

# Synthetic Data Generation with Large, Real-World Data

## Request for Solutions

**Lisa Mirel**

Statistical Advisor, National Center for Science and Engineering Statistics (NCSES) within U.S. National Science Foundation (NSF)

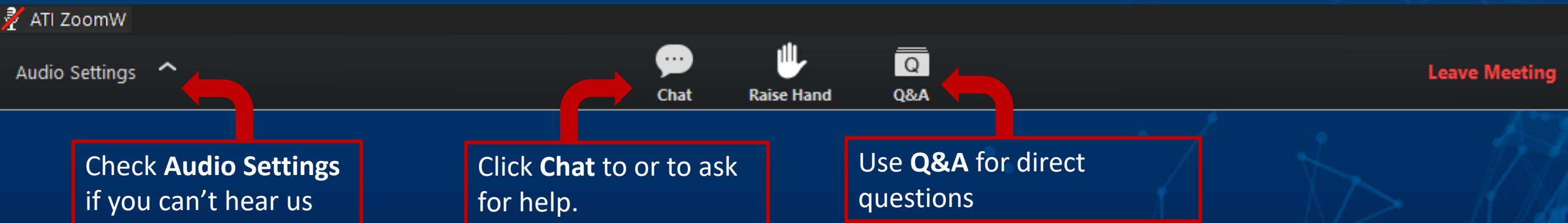


AMERICA'S DATAHUB  
CONSORTIUM



## Housekeeping Items:

- All attendees are on mute and will not be able to unmute themselves.
- Please use the “chat” function for technical difficulties only.
- Place all questions in the Q&A Box.
- Please check your audio settings if you are having difficulties hearing us.



A screenshot of a Zoom meeting toolbar with three red arrows pointing from text boxes below to specific icons. The toolbar includes a microphone icon, the text 'ATI ZoomW', 'Audio Settings' with an upward arrow, 'Chat' with a speech bubble icon, 'Raise Hand' with a hand icon, 'Q&A' with a document icon, and 'Leave Meeting' in red text.

ATI ZoomW

Audio Settings ^

Chat

Raise Hand

Q&A

Leave Meeting

Check **Audio Settings** if you can't hear us

Click **Chat** to or to ask for help.

Use **Q&A** for direct questions



AMERICA'S DATAHUB  
CONSORTIUM



# Background



AMERICA'S DATAHUB  
CONSORTIUM



# America's DataHub Consortium (ADC)

**Vision:** To be an enduring national asset, where eligible people and secure data come together for collaborative research and decision-making that will benefit the American public.

21 projects awarded since 2022



Support cutting-edge data infrastructure



Build data security capabilities to further increase privacy protections and public trust



Develop new ways of acquiring and linking data to yield valuable insights into critical issues



Provide novel and innovative analyses

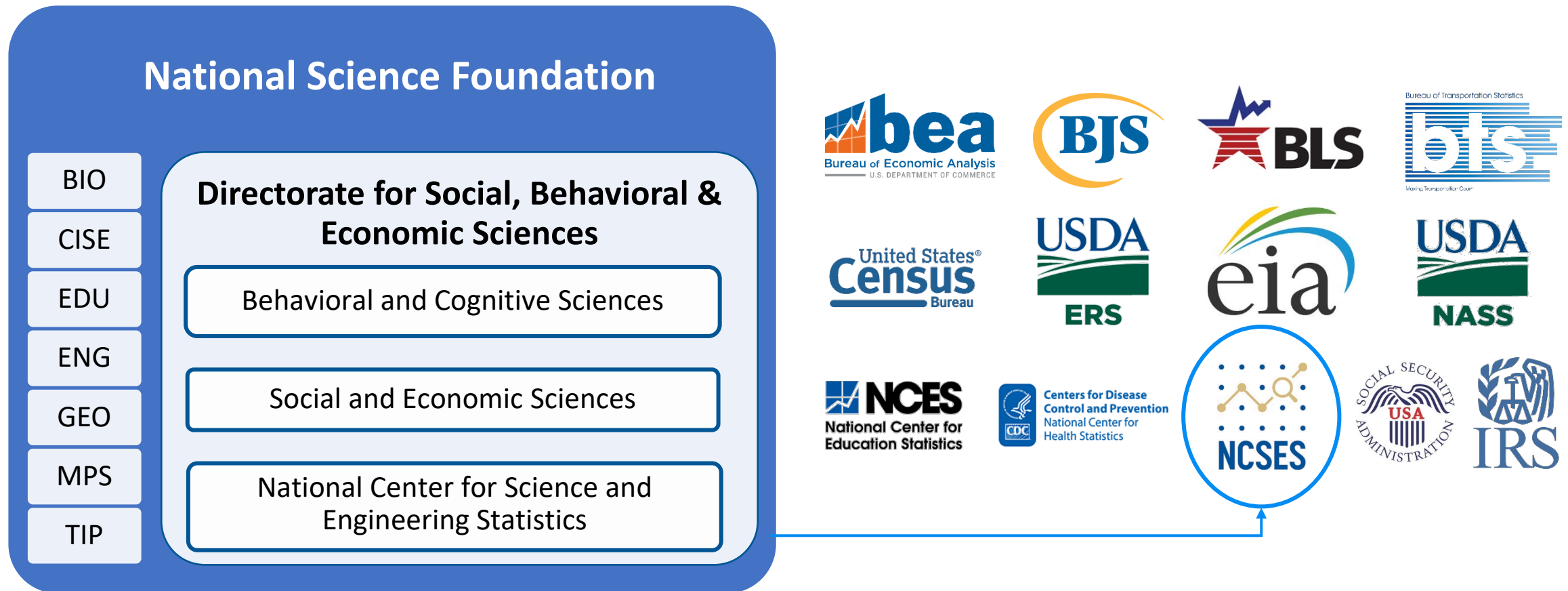


Share lessons learned for similar activities across the Federal government



AMERICA'S DATAHUB  
CONSORTIUM

National Center for Science and Engineering Statistics (NCSES) is one of 13 principal federal statistical agencies and is found within the U.S. National Science Foundation (NSF)

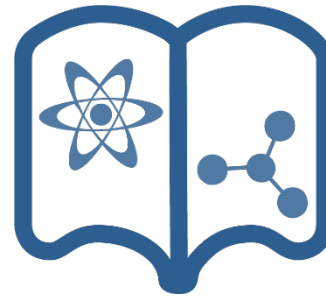


# NCSES's mission is to serve as a federal clearinghouse for objective data that provide key insights into the science and engineering enterprise

---



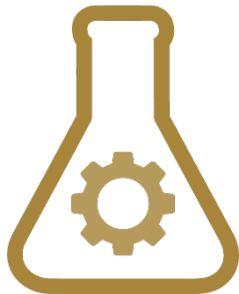
**Science & Engineering  
Workforce**



**STEM Education**



**Innovation & Global  
Competitiveness**



**Research & Development**

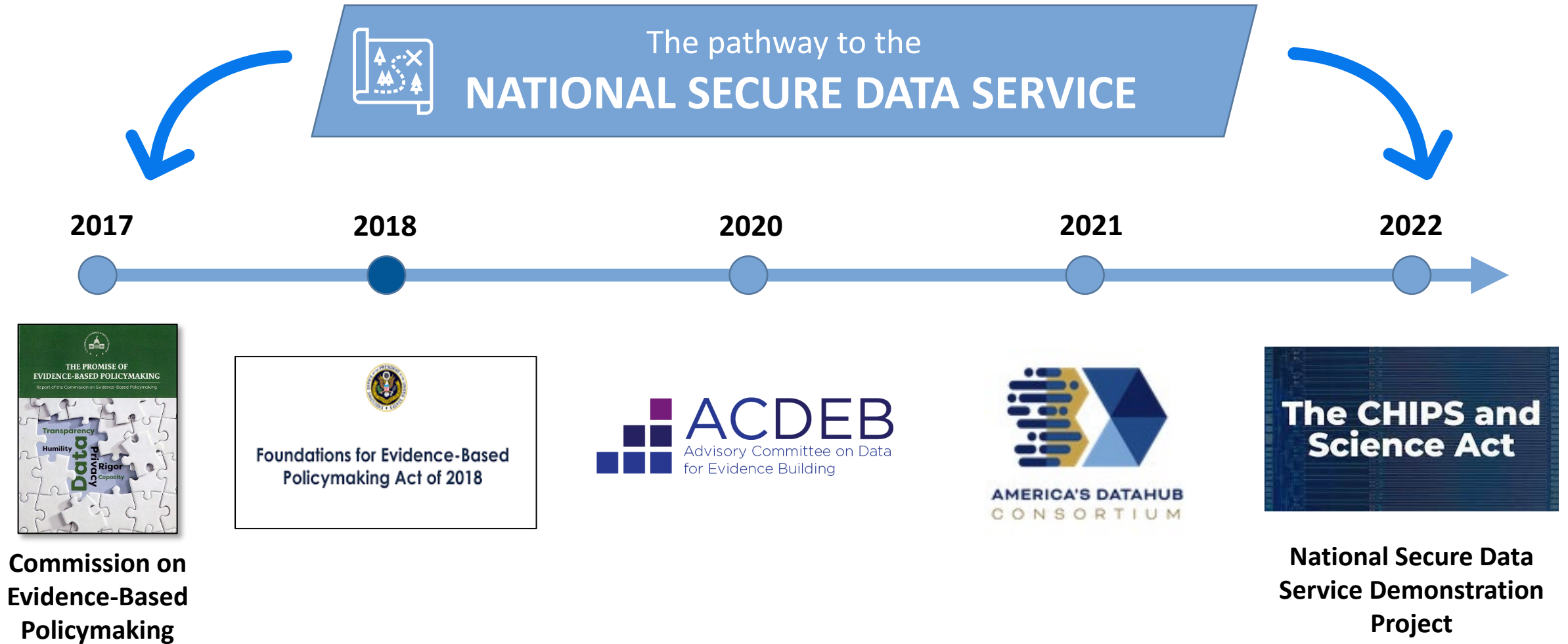


**Government Funding  
for S&E**



**Higher Ed R&D**

# The idea of a transformational National Secure Data Service (NSDS) to coordinate data linking, secure access, and support innovation has a distinguished lineage



# A vision for the NSDS

---



The NSDS is envisioned as a government-wide set of shared services. It serves as a **front door** and a central hub for users to discover shared services and resources and to utilize the NSDS data access and linkage infrastructure.



**Shared services and resources** will include a data concierge service to direct users on their evidence-building journey, toolkits for data protection and synthetic data, communities of practice where users can connect, and platforms to promote transparency concerning the use of government data.



The **data access and linkage infrastructure** will enable users to securely access, link, and analyze powerful, high value data. The NSDS access infrastructure will complement agency enclaves and the Federal Statistical Research Data Center (FSRDC) network while addressing gaps in coverage. The NSDS will also facilitate secure data linkages in support of distinct, authorized projects.



# Synthetic Data Generation with Large, Real-World Data



AMERICA'S DATAHUB  
CONSORTIUM



# Background

As demand for access to confidential federal data assets increases methods are being explored to maintain utility while protecting privacy

Use of synthetic data can reduce disclosure risk while allowing data users to access microdata for research and other statistical purposes

A synthetic dataset could provide a tiered access option that would allow researchers access to previously inaccessible data. This option could prove valuable to researchers in making maximum use of these data while enabling the government to ensure privacy.

The production of synthetic data can prove challenging and resource intensive especially with large datasets. Mitigation strategies may include the use of a super compute platform.



# Objective

- The objective of this project is to improve understanding of how synthetic data generators work with large real-world data (e.g., datasets with over 30 billion rows of data) and how they can be utilized on a large-scale compute environment
  - Case study data
  - Inform infrastructure and governance for other types of confidential data



# The following steps will be completed for synthetic data generation methods with a large RWD file: (1 of 2)

1. Conduct an initial assessment to determine what variables for a synthetic dataset will be used. This will include outreach to stakeholders and subject matter experts to identify critical variables for inclusion.
2. Decide on an open-source synthetic data generator to be utilized in the creation of the large RWD. The tool will be selected based on its ability to process large RWD and ability to be deployed in a super computing environment.
3. Once produced, evaluate the dataset to assess quality and disclosure risk. Evaluation criteria will include but not be limited to an assessment of the alignment of the synthetic data with the underlying restricted data. In addition, assessments will be made of the quality of estimates produced from the synthetic data (bias, fidelity to true data, and disclosure risk). Verification metrics will need to be developed so that researchers, data users, and other stakeholders can request them on an as needed basis. Note: no estimates based on the restricted data will be shared but rather a metric indicating alignment of the estimates (e.g., distribution comparisons, correlation heat maps, metrics on distance between synthetic and true data).



# The following steps will be completed for synthetic data generation methods with a large RWD file: (2 of 2)

4. Identify use cases for this synthetic data to assist in determining how the synthetic data could be optimally utilized by researchers, data users and other stakeholders.
5. Develop a plan for accessing the synthetic data, verification metrics, and messaging regarding this new synthetic data product.
6. Write a report outlining the considerations that are needed for synthetic data generation with large RWD and a framework of the governance considerations to create and release synthetic data. In addition, the report should address the ethical considerations around data privacy and the limitations of synthetic data.



# Information Gaps

This project will identify:

- bias in synthetic data estimates when compared to the truth data
- disclosure metrics and assessments for synthetic data created
- a framework to inform a synthetic data toolkit that will include guidance and governance for synthetic data generation with a large RWD file



# Key Evidence Building Considerations

**Key focus questions (address one or more) to assess innovation in the following areas: data acquisition, data security, data linking, privacy, and engagement:**

- Which novel techniques for data, privacy, and confidentiality protections can be used while maintaining utility for large RWD?
- Are the resulting synthetic data fit for purpose to support research, evidence building, and/or policy making?
- What mechanisms are needed to access the resulting data that uphold privacy requirements?
- How will the results of this work inform a synthetic data toolkit that will benefit the research community and guide future synthetic data generators?

# Deliverables

At a minimum, offerors will provide the following if selected for an award. Additional deliverables may be required.

- **Monthly status reports** on progress towards project objectives.
- **Monthly or bi-weekly status update meetings** with project team.
- **Quarterly lessons learned** based on the previous quarter to inform a future NSDS and the National Artificial Intelligence Research and Resource (NAIRR) pilot.
- **All code** (clearly documented): documentation of synthetic data methodology, documentation of data quality assessment, and any other documentation created under this award. Data should be made available in an agency designated repository.
- **Documentation of verification metrics** and how they will be used to show alignment with true estimates.
- **A report that outlines the framework** to inform a synthetic data toolkit with guidance for synthetic data generation with a large RWD file. The report will describe the lessons learned through this project, including but not limited to the creation of synthetic data and whether the resulting data and models fit are fit for purpose. In addition, the report will describe how this approach could inform a tiered access model and contribute to a potential NSDS and inform the NAIRR pilot.



# Questions?



AMERICA'S DATAHUB  
CONSORTIUM



# Request for Solutions (RFS) Requirements

**Mandi Ballou**

Sr. Contracts Manager

Advanced Technology International (ATI), ADC Consortium Management Firm (CMF)

[americasdatahub.org/opportunities](https://americasdatahub.org/opportunities)

*The official source of information regarding the solicitations is included in the Request for Solutions posted on the ADC website. If you act on information from other sources, it is at your own risk.*



AMERICA'S DATAHUB  
CONSORTIUM



# RFS Summary

## Project Topic

- Proposals must address the specific topic area.
- See Attachment 1 of the RFS for full topic description

## Project Awards

- It is anticipated there will be one award estimated at \$1,000,000.

## Period of Performance

- 12 Months

## One Step Process

- Offerors will submit a detailed technical and cost proposal for award evaluation.
- You do not need to be an ADC member to respond. However, if you're selected for award, you must join ADC (if not already a member).



# Full Proposal Submission

- RFS Attachment 2 includes format
  - Volume 1: Technical Proposal
    - Limited to **8 pages plus cover page**
  - Volume 2: Cost Proposal
    - No page limit
  - Submit in Word format
  - Submission form: [https://atisc.formstack.com/forms/adc\\_dg\\_rwd\\_rfs](https://atisc.formstack.com/forms/adc_dg_rwd_rfs)



# Full Proposal Cover Page

- Working title of the proposed project
- Names, phone numbers, mailing, and e-mail addresses for the principal technical and contractual POCs
- Unique Entity ID (UEI) of the Offeror (if available)
- Project partners, if any
- Date of submission
- Proprietary data restrictions, if any

# Volume 1: Technical Proposal Content

- **Executive Summary**

- **Summary Statement:** Provide a succinct statement of the aim of the project and proposed approach. In most cases, the summary statement will be no longer than a paragraph.
- **Context:** Briefly describe the current state of information and/or research in the area.
- **Proposed Approach:** Explain how the proposed approach will meet the objectives outlined in Attachment 1, result in or lead to a replicable framework that can be used to address similar issues, and inform other strategic priorities like the National Secure Data Service.



# Volume 1: Technical Proposal Content *(Continued)*

- **Statement of Work**

- **Work Scope:** Describe the work to be accomplished as part of the project, organized as it is expected to be performed. Separate the work effort into major tasks and subtasks as numbered paragraphs or in a table.
- **Deliverables:** All project deliverables should be clearly listed and described.
- **Future Phases:** Proposals may include a discussion of optional, future phases of work. The original phase or work shall in no way depend on work described under future phases to meet the program criteria.

***Do not include company-sensitive or proprietary data included in the Statement of Work***



# Volume 1: Technical Proposal Content *(Continued)*

- **Capabilities and Experience**
  - List all project partners and indicate if they are a non-traditional entity
  - Identify all key personnel and describe their roles; organize the team by organization name
  - Relate the capabilities and experience of key personnel and organizations to the project
  - Identify any supervisory relationships and the main POC check-ins during the project
  - Provide resumes (2-page max) for all key personnel in an appendix (excluded from page limit)

# Volume 1: Technical Proposal Content *(Continued)*

## • Capabilities and Experience *(continued)*

- Designate any graduate students or postdoctoral fellows to be funded by the project
  - If named, provide a biographical sketch (½ page max) of their background and research interests within resume appendix
- Describe unique capabilities that may reduce risk, duration, and/or improve financial performance
- Address any potential conflicts of interest and any proposed mitigation, and complete Exhibit 1 – Organizational Conflicts of Interest Certificate

# Volume 1: Technical Proposal Content *(Continued)*

- **Intellectual Property Rights**

- Identify limitations on any intellectual property (patents, inventions, trade secrets, copyrights, or trademarks) that will impact the performance or the Government's use of any deliverable under the project
- Describe the intellectual property in sufficient detail to describe:
  - Limitations (data assertions, potential patent licenses required by the Government, etc.)
  - Why or how the Government can accomplish the objectives with the proposed limitations



## Volume 2: Cost Proposal Content

- **Agreement Type:** Preference is firm-fixed price; Offeror to identify other preferred agreement type (e.g., cost-plus-fixed-fee) and provide rationale.
  - Agreement type will be subject to concurrence of selected offeror and Government.
- **Cost Estimate:** Account for entire cost of project, broken down for each phase of the proposed work. Contractor format for the cost estimate is acceptable.

## Volume 2: Cost Proposal Content *(Continued)*

- **Labor – Offeror only:** Describe each labor category or person with labor rate and hours.
- **Travel – Offeror only:** List number of trips and number of days, travelers, and costs per trip.
- **Team Members/Subcontractors/Consultants:** List all team member/subcontractor/consultant and associated totals.
- **Material/Equipment – Offeror only:** List all items and provide justification and basis of cost for each (i.e., catalog pricing, vendor quote, previous purchase, etc.).
- **Other Direct Costs – Offeror Only:** List all items and provide justification and basis of cost for each (i.e., catalog pricing, vendor quote, previous purchase, etc.).

## Vol 2: Cost Proposal Content *(Continued)*

- **Indirect Costs – Offeror Only:** Breakout of indirect costs; indicate if indirect rates are Government approved.
  - If approved: cite approval date and federal agency.
  - If not approved: explain how the proposed indirect rates are appropriate for pricing.
- **Profit/Fee:** Indicate any proposed profit/fee.



# Full Proposal Submission Form

## Offeror Information

Offeror Organization \*

Offeror Address

Offeror City

Offeror State

Offeror Zip Code



## Full Proposal Submission Form *(Continued)*

Are you currently a "Non-traditional Entity"? \*

☐ Yes

☐ No

"Non-traditional entity" means an entity (construed in its broadest sense to include qualified large and small businesses, universities, non-profits, philanthropic organizations, partnerships, joint ventures, and other entity forms) that is not currently performing and has not performed, for at least the three-year period preceding the solicitation of sources by NSF for the procurement or arrangement, under any NSF procurement contract or NSF instrument of financial assistance.





# Full Proposal Submission Form *(Continued)*

## POC Information

Technical POC Name\*

First Name

Last Name

Job Title\*

Technical POC Email\*

Technical POC Phone\*

Is the Contracts POC the same as the Technical POC?

☐ Yes

☐ No

# Full Proposal Submission Form *(Continued)*

## Submission

Proposal Title \*

Proposal Submittal

- ☐ I prefer to upload the proposal to this form (attachments will be unencrypted)
- ☐ I prefer to send the proposal via encrypted email to [adc-contracts@ati.org](mailto:adc-contracts@ati.org) (may be done before or after submitting this form)

Volume I: Technical Proposal Upload

No File Chosen

File names must not contain spaces or special characters

Volume II: Cost Proposal Upload

No File Chosen

File names must not contain spaces or special characters

Additional Comments



AMERICA'S DATAHUB  
CONSORTIUM

# Full Proposal Evaluation Criteria

*The criteria are listed in order of relative importance.*

- **Technical**
  - Approach: The degree to which the proposed project:
    - (i) meets the objectives outlined in (RFS) Attachment 1
    - (ii) will result in or lead to a replicable framework that can be used to address similar issues
    - (iii) demonstrates innovation
    - (iv) informs strategic priorities for a future NSDS and NAIRR
  - Teaming: The degree to which the proposed project includes a diverse team of qualified performers to include use of non-traditional entities.



# Full Proposal Evaluation Criteria *(Continued)*

- **Cost**

- The CMF will perform an analysis and will provide the results to the Government; may entail the CMF requesting additional information from the Offeror.
- The Government will determine whether the Offeror's total evaluated cost/price is fair and reasonable.



# Timeline

	Dates
Request for Solutions Release	June 18
Webinar	June 24
Teaming Speed Networking Event ( <i>next slide</i> )	June 25
Full Proposal Deadline	July 16, 3 PM ET
Offeror Notifications	July - August
Award Projects	August - September

*Any deadline updates will be communicated via email.*

# Teaming Resources

- **Teaming Speed Networking Event on June 25 1 PM ET**
- The event aims to help organizations find potential partners for this RFS. Each presenting organization will have a maximum of three minutes to highlight its capabilities and specify if they are looking to serve as a prime contractor, subcontractor, or either.
- Visit the “Events” page to learn more and to register for the event!

<https://www.americasdatahub.org/events/>

# Teaming Resources (continued)

- ADC Member Profile Database\*
  - Searchable by member demographics and capabilities, includes POC info for each member.
- Need a teaming partner outside of ADC or other resources?
  - Email [ati@govmates.com](mailto:ati@govmates.com) with who you are looking for.
  - More information is available at [govmates.com/ati](https://govmates.com/ati).

\*Requires access to Members Only website. Request access [here](#).

*Not a member, but want to access these resources? [Join today](#) — it's free!*



AMERICA'S DATAHUB  
CONSORTIUM

# Stay Engaged



Solicitation and Contract Related Questions: [ADC-Contracts@ati.org](mailto:ADC-Contracts@ati.org)



General/Membership Questions: [adc@ati.org](mailto:adc@ati.org)



Join the ADC Mailing List: <https://www.americasdatahub.org/adc-mailing-request-form/>

## Visit the ADC Website

Stay in the loop with ADC events, solicitations, project awards, news, members, and more!

[www.americasdatahub.org](http://www.americasdatahub.org)



AMERICA'S DATAHUB  
CONSORTIUM



# Questions?



AMERICA'S DATAHUB  
CONSORTIUM

