COALITION FOR ACADEMIC SCIENTIFIC COMPUTATION

For more than 30 years, the Coalition for Academic Scientific Computation (CASC) has been dedicated to advocating for the use of the most advanced computing technology to accelerate scientific discovery for national competitiveness, global security, and economic success, as well as to developing a diverse and well-prepared 21st-century workforce. Founded in 1989, CASC is an educational nonprofit 501(c)(3) organization with nearly 100 member institutions, representing many of the nation's most forward-thinking universities and research institutions.

CASC appreciates the opportunity to respond, and commends the committee on the excellent work thus far. CASC would add the following comments to the discussion, in response to each question posed.

1. ***What areas of research or topics of the 2016 Federal Big Data Research and Development Strategic Plan should continue to be a priority for federally funded research and require continued Federal R&D investments? What areas of research or topics of the plan no longer need to be prioritized for federally funded research?***

RESPONSE:
The seven strategies listed in the report remain effective priorities, though the tactics to address them have evolved over the years since the plan was released. In particular, we would stress the increasingly important role AI/ML plays in addressing Big Data problems, and suggest any revision pay more attention to this critical topic. A few comments on the individual strategies:

Strategy 1: Create next-generation capabilities by leveraging emerging Big Data foundations, techniques, and technologies.

Strategy 2: Support R&D to explore and understand trustworthiness of data and resulting knowledge, make better decisions, enable breakthrough discoveries, and take confident action.

Changes to 1 & 2: Enhance the role of AI/Deep Learning in both the R&D capabilities around Big Data.

Strategy 3: Build and enhance research cyberinfrastructure that enables Big Data innovation in support of high-level federal agency missions and policy considerations.

Changes to 3: The role of data acquisition through instruments, and the role of edge computing or "Internet of Things" devices deserves particular emphasis within the supported cyberinfrastructure.

Strategy 4: Increase the value of data through policies that promote sharing and management of data.

Strategy 5: Understand Big Data collection, sharing, and use with regard to privacy, security, and ethics.

Changes to 4 & 5: Note the increased role Controlled Unclassified Information (CUI) plays in federally supported research – and the increased cost of compliance and resulting impediments to sharing and value of data these controls may imply.

Strategy 6: Improve the national landscape for Big Data education and training to fulfill increasing demand for both deep analytical talent and analytical capacity for the broader workforce.

Changes to 6: Continued investment in this topic is the top priority of the CASC Membership.

Strategy 7: Create and enhance connections in the national Big Data innovation ecosystem.

Changes to 7: The Big Data Hub program at NSF has been somewhat under-resourced, and in particular funding for the "spokes" to connect to the hub has never materialized in a meaningful way. Consider reformulating this program.

2. What challenges or objectives not included in the 2016 Federal Big Data Research and Development Strategic Plan should be strategic priorities for federally funded Big Data R&D? Discuss what new capabilities would be desired, what objectives should guide such research, and why those capabilities and objectives should be strategic priorities?

RESPONSE:
Information theory and data architectures have evolved from the time of the last strategic plan, giving rise to new opportunities that call for cross-agency investments to maintain American competitiveness.[1]

Challenge 1: A national data storage CI that removes the conflict between data storage requirements versus data storage service funding options available at research institutions.
Challenge 2: AI reproducibility resources across different infrastructures for researchers to perform/confirm reproducibility studies to ensure the veracity of AI/ML-enabled research.[2]
Challenge 3: New research on methods and tools to reduce data burden, for example data compression.

Opportunity 1: Research and application of FAIR Digital Objects (FDO) to enable the new data-centric ecosystem[3].
Opportunity 2: Establishment of a unified data commons with streamlined resources for ML processing to enable Data-centric AI[4]

[1] https://www.science.org/doi/10.1126/science.abo5947
[2] Gundersen, O. E., Coakley, K., & Kirkpatrick, C. (2022). Sources of Irreproducibility in Machine Learning: A Review. *arXiv preprint arXiv:2204.07610*.
[3] Schultes, E., & Wittenburg, P. (2019). FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. In Y. Manolopoulos & S. Stupnikov (Eds.), *Data Analytics and Management in Data Intensive Domains* (Vol. 1003, pp. 3–16). Springer International Publishing. https://doi.org/10.1007/978-3-030-23584-0_1
[4] https://cacm.acm.org/opinion/interviews/258819-andrew-ng-calls-for-smart-sized-data-centric-solutions-to-big-issues/fulltext

3. What are emerging and future scientific and technical challenges and opportunities that are central to enabling extraction of knowledge and insight from Big Data across the data lifecycle (including capabilities for collection, storage, access, analysis, and reuse of Big Data)? Which of the challenges and opportunities are still appropriate for Federal research funding?

RESPONSE:

Challenge 1: Analysis tools and strategies are needed to leverage digital scientific artifacts beyond data that are discipline agnostic[5].

Challenge 2: There is a need for research on improved methods for discovery of data, including the development of guidance and requirements for interchangeable data formats, shared vocabularies, metadata standards, collaboratively developed ontologies, semantics preserving data integration approaches, and machine learning and inference methods.

Challenge 3: Making responsible, principled, actionable calls for the development and dissemination of data infrastructure, tools, and best practices to ensure that ethical considerations have to be an integral part of all of the steps of the data lifecycle, including data collection, data preprocessing, predictive modeling, model evaluation, and model deployment.

4. What are appropriate models for partnerships among government, academia, and industry in Big Data, and how can these partnerships be effectively leveraged to enhance innovation in Big Data R&D?

RESPONSE:

Challenge 1: Partnerships must include governance of Big Data and other digital scientific artifacts (analyses tools, analytic workflows, etc.) toward the following:

- What fraction of data to store and how long
- What metadata to store
- Templates for data use agreements
- Funding models for sustainably managing and storing the data and metadata associated with Big Data projects

Challenge 2: The principles of FAIR data use[6] and the notion of a Data Commons have become prevalent and should be explicitly incorporated into the revised strategic plan.

Challenge 3: In developing partnership guidelines, it is worthwhile to consider examples of partnership, such as MLCommons Consortia[7] and Google's Data Commons.

---

[5] Honavar, V.G., Hill, M.D. and Yelick, K., 2016. Accelerating science: A computing research agenda. A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. http://cra.org/ccc/resources/ccc-led-whitepapers/ *arXiv preprint arXiv:1604.02006*.

[6] https://doi.org/10.1038/sdata.2016.18

[7] https://mlcommons.org/en/

5. How do we nurture, develop, and enhance a diverse, inclusive, and sustainable workforce of cyberinfrastructure professionals and practitioners for Big Data R&D? What are some effective ways to broaden participation in Big Data R&D?

RESPONSE:

The ubiquitous availability of digital data in all aspects of our lives has opened up tremendous possibilities to develop a variety of careers and professional paths that are relevant to Big Data R&D:

Strategy 1: The workforce needed to address the needs of Big Data R&D requires education and training at all levels covering high schools, community and technical colleges and undergraduate and graduate programs.

Strategy 2: In addition to training of data scientists and engineers, integrating data science skills across all disciplines by making it part of general education requirements is needed, followed by discipline-focused "translational" data science courses.

Strategy 3: Big Data R&D needs a workforce with a wide range of skill levels. For example, we need a large workforce, such as data stewards, to help with data collection at the micro level, manual data cleaning and harmonization, making data at various regional levels available in a FAIR manner. With universal availability of high-speed Internet, broadening participation in regions that currently lack Big Data R&D jobs[8] is possible. A percentage of a project should be devoted data management personnel[9].

[8] https://transportation.house.gov/committee-activity/issue/infrastructure-investment-and-jobs-act#:~:text=The%20Infrastructure%20Investment%20and%20Jobs,create%20good%2Dpaying%20union%20jobs

[9] Mons, B. 2020, February 1. Invest 5% of research funds in ensuring data are reusable. *Nature*, 578: 491. DOI: https://doi.org/10.1038/d41586-020-00505-7