Agenda

1:30-2:00	**NAIRR Task Force Overview,** Lynne Parker
2:00-2:30	**Discussion: Desired Characteristics for the NAIRR,** Erwin Gianchandani
2:30-3:00	**Lessons Learned from Federal Cloud Pilots Supporting AI R&D**, Erwin Gianchandani
3:00-3:15	**Break**
3:15-4:00	**Cloud Pilot Case Study – STRIDES**, Andrea Norris & Nick Weber
4:00-4:45	**Task Force Proposed Workplan**, Lynne Parker
4:45-5:00	**Questions from Public,** Erwin Gianchandani

National AI Research Resource Task Force

LYNNE PARKER, DIRECTOR, NATIONAL AI INITIATIVE OFFICE,
WHITE HOUSE OFFICE OF SCIENCE AND TECHNOLOGY POLICY

NAIRR Task Force Mandate

Objective: to investigate the feasibility and advisability of establishing and sustaining a National Artificial Intelligence Research Resource; and to propose a roadmap detailing how such resource should be established and sustained.

Membership: The Task Force is composed of 12 members selected by the co-chairpersons of the Task Force from among technical experts in artificial intelligence or related subjects, of whom—

- 4 are representatives from government
- 4 are representatives from institutions of higher education
- 4 are representatives from private organizations

What is a National AI Research Resource?

Vision: A shared computing and data infrastructure that would provide AI researchers and students across scientific fields with access to a holistic advanced computing ecosystem. This would include:

- Secure, high-performance, privacy-preserving computing frameworks;
- High-quality, representative datasets; and
- Appropriate educational tools and user support mechanisms.

Why: democratize access to the cyberinfrastructure that fuels AI research and development, enabling all of America's diverse AI researchers to participate in exploring innovative ideas for advancing AI, including communities, institutions, and regions that have been traditionally underserved.

Implementation Plan

Congress directed the task force to include the following in it's implementation plan:

i. Appropriate agency or organization responsible for implementation and administration
ii. A governance structure.
iii. Capabilities required to create and maintain a shared computing infrastructure to facilitate access to computing resources for researchers across the country, including scalability, secured access control, resident data engineering and curation expertise, provision of curated data sets, compute resources, educational tools and services, and a user interface portal.
iv. An assessment of, and recommended solutions to, barriers to the dissemination and use of high-quality government data sets
v. An assessment of security requirements and a recommendation for a framework for the management of access controls.
vi. An assessment of privacy and civil rights and civil liberties requirements
vii. A plan for sustaining the Resource, including through Federal funding and partnerships with the private sector.
viii. Parameters for the establishment and sustainment, including agency roles and responsibilities and milestones to implement the Resource

External Consultations

(1) The National Science Foundation.
(2) The Office of Science and Technology Policy.
(3) The National Academies of Sciences, Engineering, and Medicine.
(4) The National Institute of Standards and Technology.
(5) The Director of National Intelligence.
(6) The Department of Energy.
(7) The Department of Defense.
(8) The General Services Administration.
(9) The Department of Justice.
(10) The Department of Homeland Security.
(11) The Department of Health and Human Services.
(12) Private industry.
(13) Institutions of higher education.
(14) Civil and disabilities rights organizations.
(15) Such other persons as the Task Force considers appropriate

Request for Information on the NAIRR Implementation Plan: *Public Input*

- OSTP and NSF released this RFI on July 23 (86 FR 39081)
- Comments due September 1, 2021:
  1. What options should the Task Force consider for any of roadmap elements, and why?
  2. Which capabilities and services provided through the NAIRR should be prioritized?
  3. How can the NAIRR and its components reinforce principles of ethical and responsible research and development of AI, such as those concerning issues of racial and gender equity, fairness, bias, civil rights, transparency, and accountability?
  4. What building blocks already exist for the NAIRR, in terms of government, academic, or private-sector activities, resources, and services?
  5. What role should public-private partnerships play in the NAIRR? What exemplars could be used as a model?
  6. Where do you see limitations in the ability of the NAIRR to democratize access to AI R&D? And how could these limitations be overcome?

FEDERAL REGISTER
The Daily Journal of the United States Government

Notice

Request for Information (RFI) on an Implementation Plan for a National Artificial Intelligence Research Resource

A Notice by the National Science Foundation and the Science and Technology Policy Office on 07/23/2021

This document has a comment period that ends in 35 days. (09/01/2021)

SUBMIT A FORMAL COMMENT

PUBLISHED DOCUMENT

AGENCY:

White House Office of Science and Technology Policy and National Science Foundation.

ACTION:

Request for information.

SUMMARY:

The Office of Science and Technology Policy and the National Science Foundation are issuing this Request for Information (RFI) to inform the work of the National Artificial Intelligence Research Resource (NAIRR) Task Force ("Task Force"). The Task Force has been directed by Congress to develop an implementation roadmap for a shared research infrastructure that would provide Artificial Intelligence (AI) researchers and students across scientific disciplines with access to computational resources, high-quality data, educational tools, and user support.

DATES:

To be considered, responses and comments must be received, no later than 11:59 p.m., EDT on September 1, 2021.

DOCUMENT DETAILS

Printed version: PDF

Publication Date: 07/23/2021

Agencies: National Science Foundation; Office of Science and Technology Policy

Dates: To be considered, responses and comments must be received, no later than 11:59 p.m., EDT on September 1, 2021.

Comments Close: 09/01/2021

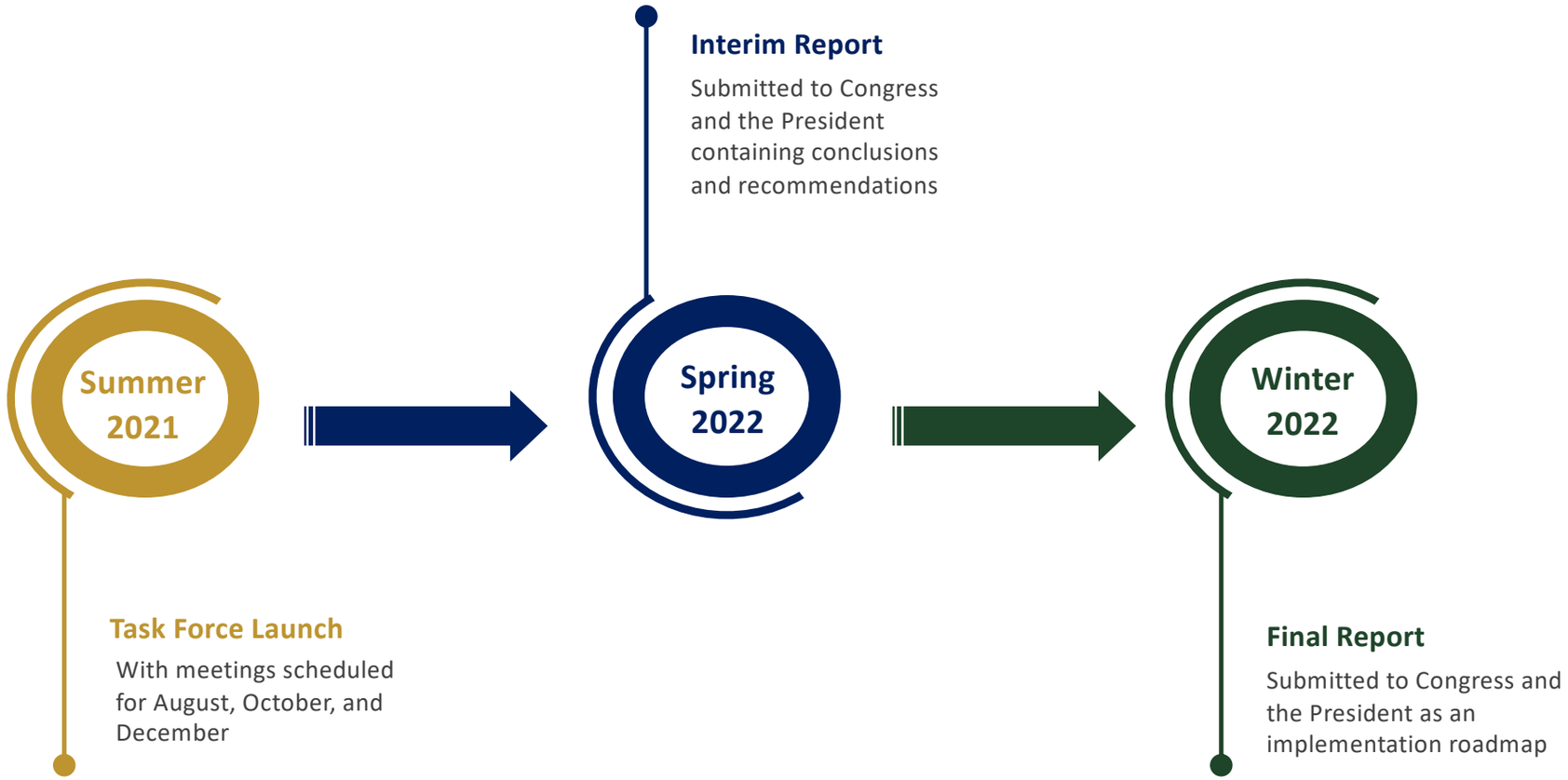
Document Type: Notice

Document Citation: 86 FR 39081

Page: 39081-39082 (2 pages)

Document Number: 2021-15660

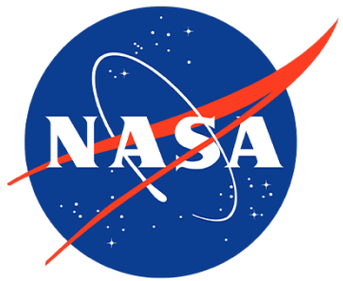
Timeline & Deliverables



Lessons Learned from Federal Cloud Pilots

ERWIN GIANCHANDANI, SENIOR ADVISOR FOR TRANSLATION, INNOVATION, AND PARTNERSHIPS, NATIONAL SCIENCE FOUNDATION

Commercial Cloud Pilots



Lessons Learned

- Access controls
- Cloud administration teams
- Skills gaps
- Impact that can be achieved

Challenges

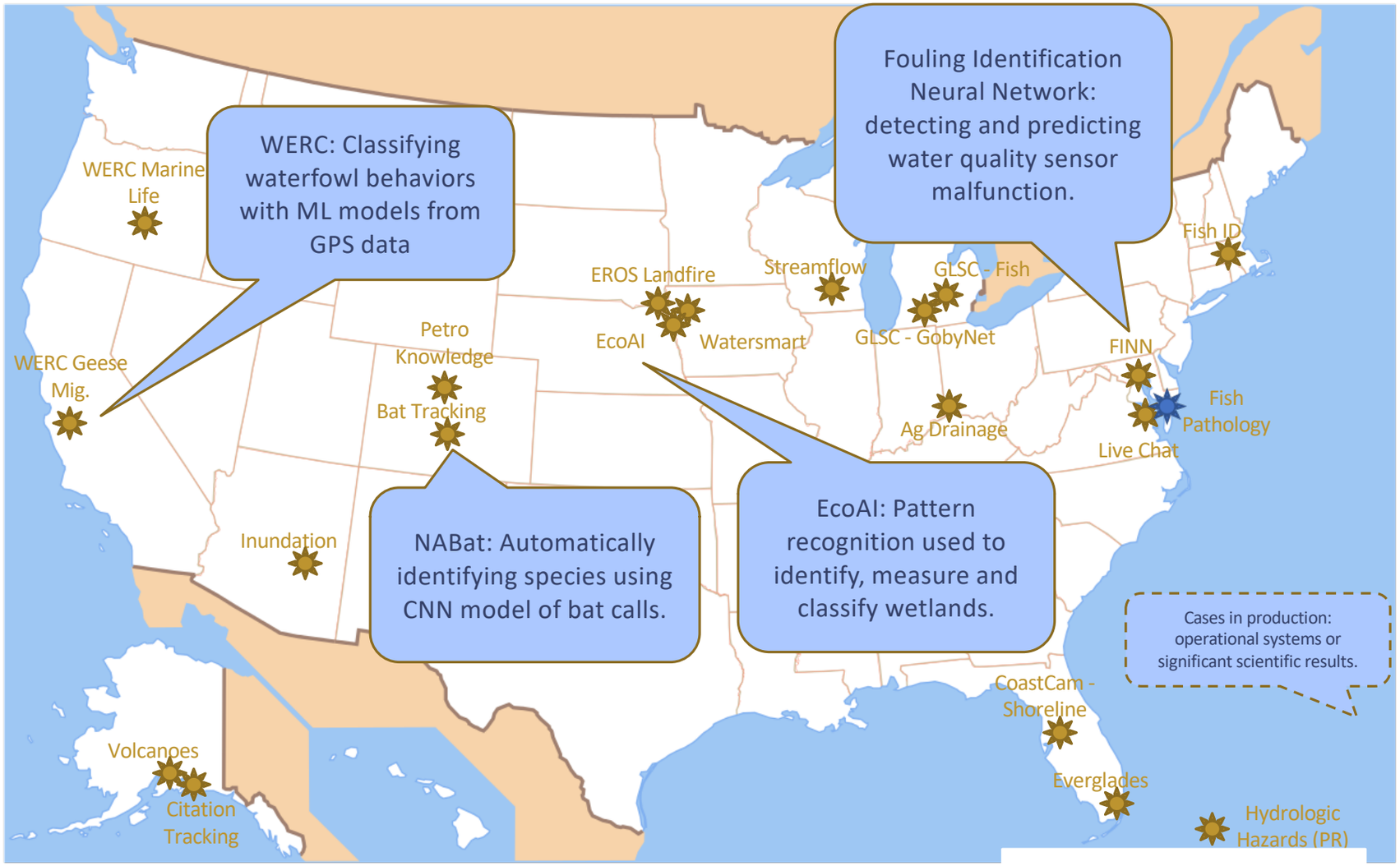
- Researcher authentication and access
- Privacy and security safeguards
- Awareness of data resources
- Billing and charging

Desired Characteristics

- Single sign-on access privileges across platforms
- Common user interfaces
- Automated self-discovery of data and other resources
- Pre-computed resources and workflows
- Further development of management and administration practices

USGS Cloud Program

- Launched in June 2020
- Dedicated AI/ML support team established in August 2020
- 29 current AI/ML use cases



NIH STRIDES Initiative

National Artificial Intelligence Research Resource Task Force

Andrea T. Norris

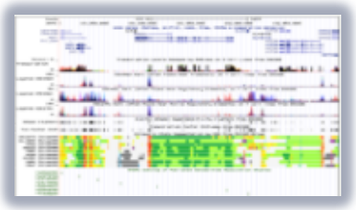
Chief Information Officer, National Institutes of Health

Director, NIH Center for Information Technology

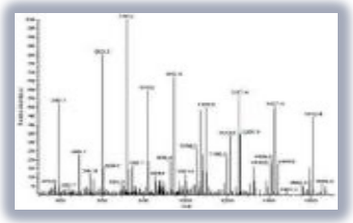
National Institutes of Health

July 28, 2021

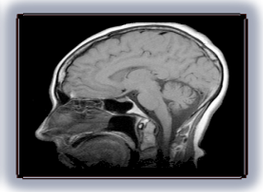
Big Biomedical Data

Genomic




Other 'Omics

Imaging


Electronic Health Records (EHR)


Exposure


Clinical

NIH Strategic Plan for Data Science

Data resource ecosystem and infrastructure modernization

Data sharing, access, and interoperability

EHR, clinical, and observational data availability enhancements

All while ensuring data confidentiality

NIH STRIDES Initiative

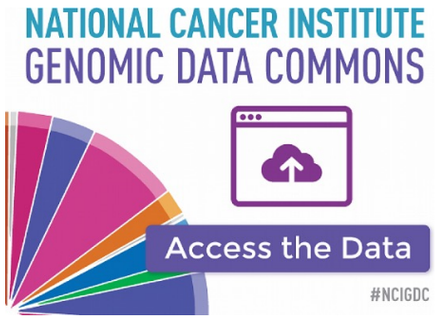
The Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability

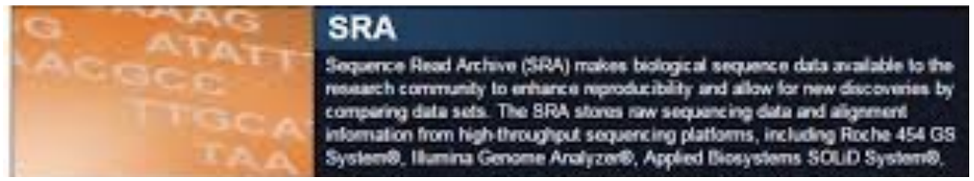
- State-of-the-art data storage and computational capabilities
- Training and education for researchers
- Innovative technologies such as artificial intelligence and machine learning

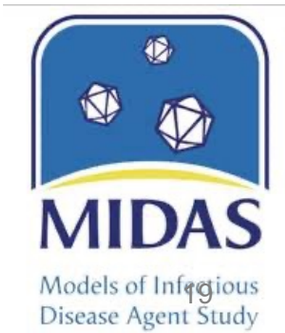
Partnerships with aws Google Cloud Microsoft Azure


Major
Research
Institutions
Enrolled
Northwestern University
Dartmouth
NYU
EMORY
THE UNIVERSITY OF CHICAGO
1890
Stanford University
MAYO CLINIC
JOHNS HOPKINS UNIVERSITY
INDIANA UNIVERSITY
MGH 1811
MASSACHUSETTS GENERAL HOSPITAL
TEXAS A&M UNIVERSITY
1876
UNIVERSITY OF CALIFORNIA SANTA CRUZ
ROSWELL PARK COMPREHENSIVE CANCER CENTER
COLUMBIA UNIVERSITY IRVING MEDICAL CENTER
Penn
UNIVERSITY of PENNS
LUX ET VERITAS
YALE
Yale University
BROWN
MICHIGAN
VE RI TAS
HARVARD MEDICAL SCHOOL
WAYNE STATE UNIVERSITY
CALIFORNIA INSTITUTE OF TECHNOLOGY
1891
Caltech
WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON
1789
Georgetown University
UNIVERSITY OF PITTSBURGH
VERITAS 1787 VIRTUS
DANA-FARBER CANCER INSTITUTE
Icahn School of Medicine at Mount Sinai

Major Research Programs Supported







Helping advance biomedical research by delivering access to industry-leading cloud providers.

The STRIDES Initiative aims to help NIH and its institutions accelerate biomedical research by reducing barriers in utilizing commercial cloud services. This initiative aims to harness the power of the cloud to accelerate biomedical discovery. NIH and NIH-funded researchers can take advantage of STRIDES benefits.

Enroll Now

Gain access to

- Discounts on partner services
- Professional services consultations
- Access to training
- Potential collaborative engagements

115
Petabytes of Data

100M
Compute Hours

>500
Research Institutions and Programs

>$19M
Cost Savings

>3500
People Trained

The Entire Sequence Read Archive (SRA) is Accessible on AWS & GCP

- **36.4 PB of public and controlled-access SRA data**, hosted by the National Center for Biotechnology Information at the National Library of Medicine, is available on Amazon Web Services (AWS) and Google Cloud Platform (GCP) via STRIDES
- Simplifying computational access to these data in conjunction with collaborative partnerships with open source & data analysis platforms **accelerates genomics research and discovery in the management of COVID-19 and beyond.**

KEY OUTCOMES OF USING CLOUD SERVICES AND TOOLS:

- A mechanism for **faster access to vital large datasets**
- Ability to **share data easily from a central location**
- **Availability of compute resources and access to data** for researchers
- **Reproducibility** of analytical processes and datasets generated

Petabase-scale sequence alignment catalyzes viral discovery

https://www.biorxiv.org/content/10.1101/2020.08.07.241729v1.full

Problem: While public sequence data represents a major opportunity for viral discovery, its exploration has been inhibited by a lack of efficient methods for searching this petabyte-scale data which is growing exponentially!

Solution: **Serratus**- a new cloud computing architecture tailored for ultra-high throughput sequence alignment at the peta base scale! The **goal** of Serratus was **to identify and share every coronavirus sequence from over 10 years of data collected by the global research community**.

Outcome: Identification of **tens of thousands of coronavirus and coronavirus-like viral alignments and family identifications made freely available to the research community to catalyze a new era of viral discovery**.

serratus.io

> We can now do this in **3-4 days instead of 12+ months** directly **as a result of the SRA data being available in the cloud**. This means we can share this data with the Covid researchers today, when it can make a difference, not a year from now. This is **important for COVID-19 now and will be important in response to the next pandemic**.

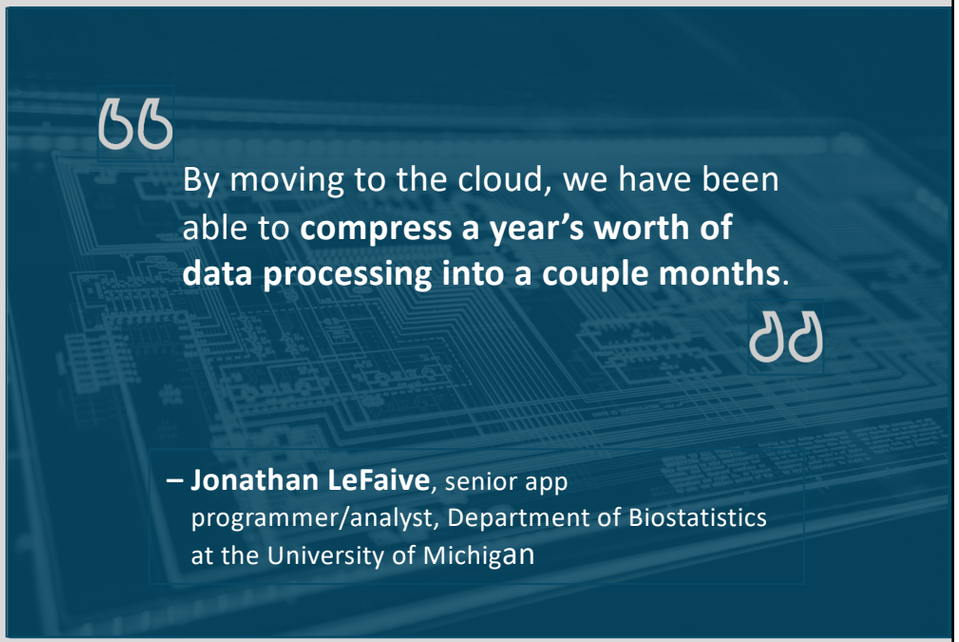
– Artem Babaian, Lead Developer at Serratus and corresponding author for publication.

National Heart, Lung, and Blood Institute

BioData CATALYST

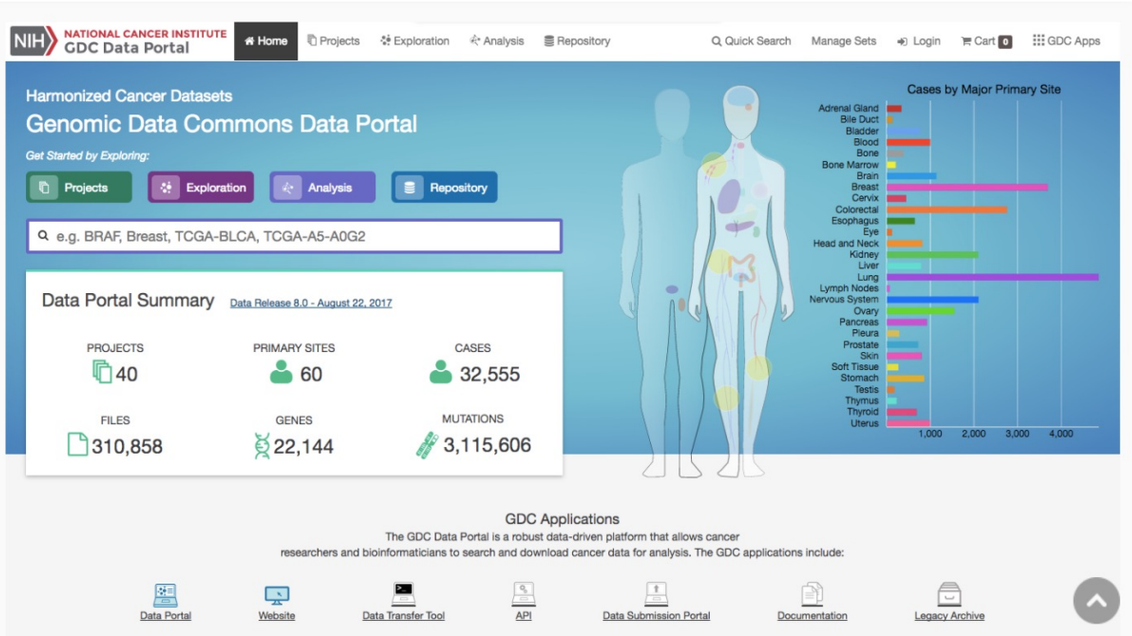
The **TOPMed Imputation Server**, which leverages TOPMed's ethnically diverse data, was **immediately popular among the research community** since it launched in May 2020.

The **STRIDES Initiative made this possible**, as it provided **access to favorable pricing and excellent engineering support** from the STRIDES Initiative partners.



The University of Michigan team manages the TOPMed Imputation Server.

Collaborative, Scalable, Reproducible, Secure

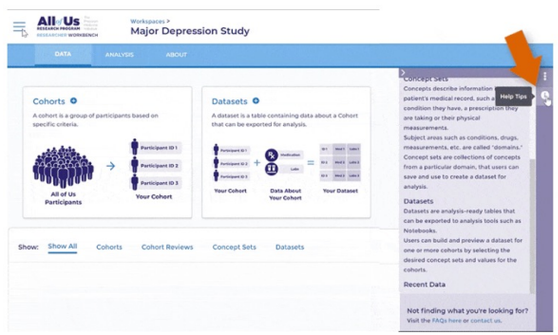


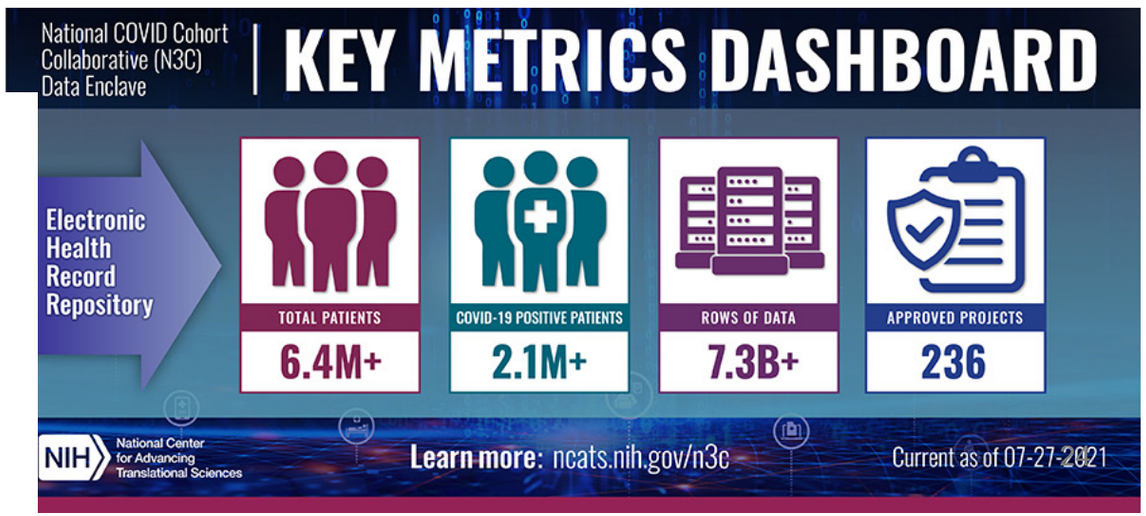
Rich Datasets
Interactive Search
Advanced Analytics and Visualization
High Quality Workflows
Broad Range of Applications and Services
Collaborative Workspaces
Documentation and Job Aids

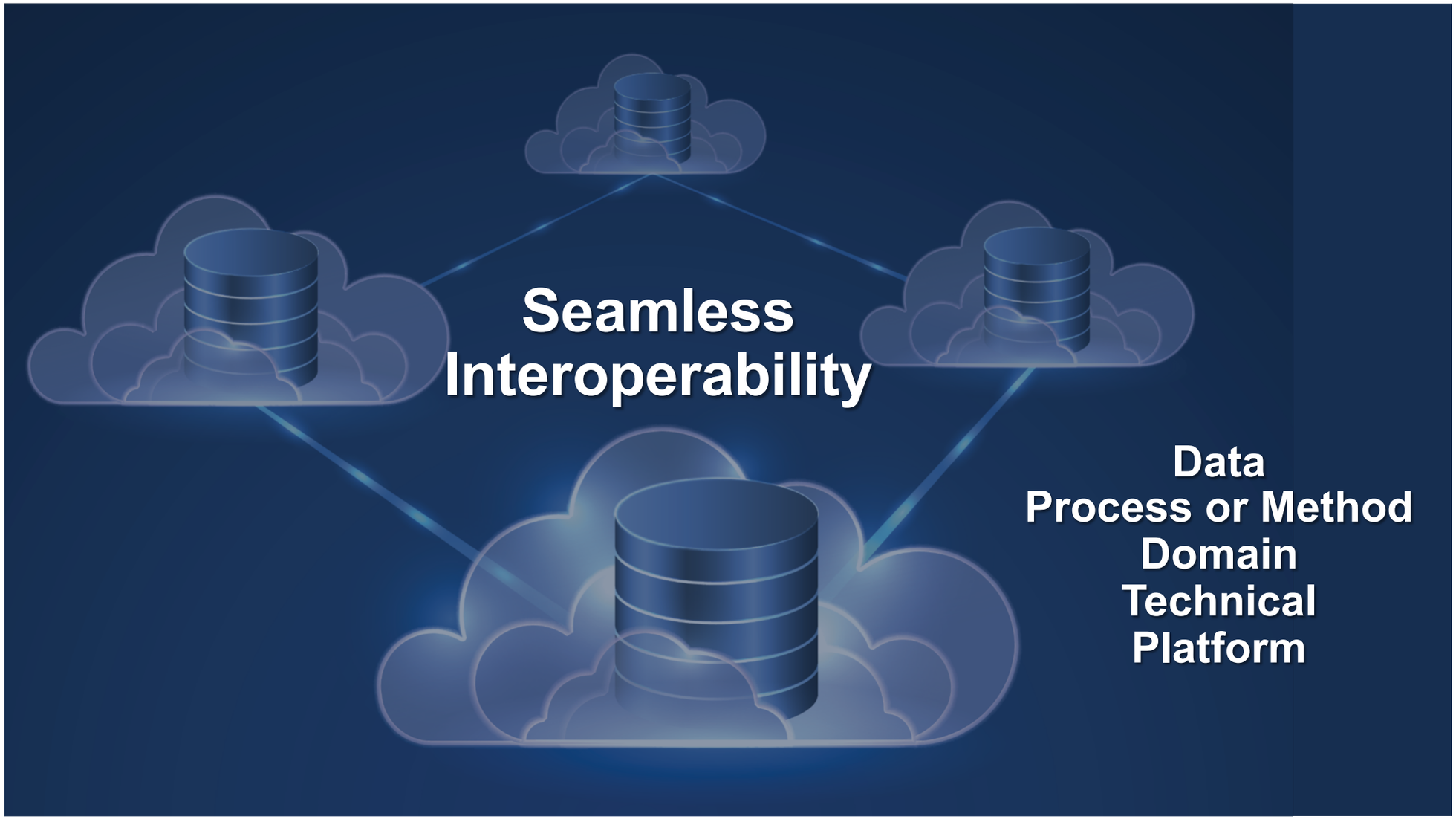
RESEARCHER WORKBENCH

The Researcher Workbench provides tools to enable powerful analysis of Registered Tier data. Registered researchers can conduct multiple research projects simultaneously and collaborate within and across teams.

LEARN ABOUT THE WORKBENCH





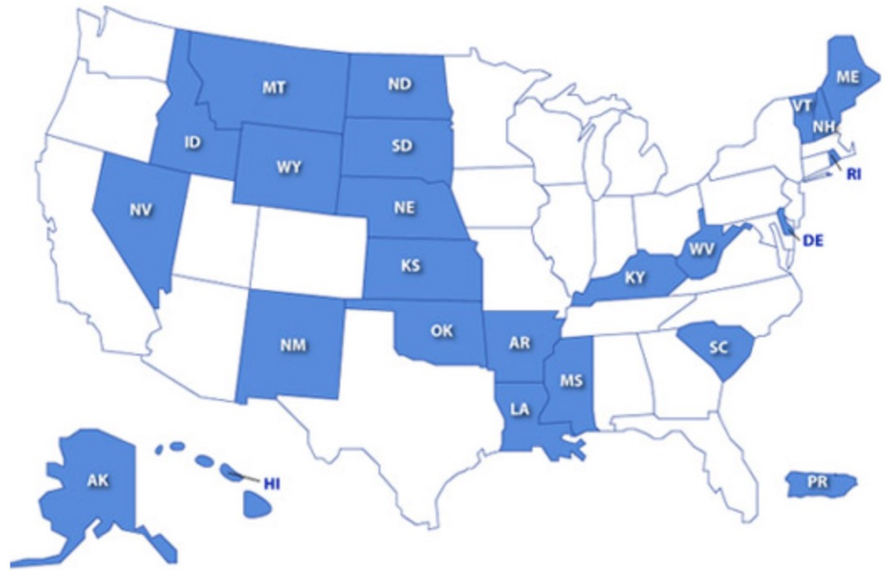

Seamless
Interoperability
Data
Process or Method
Domain
Technical
Platform

Training and Ongoing Support

- Incredible Demand for Cloud Training at all Levels
  - Different Modalities – In person, virtual, scheduled, on-demand
  - Different Venues – At NIH, at Research Institutions, at major conferences or events
- Course Offerings Range from Cloud Fundamentals to Technical Topics (e.g., Security, Networking) to Applied Topics (e.g., Big Data and Machine Learning). Progressively Complex Offerings Aligned with Researcher's Level of Expertise and Biomedical-Focused Curriculum is Most Effective.
- Continuous Outreach and Engagement are Critical
  - Webinars, Newsletters, Tech Talks, Researcher One-on-One Sessions, Success Stories, Guest Speakers, Shared Tool Kits
- Code-a-thons are Regularly Offered to Provide a Hands-On Way for Researchers, Data Scientists, and Others to Interact with the Cloud Platforms to Solve Specific Problems

Support for Major NIH Diversity & Capacity Building Programs

- Collaborative R&D projects with cloud providers to support Minority Serving Institutions and Institutional Development Award (IDeA) states
  - Proteomics pipeline development with University of Arkansas for Medical Sciences, and RNA-seq workflow training effort with University of Maine system
- Targeted training efforts at Minority Serving Institutions, including Historically Black Colleges and Universities and Tribal Colleges and Universities
- Special research credit allocations from cloud providers to help jumpstart programs from institutions typically underrepresented in computational and data-intensive research



IDeA is a program established by Congress to build faculty and research capacity in states that historically have had low levels of NIH funding.

NIH National Institute of General Medical Sciences

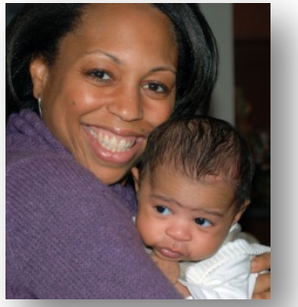
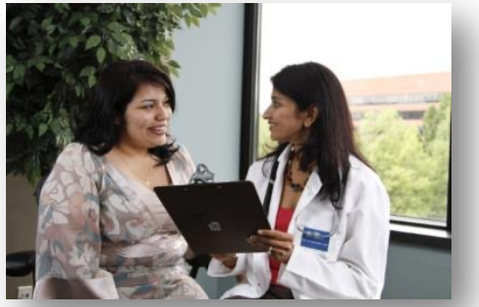
UAMS UNIVERSITY OF ARKANSAS FOR MEDICAL SCIENCES

Maine's Public Universities UNIVERSITY OF MAINE SYSTEM

- Democratization of computational research and data science
  - Leveling the playing field for underrepresented communities and institutions in biomedical research
- Cost savings and efficiencies for the research community at large
  - More usage begets more savings and greater overall discounts for all
  - New insights into cost and usage of data sets and resources to inform sustainability efforts

Connecting data, tools, resources, and researchers in new ways

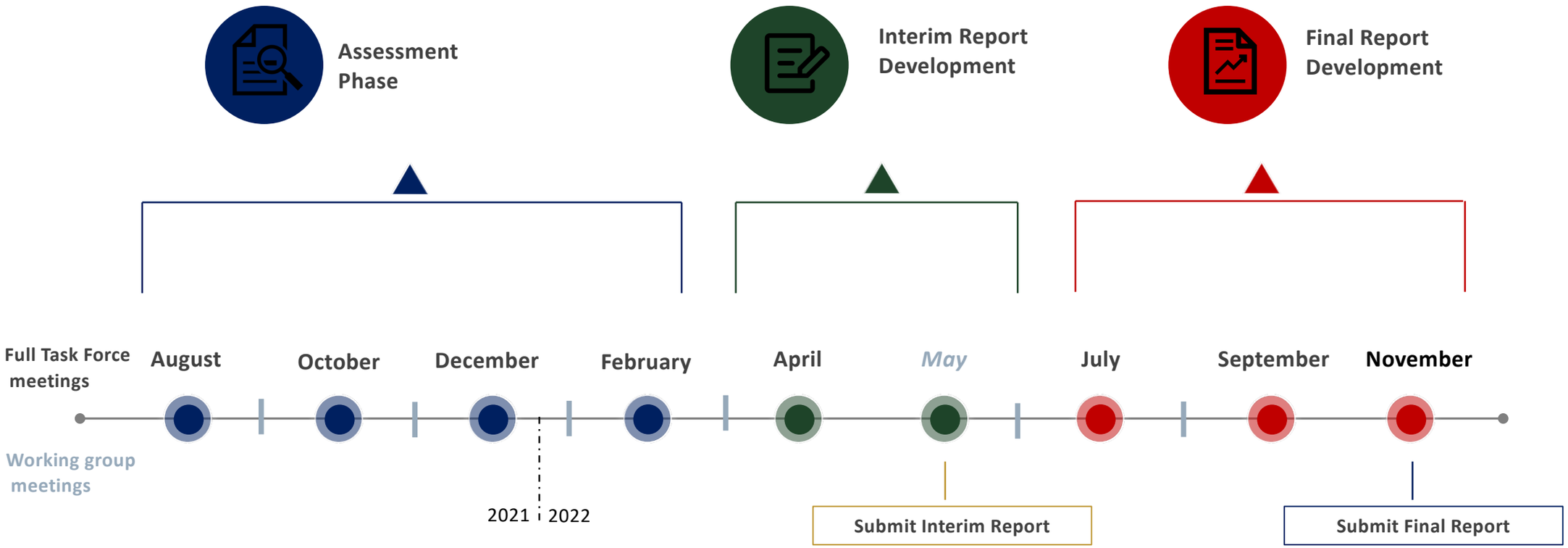
To accelerate discovery and innovation in health

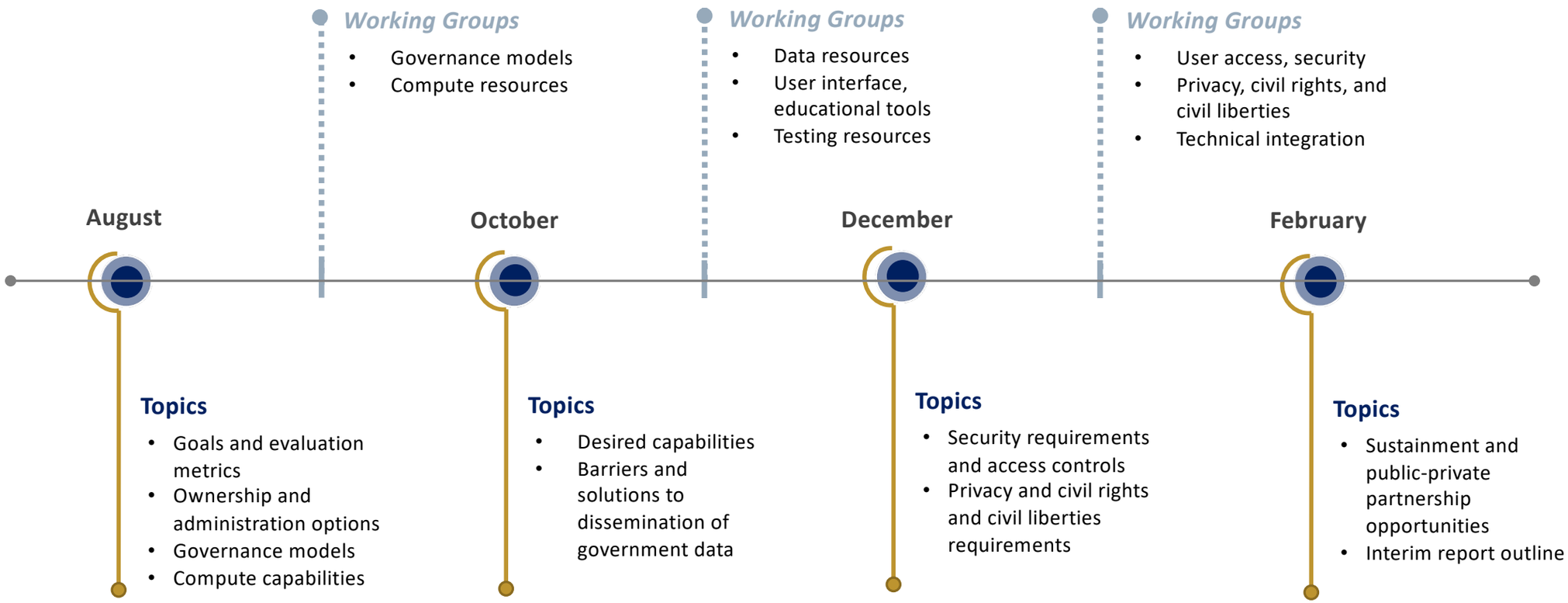
Proposed NAIRR Task Force Workplan

LYNNE PARKER, DIRECTOR, NATIONAL AI INITIATIVE OFFICE
WHITE HOUSE OFFICE OF SCIENCE AND TECHNOLOGY POLICY

Timeline


Assessment Phase
Interim Report Development
Final Report Development
Full Task Force meetings
August
October
December
February
April
May
July
September
November
Working group meetings
2021
2022
Submit Interim Report
Submit Final Report

Assessment Phase



To Dos

- Members:
  - Submit your working group preferences by August 3rd
  - Continue to send ideas on models/consultations

- Co-Chairs:
  - Send out meeting minutes
  - Prepare logistics/agenda for August meeting