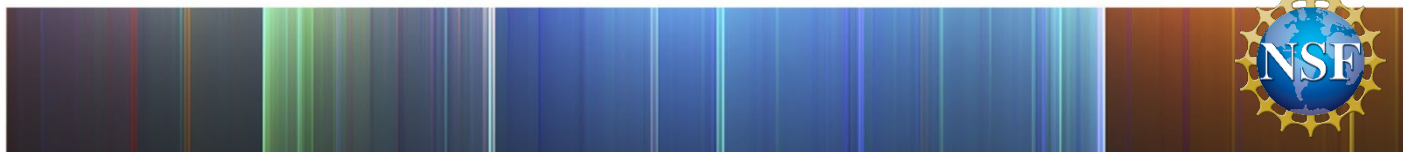
# CHE/OAC
# Joint Office Hour

Dear Colleague Letter: Pilot Projects to Integrate Existing Data and Data-Focused Cyberinfrastructure to Enable Community-level Discovery Pathways

(NSF20-085, https://www.nsf.gov/pubs/2020/nsf20085/nsf20085.jsp)

June 26, 2020



Please mute your microphone

Submit relevant questions through the chat feature

Slides will be shared in the chem listserv email next week

# NSF Office of Advanced Cyberinfrastructure (OAC)

*Foster a cyberinfrastructure ecosystem to transform science and engineering research… through Research CI and CI research*

People, organizations, and communities

Coordination & User support

Gateways, Hubs, and Services

Data Infrastructure

Software and Workflow Systems

Discover Publish Share Collaborate

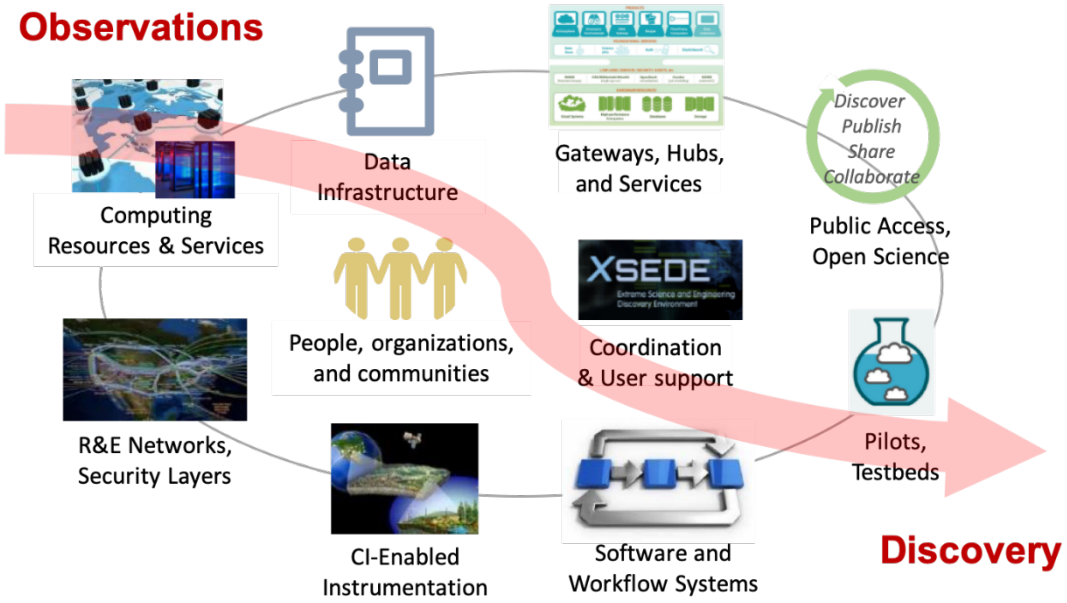
Public Access, Open Science

Pilots, Testbeds

Computing Resources & Services

R&E Networks, Security Layers

CI-Enabled Instrumentation

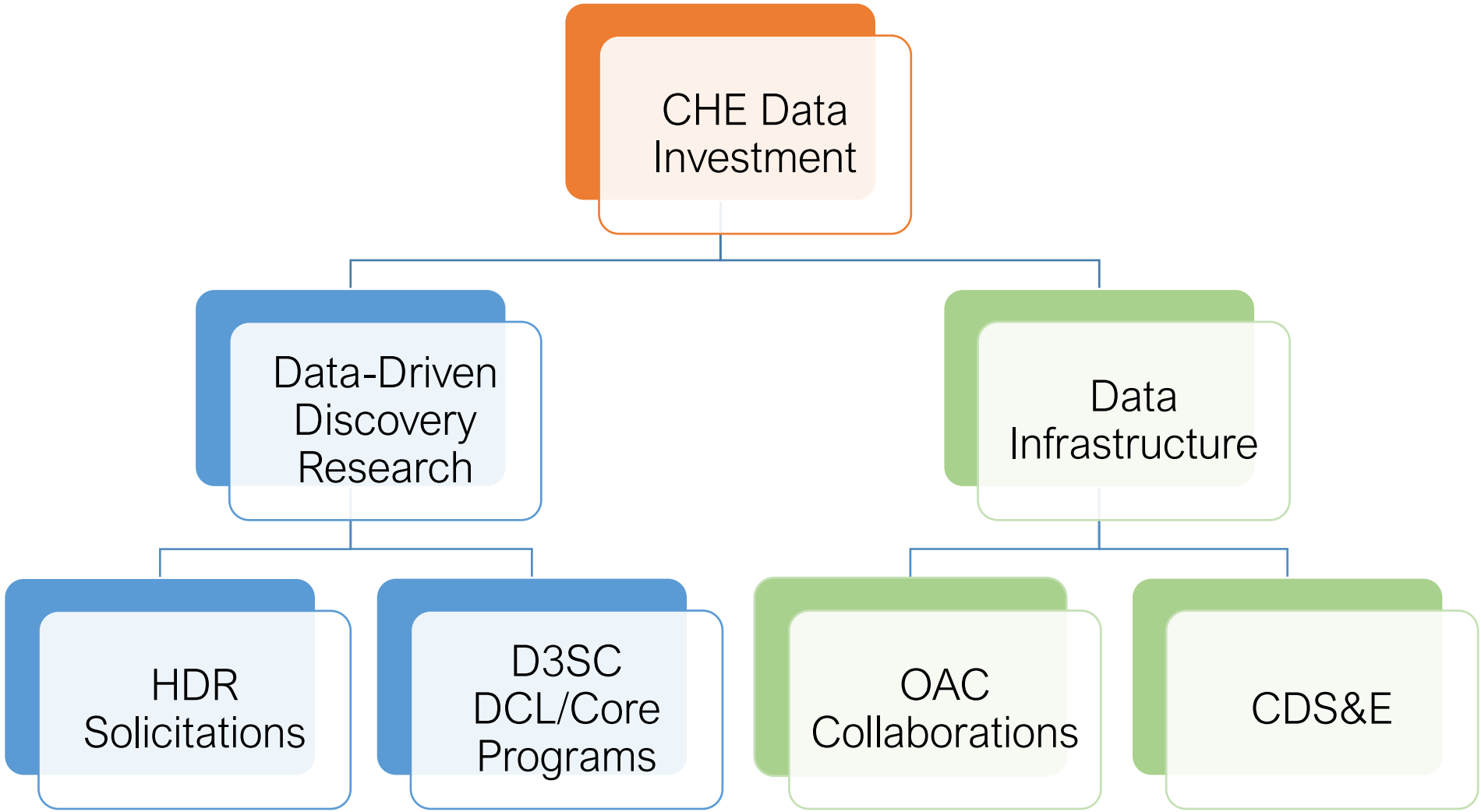



Rapid (disruptive ) changes in S&E and CI landscapes → Our cyberinfrastructure ecosystem must evolve!

NSF's vision for a National Cyberinfrastructure Ecosystem for Science and Engineering in the 21st Century
http://go.usa.gov/xm8bU

# CHE Data Investment: Connecting to Community Needs and Agency Priorities (NSF Data Roadmap)

To support research activities and infrastructure building that seek to capitalize on the data revolution and promote data-driven discoveries that advance fundamental understanding of complex chemical systems.




**Acronym Table:**

- **CDS&E:** Computational and Data-Enabled Science and Engineering
- **D3SC:** Data-Driven Discovery Science in Chemistry
- **DCL:** Dear Colleague Letter
- **HDR:** Harnessing Data Revolution
- **OAC:** Office of Advanced Infrastructure
- **RFI:** Request For Information

# Sought Input from the Community

Workshop

*CHE Data Workshops:*

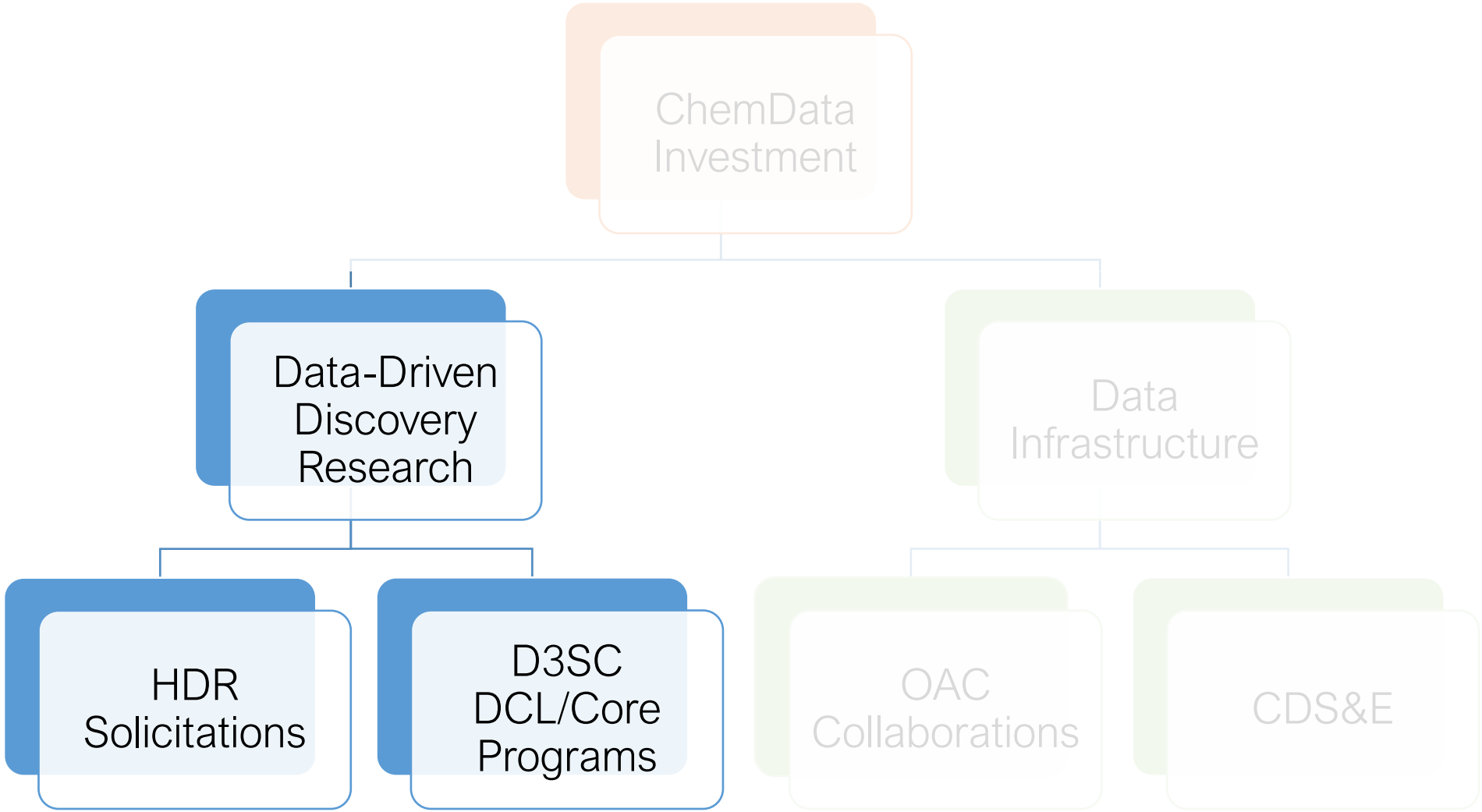
- **NSF Data-Rich Organic Chemistry Workshop,** September 11-12, 2014
- **Mass Spectrometry: Data to Knowledge,** May 11-12, 2015
- **Framing the Role of Big Data and Modern Data Science in Chemistry,** April 18-19, 2017
- **CHE/DMS Innovation Lab: Learning the Power of Data in Chemistry,** December 17-21, 2018

OAC RFI

*OAC Request for Information (RFI) DCLs:*

- **DCL: RFI on Future Needs for Advanced Cyberinfrastructure to Support Science and Engineering Research (NSF CI 2030)** (NSF 17-031)

- **DCL: RFI on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research** (NSF 20-015)




Acronym Table:

- **CDS&E:** Computational and Data-Enabled Science and Engineering
- **D3SC:** Data-Driven Discovery Science in Chemistry
- **DCL:** Dear Colleague Letter
- **HDR:** Harnessing Data Revolution
- **OAC:** Office of Advanced Infrastructure

# Data-Driven Discovery Research in Chemistry

## HDR Solicitations

*Harnessing the Data Revolution (HDR) Big Idea*:

- The foundation of data science;
- Algorithms and systems for data science;
- **Data-intensive science and engineering (NSF 19-543, 19-549);**
- Data cyberinfrastructure; and
- Education and workforce development

**Recent Awards:**

Collaborative Research: Atomic Level Structural Dynamics in Catalysts, NSF-1940263/Crozier
DELTA: Descriptors of Energy Landscape by Topological Analysis, NSF-1934725/Clark

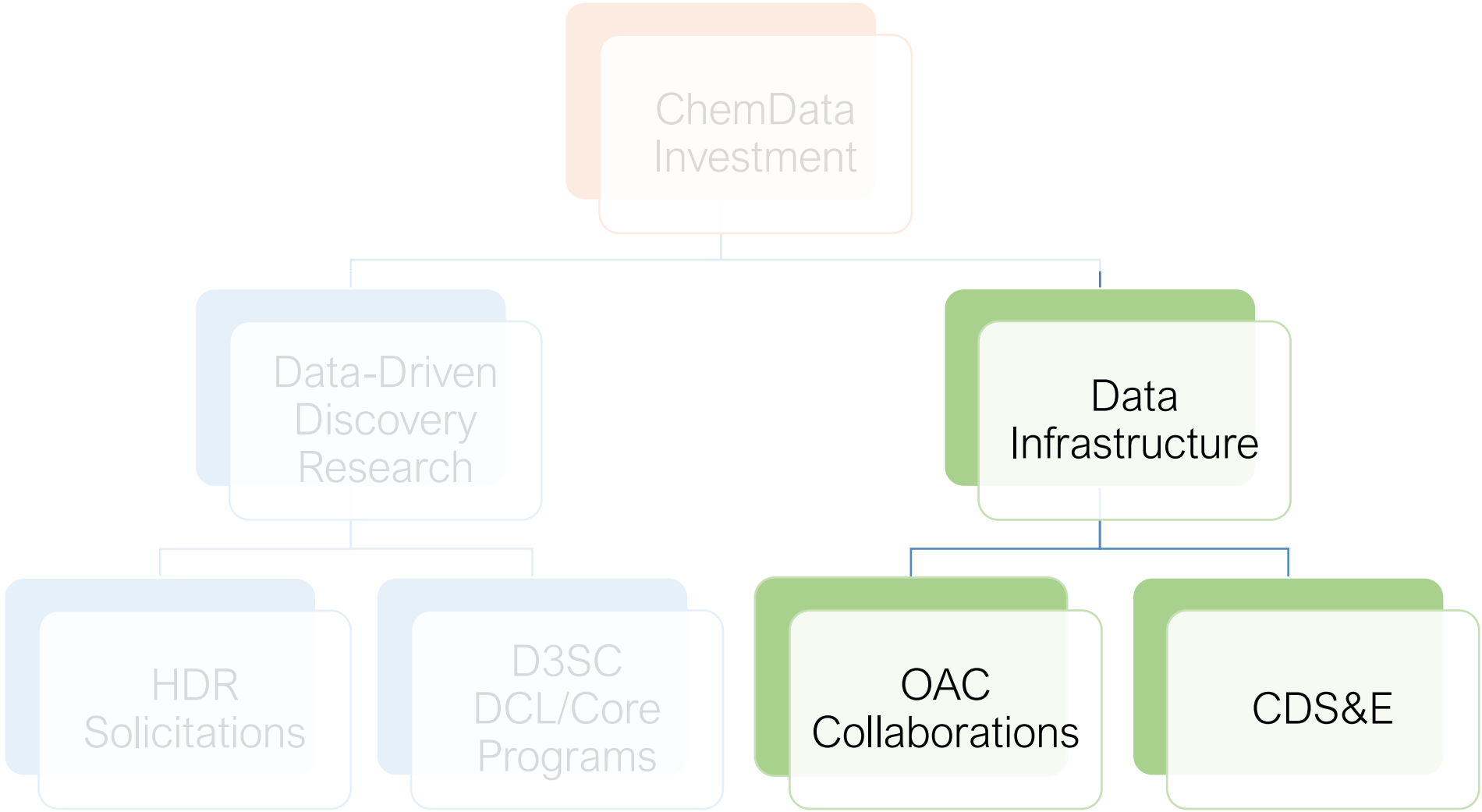
## D3SC DCL/Core Programs

*Dear Colleague Letter: Data-Driven Discovery Science in Chemistry (D3SC)* **(NSF 18-075):**

- Utilize modern data science in the context of chemical research;
- Emphasize new information from better utilization of data and how this can lead to new research directions.

**Recent Awards:**

D3SC: Discovery and Optimization of Chiral Catalysts Guided by Chemoinformatics, NSF-1900617/Denmark
D3SC: EAGER: Collaborative Research: A probabilistic framework for automated force field parameterization from experimental datasets, NSF-1738975/Shirts and NSF-1738979/Chodera




Acronym Table:

- **CDS&E:** Computational and Data-Enabled Science and Engineering
- **D3SC:** Data-Driven Discovery Science in Chemistry
- **DCL:** Dear Colleague Letter
- **HDR:** Harnessing Data Revolution
- **OAC:** Office of Advanced Infrastructure

# Data Infrastructure for Chemistry Research

OAC Collaborations

Infrastructure: Cyberinfrastructure for Sustained Scientific Innovation (CSSI):

- Data Infrastructure Building Blocks (DIBBs) and Software Infrastructure for Sustained Innovation (SI2) programs

| Recent Awards: | S2I2: Impl: The Molecular Sciences Software Institute, NSF-1547580/Crawford<br>CSSI Elements: FLARE infrastructure for reproducible active learning of Bayesian force fields for ex-machina exascale molecular dynamics, NSF-2003725/Kozinsky |
|---|---|

CDS&E

Computational and Data-Enabled Science and Engineering (CDS&E):

- New paradigms in algorithms, software design, and data techniques that impact chemistry research.

| Recent Awards: | CDS&E: Adaptive Learning for Multivariate Calibration with Big Data Attributes, NSF-1904166/ Kalivas<br>CDS&E: Development and application of computational methods: from quantum statistical mechanics calculations to data processing in Fourier transform spectroscopy, NSF-1566334/ Mandelshtam |
|---|---|

# Other Data-related Activities Relevant to Chemists

- National Artificial Intelligence (AI) Research Institutes - seeks to support the development of AI advances and AI-based tools to drive molecular discovery and identify chemical transformation pathways that support energy-efficient, sustainable chemical manufacturing

- Data Science Corps (DSC) - aims to build capacity at the local, state, national, and international levels to help unleash the power of data in the service of science and society

- Transdisciplinary Research in Principles of Data Science (TRIPODS) - supports developing the theoretical foundations of data science through integrated research and training activities;

- NSF Convergence Accelerator: AI-Driven Innovation via Data and Model Sharing – seeks to support multidisciplinary, use-inspired, research projects leading to the development of a Model Commons—for sharing data and data-driven models, for open as well as sensitive data and data-driven models.

# DCL: Pilot Projects to Integrate Existing Data and Data-Focused CI to Enable Community-level Discovery Pathways (NSF20-085)

**DCL Goal:** *bring together researchers and CI experts to develop the means of **combining existing community data resources and shared data-focused CI** into new integrative and highly performing data-intensive discovery workflows that empower new scientific pathways.*

**Sample projects include:**

1. Improving **end-to-end process** of accessing, integrating and transforming research and education data to knowledge and discovery for one or more communities;
2. Creating new **workflows** and new usage modes to address multi-disciplinary and cross-domain scientific objectives;
3. Addressing emerging **community-scale scientific data challenges** such as real-time, streaming and on-demand data access; data discovery; data fusion, integration and interoperability; data privacy concerns;
4. Enhancing the performance and robustness of **community-scale data integration and discovery** workflows such as through automated curation, end-to-end performance monitoring, provenance tracking;
5. **Federating learner data** to empower innovative assessment tools for large-scale modeling of learning gains.

**DCL <u>does NOT</u> support:**

- activities aimed at **conducting the targeted discovery-oriented research** activities themselves including but not limited to creation of datasets or creation or maintenance of databases or repositories; proposers are discouraged from including these activities in their proposals beyond a limited degree of data collection that is necessary to verify the performance of the CI.

- project ideas which **do not align with** the **cross-disciplinary** and **data-integration expectations** of this DCL.

- proposals **previously submitted** to, or otherwise designed for, those other CI programs may not be submitted to the CESER program in response to this DCL.

- **CHE Scope:**
  - to explore the scientific and technical challenges facing big data standardization, storage, dissemination and repurposing in electrochemistry, including electrosynthesis, electrocatalysis, electrochemical sensor and battery development;
  - to facilitate exploration of new research areas (as a pilot project), reuse of legacy data for emerging applications, (re)evaluation of new and previous findings, and comparison of data with future models;

- **Project Budget Range:** $300,000 to $1,500,000

- **POC PD**: CHE: Lin He and Rebecca Peebles; OAC: William Miller and Tevfik Kosar

- **Submit questions to:** CESERQueries@nsf.gov

# A few commonly asked questions

- What is the submission deadline?
- Will my research idea of … fit into this funding call?
- Do I need to have at least one chemist and one data scientist on my team to be eligible for submission?
- Will this count against my proposal submission limit to CHE?
- How will these proposals be reviewed?
- When can I start the project if awarded?
- What should be the next step? – prepare a whitepaper and contact the PDs at CESERQueries@nsf.gov