



NATIONAL SCIENCE FOUNDATION
4201 WILSON BOULEVARD
ARLINGTON, VIRGINIA 22230

NSF 17-036

Dear Colleague Letter: Data-Driven Discovery Science in Chemistry (D3SC)

Dear Colleagues:

The amount and variety of data generated in the chemical sciences, and the rate at which it is being produced, are rapidly increasing, so there is a need for corresponding growth in our ability to extract useful insight from interrelated sources. A similar need is recognized across the National Science Foundation (NSF). One example is the "*Harnessing the Data Revolution*" component in the recently-released document, *10 Big Ideas for Future NSF Investment*, which sets the goal of developing "a cohesive, national-scale approach to research data infrastructure and a 21st-century workforce capable of working effectively with data".¹ This creates an opportunity to enable the chemistry community to effectively share, mine, and repurpose its rapidly-growing chemical datasets and to apply state-of-the-art data analytics tools to expand chemical understanding.

Through this Dear Colleague Letter (DCL), the Division of Chemistry (CHE) invites submission of requests for supplements and EAGER (EARly-concept Grants for Exploratory Research) and RAISE (Research Advanced by Interdisciplinary Science and Engineering)² proposals that seek to capitalize on the data revolution. Successful proposals will emphasize *what new information can be obtained from better utilization of data (including data from multiple laboratories, techniques, and/or chemical systems), and how this can lead to new research directions*. Proposals that foster and strengthen interactions among chemists — particularly experimentalists and data scientists — to advance research goals, are strongly encouraged. Examples of possible projects include (but are not limited to) using tools of data visualization, data mining, machine learning, or other data analytics to:

- Accelerate the discovery of more efficient or selective catalysts;
- Advance the predictive design of new chemical species and/or synthetic reactions;
- Forecast synthetic conditions and predict structure/property relations based on existing chemical datasets;
- Enable real-time chemical data collection and processing for rapid identification and correlation of key events during chemical measurements;
- Identify novel ways of sharing and utilizing chemical data derived from multiple instruments, datatypes, and locations; and
- Develop innovative approaches for integrating, correlating, and analyzing chemical simulation or measurement data to provide new chemical insights.

The most competitive proposals will address how the project conceptually advances chemistry through data-enabled discovery science. Consideration of error and uncertainty analysis, recording and storing of appropriate metadata, and routes to determine the robustness and reliability of data are encouraged. Note that the construction or maintenance of large-scale databases *per se* is not the focus of this DCL, although such databases may be required as a means to the endpoint of using the data to provide insights and predictions. Proposals focused on developing cheminformatics for biomedical or materials research applications are outside the scope of this DCL. Proposals whose primary focus is on the development of general-purpose data mining or analysis algorithms not aimed at addressing a specific chemical question are more appropriate for programs supporting general tool development.³

One avenue of support will be through supplements to existing grants. Supplemental funding requests must enhance existing projects by incorporating or exploring the concepts described in this DCL. The upper limit of a supplement request in response to this DCL is \$60,000 for a maximum of twelve months.

Other mechanisms for support of work in discovery science are through the submission of EAGER⁴ and RAISE⁵ proposals. EAGER supports exploratory work in its early stages on untested, but potentially transformative, research ideas or approaches. The proposed work should be "high risk-high payoff". RAISE may also be appropriate if the proposed activities are interdisciplinary and promise transformational advances.

In all cases, Principal Investigators (PIs) are strongly encouraged to contact the cognizant program officers⁶ prior to submission to determine the appropriateness of the work for consideration. The proposal title must begin with "D3SC:". Each D3SC proposal is expected to describe how the proposed activity will lead to better utilization of existing chemistry datasets. For EAGER and RAISE proposals, the title of the proposal should have "EAGER:" or "RAISE:" specified, following the "D3SC:" designation. The PIs submitting EAGER or RAISE proposals should consider the adaptiveness and scalability as well as the broader relevance of the proposed activities to other areas of chemical research. Proposals including international collaboration are encouraged when those efforts enhance the merit of the proposed work. NSF typically supports the costs of the U.S. team and foreign partners are typically supported by their own funding agencies.

D3SC proposals and supplemental funding requests can be submitted at any time but are encouraged by **March 1, 2017, 5:00 pm, submitter's local time**, in order to ensure timely consideration. For proposals submitted on or after January 30, 2017, the general proposal guidelines in the revised Proposal & Award Policies & Procedures Guide (PAPPG) 17-1⁷ as well as those outlined in this DCL apply. Normal review guidelines for supplement, EAGER, and RAISE requests apply.

We are excited by the opportunities in the D3SC area and look forward to working with the chemistry community to develop new approaches to gain insights from existing data, as well as new experimental and theoretical results. For general questions about this DCL, email the cognizant Program Officers in CHE at ChemData@nsf.gov.

References

1. "Harnessing Data for 21st Century Science and Engineering" in 10 Big Ideas for Future NSF Investments: https://www.nsf.gov/about/congress/reports/nsf_big_ideas.pdf.
2. RAISE proposals can only be submitted after January 30, 2017, after the revised PAPPG becomes effective.
3. See solicitations for Critical Techniques, Technologies and Methodologies for Advancing Foundation and Application of Big Data Sciences and Engineering (BIGDATA, https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767), Data Infrastructure Building Blocks (DIBBS, https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504776), and Computational and Data-Enabled Science and Engineering (CDS&E, https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504813).
4. EARly-concept Grants for Exploratory Research (EAGER, https://www.nsf.gov/pubs/policydocs/pappg17_1/pappg_2.jsp#IIE2).
5. Research Advanced by Interdisciplinary Science and Engineering (RAISE, https://www.nsf.gov/pubs/policydocs/pappg17_1/pappg_2.jsp#IIE3).
6. D3SC cognizant Program Officers: Lin He (lhe@nsf.gov), David Rockcliffe (drockcli@nsf.gov), Susan Atlas (satlas@nsf.gov), and Robert Cave (rjcave@nsf.gov).
7. Proposal & Award Policies & Procedures Guide (PAPPG), January 2017: https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17001.