# Agenda

NATIONAL AI RESEARCH RESOURCE TASK FORCE: MEETING #3, October 25, 2021

| | |
|---|---|
| **11:00-11:10** | **Welcome and Administrative Remarks,** Erwin Gianchandani & Lynne Parker |
| **11:10-12:10** | **Readout and Discussion of Draft Recommendations: Compute Working Group,** Dan Stanzione |
| **12:10-1:10** | **Readout and Discussion of Draft Recommendations: Governance Working Group,** Fred Streitz |
| **1:10-1:40** | **Break** |
| **1:40-1:50** | **Scene Setter: Data Needs of the AI Community,** Daniela Braga |
| **1:50-2:50** | **Panel: Data resources,** Julia Lane<br>• Ian Foster, *Director, Data Science and Learning Division, Argonne National Laboratory*<br>• Robert Grossman, *Professor of Medicine and Computer Science, University of Chicago*<br>• Ron Hutchins, *Vice Provost for Academic Technologies, University of Virginia*<br>• Anita Nikolich, *Director of Research and Technology Innovation, School of Information Sciences, UIUC*<br>• Nancy Potok, *CEO, NAPx Consulting; former Chief Statistician of the United States*<br>• Andrew Trask, *Leader, OpenMined* |
| **2:50-3:10** | **Briefing: Testbeds as Component of the NAIRR,** Emily Grumbling & Lisa Van Pay |
| **3:10-3:40** | **Discussion: Testbeds,** Andrew Moore |
| **3:40-4:00** | **Break** |
| **4:00-4:50** | **Panel: User Resources – portal interface, educational tools,** Fei-Fei Li<br>• Tiziana Ferrari, *Director, EGI Foundation*<br>• Kimberly Greene Starks, *Global Lead, Infrastructure and Technology Strategy, IBM University Programs*<br>• Ana Hunsinger, *Vice President for Community Engagement, Internet2*<br>• Ed Lazowska*, Professor & Chair Emeritus, Paul G. Allen School of Computer Science & Engineering University of Washington* |
| **4:50-5:00** | **Questions from Public,** Erwin Gianchandani |

# Compute Resources Working Group

NAIRR TF – Interim Report Summary

# Consensus Items

- We began with the following guiding principles:
  - The NAIRR should support both Research on AI (the “Science of AI”) and Research Using AI (“AI for Science”).
  - The NAIRR should support the goal of giving US Researchers the *capability* to run the largest research problems in AI, while balancing the resources to also provide broad *capacity*, accessibility, and usability to as many researchers as possible.
  - The resource should be a federation of various compute (and data) resources, both hardware and software, and a mix of both production and experimental resources.
- We then realized even that much conversation was impossible without first better defining our terms. . .
  - …so we developed some shared definitions, and we’d encourage the task force to include/adapt these definitions for the interim report.

# Some Definitions to Clarify Recommendations

- ***Research Using AI**:*
  - AI techniques are becoming commonplace in the scientific and engineering computing workflow, and in many other applications.
  - We define "Research Using AI" to mean any R&D work that incorporates the tools and techniques of AI in producing results.
    - This may include employing trained networks, training networks to solve new problems, and any other *application* of AI that advances research, but does not directly advance the state of knowledge about AI itself (although it is likely to prove useful for motivating new directions in "Research on AI").
  - Also known as "AI for Science" or "Application-driven AI."

- ***Research on  AI***:
  - The methods, frameworks, and tools of AI are constantly evolving.
  - We define "Research on AI" as basic research that advances scientific understanding of
    - the nature of intelligence, mathematical understanding of the behavior of adaptive/autonomous systems or algorithmic understanding of techniques in the component areas of AI (which include perception, learning, planning, and robotics),
    - as well as support research into robustness, reliability, safety, security, privacy, interpretability, and equity of AI systems.
  - Translational research, such as applying Deep Learning methods to new problems, would not fall in this category.
  - Also known as "Science of AI."

# Some Definitions to Clarify Recommendations(2)

- *Production System/Resource:*
  - A production computing system or resource is defined as one in which users can run current state-of-the-art software tools and frameworks without modification and with reasonable expectations of stability and reliability.
    - Users of the system should reasonably expect accurate documentation on how to execute common use cases.
    - A system in production should be expected to have passed a set of pre-defined acceptance tests which measure performance, usability, and stability of the environment.
    - Note in this case it is the *Resource* that is production, though experimental research may be running on it.

# Some Definitions to Clarify Recommendations(3)

- *Experimental System/Resource:*
  - An experimental system or resource is exploring a new hardware or software capability, and may provide an immature or rapidly evolving environment for the user to run in.
  - Users may expect additional efforts to port applications to properly use the capabilities of the system, rather than a “turnkey” environment.
  - Changes to the system may be deployed quickly and with limited warning, and not all use cases may be well-supported.

# Recommendation #1: Nature of the Resources

- *Context: Technologies will evolve rapidly, so making vendor- or product-specific recommendations is not appropriate at this time.*
- NAIRR should consist of a *federation* of compute resources:
  - (A) Mix of Production and Experimental (Best Practice and Innovative)
  - (B) Mix of “Shared” (Commercial Cloud) and “Owned” (On-premise, at academic or government sites).
  - (C) Mix of “Core” and ”Edge” Computing Resources
  - (D) Balance of Capability and Capacity
  - (E) Co-Located with Traditional Scientific Computing Resources
  - (F) Co-Located with Data Resources, and Sufficient Network Capacity
  - (G) Zones with varying levels of security, some suitable for public data, some suitable for private data

# Recommendation #1: Discussion

- Mix of production resources:
  - Significant fraction with conventional servers with significant accelerator (e.g. GPU capability).
  - Many/most with significant high performance interconnect to support Model-Parallel runs which must span servers.
  - Combine with traditional HPC resources for mixed use cases.
- System Software:
  - Make available several “flavors”:
    - Bare metal/basic VM for the most flexible use cases.
    - Instances provisioned with a “NAIRR Application Stack” of e.g. Tensorflow, Pytorch, MX, Scilib, Numpy for normal “production runs”
    - API or “Serverless” compute – services provisioned and run by NAIRR for composable science workflows – examples might include NLP translation, OCR, Knowledge Graphs, API access to common datasets.

# Recommendation #1: Discussion

- Mix of resource homes:
    - Use Commercial Cloud for scalable capacity runs (e.g. maximum scalability for concurrent sessions) – Federate access with a Cloudbank-like model.
    - Mix in a federation of government-procured (academic/gov lab) resources, along the lines of the XSEDE Service Provider model (including procuring both experimental and production systems).
    - Run competitions regularly to award these services!
- Create a model to federate compute equipment not owned by NAIRR directly.
    - Institutional clusters or other resources.
    - In particular, enrolling edge resources – Make the NAIRR a development site for edge/datacenter hybrid applications (with sufficient dev hardware), but have most edge capability come from federating user-supplied hardware.
- In general, new persistent services developed by users should not have a *permanent* home within NAIRR.
    - However, some “promotion/graduation” process should be possible where user-created services can become part of the persistent infrastructure.

# Recommendation #2: Deployment, Scale, and Usage Models

- (A) Resources should be deployed in a phased model.
  - Not all resources acquired in the same year.
  - Periodic solicitations for new resources – new resources should come online, and old ones be retired, on a rolling basis.
    - Constantly keep "cutting edge" technology in the mix
    - Create continuity for the user community
- (B) The scale of the resources should be determined in two ways:
  - Capability – What is the largest single problem a researcher can tackle?
    - Goal: No researcher in the world should be able to run a larger model than the largest user of NAIRR.
    - Proxy: Training a nextgen GPT-scale model.
  - Capacity – How many simultaneous "typical" users/problems can the resource support?
    - Goal: With reasonable wait time, every STEM faculty/student can access a single node interactive session on the resource.
    - Proxy: Up to 10k sessions concurrently, while still supporting some capability runs.
- (C ) The operating model should support federation of user/customer-supplied computing/sensors at the edge
    - E.g. A testbed of control towers for autonomous drones flying wireless cameras should be able to interface with NAIRR to process data, even though NAIRR may not purchase the drones, the cameras, or the edge communication/compute devices.

# Recommendation #3 - Metrics

- The performance and impact of NAIRR (compute resources) should be measured in multiple dimensions:
    - (A) Countable metrics of utilization/cost/efficiency
        - # of users, % utilization, average turnaround time, # of successful jobs, cloud cost equivalent, energy efficiency etc.
        - Rankings on MLPerf/MLCommons other benchmarks
    - (B) Countable metrics that give proxies of science impact
        - # of publications, citations, patents by users, # of datasets published, datasets accessed/downloaded/reused, SW artifacts created, downloaded, reused, etc.
    - (C) Less countable metrics of transformative impact
        - Change in typical researchers workflow/scale of work
        - Adoption of produced artifacts
        - Adoption rate of new disciplines
        - Shifts in H-index of US vs. World AI publications

# Governance Model for the NAIRR

Draft Deck for Content Development

Governance WG Meeting

October 19, 2021

# This Working Group is charged with proposing answers to the following:

1. What is an optimal ownership and administration model for the NAIRR?
2. How should access to the NAIRR be governed?
3. What governance policies would need to be developed by the NAIRR?
4. What governance structures should be set up for the NAIRR?

# 1. What is an optimal ownership and administration model for the NAIRR?

## Discussed merits of various ownership models

| Approach | Pros | Cons |
|---|---|---|
| Public-private-partnership or consortium | Value driven<br>Flexible<br>Transparent<br>Sustainable | May not provide long term capacity building<br>Needs clearly defined governance/authorities<br>IP is challenging |
| NSF-style center awarded to university | Existing infrastructure for grant application and management<br>Focus on students and researchers at higher ed institutions | Little history in managing large-scale IT and data infrastructure, incl. usability and ease of access<br>No clear way of hiring needed support staff |
| New division or element under existing govt agency | Clear ownership and authority<br>Continuity | Narrow scope of single agency can introduce bias<br>Cumbersome contracting |
| | | |

# 1. What is an optimal ownership and administration model for the NAIRR?

## Recommendation 1: NAIRR would be a separate entity operated as a Federally Funded Research and Development Center (FFRDC) supported equally by DOE, NIH, and NSF (perhaps others as well)

- FFRDCs are designed to provide government with sustainable and persistent capacity to address long-term problems free from conflict of interest
- FFRDCs “operate in the public interest with objectivity and independence”
- FFRDCs have duty to accelerate commercialization for technology that is developed, collaborating (not competing) with industry
- Single entity with clear responsibility and well-defined authority owns mission space
- Reports to Board with broad representation to ensure appropriate scope

# 2. How should access to the NAIRR be governed?

**We discussed some basic principles first:**

- AI researchers go through proposal driven process with rapid reviews, relatively light-weight where possible
  - Possibly tiered proposal/reviews depending on the nature/amount of resources requested
  - Access should be as inclusive as possible
  - Special attention/consideration to underserved research communities across the country
  - Tiered access–some access heavily subsidized or entirely free, some access involves a fee structure
  - Organization can steer effort through RFP process
- Reinforcement/incentive mechanisms: those who have contributed to the public good get increased/priority access (e.g., offering curated data for all– consider 'leader board' approach)
- Consider also "pay-to-play" model for data providers, wherein entities pay to have their data included in Resource
  - Data providers gain insights from data as community makes use of it
  - Incentivizes data owners to clean/label/validate their own data, as usability becomes a value
  - Defines a source of income – could lead to financial sustainability
  - Organizations that do both can earn credit for the AI they contribute
- Who are the target users?
  - Depends upon desired impact– rejuvenate entire AI ecosystem, increase diversity/broaden participation in AI (Both goals are important– how should the NAIRR balance these goals?)
  - As many as possible should be able to contribute and leverage data

# 3. What governance policies would need to be developed by the NAIRR?

**Some fundamental principles will need to be addressed in the NAIRR Charter or developed shortly thereafter**

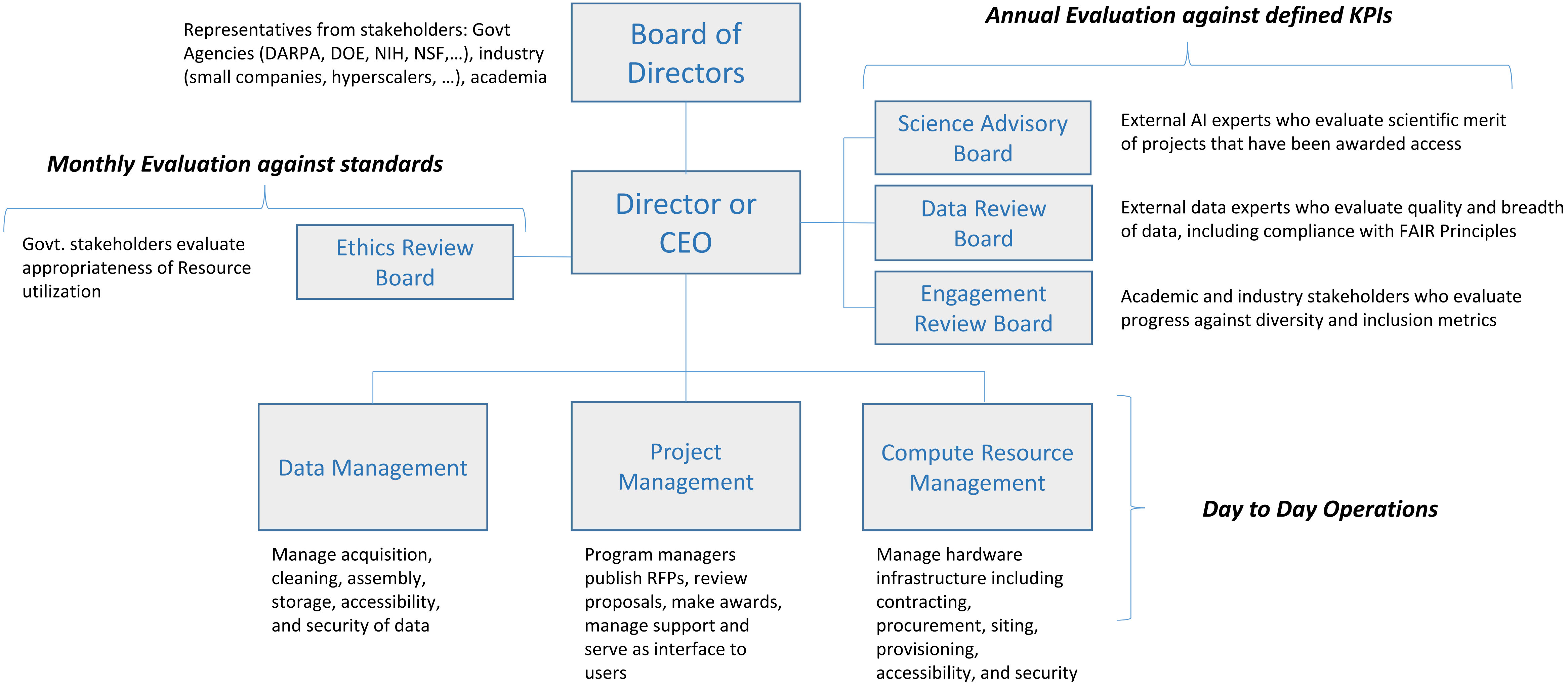
- Access rights
- Transparency and trust
  - Code of conduct, including COI
- Legal authority to protect privacy and confidentiality
- Independence
- Scalable functionality (Technical policies/standards to enable this)
- Policies that foster resource sustainability (financial)
  - Pay-to-play for certain participants? Exceptions for nonprofits, etc.
  - Incentives (see prev slide)
  - Data providers v. users
- Oversight and accountability
  - Legal/regulatory compliance
  - Policies for advisory board authority/composition/term of service
- Intergovernmental (multi-agency, state, Federal, local) support
  - Multiple agencies can support the activities
- Data policies (handling, metadata, curation)

# 4. What governance structures should be set up for the NAIRR?

**Initial thoughts:**

- Organization should have clear leadership, with defined responsibilities and authority
- Organization should be advised/overseen by board comprising the multiple stakeholders (including government, industry and academia)
- Organizational structure must support delivering an operational capability at increasing scale
- Organization should be set up as a pilot.  Even in final form, it should be able to be shut down if it fails
- To reach underserved community requires more than just access: technical support will be needed as well
    - Leads potentially to staffing/expertise issue as problem sets expand
    - Could envision a lean structure for NAIRR (resource management, logistics, awards and administration) with subject matter expertise delivered by Centers of Excellence throughout the country

# Possible Recommendation 2a: Governance Structure



Representatives from stakeholders: Govt Agencies (DARPA, DOE, NIH, NSF,…), industry (small companies, hyperscalers, …), academia

Board of Directors

Director or CEO

***Monthly Evaluation against standards***

Ethics Review Board

Govt. stakeholders evaluate appropriateness of Resource utilization

***Annual Evaluation against defined KPIs***

Science Advisory Board

External AI experts who evaluate scientific merit of projects that have been awarded access

Data Review Board

External data experts who evaluate quality and breadth of data, including compliance with FAIR Principles

Engagement Review Board

Academic and industry stakeholders who evaluate progress against diversity and inclusion metrics

***Day to Day Operations***

Data Management

Manage acquisition, cleaning, assembly, storage, accessibility, and security of data

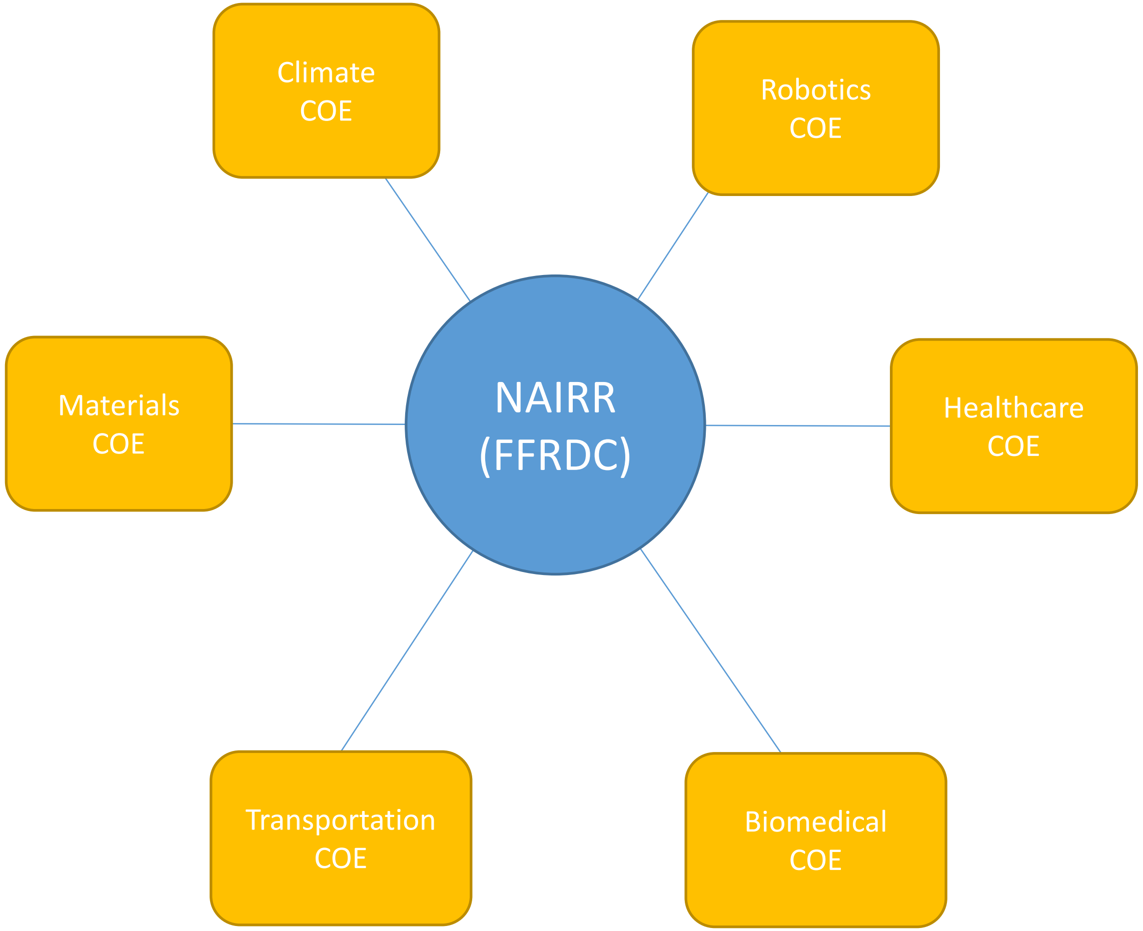
Project Management

Program managers publish RFPs, review proposals, make awards, manage support and serve as interface to users

Compute Resource Management

Manage hardware infrastructure including contracting, procurement, siting, provisioning, accessibility, and security

# Possible Recommendation 2b: Governance Structure



**Hub-and-spoke model for AI support in topical areas**

- NAIRR issues RFP on specific topic and awards COE to host institution
- COE provides support to NAIRR users in topical area, leveraging scientific expertise at the site
- COE may provide data or infrastructure
- Scalable model allows for increasing coverage of application space

***Example Centers of Excellence***

# Questions

**Several questions came up in our discussions whose answers would affect a governance model**

- What is the primary focus of NAIRR? To accelerate advances in AI for the nation? Or to address the challenges of AI in academia? (Note that these are not conflicting – we assumed the answer was "both")
- Related question: who is the target audience of users? AI researchers in the country, with a focus on underserved/underrepresented communities? Or is NAIRR a resource specifically for underrepresented AI researchers? (We assumed that reaching the broadest possible audience was a goal.)
- Is NAIRR seen as primarily a platform for providing necessary compute? Or as a platform for providing curated data sets? (Again, these are complimentary, and we assumed "both")

**The narrower the definition of focus and audience for NAIRR, the more specialized the governance structure that could emerge. We choose FFRDC as a flexible structure that could successfully accommodate any or all of the answers to these questions**

# Data Needs of the AI Community

Daniela Braga

Founder and CEO of DefinedCrowd

# We live in a world of **big** data

“The world produces 2.5 quintillion bytes a day, and 90% of all data has been produced in just the last two years.”

The World Economic Forum, The value of data | World Economic Forum (weforum.org), 2017

“463 EB* of data will be created in 2025 every day”

* 1 EB 1 billion bytes

# 80%

**of the data is unstructured and most big organizations don’t have neither the tools nor the talent in house to make sense of the data they produce.**

SMEs and citizens use often **third-party services** that collect and monetize their data, until recently without their consent. But they still opt in because **building AI is still expensive and inaccessible** for most of the world.

# The AI value chain and SMEs needs

- ✓ Data ✓ Tools ✓ Models
- ✓ Responsible AI
- ✓ Data Management
- ✓ Cloud

# Data Needs of the AI Community

## Recommendations

- Open access to trusted data per industry via Marketplaces or similar
- Create data standards to facilitate interoperability
- Certification on responsible & ethical AI
- AI literacy
- Data transparency
- Access to minority-related datasets
- Standards for data privacy

IDA SCIENCE AND TECHNOLOGY POLICY INSTITUTE

# Considerations for AI Testing Resources that could be Accessible via the NAIRR

Lisa Van Pay

Morgan Livingston

Emily Grumbling

October 25, 2021

**Science and Technology Policy Institute**

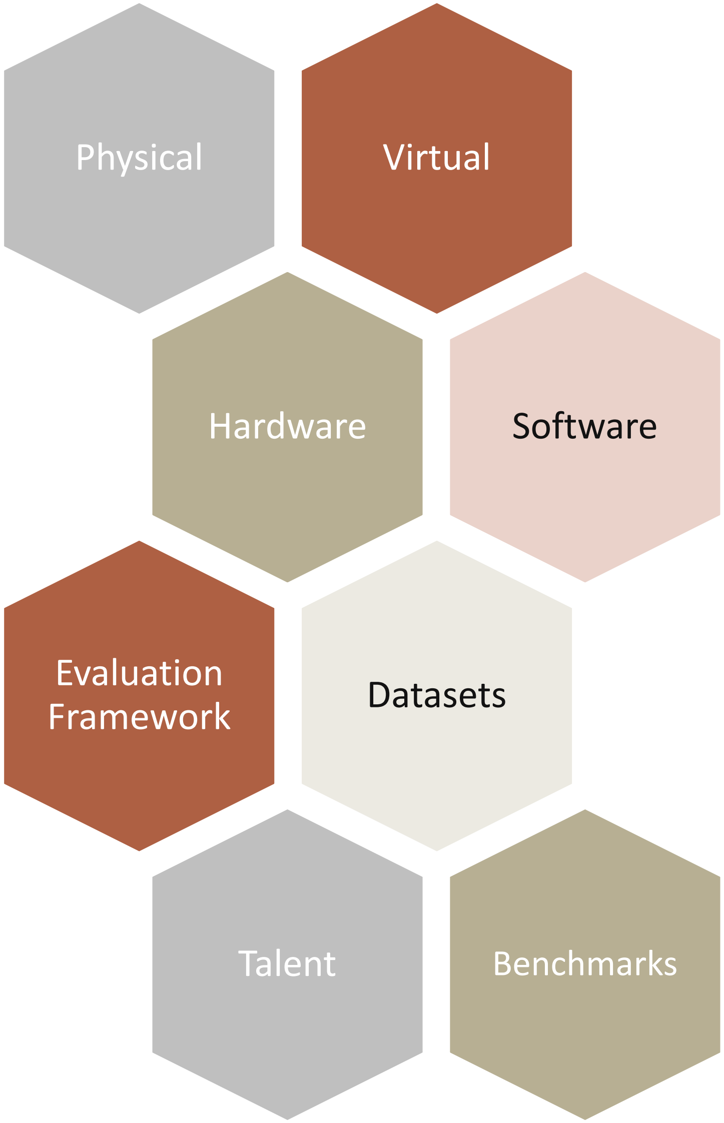
1701 Pennsylvania Ave, NW ● Suite 500 ● Washington, DC 20006-5825

# Outline and objectives

- Testing needs for AI R&D
- Existing testbeds
- Constraints and opportunities for integrating resources into NAIRR

Objective: Provide an overview of different types of AI tests and testbeds that could be associated with the NAIRR, and corresponding advantages or constraints.

# Testing and testbed resources are needed to improve AI performance and enable prototyping and deployment



***AI tests*** – Datasets and evaluations that test the performance of an AI algorithm against specified parameters and tasks.

***AI testbeds*** – Simulated, live, or blended environments that support prototyping, development, and testing of AI.

# To understand the AI testing landscape, we examined priorities for AI testing resources and existing testbeds

- Reviewed AI R&D priority documents for information on AI performance and testing
- Analyzed a set of 79 example testbeds or test resources
- Examples included private sector, academia, and federal government
- Not intended to be comprehensive

# Testbeds for fundamental AI can drive specific lines of innovation and advance trustworthy AI

## *Benchmarks to evaluate and compare AI:*

Model performance on datasets for specific use cases

Reinforcement learning performance in testbeds

Efficiency of hardware and software configurations

## *Testbeds and testing tools to develop responsible AI:*

Privacy

Fairness

Security

Explainability

# Testbeds for use inspired AI should be robust, integrated systems that align with priority use cases

Testbed types

Physical

Blended

Virtual

Used for

AI R&D

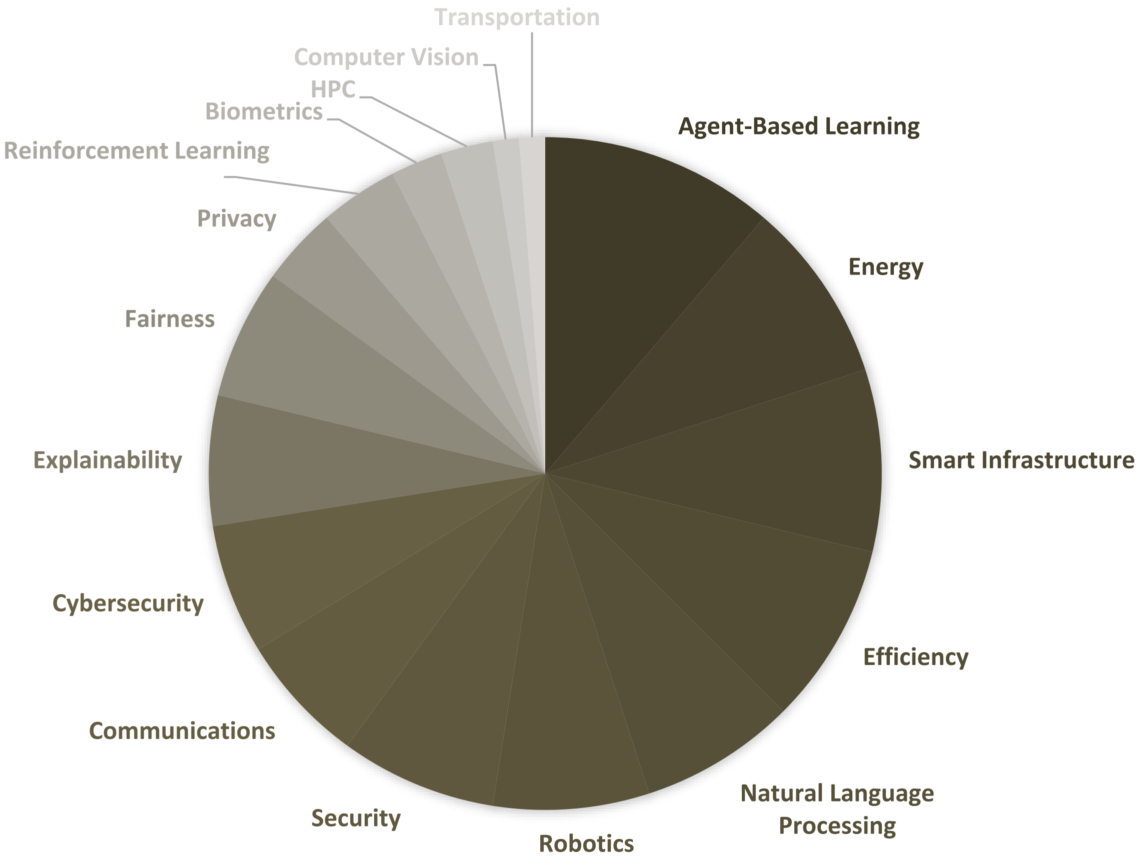
Interdisciplinary research & training

Data generation

Use cases

# A wide range of testbeds currently exist

*Profile of testbeds and tools STPI examined. These resources support overlapping and mutually reinforcing research areas:*



…with different design, management, and use case specific constraints.

# Characteristics of a testbed have implications for incorporation into the NAIRR

## Open source

*How to maximize benefits of testbeds and leverage community contributions, while protecting research?*

## User access

*How to determine which users can access testbeds? How to lower barriers to use?*

## Intellectual property

*How to respect the IP protection of the testbed, track testbed use, and protect testbed outputs?*

## Community building

*How to build a connected, multidisciplinary community using shared testbed resources?*

## Data sensitivity

*What security and privacy measures might facilitate use of sensitive, use-case specific data?*

## Evolving to stay current

*How can funding and design support connected testbeds staying up to date with user needs?*

# Testbeds can support and advance NAIRR objectives and provide opportunities to innovate

Workforce and diversity

Access to datasets

Ethical principles of AI

Community-driven challenges

# Summary and next steps

- Many testbeds and testing tools already exist
- What incentives are needed for testbed owners?
- Accessing testbeds via the NAIRR may involve limitations or tradeoffs associated with:

  -User access requirements

  -Usability

  -IP policies

  -Accessibility