LEAP Shared Experimental Facilities

"Experimental facilities" traditionally means capital equipment, plus the infrastructure and personnel necessary for operations. Befitting its transformative nature, LEAP will pioneer a new model for cyberinfrastructure that leverages cloud computing rather than equipment acquisition. This cyberinfrastructure will build upon and integrate with existing activities across all six academic institutions and two federal lab partners.

Motivation and Background: LEAP's proposed research involves massive, complex datasets and ambitious computational needs, thus requiring a robust cyberinfrastructure supporting climate data science. Earth system datasets are complex and heterogeneous, and LEAP's personnel have different levels of skill for accessing and manipulating them. Achieving LEAP's goals will require novel approaches to data and computing infrastructure: Building upon **Abernathey** and **Hamman**'s experience with the NSF- and **NASA**-supported Pangeo project, an international, open community of scientists and software engineers collaboratively developing software tools and computing infrastructure to address modern big data science challenges. Pangeo's established tools and approaches will inform LEAP's shared infrastructure. Informed by the scientific applications and datasets described in the Project Description, the following observations will govern the proposed infrastructure development: 1) data preparation is one of the most time-consuming aspects of applied ML research; 2) ML model development, fundamentally an interactive process requiring a "human in the loop" to iteratively assess and improve performance; 3) data must be stored in a central location with computing performed close to the data, because the datasets are too large to regularly download to personal computers; 4) training complex ML algorithms is most efficient on specialized computing hardware (GPUs, TPUs); and 5) open-source software and open data formats facilitate collaboration and reproducibility.

Proposed Facility – LEAPangeo: Using these guiding principles, LEAP will build a shared data and computing platform called *LEAPangeo*. This platform will use cloud computing, provided via a partnership with **Google Cloud** and **Microsoft Azure**, and allowing the required combination of high-performance mass storage (e.g., **Google Cloud Storage**, "GCS") and on-demand computing (e.g., **Google Compute Engine**, "GCE"). Datasets will be extracted from diverse sources, transformed into analysis-ready, cloud-optimized formats [38], and stored in GCS. All data will be published using open-source licenses and will be publicly available (see Ethics Plan).

The primary user interface for data analysis will be Jupyter notebooks, which have emerged as the *lingua franca* of interactive computing and data science. A cloud-based JupyterHub, running in GCS and managed via Kubernetes, will provide on-demand, interactive computing environments in close proximity to the data. The Hub will provide different software and hardware environments tailored to different analysis tasks, including ML libraries such as TensorFlow and PyTorch coupled with GPU and TPU hardware. The Dask parallel computing framework will accelerate data processing and ML model training. A cloud-based BinderHub deployment will enable reproducibility. LEAPangeo will transform the scientific process by eliminating tedious data-preparation tasks, facilitating collaboration, and enabling computational reproducibility. Team members will be able to easily share in-progress work and code, facilitating dialog between ML experts and geoscientists. This platform will also integrate with and support LEAP's education, broadening participation, and knowledge transfer initiatives. LEAPangeo's focus is on novel forms of cloud-based data sharing, analysis, and ML model development. More traditional high-performance computing (e.g., running CESM simulations) will be enabled by **NCAR**'s existing high-performance computing facility: This will not require establishing new facilities or infrastructure.

Personnel: LEAPangeo will be governed and guided by Data & Computation Director **Abernathey**, a computational physical oceanographer, open-source software advocate, and co-founder of the Pangeo project. He will supervise two full-time cloud systems engineers devoted to building and maintaining this shared experimental computational facility. A Cloud Engineer specializing in container technology, development-operations, and continuous integration/ automation will perform deployment and maintenance activities; the

incumbent will sit in **Columbia** University Information Technology (CUIT). A Scientific Data Engineer, specializing in scientific data formats, metadata management, and distributed data processing, will be responsible for construction and maintenance of the cloud-based data repository. The incumbent will sit in **Columbia**'s Data Science Institute.

Timeline:

- **Year 1:** LEAP engineers will deploy the initial LEAPangeo cloud computational platform based upon existing Pangeo best practices as documented on the Pangeo website and GitHub repositories. Throughout the first year, the primary task of the technical team will be manually ingesting datasets from their original repositories (e.g., NOAA, **NASA**, **NCAR**) into cloud storage in analysis-ready format, and creating catalogs for use by all LEAP researchers.
- **Year 2:** Building upon the experience garnered in Year 1, engineers will develop tools to automate the ingestion of datasets, creating a self-service tool that enables scientists to more easily curate their own data catalogs.
- **Years 3-5**: Engineers will continue to refine LEAPangeo's capabilities and customization in response to ongoing feedback from scientists. This scope of work will be determined by emergent data and computational challenges.

Throughout the performance period, both engineers will ensure the continuous operation of the cloud platform, provide regular maintenance and security updates, and respond to user support requests. Engineers will also offer regular training workshops to LEAP-affiliated researchers. As the LEAP performance period ends, the team will assess how to maintain operations and open LEAPangeo to the broader community, including the possibility of assessing user fees.

Open Development. With cloud computing, infrastructure *is* code. All code used to create and deploy LEAPangeo will be developed openly on GitHub, in collaboration with the broader Pangeo community, and distributed with open source licenses (see Ethics Plan). Development will follow an Agile methodology, with frequent releases and continuous integration/ deployment. Wherever possible, development will involve contributions to existing community-supported projects such as Jupyter. This strategy will mitigate risk, support sustainability of all software products, and ensure NSF's investment has the broadest possible impact.

Institutional Integration. LEAPangeo will become a new shared service, added to the portfolio of services managed and supported by the Research Computing Services (RCS) team in CUIT. The RCS team provides technology resources to support **Columbia**'s researchers and the University's mission to be a leading driver of research, innovation, and knowledge advancement. RCS is directly supported by six full-time staff while also leveraging the central infrastructure support staff of CUIT. They provide maintenance, technical support, consultation, software installation, and guidance on future computing needs.

The RCS team has expertise in building, maintaining, and expanding the University's on-premise, shared, high-performance computing infrastructure and services, which they have managed since 2009. Consistent with **Columbia**'s campuswide cyberinfrastructure plan, CUIT has been shifting to cloud resources for both functional and financial benefit. The unit boasts cloud provision staff on the infrastructure team and well as on the RCS staff to provide researchers with the expertise they need to decide where their research is best suited. The RCS team provides consulting for cloud services, including planning, onboarding, configuration, and training. With the acceleration of cloud expertise, the RCS team is poised to provide support to the LEAP effort into the future as a central service. This service will be available to all researchers, with support, administration, and training provided by RCS.

Key research stakeholder engagement and oversight of shared computational services is formalized by **Columbia**'s Research Computing Executive Committee, consisting of senior research leadership and school deans; the faculty-led Shared Research Computing Policy Advisory Committee; and the Information Technology Leadership Council, an advisory group to the Chief Information Officer. Progress on LEAP's Shared Experimental Facility will be reported at the biannual SRCPAC meetings.