Computing Community Consortium (CCC) Response to NLM Request for Information

The Computing Community Consortium (CCC) prepared this response. The mission of the CCC is to catalyze the computing research community and enable the pursuit of innovative, high-impact research. Our goal is to call attention to major research opportunities for the computing community. In our response to NLM's request for information we call attention to **promising directions for new data science research in the context of health and biomedicine.**

Holistic Sensing, Advanced Analytics and Comprehensive Decision Support

Moving the needle on healthcare costs and outcomes requires embracing this new frontier of data, analytics, mobile and in-situ interventions, and integrated human-centered systems. Improved clinical care and discovery require developing new scientific foundations that integrate new data available from the "exposome" that surrounds and influences human behavior and health (NIOSH, 2014). Public health, rooted in the need for accurate interpretation of available data and effective communication, also has much to gain from the growing availability and robustness of sensors, mobile health capabilities, data analytics, and human-system modeling.

Always available health sensing and behavioral monitoring offers the transformational possibility of advancing health and wellness that is bound by neither space nor time. This new transformation in health technologies enables observations of dynamic changes in each individual's health state, as well as key physical, biological, behavioral, social, and environmental factors that contribute to health and disease risk, anytime and anywhere that have not been possible in the past.

Although there has been a significant progress in the use of health sensors for remote care for specific diseases (e.g. heart rate monitors), future strides will come from "holistic" sensing approaches that integrate wearable, environmental, behavioral, and social network information and create new computational models that can dynamically recognize the influence of environmental factors on human physiology and behavior. Much of the current work in health sensing has focused on point solutions for specific disease conditions or states, and has largely been informed by conventional diagnostic techniques that do not take into account environmental and social information. For example, with the explosive growth of social media data, it is now possible to obtain temporally varying measurements of interpersonal exchanges, social network structures, and social capital, which can be combined with complementary physiological and psychological sensor measurements of an individual. These dynamic measurements can also reveal network and community effects in health outcomes — quantification of which has been challenging so far due to the paucity of adequate social and community-level data. Collectively this holistic picture of a person's health and health-related behaviors complements today's genomic and clinical data while honing in on the social and environmental factors that primarily contribute to disease onset and progression and that pose barriers to good health.

These approaches also aim to capture the *context* surrounding human behavior to effectively guide healthcare delivery. For example, the noise and bustle of an ICU should affect the content of ICU information displays, just as the availability of healthy food and measures of social influence should affect diabetic care and nutritional coaching. In another example, given the range of potential trigger events for PTSD, effective treatment plans for an individual would benefit from historical and real-time awareness of a person's social and environmental context. This combination of approaches, integrating information about a person's physiological and psychological state with their environmental and social context can enable the creation of targeted and personalized care to address numerous health challenges.

These approaches require addressing several challenging issues in analyzing data, designing decision-support systems, and developing strategic mechanisms for reaching informed responses from holistic sensing:

- How can we measure key environmental variables, such as exposure to cues and triggers for adverse health-related behaviors, and infer a comprehensive characterization of individual behavior?
- How can we develop models of health risks and intervention outcomes from longitudinal, multi-modal clinical, environmental, socio-demographic, and behavioral data?
- How can we optimize the measurements (e.g., clinical tests) to minimize costs and risks to patients while maximizing the utility of the measurements for diagnosis and treatment?
- How can we model the effect of environmental and social factors on behavior regulation to support behavior change and behavior maintenance?
- How do we capture and effectively model the context of and around an individual to guide "hows" of health promotion and health care delivery: e.g. how to best deliver information, how to best deliver interventions where and when they are needed, how to support caregivers, or how to best communicate with a patient?
- How can we support decision-making where prioritization of key factors, for example emotional vs. physical well-being or invasive vs. non-invasive treatment, is critical to effective health promotion and clinical care?
- How can we improve robustness and reliability of these systems operating "in the wild" outside of the traditional confines of healthcare environments?
- How can we integrate the needs and preferences of multiple stakeholders (health service providers, payers, patients, etc.) in healthcare delivery?

In turn this application-driven research exposes basic computing research challenges that are relevant across many domains including:

- Combining large-scale population "big data" with temporally dense as well as sparse individual data, controlling for biases in each to produce reliable inferences;
- Visualizing context-aware mixed-modality data analytics and high-level abstraction and summarization of large-scale, multi-modality data with uncertainties;
- Developing methods for reliable prediction and inference of health risks and intervention outcomes from multi-modal longitudinal data;
- Designing methods to extract events of interest from multi-modality data stream with varying temporal granularity, spatial irregularity, varying reliability and validity, and data incompleteness; and

- Developing provenance systems that capture both metadata and annotations of the entire data processing stage to facilitate both interpretability and comparative analysis;
- Developing infrastructure and methods that permit data access and use policy compliant integration and analyses of multi-modal, longitudinal health data;
- Developing methods and tools to for multi-stakeholder decision making;
- Developing methods and tools for effective organization and presentation of information from electronic health records in context to assist clinical decision making;
- Developing methods and tools for constructing models that enable joint-reasoning and collaborative decision-making with human-experts in the loop.

In order to realize the full potential of data to improve individual and population health outcomes, there is an urgent need to explore platforms and collaboration mechanisms that enable and promote large-scale, multi-site, reproducible studies to increase the reliability and robustness of findings.

In conclusion, these fundamental computing research challenges, when explored in the context of healthcare delivery, health and wellness, have the tremendous potential to address both long-standing national priorities in health and catalyze significant research advances in computing. Meeting these challenges requires sustained investment in basic research, while proactively integrating these visions into current NLM / NIH research programs, to address the needed research culture changes required for these efforts to have sustained impact. The CCC is available to partner with NLM to help meet these challenges to continue to strengthen and expand the scope of new data science research in the context of health and biomedicine. See the CCC's *Research Opportunities and Visions for Smart and Pervasive Health* (https://cra.org/ccc/wp-content/uploads/sites/2/2017/06/SmartandPervasiveHealth-White-Paper-June-2017.pdf) for more information.

This material is based upon work supported by the National Science Foundation under Grant No. 1136993. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Reference

National Institute for Occupational Safety and Health (NIOSH) (2014). Exposome and Exposomics. https://www.cdc.gov/niosh/topics/exposome/.

CCC
Computing Community
Consortium
Catalyst

Computing Community Consortium (CCC) Response to NLM Request for Information

The Computing Community Consortium (CCC) prepared this response. The mission of the CCC is to catalyze the computing research community and enable the pursuit of innovative, high-impact research. Our goal is to call attention to major research opportunities for the computing community. In our response to NLM's request for information we call attention to **promising directions for new initiatives relating to open science and research reproducibility.**

Accelerating Science

The emergence of "big data" offers unprecedented opportunities for not only accelerating scientific advances but also enabling new modes of discovery. Scientific progress in many disciplines is increasingly enabled by our ability to examine natural phenomena through the computational lens, i.e., using algorithmic or information processing abstractions of the underlying processes; and our ability to acquire, share, integrate and analyze disparate types of data. However, there is a huge gap between our ability to acquire, store, and process data and our ability to make effective use of the data to advance discovery. Despite successful automation of routine aspects of data management and analytics, most elements of the scientific process currently require considerable human expertise and effort. Accelerating science to keep pace with the rate of data acquisition and data processing calls for the development of algorithmic or information processing abstractions, coupled with formal methods and tools for modeling and simulation of natural processes as well as major innovations in *cognitive tools* for scientists, i.e., computational tools that leverage and extend the reach of human intellect, and partner with humans on a broad range of tasks in scientific discovery (e.g., identifying, prioritizing formulating questions, designing, prioritizing and executing experiments designed to answer a chosen question, drawing inferences and evaluating the results, and formulating new questions, in a closed-loop fashion). This calls for concerted research agenda aimed at: Development, analysis, integration, sharing, and simulation of algorithmic or information processing abstractions of natural processes, coupled with formal methods and tools for their analyses and simulation; Innovations in cognitive tools that augment and extend human intellect and partner with humans in all aspects of science. This in turn requires: the formalization, development, analysis, of algorithmic or information processing abstractions of various aspects of the scientific process; the development of computational artifacts (representations, processes, protocols, workflows, software) that embody such understanding; and the integration of the resulting cognitive tools into collaborative human-machine systems and infrastructure to advance science.

Accelerating science to keep pace with the rate of data acquisition and data processing calls for concerted research efforts that encompass both: (1) Development, analysis, integration, sharing, and simulation of algorithmic or information processing abstractions of natural processes, coupled with formal methods and tools for their analyses and simulation; and (2) Innovations in cognitive tools that augment and extend human intellect and partner with humans in all aspects of science.

Algorithmic Abstractions for Accelerating Science

The success of computational lens in shedding new light on long-standing questions in biological, cognitive, and social sciences is contributing to their transformation from descriptive sciences into predictive sciences. However, in most disciplines, this transformation is far from complete. In many areas, such abstractions are scarce. In others, the abstractions and the hypotheses that they offer have remained untested, at least in part, due in part to the limitations of our instruments of observation and experimentation and in part due to the cost and complexity of the scientific enterprise. In order for a broad range of sciences and scientists to benefit from the use of computational lens in their respective disciplines, there is an urgent need for developing, sharing, analyzing, and integrating computational abstractions or representations of the key entities, relationships, and processes of interest in the respective scientific disciplines. For example, progress in life sciences has been accelerated substantially with the emergence of gene ontology (Ashburner et al., 2000). Much work remains to be done in a similar vein in other scientific disciplines. Of particular interest are system-level, mechanistic, computational models of biological, cognitive, and social systems that enable the integration of different processes into coherent and rigorous representations that can be analyzed, simulated, integrated, shared, validated against experimental data, and used to guide experimental investigations. Such abstractions, coupled with formal methods for their analysis, can provide rich defined modeling languages with precise syntax and semantics that can be analyzed systematically and efficiently for certain properties of interest. For example, a question of interest to a cancer biologist, e.g. whether the up-regulation of genes A and B and down-regulation of gene C could possibly take a cell from a healthy state to a cancerous state can be translated into a reachability query against a model of a cell where the state of the cell encodes the expression levels of the genes. While there has been some progress in developing such abstractions for molecular and systems biology (Priami, 2009; Bernot et al., 2004; Danos and Laneve, 2004; Fisher and Henzinger, 2007), much work remains to be done, especially in relation to formalisms that allow specification of models that take into account uncertainty and variability, as well as couplings across multiple levels of abstraction, e.g., molecules, cells, tissues, organs, organisms. Similar advances are needed in other scientific disciplines. Of particular interest are formalisms for bridging models not only across levels of abstraction, but also, disciplinary boundaries, to allow studies of complex interactions, e.g., those that couple food, energy, water, environment, and people.

Cognitive Tools for Accelerating Science

In order for science to keep pace with the rate of data acquisition and data processing, there is an urgent need for innovations in cognitive tools for scientists (Saloman et al., 1991), i.e., computational tools that leverage and extend human intellect (Engelbart, 1962), and partner with humans on a broader range of tasks involved in scientific discovery (formulating a question, designing, prioritizing and executing experiments designed to answer the question, drawing inferences and evaluating the results, and formulating new questions, in a closed-loop fashion). This calls for deeper understanding formalization, and algorithmic abstractions of, various aspects of the scientific process; development of the computational artifacts (representations, processes, software) that embody such understanding; and the integration of the resulting artifacts into collaborative human-machine systems to advance science (by augmenting, and whenever feasible, replacing individual or collective human efforts). The resulting computer programs would need to close the loop from designing experiments to acquiring and analyzing data to generating and refining hypotheses back to designing new experiments.

Accelerating science calls for programs that can access and ingest information and background knowledge relevant to any scientific question. As search engines and digital libraries return more articles in response to a query than anyone can read, e.g., Google returns about 3.67 million hits for "cancer biology", there is a need for programs that can read, assess the quality and trustworthiness of, and interpret such information. With the exponential growth in scientific literature, often with conflicting scientific arguments, supported by observations of variable quality and analyses made under differing assumptions, there is a dire need for tools for managing conflicting arguments, tracking changes in the validity of the observations and assumptions that they rely on, and support justifiable conclusions. While there is considerable work on computational argumentation systems much work is needed to develop argumentation formalisms and tools that can help accelerate science. Of particular interest are expressive yet computationally tractable languages for representing and reasoning with scientific arguments, and their uncertainty and provenance.

A shift in emphasis from accelerating data collection and data processing to accelerating the entire scientific process calls for representation and modeling languages with precise formal semantics for describing, sharing, and communicating scientific observations (including measurement models) experiments, data, models, theories, conjectures, and hypotheses. The increasing reliance on cognitive tools requires that the all of these be specified in a form that can be processed by computers; and queries against them be translated into precise computational problems.

Even the relatively mundane task of data collection presents many questions including deciding which variables to measure, why, and how i.e., the instrument to use (if one exists) or to design (if need be). There is a need for languages and tools for describing the measurement process, the data models for describing observations using standard ontologies (when they exist), establishing semantics preserving mappings across data models. There is an urgent need for precise languages and tools for describing experiments, methods for quantifying the marginal utility of experiments, determining the scientific as well as economic feasibility of experiments, comparing alternative experiments, and choosing optimal experiments (in a given context). The same holds for hypotheses, conjectures, theories, scientific workflows, and other scientific artifacts.

Machine learning currently offers one of the most cost-effective approaches to constructing predictive models from data (Ghahramani, 2015; Jordan and Mitchell, 2015). However, such models are often complex hard for scientists to comprehend, and therefore to use to gain mechanistic insights into the underlying phenomena. Consider for example, a support vector machine using a non-linear kernel that predicts whether a target gene of interest is turned on or off based on the previous states of a few hundred other genes. Such a model, its high predictive accuracy, is virtually useless with regard to helping to uncover the underlying genetic regulatory network. There has been some progress in extracting comprehensible knowledge from complex predictive models (Pazzani et al., 1997). However, a significant language gap remains between model builders and model users. This language gap presents challenges in exploiting prior knowledge to guide model construction, and in interpreting predictive models produced by machine learning in advancing scientific understanding of the underlying domain. There is an urgent need for a new generation of machine learning algorithms that that can incorporate prior knowledge and constraints from a variety of sources, e.g., from physics, and produce models are expressed in forms that are easy to communicate to disciplinary scientists.

Answering complex questions increasingly requires synthesizing the findings from data from disparate observational and experimental studies to draw valid conclusions. Conclusions that are obtained in a laboratory setting may not hold exactly a setting that differs in many aspects from that of the laboratory. Often, individual studies, for practical reasons focus on the relationship between a selected set of experimental variables and a specific outcome variable. This means arriving at meaningful answers to questions of interest invariably requires synthesize the findings from multiple such studies, carried out under related, but different experimental settings, under possibly different experimental constraints (e.g., experiments that can be performed on a mouse cannot be carried out on human subjects). A great deal of work is needed to characterize the precise conditions under which findings of disparate observational and experimental studies can be synthesized, and to develop cognitive tools for synthesizing such findings.

While we have effective tools to assist scientists in routine aspects of data management and analytics, most of the other steps in the scientific process currently constitute rate limiting steps in scientific progress. These include: Characterizing the current state of knowledge in a discipline and identifying the gaps in the current state of knowledge; Generating and prioritizing questions that are ripe for investigation based on the current scientific priorities and the current state of knowledge; Designing, prioritizing, planning, and executing experiments; Analyzing and interpreting results; Generating and verifying hypotheses; Drawing and justifying conclusions; Validating scientific claims; Replicating studies; Documenting studies; Recording scientific workflows and tracking provenance of data and results; Reviewing and Communicating results; Integrating results into the larger body of knowledge within or across disciplines. Hence, accelerating science requires a rich model of the entire scientific process as well as deep knowledge of the scientific area under investigation (Honavar, 2014).

Because science is increasingly a collaborative endeavor, we need: sharable and communicable representations and processes, as well as organizational and social structures and processes, that facilitate collaborative science, including mechanisms for sharing data, experimental protocols, analysis tools, data and knowledge representations, abstractions, and visualizations, tasks, mental models, scientific workflows, mechanisms for decomposing tasks, assigning tasks, integrating results, incentivizing participants, and engaging large numbers of participants with varying levels of expertise and ability in the scientific process through citizen science (Gill and Hirsh, 2012; Bonney et al., 2014).

In conclusion, meeting these challenges requires sustained investment in basic research while proactively integrating these visions into current NLM / NIH programs. That way we can hopefully address the needed research culture changes required for these efforts to have sustained impact. The CCC is available to partner with NLM to make these promising new directions for open science and research reproducibility a success. See the CCC's *Accelerating Science: A Computing Research Agenda* white paper (https://cra.org/ccc/wp-content/uploads/sites/2/2016/02/Accelerating-Science-Whitepaper-CCC-Final2.pdf) for more information.

This material is based upon work supported by the National Science Foundation under Grant No. 1136993. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., and J.T. Eppig, J.T. (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics, 25:25–29

Bernot G, Comet JP, Richard A, Guespin J. (2004). Application of formal methods to biological regulatory networks: extending Thomas' asynchronous logical approach with temporal logic. Journal of theoretical biology. 229(3): 339-47.

Bonney, R., Shirk, J.L., Phillips, T.B., Wiggins, A., Ballard, H.L., Miller-Rushing, A.J. and Parrish, J.K., (2014). Next steps for citizen science. Science, 343(6178), pp.1436-1437.

Danos V, Laneve C. (2004). Formal molecular biology. Theoretical Computer Science. 325(1): 69-110.

Engelbart, D.C., 2001. Augmenting human intellect: a conceptual framework (1962). PACKER, Randall and JORDAN, Ken. Multimedia. From Wagner to Virtual Reality. New York: WW Norton & Company, pp.64-90.

Fisher J, Henzinger TA. (2007). Executable cell biology. Nature biotechnology. 25(11): 1239-49

Gil, Y. and Hirsh, Y. (2012). Discovery Informatics: AI Opportunities in Scientific Discovery. AAAI Technical Report FS-12-03.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. Nature, 521:452- 459.

Honavar, V. (2014). The promise and potential of big data: A case for discovery informatics. Review of Policy Research, 31(4), pp.326-330

Jordan, M.I. and Mitchell, T.M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349:255-260.

Pazzani, M.J., Mani, S. and Shankle, W.R., 1997. Beyond Concise and Colorful: Learning Intelligible Rules. In KDD (Vol. 97, pp. 235-238).

Priami C. (2009). Algorithmic systems biology. Communications of the ACM. 52(5): 80-8

Salomon, G., Perkins, D.N. and Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. Educational researcher, 20(3), pp.2-9