Federal Register Notice 86 FR 46278, https://www.federalregister.gov/documents/2021/08/18/2021-17737/request-for-information-rfi-on-an-implementation-plan-for-a-national-artificial-intelligence, October 1, 2021.

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

DISCLAIMER: Please note that the RFI public responses received and posted do not represent the views and/or opinions of the U.S. Government nor those of the National AI Research Resource Task Force., and/or any other Federal agencies and/or government entities. We bear no responsibility for the accuracy, legality, or content of all external links included in this document.

Response of Microsoft Corporation to White House Office of Science and Technology Policy and National Science Foundation RFI on the National Artificial Intelligence Research Resource

1st October 2021

Introduction

Microsoft enthusiastically supports the aims of the National AI Research Resource (NAIRR) Task Force and shares the vision of democratizing access to the evolving ecosystem of foundational computing capabilities to empower a larger and more diverse artificial intelligence (AI) research and development (R&D) community. In addition to the foundational role that compute cycles play in this R&D ecosystem, critical assets such as datasets, advanced AI tools, simulation environments, and platform models are playing an increasing role as key enablers of AI R&D.

As a global technology company and member of the larger R&D ecosystem, we have three primary recommendations for consideration by the NAIRR Task Force:

Recommendation 1: Leverage the rapidly evolving landscape of computing capabilities, advances, and practices available through commercial cloud platforms

Over the last several decades, we have experienced an increasing interdependence between advances in AI R&D and the development of AI platforms and services, with each enabled by rapidly evolving large-scale computing infrastructure, AI system software, and large-scale datasets as depicted in Figure 1. At Microsoft, AI R&D has been increasingly enabled by the co-development and use of our own AI research platforms and supported by intensive computation provided by our global network of data and compute centers.

We see the empowerment of a collaborative, global research community with expanded and democratized access to secure cloud-based compute, scalable AI infrastructure, and advanced AI platforms as both consistent with NAIRR goals D, G and H and aligned with Microsoft's mission. Microsoft research scientists and engineers have broadly and consistently engaged with the global research community through both direct collaboration and public-private partnerships. Microsoft's research and engineering teams maintain a commitment to open science and have co-authored and shared thousands of publications, many following from rich collaborations with researchers around the world. In addition to direct collaborations, Microsoft has also engaged in public-private partnerships in multiple countries including U.S.-based programs such as the COVID-19 HPC Consortium[1], The National Science Foundation's Cloudbank[2] program, and the National Institute of Health's STRIDES[3] program. These programs demonstrate the potential for effective partnerships across the public and private sectors to support the advancement of AI R&D.

Recommendation 2: Create a resource framework that offers large-scale infrastructure, AI system software, data, and platform models to support the AI R&D community

Consistent with NAIRR goals D, F and H, a NAIRR solution should enable a broad range of research workflows spanning many different disciplines and areas of study and support effective, efficient, and

secure execution of the associated computing workloads. Just as software development has come a long way from assembly language, computing resources that support scientific research computing are evolving into higher-level building blocks and tools. While large-scale and high-performance infrastructure provides the lowest common denominator for advancing research computing, higher-level tools in the AI technology stack (e.g., as depicted in Figure 1) aim to automate essential elements of current and anticipated future workflows relevant to scientific research. Given the heterogenous landscape of AI services, this broad range of workflows will be best supported by leveraging services from multiple providers that have the capacity to offer these capabilities for research. Such a resource would benefit the AI R&D community by enabling broad access to ongoing innovation, unique breakthroughs, and distinctive services being created across the public and private sectors. The need to increase access to these services has been illustrated by the recent emergence of large pre-trained neural models as platforms for both AI research and application development. A research paradigm has become central in AI R&D where fixed large-scale *platform models* are adapted via "fine-tuning" procedures to develop capabilities for specific tasks.

Commercial cloud capabilities, augmenting on-premises infrastructure, could enable a robust resource framework with minimal or no queue times and allocation limits for users. Historically, major technology shifts have occurred every 2-4 years as evidenced by recent advancements in networking, security and encryption technologies, and personal assistants. Modern commercial cloud capabilities offer the potential to enable significant advancements across the AI technology stack to occur on the order of months. A NAIRR solution should have sufficient agility to support the incorporation of new AI advancements as they become available for use. The ability for the broader research community to keep pace with these advancements will, in turn, help accelerate innovation across the AI technology stack.

Recommendation 3: Harness the creation of a National AI Research Resource to advance U.S. workforce development goals

The U.S. government has articulated a need to significantly ramp up the capacity to develop and skill a diverse next generation of AI researchers and engineers[4]. In reference to the NAIRR goals D and H, we believe that industry can play an important role in advancing educational tools and services, and overall, AI workforce readiness. Shared infrastructure between industrial and academic research, such as in Microsoft's AI and IoT Insider Labs[5], will improve collaboration across sectors in advancing technology to address societal challenges at scale while enhancing the transparency and reproducibility of research breakthroughs and strengthening the research to production pipeline. Academia can leverage the industry connection to enhance relevant curricula[6] with state-of-the-art computing capabilities that are already ubiquitous in industry. By engaging industry, including large corporations, small and medium technology companies, and early-stage startups, the responsibility of catalyzing robust research to innovation pipeline will be broadly shared.

The remainder of the document provides further detailed ideas on how we can accomplish these three recommendations in collaboration with the broader research community.

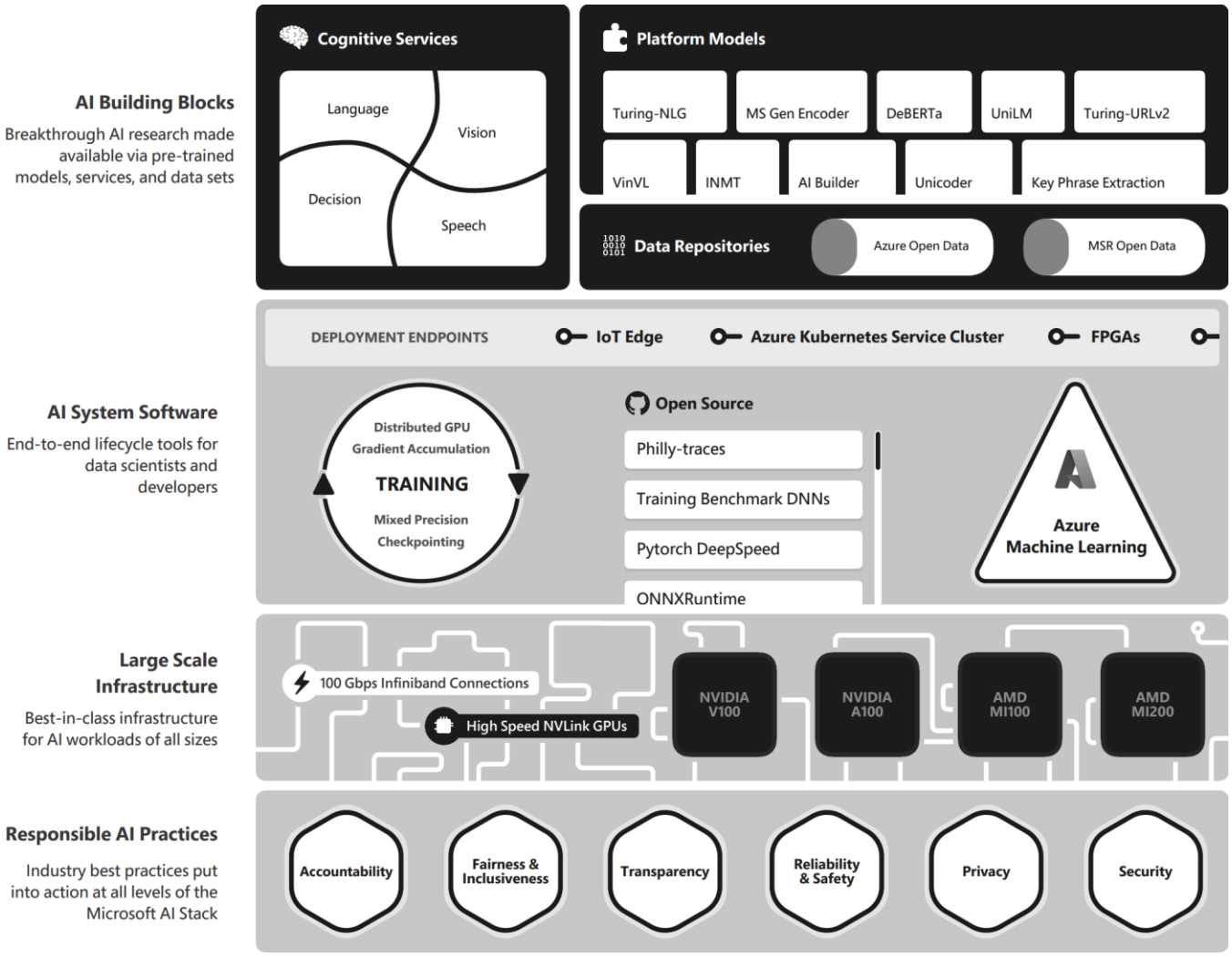

Figure 1. AI Technology Stack, illustrated by representative Microsoft services[13,14,20,24,25]

Recommendation 1: Leverage the rapidly evolving landscape of computing capabilities, advances, and practices available through commercial cloud platforms

With reference to Question 2: *Which capabilities and services provided through the NAIRR should be prioritized?*

It was recently stated by panelists at the Computing Research Association (CRA) led Virtual Roundtable on Best Practices on using the Cloud for Computing Research[7] that research created in education institutions will not keep up with 21st century advances if it doesn't take advantage of the enormous capacity, rich software, and hardware infrastructure that the commercial cloud offers. The NAIRR should accommodate a variety of research and simulation workloads (as illustrated in Figure 2) across systems, platforms, and resources. In general, research computing spans a very diverse set of workloads including core topics in AI and broader computer science as well as in a broad spectrum of areas across sciences

and engineering. Topics of particular interest cross a broad swath of frontier technologies and research with distinct compute needs. As examples, finite element analyses for computational fluid dynamics typically require high memory and core while climate modeling typically requires ensemble calculations. Given the ubiquity of AI across many areas of research, a robust NAIRR solution is likely to require similar breadth and diversity.

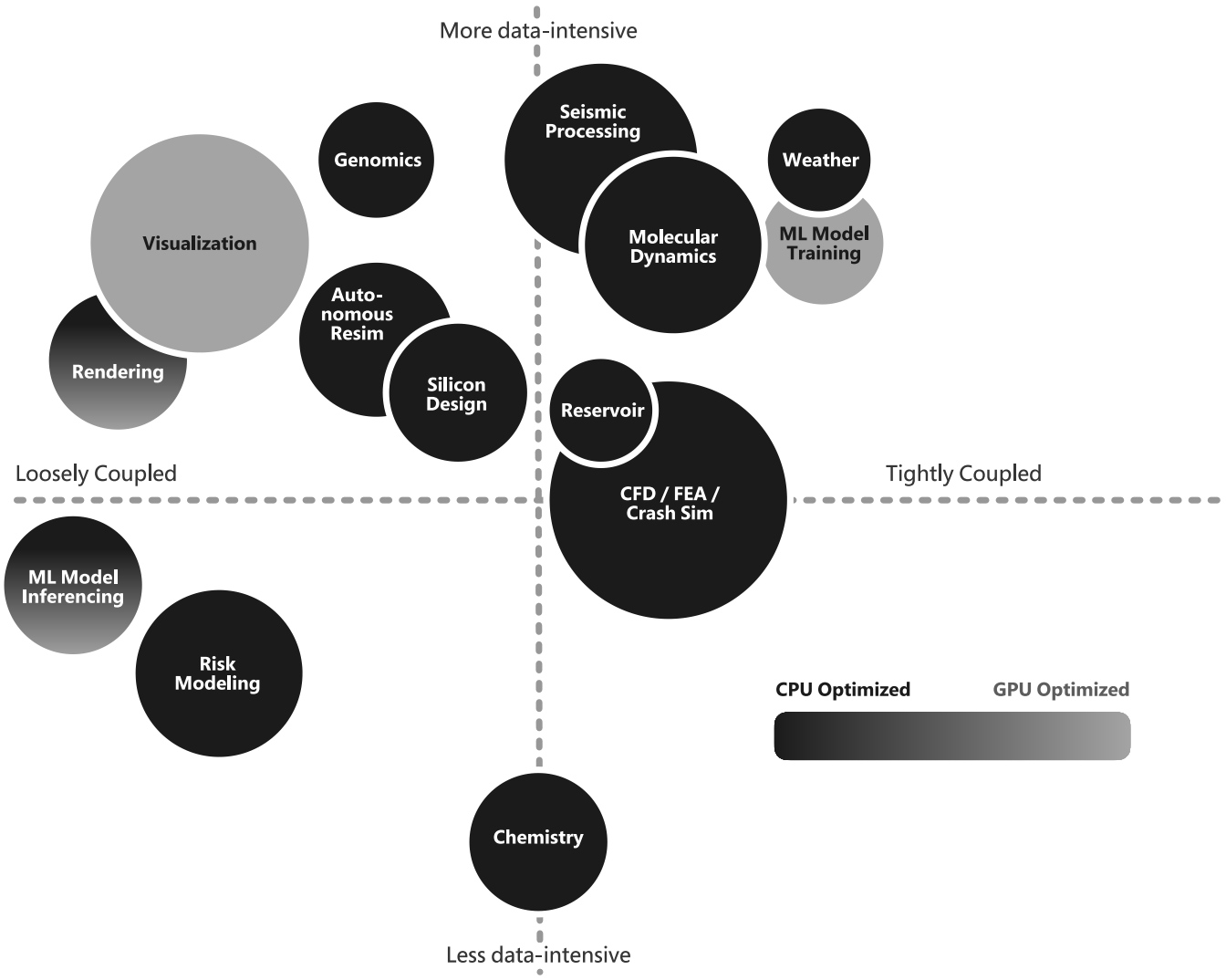

Figure 2. Cloud Compute Workload Mapping[26]

With reference to Question 4: *What building blocks already exist for the NAIRR, in terms of government, academic, or private-sector activities, resources, and services?*

Depending upon the specific computing need, a single subscription to a commercial cloud account could enable a researcher to access a broad range of compute resources optimized for performance and efficiency for the relevant workloads. For example, Azure's VM instances with InfiniBand-enabled clusters provide significant performance advantages for running tightly coupled high-performance compute workloads. Figure 3 provides an 'at a glance' reference to the Azure VM families. In addition, VMs that incorporate

confidential compute capabilities offer the ability to build secure enclave-based applications within trusted execution environments.

As climate change and health concerns continue to impact society, they represent an increasing influence on computing needs associated with government-funded research. Hence, the workloads and datasets that support these research areas are expanding beyond traditional modeling and simulation.

Azure can be connected to datasets hosted in federal agencies and if needed, secured by a Virtual Private Network tunnel. Azure ExpressRoute is a service that enables users to create private connections between the datacenters that cloud servers are hosted in, and infrastructure that are on premises or in a co-location environment with dedicated circuit available from 50 Mbps to a 100 Gbps pair.

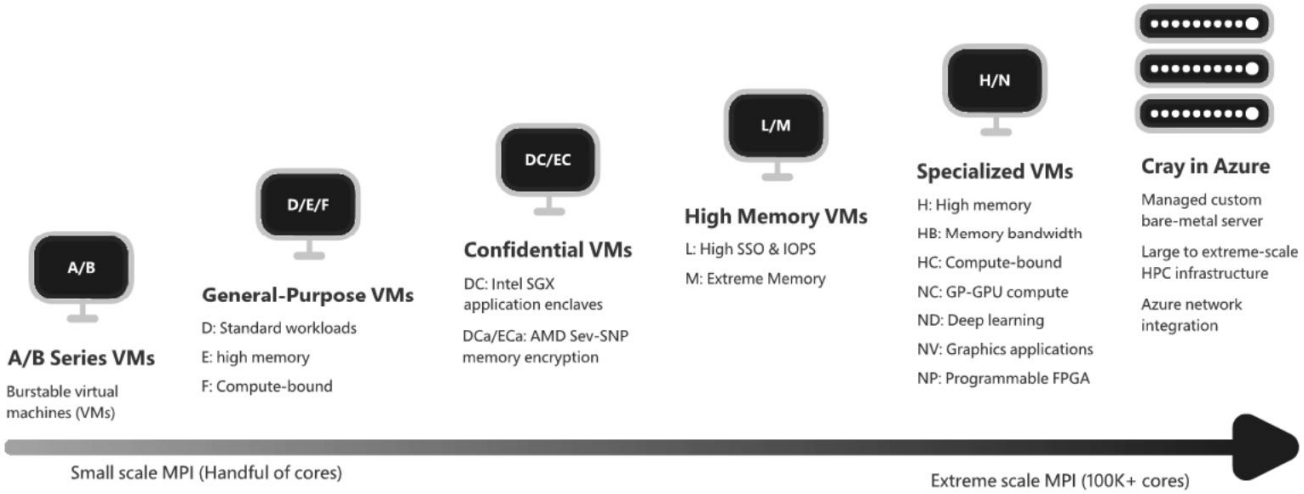

Figure 3. Azure Virtual Machine Groups[27]

Given the increasing scale of commercial cloud, cybersecurity represents a growing need and area of continuous innovation and advancement. For instance, Azure ingests trillions of security signals each day and combines security, compliance, identity, and management as an interdependent whole. The application of AI to the analysis of these signals enhances the ability to identify, detect, protect, and respond in real-time to known and emerging cybersecurity threats.

Azure is accredited[22] at FedRAMP Moderate and High, DoD Impact Levels (IL) 2,4,5, and 6, and meets ITAR, CJIS, DFARS, and NIST 800-171 requirements. Additionally, Microsoft complies with NIST SP 800-53 and is aligned with NIST SP 800-161 supplemental guidance.

With reference to Question 5: *What role should public-private partnerships play in the NAIRR? What exemplars could be used as a model?*

NAIRR/NSF-funded researchers should ideally have sufficient understanding of requirements associated with relevant workflows so that they are equipped to evaluate or take full advantage of a cloud-supported NAIRR solution. NAIRR should work toward a model that facilitates this level of understanding to help ensure that each workflow runs with high efficiency on readily available resources.

In a cloud-computing setting, each scientific workflow is, in effect, its own benchmark. Benchmarking the workflow in the target environment enables in-situ tuning and optimization with respect to

performance and cost. While individualized benchmarks take time on the front-end, they eliminate the benchmark-to-execution step and can inform a framework of pre-computed simulation environments.

Programs such as NSF Cloudbank and NIH Strides represent emerging efforts leveraging public-private partnerships to promote a culture of government-funded research supported by cloud compute capabilities. Outside the US, recent examples include the UK government's billion-dollar investment through which the UK Met Office and Microsoft partnered to build the world's most powerful weather and climate forecasting supercomputer,[16] the establishment of a government-funded public-private 66,000m^2 AI innovation hub in China[17], and the projects undertaken by the Swedish National Center for AI in partnership with the private sector[18].

Recommendation 2: Create a resource framework that offers large-scale infrastructure, AI system software, data, and platform models to support the AI R&D community

The Fourth Industrial Revolution has resulted in widespread application of machine learning and AI technologies across a broad range of industrial automation use-cases (eg. as depicted in Figure 4). Emerging capabilities like Github Copilot[9] are accelerating the development and application of new capabilities by enabling developers to automatically identify relevant code written by other developers with the same intent, code search systems can help automatically retrieve relevant code based on natural language queries. This AI-enabled code intelligence has been the intent behind open benchmark datasets like CodeXGLUE[10] and developer tool enhancements powered by Github coPilot that aim to empower the 23 million+ developer community.

AI and machine learning aren't new concepts, and many of the theories have been unchanged for decades, but recent technological advances have accelerated AI innovation. These advances include large-scale

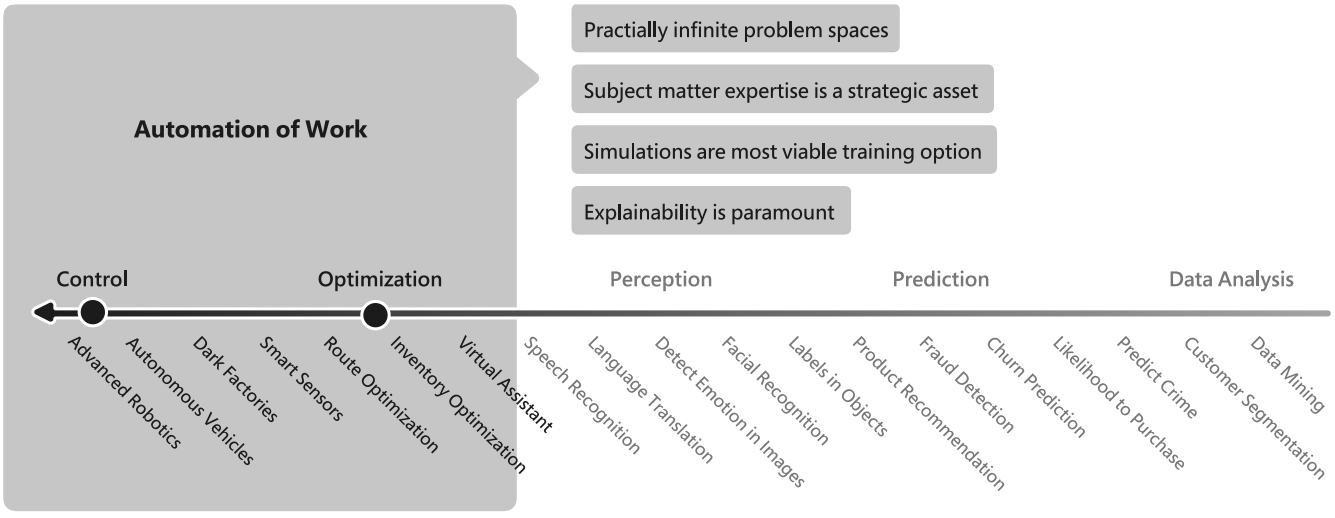

Figure 4. AI Use-Cases in Industry[8]

infrastructure such as AI supercomputing[11], and advanced storage and bandwidth capabilities to enable more accurate and useful algorithm predictions. There is a clear need for higher-level AI services beyond the CS and industrial automation community to computational scientists and other advanced practitioners.

In reference to Question 2: *Which capabilities and services provided through the NAIRR should be prioritized?*

Recognizing that the NAIRR intends to serve communities across a diversity of scientific disciplines, it is important to acknowledge the need for higher-level AI services to facilitate access to AI resources by research communities. As illustrated in Figure 1, these resources fall under the following broad categories:

A. AI System Software

Training a large neural model requires several components such as deep learning optimization strategies, efficient and scalable runtime engine, and a service to manage the experiments and the distributed training infrastructure. Workflows include model design, distributed training, mixed precision, gradient accumulation, and checkpointing.

ONNX Runtime is an open-source engine to support the highly efficient high performant training and inference of AI models in a framework-agnostic manner across a range of hardware. This runtime brings together efficient implementations of the mathematical operations underpinning the deep learning algorithms and the training optimizations from various Microsoft capabilities into one integrated package.

Azure Machine Learning (Azure ML) is the end-to-end machine learning development lifecycle that enables efficient building, training, and deployment of machine-learned models at scale. Azure ML enables team collaboration with experiment tracking, model performance metrics collection, industry leading MLOps, i.e., DevOps for machine learning. Azure ML Service supports model training and deployment across all major deep learning frameworks and runtimes including the ONNX runtime, leveraging the Azure AI infrastructure including large GPU clusters.

Today we're seeing the most sweeping changes in data management since the relational database revolution of the '70s and '80s. These advances motivate significant changes in how researchers engage with the next generation of data management platforms.

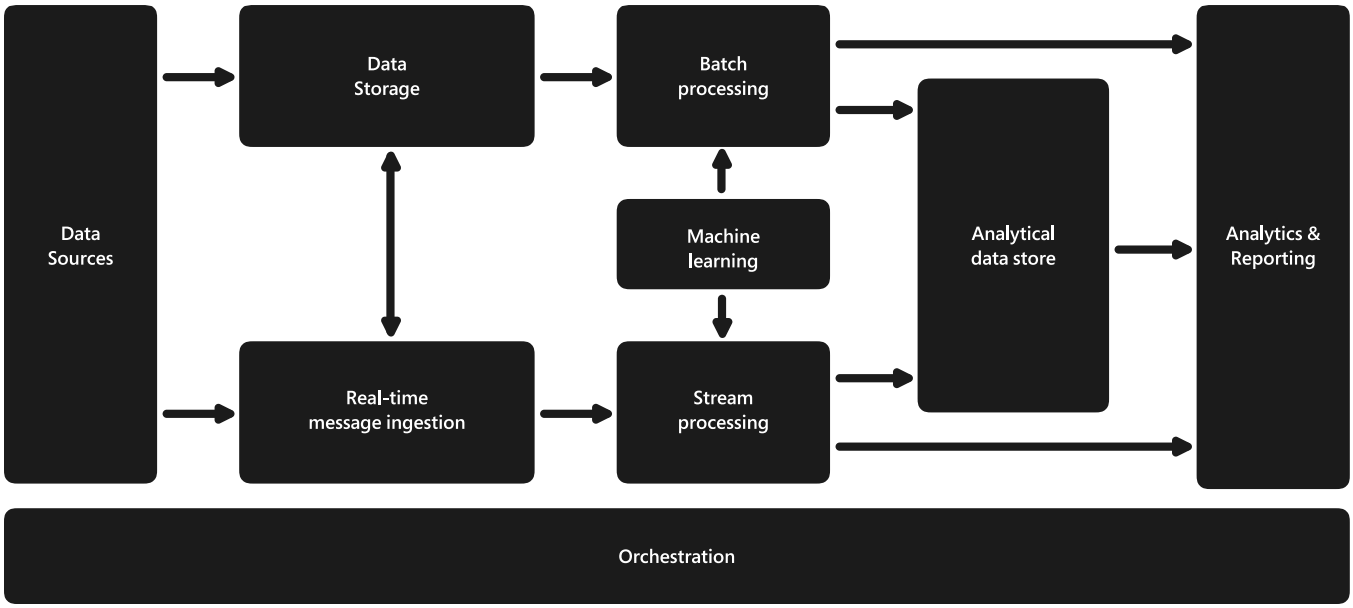

Figure 5. Cloud based Modern Data Platform Reference Architecture[23]

The components of Microsoft's Modern Data Platform such as shown in Figure 5 are designed to address such demands.

A modern data platform architecture is designed to handle ingestion, storage, processing, and analysis of data that is too large or complex for traditional database systems. In addition to size (volume), the types of data (variety), speed of the data (velocity) and inconsistencies in the data (veracity) is also considered. Azure supports several "Big Data" deployment architectures, to include the Lambda (hot and cold data processing) and Kappa (single path stream processing) architectures which provide a consistent way to ask questions of data in motion or data at rest. Azure also supports multiple other products build to simply the management of AI/ML processes such as Data Science Virtual Machine (DSVM), Azure HDInsight, Azure Databricks, Azure Synapse Analytics and Azure Machine Learning Service (Azure ML).

B. Models that have been pre-trained with massive data

Following the trend that larger natural language models lead to better results, we are building upon the strong collaboration across research and engineering to achieve new modelling techniques for language breakthroughs. Language generation models are currently used in many natural language scenarios such as word prediction, sentence completion, text summarization and are an important research area in the future of AI. Microsoft developed a family of large-scale natural language processing models, a few of which are listed in Figure 1, and in 2020, announced Turing model[13] for natural language generation (NLG), the largest model ever published at 17 billion parameters, outperforming the state of the art on a variety of language modeling benchmarks while accomplishing numerous practical tasks such as summarization and question answering.

The deep learning library behind the model, DeepSpeed was made open source to make distributed training of large models easier. DeepSpeed contains the Zero Redundancy Optimizer (ZeRO) for training models with 100 million parameters or more at scale. Both DeepSpeed and ZeRO are available to researchers, because training large networks like those that utilize the Transformer architecture can be expensive and can encounter issues at scale.

Through the Turing Academic program[12] Microsoft provided access to the Turing family of models, to support general and open research, including efforts aimed at advancing principles of learning and reasoning, exploring novel applications, and pursuing better understanding of challenges and opportunities regarding the ethical and responsible use of large-scale neural language models.

The initial results of the program have resulted in research collaborations with several academic research groups that explore research areas across a broad range of AI research problems such as to 1) transfer learning in medical notes 2) understand and quantify biases in models 3) identify linguistic markers that exacerbate bias 4) distinguish between real and fake events.

The program has shown promise as something that can have far reaching impact in examining the unintended impacts of language model-based AI technologies. Eventually the language model approach of how AI is developed from narrow, custom models to multi-purpose, massive models is expected to be generalized to multi-modal text, video, voice data. Research investment via a convening of public and private sector stakeholders is critical to balance the economic incentives that make the resultant AI technologies inevitable in use by commercial interests.

In the exploration of language models, as well as other forms of AI, it is important to consider the responsible design, development, and deployment of those technologies. Overall, there is growing investment[20] on the responsible development and fielding of AI systems, including developing

accountability and governance across industry. Efforts on responsible AI innovation are focused on how to operationalize the following principles:

- **Fairness.** AI systems should treat all people fairly
- **Reliability & Safety.** AI systems should perform reliably and safely
- **Privacy & Security.** AI systems should be secure and respect privacy
- **Inclusiveness.** AI systems should empower everyone and engage people
- **Transparency.** AI systems should be understandable
- **Accountability.** People should be accountable for AI systems

Moving forward it will be important to seek scalable, research-based methods to advance these principles, such as outlined in the NSCAI report[14] that presents key considerations for the responsible development of AI. Measurement and evaluation tools to assess the technical requirements for responsible AI, for example areas like fairness and accuracy, will contribute to this goal.

In addition, several AI benchmark datasets and models are such as MS-MARCO for large scale reading comprehension and question answering, MT-DNN for natural language understanding, XGLUE for cross-lingual evaluation benchmark XGLUE are available under the AI for Scale research[15] effort.

Recommendation 3: Utilize the creation of a national AI research resource to advance U.S. workforce development goals

With reference to Question 5: *What role should public-private partnerships play in the NAIRR? What exemplars could be used as a model?* And question 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?*

The research communities' experiences over the past two decades indicates that AI research is currently dependent on *both* pioneering researchers and advanced engineers. We expect this situation will continue through the next few decades of research in AI. In our AI research efforts and in working with academic researchers in AI, we have periodically seen a limitation to the pace of research can be the researchers' ability to employ state-of-art computing systems and techniques. These are techniques that are enabled by collaborating engineers, and in contrast, the researchers who are able to rapidly and efficiently leverage their computing resources have some advantages over peers who do not.

Academia provides a unique and foundational educational environment that prepares individuals for a lifetime of learning. It develops the emerging workforce by fostering their abilities to think critically, to grow techniques and capacity to learn; to approach new problems creatively, and to understand and evaluate conflicting ideas and information. In doing so, academia creates the knowledge foundations required for most researchers and engineers that are responsible for the next wave of AI innovation in the US. Academic learning content and materials take a relatively long period of time to develop, evolve in an iteratively manner, and are vetted by peers and communities of educators.

We also recognize that undergraduate and graduate curriculums have more content and material than they can practically accommodate in their degree programs. And we realize that it is not appropriate to re-orient degree programs towards courses and work dedicated to the development of skills on specific tools and platforms. However, to reach their full potential throughout their career, researchers and engineers do need to develop some explicit skills to understand and employ the continually changing AI systems, techniques, and tools. These are skills that Industry is well-equipped to provide.

It will also take a combined effort of academia, industry, and the government to increase the diversity of AI researchers and engineers. We can only accomplish this through sustained initiatives that impact students through their educational journey and professionals through their careers.

NAIRR should embrace the critical, complementary roles that both academia and industry play in educating and skilling researchers and engineers. Building, operating, and using the next evolution of compute and data resources will require foundationally sound and continually skilled workforce. Together, academia and industry will provide institutional capacity and commitment to adequately staff, in both quality and quantity, the workforce essential to research workflows.

Microsoft has a strong history for supporting the development of future-ready skills in collaboration with the academic community through co-curricular programs like Microsoft's AI Business School and Microsoft Learn for Educators. AI Business School[21] (AIBS) is a master class curriculum designed to build knowledge and confidence in AI. It provides comprehensive training for non-technical learners spanning critical topics like strategy, culture, responsible AI, scale AI, AI for business users, and AI technology for leaders. Its content has been adopted widely by colleges and universities across the globe.

Our experiences with Microsoft Learn for Educators[19] (MSLE) is anchored in the belief that higher education intuitions and faculty members play a pivotal role in empowering students for future success. As such, MSLE supports students, faculty members, and higher education institutions with free curriculum, training, and tools for teaching technical skills. MSLE supports faculty at colleges, universities, and community colleges who want to help build their students' skills in technical topics like cloud computing, AI, data engineering and security, so that they can attain industry-recognized certifications that prepare them for their future careers. Microsoft Learn for Educators currently trains and provides technical curriculum to over 3500 faculty members at 450 colleges, universities, community colleges, and polytechnics across 85 countries and in 9 languages. In fiscal year 2021 alone this Microsoft program helped higher education institutions provide cutting-edge technical skilling learning experiences to over 60,000 students.

Summary

Advancing AI in a manner that is trustworthy, ethical and benefits the whole of society will require participation and collaboration across a wide range of scientific disciplines, institutions, and sectors. Microsoft is aligned with the NAIRR task force vision of democratizing access to the cyberinfrastructure that fuels AI research and development and looks forward to continuing to participate in upcoming forums related to the mission.

References

[1] https://www.hpcwire.com/2021/03/26/the-covid-19-hpc-consortium-looks-ahead-to-a-national-strategic-computing-reserve/

[2] https://www.cloudbank.org/about

[3] https://www.nih.gov/news-events/news-releases/nih-expands-biomedical-research-cloud-microsoft-azure

[4] https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence

[5] https://www.microsoftiotinsiderlabs.com/

[6] https://www.businesswire.com/news/home/20210114005727/en/