Data-Driven Discovery of Models (D^3M) Questions and Answers

(As of June 23, 2016)

Q24: We are planning on submitting abstracts to the D3M program, and currently have two fundamentally different ideas for addressing two different problems in the TA1 space. One idea is on developing anomaly detection primitives and the other is on primitives for modeling heterogeneous temporal sequences. Can we submit two abstracts - one for each idea - even if both address the same TA?

A24: Hello, we encourage you to submit one abstract with both sets of primitives.

Q23: Can an organization submit multiple proposals?

A23: Proposers may bid on any technical area individually or any combination of technical areas thereof. Proposers who include multiple technical areas in their proposal should describe the work in each technical area separately and should provide separate costing for the work proposed in each technical area. This is required to enable DARPA to select tasks in one or more technical areas when work is proposed in multiple technical areas.

Q22: Can an organization submit multiple abstracts?

A22: Yes.

Q21: How many awards are expected for contracts? How many for grants? Are there any set asides for small business?

A21: Per the BAA, multiple awards are anticipated. Per the BAA, no grants will be awarded. There are no asides for small business.

Q20: The program manager mentioned 5 to 6 weeks of travel to DARPA/Virginia. Was that per year?

A20: Yes, performers will be required to send integration programmers to support integration and evaluation.

Q19: What size ($) will the awards be?

A19: No specific size. Depends on the proposed effort and selectability.

Q18: Is it okay for proposals to respond to all 3 TAs?

A18: Yes. See BAA for more details.

Q17: Is there a requirement to only work with open data? Must the tools be open source?

A17: Yes, for open data and open source tools. If there is a really compelling reason for tools not to be open source, it should be explained in submitted proposal for consideration.

Q16: Can the proposal be attached to a specific application scenario.

A16: No.

Q15: The program seems to be focused on discrete modeling of finite data sets. Is it in scope to consider streaming data or other types of problems with ever changing ground truth and/or practically infinite data?

A15: We assume there is a well-formed outcome. There are no restrictions on data.

Q14: Is there any scope for proposing new modeling primitives? Is there any interest in new learning methods for discovering models beyond what is available in state of art?

A14: Yes and yes. This is the purpose of TA1.

Q13: In TA3, can humans also provide knowledge to the system e.g., causal information, physics first-principle models? How is TA1 supposed to differ from the current work in AutoML?

A13: Yes, on TA3. TA1 is about new primitives and making them discoverable directly from data.

Q12: Does the program envision a 'pipeline' view of the composition, or is a more 'joint inference' view is also appropriate?

A12: Both. Ideally, both classes of model compositions are made possible in TA2 as part of D^3M.

Q11: Is the model limited to machine learning models?

A11: Program interest is in all empirical models derived from data.

Q10: What is the anticipated domain of expertise of subject matter experts?

A10: Domains expected include sociology, biology, intelligence analysis, and physical sciences.

Q9: Do the primitives need to be generic, or they can be specific to certain domains e.g., social media, health?

A9: They could be specific. Data transforming/cleaning primitives are often to be domain specific. When possible, we prefer generality (i.e. if a domain-specific primitive can be generalized and applied to multiple domains, we prefer the more general version).

Q8: What do you see are major differences between XDATA and D^3M?

A8: TA1 has XDATA elements but with discoverability as a constraint. TA2 and TA3 have no analog in XDATA. XDATA built a library of components for data science. A programmer was required to build specific models to solve specific problems. D^3M automates the construction of these models.

Q7: Could you clarify the role of the Government team vs. TA1/TA2/TA3 performers in terms of who is responsible for determining how to specify TA1 components (characteristics, data requirements etc.)?

A7: The government team will build corpus for performers to mine, construct a baseline library for TA1 from publicly available components, construct an integration platform, and perform technology evaluation. TA1 performers will extend the baseline library, build

automated composition methods in TA2 and design user-facing interactions to formalize modeling outcomes and curate models in TA3.

Q6: Chris Mattmann (from the government team presentation) talked about using mechanical Turk for eliciting ground truth. How does this relate to the work scope for D^3M?

A6: It is only relevant for collection of the corpus and evaluation?

Q5: What aspects of the API development will be the performer's responsibility, and which will be that of the GFI teams?

A5: Performer's responsibility is to create new primitives and making them discoverable. They will also work with the government team to define the API for integration and evaluation.

Q4: D^3M is supposed to solve easy problems so that data scientists can be occupied with hard problems. Where do you draw the line between easy and hard? How hard are the problems that D^3M is expected to solve?

A4: It is a secondary goal. The primary target for D^3M is to solve all problems without data scientists.

Q3: How do you reconcile the TA3 goal of explaining models with the TA2 goal of composing models that exceed human comprehension?

A3: The explanation should be good enough for curation. In general, this may not require deep understanding of how data is processed internal to a model. See BAA for more details on TA3.

Q2: How is TA3 going to differ from an explainable AI program?

A2: The goal here is curation and control of TA2 processes.

Q 1: Will performers have input on what the evaluation problems are?

A1: Yes, but the stakeholders/users will have more input.