Agenda

NATIONAL AI RESEARCH RESOURCE TASK FORCE: MEETING #4, December 13, 2021

11:00-11:10	**Welcome and Administrative Remarks,** Lynne Parker & Manish Parashar
11:10-11:50	**Overarching NAIRR Vision,** Lynne Parker
11:50-12:50	**Readout and Discussion of Draft Recommendations: User Resources Working Group,** Fei-Fei Li
12:50-1:50	**Readout and Discussion of Draft Recommendations: Testbed/Testing Resources Working Group,** Andrew Moore
1:50-2:20	**Break**
2:20-3:20	**Readout and Discussion of Draft Recommendations: Data Working Group,** Daniela Braga & Julia Lane
3:20-4:20	**Panel:** • Solon Barocas, *Principal Researcher, Microsoft Research; Adjunct Assistant Professor, Information Science, Cornell University* • Lujo Bauer, *Professor, Electrical & Computer Engineering and Computer Science, Carnegie Mellon University* • danah boyd, *Partner Researcher, Microsoft Research; and Founder/President, Data & Society* • Deborah Raji, *Fellow, Mozilla Foundation* • Nicol Turner Lee, *Senior Fellow and Director of the Center for Technology Innovation, Brookings Institution* • Hannah Quay-de la Vallee, *Senior Technologist, Center for Democracy and Technology*
4:20-4:40	**Break**
4:40-4:55	**Briefing: Security and Access Control Considerations for the NAIRR,** Emily Grumbling & Morgan Livingston, STPI
4:55-5:20	**Discussion: Usable Security and User Access Controls,** Elham Tabassi
5:20-5:45	**Briefing: Technical Integration,** Mike Norman
5:45-5:55	**Questions from Public,** Lynne Parker
5:55-6:00	**Closing Remarks,** Manish Parashar

Framing the NAIRR Vision

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

Overarching Vision & Objectives

- **Foundation:** The National AI Research Resource (NAIRR) is envisioned as a shared computing and data infrastructure that would provide a diverse set of researchers and students across a range of fields with access to a holistic ecosystem of resources to fuel AI discovery and innovation.

- **Objectives:** Strengthen the U.S. AI innovation ecosystem and support economic growth by:
    i. Lowering barriers to entry – in terms of access to computational and data resources and providing user training and support;
    ii. Promoting AI skills and knowledge, thereby laying the foundation for the AI workforce;
    iii. Broadening participation to include all segments of our diverse nation; and
    iv. Supporting innovative and novel efforts in foundational and use-inspired AI research including the adoption of AI across all fields of science and engineering and all sectors of the economy.

Why do we need a NAIRR?

- **To unlock innovation**: AI holds the potential to positively impact science, the economy, national security, and society
- **To overcome the "resource-divide"**: today access to computational and data resources are primarily limited to the large private sector firms and well-resourced universities.
  - Traditionally underserved communities lack sufficient representation and pathways to participation in the field, contributing to cases of bias in AI tools and approaches.
- **To cultivate diversity in and of AI**: Expansion of access will broaden the range of researchers involved in AI, growing and diversifying approaches to and applications of AI.
  - Increase opportunities for research in critical areas such as auditing, testing and evaluation, bias mitigation, security, etc.

The National AI Initiative

- Congress established the NAIRR Task Force to take on the time and issue-bounded task to build an implementation plan that describes a national solution to expand access to data and computing resources for AI R&D. *The Task Force started its work with this charge.*

- The NAIRR Task Force is one component of a government-wide initiative to:
  - Invest in AI research and development;
  - lead the world in the development and use of trustworthy AI systems across public and private sectors; and
  - prepare the present and future U.S. workforce for integration of AI systems across all sectors of economy and society.

- A National AI Advisory Council is in the process of being established to advise the National AI Initiative and the President on the broader portfolio of AI priorities and activities.

- The NAIRR Task Force will conclude its work after submission of its final report; the Task Force itself will not execute any of its recommendations nor be involved in the administration of a future NAIRR.

Composition & Resources

- **Composition**: Federated ecosystem of diverse resources, knitting together existing resources and creates new resources, bringing them into an accessible cyberinfrastructure.
  - Public- and private-sector resources
  - Take advantage of campus-level, region-level, and national-level resources
  - Possess attributes of transparency and trust; security and robustness; accessibility; independence; scalable functionality; sustainability; and oversight and accountability.

- **Resources:**
  - Computational Resources
  - Datasets
  - AI testbeds
  - Software tools
  - User support
  - Unified interface

User Base: Baseline

- **User Profiles:**
  1. AI Researchers
  2. Researchers conducting use-inspired AI research and using AI to advance other fields
  3. Students

- **As a baseline, the NAIRR should be open to U.S.-based AI researchers and students at:**
  - — Academic institutions;
  - — Non-profit research organizations;
  - — National laboratories;
  - — Federally-funded research and development centers; and
  - — Startup companies or research organizations that have been awarded Federal Small Business Innovation Research (SBIR) or Small Business Technology Transfer (STTR) grants.

- **Access based on the assumptions:**
  - — Users of NAIRR computational resources would need to pass a proposal evaluation process
  - — Projects would be open and publishable, pursuant to relevant open-source policies and practices of funding agencies
  - — Access to all resources subject to clear use policies and user agreements

For Discussion: Additional User Base Considerations

- **Additional Users:**
  - — Early-stage startups;
  - — Private-sector researchers whose companies have made resources available to the NAIRR; and/or
  - — Private-sector researchers who pay fees.

- **Proprietary Research:**
  - — A small percentage of the resources could be used to support proprietary research.

- **Governance Models:**
  - — A requirement that overall allocations to industry do not exceed a certain percentage (say, 20%) of the total available resources;
  - — A requirement that private-sector users provide resources or pay fees in order to access the NAIRR;
  - — Different policies for different types of resources:
    - i. Access to computational resources would be provided to the agreed upon user base (above);
    - ii. Non-sensitive data resources could be as open as possible, to both the researchers and data providers, without a proposal review process; and/or
    - iii. Testbed resources could be provided as open access, although potentially with a user fee.

For Discussion: Resource Allocation and the NAIRR Governance Body

- **Options to consider:**
  - —Access to the NAIRR resources could be determined solely by Federal funding agencies;
  - —A separate NAIRR governance body could govern allocation through a proposal process independent of other Federal funding processes;
  - —In a hybrid approach, access by Federal research grantees would be managed by the Federal funding agencies, and the NAIRR governance body would manage an independent review and allocation process for researchers without current Federal funding; and/or
  - —The NAIRR governance body could provide access to NAIRR resources to private-sector users through a proposal review process, if they provide resources or pay fees for access.

Additional Areas for Discussion

- Any other areas that warrant discussion among the full group?

- Please send feedback to us on any additional aspects of this vision by January 12, 2022

Goal: Recommend a User Portal Strategy for NAIRR

WORKING GROUP MEMBERS: MICHAEL NORMAN, FRED STREITZ, FEI-FEI LI

Outline

- Mission Statement
- Design Criteria
- Key Recommendations
- Pitfalls, Issue Areas
- Specific responses to charge questions

Mission Statement/Charge

- Provide recommendations on an integrated user interface and advice on training and support for NAIRR users
- 5 specific charge questions at the end of this deck

Design Criteria

- Integrated (services, resources, data, training material, ...)
- User-centered design
  - Multi-stakeholder: sponsor, researcher, educator, resource provider
  - Enable collaboration
  - Ease of use
  - Flexibility: support cloud & edge
  - Account management
- Scalability and extensibility: data & compute
- Speed
  - Access
  - Compute
  - Service
- Training & support
  - Multi-dimensional, expanding topic space
  - Training: Meet users where they are
  - Support: Tiered support

Key Recommendations

- Leverage user portal concepts from existing state-of-the art approaches
  - Need to evaluate cost-effectiveness of building NAIRR user portal in-house versus acquiring commercially
  - Recommend to outsource the design, construction, and maintenance of NAIRR user portal to a commercial entity that has successfully done this
    - NAIRR defines the design criteria, and puts out a bid
    - We should not build this in house
- NAIRR will identify and curate appropriate education and training materials for different skill levels leveraging content from NAIRR resource providers. Content will be delivered through the portal.

Pitfalls, Issue Areas

- Watch out for the fast-pace of the field, extensibility is critical
  - More and more compute resources and datasets
  - Build with this in mind
- Robust governance, allocation and monitoring models for everyone, especially traditionally underserved and underrepresented communities
- SECURITY!
  - Security of operations and data

Predecisional

Q1 - How should the portal interface be designed and constructed to maximize usability and intuitive navigation by users?

- Lots of good lessons/ideas from XSEDE, CloudBank, Rescale* and CoreScientific*
  - (*) The WG has interviewed two companies specialized in hybrid cloud user platform companies to learn about the current industry efforts
- Distinction between information serving and job submission portals

  →NAIRR portal should do both
- User should be able to select an AI application, compute resource, and data source(s) from a list and build, launch, and monitor the job on either on-prem HPC or public cloud with a common interface
  - Both Rescale and CoreScientific Plexus portals already do this
  - See design criteria for additional requirements

Q2 - What level of technical support should be made available to first-time users of the NAIRR? To prior users of the NAIRR? For example, what sort of coaching and guidance should be provided to researchers on how to structure and run their (potentially distributed) experiments on NAIRR resources?

- On the portal site:
  - “How do I use the portal?” documentation
  - Self-help/self-paced tutorials
  - Q/A user forum (e.g., Concourse)
- Training programs
  - Focus on first time and traditionally underserved
- Tech support
  - Frontline support: build into the portal
  - Advanced/devops support: included in resource provider contracts
- Agreed on not focusing on domain expertise
  - Governance model mentions a path to access domain experts outside of NAIRR staff

Q3 - What types of educational tools and resources should be included in the NAIRR?

- If “education” means general education of expertise, then it’s outside of NAIRR’s mission
- NAIRR platform can provide facilitatory functions for educational efforts
  - Part of the application stack can provide an educational platform on NAIRR resources. (e.g. Berkeley data stack on CloudBank)
  - NAIRR does not provide discipline-specific educational content
- A platform for community building among PIs, students, researchers
  - E.g. chat functions, meeting rooms, forums, etc.

(relate to s8)

Q4 - What kind of staffing will be necessary to support the provision of educational, training, and other user resources?

- NAIRR portal DevOps/maintenance will be outsourced
- FT employees at the NAIRR Management Entity (e.g. FFRDC)
  - In house management and operations
  - In house training experts
  - Community coordinator(s)
- Part-time contractors
  - Some from resource providers

Q5 - Portal access to data

- YES!
- Data is first-class citizen, part of the integrated portal
- Includes
    - Resource attached data
    - Free standing data
    - Searchable data catalog

Predecisional

Proposal for NAIRR Testbed Resources

NAIRR TESTBED WORKING GROUP (WG)

ANDREW MOORE, WG LEAD

OREN ETZIONI

ELHAM TABASSI

The Working Group considered two main questions

Should NAIRR spend resources on testbeds?

If so, what is the most effective way to spend those resources?

Recommendation: The NAIRR should provide access to testbeds.

Testbeds offer the potential to:

Accelerate research (e.g., as occurred with ImageNet, MLPerf, DARPA AV Challenges, others)

Support equity for researchers by providing testing data and infrastructure that would be expensive to maintain independently

Ensure quality testing resources for accurate/standard measures of performance

Inspire the next generation of researchers through use-inspired AI testbeds

Recommendation: The NAIRR should spend 5% to at most 10% of its resources on testbeds.

Assuming an annual NAIRR budget of $20M

~$1.25M for testbeds

- 2-3 full-time staff
- $200K outsourced web development for a live testbed catalogue
- Additional staff of engineers/ data scientists to build testbeds identified by NAIRR governance

NAIRR Testbed Office duties:

- Maintain a world-class catalogue of testbeds with pointers to those resources
- Develop some testbeds that are not otherwise available
- Advocate for consistency among testbeds to enable reuse of infrastructure

Simple Taxonomy of Testbeds

Comparison testbed - allows researchers to measure effectiveness of new engineering, math, or algorithmic developments

Validation testbed – allows governance bodies to decide whether an end-to-end system is mature enough to move to advanced development stages

Tentative Recommendation:

The NAIRR should focus solely on comparison testbeds, in alignment with NAIRR's research focus, leaving validation testbeds to NIST or industry consortia.

Types of Comparison Testbeds

1. Open Book Modeling competitions
2. Closed Book Modeling competitions
3. Simulated Perceive-Decide-Act competitions
4. Real-World Perceive-Decide-Act competitions

Recommendation: NAIRR should devote some resources to cataloguing all four types and promoting their accessibility and use.

Recommendation: NAIRR should provide additional infrastructure to support types 1, 2, and 3, but de-emphasize type 4.

Additional Recommendations:

Federal agency RFPs that require testbeds should be coordinated through the NAIRR testbed office to avoid duplication of efforts

Proposal for NAIRR Data Resources

NAIRR Data Working Group (WG)

Daniela Braga, co-lead

Julia Lane, co-lead

Mark Dean

Dan Stanzione

NAIRR Roadmap must include two data-related elements

- (D) 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.
- (E) An assessment of, and recommended solutions to, barriers to the **dissemination and use of high-quality government data sets** as part of the National Artificial Intelligence Research Resource.

The WG considered several scoping questions

- How should the NAIRR facilitate user access to existing data repositories and resources?
- What curated data sets should be provided through the NAIRR? What search and other capabilities should be designed into the NAIRR?
- What types of government data should be made available through the NAIRR?
- How should dissemination be managed responsibly?
- How should access to data resources be managed?
- How should privacy, civil liberties, and civil rights considerations be designed into the NAIRR data resources as a first principle?
- How should data resources interact with the compute resources discussed on 10/25?
- What data engineering and curation expertise will need to be included in the NAIRR workforce?
- How should specific use cases drive any of the above?

The WG gathered information three ways

- Considered four potential use cases/data types
  - Transportation system data
  - Social/economic data
  - Contact centers
  - Natural hazards
- Drew on input from six external experts (briefers at 10/25 Task Force meeting)
  - Ian Foster, Director, Data Science and Learning Division, Argonne National Laboratory; Professor of Computer Science, University of Chicago
  - Robert L. Grossman, Professor of Medicine and Computer Science, University of Chicago
  - Ron Hutchins, Vice Provost for Academic Technologies, University of Virginia
  - Anita Nikolich, Research Scientist and Director of Research and Technology Innovation, University of Illinois at Urbana-Champaign
  - Nancy Potok, CEO, NAPx Consulting; former Chief Statistician of the United States
  - Andrew Trask, Leader, OpenMined
- Considered public input in response to request for information (RFI)

Predecisional

The committee considered four canonical use cases

Transportation data

- Includes video, image, GPS, human reaction, and sensor data
- Challenges: Data curation for effective machine learning is very challenging, especially for use across differing modalities.

Social data

- Includes data about humans and organizations, gathered through surveys or digital exhaust
- Challenges: Multiple data management, curation, and governance issues exacerbated by privacy and resulting access constraints which limit the transparency, reproducibility and replicability of research and introduce significant potential for bias

Contact center data

- Includes multilingual voice or text interactions (via phone, chat, or email) between customers and customer support agents (public or private sector)
- Challenges: Personally Identifiable Information (PII), usually present in the beginning of the interaction, when the client has to provide name, address, account number to be identified.

Natural hazards data

The three phases of hazard response each generate massive amounts of data suitable for use-inspired AI R&D:

- Includes: Forecast/prediction prior to the event – data include environmental conditions (e.g. available fuel for wildfires, wave height on the ocean, seismic readings, atmospheric conditions, etc.) or from simulations (predictions of storm paths, earthquake forecasts, etc.); Assessment during and in the immediate aftermath of the event– data include “conditions on the ground”, most frequently geo-located images from reconnaissance teams, UAVs, aircraft, or satellites; Mitigation in the longer recovery period- simulation and experimental data inform infrastructure improvements and new building codes
- Challenges: Curation and integration

The committee identified three main types of government data to make available through the NAIRR

Statistical data

- Decennial census and many other important surveys
    - Use subject to CIPSEA; U.S.C. Title 13
- Federal tax information
    - Use subject to U.S.C. Titles 13 and 26

Administrative data

- Programmatic and transaction data (e.g., TANF/SNAP/WIC/CDC/HHS/VA/DOD/FAA/DOJ/ DOT/USDA/HUD)

Data generated by federally funded research

- NASA/FEMA/NOAA/NSF/NIH
- Supports the Foundations for Evidence-Based Policymaking Act of 2018

Overarching Finding: The full value of AI is often not realized without high quality, trusted, dense, transparent data

For example

1. In social data, badly trained criminal justice algorithms and lack of transparency can lead to social harm
2. For customer care data, underrepresentation of certain demographic groups in training data leads to false positives and resultant inequitable treatment
3. In natural hazards data, sparse data leads to poor simulations of the impact of hurricane surges
4. For self-driving cars, sparse data on rare events can have significant negative impact

Specific findings

- For most research domains, data is highly distributed, not often discoverable, and seldom reusable
- Data are extremely heterogeneous within and across domains (e.g., video, voice, text, image, sensor, GPS), and poorly documented and curated. The sheer volume and variety of data of interest will make it impossible for the NAIRR to curate it all
- While a wide variety of data types are of potential interest; many come unstructured, but only provide value if properly labelled. Labeling/tagging/annotation are difficult to automate and require significant hours of expert analysis. Data labeling and curation standards are generally evolving, limited, absent, or inconsistently adopted
- Research data curation relies on communities of expert researchers (academic or commercial); in some areas, standards are not established or adopted
- Transfer of large data sets can be expensive in the commercial cloud
- Privacy is not an absolute – it is contextual – and privacy risks should be assessed in the context of the value proposition
- The work of NAIRR could complement the work supporting the implementation of the Foundations of Evidence-based Policymaking Act

NAIRR Roadmap must include two data-related elements

- (D) 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.
- (E) An assessment of, and recommended solutions to, barriers to the **dissemination and use of high-quality government data sets** as part of the National Artificial Intelligence Research Resource.

D. Goal for NAIRR data resource infrastructure

Provide high quality, trusted, transparent and dense training and test data that have been curated, recognized, and validated by a diverse community.

The NAIRR will achieve this goal by treating data as a first-class asset and institutionalizing:

1. **Trust**: Being transparent in its methodology and protecting privacy and confidentiality while maintaining quality and value
2. **Curation**: Bringing together a diverse constituency of expert contributors
3. **Validation**: Creating incentives for good data practices by the community (inspired by Kaggle)
4. **Discoverability**: Using ML tools and expert engagement to find how data are used to answer what questions

D. General Recommendations

The NAIRR should include:

1. A technical infrastructure to host data in secure facilities so that costs are minimized, and access is maximized
2. An access infrastructure that is networked to enable the domain-specific heterogeneity of data structures to be addressed
3. Trained NAIRR staff that support diverse community data curation, linkage, and validation activities
4. Training programs to develop a diverse AI workforce, foster innovation, and create community driven value
5. A search and discovery platform so that knowledge about data use, users, and value can be identified, leveraged, and replicated

D. 1a. Hosting infrastructure

Recommendation: The NAIRR should coordinate a network of trusted data/compute providers and hosts for a transparent and responsible data marketplace.

- Data heterogeneity means that the NAIRR should support multiple partner sites (e.g., compute centers or other partners) but that data can be co-located and combined either at NAIRR or a partner site. . Data providers within the resource will serve specific domains but can learn from each other
- Data scale means that NAIRR will need to provide the processing capability to support machine learning, modeling, simulation and testing processes
- Establishing a data marketplace means that search and discovery tools will need to be deployed and data sharing incentives established

D1b. Access infrastructure

Recommendation: Access to data should be tiered, controlled by the data providers, and provided through the same portal through which compute resources are provided.

- Data access rights and pricing levels should be tiered by institution type, research context, and data sensitivity (accounting for legal requirements), with appropriate user agreements, training, and credentialing/authentication.
- Organizations should be charged for access at a reasonable rate in alignment with NAIRR goals

D1c. Privacy and confidentiality

Recommendation: Follow the “Five Safes” framework for decision-making (safe projects, people, data, settings, and outputs).

- Confidential data should be protected by restricting access to authorized users rather than by applying privacy-preserving technologies, because while privacy-preserving technologies hold promise, they are not currently able to address quickly changing data structures, and are not timely or robust enough to ensure quality analysis in most cases
- The federal government should invest in R&D in the area of privacy-preserving ML methods, advancing the technologies to replace PII in data sets, and improving tools and environments to process data with PII in a secured way
- The federal government should invest in R&D in the area of data ethics, contextual privacy, and informed consent

D2a. Data governance

Recommendation: The NAIRR community and leadership should establish and periodically update policies and governance entities that address data quality and use.

- NAIRR should produce transparent policies and oversight protocols for the AI community to inform practices such as: data version control, provenance, combined/derived datasets, curation standards, standards development and enforcement, access rights, and pricing tiers.
- NAIRR should establish governance boards to oversee compliance with data policies. Those boards should include community members or independent oversight

D2b. Biases, civil liberties, and civil rights

Recommendation: The NAIRR should establish governance policies, fund oversight entities and evaluations, and enable public engagement to promote transparency and reduce bias and potential harms associated with NAIRR data use.

- An internal governance board should consider potential harms of NAIRR data use and research outputs to inform decision-making about NAIRR assets, tools, policies, data types, and uses
- Data policies and standards should address the need for diversity and representation in data sets and among data curators and AI researchers
- NAIRR resources should include tools for assessing bias in data and models
- The federal government should support R&D on bias and fairness in data-driven AI
- The NAIRR should fund an ombudsman’s office to engage with the public and hear concerns over potential harms
- The NAIRR should fund ongoing, independent monitoring of real or potential harms associated with NAIRR data use and AI models, with transparent public reporting

D2c. Data contributions

Recommendation: The NAIRR ecosystem should incentivize contribution of high quality data for AI R&D.

- People who bring data that are useful and used should receive recognition and credits with which to access NAIRR resources
- Data use and impacts should be measured, monitored, and displayed on a NAIRR leaderboard
- Contributed data must meet a minimum quality standard, with standards developed by the NAIRR community

D3a. Curation approach

Recommendation: The NAIRR should provide infrastructure and staff support for provider, host, or community data curation and incentivize community driven improvements to data quality.

- The NAIRR should support community workshops (analogous to IEEE's; the Bermuda conference) with express remit to establish standards that are required for putting data in NAIRR
- Contributors to establishment and/or ongoing evolution of NAIRR data standards should get credits for compute or data access, or citations in the leaderboard
- Sufficient curation is required to allow for the combination of like datatypes across research groups – with some (likely automated) data quality checks to remove bad data from analyses.
- NAIRR should publish standards and provide tools and support for data ingestion, validation of compliance with standards, and QA/QC, but should not curate data itself.

D3b. Technical support

Recommendation: Substantial resources should be dedicated to technical support staff

- Data users, contributors, and curators will require support to understand and meet technical standards and to ensure rigorous data use
- Data heterogeneity and complexity means that there will be substantial tacit knowledge embedded in support staff. Attracting and retaining skilled staff with appropriate compensation packages will be essential

D4. Training

Recommendation: Substantial resources should be devoted to the establishment of training programs on NAIRR data policies, use, and curation.

- NAIRR should make some open data available to all for education/training purposes
- Data users, contributors, and curators will require training to understand and meet technical standards and to ensure rigorous data use
- Training can act as a catalyst to engage and expand diverse constituencies

D5. Data search and discovery

Recommendation: The NAIRR should establish a transparent data marketplace by creating a search and discovery platform that uses AI to find the most valuable and relevant data for researchers -

- NAIRR should be charged with establishing a value ecosystem around data that can be used for AI
- NAIRR should seed the marketplace with core datasets, to jump-start the value discovery. It should use digital tools to identify key datasets in each area, partner with academic and local communities to add data and knowledge and engage governors, legislators, and local decision makers to ensure that the marketplace demonstrates value

NAIRR Roadmap must include two data-related elements

- (D) 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.
- (E) An assessment of, and recommended solutions to, barriers to the **dissemination and use of high-quality government data sets** as part of the National Artificial Intelligence Research Resource.

Predecisional

E. Barriers and Solutions to access to and use of government data

1. Legal barriers to data sharing

Barriers: Lack of uniform guidance and interpretations; no clear compliance standard for compliance so legal interpretation is varied

Solutions: Standardized legal guidance; standardized technologies FedRAMP; FISMA;

2. Privacy and confidentiality protections

Barriers: Privacy concerns have limited access to data and resulted in reduced value of government data

Solution: Privacy and value should be jointly determined by an open and transparent process. Access to researchers should be integral to producing value

3. Government Workforce capacity

Barriers: Government pay scales are inadequate to attract and retain high quality data scientists

Solution: Either establish FFRDC that can pay market wages or establish separate pay scale

Definitions

- **Deidentification**- The process used to prevent personal identifiers—both direct and indirect—from being revealed and/or connected with other person specific information. This can include replacement of PII with hashed identifiers
- **Anonymization**- Sub-category of de-identification where direct and indirect personal identifiers have been removed such that the person can never be identified and re-identified. True data anonymization is believed to be impossible given the potential correlation with other data sources and the level of data needed to be retained to make the data useful.
- **Five Safes**- A framework for helping make decisions about making effective use of data which is confidential or sensitive. It is mainly used to describe or design research access to statistical data held by government agencies, and by data archives. The five dimensions for problem solving include: projects, people, settings, data and outputs. (Wikipedia)
- **Standards**- Standards provide consistency in the creation, use and testing of information and information systems.
- **Curation**- Organization, maintenance, and management of a collection of one or more datasets to support their use by specific groups and/or systems. Organization of a dataset implies the process of making the data easily accessible, consistent in format and structure, and viable for use in its intended purpose.
- **Ethics**- Often defined as a set of moral principles or values. But these often vary by society or organization. Ethics in computational systems could be defined as: “do no harm”, “protect from harm”, and “equal treatment of all” individuals, groups, society, and information.
- **Bias**- Bias in datasets usually occurs when the bias of the person or group collecting the data “leaks” or is “injected” into the dataset via the data collection process, either on purpose or unconsciously. Societal norms and expectations can also create biased data, as can incomplete data sets that underrepresent or misrepresent certain groups. How, where, when, from who and what attributes are included in the data collected can create bias. Biased data can create bias in associated algorithms and processes.
- **Privacy vs. confidentiality**- 1) Privacy is defined as the appropriateness of information flows, which 2) is defined by contextual norms governing particular settings (contexts) in which information is transmitted (Nissenbaum)
- **Informed consent** Is not practicable, because privacy information is either comprehensive or comprehensible but not both (Nissenbaum)

Predecisional

IDA | SCIENCE AND TECHNOLOGY POLICY INSTITUTE

Context on NAIRR Security Requirements and Access Controls

Emily Grumbling

Morgan Livingston

Lisa Van Pay

December 13, 2021

Science and Technology Policy Institute

1701 Pennsylvania Ave, NW ● Suite 500 ● Washington, DC 20006-5825

The NAIRR Roadmap must include:

> "An assessment of security requirements associated with the National Artificial Intelligence Research Resource and its management of access controls"
>
> - P.L. 116-283 5106(b)(F)

Goal: Frame key issues for TF consideration

1) Security Threats, Risks, and Tradeoffs
2) Security Requirements & User Access Controls
3) Implementation of Security Requirements in a NAIRR

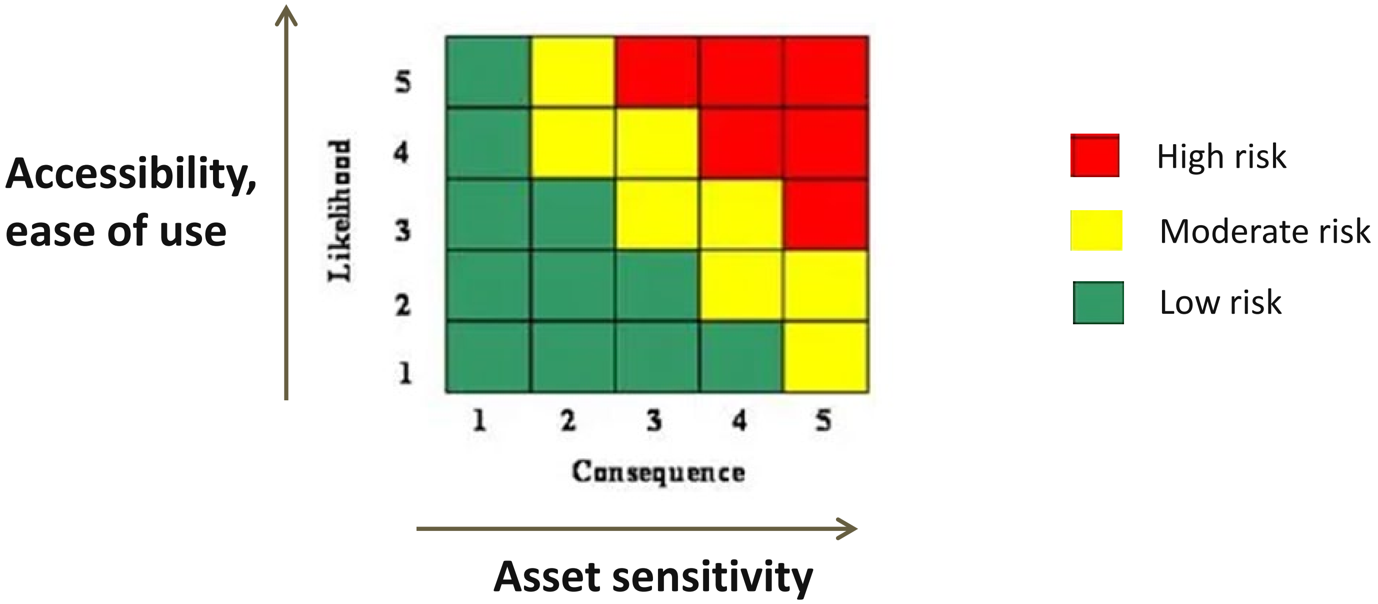
The NAIRR will face threats to confidentiality, integrity, and availability of assets/resources

Standard cybersecurity threats

AI/ML-specific threats

Potential consequences

Security is about understanding and managing risk



	1	2	3	4	5
5	🟩	🟨	🟥	🟥	🟥
4	🟩	🟨	🟨	🟥	🟥
3	🟩	🟩	🟨	🟨	🟥
2	🟩	🟩	🟩	🟨	🟨
1	🟩	🟩	🟩	🟩	🟨

🟥	High risk
🟨	Moderate risk
🟩	Low risk

Breaches will be inevitable

- Plan for both mitigation and response/recovery
- Apply existing frameworks (NIST, Trusted CI, Five Safes)

NAIRR security requirements will depend on system assets, owners, and intended uses

Policies & governance

System & security architecture

Physical and personnel security

User access controls

Access controls involve policy, technical, and usability considerations

- Tiered access
- Least privilege approach
- Role Based vs. Attribute Based Access Control models
- Using third party sign in manager
- Multifactor authentication
- Ease usability by using university sign on credentials
- Require security training

Privacy and security preserving ML methods have limitations and tradeoffs

Method	***Tradeoffs***
Differential Privacy	Accuracy, fairness, usability
Synthetic Data	Requires design and curation
Federated Learning	Usability, insufficient to protect privacy on its own
Homomorphic Encryption	Computationally costly, limited computations supported
Garbled Circuits	Computationally costly, limited computations supported
Secret Sharing	Computationally costly, high communication overhead
Secure Processors	Costly, slow to set up, requires trusted hardware

- Likely not universal solutions
- Require expertise to implement and maintain
- NAIRR can support R&D on these and new methods

A federated resource model would require clearly defined roles and responsibilities

Management of resource components

Selection of partners

Security requirements for partnership

Liability management

Policy enforcement for NAIRR system

Validation of resource integrity

Consider implementing the NAIRR first as a pilot

The NAIRR TF can look to approaches taken by other research resources

Scientific Computing and Data Storage/Sharing

Open Science Grid (OSG)

Distributed high-throughput computing

eXtreme Science and Engineering Discovery Environment (XSEDE)

Federated high-performance computing

Sensitive Administrative Data and Analytics Tools

Coleridge Initiative Administrative Data Research Facility

Virtual platform for approved research access to government microdata in AWS GovCloud

Institute for Research on Innovation & Science (IRIS)

Clearinghouse/research platform for member university administrative data

User access controls are tailored for each resource

- *Access portal management*
- *Research asset ownership/operation*
- *User authorization/allocation decisions*
- *ID management/authentication*
- *Security operations*

NAIRR Security Questions

- What are the security tradeoffs of NAIRR components?
- What do acceptable protections look like?
- What will the core NAIRR entity be responsible for?
- How will NAIRR manage liability and compliance issues?
- How might NAIRR support R&D on AI and security?

Technical Integration – a conversation starter

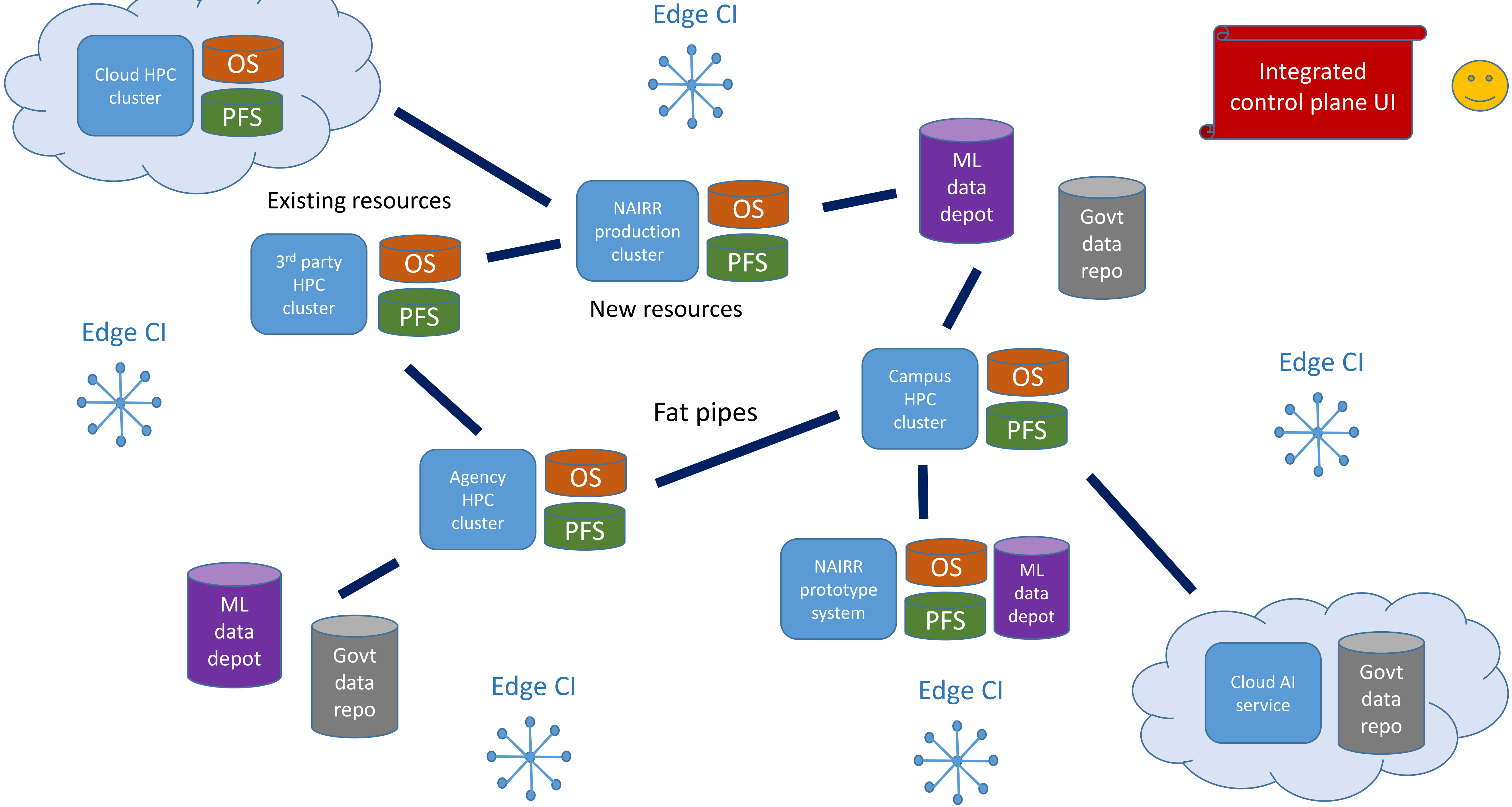
Mike Norman

UC San Diego

Going in assumptions

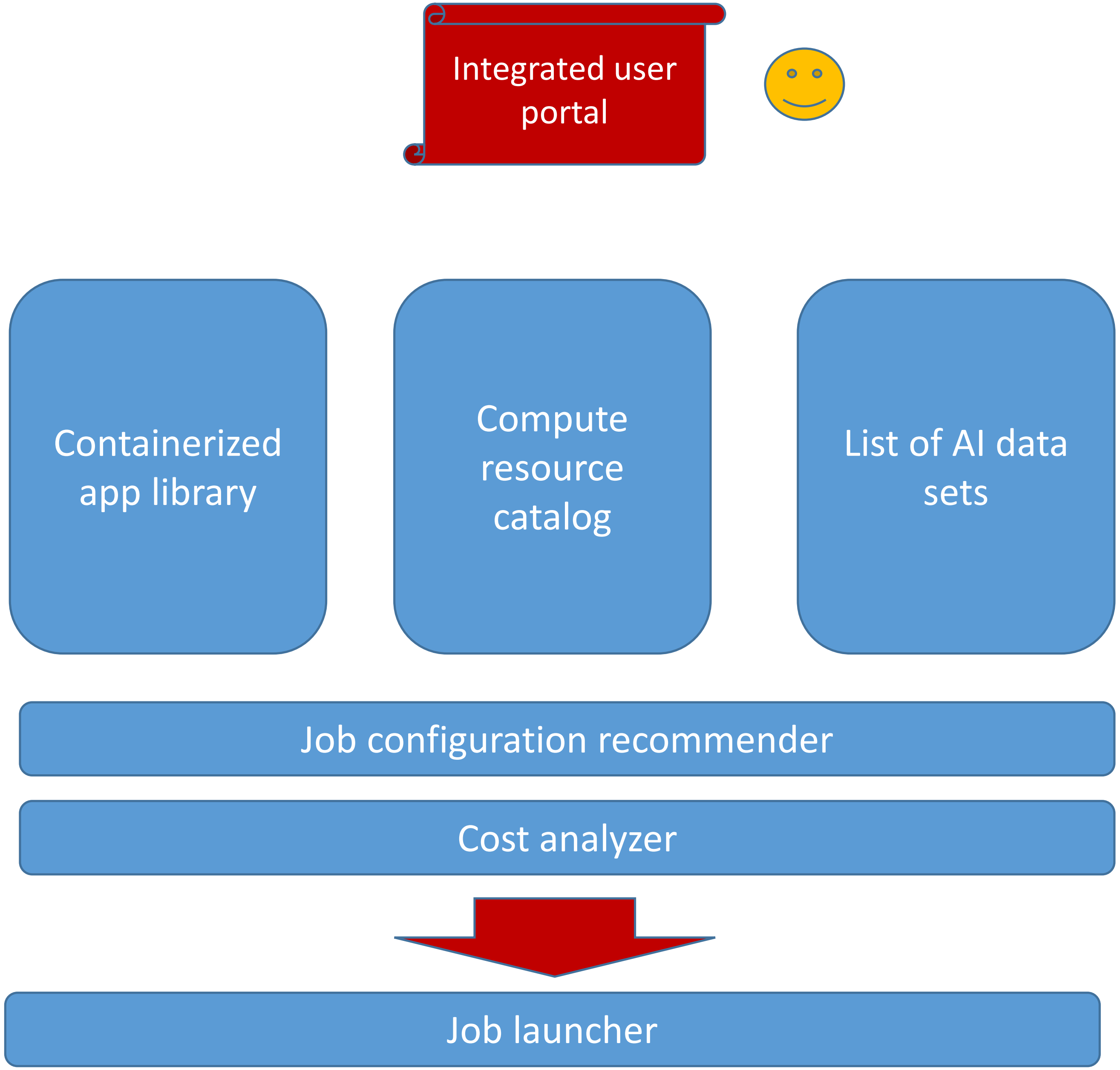
- NAIRR will be part of an integrated national CI ecosystem that is constantly evolving
- There may be one or more NAIRR-inspired resources (compute, data) in the ecosystem (production, experimental)
- NAIRR will leverage ongoing and planned federal agency investments in integrated CI and interoperate with them
- Other agencies/organizations may wish to contribute specific/unique resources to NAIRR (e.g., data) which will need to be integrated
- Integration of AI/ML data repositories and edge computing resources represent a new element to the CI ecosystem
- NAIRR users should be able to select their AI application, compute resource, and data source(s) from a list and launch and monitor jobs from a portal providing a uniform, integrated view with a minimum of effort

Predecisional



User experience

- User picks an app from a list
- User picks a compute cluster (real, virtual) from list
- User picks a dataset from a list
- Launches the job after approving cost estimate
- The portal working group has seen production commercial examples of this



Integrated user portal

Containerized app library

Compute resource catalog

List of AI data sets

Job configuration recommender

Cost analyzer

Job launcher

Integration topics

- Technical integration of compute resources
- Technical integration of data resources
- Technical integration of edge resources
- Technical integration of allocations and usage reporting
- Technical integration of training resources

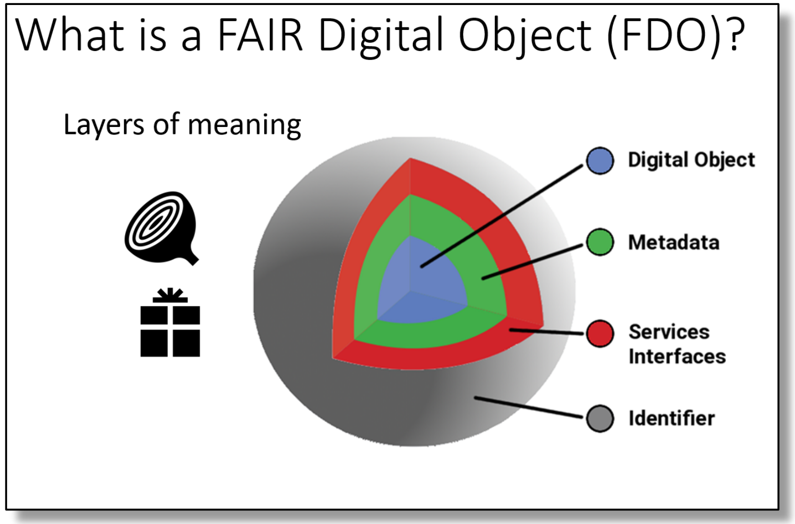
Predecisional

Technical integration of compute resources

- Abstraction: job scheduler locates and acquires compute nodes in a cluster, checks allocation balance, and executes job if OK. Can schedule interactive nodes.
- Current status: mature, in production (dozens of options)
- Examples:
  - XSEDE, OSG, DOE HPC centers, public cloud HPC services
- Key integration points
  - Allocation database
  - User authentication
  - Pre-installed applications
  - Access to high performance storage
- S-O-A
  - CILogon (XSEDE, CloudBank)
  - flexible Slurm+Kubernetes cluster partitioning for hybrid HPC+AI workloads (Expanse)
  - scheduler abstraction + adapters

Technical integration of data resources

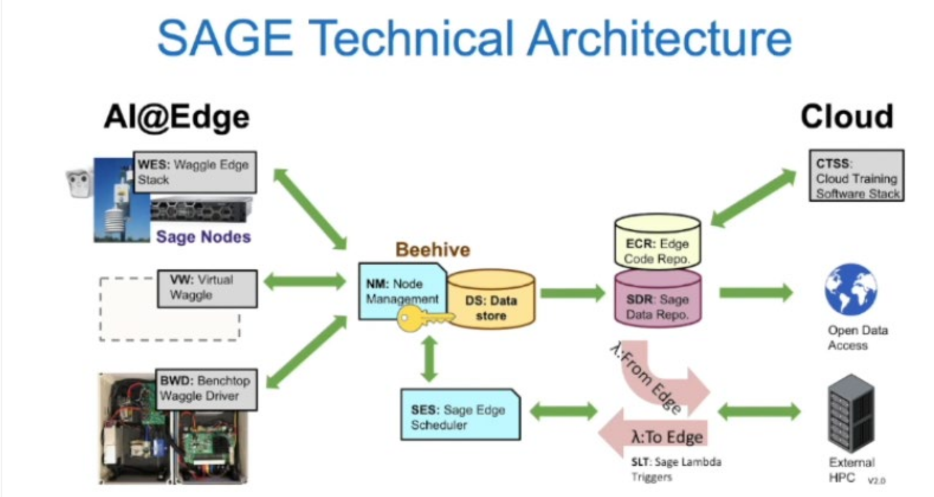
- The Dream: FAIR digital object located through search of online catalogs and moved from AI data depot to compute resource automatically by a “data delivery service”
- Current status: immature and fragmented
  - All done manually now
  - fledgling R&D efforts, demonstration projects
- Examples
  - MLCommons.org, DataCommons.org,
  - FAIR data points, FAIR DO forum,
  - Google data search, GeoCodes (EarthCube)
- Key integration points
  - Schema.org, S3 API, storage+network fabric, catalog search, storage allocations, access privileges, provenance/version metadata



What is a FAIR Digital Object (FDO)?

Technical integration of edge resources

- The Dream: Edge resources (sensors+accelerators) are discoverable, programmable, schedulable through a standardized “edge software stack” which also integrates with “continuum” (HPC+cloud) resources.
- Current status: R&D pilot projects, emergent commercial offerings
- Examples: SAGE, Cox Edge
- Key integration points
  - Edge device
  - Edge scheduler
  - Edge data repo
  - Edge code repository
  - Edge communications protocol
  - Edge-continuum integration
  - Composable systems

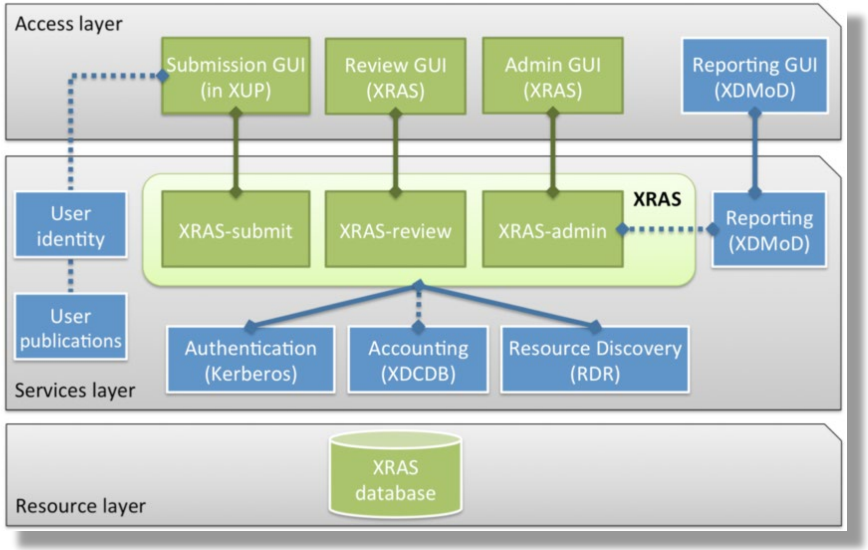


The architecture for the Sage project. Image courtesy of Pete Beckman.

Resource allocations and usage reporting

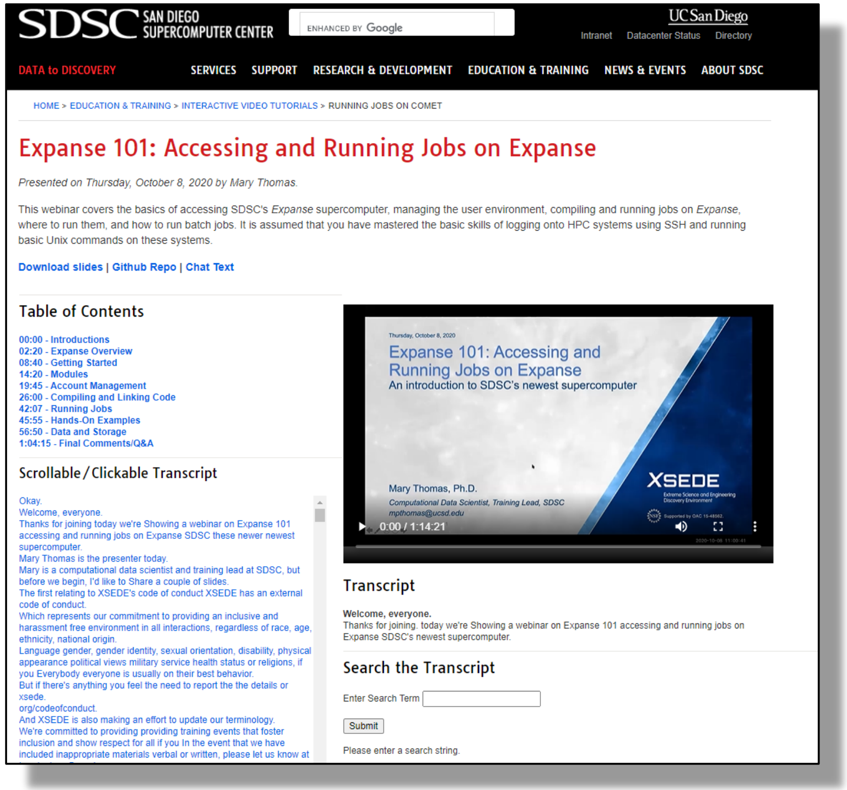
- Abstraction: financial accounting system that stores and displays account balances and updates balances due to deposits, withdrawals, and spending. Can be $, SUs, node-hrs, GB-yrs, or all the above.
- Current status: mature and in production
- Examples: XRAS, CloudBank
- Key integration points
  - Central database
  - Account management
  - User/group management
  - Resource discovery
  - Resource usage reporting
  - Reporting screens

XSEDE Resource Allocation System (XRAS)



Training materials technical integration

- Abstraction: *How do I use the resources?* training materials developed and curated by training experts at the resource provider sites, but centrally accessible through a user portal. *How do I use the portal?* Help system built into the portal.
- Current status: mature and in production
- Examples:
  - XSEDE user portal, CloudBank
  - Cloud provider training resources
- Key integration points
  - Curated training catalog
  - Self-paced tutorials
  - Web pages, github repos, YouTube
  - Searchable webinars



HOME > EDUCATION & TRAINING > INTERACTIVE VIDEO TUTORIALS > RUNNING JOBS ON COMET

Expanse 101: Accessing and Running Jobs on Expanse

Presented on Thursday, October 8, 2020 by Mary Thomas.

This webinar covers the basics of accessing SDSC's *Expanse* supercomputer, managing the user environment, compiling and running jobs on *Expanse*, where to run them, and how to run batch jobs. It is assumed that you have mastered the basic skills of logging onto HPC systems using SSH and running basic Unix commands on these systems.

Download slides | Github Repo | Chat Text

Table of Contents

00:00 - Introductions
02:20 - Expanse Overview
08:40 - Getting Started
14:20 - Modules
19:45 - Account Management
26:00 - Compiling and Linking Code
42:07 - Running Jobs
45:55 - Hands-On Examples
56:50 - Data and Storage
1:04:15 - Final Comments/Q&A

Scrollable / Clickable Transcript

Okay.
Welcome, everyone.
Thanks for joining today we're Showing a webinar on Expanse 101 accessing and running jobs on Expanse SDSC these newer newest supercomputer.
Mary Thomas is the presenter today.
Mary is a computational data scientist and training lead at SDSC, but before we begin, I'd like to Share a couple of slides.
The first relating to XSEDE's code of conduct XSEDE has an external code of conduct.
Which represents our commitment to providing an inclusive and harassment free environment in all interactions, regardless of race, age, ethnicity, national origin.
Language gender, gender identity, sexual orientation, disability, physical appearance political views military service health status or religions, if you Everybody everyone is usually on their best behavior.
But if there's anything you feel the need to report the the details or xsede.
org/codeofconduct.
And XSEDE is also making an effort to update our terminology.
We're committed to providing providing training events that foster inclusion and show respect for all if you In the event that we have included inappropriate materials verbal or written, please let us know at

Thursday, October 8, 2020
Expanse 101: Accessing and Running Jobs on Expanse
An introduction to SDSC's newest supercomputer
Mary Thomas, Ph.D.
Computational Data Scientist, Training Lead, SDSC
mpthomas@ucsd.edu
XSEDE
Extreme Science and Engineering Discovery Environment
Supported by OAC 15-48562.
2020-10-08 11:00:41

Transcript

Welcome, everyone.
Thanks for joining. today we're Showing a webinar on Expanse 101 accessing and running jobs on Expanse SDSC's newest supercomputer.

Search the Transcript

Enter Search Term

Please enter a search string.

Discussion prompts

- How should compute resources be denominated? Key integration decision vis a vis cloud versus on-prem.
- Data integration is the key challenge and essential for NAIRR success
  - How visionary should NAIRR be?
  - e.g., should NAIRR embrace/co-fund emerging standards like FAIR data objects?
  - Or, do we settle for the status quo (a hodgepodge)?
- Buy versus build decision.
  - Should NAIRR outsource the construction and operation of the user portal to a commercial entity?
- How can/should NAIRR drive R&D innovation in edge computing?
  - E.g., how much resource should NAIRR invest in edge computing software stack standardization and deployment?
- What are the technical building blocks for enabling accountability?