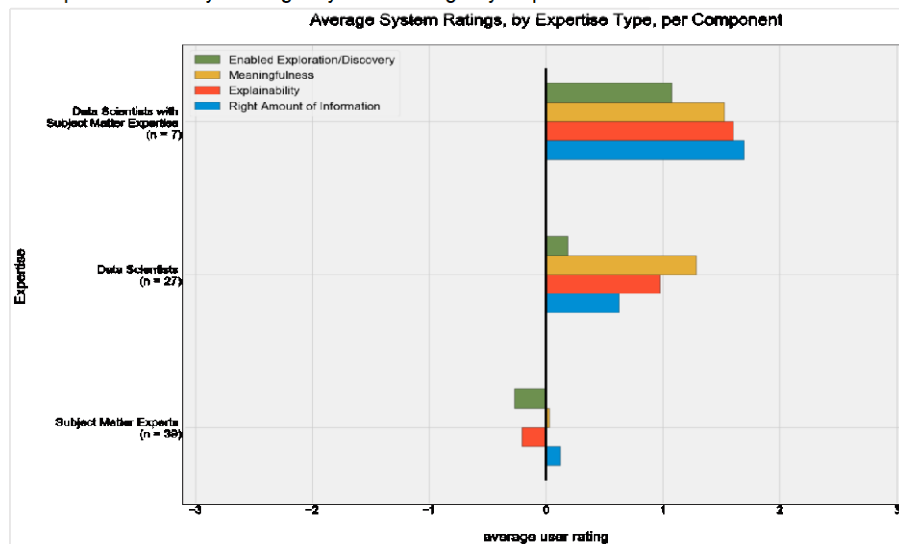
Notably, we detected only a marginal difference between SMEs' and data scientists' abilities to explain how a system generated its results. In effect, when SMEs could get a TA3 system to generate results, **SMEs and Data Scientists were both capable of describing the results that TA3 systems generated, but SMEs could not understand what these results meant.**

Not all systems performed equally well. In particular, we found that System #4 significantly underperformed System #2 in its ability to help users discover new models. Small sample sizes precluded us from detecting significant differences along other dimensions of system performance.

System Scores at-a-Glance

An overview of key system results based on Likert responses from the protocol is shown in Graph 1 and Figure 1 below. A detailed reporting of all quantitative results can be found in [Evaluation Outcomes](#).

Graph 1: Summary Average System Ratings by Expertise

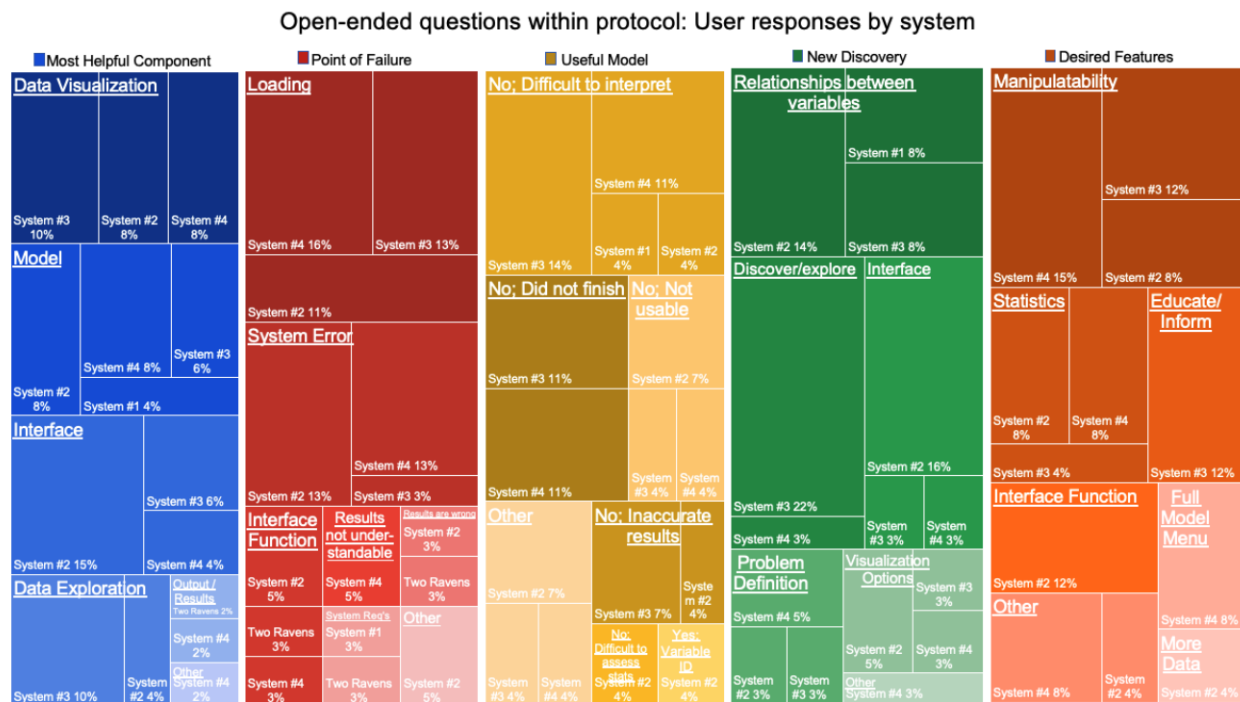


www.parenthetic.io

EXECUTIVE SUMMARY

In addition to Likert responses, the protocol contained seven (7) open-ended questions to elicit feedback on users' experiences. Responses were evaluated for common themes and independently coded. We summarize responses to five of the seven questions below. A detailed reporting of all qualitative results can be found in [Qualitative Findings](#).

Figure 1: Summary of open-ended responses to protocol questions.



EXECUTIVE SUMMARY

Conclusions

Although the overall number of users who completed system testing was small (n=41), we can make the following general conclusions:

- **User Attitudes About Testing Tasks**

- Data science users were more likely than SMEs to successfully complete their testing tasks. This result may be due to variance in availability to test systems.
- Overall, users who completed testing found the protocol and workflows easy to follow. Yet, a majority of these same users felt lost or confused while completing their tasks, and had difficulty understanding their tasks - regardless of their assigned system. In short, users could follow the directions provided by performers, but they didn't know *why* they were doing what they were doing.
- Users who were unable to complete testing found the system training documentation to be insufficient, to the point of limiting their ability to carry out their tasks. These users also felt unable to complete their assigned tasks within the scheduled time-frame.

- **System Functionality**

- Systems #1 and #2 were the only systems able to host users across all five datasets. Other systems were either unable to work with specific datasets (System #1 only worked with one of the five datasets, System #4 worked with four of the five datasets), or users were left waiting a long time for systems to load data augmentation, exploration, or modeling tasks.
- While systems undergo suitability testing before being put in front of users, system stability is often fragile. The capabilities that users required were often either not available or not functioning.⁵

System Functionality

"Getting a bad gateway, whole page went down, can't access it at all"

- Data Science user

"The system did not provide me with a solution and I waited for over 30 minutes to let it run"

- SME user

"First it failed when augmenting the data, then it failed when it got stuck on 'discovered solutions' under the results tab"

- Data Science user

"I never got a chance to see a single model produced."

- Data Science user

⁵ Parenthetic is unable to provide insight about whether these issues are related to system status or the evaluation environment.

EXECUTIVE SUMMARY

- **Data Problem Exploration**

- Overall, users seek to explore their data with the visualization and exploration features. Three of the four systems provided users the ability to explore their data and problems of interest. System #4 is unable to support the SME user cohort in this endeavor.

- **System Utility and Explainability**

- In general, users perceive AI as very powerful, but, when using it, they often feel that their expectations are not satisfied. Often, they are unwilling to adapt to the process/tool.
- Systems are focused on communicating *how* a model works instead of *why it generated the results that it did*. Users across all systems reported feeling unable to assess model usability. These models therefore lacked *explainability* - which is their key system requirement.
- User ratings appear to run on a continuum based on type of expertise:

Figure 2: Summary of open-ended responses to protocol questions.



Data/Problem Exploration

"I loved the "explore" function, though I wish it could have produced the graphs I wanted and that the graphs could have been used to include/exclude data from the final analysis."

- SME user

"Being able to play with the data by testing the models. Having the information stored after outputting a model and then tweaking it was really useful for tracking how effective my adjustments were"

- SME user

System Utility and Explainability

"I would have liked in the results view to have more interactive options in comparing the predictive power of variable I could self-assign." - SME user

"I was unable to assess how usable/workable this model was and unsure how to compare it with others including the individual specific variable factors in the Results screen."

- Data Science user

"I can't think of a circumstance where predicting the type of event and having the given variables available would be useful."

- Data Science user

"I interpret this as "there is a lot of variation", so these factors cannot be reliably used to make predictions."

- SME user

EXECUTIVE SUMMARY

System Recommendations

System #1

With only 4 users testing System #1 we are unable to draw any specific conclusions. However, the team should investigate data ingest issues that caused System #1 to only work with the ACLED dataset (a tabular dataset and classification problem).

System #2

System #2 received mostly positive results. We recommend system improvement to enable SMEs to derive meaningful results from system output. This issue can be addressed with improved training and a user interface (UI) that is sufficiently flexible to support different user profiles. For example, some users want to see features that are currently hidden on the UI, and want to hide the more advanced features that are currently visible. Other users want to see the opposite.

System #3

System #3 received positive ratings overall, and the highest ratings from both SME and data scientist users. However, System #3 users often reported that they were unable to finish due to long load times, errors, and other functionality issues. These users also specifically sought additional education or information about the system and the ability to manipulate data. This system requires improvements to the data ingest process. We also recommend this team pivot to the more complex tasks of improved system utility and model explainability.

System #4

System #4 received mostly negative ratings and requires significant improvements to meet the goals of the Program. Overall, System #4 underperformed other systems in model discovery. Furthermore, SMEs using System #4 especially reported that they were overwhelmed by information when compared to data scientists. The team could benefit from more frequent user feedback, and finding ways to meet the needs of non-data scientist users.

EXECUTIVE SUMMARY

Future Testing Recommendations

In Phase 2, TA3 performer teams have been directed to address underspecified and unsolved problems, improve data ingest capabilities by extending the types and quality of data their systems can handle, improve model explainability to SME audiences, and improve or add model deployment capabilities.

It is clear that performer teams need to find better ways to communicate the *why* to their systems' users, instead of exclusively focusing on the *how*. We know that humans of all levels of expertise are better able to derive meaning when they are given the bottom-line meaning (gist) of the output. Users are unable to do this if they are bombarded with a lot of verbatim information and scores, with no way to connect it to their questions of interest. Whereas the role of the algorithm is to order information, and rank score output, the user's role is to find the meaning and explain it to others. In the next Phase of the Program we seek to connect these two pieces and support the teams as they develop tools that will measurably benefit end users.

We propose the following evaluation updates to meet Phase 2 goals:

1. Provide Frequent Feedback

First and foremost, performer teams can benefit from **direct, timely, and frequent feedback on their tools from real world users across the range of expertise**, and iterative updating of their systems using that feedback.

2. Increase Accessibility

Getting systems "user-ready" for evaluation requires a lot of time and resources. Sequestering data from the performers serves a purpose, but one which can likely be accomplished in other ways. We can ensure system flexibility by **giving performer teams rapid and repeated access to a variety of domains, users, and datasets**.

3. Streamline Testing Processes

Another weakness is that users are unable to complete testing as quickly as desired. Recruiting users with the right background, or who are volunteering their time as a government SME, is a necessary part of the process. But even a moderate length of time between sign-up and task initiation leads to significant user drop-out.

- a. We will **adapt our training and evaluation procedures**, including revising the user protocol, to ensure a lower level of complexity and more streamlined user experience.
- b. We can **treat transition partner users differently than users from other sources**, and not overlap their activities. This would allow us to get a broader perspective, and faster, more direct user feedback to the performer teams. Proposed process:
 - i. For transition partner users: these SMEs are often small in numbers and are a better fit for direct focused **workshops or experiments** with hands-on training.
 - ii. For non-transition partner users: continue online **testing with a broad user base** to decrease the time it takes to get users through the sign-up to testing pipeline. Test new features and capabilities with these users first, so transition partners are accessing vetted systems. These online tests can also be used to hone instruments that will ultimately be used to assess system-user interaction quality.

EXECUTIVE SUMMARY

4. Better Characterize User Base

These shifts can help performer teams get feedback more readily, but that feedback must directly improve the current system weaknesses related to providing users with meaningful and explainable results – a prerequisite to a relevant data modeling and analysis experience. To meet that objective:

- a. **We will update our methods for characterizing the traits of our user base, and the distribution of expertise across different user types.** These insights will allow us, in conjunction with performer teams, to develop support systems and training that align with the various user traits, skills and perspectives and increase the likelihood of successful technology transition. It will also have the added benefit of providing the performer teams with different perspectives on how their tools can be used and potentially foster new areas of discovery.

FULL REPORT

BACKGROUND

About Us

Parenthetic works on research and development projects in both the public and private sectors focused on new methods or technology for information and communication. We translate end user requirements to technical teams and similarly translate technology capabilities and features to targeted end users. We maintain—and continue to grow—a database of “users” from across public, private, and academic institutions that are available for software testing and evaluation. This community spans many domains and levels of expertise, but users have one commonality - interest in new technology that optimizes human-machine interaction in their domain.

We develop operationally-relevant challenge problem sets and evaluation metrics that are rooted in real-world data and circumstances. Our goal is to inform and push the development of capabilities that will bring high fidelity to addressing national security interests.

On the D3M evaluation team, Parenthetic focuses exclusively on assessing TA3 system performance in supporting SMEs' creation of data-driven models and understanding model output. We want to do more for the end user than point them at an overabundance of output or alert them to what is in that pile. For machine learning tools to assist the user and provide them with explainable and meaningful results, the tools have to enable them to connect the dots, and allow them to answer their questions using their contextual knowledge and expertise. We are supported in this effort by Professor David Broniatowski, Director of the Decision Making and Systems Architecture Laboratory at the George Washington University (GWU) School of Engineering and Applied Science, and GWU doctoral candidate Lydia Gleaves.

For D3M, Parenthetic:

- Provides guidance to the evaluation team on the role of subject matter expertise in TA3 testing and related metrics, including:
 - Human-in-the-loop (HITL) practices
 - How domain experts understand model output; and how model output can be meaningful to experts
 - AI explainability for non-computer scientist communities
- Meets regularly with transition partners to develop TA3 evaluation challenge problems, identify relevant datasets and users, and work towards adoption of developed technology
- Recruits SMEs and data science users for TA3 system testing and evaluation timelines
- Develops targeted user training
- Creates and updates evaluation protocols
- Facilitates information exchange between users and technology developers
- Ensures TA3 testing compliance with IRB requirements
- Analyzes TA3 testing results and provides user-focused conclusions and recommendations for evaluation team

BACKGROUND

Evaluation Team

For this evaluation, Parenthetic worked in conjunction with the following members of the evaluation team:



Data Machines Corp Computing and Infrastructure

Data Machines Corporation was responsible for system staging; hosting all user interactions and system deployments in a cloud computing environment; supporting user troubleshooting, bug reporting and resolution; and collecting user model output.



MIT Lincoln Laboratory Dataset Management

MIT Lincoln Laboratory was responsible for managing the selection and preparation of relevant datasets for TA3 testing.



Program SETA Lead for Evaluation Team and Dry Run Planning

The program SETA developed and implemented a dry-run plan for evaluating technical system readiness for participation in the evaluation.

www.parenthetic.io

EVALUATION ELEMENTS

The primary focus of TA3 systems is to support SMEs with no data science expertise in automatically building empirical models and understanding their output. Specific outcomes of interest to DARPA as they pertain to this overall goal include:

01 Problem Formulation

Before model-building, can the user explore data, features, and engineer a problem?

02 Feature Importance

Does the system allow the user to determine feature importance in a model?

03 Prediction Performance

Does the system inform the user about prediction performance?

04 Model Explainability

Does the system explain why a model performs well or fails?

05 Model Selection

How well does the system guide a user in selecting a model?

To help answer those questions, we developed a test protocol to elicit user input in these areas.

Test Protocol

Our protocol was designed to index four key concepts for performer systems (see Table 2; full evaluation protocol in [Technical Appendix](#)). To test the validity of these concepts, we developed several questions for each one (both forward- and reverse-coded) and randomized the order of their presentation to users.

Pilot results indicate that these scales are highly reliable (see Table 2), meaning that future iterations of this protocol can be significantly shortened while enabling us to gather the same information.

Table 2: Protocol Key Concepts

Concept	Description	Example Protocol Question	Reliability of all items in this concept (Cronbach's α)
Meaningfulness of output	Systems provide output to the user that is contextualized, comprehensible, and capable of being expressed by the user to decision-makers. Users will be able to do something different as a result of the system output.	I can make sense of what the system's results are saying	0.97
System utility	Systems are able to provide output that is accurate and fit for purpose in answering the question(s) of interest	I am confident in my decision	0.97
Relevance	Systems allow the user to hunt through, edit, explore, munge the data to be sure they have what they need, and remove what they don't need to answer the question(s) of interest. Systems are able to provide added value by exposing previously hidden insights.	The model produced features which I deemed important	0.97
System functionality	Systems are operational and produce valid output.	The system helped me to hone the question(s) that I was asking	0.98

EVALUATION ELEMENTS

System Readiness

All TA3 systems underwent a system suitability assessment prior to user assignment. For this evaluation, the process entailed the following steps:⁶

1. **TA3 performer teams** performed preliminary testing using seed datasets and submitted their results. Systems passing this stage moved on to step 2.
2. **DMC** carried out smoke testing of systems using sequestered datasets. Systems passing this stage moved on to step 3. Systems that functioned on some datasets but not others were still able to move forward for evaluation on the subset of datasets that they passed.
3. This consists of two separate but simultaneous sub-steps:
 - a) The Program SETA carried out a detailed system review on factors such as accuracy of datamart datasets, reset buttons causing issues on the backend, and appropriateness of model explanations relative to the dataset.
 - b) **Parenthetic** conducted a review from the user perspective and ensured there were no major stumbling blocks to a user successfully completing the system test. At the end of this stage, recommendations were made to DARPA on which TA2/3 system combination for each sequestered dataset should be presented to users.
4. Systems making it to this stage were deemed suitable for presentation to users. TA3 software prototypes that made it past the “user-readiness” stage were tested by users while following a test protocol.

⁶ This process was cumbersome and will be streamlined over the coming months.

EVALUATION ELEMENTS

User Recruitment and Expertise

Throughout the life of the Program, there have been many opportunities for TA3 teams to interact directly with SMEs from across domains. We do this to help teams get a better perspective on how SMEs apply their insight, a step on the path to addressing current system shortcomings in model meaningfulness and explainability for all users. SME skills typically fall into the following categories:

Definite:

- Domain expertise
- Utilize a specified method, process or workflow to move from research/operational question to conclusion
- Able to explain their method and conclusions
- In-depth knowledge about their data and the questions they are seeking to answer within it

Maybe

- Statistical and data modeling expertise

Unlikely

- Able to convert their questions into data modeling or machine learning (ML) tasks

Parenthetic recruited participants for this evaluation through our online sign-up page at www.parenthetic.io/studies. Users were recruited based on interest and experience in specific domains. The domains and associated datasets used in TA3 evaluations were selected through criteria - such as availability and fit for purpose - developed by MIT LL and the evaluation team. Transition partner domains, datasets and challenge questions of interest were developed in close collaboration between transition partner representatives, the D3M SETAs, MIT LL and Parenthetic.

Enrollees from transition partner domains were recruited through designated “gatekeepers” who provide contact information for interested participants. Non-transition partner enrollees were recruited or selected from our larger user base on the basis of interest and experience. These users received additional screening questionnaires to ensure they represented the targeted end user population.

After sign-up, enrollees completed an informed consent form, scheduling form, and personality questionnaire - to gauge individual traits that may provide context about their expertise and how they synthesize information and visual data. All users were provided a unique id, testing schedule, and access to the online testing portal where they could access all of their evaluation tasks.

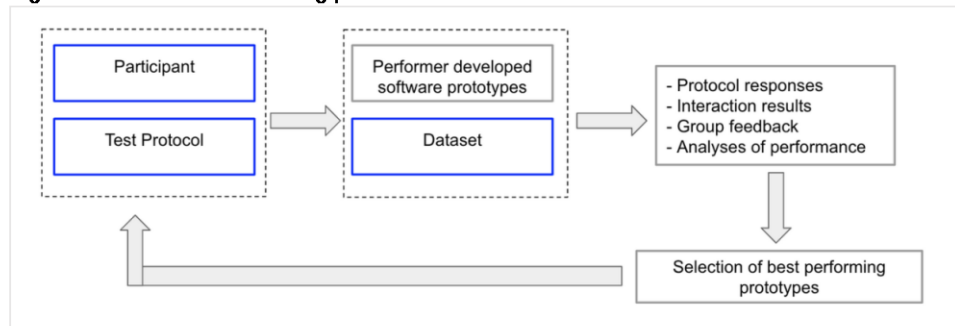
Users were assigned two TA3 systems and asked to review corresponding training material, developed by the performer teams. Once they completed their document review, users could access and test their two systems at their own convenience during a pre-scheduled week. Upon test completion, users filled out a post-test questionnaire designed to evaluate user protocol responses, model output scores, post-test questionnaire responses, and feedback about their experience through online group discussion. Users that did not complete their assigned tests were requested to fill out a feedback questionnaire. A full “user run” is user completion of their two assigned tests. Participants received payment for their time unless they were USG employees and opted out of payment.

EVALUATION ELEMENTS

Tasks

Users were asked to assess prototype software applied to a dataset in their area of expertise while following a common test protocol.

Figure 3: D3M TA3 User-testing process



* blue-lined boxes and dotted line containers highlight user-activity functions

Datasets

A variety of datasets were used across five domains for this evaluation. These datasets are described below. Three of these domains and associated datasets are linked to D3M transition partners, and are denoted with *.

ACLED

The Armed Conflict Location & Event Data Project (ACLED) is a conflict collection, analysis and crisis mapping project. ACLED collects the dates, actors, types of violence, locations, and fatalities of all reported political violence and protest events across Africa, South Asia, South East Asia, the Middle East, Europe, and Latin America. This includes events that occur within civil wars and periods of instability, public protest and regime breakdown. ACLED's aim is to capture the forms, actors, dates and locations of political violence and protest as it occurs across states.

Hate Crimes

FiveThirtyEight's hate crimes dataset combines data from a variety of sources, including the Kaiser Family Foundation, Census Bureau, the United States Elections Project, the Southern Poverty Law Center, and the FBI. The dataset includes key demographics, election shares, and hate crime rates for each US state. These data were described in a story on the FiveThirtyEight website, which also allows readers to investigate the story and make their own connections regarding the prevalence of hate in the United States.

TERRA*

Users were given the option to work with one or both of the following datasets:

1. **Plant Heights:** This dataset contains phenotypic features (i.e. canopy height measurements) for several genetic variants (cultivars) of the sorghum plant. The plant heights were measured by laser scanners over the course of 133 days. Each plot (site) contains a timeseries of height measurements (in cm) taken over the 133 day period.

EVALUATION ELEMENTS

2. **Plant Leaf Angles:** This dataset contains leaf angle measurements for several genetic variants (cultivars) of the sorghum plant. The measurements were measured over the course of 133 days. Each plot (site) contains a timeseries of leaf angle measurements (in normalized degrees taken over the 133 day period). The task is to predict the last leaf angle measurement of each plot/site.

Dataset #1

This dataset's details have been withheld from this report due to transition partner MOU.

Dataset #2

This dataset's details have been withheld from this report due to transition partner MOU.

System Descriptions

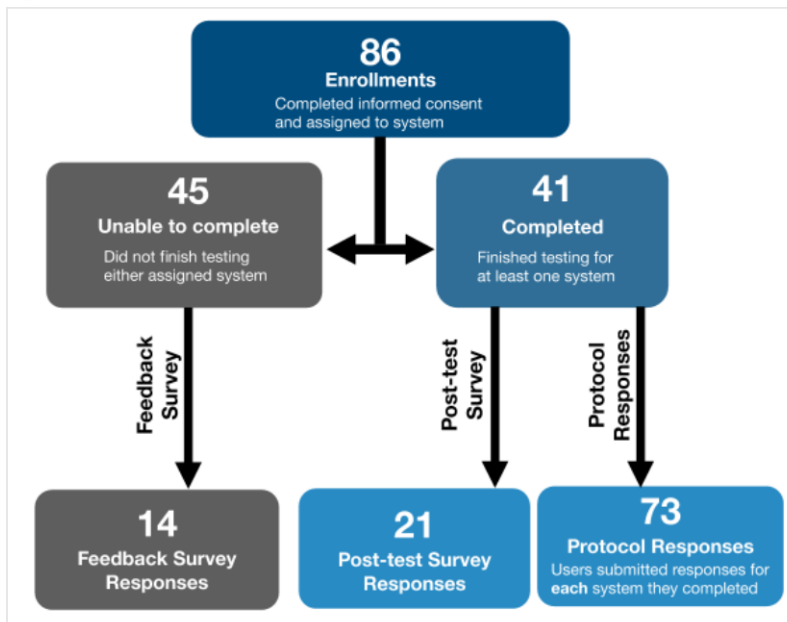
Four performer teams submitted systems that passed the readiness stage. Descriptions of each of these four systems (provided by each TA3 performer) are below:

1. **System #1:** System #1 will enable SMEs with no data science training to dynamically derive, understand and apply data-driven models in answering mission critical questions. This approach augments human intellect with data discovery, enrichment, analytic model recommendation, and automated visualization to discover observable relationships in data, empowering intelligence action.
2. **System #2:** System #2 provides a fundamentally re-thought data analytics stack: domain experts without programming experience or background in statistics can now explore and mine data through an intuitive user interface, and effortlessly build, analyze, and evaluate machine learning pipelines.
3. **System #3:** System #3 is a data-driven web application designed to bring researchers to insights fast. It facilitates intuitive machine learning, model discovery, and data exploration for researchers. As System #3's intelligent back-end automatically seeks interesting relationships in the data and builds models to predict outcomes, researchers impart substantive knowledge about their data and own research questions to guide the automated generation of AI assistance for data analysis in an interactive paradigm we call *human-guided machine learning*.
4. **System #4:** System #4 is designed to support the model building process and curation of ML data processing pipelines generated by AutoML systems. System #4 also integrates visual analytics techniques and allows users to perform interactive data augmentation and visual model selection.

System #4 was operational on only one evaluation dataset (ACLED). System #1 worked with four of the five datasets, and System #2 and System #3 were able to work for all five of the evaluation datasets.

User Participation

Eighty-six (86) users were recruited to participate in user testing across the five domains. 41 (48%) users completed both tasks. Of these 41 users, 23 (56%) were SMEs, 14 (34%) were data scientists, and 4 (10%) were data scientists with subject matter expertise. The remaining 45 users failed to carry out all assigned tasks.

Figure 4: Breakdown of Fall 2019 TA3 evaluation user base and associated tasks

EVALUATION ELEMENTS

Sample Demographics

Recruited participants were asked via pre-test questionnaires and screeners to provide their demographic and expert backgrounds. Of the 34 users who also responded to post-test feedback questionnaires, 16 (47%) users provided demographic information and 33 (97%) provided their areas of expertise. The demographic and expert breakdown of these users is as follows:

- Age: Of the 16 users who responded, the average age was **33 years** with a range of 20-74 years.
- Gender: Of the 16 users who responded, **15 (94%) were male**, and one was female.
- Expert Domain: Of the 34 users who responded, **nine (26%) were analysts from Transition Partner #1**
- Race: Of the 16 users who responded, **9 (56%) were white**
- Expertise: Of the 33 users who responded, **42% had expertise in data science**; similarly, **42% had expertise in social science** (see Table 3).

Table 3: Breakdown of user expertise by area

Area of Expertise	Total	Total
N=33	Frequency	Percent*
Applied sciences	2	6%
Data science	14	42%
Life sciences	4	12%
Mathematics	4	12%
Computer sciences	2	14%
Social sciences	14	42%
*Note: Respondents could select more than one area of expertise; percent total is greater than 100		

EVALUATION ELEMENTS

Analysis of Protocol

We created protocol questions to measure how well the performer systems meet the key concepts of meaningfulness, usefulness, relevance, and functionality. To score the systems, protocol items were converted to numeric values. Negatively-framed questions (e.g., "I did NOT understand the model output") were reverse-coded, so that positive experiences always resulted in a positive value. In order to summarize protocol responses, we conducted a principal component analysis (PCA) of protocol results (see [Appendix G](#) for detailed protocol mappings). We found that systems could be evaluated along 4 factors that emerged organically from protocol data (Table 4), explaining 71% of the variance in user responses. Next, we constructed scales by retaining those items that loaded uniquely above a value of 0.4 on each factor and removing items in a stepwise manner to maximize reliability (Cronbach's α). Finally, we averaged all protocol items that remained for each factor. These scales were highly reliable, and captured several concepts that were key to D3M's aims.⁷ Therefore, the remainder of our analysis uses scales derived from these factors to score each system. Specifically, to calculate the overall system ratings, all of the items that loaded on each factor were averaged.

Table 4: Factors used to evaluate systems

Factor	Description	Best-fitting protocol question	Reliability of all items in this factor, post-analysis (Cronbach's α)
Enabled Exploration/ Discovery	Systems enabled users to discover new insights	The system made it easy to find new sources of relevant data	0.99
Meaningfulness	Systems provided output that users found meaningful, made sense, and they could explain. System answered "why"	I CANNOT explain what these results mean	0.97
Explainability	Systems did not provide explainable output. System answered "how"	The system did NOT allow me to see how it made predictions	0.94
Right Amount of Information	Systems gave users the right amount of information	The system gave me an overwhelming amount of information	0.89

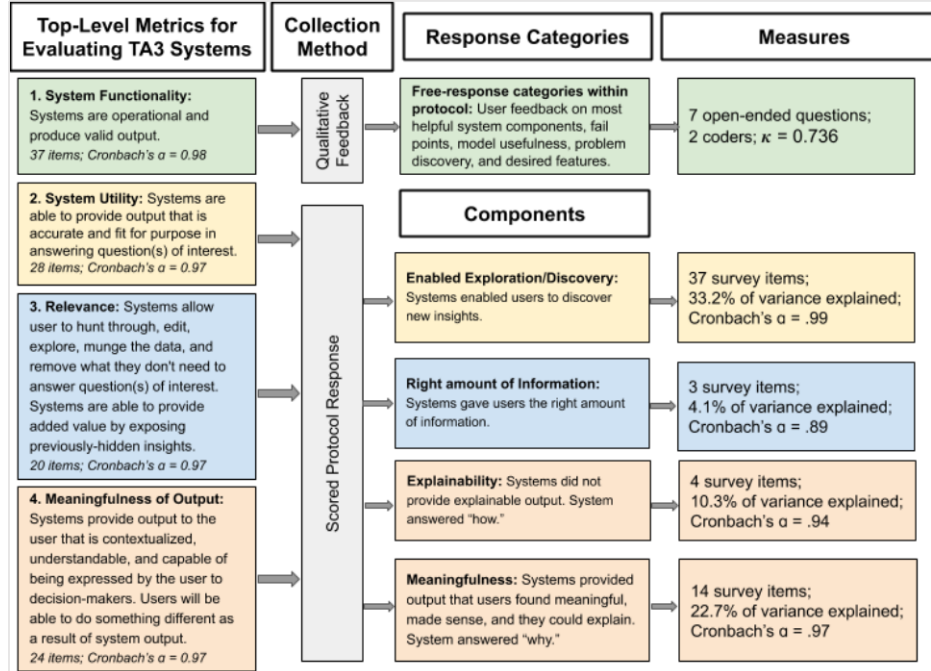
⁷ Most of these factors were distinct, yet correlated. Enabled Exploration/ Discovery was correlated with: meaningfulness, $r(73) = 0.74$, $p < 0.001$; explainability $r(73) = 0.25$, $p = 0.03$; right amount of information $r(73) = 0.60$, $p < 0.001$. Meaningfulness was correlated with explainability $r(73) = 0.41$, $p < 0.001$; right amount of information, $r(73) = 0.62$, $p < 0.001$. Explainability was marginally correlated with right amount of information, $r(73) = 0.23$, $p = 0.05$.

EVALUATION OUTCOMES

When combined with free-response questions, the overall analysis process is described below.

Figure 5: Process Model

Note: Item boxes are color coded to match themes. Components map to multiple top-level metric themes.



Overall System Performance

A series of one-way ANOVAs were conducted to compare the effect of the system being used on the components of utility. System #1 was not included in these analyses due to its low number of samples (System #1 was not functional on all but one dataset).

Systems differed from one another along the factor indexing the Enabling Exploration/Discovery component at the $p < .05$ level for the three conditions [$F(2, 66) = 3.80, p = 0.03$].

Post hoc comparisons using the Tukey HSD test indicated that the mean score for System #2, $M = 0.734, SD = 1.424$, was significantly different from the mean score for System #4, $M = -0.136, SD = 1.252, p = 0.047$.

System #3, $M = 0.686, SD = 0.891$ marginally outperformed System #4 along this dimension, but this difference was not significant after controlling for multiple comparisons, $p = 0.06$. We did not detect a significant difference between System #2 and System #3 along this dimension, $p = 0.99$.

We did not detect significant differences between systems along the Meaningfulness component at the $p < .05$ level, $F(2, 66) = 1.10, p = 0.34$, the explainability component, $F(2, 66) = 2.10, p = 0.130$, or the Right Amount of Information component, $F(2, 66) = 2.463, p = 0.093$. Finally, we were unable to test System #1's performance relative to the other systems due to small sample size ($n=4$) and the fact that System #1 was only operational on ACLED data.

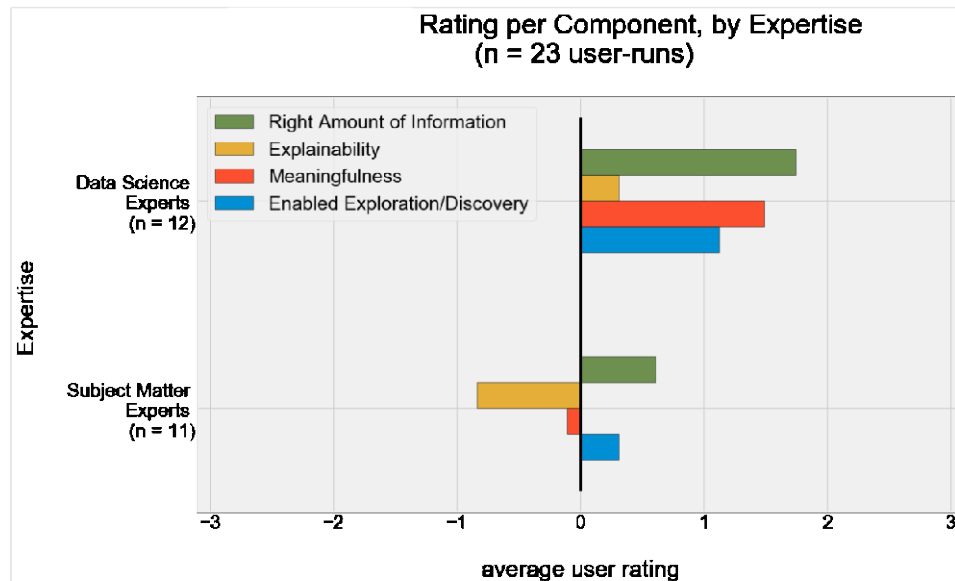
EVALUATION OUTCOMES

Protocol Results by Type of Expertise

The below graphs illustrate how well each system supports users based on the type of their expertise – either a) subject matter expertise in the given domain (and dataset used), or b) data science expertise. Overall, we found that data scientists found systems to be more useful for exploration, $t(71) = 2.52, p = 0.01$, more meaningful, $t(71) = 4.08, p < 0.001$, and were less overwhelmed by information, $t(71) = 4.05, p < 0.001$, when compared to SMEs. We detected only a marginal difference between SMEs and data scientists when evaluating system explainability, $t(71) = 1.90, p = 0.07$. These results suggest that systems may have prioritized describing algorithmic details over bottom-line meaning, leaving SMEs – who lacked the data science expertise to interpret these algorithms – confused and unable to assess or explain the model output. Below, we drill down to describe the performance of each system along each of these dimensions.

System #2

Graph 2: System #2 Rating by Expertise

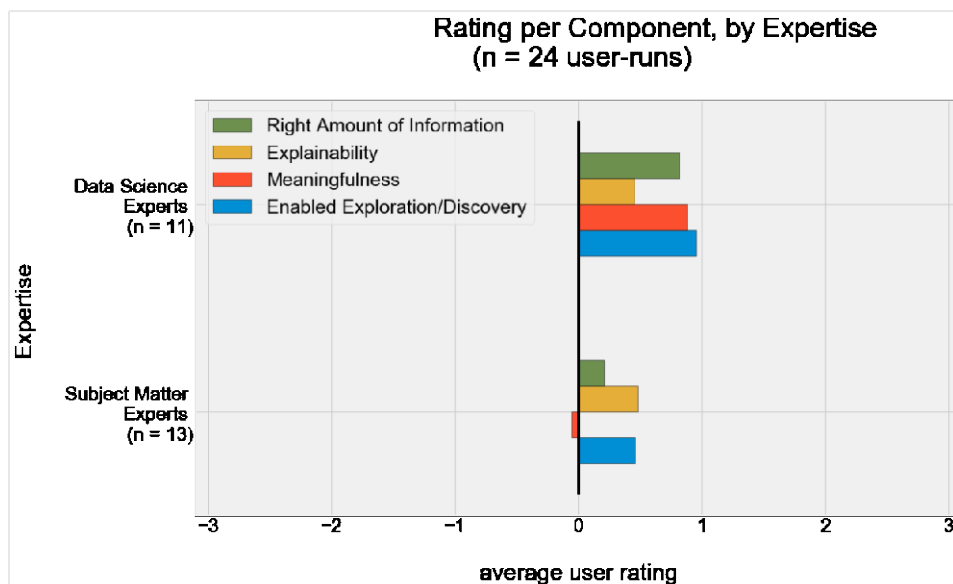


Data scientists found System #2's output significantly more meaningful, $t(21) = 2.79, p = 0.01$, than SMEs. SMEs were marginally, although not statistically significantly, more likely to become overwhelmed by information, $t(21) = 2.02, p = 0.06$, while using System #2 and to find the system's results explainable, $t(21) = 1.90, p = 0.07$. No significant differences were found between SMEs and data scientists in terms of data/problem exploration $t(21) = 1.40, p = 0.17$.

EVALUATION OUTCOMES

System #3

Graph 3: System #3 Rating by Expertise

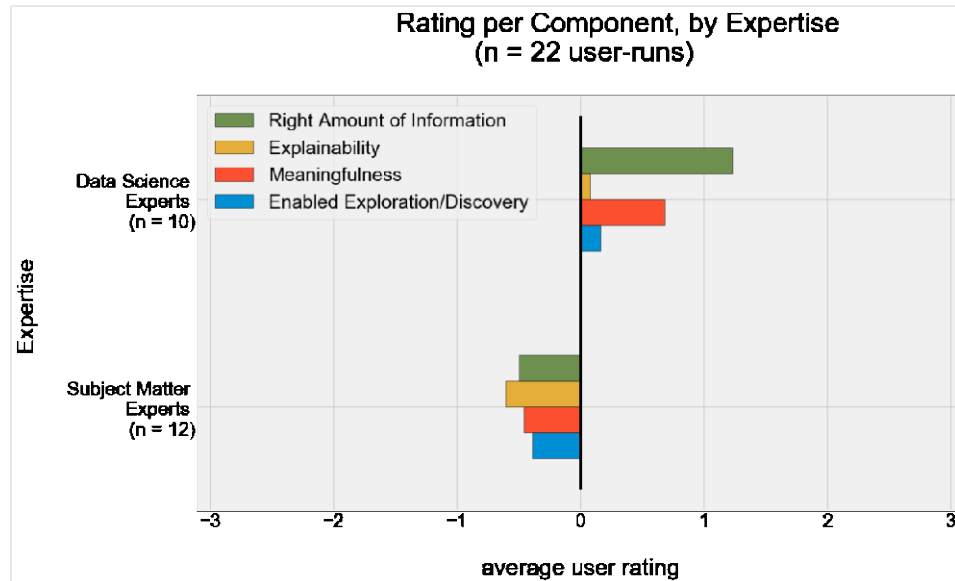


No statistically significant differences were found between SMEs and data scientists when evaluating System #3 in terms of enabling exploration/discovery, $t(22) = 1.38$, $p = 0.18$; meaningfulness, $t(22) = 1.71$, $p = 0.10$; explainability, $t(22) = -0.05$, $p = 0.96$; or having the right amount of information, $t(22) = 1.04$, $p = 0.31$.

EVALUATION OUTCOMES

System #4

Graph 4: System #4 Rating by Expertise

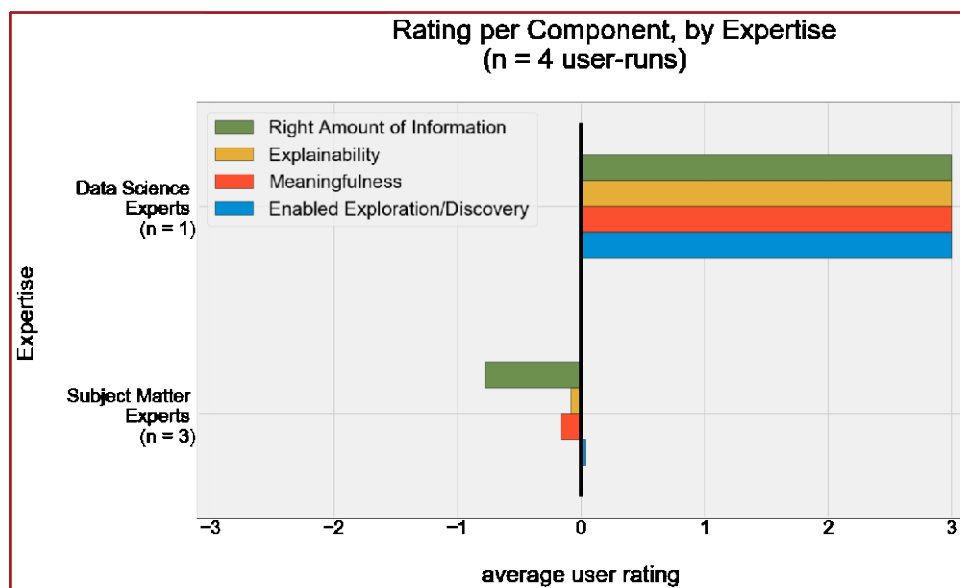


Data scientists were less likely than SMEs, $t(20) = 3.27, p = 0.004$, to find themselves overwhelmed by information when using System #4 and found its output marginally, although not significantly, more meaningful, $t(20) = 1.85, p = 0.08$. No significant difference was found between SMEs and data scientists when evaluating problem/data exploration, $t(20) = 1.03, p = 0.31$, or system explainability, $t(20) = 1.09, p = 0.29$. Notably, SMEs rated System #4 negatively on average along all dimensions, whereas data scientists rated all other systems positively on average.

EVALUATION OUTCOMES

System #1

Graph 5: System #1 Rating by Expertise



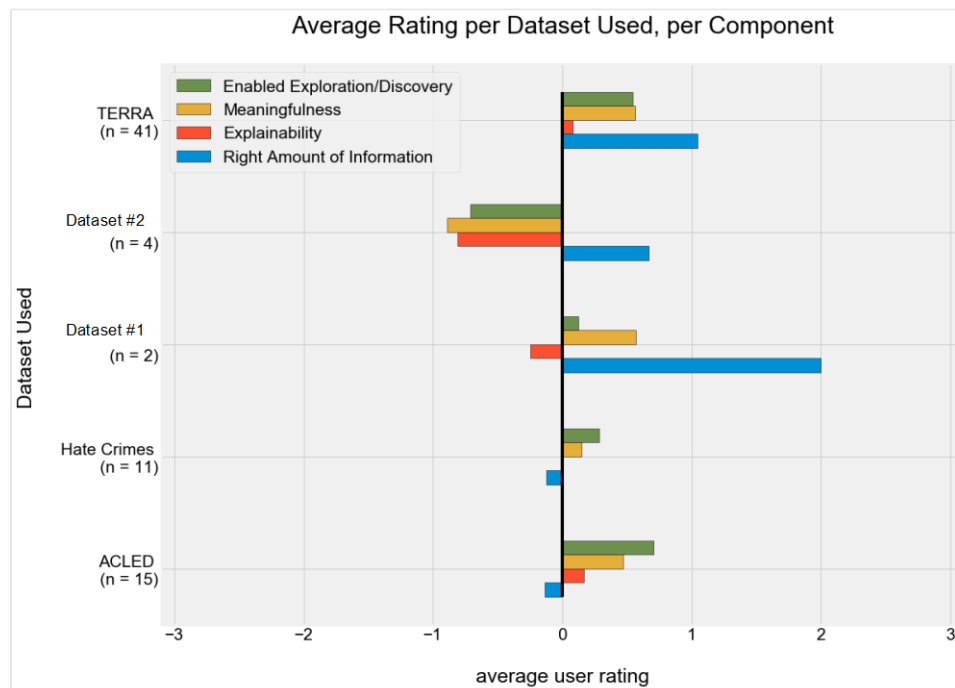
No statistical tests were run comparing data scientists' and SMEs' responses to System #1 due to small sample sizes and ability to run on only one dataset. The one data science user rated all four systems positively. Note that all output for System #1 in this document is highlighted in red to alert the reader to the low user numbers.

EVALUATION OUTCOMES

Summary Protocol Results by Dataset

The below graph shows component ratings by dataset for all systems combined. This visualizes how each group of users felt about systems as a whole. Of note, only the TERRA dataset was tested with sizeable samples.

Graph 6: Average Rating by Dataset - All Systems Combined



These results show a mix of ratings across the tested domains. The ACLED and TERRA users rated their systems higher across all components. The few Dataset #2 users who completed testing had mostly negative ratings of systems' capabilities. With only two Dataset #1 users completing testing, we do not recommend making any conclusions about their results.

These results show that System #1 and System #2 receive the highest ratings from users across datasets. However, the System #1 result cannot indicate any specific finding or recommendation due to the low number of users, and System #2 received low ratings from Dataset #2 users, who reacted negatively to all systems. System #3 performed well with the TERRA and Hate Crimes users - the other datasets had too few users to conclude. System #4 shows more mixed results.

EVALUATION OUTCOMES

Qualitative Findings

The protocol contained seven (7) open-ended questions to elicit additional feedback on users' experiences. Seventy-nine (79) respondents submitted feedback; however, not every respondent answered all seven questions. Responses were independently coded by two researchers and analyzed for intercoder reliability ($\kappa = 0.736$ – indicating substantial agreement). With the exception of System #1, for which there were only four responses, the number of responses was comparable across systems: 25 System #2 respondents, 26 System #3 respondents, and 24 System #4 respondents. Dominant themes within five primary areas of interest are described below:

- **Most Helpful Component.** Users were asked if there was one specific component of this system they found to be the most helpful aspect as they carried out their tasks.⁸
 - On average, users identified systems' abilities to **visualize data**, provide **models** and offer workable user **interfaces** as the most helpful components (27%, 27%, and 25% respectively).
- **Point of Failure.** Users were asked whether the system failed at any point during their test.⁹
 - Across systems, users identified **loading** issues and unknown **system errors** as the most common points of failure within systems (39% and 29% respectively).
- **Useful Model.** Users were asked if they felt the system provided a useful model.¹⁰
 - Across systems, the most dominant response among users was that they had **difficulty interpreting models** the systems provided (32% of all responses).
 - **Systems #3 and #4** users reported that they were **not able to finish** due to long load times, errors, or other issues, and they were thus unable to assess model usability. Notably, this theme emerged as the second most-dominant response across all systems (21%), despite only System #3 and #4 respondents identifying it.
- **New Discovery.** Users were asked if any features of the system were helpful in discovering new and interesting problems.¹¹
 - Across systems, users identified the ability to explore **relationships between variables**, discover or **explore data generally**, and use a functioning **interface** as helpful features in discovering new problems (30%, 24%, and 22% respectively).
- **Desired Features.** Users were asked if there were any system features they wish they had in their tested systems, that might have been helpful in identifying new tasks.¹²
 - The ability to **manipulate data** was the most dominant theme across all respondents (35%). This was the most dominant response among System #4 respondents (15% of all responses; 40% of System #4 responses).

⁸ There were 48 responses with the following breakdown: System #1 - 2, System #2 - 17, System #3 - 17, and System #4 - 12.

⁹ There were 38 responses with the following breakdown: System #1 - 1, System #2 - 14, System #3 - 9, and System #4 - 14).

¹⁰ There were 28 responses with the following breakdown: System #1 - 1, System #2 - 8, System #3 - 11, System #4 - 8.

¹¹ There were 37 responses with the following breakdown: System #1 - 3, System #2 - 14, System #3 - 14, System #4 - 6.

¹² There were 26 responses with the following breakdown: System #1 - 0, System #2 - 9, System #3 - 7, System #4 - 10.

EVALUATION OUTCOMES

Post-test User Feedback

Following each testing period, users who completed all tasks received a post-test questionnaire. Users who either didn't start or complete testing were sent a feedback questionnaire. Twenty (20) users responded to the post-test questionnaire, and 14 responded to the feedback questionnaire.

On average, users who completed testing tended to be younger than those who did not complete testing. The average age of users who completed tasks and responded to the questionnaire was 29 years, with a range of 20-51 years, compared to an average age of 45 years—with a larger range of 26-74 years—for non-completers. Specifically, the average age of those who completed testing ($M=28.8$, $SD=6.18$) was significantly lower than the average age of those who did not complete testing ($M=45$, $SD=20.6$; $t(14)=-2.25$, $p=0.02$). Additionally, the proportion of users who identified as data science experts was significantly higher among those who completed ($M=0.7$, $SD=0.47$) compared to those who did not ($M=0.00$, $SD=0.00$; $t(32)=5.55$, $p=0.00$).

When taking data science expertise and age into account with which systems users were testing, data science expertise was associated with an increased likelihood of completion ($\beta=0.52$, $p=.027$). This model accounts for 72% of the variance in the data (see appendix). Additionally, two domains had significant effects on users' likelihood of completion. Dataset #2 users were less likely to complete [$M=.11$, $F(1,32)=16.36$, $p=0.00$], while TERRA users were more likely to complete [$M=.88$, $F(1,32)=13.83$, $p=0.001$] (see appendix). Notably, 88% of TERRA users also had data science expertise. Despite limitations due to a small sample of respondents, these findings suggest that data science expertise—more-so than system or domain—is a primary factor in users' ability to successfully complete their tasks.

Users who completed testing

Overall, the majority of the 20 users who completed testing across systems and submitted feedback either agreed or strongly agreed that both the protocol and workflows were easy to follow (12 [60%] and 18 [90%] users, respectively). One difference worth noting, however, is that half of users who completed System #3 as their first system (four of eight) did *not* think the protocol was easy to follow. This is likely due to the length of the protocol and potential clash between it and the System #3 onscreen help process.

Most users reported that they learned some aspect of data exploration through testing, including: separating data of interest, identifying variables of interest, performing statistical analyses, transforming data, and interpreting trends. Only 4 users (20%) reported that they did not learn anything related to data exploration while taking part in this study. Similarly, most users (65%) reported that they gained a better understanding of how to visually explore their data. A smaller portion of users reported that they gained insight into how their expertise can improve algorithms for use in their field (40%), how visual models are built (45%), better ways to mining their data (30%), how predictive models are built (25%), an appreciation for data augmentation (35%), and ways they could use data modeling in their work (35%).

A majority of users (12 users; 60%) who completed testing felt **lost or confused** while completing their tasks, and 40% of users reported confusion specifically with **understanding their tasks**. However, which system they were testing did *not* have a significant effect on whether they were more likely to report feeling lost or confused, indicating room for improvement in the user preparation and training process.

EVALUATION OUTCOMES

Users who did not complete testing

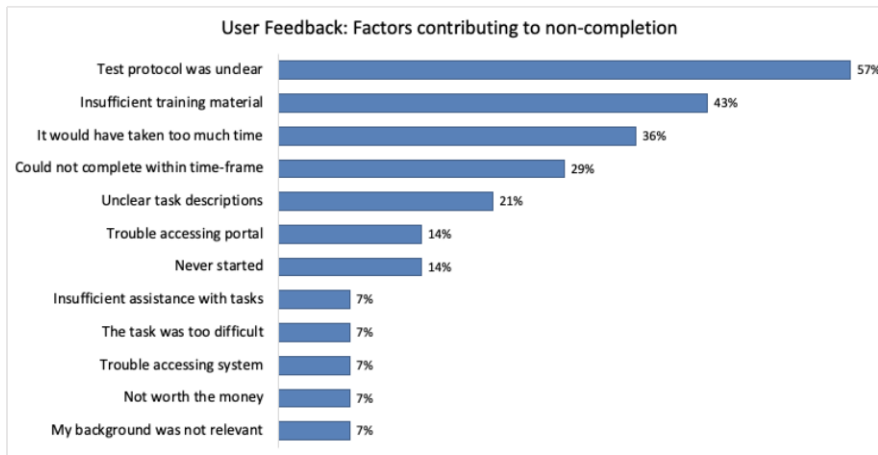
Of the 14 users who did not complete testing *and* responded to the feedback questionnaire, two never started. Among those who did not complete testing, most users had expertise in the social or life sciences (70%), and none had expertise in data science. By the end of the testing period, most (11 users; 78%) started testing their first assigned system, and six (43%) started testing their second assigned system.

A majority of users who did not complete testing felt the **system training documentation was insufficient** and limited their ability to carry out their tasks (10 users; 71%), and eight users (57%) felt the **test protocol was unclear** or confusing. While providing additional comments, four users (29%) also identified a general lack of clarity regarding system training and/or **not understanding what was expected** of them.

Time was also a major contributing factor for most users who did not complete testing, with eight people (57%) reporting that the tasks either would have taken too much time or that they could not complete the tasks within the scheduled time-frame.¹³

Within this limited sample, system assignments did not have a significant effect on whether users were more likely to report insufficient system training documentation or concern with time commitments.

Graph 7: Summary of factors contributing to task non-completion



¹³ While the amount of time each user had to complete their two test runs was not limited, users were asked to participate within the week-long timeframe they selected during the sign-up process.

LESSONS LEARNED

The fall evaluation helped further refine the overarching programmatic lessons to date. Systems are not currently aligning with SME needs - they prioritize algorithmic factors, while the users want meaningfulness and the ability to explore. SME users, in particular, are left feeling like a “button pusher” for the system to build a model, unable to derive meaning from system results and unable to assess or explain the model output. To address this, it is recommended that Performer teams improve techniques for communicating bottom-line meaning to users. Without insight from model output, users are left having to fill in too much information, all while using a process that is unfamiliar to them.

Users are eager to learn about how a model is built, what features are the most powerful and why, and what the output is based on. But, if the bar is too high, users will not adopt the software. They want to be able to explore and discover, and they don't want to be overloaded with information. They are looking for concise explanations for *why* the system did what it did (instead of *how*)- whether during data exploration, model building, or reviewing output and results - and they need it to be meaningful, so they can make their conclusions and explain to others. System output and scores should be communicated in meaningful terms, and tied to the original question.

FUTURE CONSIDERATIONS

To meet program goals, Parenthetic plans to:

1. Provide performer teams with more rapid feedback from a large array of subjects across different formats, such as small groups interactions, interviews, and workshops.
2. Use small group events to identify factors that require in-depth investigation, and follow up with periodic online system testing or questionnaire - testing can be staged with a dataset from a domain (not dataset) known to the performer teams, and equal numbers of SME and data science evaluators (ensuring replicability).
3. Expand the topic space and bring in new users, add new domains and dataset types (to address data ingest issues, model explanation and model deployment).
4. Improve testing procedures to make certain that users are directly focused on the strengths and weaknesses of the systems rather than the testing harness.
5. Find the right dataset balance so that teams have access to a variety of problems and data - ensuring system flexibility - while also allowing them enough time with open datasets to understand operational conditions (addressing model explanation).
6. Test with equal representation of user types—e.g., SMEs, Data Scientists, and Data Scientist SMEs— and gain a detailed understanding of the traits and characteristics of each group (addressing model explanation).
7. Develop standard user profiles indexing the level of data science expertise (ie: low vs high) to be used for evaluations (addressing model explanation).

APPENDICES

Appendix A

Independent Institutional Review Board

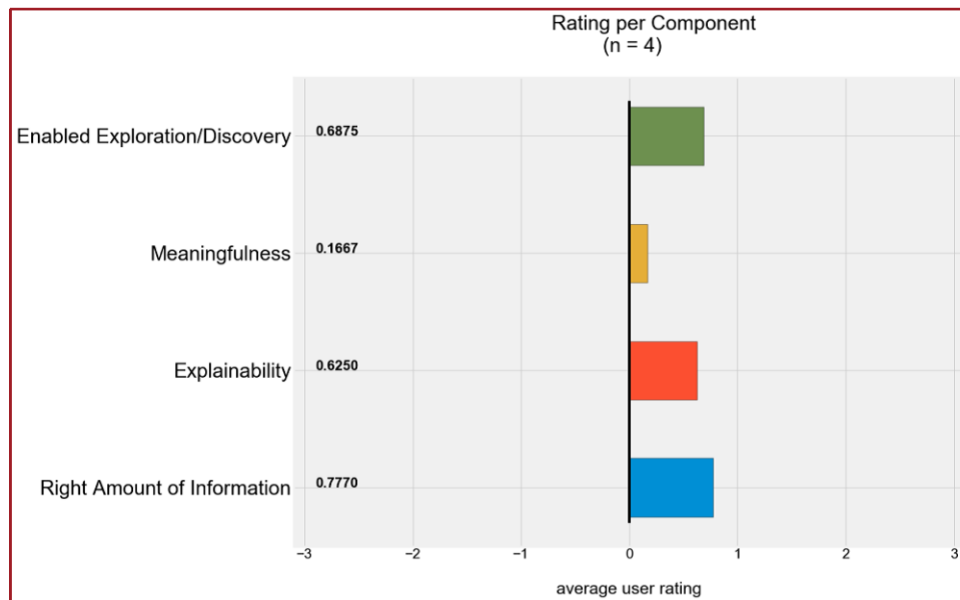
On June 20, 2019, Parenthetic received expedited human subjects research approval from IntegReview to carry out our evaluation work on D3M. Based on this approval, Parenthetic also received concurrence through the Air Force Research Laboratory on August 8, 2019 (documents available upon request). The approval covers all procedures and documentation used in the D3M user evaluations, including the test protocol.

Appendix B

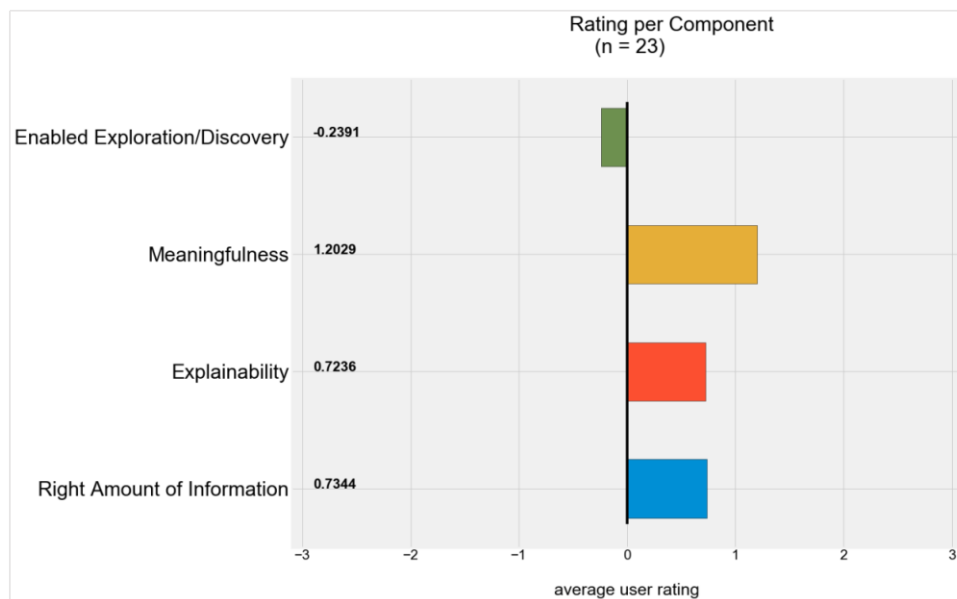
System Summary Scores by Component

Graph 8: System #1 Ratings by Component

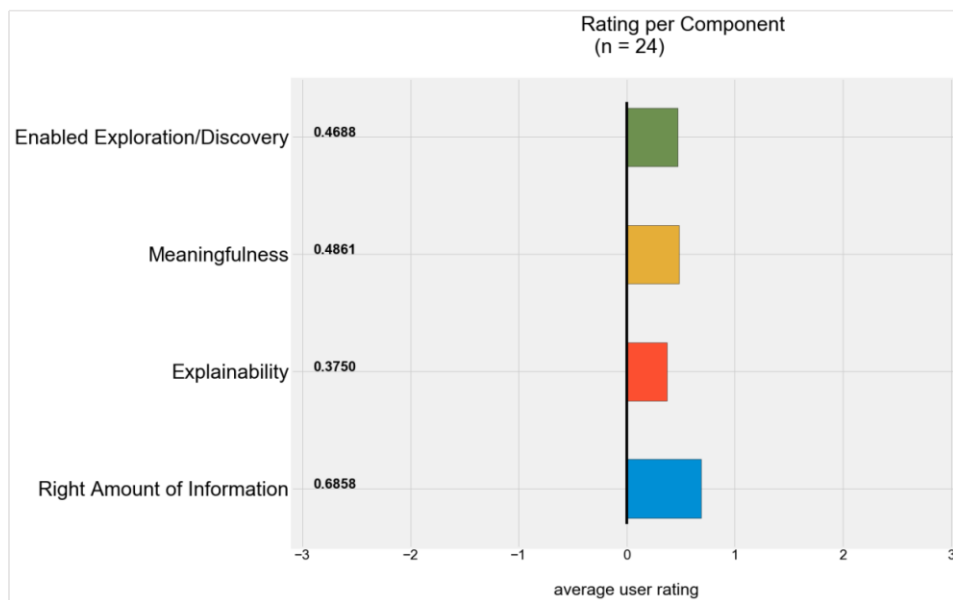
System #1

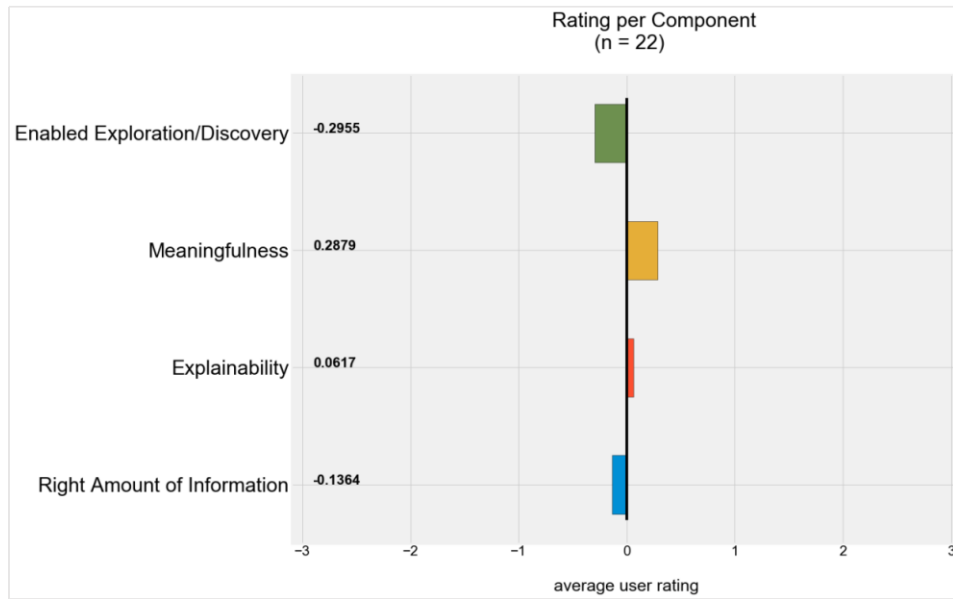


Graph 9: System #2 Ratings by Component

System #2

Graph 10: System #3 Ratings by Component

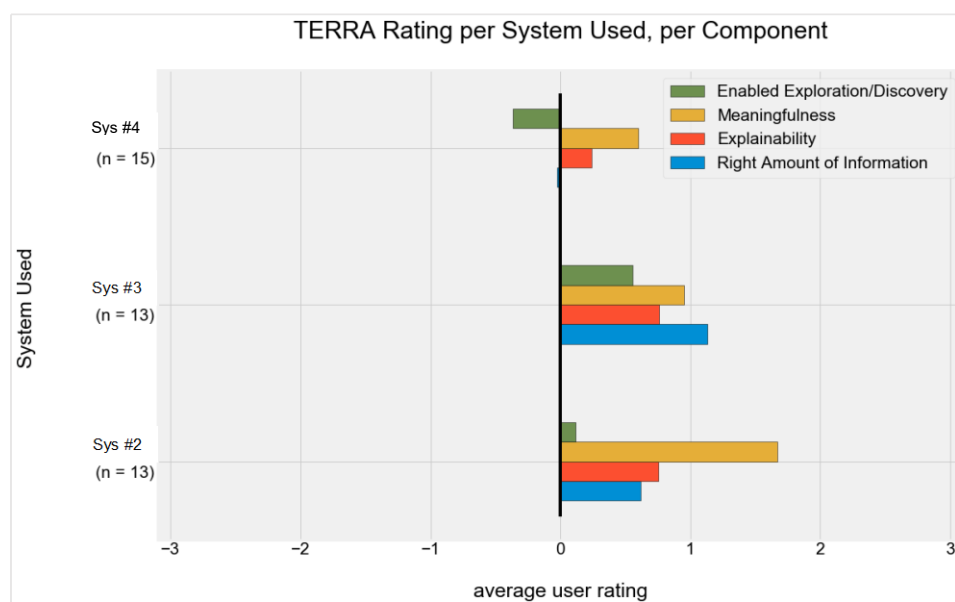
System #3

Graph 11: System #4 Ratings by Component**System #4**

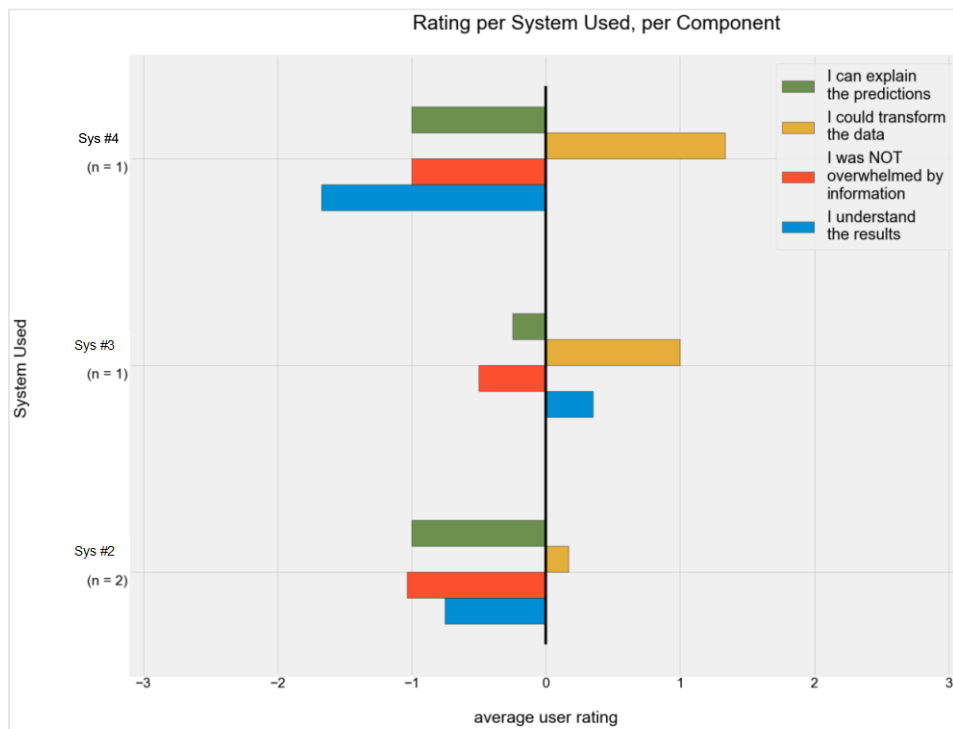
Appendix C

Protocol Output Scores by Dataset

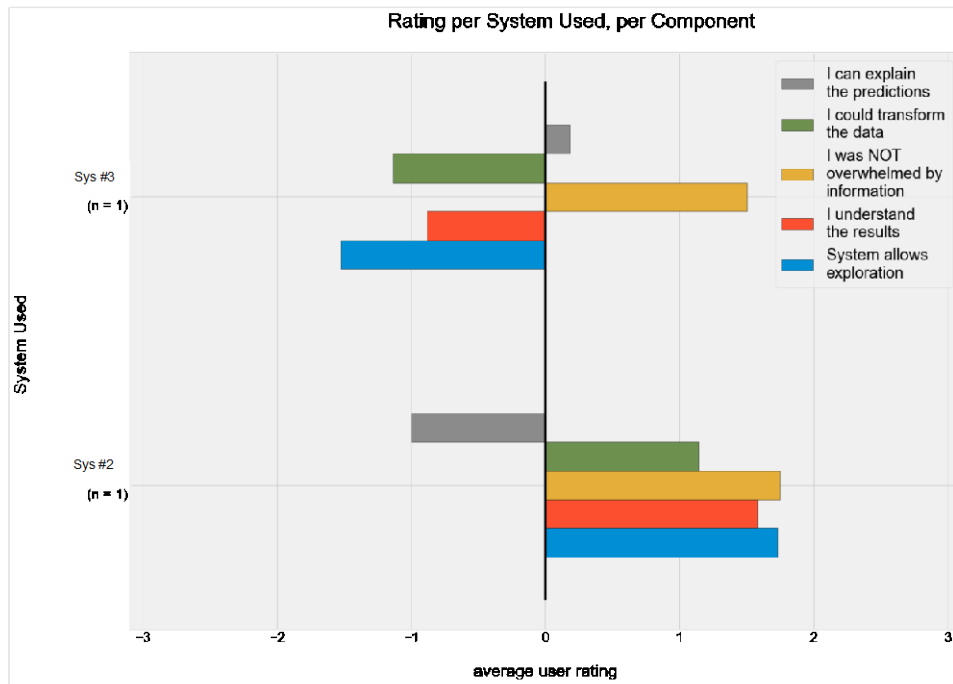
Graph 12: System Ratings by TERRA users



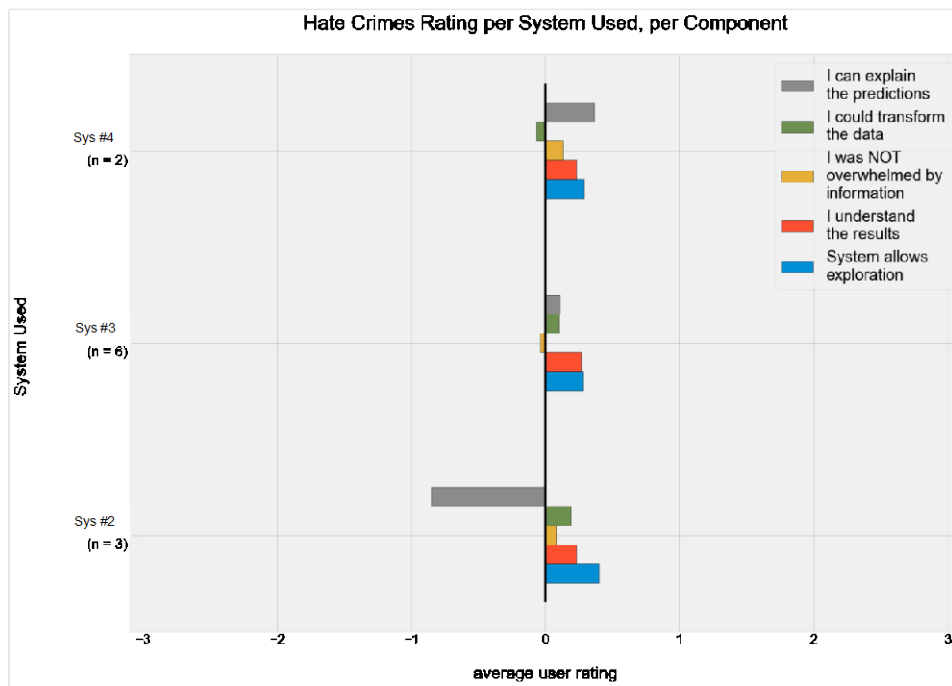
Graph 13: System Ratings by Dataset #2 users



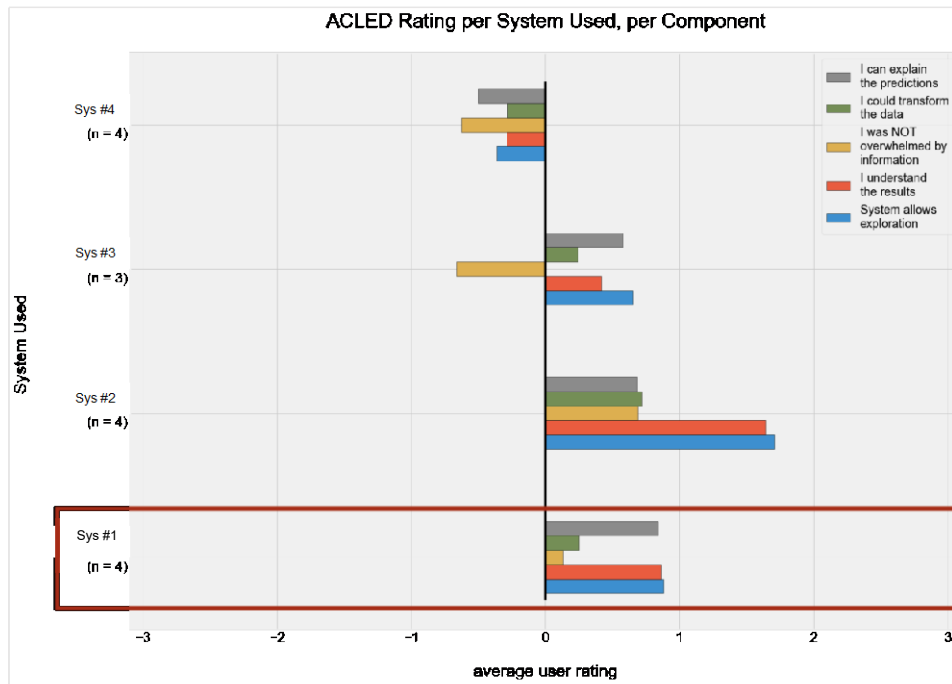
Graph 14: System Ratings by Dataset #1 users



Graph 15: System Ratings by Hate Crimes users



Graph 16: System Ratings by ACLED users

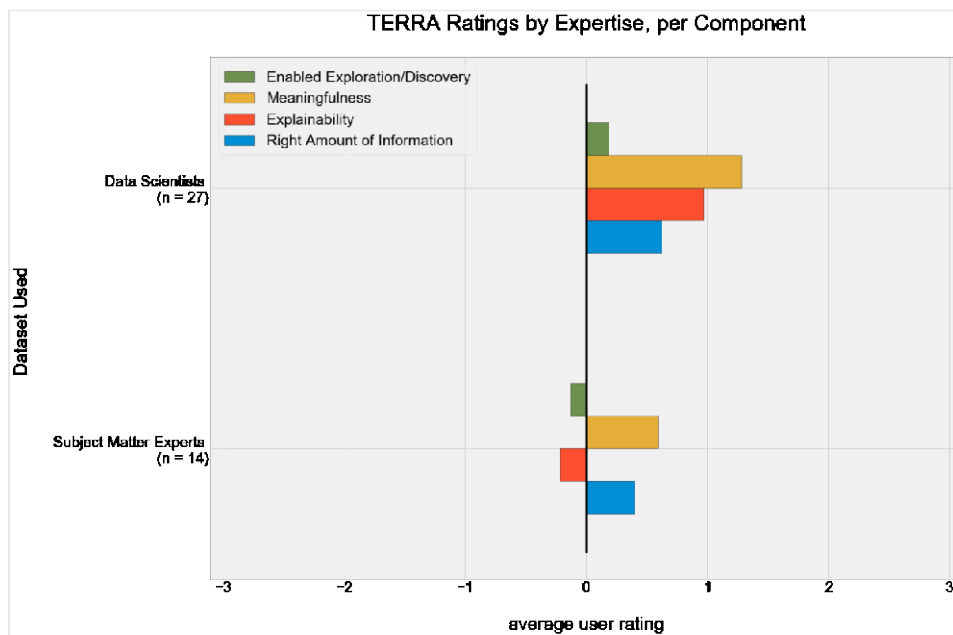


Appendix D

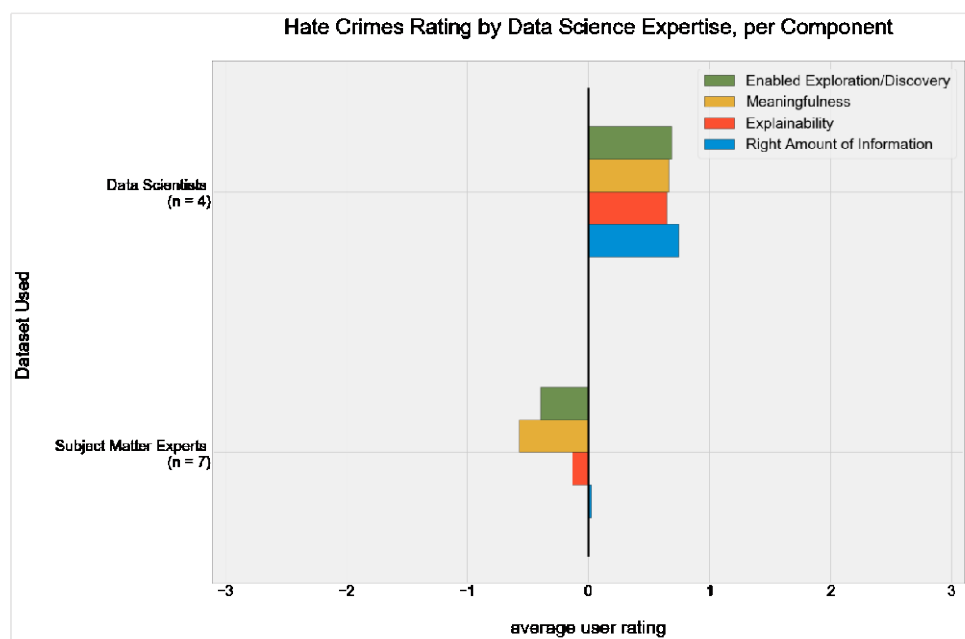
Protocol Output Scores by Type of Expertise

Below graphs show TERRA, Hate Crimes, and ACLED user ratings of systems broken down by type of expertise. The users from both Datasets #1 and #2 were domain experts, there were no data science users, so results are not shown for these datasets.

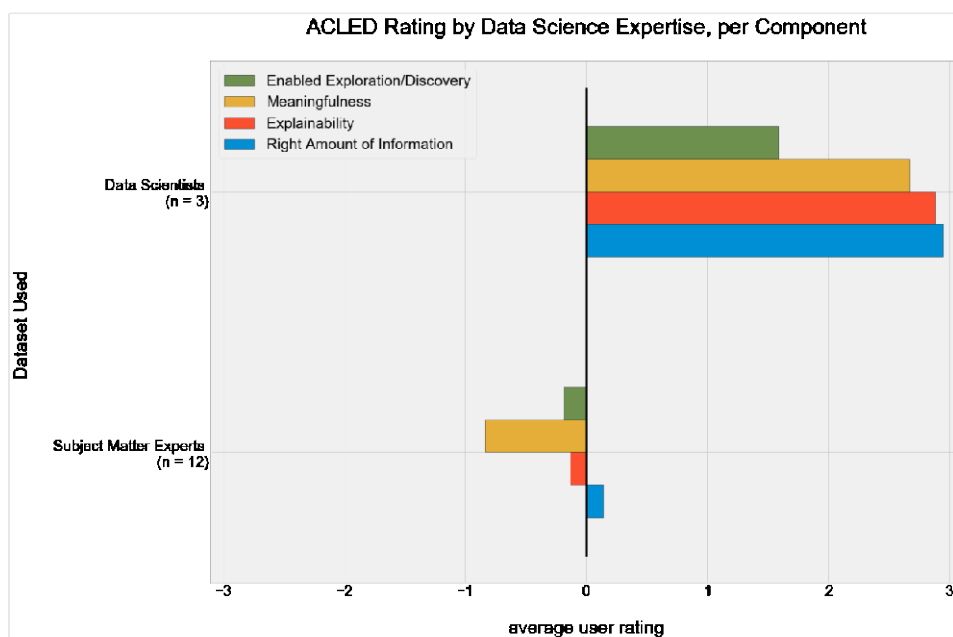
Graph 17: TERRA Users' System Ratings by Expertise



Graph 18: Hate Crimes Users' System Ratings by Expertise



Graph 19: ACLED Users' System Ratings by Expertise



Appendix E

Statistical output for user feedback questionnaire

Regression: Completion status	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		
	B	Std. Error	Beta			Lower Bound	Upper Bound	
(Constant)	0.776	0.51		1.521	0.163	-0.378	1.93	
System #2	-0.293	0.266	-0.264	-1.102	0.299	-0.896	0.309	
Sys1_System #3	-0.207	0.268	-0.231	-0.771	0.46	-0.814	0.4	
Sys2_System #3	0.274	0.224	0.293	1.221	0.253	-0.233	0.781	
Sys2_System #4	0.241	0.232	0.258	1.035	0.328	-0.285	0.766	
DataSciExpert	0.52	0.197	0.581	2.636	0.027	0.074	0.966	
Age	-0.012	0.01	-0.331	-1.208	0.258	-0.033	0.01	
a Dependent Variable: Complete								
Model Summary								
R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
.845a	0.715	0.525	0.308	0.715	3.76	6	9	0.037
a Predictors: (Constant), Age, Sys2_System #3, Sys1_System #2, DataSciExpert, Sys2_System #4, Sys1_System #3								
b Excluded Variables: Sys1_System #4, Sys2_System #2								

Group Statistics					
	Comple e	N	Mea n	Std. Deviation	Std. Error Mean
Age	1	12	28.83	6.177	1.783
	0	4	45	20.575	10.288
DataSciExpert	1	20	0.7	0.47	0.105
	0	14	0	0	0

Independent Samples T- Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2- tailed)	Mean Differenc e	Std. Error Differenc e	95% Confidence Interval of the Difference	
									Lower	Upper
Age	Equal variances assumed	6.906	0.02	-2.549	14	0.023	-16.167	6.343	-29.77	-2.563
	Equal variances not assumed			-1.548	3.182	0.214	-16.167	10.441	-48.344	16.01
DataSciExpert	Equal variances assumed	69.176	0	5.545	32	0	0.7	0.126	0.443	0.957
	Equal variances not assumed			6.658	19	0	0.7	0.105	0.48	0.92

Mean Completion Status by Domain*				ANOVA:					
	N: # Respondents	Mean Completion Rate	% respondents with data science expertise:		Sum of Squares	df	Mean Square	F	Sig.
Domain_ ACLED	4	0.75	0%	Between Groups	0.051	1	0.051	0.468	0.499
				Within Groups	3.479	32	0.109		
				Total	3.529	33			
Domain_ HateCrimes	4	0.25	0%	Between Groups	0.222	1	0.222	2.151	0.152
				Within Groups	3.307	32	0.103		
				Total	3.529	33			
Domain_ Dataset #2	9	0.11	0%	Between Groups	2.239	1	2.239	16.364	0.000
				Within Groups	4.379	32	0.137		
				Total	6.618	33			
Domain_ TERRA	16	0.88	88%	Between Groups	2.556	1	2.556	13.831	0.001
				Within Groups	5.914	32	0.185		
				Total	8.471	33			
*Note: No questionnaire respondents were assigned Dataset #1's domain									

ANOVA Tables, System Effects on Component Ratings

ANOVA (Enabled Exploration/Discovery)

Source	df	Sum of squares	Mean squares	F	Pr > F
Between Groups	2	10.752	5.376	3.705	0.030
Within Groups	66	95.776	1.451		
Corrected Total	68	106.529			

Tukey Test (Enabling Exploration/Discovery)

Contrast	Difference	Standardized difference	Critical value	Pr > Diff	Significant	Lower bound (95%)	Upper bound (95%)
System #2 vs System #4	0.871	2.424	2.398	0.047	Yes	0.009	1.732
System #2 vs System #3	0.049	0.138	2.398	0.990	No	-0.794	0.891
System #3 vs System #4	0.822	2.312	2.398	0.061	No	-0.030	1.675

ANOVA (Meaningfulness)

Source	df	Sum of squares	Mean squares	F	Pr > F
Between Groups	2	4.936	2.468	1.101	0.339
Within Groups	66	147.988	2.242		
Corrected Total	68	152.925			

ANOVA (Explainability)

Source	df	Sum of squares	Mean squares	F	Pr > F
Between Groups	2	8.501	4.250	2.10	0.130
Within Groups	66	133.303	2.020		
Corrected Total	68	141.804			

ANOVA (Right Amount of Information)

Source	df	Sum of squares	Mean squares	F	Pr > F
Between Groups	2	10.551	5.275	2.463	0.093
Within Groups	66	141.336	2.141		
Corrected Total	68	151.887			

Appendix F

Protocol Item-Concept Mappings

In the table below, each item is represented in its short form (full item text is shown further below in appendix). The nine right columns represent the nine major concepts captured by the protocol, with an “x” marking those captured by each item:

1. Meaningfulness
2. Usefulness
3. Relevance
4. Functionality
5. Enabling Exploration/Discovery
6. Meaningfulness-2
7. Explainability
8. Right Amount of Information

Item Short Form	M	U	R	F	EE/D	M-2	E	RAI
System helped find new relevant data		x	x	x				
System didn't help find new relevant data		x	x	x				
System didn't show something new, novel								
System helped look at data in new way								
System helped ask novel questions								
System didn't provide fresh perspective								
System didn't help identify many new hypotheses								
System helped think of new data sources				x				

Item Short Form	M	U	R	F	EE/D	M-2	E	RAI
System didn't help think of question in new way								
System didn't help find new data		x	x	x				
System helped think of something new								
System helped identify many new hypotheses								
System allows investigation								
System didn't help look at data in new way								
System gave no new insight		x						
System provided fresh perspective								
System didn't help ask novel questions								
System didn't help think of new data sources				x				
System helped think of question in new way								
System helped find new data		x	x	x				
System didn't help think of something new								
Couldn't explore new questions								
System gave new insight		x						
System didn't help investigate new ideas or paths								

Item Short Form	M	U	R	F	EE/D	M-2	E	RAI
System helped make new connections								
Could explore new questions								
System didn't help make new connections								
System didn't help formulate question				x				
System helped formulate question				x				
System didn't make good use of new data sources				x				
System helped identify interesting hypothesis								
System made good use of new data sources				x				
System didn't help pick relevant information			x	x				
System helped investigate new ideas or paths								
System showed novel relationships in data								
System showed no novel relationships in data								
System does not allow investigation								
System helped hone question				x				
System helped pick relevant information			x	x				

Item Short Form	M	U	R	F	EE/D	M-2	E	RAI
System didn't help identify interesting hypothesis								
System didn't help hone question				x				
System helped remove extraneous information			x	x				
System showed something new, novel								
Visualizations didn't help understand task				x				
System helped identify relevant prior data and trends			x	x				
Could look for what user wanted in the data		x	x	x				
System didn't help remove extraneous information			x	x				
Wasn't able to investigate model performance with prior data	x							
Unable to transform data		x	x					
System provided information needed		x	x	x				
System didn't help identify relevant prior data and trends			x	x				
System helped focus on what mattered				x				
System didn't indicate important features	x		x	x				
System didn't help focus on what mattered				x				

Item Short Form	M	U	R	F	EE/D	M-2	E	RAI
Didn't use model output in decision	x	x						
Wasn't able to predict future model performance using prior data	x							
Able to transform data		x	x					
Was able to predict future model performance using prior data	x							
System indicated important features	x		x	x				
Used model output in decision	x	x						
Model produced user-important features			x					
Output not aligned with intuition		x						
Couldn't look for what user wanted in the data		x	x	x				
Got less information than needed	x	x		x				
Understood individual predictions	x							
Output aligned with intuition		x						
Visualizations helped understand task				x				
System told user what they already knew		x						
Confident in results		x						
Can't explain results' meaning	x							
Can make sense of results	x							

Item Short Form	M	U	R	F	EE/D	M-2	E	RAI
Can explain results' meaning	x							
Results don't make sense	x							
Can't make sense of results	x							
Results make sense	x							
Confident in decision		x						
Not confident in decision		x						
Not confident in results		x						
System didn't provide information needed		x	x	x				
Given right amount of information	x			x				
Can't explain reasoning behind predictions	x			x				
Can explain reasoning behind predictions	x			x				
Results not trustworthy		x		x				
Got exactly the information needed	x	x		x				
Model didn't produce user-important features			x					
Didn't understand individual predictions	x							
Was able to investigate model performance with prior data	x							
System didn't show how it made predictions	x							

Item Short Form	M	U	R	F	EE/D	M-2	E	RAI
System showed how it made predictions	x							
System didn't explain particular prediction	x							
System explained particular prediction	x							
Overwhelmed by information	x			x				
Given too much information	x			x				
Not overwhelmed by information	x			x				
Got more information than needed								
Indicate user's task								
Features system expressed as important, according to user								
Necessary factors for task								
System-generated features match user-generated								
Meaning user derives from system output								
System's output in user's words								
User's decision per output								
Uses formal method to carry out work								
Uses ML, AI techniques to carry out work								
System provided useful model								

www.parenthetic.io

Item Short Form	M	U	R	F	EE/D	M-2	E	RAI
Thought of task that could not specify for system								
System failed at some point								
System flagged problem in data								
System lacked feature that would have been helpful								
System helped discover new, interesting problems								

Full Protocol

Item Short Form	Code	Item Full Text	Mean (SD)
System helped find new relevant data	+	The system made it easy to find new sources of relevant data	0.219 (1.521)
System didn't help find new relevant data	-	The system did NOT make it easy to find new sources of relevant data	0.260 (1.616)
System didn't show something new, novel	-	The system did NOT show me something new	0.562 (1.518)
System helped look at data in new way	+	The system helped me to look at my data in a new way	0.603 (1.431)
System helped ask novel questions	+	The system helped me to ask new and interesting questions	0.411 (1.526)
System didn't provide fresh perspective	-	The system did NOT help provide a fresh perspective	0.479 (1.634)
System didn't help identify many new hypotheses	-	The system did NOT help me to identify very many new, interesting hypotheses	0.274 (1.635)
System helped think of new data sources	+	The system helped me think of new data sources to use in my work	0.466 (1.501)
System didn't help think of question in new way	-	The system did NOT help me to think of the question(s) I work on in a new way	0.411 (1.517)
System didn't help find new data	-	The system did NOT help me to find new sources of data easily	0.233 (1.55)
System helped think of something new	+	The system helped me to think of something new	0.507 (1.573)

www.parenthetic.io

Item Short Form	Code	Item Full Text	Mean (SD)
System helped identify many new hypotheses	+	The system helped me to identify many new, interesting hypotheses	0.247 (1.498)
System allows investigation	+	The system allowed me to investigate the data in new ways	0.753 (1.535)
System didn't help look at data in new way	-	The system did NOT help me to look at my data in a new way	0.562 (1.518)
System gave no new insight	-	The system did NOT provide me with new insight	0.425 (1.572)
System provided fresh perspective	+	The system helped provide a fresh perspective	0.521 (1.642)
System didn't help ask novel questions	-	The system did NOT lead me to ask new and interesting questions	0.534 (1.582)
System didn't help think of new data sources	-	The system did NOT help me think of new data sources to use in my work	0.247 (1.656)
System helped think of question in new way	+	The system helped me to think of the question(s) I work on in a new way	0.548 (1.564)
System helped find new data	+	The system helped me to find new sources of data easily	0.356 (1.522)
System didn't help think of something new	-	The system did NOT help me to think of something new	0.548 (1.59)
Couldn't explore new questions	-	The system did NOT allow me to explore new questions as I discovered them	0.699 (1.459)
System gave new insight	+	The system provided me with new insight	0.384 (1.672)

Item Short Form	Code	Item Full Text	Mean (SD)
System didn't help investigate new ideas or paths	-	The system did NOT help me to investigate new ideas or paths	0.671 (1.463)
System helped make new connections	+	The system helped me make connections I had not previously thought of	0.466 (1.537)
Could explore new questions	+	The system allowed me to explore new questions as I discovered them	0.534 (1.573)
System didn't help make new connections	-	The system did NOT help me make connections I had not previously thought of	0.479 (1.625)
System didn't help formulate question	-	The system did NOT help me to better formulate my question(s)	0.192 (1.639)
System helped formulate question	+	The system helped me to better formulate my question(s)	0.493 (1.6)
System didn't make good use of new data sources	-	The system did NOT make effective use of new sources of data	0.479 (1.386)
System helped identify interesting hypothesis	+	The system helped me to identify at least one interesting hypothesis to test	0.822 (1.494)
System made good use of new data sources	+	The system made effective use of new sources of data	0.288 (1.514)
System didn't help pick relevant information	-	The system did NOT help me to decide what information was relevant	0.288 (1.791)
System helped investigate new ideas or paths	+	The system helped me to investigate ideas or paths that I wouldn't usually investigate	0.356 (1.576)

Item Short Form	Code	Item Full Text	Mean (SD)
System showed novel relationships in data	+	The system showed me new and interesting relationships in the data	0.548 (1.537)
System showed no novel relationships in data	-	The system did NOT show me new and interesting relationships in the data	0.507 (1.556)
System does not allow investigation	-	The system did NOT allow me to investigate the data in new ways	0.493 (1.547)
System helped hone question	+	The system helped me to hone the question(s) that I was asking	0.438 (1.598)
System helped pick relevant information	+	The system helped me to decide what information was relevant	0.589 (1.632)
System didn't help identify interesting hypothesis	-	The system did NOT help me to identify at least one interesting hypothesis to test	0.781 (1.575)
System didn't help hone question	-	The system did NOT help me to hone the question(s) that I was asking	0.384 (1.551)
System helped remove extraneous information	+	The system helped me to remove extraneous information	0.000 (1.616)
System showed something new, novel	+	The system showed me something new that I wouldn't have thought to look for	0.438 (1.554)
Visualizations didn't help understand task	-	The system's visualizations did NOT help me to understand the task at hand	0.507 (1.78)

Item Short Form	Code	Item Full Text	Mean (SD)
System helped identify relevant prior data and trends	+	The system helped me understand which prior data and trends were relevant	0.425 (1.649)
Could look for what user wanted in the data	+	The system allowed me to look for what I wanted in the data	0.493 (1.765)
System didn't help remove extraneous information	-	The system did NOT help me remove extraneous information	0.137 (1.627)
Wasn't able to investigate model performance with prior data	-	The system did NOT allow me to investigate how well the model performed in finding relationships in prior data	0.178 (1.644)
Unable to transform data	-	The system did NOT allow me to transform the data to my liking	-0.055 (1.649)
System provided information needed	+	The system was able to provide me with information that I wanted	0.438 (1.748)
System didn't help identify relevant prior data and trends	-	The system did NOT help me understand which prior data and trends were relevant	0.342 (1.53)
System helped focus on what mattered	+	The system helped me focus on what mattered to perform the task	0.671 (1.463)
System didn't indicate important features	-	The system did NOT indicate which features were important	0.192 (1.846)
System didn't help focus on what mattered	-	The system did NOT help me focus on what mattered to perform the task	0.534 (1.573)
Didn't use model output in decision	-	I did NOT use the model to help make my decision	0.247 (1.77)

www.parenthetic.io

Item Short Form	Code	Item Full Text	Mean (SD)
Wasn't able to predict future model performance using prior data	-	The system did NOT help me to use information about relationships based on prior data to make predictions about future model performance	0.521 (1.492)
Able to transform data	+	The system allowed me to transform the data to my liking	0.178 (1.61)
Was able to predict future model performance using prior data	+	The system helped me to use information about relationships based on prior data to make predictions about future model performance	0.274 (1.652)
System indicated important features	+	The system indicated which features were important	0.151 (1.8)
Used model output in decision	+	I used the model output to help make my decision	0.260 (1.788)
Model produced user-important features	+	The model produced features which I deemed important	0.616 (1.729)
Output not aligned with intuition	-	The system's output did NOT align with my intuition	0.740 (1.375)
Couldn't look for what user wanted in the data	-	The system did NOT allow me to look for what I wanted in the data	0.219 (1.627)
Got less information than needed	-	The system gave me less information than I needed to complete the task	0.521 (1.684)
Understood individual predictions	-	The model allowed me to understand individual predictions based on these data	0.219 (1.702)
Output aligned with intuition	+	The system's output aligned with my intuition	0.603 (1.431)

Item Short Form	Code	Item Full Text	Mean (SD)
Visualizations helped understand task	+	The system's visualizations helped me to understand the task at hand	1.014 (1.495)
System told user what they already knew	+	The system told me what I already knew	0.411 (1.278)
Confident in results	+	I have confidence in the system's results	0.849 (1.578)
Can't explain results' meaning	-	I CANNOT explain what these results mean	0.466 (1.915)
Can make sense of results	+	I can make sense of what the system's results are saying	0.630 (1.728)
Can explain results' meaning	+	I can explain what the system's results mean	0.438 (1.856)
Results don't make sense	-	The system's results do NOT make sense to me	0.301 (1.831)
Can't make sense of results	-	I can NOT make sense of what the system's results are saying	0.329 (1.93)
Results make sense	+	The system's results make sense to me	0.493 (1.908)
Confident in decision	+	I am confident in my decision	0.425 (1.545)
Not confident in decision	-	I am NOT confident in my decision	0.260 (1.716)
Not confident in results	-	I do NOT have confidence in the system's results	0.671 (1.633)
System didn't provide information needed	-	The system was NOT able to provide me with information that I wanted	0.192 (1.823)

Item Short Form	Code	Item Full Text	Mean (SD)
Given right amount of information	+	The system gave me the right amount of information needed to make sense of it	0.384 (1.697)
Can't explain reasoning behind predictions	-	I CANNOT explain the reasoning behind the predictions the model made	0.055 (1.914)
Can explain reasoning behind predictions	+	I can explain the reasoning behind the predictions the model made	-0.014 (1.752)
Results not trustworthy	-	The system did NOT give me trustworthy results	0.712 (1.585)
Got exactly the information needed	+	The system gave me exactly the information that I needed to complete the task	-0.096 (1.717)
Model didn't produce user-important features	-	The model did NOT produce features which I deem important	0.534 (1.659)
Didn't understand individual predictions	-	The model did NOT allow me to understand individual predictions based on these data	0.178 (1.67)
Was able to investigate model performance with prior data	+	The system allowed me to investigate how well the model performed in finding relationships in prior data	0.452 (1.599)
System didn't show how it made predictions	-	The system did NOT allow me to see how it made predictions	0.247 (1.64)
System showed how it made predictions	+	The system allowed me to see how it made predictions	0.137 (1.694)
System didn't explain particular prediction	-	The system did NOT explain to me how I got a particular prediction	-0.123 (1.675)

www.parenthetic.io

Item Short Form	Code	Item Full Text	Mean (SD)
System explained particular prediction	+	The system explained to me how I got a particular prediction	-0.151 (1.713)
Overwhelmed by information	-	The system gave me an overwhelming amount of information	0.589 (1.674)
Given too much information	-	The system gave me too much information to make sense of it	0.877 (1.624)
Not overwhelmed by information	+	The system did NOT give me an overwhelming amount of information	0.438 (1.779)
Got more information than needed	-	The system gave me more information than I needed to complete the task	-0.356 (1.628)
Indicate user's task	n/a	Please indicate the task that you would like the system to help you to perform today: Here, we use the term "task" to refer to the specific way in which you are expecting the system to help you to perform your work. For example, this may be something like "help me answer a question about X" or "validate my data."	—
Features system expressed as important, according to user	n/a	Which features did the system express as most important?	—
Necessary factors for task	n/a	Please list any factors or items that you need to know in order to perform this task. Only list as many as you feel you need in order to perform this task. - Selected Choice	—

Item Short Form	Code	Item Full Text	Mean (SD)
System-generated features match user-generated	+	How well do the items from the system match the items that you generated?	1.411 (0.955)
Meaning user derives from system output	n/a	What do you think this output means?	—
System's output in user's words	n/a	Please describe the system's output in your own words.	—
User's decision per output	n/a	What decision, if any, would you recommend based on this output?	—
Uses formal method to carry out work	+	Do you currently use a formal method to carry out your work? - Selected Choice [No : 0, Yes : 1]	
Uses ML, AI techniques to carry out work	+	Do you currently use machine learning or other AI techniques to carry out your work? - Selected Choice [No: 0, Yes : 1]	
System provided useful model	+	Do you feel that the system provided a useful model? - Selected Choice [No: 0, Yes : 1]	0.630 (0.486)
Thought of task that could not specify for system	+	Were there tasks that you thought of, and wanted to specify, but could not? - Selected Choice [No : -1, Yes : 0]	-0.247 (0.434)
System failed at some point	+	Did the system fail at any point during your test? - Selected Choice [No : -1, Yes : 0]	-0.479 (0.503)
System flagged problem in data	+	Did the system flag any problems in the data (e.g., rows with missing values, extreme outliers, logically impossible values)? - Selected Choice [No : 0, Yes : 1]	0.096 (0.296)

Item Short Form	Code	Item Full Text	Mean (SD)
System lacked feature that would have been helpful	+	Are there any system features that you wish you had, but did not have, that might have been helpful in identifying new tasks? - Selected Choice [No : -1, Yes : 0]	-0.342 (0.478)
System helped discover new, interesting problems	+	Were any features of the system helpful in discovering new and interesting problems? - Selected Choice [No : 0, Yes : 1]	0.534 (0.502)

PCA Factor Loadings

	Enabled Exploration/Discovery	Meaningfulness	Explainability	Right Amount of Information
% Variance Explained	33.2%	22.7%	10.3%	4.1%
Item Short Form				
System helped find new relevant data	0.82			
System didn't help find new relevant data	0.80			
System didn't show something new, novel	0.79			
System helped look at data in new way	0.79			
System helped ask novel questions	0.79			
System didn't provide fresh perspective	0.79			
System didn't help identify many new hypotheses	0.79			
System helped think of new data sources	0.79			
System didn't help think of question in new way	0.78	0.42		
System didn't help find new data	0.78			
System helped think of something new	0.78			
System helped identify many new hypotheses	0.78			
System allows investigation	0.78			
System didn't help look at data in new way	0.77			

	Enabled Exploration/Discovery	Meaningfulness	Explainability	Right Amount of Information
System gave no new insight	0.76	0.47		
System provided fresh perspective	0.76			
System didn't help ask novel questions	0.75			
System didn't help think of new data sources	0.75			
System helped think of question in new way	0.75			
System helped find new data	0.75			
System didn't help think of something new	0.75			
Couldn't explore new questions	0.75			
System gave new insight	0.75	0.50		
System didn't help investigate new ideas or paths	0.74			
System helped make new connections	0.74	0.41		
Could explore new questions	0.74			
System didn't help make new connections	0.74			
System didn't help formulate question	0.73	0.41		
System helped formulate question	0.73			
System didn't make good use of new data sources	0.71			
System helped identify interesting hypothesis	0.71			

	Enabled Exploration/Discovery	Meaningfulness	Explainability	Right Amount of Information
System made good use of new data sources	0.69			
System didn't help pick relevant information	0.69		0.41	
System helped investigate new ideas or paths	0.68			
System showed novel relationships in data	0.68	0.41		
System showed no novel relationships in data	0.67	0.53		
System does not allow investigation	0.67			
System helped hone question	0.67	0.53		
System helped pick relevant information	0.64			
System didn't help identify interesting hypothesis	0.64			
System didn't help hone question	0.63	0.50		
System helped remove extraneous information	0.61			
System showed something new, novel	0.59			
Visualizations didn't help understand task	0.59	0.48		
System helped identify relevant prior data and trends	0.59		0.47	
Could look for what user wanted in the data	0.58	0.66		

	Enabled Exploration/Discovery	Meaningfulness	Explainability	Right Amount of Information
System didn't help remove extraneous information	0.58			
Wasn't able to investigate model performance with prior data	0.55	0.48		
Unable to transform data	0.54			
System provided information needed	0.54	0.67		
System didn't help identify relevant prior data and trends	0.54		0.44	
System helped focus on what mattered	0.53	0.53		
System didn't indicate important features	0.53		0.60	
System didn't help focus on what mattered	0.52	0.51		
Didn't use model output in decision	0.51	0.64		
Wasn't able to predict future model performance using prior data	0.49	0.50	0.44	
Able to transform data	0.49	0.52		
Was able to predict future model performance using prior data	0.49	0.48		
System indicated important features	0.47		0.64	
Used model output in decision	0.47	0.69		
Model produced user-important features	0.46	0.57		
Output not aligned with intuition	0.46	0.49		
Couldn't look for what user wanted in the data	0.45	0.56		

	Enabled Exploration/ Discovery	Meaningful- ness	Explainability	Right Amount of Information
Got less information than needed	0.44	0.58		
Understood individual predictions	0.43	0.53	0.46	
Output aligned with intuition	0.43	0.49	0.43	
Visualizations helped understand task	0.42			
System told user what they already knew	0.42			
Confident in results	0.42	0.66		
Can't explain results' meaning		0.85		
Can make sense of results		0.83		
Can explain results' meaning		0.83		
Results don't make sense		0.81		
Can't make sense of results		0.81		
Results make sense		0.81		
Confident in decision		0.77		
Not confident in decision		0.76		
Not confident in results		0.76		
System didn't provide information needed		0.69		
Given right amount of information		0.68		
Can't explain reasoning behind predictions		0.67	0.48	
Can explain reasoning behind predictions		0.64	0.49	

www.parenthetic.io

	Enabled Exploration/Discovery	Meaningfulness	Explainability	Right Amount of Information
Results not trustworthy		0.63		
Got exactly the information needed		0.61		
Model didn't produce user-important features		0.57	0.43	
Didn't understand individual predictions		0.53	0.51	
Was able to investigate model performance with prior data		0.50		
System didn't show how it made predictions			0.75	
System showed how it made predictions			0.70	
System didn't explain particular prediction			0.70	
System explained particular prediction			0.68	
Overwhelmed by information				0.86
Given too much information				0.80
Not overwhelmed by information				0.77
Got more information than needed				-0.67
System lacked feature that would have been helpful				0.4229

Protocol Items per Subscale/Component

Enabled Exploration/Discovery	
System helped find new relevant data	The system made it easy to find new sources of relevant data
System didn't help find new relevant data	The system did NOT make it easy to find new sources of relevant data
System didn't show something new, novel	The system did NOT show me something new
System helped look at data in new way	The system helped me to look at my data in a new way
System helped ask novel questions	The system helped me to ask new and interesting questions
System didn't provide fresh perspective	The system did NOT help provide a fresh perspective
System didn't help identify many new hypotheses	The system did NOT help me to identify very many new, interesting hypotheses
System helped think of new data sources	The system helped me think of new data sources to use in my work
System didn't help find new data	The system did NOT help me to find new sources of data easily
System helped think of something new	The system helped me to think of something new
System helped identify many new hypotheses	The system helped me to identify many new, interesting hypotheses
System allows investigation	The system allowed me to investigate the data in new ways
System didn't help look at data in new way	The system did NOT help me to look at my data in a new way

www.parenthetic.io

System provided fresh perspective	The system helped provide a fresh perspective
System didn't help ask novel questions	The system did NOT lead me to ask new and interesting questions
System didn't help think of new data sources	The system did NOT help me think of new data sources to use in my work
System helped think of question in new way	The system helped me to think of the question(s) I work on in a new way
System helped find new data	The system helped me to find new sources of data easily
System didn't help think of something new	The system did NOT help me to think of something new
Couldn't explore new questions	The system did NOT allow me to explore new questions as I discovered them
System didn't help investigate new ideas or paths	The system did NOT help me to investigate new ideas or paths
Could explore new questions	The system allowed me to explore new questions as I discovered them
System didn't help make new connections	The system did NOT help me make connections I had not previously thought of
System helped formulate question	The system helped me to better formulate my question(s)
System didn't make good use of new data sources	The system did NOT make effective use of new sources of data
System helped identify interesting hypothesis	The system helped me to identify at least one interesting hypothesis to test
System made good use of new data sources	The system made effective use of new sources of data

System helped investigate new ideas or paths	The system helped me to investigate ideas or paths that I wouldn't usually investigate
System does not allow investigation	The system did NOT allow me to investigate the data in new ways
System helped pick relevant information	The system helped me to decide what information was relevant
System didn't help identify interesting hypothesis	The system did NOT help me to identify at least one interesting hypothesis to test
System helped remove extraneous information	The system helped me to remove extraneous information
System showed something new, novel	The system showed me something new that I wouldn't have thought to look for
System didn't help remove extraneous information	The system did NOT help me remove extraneous information
Unable to transform data	The system did NOT allow me to transform the data to my liking
Visualizations helped understand task	The system's visualizations helped me to understand the task at hand
System told user what they already knew	The system told me what I already knew
Meaningfulness	
Can't explain results' meaning	I CANNOT explain what these results mean
Can make sense of results	I can make sense of what the system's results are saying
Can explain results' meaning	I can explain what the system's results mean
Results don't make sense	The system's results do NOT make sense to me

Can't make sense of results	I can NOT make sense of what the system's results are saying
Results make sense	The system's results make sense to me
Confident in decision	I am confident in my decision
Not confident in decision	I am NOT confident in my decision
Not confident in results	I do NOT have confidence in the system's results
System didn't provide information needed	The system was NOT able to provide me with information that I wanted
Given right amount of information	The system gave me exactly the information that I needed to complete the task
Results not trustworthy	The system did NOT give me trustworthy results
Got exactly the information needed	The system gave me exactly the information that I needed to complete the task
Was able to investigate model performance with prior data	The system allowed me to investigate how well the model performed in finding relationships in prior data
Explainability	
System didn't show how it made predictions	The system did NOT allow me to see how it made predictions
System showed how it made predictions	The system allowed me to see how it made predictions
System didn't explain particular prediction	The system did NOT explain to me how I got a particular prediction
System explained particular prediction	The system explained to me how I got a particular prediction

Right Amount of Information	
Overwhelmed by information	The system gave me an overwhelming amount of information
Not overwhelmed by information	The system did NOT give me an overwhelming amount of information
Got more information than needed	The system gave me more information than I needed to complete the task

TECHNICAL REPORT

DARPA D3M TA3 User
Bi-Annual Evaluation Brief

Fall 2019

PARENTHETIC
INFLUENCE INTELLIGENTLY

6.2 Summer 2020 Bi-Annual Evaluation Brief

Alexandria, VA
703.406.8730
www.parenthetic.io

Contract No.: FA8750-19-C-1522
Performing Organization: Parenthetic
Deliverable: 0001AD Technical Report – Bi-Annual Evaluation Brief

PARENTHETIC

TECHNICAL REPORT

DARPA D3M TA3 User
Bi-Annual Evaluation Brief

Summer 2020

Prepared for:
DARPA Program Manager

Submission date:
October 16, 2020

Submitted by:
Parenthetic
Catherine Tejeda
Reva Schwartz
Michael Liang

TABLE OF CONTENTS

Executive Summary	3 - 5
Evaluation Outcomes	7 - 12
Recommendations	13 - 17
Lessons Learned	18
Future Considerations	18
Evaluation Elements	19 - 25
Background	26 - 27
Appendices	
Appendix A – Score sheet & Rubric	28
Appendix B – User Resource Guide	29 - 38
Appendix C – System Submission Procedure	39

EXECUTIVE SUMMARY

This report serves to document the DARPA D3M program's Summer 2020 internal evaluation. For the D3M Summer 2020 evaluation, systems were assessed by an internal group from the D3M evaluation team in order to prepare for future transition partner engagements during the final stages of the program. The evaluation was split into two distinct phases:

1. Phase 1 – Transition Partner Use Case Evaluation

Each TA3 performer team demonstrated their system's ability to support the unique use case assigned to them. The unique use cases for each system are listed in Table 1:

Table 1: Unique use cases per TA3 system

TA3 Team System	Use Case Domain
System #1	Active Search Remote Sensing Locust Breeding
System #2	Integrated Food Security Phase Classification (IPC) Analysis Food Insecurity
System #3	Stabilization Conflict Analysis

2. Phase 2 – Common Problem Types Evaluation

A total of 10 users assessed TA3 performer-built systems for factors related to functionality and usability. For this evaluation, we captured user responses via a scoring rubric and scoresheet. We collected the following input from each user: 1) user feedback scores based on various workflow tasks and 2) detailed user notes containing context supporting their scores. A detailed review of results can be found in the body of this report.

Three (3) TA3 systems were evaluated using the following datasets spanning the five (5) problem types that systems are expected to support:

Table 2: Datasets used in evaluation

Dataset	Relevant Transition Partner	Binary Classification	Multiclass Classification	Regression	Time Series Forecasting
Dataset #1	Transition Partner #2	X			
Synthetic Dataset	N/A	X			
Dataset #2	Transition Partner #1		X		
Cancer Mortality Rate	N/A			X	
Sorghum Plant Growth	TERRA-REF				X

EXECUTIVE SUMMARY

Users for this phase of the evaluation were sourced from members of the D3M government team which includes personnel from DARPA, NASA Jet Propulsion Laboratory (JPL), Data Machines Corp., and Parenthetic. This set of users have a range of experience with data science, from novice to advanced, and cover the full spectrum of familiarity with the evaluated TA3 systems, with some having interacted with the tools repeatedly over the life of the D3M program and others seeing the systems for the first time.

Key Findings and Recommendations

The evaluation's key findings vary by system, and can be found in more detail in the section titled Evaluation Outcomes. Overall, the TA3 systems do not fully satisfy the needs of desired users, particularly more prolific data science users. Table 3 shows a high-level look at how the TA3 systems performed across the various workflow stages.

Table 3: TA3 system performance across workflow stages (overall average scores*)

Stage	System #1	System #2	System #3
Load	2.3	3.8	2.3
Explore	2.0	2.9	2.6
Prepare	2.1	2.5	2.0
Model	1.9	2.8	1.8
Explain	1.3	2.2	1.9
Deploy	1.9	2.3	1.1
Overall	1.9	2.7	1.9

*Scoring system: 0=Not Supported, 1=Insufficiently Supported, 2=Minimally Supported, 3=Adequately Supported, 4=Exceptionally Supported. Color coding is to aid score differentiation where green represents higher scores, red represents lower scores, and orange represents mid-range scores.

An overview of system findings includes:

- **System Messaging:** TA3 systems must improve how they communicate with users. This can be done with displaying accurate progress bars, run-time timers and other system prompts. Without this information, the systems place users in ambiguous situations where they are unsure if a system is processing a request or experiencing a freeze or bug. This has the effect of causing users to be unsure of what to do – for example, some users restarted their browsers thinking the systems had stalled.
- **User Guidance:** Users appreciate the level of freedom that the systems allowed them, but at times struggled with how to best approach their analysis given the array of options and functions presented in each system. Systems need to find a balance between user freedom and system guidance, especially when considering the potential diversity of data science expertise among users. This requirement for level of guidance spans the entire user experience spectrum - from helping users best prepare their data for modeling to understanding model outputs.
- **Explainability:** Every system overwhelmed users to some degree with information overload. Often systems placed the burden of interpreting large amounts of information on the user rather than helping users to connect the dots, leading to insights they would not have discovered otherwise. Requiring users to sift through various outputs and results is counterproductive and antithetical to the ultimate role ML systems are supposed to play:

helping those with minimal data science knowledge construct and interpret models without a data scientist by their side.

- **Quality of Life:** Each system could consider a number of quality of life (e.g., aspects of design that ensure users have a smooth experience) adjustments to enhance the overall user experience, such as:
 - Adding tooltips – For example, providing more detailed data over charts or over buttons to more clearly indicate functionality.
 - Reducing redundancy – For example, allowing users to conduct actions in batches rather than requiring repetitious and tedious steps.
 - Automating processes – For example, automatically identifying outliers based on user defined parameters or removing features with no data.

FULL REPORT

EVALUATION OUTCOMES

Summary Protocol Results by Dataset

This first section shows the distribution of user scores across each evaluation dataset per TA3 system. Users also included detailed notes about their experience with each system, providing more context. Outcome tables list the average user scores. The number of testers varies across each system and dataset, mostly due to tester availability. For more details on the evaluation methodology, please see the section on Evaluation Elements beginning on page 20.

System #1

Table 4: System #1 scores* across workflow stages and datasets

Stage/Dataset	Dataset #1	Dataset #2	Cancer	TERRA	Synthetic	Overall
Load	2.3	1.9	2.6	2.7	1.9	2.3
Explore	1.8	1.9	2.6	2.6	1.4	2.0
Prepare	2.0	1.9	2.4	2.4	1.6	2.1
Model	2.2	2.1	2.3	1.3	1.4	1.9
Explain	1.6	1.4	1.9	0.7	1.0	1.3
Deploy	2.3	2.3	2.6	0.9	1.4	1.9
Overall	2.1	1.9	2.4	1.8	1.5	1.9

*Scoring system: 0=Not Supported, 1=Insufficiently Supported, 2=Minimally Supported, 3=Adequately Supported, 4=Exceptionally Supported. Color coding is to aid score differentiation where green represents higher scores, red represents lower scores, and orange represents mid-range scores.

Results in Table 7 indicate the following key points:

- Overall, users rated System #1 the strongest for the first workflow stages (Load through Prepare) and the weakest for Model through Deploy.
- In particular, the system struggled at the Explain stage as users felt they were not properly guided on how to interpret the system-generated results – an area where all the systems fell short.
- System #1 scored highest with the Cancer Mortality Rate dataset (regression).
- Users appeared to struggle the most with the Synthetic (binary classification) dataset, likely attributable to the artificial nature of the data which contains no semantics, meant to mimic a transition partner's data.
- Users appeared to face particularly difficulties with the TERRA (time-series forecasting) dataset. Once they reached the Model stage, users struggled to produce a model – with a number of users reporting that the system ran a long time or in some cases, ran indefinitely.
- Although System #1 generally received higher scores for the Load stage, the lower scores for Dataset #2 and the Synthetic dataset are attributable to the size of these datasets which resulted in the system truncating the data significantly.

EVALUATION OUTCOMES

Below is a detailed summary that incorporates evaluation users' qualitative feedback about System #1, providing more context to the scores above:

- **Load** - Users found this to be among the more intuitive loading experiences since it's front and center upon system startup. The most reported issue was the dataset size restrictions, which limits the number of rows imported (to only 25,000 rows). The limitations are likely to bolster
- system performance, but several users flagged this as a major issue, compounded by the system not clearly communicating that the system truncated the dataset (some users did not even notice). There are also concerns with load times which will need to be addressed, especially once the dataset size restrictions are removed.
- **Explore** - Users appreciated the simple histograms that gave them a clear look at the features, as well as being able to reclassify data types easily. The main issue for this stage revolves around the workflow as some users found it unintuitive that the system forced them to select a target variable before they were able to explore the datasets in more depth. Users typically expect the workflow order to be "Load > Explore > Prepare," but System #1's interface mimics a workflow more analogous to "Load > Prepare > Explore." This resulted in some users having to backtrack and adjust their analyses after discovering more details about the data.
- **Prepare** - Users found this stage simple when it came to adding or removing features for model construction. The system clearly highlighted feature importance. The system's simplicity also means there are a number of limitations as well. For example, users were unsure of how to transform variables and identify outliers. For those that were able to discover how the system handles outlier detection, they reported that removing outliers was often cumbersome and imprecise.
- **Model** - The system provided users with various and easily adjustable parameters; however, the system didn't always adhere to these parameters. For example, this system consistently ran substantially longer than the allotted time threshold. More advanced users requested the ability to control parameters outside of the default setting - such as switching the metric used from the default "mean absolute error."
- **Explain** - The color-coded display highlighting feature importance was well-received by users, but overall, users felt that the system struggled at this stage the most. The system generates large amounts of information and places the burden of interpreting the output entirely on the user - a particularly challenging task for users with more limited data science skills. Additionally, outputs were missing various key pieces of information that more advanced users would have liked to see, such as beta coefficients, significance indicators and confidence intervals. Users were particularly confused on how to interpret results for the TERRA dataset (time-series forecasting).
- **Deploy** - Users found this workflow very intuitive with the system prompting them each step of the way on how to apply their models to new datasets. The main issue here is users did not understand how to save or download their predictions.

www.parenthetic.io

EVALUATION OUTCOMES

System #2

Table 5: System #2 scores across workflow stages and datasets

Stage/Dataset	Dataset #1	Dataset #2	Cancer	TERRA	Synthetic	Overall
Load	3.9	4.0	3.9	3.7	3.8	3.8
Explore	2.4	3.0	3.1	3.0	3.0	2.9
Prepare	2.1	2.9	2.4	2.4	2.6	2.5
Model	2.9	3.0	2.9	2.4	2.8	2.8
Explain	2.0	2.1	2.4	2.1	2.2	2.2
Deploy	2.6	2.6	2.9	0.7	2.6	2.3
Overall	2.6	2.9	2.9	2.4	2.8	2.7

*Scoring system: 0=Not Supported, 1=Insufficiently Supported, 2=Minimally Supported, 3=Adequately Supported, 4=Exceptionally Supported. Color coding is to aid score differentiation where green represents higher scores, red represents lower scores, and orange represents mid-range scores.

Results in Table 8 indicate the following key points:

- System #2 received the highest overall scores of the three systems evaluated.
- System #2 scored extremely well for the Load stage, attributed to its intuitive data import experience.
- Like the other TA3 systems, System #2 users required more guidance than was available for interpreting results in the Explain stage.
- Overall, System #2 received consistent scores across all the datasets except TERRA (time-series forecasting) as users did not fully understand how to deploy a forecasting model to new data.

Below is a detailed summary that incorporates evaluation users' qualitative feedback, providing more context to the scores above:

- **Load** - System #2 excels in loading datasets with most users finding it easy and fast. The ability to drag and drop files directly into the system was particularly intuitive.
- **Explore** - Data exploration was another strong point among users who felt that the system empowered them to freely explore the datasets. However, users struggled more with Dataset #1 dataset due to the text-based nature of the data, thus limiting System #2's data visualization capabilities. There are also opportunities for minor user experience improvements but these are more quality of life enhancements rather than functional hurdles – for example, adding the ability for users to manually adjust the scales of the histograms.
- **Prepare** - This is where we begin to see more of a distinction between the users who had strong data science and coding skills. Advanced users found transformations easy while others struggled, finding the process very confusing.

EVALUATION OUTCOMES

- **Model** - Users were impressed by the system's model construction speed and generally simple way of setting up models, especially with the aid of the wizard assistant, which the novice data science users appreciated. However, users wanted the ability to set more parameters such as processing time, number of models returned, etc. Also, users without data science background who did not notice the wizard were confused about both how to set model parameters, and why they would need that capability in the first place.
- **Explain** - This is the area users struggled the most with System #2 (as well with the other two TA3 systems), particularly among novice users. Whereas novice users required more context and guidance on how to interpret outputs, the advanced users were more interested in adjusting the underlying elements to increase their confidence in the results. The inability to fully understand processes "under the hood" hindered efforts to do this.
- **Deploy** - Users found applying models to new data intuitive, but many were unsure of how to download or export their results. The time-series forecasting dataset caused users to be particularly left in the dark while interacting with this system.

EVALUATION OUTCOMES

System #3

Table 6: System #3 scores across workflow stages and datasets

Stage/Dataset	Dataset #1	Dataset #2	Cancer	TERRA	Synthetic	Overall
Load	2.5	1.9	2.8	2.6	1.7	2.3
Explore	2.5	2.6	2.6	2.5	2.8	2.6
Prepare	2.0	2.4	2.0	1.1	2.6	2.0
Model	2.2	2.1	1.9	1.1	1.9	1.8
Explain	2.5	2.1	1.9	0.9	1.9	1.9
Deploy	1.0	0.9	1.1	0.9	1.6	1.1
Overall	2.1	2.0	2.0	1.5	2.1	1.9

*Scoring system: 0=Not Supported, 1=Insufficiently Supported, 2=Minimally Supported, 3=Adequately Supported, 4=Exceptionally Supported. Color coding is to aid score differentiation where green represents higher scores, red represents lower scores, and orange represents mid-range scores.

Results in Table 8 indicate the following key points:

- System #3 received the highest marks for the Explore stage as it offers an abundance of data visualization options.
- In general, users were able to easily import the datasets, but there were performance issues with Dataset #2 and the Synthetic dataset, likely due to their large size.
- System #3 received particularly low scores for the TERRA (time-series forecasting) dataset. This was due to a system bug that prevented most users from moving forward in the workflow (discussed in more detail in the section below).
- As with other systems, System #3 overwhelmed users with information without properly explaining or guiding them to key insights in the model output.
- Users rated the Model stage low due to the system not providing context for the three solvers available to users to choose from.
- Users rated System #3 the lowest for the Deploy stage across all datasets as the system provided little guidance on how to apply a model to new data.

Below is a detailed summary that incorporates evaluation users' qualitative feedback, providing more context to the scores above:

- **Load** – Several users found the load stage to be confusing in System #3. A major component of this is attributed to the fact that upon system startup, users are brought to the Model mode instead of Dataset mode. Even once they do navigate to the Dataset mode, the process for uploading data is not intuitive. For example, requiring users to enter a dataset name before being able to upload a file is confusing and unlike most typical file import methods. This issue is compounded by the lack of system feedback where clicking the Upload icon results in nothing happening, when most users expect at least a system prompt that alerts them to a missed step.

EVALUATION OUTCOMES

- **Explore** - Data exploration was the strongest area of performance in System #3. Users felt it was easy to explore their data, especially with the various visualization options offered. There are a number of missing functions that users would have liked to see though (or were unable to discover), such as an easier way to detect outliers, the ability to interact with the various data visualizations versus being view-only, and more ways for them to directly interact with their dataset to better understand it.
- **Prepare** - Users found the pebble interface to be a very intuitive feature and an aesthetically pleasing way to visualize models. However, there is room for improvement on the user experience as the system seems designed to assume that users know exactly what they want to do and how to do it. Novice users were confused on how to get started without significant guidance provided by training materials and developer video walkthroughs. There are also several quality of life improvements that can be made. For example, adding features one-by-one to the workspace is tedious, as well as reclassifying them individually. Also, the system does not provide a way to delete single pebbles, instead requiring users to delete everything at once, causing significant frustration among many users. The low system scores for the TERRA dataset are attributed to many users struggling to assign the date feature as "temporal," thus impacting their ability to test the system any further beyond this stage.
- **Model** - Users noted the various options to adjust parameters and the system's fast speed in returning models. However, users expected some basic features which were missing, like the ability to stop/pause the model construction process. The chief concern is the system not providing additional context about the model construction process. Users had no guidance as to why they should choose one "solver" over the others, or how each "solver" approached model construction differently. This lack of information likely had a cascading effect, creating more confusion in subsequent steps as the system provided little transparency as to how it reached its results.
- **Explain** – The system recommended a model with supporting confusion matrices – a feature no other system provided. However, the amount of information presented in the output was overwhelming for many users, particularly more novice users, since the system placed the responsibility of interpretation entirely on them.
- **Deploy** - This was the weakest area across the board for all users. Many were unsure how to apply a model to a new dataset and those who were able to figure it out were confused by the output itself.

RECOMMENDATIONS

The following section lists a number of recommendations for each system based on the evaluation results. Note that these recommendations, along with a full list of bugs and issues have been shared with the TA3 performer teams to expedite their ability to iterate on the systems in preparation for transition partner engagements.

System #1

Table 7: Recommendations for System #1 by workflow stage

Stage	Recommendation
Load	☐ Raise the dataset size limit. ☐ Add a progress bar so users have feedback that the system is processing the dataset import request. This allows users to stay informed that the system acknowledged their request. A “processing timer” would have a similar effect. ☐ Redesign the loaded datasets to a more intuitive menu. The current design requires users to scroll tediously to browse through datasets. For example, consider only displaying a datasets summary information once a user clicks on the dataset name rather than including it by default, taking up valuable screen real estate. ☐ Load times need to be shortened significantly. This is particularly an issue considering only a fraction of the largest datasets were imported due to the size restrictions in place during the evaluation.
Explore	☐ The transition from importing data to immediately selecting a target variable is not intuitive for how many analysts approach solving problems. Users typically prefer the option to explore their data before deciding on their target and predictor variables . ☐ Redesign the variable list on the current interface so users do not have to scroll excessively. The search bar is a useful feature, but only when a user is familiar with a dataset. A sorting feature would be a good step in the right direction (e.g., by name, by data type, etc.).
Prepare	☐ Consider designing a wizard or model preparation assistant to guide more novice users through the model preparation stage. Question prompts should avoid data science jargon – for example: How many possible options are there for the target variable? - o 2 (e.g., Yes or No) - o More than 2 (e.g., Red, White or Blue) The system could then set the model to binary or multiclass classification based on the user's selection.
Model	☐ Add a run-time display to provide further feedback to users that the system is processing and constructing the model. The current progress bar often stops in place for long durations, causing users to suspect that the system has frozen. ☐ Fix the stop button as it currently has no functionality.
Explain	☐ The system needs to do a better job of helping novice users to interpret the output, and rely less on data science jargon. The current system places a heavy burden on users to interpret the model results on their own.
Deploy	☐ Add the capability to save or download prediction results, and provide a way for users to clearly surface this feature.

RECOMMENDATIONS

System #2

Table 8: Recommendations for the System #2 system by workflow stage

Stage	Recommendation
Load	☐ Add a sound effect or more clear animation upon successful data import to more clearly communicate when the data loading has completed. ☐ Add the ability to import multiple datasets at once via drag and drop.
Explore	☐ Improve the usability on the canvas which includes adding: - o A “clear all” option rather than requiring users to delete elements individually. - o A zoom in/out capability to allow for easier navigation, especially as the canvas becomes more crowded. - o An “auto-snap” feature to allow for easier organization of canvas elements. ☐ Prompt users with an alert when they attempt to create an analysis that does not make sense (e.g., including a feature with no data). ☐ Allow the system to automatically identify empty features and prompt the user to remove from the dataset. ☐ Add a functionality that allows users to easily see summary statistics of features. For example, add a “statistics” operator, a hover-over tooltip capability, or a sidebar that slides in/out when a user selects a feature. ☐ Clearly communicate that dataframe colors exist simply to aid differentiation and have no functional use beyond visual distinction. ☐ Consider adding a correlation wizard to help more novice users. For example, the wizard can aid users on whether their analysis is best done using Pearson, Kendall or Spearman correlations – and even inform them about how these options affect their model.
Prepare	☐ Allow users to apply data transformations across dataframes to reduce redundancy. ☐ Create wizards for transformations and the forecasting operator to guide more novice users. ☐ Automatically flag potential data quality issues and surface these to the user. Currently, the system requires users to create individual histograms to check each feature which is tedious and repetitive. A potential workaround could be adding a “create all histograms” function. ☐ Explore a menu-based or point-and-click functionality to filter data. The current process is not intuitive and can make managing a high volume of dataframes cumbersome. This capability exists to an extent in other areas of the system. For example, users can already filter within the correlation operator via a filter tab.

RECOMMENDATIONS

Table 8: Recommendations for the System #2 system by workflow stage (continued)

Stage	Recommendation
Model	□ Surface the model construction wizard more effectively since many users found it very useful, particularly more novice users. □ Add more adjustable model parameters such as processing time and number of models to return.
Explain	□ Add hover tooltips to the Explainer to provide better guidance and more context to users (e.g., confidence intervals and p-values). □ Add hover details to feature importance bar graphs to give more information beyond just the visualization of feature rankings. □ Provide links or hover tips for key data science concepts and terms that direct users to a glossary so they can self-educate. If implemented, this feature should have a toggle switch so advanced users can disable it.
Deploy	□ The “Operator” verbiage in the training materials may confuse users as they may look for the Executor Operator in the main Operator drop-down menu, not knowing that it is a feature embedded within an existing operator. □ Surface the download/export functionality more clearly. Consider a pop-up tooltip after a user generates their first model(s).

RECOMMENDATIONS

System #3

Table 9: Recommendations for System #3 by workflow stage

Stage	Recommendation
Load	☐ Add a progress indicator so that users get feedback that the system is processing their data import request. The lack of a progress bar is exacerbated by the slow loading times which makes it unclear if the system is actively working or not. ☐ Data mode should be the first screen shown, especially for first-time users. Currently, system startup brings users to Model mode which is very confusing to new users. Additionally, the data upload option should be placed in a more intuitive location as it is currently embedded among other buttons in a seemingly random order. ☐ Remove the requirement that forces a user to name a dataset before allowing them to upload a file. The system should assign a default dataset name based on the file name, and allow the user to rename if desired. At the very least, the system should prompt the user to enter a dataset name rather than doing nothing when the user clicks “upload.” ☐ In Model mode, the system should be disabled from pre-loading random pebbles upon successful data import. ☐ Allow users to delete previously loaded datasets.
Explore	☐ Add the ability to easily view high-level dataset details (e.g., number of rows, columns, etc.). ☐ Add context for visualization options, highlighting the advantages of various charts to aid users in choosing the optimal visual. ☐ Recommend optimal charting or other visualizations based on user-selected features (e.g., data type, relationship pattern, etc.). ☐ Communicate the meaning of chart outlines. Currently, the system outlines available charts in green and unavailable charts in red which is ambiguous since the system never teaches the user the meaning of this color code. Alternatively, consider not showing incompatible charts at all to minimize displaying unnecessary information the user. ☐ Add a mouse hover-over ability for charts to provide more details on specific data points. ☐ Redesign the Explore > Variables view. Users do not find it useful to view the list of variables as tile images. The ordering is also unintuitive as it is based on the original file's structure. Add sorting ability to allow users more flexibility in how they want to view the variable list (e.g., alphabetically or by data type). ☐ Enable the system to automatically clean a dataset of empty features to save the user time from exploring relationships that will provide erroneous results. ☐ Allow users to manually change variable names.

RECOMMENDATIONS

Table 9: Recommendations for System #3 by workflow stage (continued)

Stage	Recommendation
Prepare	☐ Update the legend to include all color options as it currently does not include what orange means. ☐ Add a “select all” or “add all” option for adding features to Model mode as users found adding them individually to be tedious. ☐ Add ability to remove individual pebbles instead of requiring users to delete everything. ☐ Due to the number of buttons, add a hover tooltip to provide brief descriptions for every button. ☐ Redesign the method for assigning variable characteristics as it currently requires a lot of repetitive actions. ☐ Consider adding a wizard to help more novice users optimally set up their models. ☐ Better surface that next steps in the workflow are highlighted in green. Most users did not notice this feature.
Model	☐ Add feedback such as a progress bar or run-time information so users know that the system is actively generating models. ☐ Provide descriptions for each solver. Users are currently provided no information as to what makes each solver unique. ☐ Fix the “time bound” parameter as the system often runs beyond the desired time limit.
Explain	☐ Improve guidance and presentation of information so that users are provided with meaningful outputs and context to help them decipher the results. ☐ Limit the amount of information presented to users as users were often overwhelmed.
Deploy	☐ Add progress indicators so users receive feedback that the system is processing their requests ☐ Streamline the deploy workflow process.

LESSONS LEARNED

The main purpose of the Summer 2020 evaluation was to provide an internal “gut check” of the state of each TA3 system as the program prepares to engage with various transition partners. Compared to the previous larger-scale user evaluation conducted in Fall 2019, the systems have made progress, but it is clear that there is still significant room for improvement. Of particular concern are issues previously identified that still persist such as the lack of explainability regarding system processes and results – many users continue to struggle with how the systems reach results and how to interpret those results as well. Until this is resolved, this gap in system capabilities and user needs will likely prove to be a substantial hurdle to transition efforts. A key objective of transition partner engagements will be to solicit feedback on how to improve the systems’ ability to explain their processes.

FUTURE CONSIDERATIONS

Originally, this phase of the D3M program would see transition partner users immersed with the TA3 systems via in-person training sessions and workshops, allowing for face-to-face interactions so performer teams can collect feedback in the most direct manner possible. However, due to Covid-19 restrictions on travel and in-person meetings, these engagements have shifted towards virtual demonstrations which inherently limits the type of interaction originally envisioned.

Taking the above into consideration, Parenthetic will continue to further program goals by:

1. Coordinating virtual engagements with transition partners, providing remote support and soliciting feedback to ensure performer teams have the insights necessary to further develop their systems with the end user in mind.
2. Continuing to test the systems directly and provide performer teams with quick turnaround feedback. This includes helping performer teams refine their training materials and other educational resources.
3. Organizing and facilitating additional internal evaluations as needed.
4. Identifying relevant datasets and use cases that may appeal to potential transition partners.

EVALUATION ELEMENTS

Evaluation Protocols

Phase 1 - Transition Partner Use Case Evaluation

During the D3M Summer 2020 Evaluation Workshop (July 6 – July 24, 2020), all TA3 performer teams were tasked to prepare their systems for a one-hour demonstration, showcasing their ability to support specific use case challenge problems relevant to transition partner domains. The demonstrations took place on July 22-23, 2020 through a virtual conference call. Prior to the scheduled demonstration, TA3 teams provided a brief to the D3M Program Manager and SETA outlining the various system capabilities to be showcased which included:

- ☐ System strengths and weaknesses
- ☐ Supported problem types
- ☐ Examples of real-world users currently using the system
- ☐ List of technical accomplishments
- ☐ Plans for final year of the program

Each TA3 system was assigned specific use cases pertinent to various transition partners as follows:

Table 10: Unique use cases per TA3 system

TA3 Team System	Use Case Domain
System #1	Active Search Remote Sensing Locust Breeding
System #2	Integrated Food Security Phase Classification (IPC) Analysis Food Insecurity
System #3	Stabilization Conflict Analysis

The audience for the TA3 transition partner use case demonstrations included the following:

- ☐ DARPA Program Manager
- ☐ DARPA SETA
- ☐ Various evaluation team stakeholders:
 - o MIT Lincoln Labs
 - o Data Machines Corp
 - o JPL
 - o Parenthetic

EVALUATION ELEMENTS

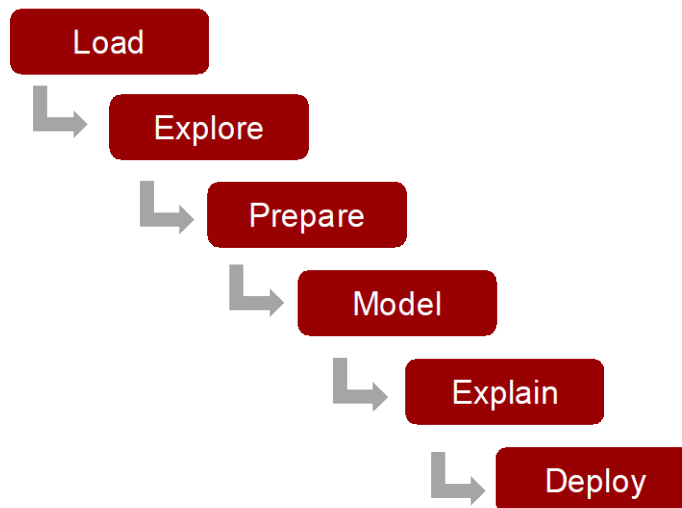
Phase 2 – Common Problem Evaluation

Immediately following the Summer Evaluation Workshop, all TA3 systems were evaluated for their ability to support the standard set of common problem types (regression, binary classification, multiclass classification and time series forecasting) across a variety of datasets. These datasets are relevant to transition partner domains. These datasets are described in more detail in Table 6 of this report (page 14).

User Workflow

Users were asked to assess each of the three TA3 systems across five (5) datasets following the general workflow below that focuses on a number of core capabilities the systems are expected to support:

Figure 1: General workflow of core capabilities each TA3 system is expected to support



1. **Load** – User can flexibly upload datasets from local computer, web, databases, etc. Has features to easily use D3M, OpenML, Kaggle, or other data repositories.
2. **Explore** – System allows user to explore data, statistics, and view single and multiple variable relationships. User can easily see important characteristics in data such as correlations and outliers.
3. **Prepare** – User can easily add or delete features and support feature engineering using Datamarts or other related services. User can address issues with data such as outliers and resolve data issues prior to constructing models.

EVALUATION ELEMENTS

4. **Model** – User can define budget of AutoML system such as time to search, cost constraints, number of models, etc. User can reliably stop, restart, or start search over when constructing models.
5. **Explain** - System automatically explains performance of models in clear terms and visuals. Feature importance and relationships between features and target variable are clearly explained. User can compare models, and analyze the actual pipelines produced prior to model selection. System recommends best model to deploy based on performance, complexity, and other factors relevant to recommending best model.
6. **Deploy** – Selected models can be easily deployed so that new data can be applied to the model, with explanation of results.

EVALUATION ELEMENTS

Evaluation Rubric

The general workflow introduced in the previous section was broken down into more discrete components to aid users in evaluating each system's core capabilities. Users were tasked to score each of these components on a five-point scale which were then aggregated to create a scoresheet for each system per problem type.

The general workflow was broken down into more distinct stages as shown in Table 4. Note that this list was not meant to be comprehensive as users were given freedom to explore the systems thoroughly and provide notes to describe other factors they considered when scoring the systems.

Table 11: Breakdown of key workflow stages to aid users

Workflow Stage	Key Component
Load	Enables user to easily load new datasets
	Uploads datasets from local computer
	Provides additional options to upload datasets (e.g., Internet, Kaggle, etc.)
Explore	Enables user to easily explore the data
	Views single variable relationships
	Views multivariable relationships
	Explores correlations between features
	Identifies outliers in the data
Prepare	Displays an appropriate amount of information
	Enables user to easily prepare data for model construction
	Allows for addition/deletion of features
	Supports feature engineering using Datamart or other related services
Model	Allows user to address outliers or other data issues prior to constructing models
	Enables user to easily construct models
	Allows user to define parameters for model construction (e.g., processing time, number of models, etc.)
Explain	Allows user to stop, restart or reset model construction
	Explains model feature importance
	Explains the relationships between features and target variables
	Allows user to compare models prior to model selection
Deploy	Recommends the best model based on user preferences
	Allows models to be easily applied to new data
	Provides user with clear results in deployed settings

EVALUATION ELEMENTS

The scale used for scoring the systems is shown in Table 5. Users scored each system on each stage with this rubric.

Table 12: Rubric for evaluating the TA3 systems

Score	Threshold
0	Not Supported (i.e., capability does not exist or user is unable to find/use the functionality)
1	Insufficiently Supported (i.e., capability exists but it not useful or effective)
2	Minimally Supported (i.e., capability is supported but needs improvement)
3	Adequately Supported (i.e., capability is useful and effective)
4	Exceptionally Supported (i.e., capability is unique, highly useful and effective)

EVALUATION ELEMENTS

Datasets

A variety of datasets were used across several domains for this evaluation. These datasets are described in Table 6. Three of these domains and associated datasets are linked to D3M transition partners and are labeled appropriately.

Table 13: Descriptions of datasets used in the TA3 evaluation

Dataset Name	Transition Partner	Description
Dataset #1	Transition Partner #2	This dataset's details have been withheld from this report due to transition partner MOU.
Dataset #2	Transition Partner #1	This dataset's details have been withheld from this report due to transition partner MOU.
Cancer Mortality Rate	N/A	This dataset contains cancer incidence and mortality rate data per county sourced from the National Cancer Institute. It includes additional socio-economic data per county from the U.S. Census Bureau.
Sorghum Plant Growth	TERRA-REF	This dataset contains phenotypic features (i.e. canopy height measurements) for several genetic variants (cultivars) of the sorghum plant. The plant heights were measured by laser scanners over the course of 133 days. Each plot (site) contains a timeseries of height measurements (in cm) taken over the 133-day period.
Synthetic Dataset	N/A	This dataset was synthetically generated by MIT Lincoln Labs to have certain characteristics conducive for AutoML processes and techniques. It includes 44 predictor variables and a target variable designed to mimic a binary classification problem.

EVALUATION ELEMENTS

System Descriptions

Three (3) performer teams submitted systems that passed the readiness stage per the submission checklist designed by Data Machines Corp (see Appendix C). Descriptions of each of these systems (provided by each TA3 performer) are below:

1. **System #1:** System #1 will enable SMEs with no data science training to dynamically derive, understand and apply data-driven models in answering mission critical questions. This approach augments human intellect with data discovery, enrichment, analytic model recommendation, and automated visualization to discover observable relationships in data, empowering intelligence action.
2. **System #2:** System #2 provides a fundamentally re-thought data analytics stack: domain experts without programming experience or background in statistics can now explore and mine data through an intuitive user interface, and effortlessly build, analyze, and evaluate machine learning pipelines.
3. **System #3:** System #3 is a data-driven web application designed to bring researchers to insights fast. It facilitates intuitive machine learning, model discovery, and data exploration for researchers. As System 3's intelligent back-end automatically seeks interesting relationships in the data and builds models to predict outcomes, researchers impart substantive knowledge about their data and own research questions to guide the automated generation of AI assistance for data analysis in an interactive paradigm we call *human-guided machine learning*.

User Participation

Users for Phase 2 of the Summer 2020 evaluation were sourced from the program's evaluation team. Users spanned the spectrum in terms of data science experience and familiarity with the systems being evaluated. Each TA3 performer team provided training materials for users to self-train ahead of system testing. These materials focused on the general analytical workflow highlighted previously in this section and included written instructions as well as pre-recorded video tutorials. Users were also given the option to spectate the TA3 unique use case demonstrations, which occurred during the D3M Summer Workshop, to familiarize themselves with the systems.

BACKGROUND

About Us

Parenthetic works on research and development projects in both the public and private sectors focused on new methods or technology for information and communication. We translate end user requirements to technical teams and similarly translate technology capabilities and features to targeted end users. We maintain—and continue to grow—a database of “users” from across public, private, and academic institutions that are available for software testing and evaluation. This community spans many domains and levels of expertise, but users have one commonality - interest in new technology that optimizes human-machine interaction in their domain. We develop operationally-relevant challenge problem sets and evaluation metrics that are rooted in real-world data and circumstances. Our goal is to inform and push the development of capabilities that will bring high fidelity to addressing national security interests.

On the D3M evaluation team, Parenthetic focuses exclusively on assessing TA3 system performance in supporting SMEs’ creation of data-driven models and understanding model output. We want to do more for the end user than point them at an overabundance of output or alert them to what is in that pile. For machine learning tools to assist the user and provide them with explainable and meaningful results, the tools have to enable them to connect the dots, and allow them to answer their questions using their contextual knowledge and expertise. We are supported in this effort by Professor David Broniatowski, Director of the Decision Making and Systems Architecture Laboratory at the George Washington University (GWU) School of Engineering and Applied Science, and GWU doctoral candidate Lydia Gleaves.

For D3M, Parenthetic:

- Provides guidance to the evaluation team on the role of subject matter expertise in TA3 testing and related metrics, including:
 - Human-in-the-loop (HITL) practices
 - How domain experts understand model output; and how model output can be meaningful to experts
 - AI explainability for non-computer scientist communities
- Meets regularly with transition partners to develop TA3 evaluation challenge problems, identify relevant datasets and users, and work towards adoption of developed technology
- Recruits SMEs and data science users for TA3 system testing and evaluation timelines
- Develops targeted user training
- Creates and updates evaluation protocols
- Facilitates information exchange between users and technology developers
- Ensures TA3 testing compliance with IRB requirements
- Analyzes TA3 testing results and provides user-focused conclusions and recommendations for evaluation team

BACKGROUND

Evaluation Team

For this evaluation, Parenthetic worked in conjunction with the following members of the evaluation team:



Data Machines Corp **Computing and Infrastructure**

Data Machines Corporation was responsible for system staging; hosting all user interactions and system deployments in a cloud computing environment; supporting user troubleshooting, bug reporting and resolution; and collecting user model output.



MIT Lincoln Laboratory **Dataset Management**

MIT Lincoln Laboratory was responsible for managing the selection and preparation of relevant datasets for TA3 testing.



Program SETA **Lead for Evaluation Team and Dry Run Planning**

The Program SETA developed and implemented a dry-run plan for evaluating technical system readiness for participation in the evaluation.

www.parenthetic.io

APPENDICES

Appendix A – Scoresheet & Rubric

To streamline and standardize the TA3 evaluation process, the following scoresheet and rubric was distributed to all users. This ensured that all users were assessing the systems using identical baseline criteria and scoring system.

Sample Scoresheet		
	Score (0-4)	Detailed Notes
System		
Problem (Dataset)		
Load		
Explore		
Prepare		
Model		
Explain		
Deploy		
Overall		

Scoring Rubric	
Score	Threshold
0	Not Supported (i.e., capability does not exist or user is unable to find/use the functionality)
1	Insufficiently Supported (i.e., capability exists but is not useful or effective)
2	Minimally Supported (i.e., capability is supported but needs improvement)
3	Adequately Supported (i.e., capability is useful and effective)
4	Exceptionally Supported (i.e., capability is unique, highly useful and effective)

Appendix B - User Resource Guide

To prepare users for the evaluation effort, the following User Resource Guide was provided to each user as an aid with important information including evaluation instructions, system access, scheduling, training materials and other resources.

D3M TA3 Summer 2020 Formal Evaluation User Resource Guide

Overview

The following live document serves as an aid for test users of the Data-Driven Discovery of Models (D3M) program's TA3 Summer Evaluation taking place in August 2020. The goal of this evaluation is to assess the program's three TA3 systems' ability to support a common set of machine learning problem types. Users are tasked to evaluate these problem types across various datasets that are relevant to the program's transition partners. This evaluation serves as a critical element that will inform the next phase of the program which will focus on direct engagements with transition partners.

This document should provide you with all of the information you need to complete this task. However, feel free to reach out to Michael Liang (mliang@parenthetic.io) if you have any additional questions about the evaluation process.

Note that this document has a number of important links so you can find a Quick Links section at the very end for easier reference.

Program Introduction

The DARPA D3M program automates the methods in data science to create empirical models of real, complex processes. D3M enables domain experts to make predictions from data without the need for data scientists. The program accelerates scientific discovery and intelligence analysis by automatically searching the complex model space and discovering and explaining models to users.

Tasking

Users are tasked with evaluating three (3) TA3 systems based on a general analytical workflow, assessing each system's ability to support the user in:

1. **Loading** new data (e.g., training data) into the system
2. **Exploring** data to identify relationships and important elements of the data such as identifying outliers
3. **Preparing** data for model construction such as removing outliers
4. **Modeling** based on user-defined parameters
5. **Explaining** model performance in clear terms to the user
6. **Deploying** models against new data (e.g., test data) and explaining the results

This workflow will be applied to each dataset and system. Ultimately, each user is expected to conduct fifteen (15) walkthroughs (3 systems x 5 datasets). Users will rate each system using a scoresheet and rubric, providing detailed notes to support their ratings.

www.parenthetic.io

Note that each dataset has two files:

1. **Training file:** This is the data you will use for workflow categories spanning Load to Explain. It is the data that you are “training” each system to learn about.
2. **Testing file:** This is the data you will use to apply the models you create from the training data. This will help you validate just how deployable your model is when applied to new data.

Schedule

Tuesday (7/28): User Kickoff Call

Friday (7/31): TA3 systems submitted for evaluation and training materials released to users

Wednesday (8/5): Testing start date

Friday (8/14): Testing end date

TBD: Users meet to discuss evaluation outcomes

Testing is scheduled to begin by August 5, 2020. TA3 systems will submit their systems by July 31st at which systems will have to go through basic validation before deemed ready for user testing. User testing can begin as soon as systems are through the vetting process so users may start earlier if this is the case. Regardless of the starting date, all evaluations should conclude no later than August 14, 2020.

Communications

There will be two dedicated Slack channels set up during the test. One will include users and developers where developers will make themselves available to address any questions or issues users may experience with the systems. The second channel will be for only users, allowing for more candid discussions related to the evaluation

Slack Workspace: datadrivendiscovery

Slack Channel: summer2020ta3eval

Slack Workspace: datadrivendiscovery

Slack Channel: summer2020ta3eval_users only

If you do not have access by the time you have begun testing, please reach out to Michael Liang (mliang@parenthetic.io).

Training

Each TA3 team will provide training documentation as well as walkthrough videos to show users basic workflows within each system for two of the evaluation datasets.

Evaluation Scoresheet

We have set up an evaluation scoresheet on Google Sheets, accessible [here](#). This will be your primary workspace to document your evaluation scores and detailed notes. Each user has their own designated tab to record their information, but feel free to download a copy to your local machine if you'd prefer to work offline.

Each user's tab has four (4) types of boxes:

Grey: Instructions for filling out scoresheets

Green: Breakdown of the 5-point scoring rubric

Red: List of key features to keep in mind. This breaks down the six (6) workflow categories into subcategories. This is not an exhaustive list, but serves as a baseline to give users an idea of key features to look out for during their evaluation of each category.

Blue: Scoresheets – these are the only tables you are expected to populate. Note that there are fifteen (15) individual scoresheets on each tab to account for the 15 walkthroughs. Be sure to label each with the appropriate system and problem type.

Systems Access

In order to access the systems, you will first need a set of credentials for data.drivendiscovery.org which you should have received in your email.

Unique system links for each user, including credentials are provided below. These will need to be entered on the screen after you enter your D3M credentials.

User	System #1 URL	System #1 credentials	System #2 URL	System #2 credentials	System #3 URL	System #3 credentials
Tester 1	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 2	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 3	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 4	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 5	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 6	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 7	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 8	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 9	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 10	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 11	Link	n/a	Link	test/test	Link	test_user/test_user

TA3 Systems

Below are details of each TA3 system:

System	Team	
System #1	Performer Team #1	
System Description System #1 will enable SMEs with no data science training to dynamically derive, understand and apply data-driven models in answering mission critical questions. This approach augments human intellect with data discovery, enrichment, analytic model recommendation, and automated visualization to discover observable relationships in data, empowering intelligence action.		
System #2	Performer Team #2	
System Description System #2 provides a fundamentally re-thought data analytics stack: domain experts without programming experience or background in statistics can now explore and mine data through an intuitive user interface, and effortlessly build, analyze, and evaluate machine learning pipelines.		
System #3	Performer Team #3	
System Description System #3 is a data-driven web application designed to bring researchers to insights fast. It facilitates intuitive machine learning, model discovery, and data exploration for researchers. As System #3's intelligent back-end automatically seeks interesting relationships in the data and builds models to predict outcomes, researchers impart substantive knowledge about their data and own research questions to guide the automated generation of AI assistance for data analysis in an interactive paradigm we call <i>human-guided machine learning</i> .		

Problem Types

Each TA3 system is expected to support the analysis for the following problem types which are associated with five (5) specific datasets, listed below:

Dataset	Relevant Transition Partner	Binary Classification	Multiclass Classification	Regression	Time Series Forecasting
Dataset #1	Transition Partner #2	X			
Synthetic Dataset	Transition Partner #1	X			
Dataset #2	Transition Partner #1		X		
Cancer Mortality Rate	N/A			X	
Sorghum Plant Growth	TERRA				X

Below is a brief description of each problem type:

1. **Binary Classification:** Involves classifying elements into two (2) distinct groups based on certain properties
2. **Multiclass Classification:** Similar to binary classification but involves classifying elements into three (3) or more distinct groups
3. **Regression:** Involves predicting the outcome of an event based on the relationship between variables in a dataset
4. **Time Series Forecasting:** Involves the analysis of temporal data (e.g., data collected at different points in time)

Datasets

The datasets are all hosted [here](#)

Here is a [link](#) to MIT's slides providing details on each dataset which was shared during the Evaluation User Sync call on Tuesday, July 28th. Below are detailed summaries of each dataset for quicker reference:

1. News Events

Dataset Name: Dataset #1

Summary: Consists of news articles tracking certain news events

Problem Type: Binary classification

Transition Partner: Transition Partner #2

Contents: Each row represents a single article and column variables include details about the event (e.g., event type, date, location, etc.).

Target variable: "articleofinterest" - Indicated if the article in the row is relevant to an analyst or not (0=not relevant, 1=relevant)

Research questions include:

- Is it possible to create a model to identify articles of interest?

2. Dataset #2

Summary: Contains administrative data extracted from an electronic filing and application processing system.

Problem type: Multiclass classification

Transition Partner: Transition Partner #1

Contents: Each row represents an application column variables covering details about the employer, employee, business, etc.

Target variable: "Case_Status" - Application status (Certified, Certified-Withdrawn, Denied and Withdrawn)

Potential user questions to ask:

- What factors are most predictive of visa issuance or denial?

3. Cancer Mortality Rate

Summary: Contains cancer incidence and mortality rate data per county.

Dataset Name: LL1_Cancer_Mortality_Prediction

Problem Type: Regression

Transition Partner: Not Applicable

Contents: Each row represents a single county and column variables represent socio-economic data such as income, education, demographics, etc. Column descriptions can be found [here](#) (see Cancer Mortality Rate tab).

Target variable: “target_deathrate” - Mean per capita (100,000) cancer mortalities

Research questions include:

- What factors are most correlated with cancer incidence in a county?
- What factors are most predictive of the rate of deaths due to cancer in a county?

4. Sorghum Plant Growth

Summary: Consists of plant growth data between April-August 2017 where canopy heights were measured daily.

Dataset Name: LL1_terra_canopy_height_long_form_s4_80_MIN_METADATA

Problem Type: Time Series Forecasting

Transition Partner: TERRA-REF Project

Contents: Each row represents a daily measurement and column variables cover additional details such as cultivar and day. Column descriptions can be found [here](#) (see Sorghum Plant Growth tab).

Target variable: “canopy_height” - Measured height of the plant in centimeters

Research questions include:

- Can you forecast end-of-season growth based on early season growth?
- Can you increase forecast accuracy by augmenting the model with weather data (e.g., temperature, precipitation, etc.)?

5. Synthetic Dataset

Summary: A dataset that was synthetically generated to have certain characteristics, acting as a proxy for Transition Partner #1's internal data.

Dataset Name: LL1_MITLL_synthetic_vora_E_2538

Problem Type: Binary classification

Transition Partner: Transition Partner #1

Contents: Each row represents a case with the column variables representing 44 predictor variables. The data mimics the samples, features and class imbalances of the actual transition partner data. There are no semantics associated with the column variables.

Target variable: "y"

Research questions include:

- Can you accurately classify the target variable into either outcome (values of 0 or 1)?

Quick Links

1. [Evaluation Scoresheet](#)
2. [Datasets](#)
3. **Training Materials**

System	Training Documentation	Video Walkthrough
System #1	System #1 Guide	System #1 Walkthrough
System #2	System #2 Guide	System #2 Walkthrough
System #3	System #3 Guide System #3 Basic Training Flowchart	System #3 Walkthrough

4. System Links

User	System #1 URL	System #1 credentials	System #2 URL	System #2 credentials	System #3 URL	System #3 credentials
Tester 1	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 2	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 3	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 4	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 5	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 6	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 7	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 8	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 9	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 10	Link	n/a	Link	test/test	Link	test_user/test_user
Tester 11	Link	n/a	Link	test/test	Link	test_user/test_user

Appendix C - System Submission Procedure

In order to be deemed ready for evaluation, each TA3 team had to follow the procedure below when submitting their system for evaluation. This procedure was designed and overseen by DataMachines Corp.

Submission Procedure for TA3 Summer 2020 Evaluation

1. Create a new project under your subgroup here: <https://gitlab.datadrivendiscovery.org/ta3-submissions>.
 - a. Name your project Summer2020Evaluation
2. Build and push your image(s) to the repository registry created above
3. Upload the Dockerfile you used to build your image(s) to your project repository
4. Upload your ta3.yaml file (if applicable) to your project repository
5. Upload training materials
6. Create a README file with any instructions necessary to run your system
7. Test deploying your system on the Kubernetes Performers Server

Checklist

- ☐ You've included detailed instructions for starting your system
- ☐ You/Your partner's TA2 pod is listening on port 45042
- ☐ You've exposed port 8080
- ☐ You use your partners TA2 submission repo image (not a privately hosted image) in your ta3.yaml
- ☐ Your docker image is hosted on the D3M repository (<https://gitlab.datadrivendiscovery.org/ta3-submissions>)
- ☐ You tested your system on the Kubernetes Performers Server

PARENTHETIC
INFLUENCE INTELLIGENTLY

7.0 LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

ACLED	Armed Conflict Location & Event Data
Cronbach's α	The most common test score reliability coefficient
D3M	Data-Driven Discovery of Models
DARPA	Defense Advanced Research Projects Agency
DoD	U.S. Department of Defense
H1-B	A visa in the U.S. that allows U.S. employers to temporarily employ foreign works in specialty occupations
HITL	Human in the Loop
I2O	DARPA Information Innovation Office
IRB	Institutional Review Board
JPL	NASA Jet Propulsion Lab
ML	Machine Learning
SETA	Systems Engineering and Technical Assistance
SME	Subject Matter Expert
TA	Technical Area
TERRA	Transportation Energy Resources from Renewable Agriculture
UI	User Interface
USG	United States Government