

Reference ID: 11227025814\_Chaturvedi

---

**Reference ID:** 11227025814\_Chaturvedi

**Submission Date and Time:** 12/16/2019 4:35:49 PM

This contribution was submitted to the National Science Foundation in response to a Request for Information, <https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp>. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

*Consent Statement:* "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publicly available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

**Consent answer:** I consent to NSF's use and display of the submitted information.

#### **Author Names & Affiliations**

Submitting author: Alok Chaturvedi - Purdue University

**Additional authors:** None

**Contact Email Address** (for NSF use only): (hidden)

#### **Research domain(s), discipline(s)/sub-discipline(s)**

self-organizing systems; Generative AI; Machine learning; auto-scaling virtual brain model

#### **Title of Response**

Self-configuring Infrastructure for generative analytics

#### **Abstract**

Experts have traditionally built models to study narrow problems in specific domains. While these models are insightful, they partially address the diverse issues related to complex, real-world problems such as emergency response, global corporate strategy, and military campaign planning. These real-world problems require multi-disciplinary thinking, multi-scale forward and inverse representations of

problems; multiple analytical points of view; massive numbers of entities representing multiple sides with diverse interests; and emergent interactions and behaviors. The traditional approach to building comprehensive, requirements-driven models does not work for such problems. Knowledge Portal (KP) is a data-intensive cyberinfrastructure that builds a repository of data, models, scenarios, experience, and experiments through generative, adversarial, and automated integration by taking direct advantage of the collective knowledge of the community and AI. This architecture will support integration in a seamless and automated manner, allowing users to “plug-and-play” models and data into scenarios. Additionally, the architecture supports a community-like ability to develop, submit, and share data, models, scenarios, and experiences. KP provides mechanisms of sharing knowledge components as well as provide access restriction and control to protect proprietary data and provide metering for usage-based revenue sharing.

**Question 1 (maximum 400 words) – Data-Intensive Research Question(s) and Challenge(s).** Describe current or emerging data-intensive/data-driven S&E research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.

There are several challenges related to data-intensive research. These challenges include too much data, too little data, data in different scales and formats, data generated by phenomena with unknown understanding, different modeling paradigms, various assumptions and points of view (perspective), different levels of analysis, questionable rating methods, and deliberate misinformation. Further, users of these data and models from open source repositories (Github, Bitbucket, etc.) have limited knowledge of data science, machine learning and AI, domain knowledge of diverse disciplines, analysis, and interpretations. For example, if one tries to combine data from neuroscience and economics, proper adjustments for scale has to be made; constructing a scene before and after a crime from images from different camera angles and times of shots will require serious analytics consisting of points of view and temporal analysis, resolution matching, and forward and backward simulations, etc. Therefore, a cyberinfrastructure is needed that can support a non-expert user to act and behave like an expert in multiple disciplines. Such architecture should allow users to optimally integrate models and data to create scenarios and experiments by taking direct advantage of generative AI and collective knowledge of the community of diverse users in a seamless and automated manner. Users may “plug-and-play” models and data, seeing how inclusion and exclusion of certain information affects results. Additionally, the architecture must support a community-like ability to develop, submit, rate, and share data, models, scenarios, and experiences.

**Question 2 (maximum 600 words) – Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s).** Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing,

data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of data heterogeneity, data integration and interoperability? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?

The core technology of KP may consist of two primary components – representation of Knowledge Component and Knowledge Integration Backplane (KIB). An open modeling framework for managing Knowledge Components. A knowledge component is encapsulated through AI-driven Universal-Lexicon Descriptors that handle the semantic and syntactic translation of inputs, allowing otherwise heterogeneous components seamlessly interact with one another. An intelligent Knowledge Integration Backplane (KIB) that forms a composition framework that facilitates seamless integration of generative components. KIB can be conceptualized as a software analogous to a computer motherboard, through which anyone component can connect and interact with any other component connected on the bus. KP's open Modeling Framework enables users to develop and integrate models for diverse purposes (Economics, Psychology, Political Science, etc.) using different modeling paradigms (Agent-Based Modeling, System Dynamics, Differential Equation, supervised and unsupervised, reinforcement, and deep learning methods, etc.). Models of different disciplines may use these constructs to specify Universal-Lexicon Descriptors (ULDs) that fully define the syntax, granularities, and semantics of all data produced and consumed by the model in a standardized format. The ULDs are used to integrate multiple models together, by Knowledge Integration Backplane. Including a new model in an executable scenario, need to follow a general, three-step process. First, the nature of the model is defined, including model logic, inputs and outputs, and the temporal and spatial nature of the model. Models' specifications are identified independently of all other models and data components, enabling the model to be integrated with numerous data sources and models that fulfill the model's input requirements. Second, the ULD for a model is generated, specifying the syntax, granularity, and semantics of each of the model's inputs and outputs. The semantics identified within the ULD relate to concepts in a Domain-Specific Ontology that is used for integration. Third, a model is integrated with other models and data sources in a scenario and for execution. The KP Modeling Framework provides ease-of-use features that make model development accessible to a broader community of skill sets and speeds the development process. The features of the Modeling Framework enable various model development tools as well as diverse skill sets to help generate and interact with models in KP. Through the Modeling Framework, models are specified declaratively in a common markup language that is independent of a programming language used for implementation, separating model constructs from the model logic. Knowledge Integration Backplane (KIB) provides the framework for generative and automated integration of models and data, which, when combined with user selections and parameterizations, enables the construction of customized executable scenarios. Integration through KIB enables data sources and models to be mixed and matched interchangeably. KIB may use a mechanism that ensures semantic consistency among heterogeneous models as well as simply linking together data feeds. Additionally, the ontological basis of integration in KIB enables self-assembly of generative components for the rapid and extensible construction of scenarios. KIB allows a scenario to

be composed of models in an incremental fashion. A user creates a simplistic scenario by composing a model with data sources that fulfill the model's input requirements to test or better understand the nature of a particular model under certain conditions. A user can then proceed to augment static data sources with models, executing each revision within KP. Once components are composed into scenarios in KIB, the data exchange that occurs during execution occurs in the Cloud. Scenarios once executed, any portion of results are saved in the Repository for incorporation in another scenario as a data source or sharing experiences with the community.

**Question 3 (maximum 300 words) – Other considerations.** Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.

KP Sustainability KIB may facilitate access control and metering of components at the individual level. As a user creates an element, that individual can set access permissions and billing rates for access to that particular component. All KP users may be assigned security profiles. Enforcement of security occurs at two points. During scenario construction, any selected component and its references are scanned for discrepancies in security access, informing the user if a component cannot be accessed. Once a session to KP, all accessed components are rechecked, flagging any inaccessible components and suggesting alternative paths or terminating execution. This approach allows users to automatically alter the set of requested components both before and during scenario execution, avoiding wasted time and resources. One key source of revenue in KP is generated through usage-based fees metered at the individual component level. Revenue sharing for a contributed component is produced by two means: accesses either directed by a user or through Nth-order references by other components. In either case, revenue is metered by the number of executions, accesses to the component, and component types. For example, models are measured by the number of executions while the number of reads measures data sets. The metered rates are also applied during scenario construction and prior to execution, keeping users informed of the cost incurred by a scenario run. Revenues from this channel are divided among the creator of the component and the platform. To enable this, KP tracks usage of every component through source tags that tie the component back to its original creator. As a user builds a scenario that is using metered components, he or she is presented with an estimated cost for executing an excursion based on that scenario.

-- End Submission --