Agenda

NATIONAL AI RESEARCH RESOURCE TASK FORCE: MEETING #2, August 30, 2021

11:00-11:10	**Welcome and Administrative Remarks,** Erwin Gianchandani
11:10-11:35	**The National AI Initiative and a baseline vision for the NAIRR,** Lynne Parker & Erwin Gianchandani
11:35-12:35	**Panel: Value Proposition and Intended Outcomes of a NAIRR** • Damian Clarke, *Chief Information Officer and Computer Science Faculty, Alabama A&M University* • James Deaton, *Executive Director, Great Plains Network* • Deborah Dent, *Chief Information Officer, Jackson State University* • Tripti Sinha, *Assistant Vice President and Chief Technology Officer, University of Maryland and Executive Director of MAX* • Talitha Washington, *Director, AUC Data Science Initiative*
12:35-1:00	**Discussion: Defining the value proposition and intended outcomes of a NAIRR,** Lynne Parker
1:00-1:30	**Break**
1:30-1:50	**Presentation: Ownership, governance and administration options,** Emily Grumbling & Lisa Van Pay
1:50-3:00	**Panel:** • Sharon Broude Geva, *Director for Innovation and Computational Research, University of Michigan* • Manish Parashar, *Office Director, Office of Advanced Cyberinfrastructure, National Science Foundation* • Gina Tourassi, *Director*, *National Center of Computational Sciences and the Oak Ridge Leadership Computing Facility, ORNL* • John Towns, *National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign* • Frank Würthwein, *Interim Executive Director, San Diego Supercomputer Center*
3:00-3:30	**Break**
3:30-4:30	**Discussion: Compute Capabilities,** Dan Stanzione
4:30-4:45	**Working Group Expectations,** Lynne Parker
4:45-5:00	**Questions from Public,** Erwin Gianchandani

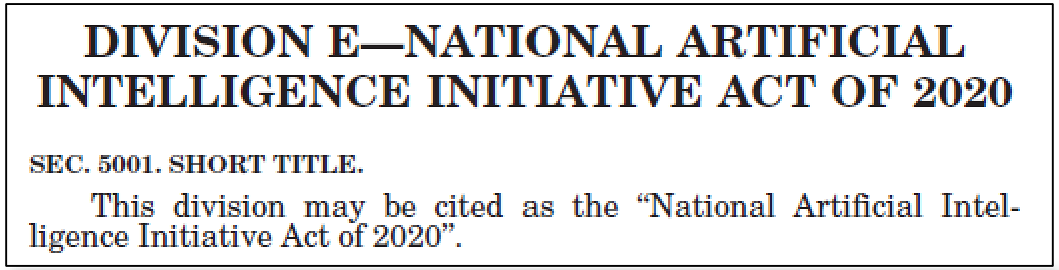
National AI Initiative

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

National AI Initiative Act of 2020 (NAIIA)

Became law on January 1, 2021

As part of the "*William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021",*
H.R. 6395, Division E.



DIVISION E—NATIONAL ARTIFICIAL INTELLIGENCE INITIATIVE ACT OF 2020

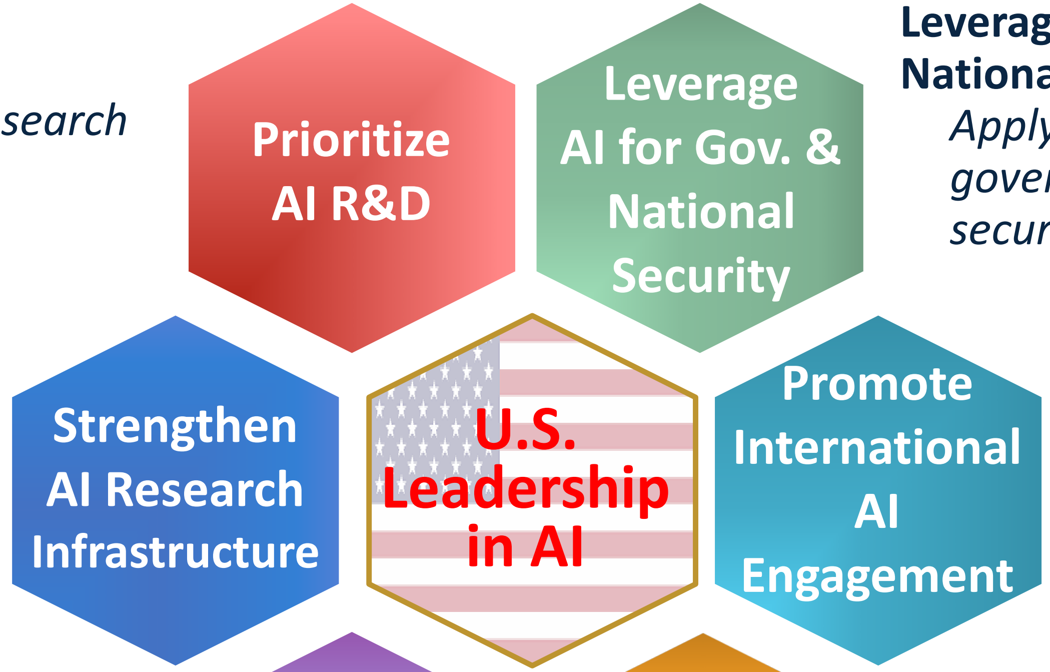
SEC. 5001. SHORT TITLE.

This division may be cited as the "National Artificial Intelligence Initiative Act of 2020".

Bipartisan legislation defining **National AI Initiative**, with purpose of:

- Ensuring continued U.S. leadership in AI research and development (R&D);
- Leading world in development and use of trustworthy AI systems in public and private sectors;
- Preparing present and future U.S. workforce for integration of AI systems across all sectors of economy and society; and
- Coordinating AI research, development, and demonstration activities among civilian agencies, Department of Defense, and Intelligence Community to ensure that each informs work of the others.

National AI Initiative


Prioritize AI R&D
Leverage AI for Gov. & National Security
Strengthen AI Research Infrastructure
U.S. Leadership in AI
Promote International AI Engagement
Advance Trustworthy AI
Train AI-Ready Workforce

Prioritize AI R&D
Grow and sustain U.S. research leadership and capacity

Leverage AI for Government and National Security
Apply AI to improve provision of government services and national security

Strengthen AI Research Infrastructure
Enhance access to high quality data, models, and computing resources

Promote International AI Engagement
Engage with like-minded allies to promote a global AI environment supportive of democratic values

Advance Trustworthy AI
Modernize governance and technical standards for AI-powered technologies, protecting privacy, civil rights, civil liberties, and other democratic values

Train AI-Ready Workforce
Provide AI-ready education at all levels: K-12, college, re-training, re-skilling, R&D workforce

National AI Advisory Committee

Will advise President and National AI Initiative Office on:

- State of U.S. competitiveness and leadership in AI
- Progress made in implementing Initiative
- State of AI science
- AI and U.S. workforce issues
- How to leverage Initiative resources to streamline and enhance government operations
- Need to update the Initiative
- Balance of activities and funding across Initiative
- Whether strategic plan is helping U.S. leadership in AI
- Management, coordination, and activities of the Initiative
- Whether ethical, legal, safety, security, and other societal issues of AI are adequately addressed by the Initiative
- Opportunities for international collaboration with strategic allies on AI
- Accountability and legal rights, including oversight
- How AI can enhance opportunities for diverse geographic regions of the U.S.

Framing the NAIRR Vision

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

What are the objectives of establishing a NAIRR?

- The strategic objective of a NAIRR would be to strengthen the U.S. AI innovation ecosystem by both (i) supporting fundamental AI research and (ii) increasing the number and diversity of AI researchers and organizations. It would do so by:
  - ➢Lowering barriers to entry
  - ➢Supporting innovative and novel efforts in AI research and the broad adoption of AI
  - ➢Reinforcing the viability of academic career paths in AI
  - ➢Advancing the development and training of the AI workforce

Why do we need a NAIRR?

- AI holds the potential to impact science, the economy, national security, and society
- Overcoming the “compute-divide”: today access to computational and data resources are primarily limited to the large private sector firms and well-resourced universities
- Expansion of access will broaden the diversity of researchers involved in AI, expanding approaches to and applications of AI

Fundamental Questions

- What are the metrics of success?
- Who are the intended users?
- How will access be adjudicated and finite resources allocated to a diverse group of users in an equitable manner?
- What capabilities will be provided?
- How will the resources come together to create the NAIRR?
- How will users access the NAIRR?
- How will the NAIRR be funded and managed?
- How will the NAIRR address concerns around the ethical and responsible development of AI?
- What are other associated issues?

SCIENCE AND
TECHNOLOGY
POLICY INSTITUTE

Considerations for NAIRR Governance and Administration

Emily Grumbling

Lisa Van Pay

Morgan Livingston

August 30, 2021

Science and Technology Policy Institute

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

Outline and objectives

- Legislative requirements
- Ownership
- Administration
- Governance

Objective: Develop a general understanding of different types of ownership, governance, and administrative options for the NAIRR, and associated advantages or constraints.

The National AI Initiative Act outlines elements that must be included in roadmap and implementation plan

(1) IN GENERAL.—The Task Force shall develop a coordinated roadmap and implementation plan for creating and sustaining a National Artificial Intelligence Research Resource.

(2) CONTENTS.—The roadmap and plan required by paragraph (1) shall include the following:

A. Goals for establishment and sustainment of a NAIRR, and metrics for success.

B. A plan for **ownership and administration** of the National Artificial Intelligence Research Resource, including

i. an appropriate agency or organization responsible for the implementation, deployment, and administration of the Resource; and

ii. a governance structure for the Resource, including oversight and decision-making authorities.

C. A model for **governance and oversight** to establish strategic direction, make programmatic decisions, and manage the allocation of resources;

(H.R. 6395 Sec. 5106(b)(2))

Input from the first NAIRR Task Force (TF) meeting and TF co-chairs was used to scope this research

- Reviewed past studies and examples of research resources
- Drawn largely from examples focused on HPC
- Identified range of options
- Not intended to be comprehensive

Research resources generally fall under one of 3 ownership categories

Federal government

Academic or private sector organization

Partnership or consortium

“[R]esponsibility and accountability for the implementation, deployment, and ongoing development of the National Artificial Intelligence Research Resource, and for providing staff support to that effort” (HR 6395)

Each type of ownership carries implications for use and management of the resource

- Federal government
- Academic or private sector organization
- Partnership or consortium

Funding

Staffing

Accountability

Intellectual property

New or existing infrastructure

Singular or federated resources

Elements of the resource itself carry implications for administration and governance

Constraints on use

Single sign-on

Remote or physical access

Sensitivity of data

New or existing infrastructure

Singular or federated resources

Access to limited resources can be managed via deliberate allocation methods

- Baseline user eligibility
- Prioritization by intellectual merit and broader impacts of research
- Standard units of allocation and usage caps
- First-come, first-served scheduling
- Fee-for-access

Open question: what allocation methods enable equitable access?

Governance structures are variable, often aligning with owner’s organizational structure

Centralized

Hierarchical

Phased

Decentralized

Representative

Evolving

Governance involves strategic planning, operational decision-making, and oversight

Resource governance spans a range of functions

Visioning & strategic planning

Leadership & decision-making

Advising

Coordination and communication

Technical design and operation

Oversight & accountability

Governance principles can be upheld and enforced through policies and documents

- Strategic documents
  - Charter or documents of incorporation
  - Vision, mission, and strategic or business plan
- Partner agreements
- End user agreements
- Code of conduct
- Technical standards & practices
- Legal, regulatory, and ethics policies

Built-in oversight tools and mechanisms support progress and accountability

Periodic evaluation

Regular surveys of users and staff

Design choices that enable metrics collection

Transparent reporting

Summary and next steps

- Many options for ownership, governance, and administration
- Design should be informed by decisions regarding:
  - Desired impacts of NAIRR
  - Target user base and research needs
  - Resource components included

National AI Research Resource Task Force: Compute Options

8/30/21

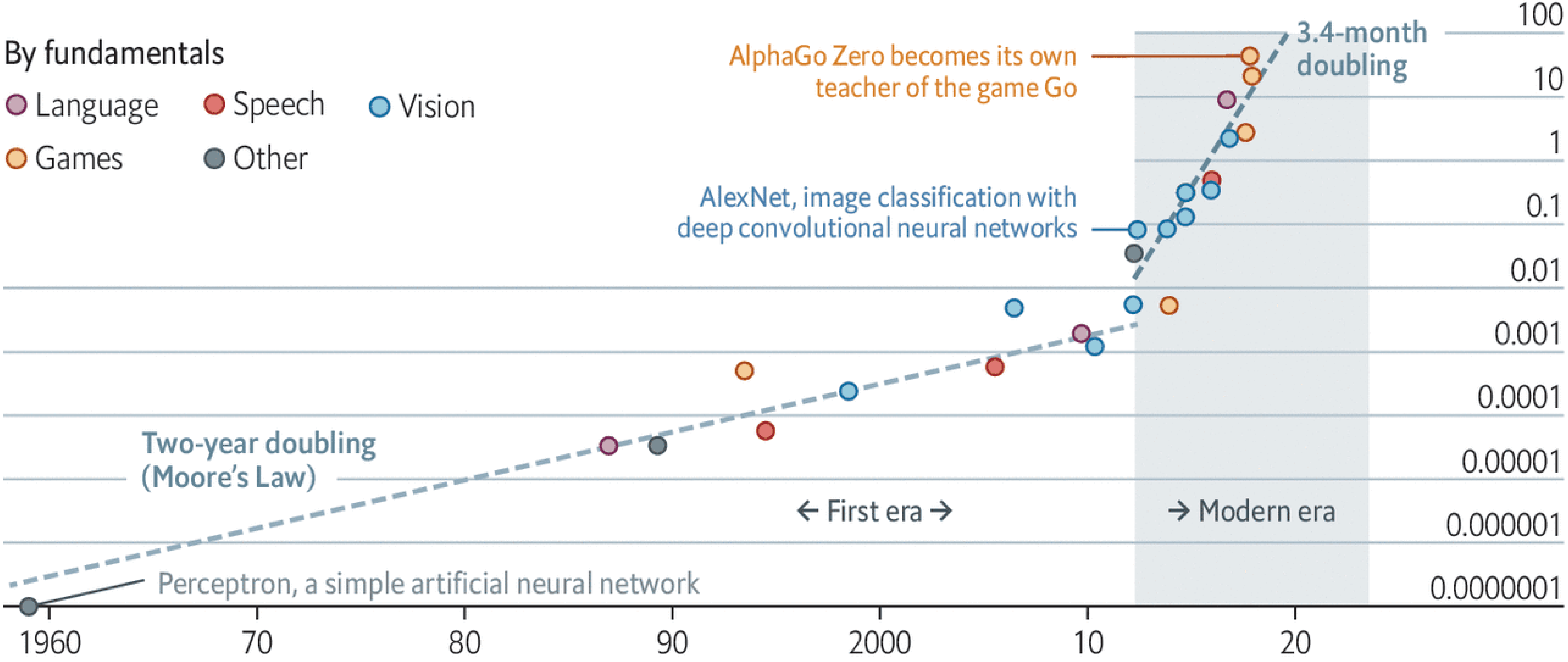
Compute Resource Working Group – Some Big Questions

- How do we determine what an “appropriate” size is?
- What do we do when that isn’t enough?
- How do we determine what the right mix of resources is?
    - Software/Workflow match to Hardware
    - Dedicated or not?
    - Testbed vs. Production
    - Who makes architecture decisions?
- Co-Location with Data
- Co-Location with Simulation/other Computing
- *What are our Metrics for evaluating all of these things???*

Deep and steep

Computing power used in training AI systems

Days spent calculating at one petaflop per second*, log scale



Source: OpenAI

*1 petaflop=10^{15} calculations

The Economist

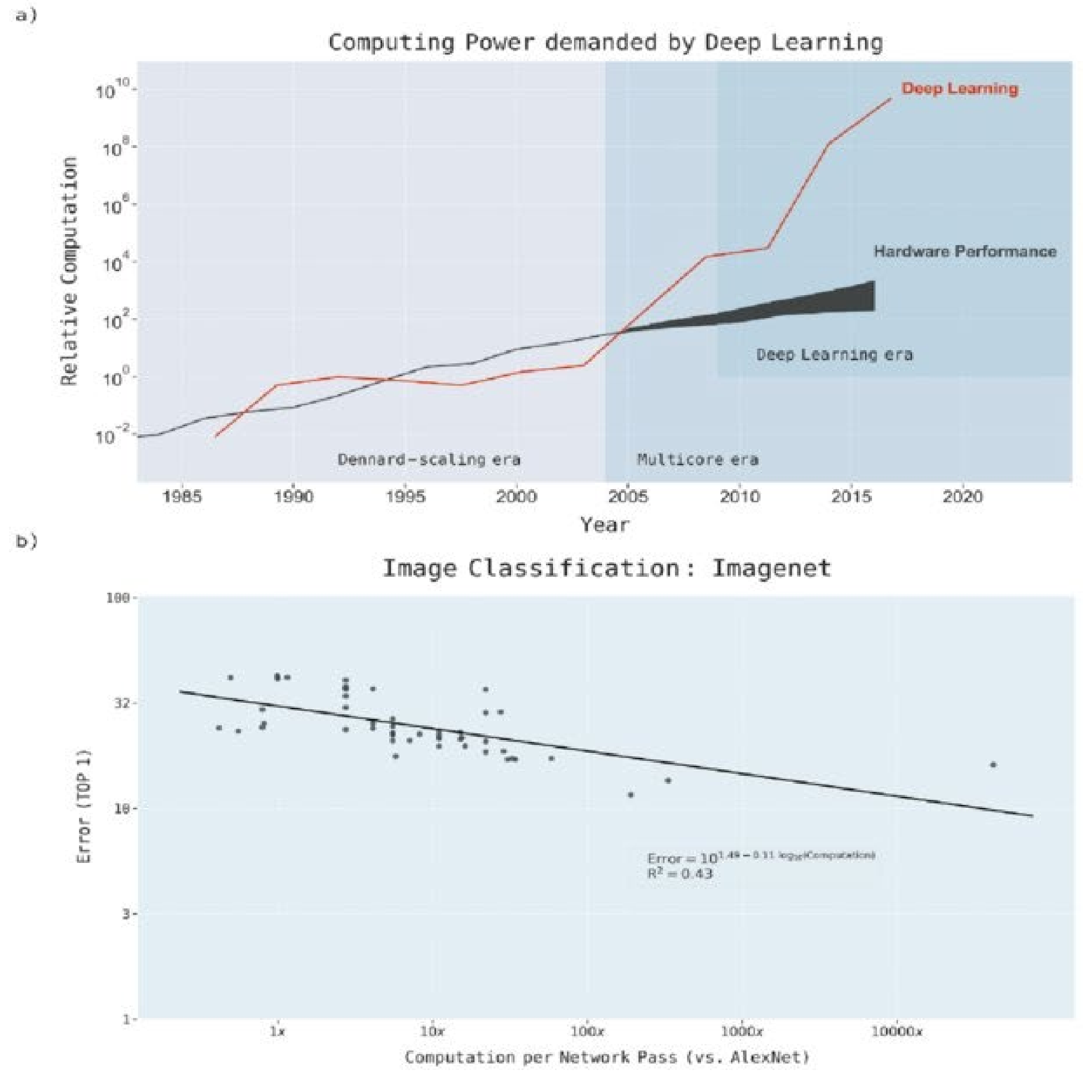


Figure 2: Computing power used in: **(a)** deep learning models of all types [4] (as compared with the growth in hardware performance from improving processors[23], as analyzed by [39] and [56])[8], **(b)** image classification models tested on the ImageNet benchmark (normalized to the 2012 AlexNet model [52]).

- From “The Computational Limits of Deep Learning” , Thompson et al

How do we determine what an “appropriate” size is?

- Determining the amount of computing for a given AI task is notoriously difficult.
  - Within Deep Learning in particular:
    - Inference is predictable.
    - Parameter size gives some idea of the *max* memory to train a model, and training data gives some notion of runtime length
    - But required training to converge the model to acceptable error rates is hard to predict.
- In other computing research infrastructures:
  - Allocations are made by number of compute hours required.
  - Past simulations are predictive of time required for future ones, so peer-review of the scope of requests is possible.
  - Codes are often well-known, so effectiveness of the requested time can somewhat be judged for peer review.
  - Anecdotally (within, for instance, NSF XSEDE) DL-oriented requests can’t be allocated via this process.
- There will be constraints of reasonable budget, but how do we estimate the size of the need? How can we translate that need into an amount of resource?
  - Estimating the need will probably require a better description of the audience/user base for the Resource.
  - Justifying the amount of demand properly will likely play a role in budget decisions.
- Operational model decisions may allow the Resource to grow.

What do we do when that isn’t enough?

- It is almost inevitable, barring budget miracles, that demand for the resource will outstrip supply.
- Do we provide the highest performing resources, and force users to adapt software/workflow?
- Or do we focus on usability, and perhaps sacrifice performance?
    - Do we worry about measuring user effectiveness and match to compute hardware?

How do we determine what the right mix of resources is?

- Software/Workflow match to Hardware
  - Not all AI runs equally well (or at all) on all platforms
  - How do we decide what tools/workflows to support? How heterogeneous do we want the resources to be?
    - Tradeoff scale for heterogeneity?
    - Tradeoff simplicity of use/programming for more initial compatibility?
- Dedicated or not
  - Dedicated Resources, or Multi-Tenancy?
    - How does data protection play into this? Secure/classified resources?
    - The "owned" model, e.g. National Labs LCF vs. the "shared" model, e.g. Commercial Clouds? Some Mix?

How do we determine what the right mix of resources is?

- Who makes architecture decisions?
- Testbed vs. Production?
  - There are numerous possible architectures for right now (GPUs, CPUs, FPGAs, and many kinds of those) and for the future (literally dozens of AI-specific chips coming online).
  - How do we decide what to deploy to do work ”right now”?
  - How much do we set aside for testbeds, experiments to let both the Resource and the AI Software stack evolve?
  - Is this centrally controlled? Do we open solicitations for specific kinds of resources, or broad calls and let proposing providers offer a mix of hardware solutions?

The Landscape of AI Computing Resources

- There are many types of AI that are *not* Deep Learning, but undoubtedly, DL is the dominant *computational* consumer at the moment.
  - (Let’s just stipulate that there are other modes to consider, and that methods will change over time).
- For DL training, the dominant platform today is GPUs.
  - There remains tension into how much shared vs. distributed memory is required, what are the algorithms for model parallel training vs. scaling single systems, etc.
- Inference is more of a mix of devices.
- There are numerous emerging devices and accelerators, such as Google’s TPU. Startups in this area have attracted many billions in capital, and have different approaches, for instance:
  - Cerebus
  - GraphCore
  - SambaNova
  - Grok
  - NextSilicon
  - Habana
- Most focus on lower-precision operation acceleration.   Most focus on a more dataflow-oriented architecture.   Major differences in approach on size of wafer, programming model, software enablement.

The Landscape of AI Computing Resources

- Many resources have been deployed with an “AI”-lean among computational facilities:
    - Commercial Clouds – Mostly GPU, but the only way to get to TPU and some other technologies.
    - National Supercomputer Facilities:
        - DOE -- Frontier, Aurora, Polaris, El Capitan.
            - GPUs from AMD, Intel, and NVIDIA
        - NSF
            - A mix of production (GPU, ARM) and testbed systems (Cerebus, etc.).
        - Japan – Fugaku
            - ARMs with extension for lower precision computation
- Note most of these are “general-purpose” computers, often with GPUs added.
    - Likely because of both hybrid and other workflows,
    - but mostly because of reliable, re-purposable software stacks!
    - This lesson has been learned many, many times in multiple contexts through the history of computing. . .

Co-Location with Data and Other Computing

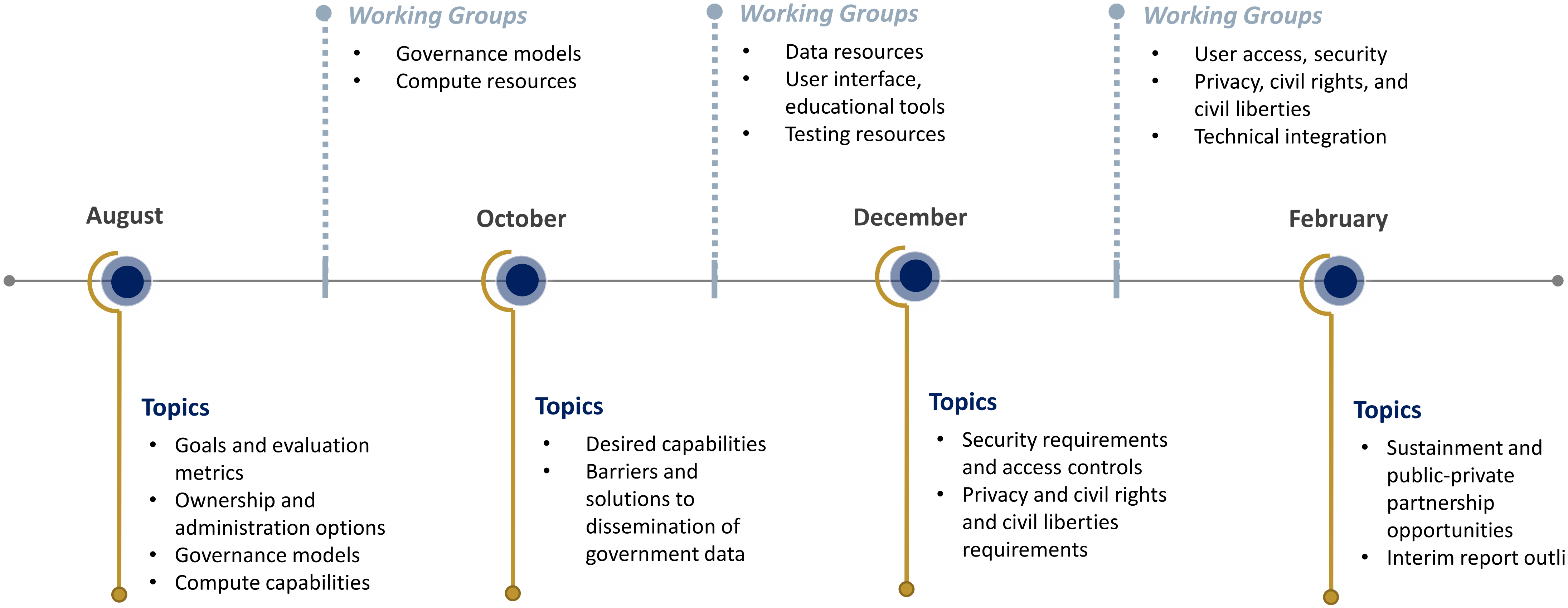
- What is our strategy for co-location?
  - Training datasets are huge, and likely to grow larger.
  - Moving data between providers is a large cost in commercial clouds
  - If we have a variety of resource providers, do we have permanent storage associated, staging space, etc.? Do we constrain where things can run by the data?
  - How much networking do we provision with compute to handle this?
- Similarly, many AI workflows couple tightly to massive simulation runs
  - Often, tens of thousands of simulations (or more) to generate training data for a single model.
  - Do we co-locate resources for AI with resources for simulation?
  - If not, data will have to move, exacerbating the problem above.

Discussion

Working Group Expectations

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

Assessment Phase



August

Topics

- Goals and evaluation metrics
- Ownership and administration options
- Governance models
- Compute capabilities

Working Groups

- Governance models
- Compute resources

October

Topics

- Desired capabilities
- Barriers and solutions to dissemination of government data

Working Groups

- Data resources
- User interface, educational tools
- Testing resources

December

Topics

- Security requirements and access controls
- Privacy and civil rights and civil liberties requirements

Working Groups

- User access, security
- Privacy, civil rights, and civil liberties
- Technical integration

February

Topics

- Sustainment and public-private partnership opportunities
- Interim report outline

Working Groups

Format:

- Working groups leads have the responsibility to set up and lead the discussions.
- Working groups can decide meeting frequency and how you want to manage your collaboration.
- Working groups are free to consult additional experts

Task:

- Develop recommendations to propose for consideration by the full group at the October meeting
- Provide a briefing at the October meeting summarizing the proposed recommendations and rationale for how they were reached

Working Groups: Baseline Questions

Governance Working Group

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

Compute Working Group

- What compute capabilities should the NAIRR include?
- How should access to these compute resources be managed through the NAIRR?
- Where should existing computing resources be leveraged and what new resources (if any) should be created?

Paper Process: Building the NAIRR Vision

- Record the growing consensus built over the course the past two meeting on the topics of value proposition, user base, and intended outcomes.

- Send around for comment, edit, and iteration among Task Force members.

- Enable all Task Force members to provide their input while simultaneously informing the deliberations of the working groups.