Scientific Integrity in Using Data for Evidence-Building

Lisa B. Mirel

FCSM Research and Policy Conference, October 23, 2024

NATIONAL CENTER FOR SCIENCE AND ENGINEERING STATISTICS
NATIONAL SCIENCE FOUNDATION

Outline

- Scientific Integrity
- Scientific Integrity in Shared Service Environment
- Case Study
- Lessons Learned
- Final Thoughts

Scientific Integrity

- Use established scientific methods
- Disseminate objective information
- Shield products from inappropriate political influence

Scientific Integrity: Using Data for Evidence Building

- Shared service ecosystem
    - A model that streamlines and innovates data sharing and linking to enable using data for evidence building and decision-making at all levels of government and in all sectors
- One model is the formation of a National Secure Data Service (NSDS)

CHIPS and Science Act Requirements (§10375)

Calls for a 5-year demonstration project to develop, refine, and test models to inform the full implementation of an NSDS.

The NSDS is envisioned as a set of shared services and a government-wide data linkage and access infrastructure to support **evidence building**.

Requires consultation with the director of OMB, the National Artificial Intelligence Research Resource (NAIRR), and alignment with the Advisory Committee on Data for Evidence Building (ACDEB) recommendations.

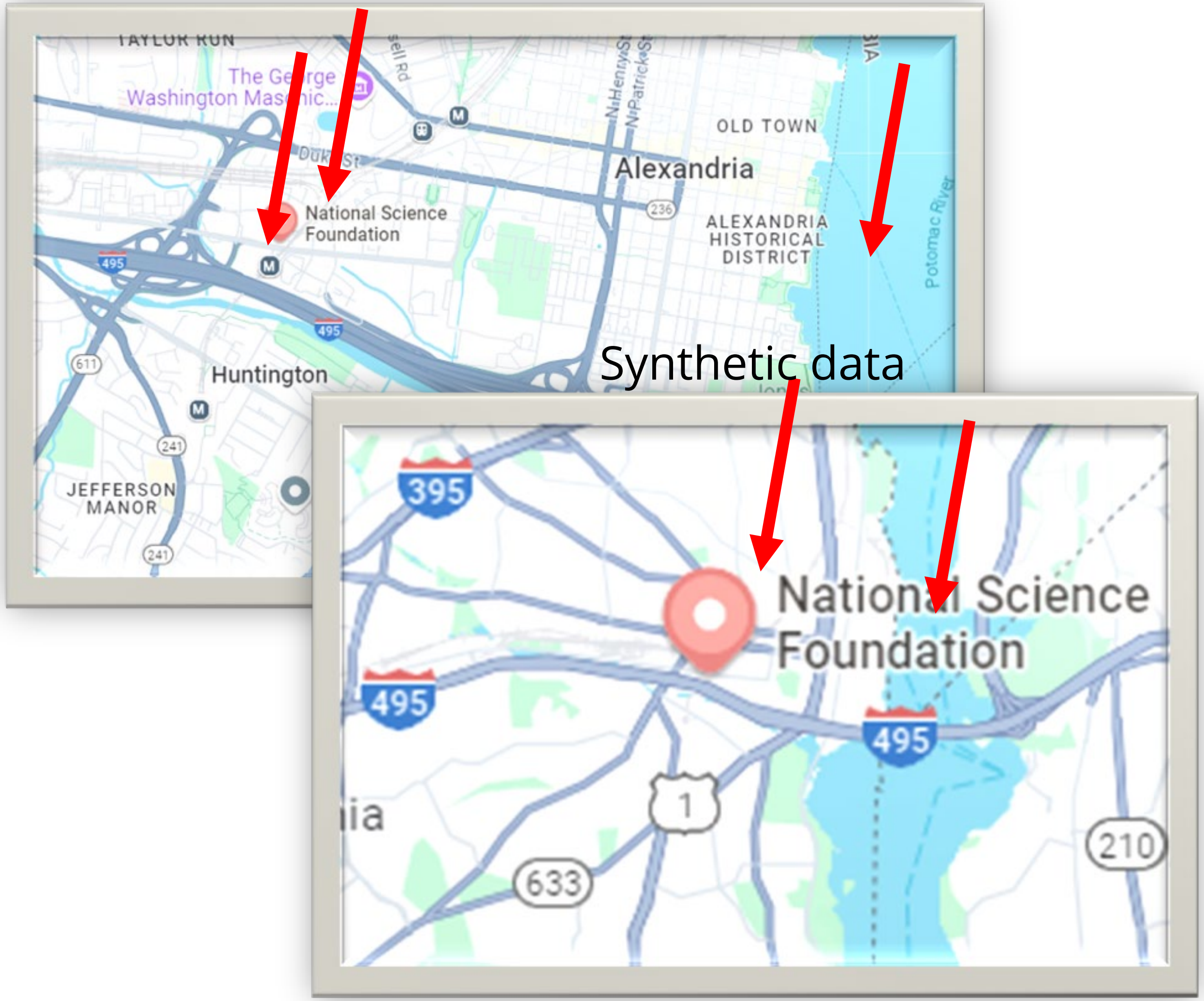
The NSDS Demonstration Project will be implemented by the National Center for Science and Engineering Statistics (NCSES).

Scientific Integrity in a Shared Services Environment

- NCSES and our partners have undertaken a series of projects designed to build foundational systems and processes to inform a future NSDS that incorporate the standards and principles of scientific integrity
  - **Conducting feasibility studies**: concierge services, federated data usage platform
  - **Testing use cases**: generating synthetic data, performing privacy preserving record linkages
  - **Building infrastructure**: establishing website (.Gov), secure compute environment testbed
- Each project documents lessons learned

Creating Synthetic Data Files

- Generated using statistical or machine learning techniques
- Fully synthetic data do not contain the exact records from the original data but are informed by them
- Have similar in properties as the original data in aggregate statistics
- Can potentially be shared and released while balancing utility and privacy concerns


Restricted data

Synthetic data
TAYLOR RUN
The George Washington Masonic...
Duke St
National Science Foundation
Alexandria
OLD TOWN
ALEXANDRIA HISTORICAL DISTRICT
Potomac River
495
611
Huntington
241
JEFFERSON MANOR
236
395
National Science Foundation
495
495
1
633
210

Case Study

- Survey of Earned Doctorates (SED): annual census, conducted since 1958, of all individuals receiving a research doctorate from an accredited U.S. institution in a given academic year
- Results are used to assess characteristics of the doctoral population and trends in doctoral education and degrees
- Several ways to access SED data
  - Data tools
  - Restricted use data access system (RDAS)
  - Restricted use license agreement
- No public use micro data file

SED
Survey of Earned Doctorates

Increasing Access with Fully Synthetic Data

- Pilot innovative methods to create public-use synthetic data, which are based on the true restricted SED data

- Conduct outreach with stakeholders to learn their needs

- Create publicly available synthetic dataset
  - Conduct comparison analyses
  - Establish a verification system

Case Study: Generate Synthetic Data

Objectives:

- Generate synthetic SED data, using open-source code (*scientific integrity: using established scientific methods*)
- Assess fidelity to the truth, document methodology and communicate analytic considerations (*scientific integrity: dissemination and communication of objective information*)
- Create verification metrics (*scientific integrity: shield from inappropriate political influence*)

Synthetic Data: Opportunities and Challenges

Opportunities:

- A tiered access approach expands data access and research potential
- Can be used to address emerging policy questions and develop hypotheses

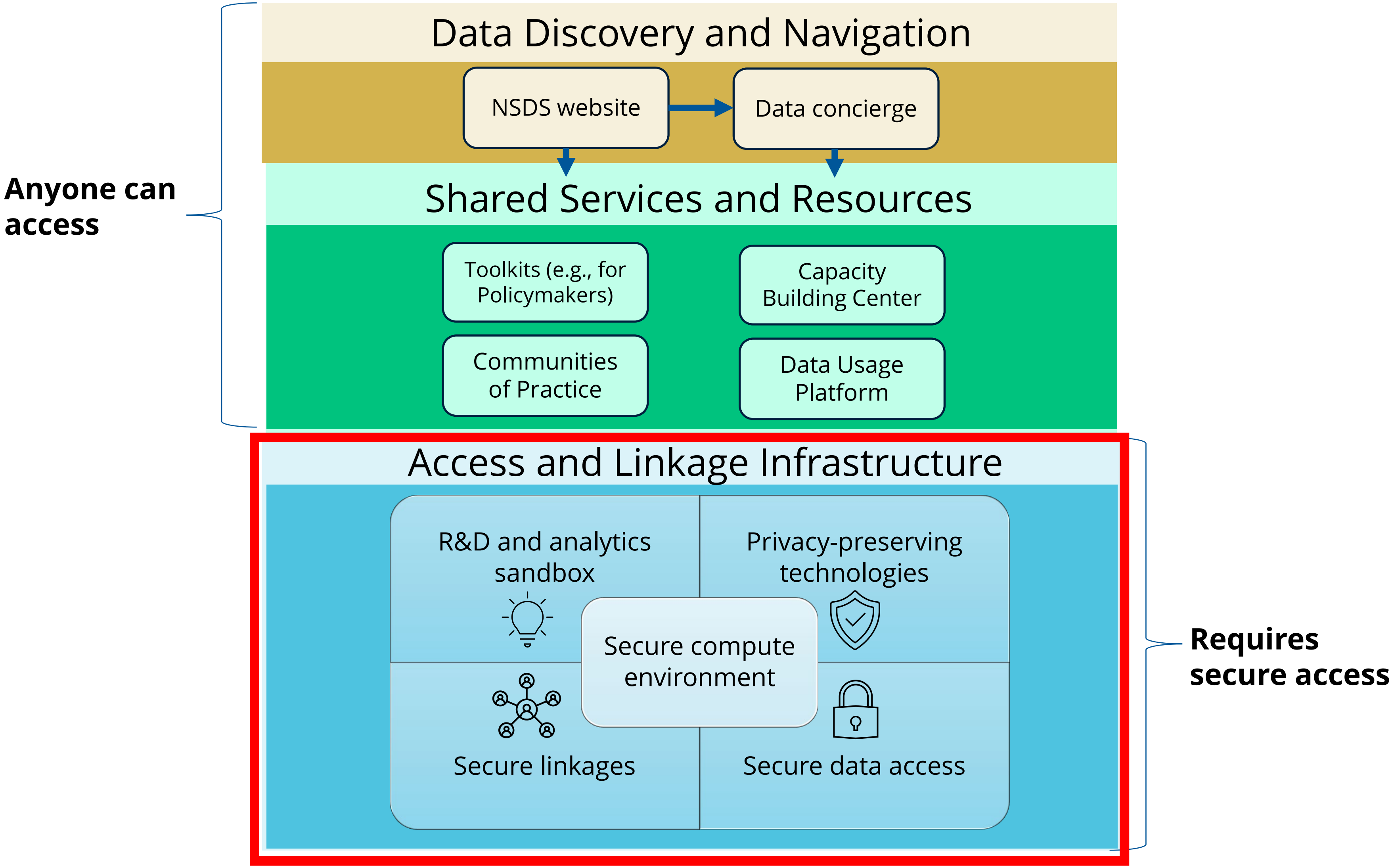
Challenges:

- Ensuring methodologies produce analytically sound estimates
- Communicating the importance of verifying the results
- Continually evaluating and improving a generator to develop new synthetic data files as new data are collected

Lessons Learned to Date: Synthetic Data

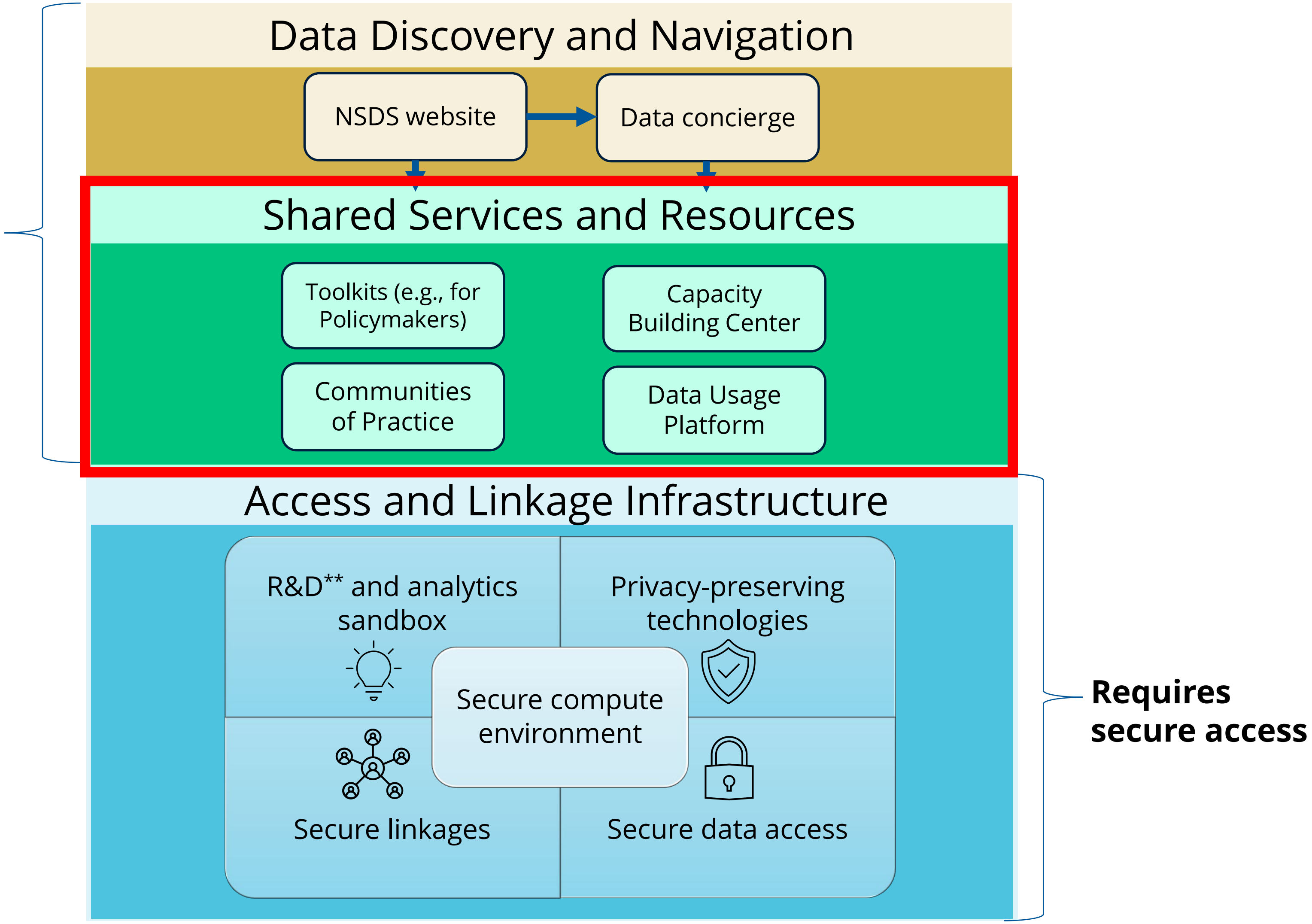
- Communication with key stakeholders to understand needs and uses of synthetic data file is critical
- Important to develop understanding of the processes and technical infrastructure needed to generate synthetic data
- Detailed documentation about methodology and appropriate uses of the synthetic data file is essential to support its use
- Disclosure risk assessment of synthetic data should incorporate multiple sources of available data

Building a Future NSDS

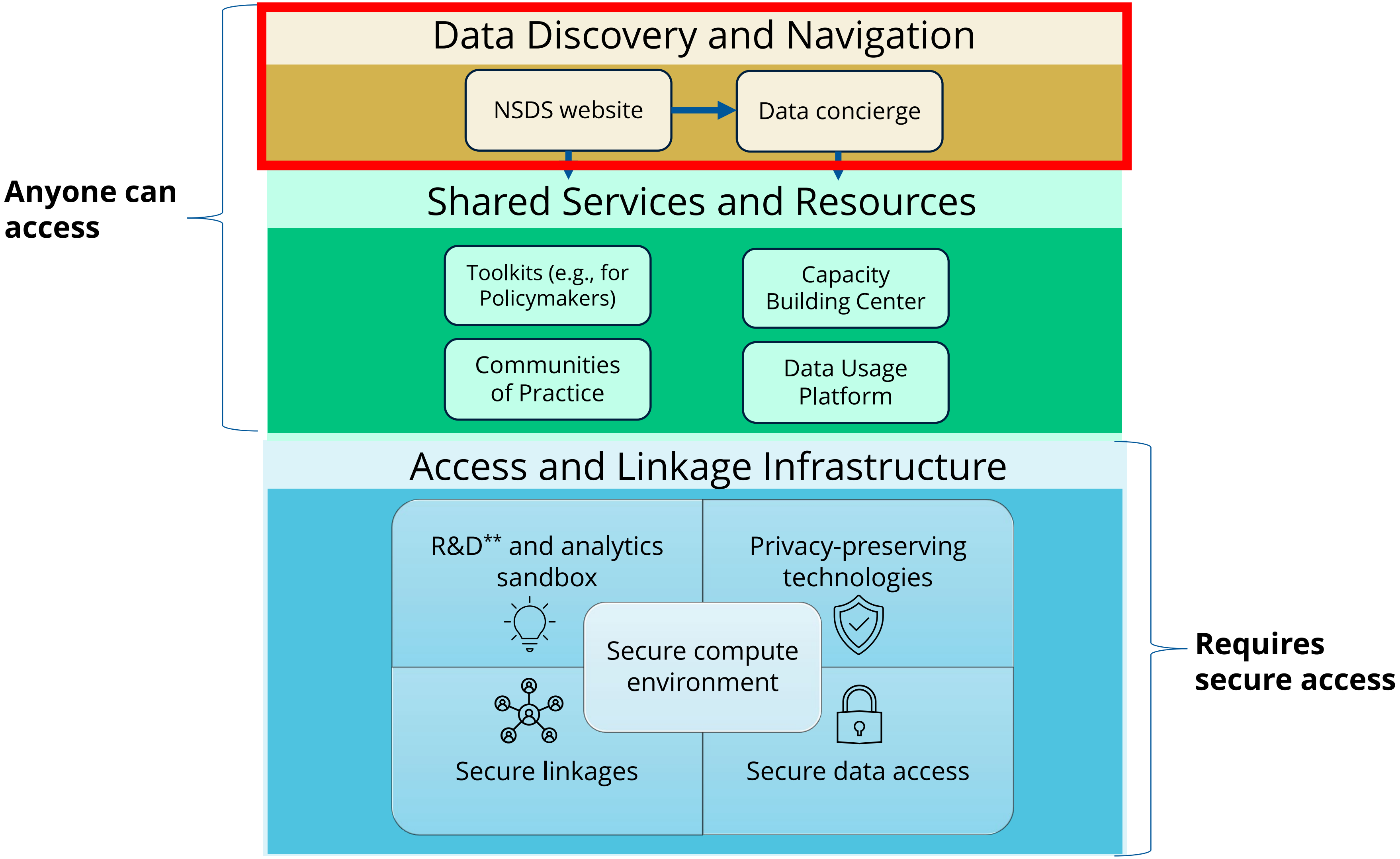


Anyone can access
Data Discovery and Navigation
NSDS website
Data concierge
Shared Services and Resources
Toolkits (e.g., for Policymakers)
Capacity Building Center
Communities of Practice
Data Usage Platform
Access and Linkage Infrastructure
R&D and analytics sandbox
Privacy-preserving technologies
Secure compute environment
Secure linkages
Secure data access
Requires secure access

Building a Future NSDS


Anyone can access
Data Discovery and Navigation
NSDS website
Data concierge
Shared Services and Resources
Toolkits (e.g., for Policymakers)
Capacity Building Center
Communities of Practice
Data Usage Platform
Access and Linkage Infrastructure
R&D** and analytics sandbox
Privacy-preserving technologies
Secure compute environment
Secure linkages
Secure data access
Requires secure access

Building a Future NSDS



Anyone can access

Data Discovery and Navigation

NSDS website

Data concierge

Shared Services and Resources

Toolkits (e.g., for Policymakers)

Capacity Building Center

Communities of Practice

Data Usage Platform

Access and Linkage Infrastructure

R&D** and analytics sandbox

Privacy-preserving technologies

Secure compute environment

Secure linkages

Secure data access

Requires secure access

Final Thoughts

- A successful, future NSDS, that supports using data for evidence building, will support projects that maintain scientific integrity and model ways to ensure the standards and principles are upheld
- Continuous need for additional use cases to further the success of building a shared service ecosystem

Questions?

Lisa B. Mirel

Email address: lbmirel@nsf.gov

https://ncses.nsf.gov

@NCSESgov

Appendix

What is Scientific Integrity?

FCSM Framework for Data Quality:

Scientific integrity refers to an environment that ensures adherence to scientific standards and use of established scientific methods to produce and disseminate objective data products and one that shields these products from inappropriate political influence.

NSF:

Scientific integrity is the adherence to professional practices, ethical behavior and the principles of honesty and objectivity when conducting, managing, using the results of and communicating about science and scientific activities. Inclusivity, transparency and protection from inappropriate influence are hallmarks of scientific integrity.