Request for Information (RFI)
DARPA-SN-17-32
Foundations for Generalizable Representation and Automated Extraction of Knowledge

Responses Accepted: Until 1:00 PM (Eastern) on April 14, 2017
Point of Contact: Dr. John Paschkewitz, Program Manager, DARPA/DSO
Email Address: DARPA-SN-17-32@darpa.mil

The Defense Advanced Research Projects Agency (DARPA) Defense Sciences Office (DSO) is requesting information on methods for representation and extraction of knowledge in scientific data analysis and modeling to support computational reasoning over diverse sources of evidence. Scientific discoveries rely on piecing together evidence from many sources—including mathematical, computational, experimental, and observational—spanning multiple modalities and scales. Because of the diversity of the data types and reasoning involved, this fusion generally only occurs in the brains of researchers, resulting in fragmented understanding and leading to a significant bottleneck to new discoveries. New representations and algorithms for connecting such heterogeneous forms of scientific information would greatly accelerate the research process.

The DARPA/DSO Simplifying Complexity in Scientific Discovery (SIMPLEX) program has developed a variety of tools and methods to aid scientific research exploiting multimodal data. While SIMPLEX has led to new capabilities in a number of domains, DSO is looking to extend and conceptually unify these results, which span both general purpose abstract methods and domain-specific application tools. The apparent tension between domain-specific tools requiring tuning to specific problems and general purpose abstract methods that require idealization of the data stems from an underlying lack of a properly rigorous mathematical framework for automatic knowledge extraction, which is the focus of this RFI.

To foster applicable responses, definitions of key terms are provided below:

Knowledge: The ability to "answer questions" about the underlying processes that are generating the data. A formal foundation for automated knowledge extraction should unify the relationship between data, the process by which it is produced and an efficient description of its defining intrinsic properties in a sufficiently expressive and manipulable representation.

Knowledge Representation: An efficient encoding of the instructions that will enable purposeful action based on the behavior (past, present and future) of the process/system of interest. This representation may exist monolithically or be distributed. In the latter case, methods for effective local reconstruction are part of the representation.

Semantics: The application of context to data. "Context" means choice of a model hypothesized to have produced the given data. Since this process is stochastic in nature, having fixed context, one can speak of semantic information. Semantic markup then is the "annotation" of data that affects association with context.

Datafication: Encoding domain-specific information into a representation, using a process of associating data to (parts of) another structure, usually geometric or perhaps topological. The existing relationships between the component parts of the underlying structure (e.g., proximity) can be hypothesized to correspond to correlations in the data. Conversely, the data can be assumed to contain relevant, and perhaps even latent, information about the structure itself.

Some outstanding challenges requiring new formal representations of knowledge include the following:

- Analysis of inexact semantic alignment between models, datasets, or paradigms. For example, a property value measured by experiment may systematically differ from the value computed by a model (or from values measured by another experimental method). Different models or analyses may make slightly different assumptions about the same underlying system. In such cases, naïve integration may produce inconsistencies in the result, while treating the items as distinct may thwart any possibility for integrated analysis. Therefore, reconciling these sorts of conflicts is critical for practical use and may require novel integration of soft constraints and learning.
- Automatic discovery of behavioral or temporal attributes in dynamic data at multiple levels of abstraction. For example, while it is straightforward to determine basic rules or structures of simple behaviors (e.g., determining the rules of an unknown game from video footage), it is much harder to determine higher-level tactics or cross-scale interactions, and harder yet to determine strategy or causal mechanisms using conventional representations.

Information regarding approaches of interest to DARPA may include, but is not limited to the following:

- Topological methods: For example, datafication can be interpreted in topological terms as defining a sheaf of sets over a topological space. These sets may have additional (algebraic) structure, in which case one may get a sheaf of Abelian groups, rings, vector spaces, etc. The value of this sheaf theoretic formulation is twofold: first, data consistency is tautologically enforced; second, appropriate formulations allow for the use of (co)homological methods, which are/can be formal and computational, to answer questions about the feasibility of taking "datafied" local information and producing a globally consistent extension.
- Probabilistic program induction[1]
- Differentiable, stochastic logic
- Distributed representations[2]
- Semantic information theory
- Representation by reference: The notion of using associations or connections between

[1] For example, recent work by Tenenbaum and co-workers (http://web.mit.edu/cocosci/Papers/Science-2015-Lake-1332-8.pdf) but note that this application is not of interest as noted below.
[2] For example, R.V. Guha, "Towards a Model Theory for Distributed Representations," https://arxiv.org/pdf/1410.5859.pdf

entities (not necessarily causal) as a method for unique identification. For example, an approach is using labeled and decorated graphs so that an entity or "concept" is uniquely represented by the labeled and/or decorated set of vertices/edges to which it is adjacent.[3]
- Datafication (as defined): Data ingestion, metadata extraction and semantic markup as manifested in a principled framework for knowledge representation and extraction.

Conventional methods for data classification, image segmentation, and data fusion are explicitly not of interest, nor are incremental extensions of current approaches. The following are also not considered to be of interest:

- Methods that are specific to a particular modality (e.g., images, text)
- Development of representations without corresponding analytic innovations
- Work that is primarily software engineering/integration
- Representations that simply store significant portions of the informative features as attributes or metadata without a means to include them in computational analysis.
- Work primarily focused on representing "common sense" (e.g., Artificial Intelligence [AI]) vs. complex domain-specific knowledge as occurs in research.
- AI research for its own sake (rather than as a means to scientific and engineering data analysis and modeling).
- Methods that exclusively deal with human-curated, post-publication data (vs. research data)
- Methods that cannot handle quantitative data

DARPA is interested in information pertinent to the following topics, as applicable:

- Representation: Define your proposed representation and the kinds of data, models, and other domain information it can support. What advantages does this provide compared with other representations? What kinds of details does it not capture, and how does this impact its usefulness to research applications?
- Mathematical & Algorithmic foundations: What assertions involving postulates and hypotheses can be formally verified and validated? What are the associated computational complexity and decidability properties of the algorithms associated with the representation?
- Analysis & Reasoning: Describe the kinds of analysis you can perform using multimodal data with the chosen representation. How will the algorithms make use of the rich structure of the data (i.e., the informative features that domain experts use to reason about the implications of the data)?
- Application domains: Identify at least two distinct research areas the approach will support and where datasets will be obtained to demonstrate value. These datasets should have research value in their own right, not simply for showcasing chosen methods. Identify a set of meaningful quantitative metrics to evaluate the performance of methods associated with specific tests of hypothesis generation. What impact will it make in those domains? What research questions does it answer?

[3] For example, R.V. Guha & V. Gupta, "Reference by Description," https://research.google.com/pubs/pub44679.html

SUBMISSION FORMAT
Respondents to this RFI are encouraged to be as succinct as possible, while also providing actionable insight. Responses are limited to 7 pages (1 page cover sheet + 5 pages text + 1 page bibliography/references) Format specifications for responses include 12-point font, single-spaced, single-sided, 8.5 by 11-inch paper, with 1-inch margins in MS Word or Adobe PDF format.

Respondents are responsible for clearly identifying proprietary information. Responses containing proprietary information must have each page containing such information clearly marked with a label such as "Proprietary" or "Company Proprietary." DO NOT INCLUDE ANY CLASSIFIED INFORMATION IN THE RFI RESPONSE.

SUBMISSION INSTRUCTIONS AND CONTACT INFORMATION
All responses to this RFI must be emailed to DARPA-SN-17-32@darpa.mil. Responses will be accepted any time from the publication of this RFI until 1:00 PM (Eastern) on April 14, 2017. Early responses are encouraged.

All technical and administrative correspondence and questions regarding this RFI should also be sent to the same email address. Emails sent directly to the Program Manager may result in delayed/no response.

ELIGIBILITY
DARPA invites participation from all those engaged in related research activities and appreciates responses from all capable and qualified sources including, but not limited to, universities, university-affiliated research centers (UARCs), Federally-Funded Research and Development Centers (FFRDCs), private or public companies and Government research laboratories.

DISCLAIMERS AND IMPORTANT NOTES
- This is an RFI issued solely for information and new program planning purposes; it does not constitute a formal solicitation for proposals. In accordance with FAR 15.201(e), responses to this RFI are not offers and cannot be accepted by the Government as such.
- Responses do not bind DARPA to any further actions related to this topic including requesting follow-on proposals from respondents to this RFI.
- Submission is voluntary and is not required to propose to a subsequent Broad Agency Announcement (BAA) (if any) or other research solicitation (if any) on this topic.
- DARPA will not provide reimbursement for costs incurred in responding to this RFI.
- Respondents are advised that DARPA is under no obligation to acknowledge receipt of the information received or provide feedback to respondents with respect to any information submitted under this RFI.
- DARPA will disclose submission contents only for the purpose of review. Submissions may be reviewed by the Government (DARPA and partners); Federally Funded Research and Development Centers (FFRDCs); and Scientific, Engineering and Technical Assistance (SETA) support contractors.