COLUMBIA UNIVERSITY

COMMITTEE ON RESEARCH COMPUTING and DATA INFRASTRUCTURE

June 2024

Hod Lipson (School for Engineering and Applied Sciences, co-chair)

Darcy Peterka (Zuckerman Inst for Mind, Brain, and Behavior, co-chair)

Timothy Berkelbach (Department of Chemistry, Arts & Sciences)

Seth Cluett (Computer Music Center, Arts & Sciences/SoA)

Róisín Commane (Department of Earth & Environmental Sciences in Arts & Science/LDEO)

Siddhartha Dalal (SPS and Department of Statistics in Arts & Science)

George Hripcsak (Biomedical Informatics, CUIMC)

Despina Kontos (Department of Radiology, VP&S, CRIO for CUIMC)

Wojciech Kopczuk (Economics in Arts & Sciences and SIPA)

Laura Kurgan (GSAPP and Center for Spatial Research)

Ciamac Moallemi (Graduate School of Business)

Robert Pincus (Lamont-Doherty Earth Observatory)

Julien Teitler (Social Work and Columbia Population Research Center)

Ex Officio

Maneesha Aggarwal (CUIT), Robert Cartolano (Libraries)

Gaspare S LoDuca *(CIO, VP of IT, CUIT), Victoria Hamilton (EVPR)*

Alexander Urban (Chemical Engineering, SEAS, Chair of Shared Research Computing Policy Advisory Committee (SRCPAC)), Jeannette M. Wing (EVPR, Computer Science)

Table of Contents

Table of Contents 2

Executive Summary 2

Introduction 6

From Mainframes to Datacenters to Generative AI 8

Why existing solutions do not satisfy our growing AI needs 13

Outline of methodology 15

Survey Results 16

Basic Infrastructure 16

Existing High-Performance Computing 17

New Needs for AI 18

The research computing “pyramid of needs” 19

Summary of emerging needs 20

Recommendations 24

1. Create The Discovery Accelerator 24

2. Enable all disciplines to engage in Research Computing 30

Governance 33

Conclusion: A bifurcation of futures 34

References 36

Appendix 44

Acknowledgements 44

Committee members 45

Peer interest and investment in HPC/Compute and AI 46

Survey 47

School of Engineering and Applied Sciences (SEAS) 49

Key observations 50

Graduate School of Architecture, Planning and Preservation (GSAPP) 52

Columbia Population Research Center 56

Lamont campus perspective on research computing survey 59

Arts & Sciences - Sciences response 61

Arts & Sciences Humanities and the School of the Arts 64

Social Sciences (mainly economics and political science) 67

Zuckerman Institute 69

Research Computing Needs at Columbia Business School 73

Faculty Input on Computing Needs - Law School 75

Faculty Input on Computing Needs - School of Professional Studies (SPS): 76

CUIMC Faculty Response on Research Computing 78

CUIT Point of View 83

Executive Summary

The core mission of Columbia University is to create, curate, and disseminate knowledge. For the nearly three centuries that Columbia University has existed, knowledge has been derived from the intellectual pursuits of human scholars working alone and in teams. However, in the past decade, and accelerating in the past few years, a new epistemological paradigm has emerged: Scholars, artists, scientists and engineers can magnify and amplify their ability to create and discover new knowledge by collaborating with new forms of Artificial Intelligence (AI). We use the term AI loosely to encompass all forms of current and future high-performance computing tools that assist, augment and accelerate knowledge discovery and generation.

Scholars debate to what degree AI will affect different fields and how soon, how academic roles will change, and in what ways the technology will parallel or differ from human ingenuity, limits and constraints. However, one trend is clear: Just as automation technology has changed industry in the past, automation is now poised to revolutionize academic pursuits. And just as not every manufacturer survived automation, not every academic institution will survive and thrive in the impending AI transformation. AI technology is not merely hype, nor has its full potential revealed itself yet. We are only at the beginning.

With this opportunity also comes great responsibility. We would argue that Columbia University is particularly well poised to play a key role in this transformation, specifically because it can bring to bear many pillars of disciplinary excellence. Columbia University must fully embrace and engage with AI across all of its fields, disciplines, and pursuits, not only to stay competitive and thrive, but critically to lead the nation and the world in the responsible alignment and application of these new technologies towards society and human values.

Artificial intelligence offers opportunities to alleviate or solve grand challenges in climate, energy, food, water, health, shelter, privacy and security, to name a few. But this technology can also be used ineffectively or worse, nefariously; opportunities will be lost and harm will be done. Academic researchers have the ethical responsibility to engage and lead the beneficial development of this technology for society, and to balance competing commercial and governmental forces. Few institutions compare to Columbia in its ability to holistically scrutinize the potential and peril of AI and recommend fruitful paths forward.

To fulfill our responsibility and serve effectively, academic scholars and researchers must possess both **independent access to the technology itself,** as well as **the necessary skills to use it and direct its trajectory**. This report outlines what Columbia University needs to achieve both these goals.

The Opportunity:

- High Performance Computing and AI (HPC+AI) are profoundly altering the opportunities and risks of all research universities. Unlike the past, when HPC was the domain primarily of science and engineering, HPC+AI will cut broadly across all disciplines of scholarship and discovery including arts, humanities, law, policy, journalism and more. HPC+AI will transform research, scholarship, and teaching. A significant, early investment in this emerging area will make Columbia exceptional. Columbia's broad and deep faculty knowledge, research portfolio, and data have long been a great strength. When paired with significant HPC+AI, tremendous opportunities for discovery and societal impact should emerge.
- Columbia needs extraordinary HPC to attract outstanding faculty as well as retain existing faculty.
- Our peers recognize the opportunity and threat investing significantly in staff, capital expenditures and licensing deals with HPC+AI vendors. While the exact spending on HPC/AI infrastructure and support vs. total investments makes direct comparisons of these numbers difficult, there are large investments in this space (e.g. MIT $1B, USC $1B, UPenn $750M) [1–7] that will empower these peer universities to recruit and retain talent. Faculty and students alike will gravitate to these investments and opportunities.

The Recommendation:

- **Recommendation 1:** Create the "Discovery Accelerator" facility that will enable and encourage all CU researchers to take advantage of HPC+AI, subsidizing usage at low or no direct cost, particularly at the beginning. This includes compute and data infrastructure, as well as operational support. Critically, this is not a "one-and-done" investment but needs to be structured and budgeted for continual upgrades.
- **Recommendation 2:** Facilitate educational and technical services that will catalyze development, understanding, exploration, and adoption of HPC+AI opportunities across

the University, allowing Columbia to broadly and rapidly advance—including in disciplines that currently lack in-house expertise.

The university must undertake both prongs of this strategy rapidly to catalyze a virtuous cycle whereby our researchers demonstrate leadership which in turn attracts talent, research awards, and reputation. Conversely, playing catch-up will cost more and deliver significantly less. This is truly a situation in which “those who hesitate are lost.”

The Implementation:

- In the next decade, investing approximately $***25M per year on computational infrastructure and $5M per year on training*** *(~$300M total over* a decade) to realize this vision. Such an investment strongly aligns with university strategic priorities of academic and operational excellence in areas including climate, artificial intelligence, mental health, and probably any and all future initiatives[8]. We believe that this investment will return increased research productivity, talent recruitment and retention, increased proposal competitiveness, and strategic fundraising campaigns for impact-driven donors. If we assume roughly ~$1B in research expenditures a year, this investment represents a small fraction of our total sponsored projects spending and will create competitive advantage for awards. Additional cost savings opportunities exist through mindful consideration of local power consumption or space (e.g., Local Law 97)[9,10]. This investment roughly divides between infrastructure ($250M) and training ($50M).
- A center or similar structure, led by established researchers in the field, should oversee the facility and staffing, engage with campus faculty and leadership, and coordinate with OAD for fundraising to support the facility. A faculty-led steering committee, similar in design to the current SRCPAC, will work with leadership and CUIT to direct purchasing and allocations to balance needs and access.

Introduction

In April 2023, Vice President and CIO *Gaspare LoDuca* and Executive Vice President for Research *Jeannette Wing* commissioned a task force of faculty from all schools and colleges of Columbia University to submit a report outlining proposed strategies for the university to address its growing reliance on research computing.

The charge, provided in the appendix of this report, is "to recommend a strategic plan for the University's future computational and data infrastructure for research." The committee considered all the major elements of this plan, including but not limited to:

- Computing resources
- Data resources for analysis, sharing, storage, archiving, privacy and security
- Technology skills required
- Policy impacts
- High level cost implications

Even a year or two ago, it would have been difficult to predict that we would now have machines capable of holding deep conversations in natural language, creating novel artwork, writing prose and poems, composing music, discovering new materials, controlling robots, crafting new proteins, and diagnosing disease. The pace of progress and change is startling. Just as the AI models this year far exceed the capabilities of last year, the AI next year will exceed the capacity and potential of today's AI.

Therefore, this report seeks not to simply describe current research computing needs, nor to detail the necessary existing infrastructure maintenance, upgrades, and services to bring fast networking, adequate wifi, and powerful laptops to every researcher. Instead we aim more ambitiously to:

1. Anticipate future computing needs of all scholarly fields, allowing scholars to flourish in an era where scholarly pursuit will likely be enabled (if not driven) by Generative AI, big data, and computational discovery and creation.
2. Encourage Columbia researchers to leverage and engage computational resources to remain leaders in their chosen fields in the face of accelerating HPC and AI-driven progress (and competition).

The urgency of an institutional strategy to adapt to the fast pace of AI and HPC is summarized by Liu and Jagadish in their recent Harvard Data Science Review[11] as follows:

> *"The scale and speed of the Generative AI revolution, while offering unprecedented opportunities to advance science, is also challenging the traditional academic research model in fundamental ways. The academic research model and academic institutions are not set up to be nimble in the face of rapidly advancing technologies, and the task of adopting such new technologies usually falls on individual researchers. Excitement about the opportunities that Generative AI brings is leading to a rush of researchers with various levels of technical expertise and access to resources to adopt this new technology, which could lead to many researchers "reinventing the wheel" and research outcomes lacking in ethics, rigor and reproducibility. This problem not only applies to Generative AI, but could also be true for other upcoming and similarly disruptive technologies. We argue that the current norm of relying on individual researchers for new technology adoption is no longer adequate. It is time that academic institutions and their research organizations [...] develop new mechanisms to help researchers adopt new technologies, especially those that cause major seismic shifts such as Generative AI. We believe this is essential for helping academic researchers stay at the forefront of research and discovery, while preserving the validity and trustworthiness of science."*

Though this committee's mandate specifically excluded studying the use of computing for education, teaching, and administration, in fact, borders can –and should – be porous. Using computing to further research (including the field of research computing itself) and teaching students how to use computing to advance their understanding and/or to do research, inevitably interact and overlap. For example, many undergraduate and graduate students participate in advanced research as part of their overall education and experience at the University, and the curriculum and course work should provide a strong foundation in the principles and practice such that our students can become leaders in their respective domains. Training students in AI/HPC-enabled research will require ready access to advanced computing resources, as well as sufficient and appropriate training to use these tools effectively.

Beyond direct research and coursework, teaching students to use and interact with advanced AI and simulation tools, as well as interpret and evaluate their output has value and impact for myriad future careers beyond the classroom, and we would argue, for society as a whole. We want - *and* *will need* - our lawyers, journalists, entrepreneurs, business executives and social workers to have a well-grounded understanding of and appreciation for the power and dangers of generative AI.

From Mainframes to Datacenters to Generative AI

In order to understand why Research Computing now requires different thinking and why we cannot simply expand existing systems incrementally, it helps to understand the **evolution of computing** over the past few decades in the context of academic research.

Shared-use **Mainframes** dominated research computing in the 1970's and 1980's. Access to mainframe computers enabled the new internet, as well as large-scale computational simulations in the physical sciences, engineering analyses such as structural finite elements, computer aided design models for architecture, expert systems for medical diagnostics, and statistical data analysis —- a precursor of today's machine learning algorithms. However, these uses were mainly limited to experts in science and engineering fields.

As computer technology advanced in the 1990's and 2000's, personal computers became increasingly powerful and commoditized. Many researchers now had ready access to desktop workstations. The next phase of **shared** infrastructure shifted to large scale data storage and processing that could not be done locally. New **Data centers** offered centralized and secure handling of large amounts of data, which were too cumbersome, sensitive, or inefficient to be handled on personal computers. Data centers consisted mostly of single-core or few-core computers connected in clusters, running CPU-dominated database applications, large scale simulations, and statistical analyses. Applications like AI-driven heuristic search for supply-chain optimization or climate simulations required thousands of computers running in tandem, with access to common databases. However, this phase too, was mostly exploited by science, engineering, and medicine.

The 2010's introduced a new twist to shared computing in the form of machine learning using thousands of parallel cores (known as GPUs). Until 2012, not even the largest data center with the most sophisticated machine learning algorithms could reliably perform what we now consider relatively straightforward tasks, such as automatically telling the difference between a photo of a cat and a dog. The introduction of deep learning, a new architecture of machine learning based on neural networks that underlies most forms of AI today, removed that barrier.

Since 2012, the world has learned how to build *scalable AI* **– These are AI systems that can keep growing in size and improving in their capacity to learn every generation.** While we can anticipate further significant innovation and improvements in the algorithms, models, and even architectures, a fundamental and critical bottleneck has now become access to compute hardware and data[12]: The AI race began.

In 2017, yet another AI architecture emerged in the form of Transformers – a new Deep Learning architecture that unlocked generative and creative tasks (aka. Generative AI). Machines could be taught (trained) to generate new content that is loosely inspired by past content. Since 2017, generative AI systems have continued to grow at an astonishing rate and have made their way into nearly every field. In the past few years alone, AI has emerged with capabilities in some domains that appear to rival human abilities in both analysis and synthesis. Newer architectures that outperform transformer architectures are already on the horizon as we write this report.

Given these new AI abilities, the third wave of research computing that is sweeping across almost all disciplines today focuses on **accelerated idea generation**, not just analysis. Unlike the previous generations of computing, generative AI appears to be accelerating discovery and ideation also in **non-technical fields**.

New generative AI tools will change the way that medical diagnostics is performed[13–15], accelerate the ability of journalists to parse news in real time[16], enable legal scholars to interrogate and query unimaginably large corpuses of text[17–22], planners and policy makers to probe vast amounts of urban data, architects to *optimize* the performative qualities of large building complexes and artists to generate visual art and musicians to compose music[23–26], just to name a few fields experiencing dramatic transformation.

Materials scientists are deploying generative AI to help discover new materials at a rate that is orders of magnitude faster than what these experts could discover manually just a few years ago[27–29]. Researchers are predicting previously unknown protein structures predicted with high accuracy[30–34], and deploying AI to design novel proteins and molecules with specific functions. Researchers also deploy these tools to denoise, restore, and augment images and signals, enabling new observations. Companies and teachers are exploring new forms of personalized education now that machines can communicate fluently in natural language and across multiple human languages and dialects. There is also considerable focus on detecting, identifying, and classifying emotion from speech and video[35–40]. These trends will continue to accelerate as computing power continues to grow exponentially (Fig 1) and AI models continue to grow in capacity (Figs 2, 3), soon reaching into the physical world by controlling robots.

AI will change the very nature of the way we will **conduct and communicate academic exploration** across every discipline. While today's language translation is far from perfect, translation performance by objective metrics is improving every generation of AI. It is only a matter of time before automated language translation and generation will exceed the performance of the average human translator. We believe every field at Columbia needs access to computational resources and training in how to leverage generative AI and modern compute and data infrastructure.

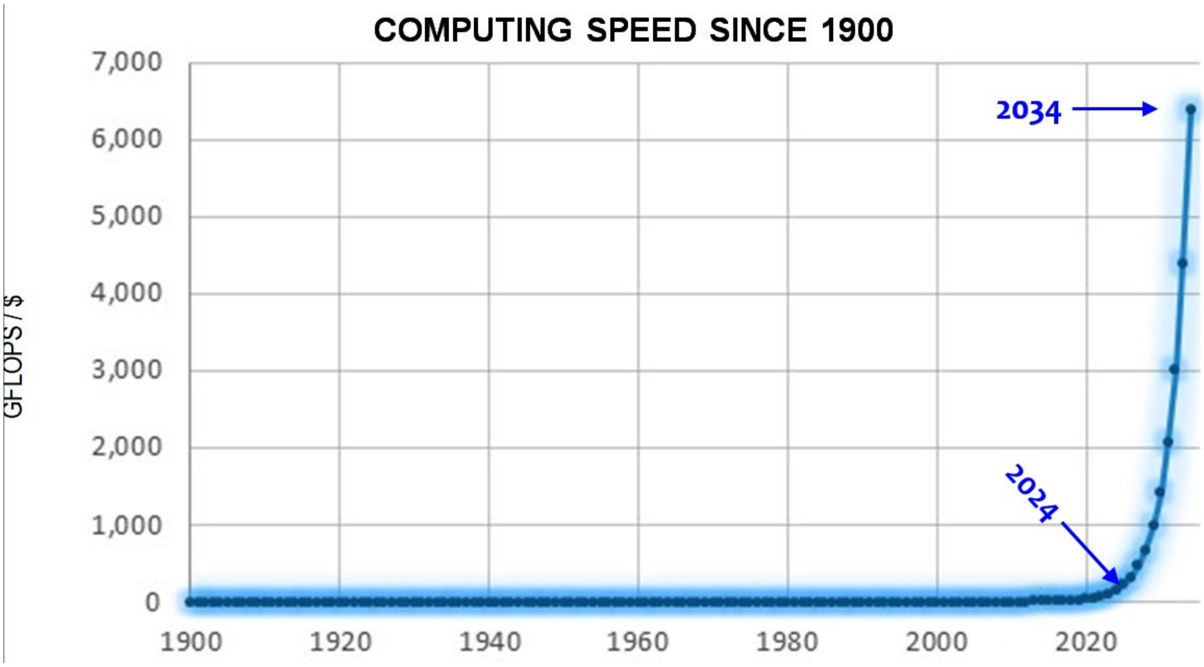


Fig 1. Expected growth in computing power measured in GFLOPS/$ in the next decade, based on continuation of existing exponential trends. Shown on linear scale to emphasize actual expected performance gains during the period covered by the recommendations proposed in our report.

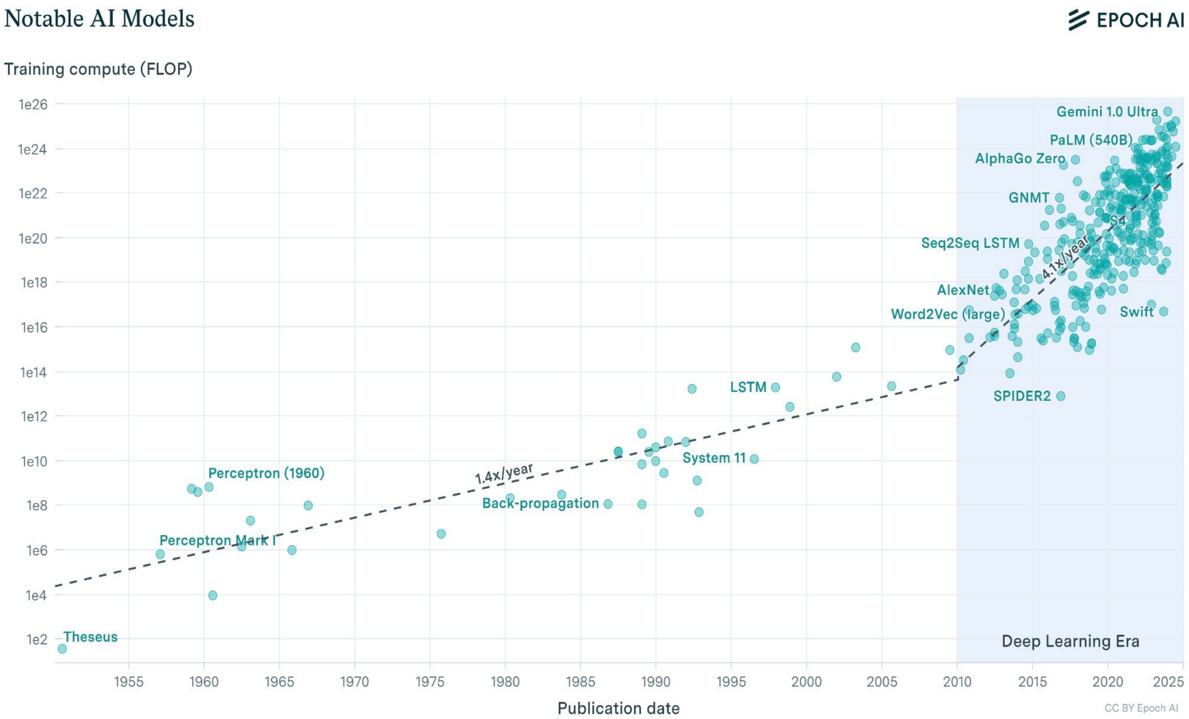
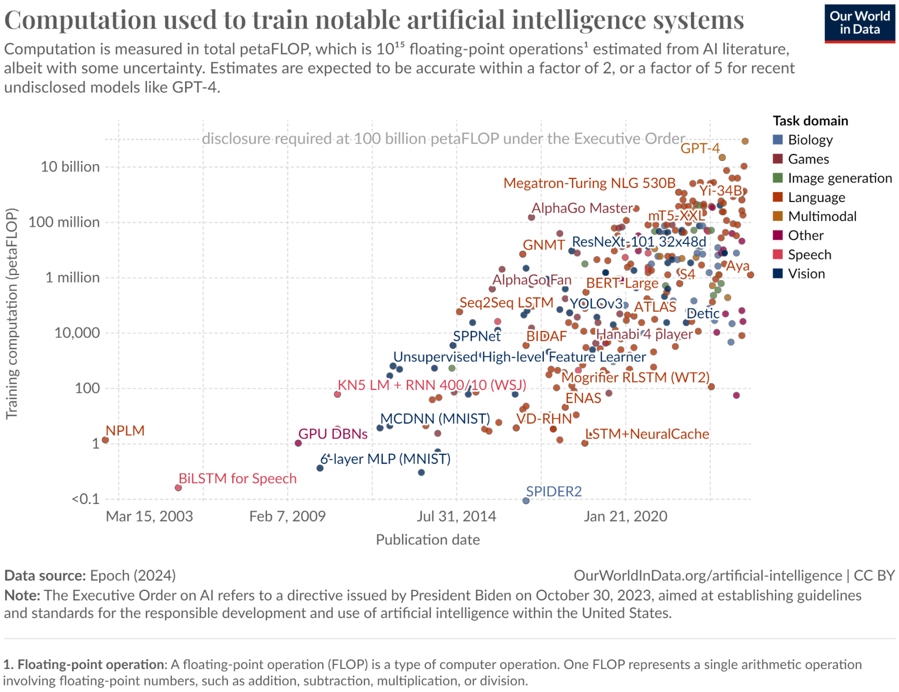


Fig 2. Growth of AI models in the past decade. Figure from Epoch AI[41] (log scale) shows a doubling approximately every six month



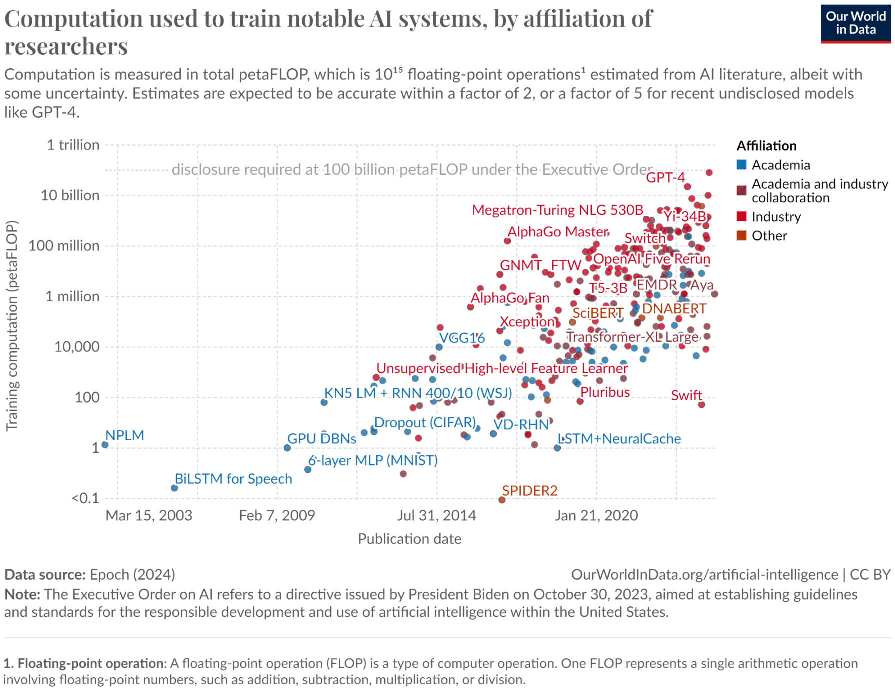
Computation used to train notable artificial intelligence systems

Computation is measured in total petaFLOP, which is 10^{15} floating-point operations[1] estimated from AI literature, albeit with some uncertainty. Estimates are expected to be accurate within a factor of 2, or a factor of 5 for recent undisclosed models like GPT-4.

Data source: Epoch (2024) OurWorldInData.org/artificial-intelligence | CC BY

Note: The Executive Order on AI refers to a directive issued by President Biden on October 30, 2023, aimed at establishing guidelines and standards for the responsible development and use of artificial intelligence within the United States.

[1] **Floating-point operation**: A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.



Computation used to train notable AI systems, by affiliation of researchers

Computation is measured in total petaFLOP, which is 10^{15} floating-point operations[1] estimated from AI literature, albeit with some uncertainty. Estimates are expected to be accurate within a factor of 2, or a factor of 5 for recent undisclosed models like GPT-4.

Data source: Epoch (2024) OurWorldInData.org/artificial-intelligence | CC BY

Note: The Executive Order on AI refers to a directive issued by President Biden on October 30, 2023, aimed at establishing guidelines and standards for the responsible development and use of artificial intelligence within the United States.

[1] **Floating-point operation**: A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

Fig 3. Computational effort per AI model. Growth of AI models in the past decade (log scale)[42,43]

Why existing solutions do not satisfy our growing AI needs

Currently, researchers can endeavor to meet their growing research computing demands in four different ways: (a) On-demand cloud services, (b) faculty-sponsored, shared high-performance computing clusters, such as Terremoto, Ginsburg, and Insomnia managed by the Shared Research Computing Policy Advisory Committee (SRCPAC), (c) Government supported federal and state compute resources, and (d) ad-hoc “under-the desk-and-in-the-closet” server clusters, which charitably can be called *laissez-faire* “edge computing”.

Many hoped that **cloud services** would provide a long-term elastic solution to growing computational needs. Cloud services sound ideal: they take care of the effort, cost and overhead of managing, maintaining, and upgrading equipment, as well as the potential for efficient sharing across fluctuating needs of multiple users. However, in practice reliance on cloud computing resources can often be risky, even for large commercial entities and other universities with deeper pockets than Columbia. Commercial cloud computing resources are expensive, and are sporadically limited, based on fluctuating industrial and commercial needs[44–47], leaving academia increasingly dependent on industry. As an indication that shared cloud resources are not sufficient, note that commercial entities are investing **billions** of dollars in their private wholly-owned resources[48–50].

When used only intermittently, commercial cloud services can be cost-efficient. Similarly, they may remain a preferred solution for peak use for some applications. However, under sustained high-utilization, commercial cloud services are substantially more expensive per core-hour than comparable private resources. Charging by the hour also makes it difficult to plan for open-ended research, and disincentivizes exploratory research into new areas and potential applications, especially for fields that are not traditionally strongly funded. Standard cloud computing with metered rates also presents significant risks for research continuation during periods of variable or otherwise insecure funding over University-owned resources.

The committee considered the possibility of recommending a policy change to eliminate overhead charges on cloud services, in order to make them more competitive compared with capital purchases, which have additional costs that are borne by the University. However, many counter arguments were presented as well. Efficient broad use of the cloud likely requires additional administration, personnel, resources, and network infrastructure and support, that are general to the University, rather than specific research projects, which is nominally the purpose of indirect costs. And critically, it is exactly because cloud services do not contribute to the University's capital accretion of computation resources that they *should* be charged overhead. We would like to encourage faculty to invest in capital accretion of computation resources, and therefore did not recommend that policy change.

A second existing approach is to deploy large shared computing clusters such as those Columbia has successfully deployed over the past decade. Given its overall high utilization, this infrastructure has proven to be cost-effective relative to similar cloud services. However, access to clusters such as SRCPAC has remained relatively limited to advanced users, due to the high upfront cost of joining the system, and advanced skills required to utilize the system. A typical single server will cost $21K-$43K (2023) and last for five years. Similar clusters exist in other schools such as ZMBBI, Computer Science, etc. We see SRCPAC and other department-level clusters as precursors to the Discovery Accelerator facility recommended by this report. Training for SRCPAC is severely oversubscribed, and does not span the range of needs required for the cross-discipline success, hence our recommendation 2, amplified in the following section.

Finally, many, if not most, faculty resort to ad-hoc edge computing, whether in the form of laptops, desktop computers, and high-end workstations, servers scattered across labs, closets, under desks and on shelves. Such solutions offer researchers guaranteed availability, relative cost efficiency (often consumer-grade commoditized components), administrative flexibility, and permanence beyond any specific project. Institutionally, however, the hidden costs are significant. Edge resources can be difficult to share and manage, they rarely have very high utilization rates, often lack coordinated back-up and archiving, and introduce numerous security vulnerabilities. They also incur energy costs from the associated power and cooling, and their delocalization limits targeted investments towards high efficiency cooling. Further, NYC’s new climate related Local Law 97 will soon increase significantly from fines related to excess energy use in buildings. Finally, local systems create heat and noise that adversely affects the working conditions of nearby researchers.

We emphasize that there is widespread recognition and enthusiasm for State and National-level government sponsored AI infrastructure such as Empire AI (which we also strongly support), however, many questions remain. First, we do not yet know the scale and timelines, leaving the University vulnerable and without sufficient computing power in the interim. Second, given that these centers will serve so many constituents, of varying resources and needs, it is likely that the allocations guidelines will include specific regional and national research priorities, and will also include important and necessary considerations of equity and geographic diversity, to ensure that governmental HPC-AI resources do not further exacerbate existing disparities in state and federal support.

Thus, while it is clear that any government investments in state or national AI-infrastructure will aid and support the University and its research, we predict very high demand for these limited resources and expect rapid growth of need. Therefore, we believe that **academic institutions that aim to lead in AI-enabled research, must be able to satisfy many of their own computation needs <u>independently and sustainably</u>.**

Outline of methodology

The committee progressed according the following phases

1. Establishing the general landscape of options and levels available to the university, ranging from creating new services, expanding existing services, or outsourcing services
2. Surveying and interviewing faculty, students and staff to understand where the computational needs of their fields are heading in the next decade
3. Deciding on a set of long-term goals and short-term actions given these objectives and options

From April 2023 until the submission of this report, the committee has met regularly roughly twice monthly, excluding summer and winter breaks. The original committee was expanded to invite representatives from all schools.

After the first initial meetings debated the perimeter of the study, the committee designed a survey to capture needs and thoughts of as many stakeholders as possible. We endeavored to keep the survey as short and as open-ended as possible and to deliver it to many researchers at the university, including faculty, research staff, and postdocs. The survey as well as the raw anonymized responses are provided in the appendix. The goal of the committee was not to address all issues raised in the survey; rather the survey was meant to ensure that the needs and viewpoints of as many stakeholders as possible were being considered.

Following the survey, the committee asked representatives from each school and college of the university to perform their own analysis, write an interim report, engage in follow-up interviews and in-person discussions, and make recommendations suitable for their unit. These individual summaries are included in the appendix of this report with the individual unit and authors of those analyses listed. These interim reports do not represent the official position of those schools, but merely a subjective point of view offered by the school's representative based on interaction and interviews with their colleagues, review of the survey results, and personal familiarity with their fields. Some schools and representatives did not contribute a position.

Finally, after all interim reports and recommendations were completed, the committee set out to identify concrete recommendations for consideration by the university. These recommendations were discussed within and outside the committee, and feedback from various stakeholders were considered and debated.

Survey Results

Both quantitative (survey) and qualitative (interview) tools employed by this committee provided discipline-specific metrics regarding research computing. The appendix lists all survey results (anonymized) reflecting feedback from approximately 185 faculty, and 50 postdocs/associate research scientists (in addition to the valuable feedback and contextualization provided by committee members). Here, we organize the feedback into three areas: Basic Infrastructure, Existing High-Performance Computing, and New Needs for AI.

Basic Infrastructure

The survey exposed an overwhelming perception that Columbia University lacks the campus-wide basic infrastructure both to support current use and to accommodate inclusion of the diverse user-base anticipated by the rise of AI across all disciplines at the University. Responses to our survey highlighted significant infrastructure disparities across buildings, departments, and schools—including basic needs like reliable internet connections. While some differences in computing infrastructure are expected across fields, modern network infrastructure, connectivity, and training are essential now. These structural deficiencies inhibit teaching and operations of University functions, and fundamentally impede research effectiveness and productivity both within and beyond the scope of this committee. Moreover—in research especially—data, education, collaboration, and communication are all critical currencies whose importance will expand exponentially in the near future. It is incumbent on the University to remedy these structural deficiencies to ensure that the benefits of our proposal are broad, accessible, and equitable across the University. Specific areas identified through our survey and discussions include:

- Poor and inconsistent networking (e.g., insufficient wifi, low bandwidth, and bad remote access limit)
- Poor communication and ineffectively structured website information about free or low cost access to existing computer resources
- Existing faculty personal computer purchasing mechanisms and funding structures are insufficient to promote transition to research computing across disciplines
- Training is either non-existent, not available for users at different levels of experience, or poorly advertised
- Slow IT responses and expensive solutions whose cost falls back to departments
- Inefficient bureaucratic processes regulating data acquisitions and access to systems
- Lengthy and laborious process for providing access to our systems to outside collaborators
- Lack of site licensing for commonly used productivity software and services such as OneDrive, Zoom, Adobe, etc.

Existing High-Performance Computing

For existing HPC, our survey shows that Columbia researchers either use “local” clusters, which may belong to one research group or a department, and/or the University’s shared HPC clusters, governed by SRCPAC. Many survey responses described concerns about the existing shared HPC resources, which can prompt the decision to use local clusters. More flexibility and better support would entice many users to join in shared computing, which is more efficient in terms of energy, space, and cluster usage. For context, the existing HPC clusters are purchased by individual faculty members who can now purchase between a quarter and multiple nodes roughly four times each year. Typical options include: a standard memory CPU node, a high-memory CPU node, and a GPU node with different types of GPUs. Currently, one standard node costs about $16,000, with the other nodes more expensive. (This exact model is changing slightly at the time of writing, with the introduction of the Insomnia cluster). The survey demonstrated consensus that access to this current HPC system introduces structural frictions that reduce research effectiveness, some of which are real, others perceived. In preparation for an explosive growth in University research computing, the survey produced valuable insights into issues with our recent and existing HPC. Specific concerns include:

- Cost is perceived as neither competitive nor transparent
- Lack of training, so people do not know how to make the most of existing HPC
- More flexible purchasing plans to lower the barrier to entry and/or distribute high one-time costs over 3-5 year grant cycles
- Greater selection, e.g., centralized specialty clusters in addition to “generic” HPC
- Dissatisfaction with queue times or time limits for big jobs
- Lack of experts that can facilitate running large jobs
- Lack of access to (or high cost of) large RAM systems (1-2 TB)
- Continued access for recently graduated students, postdocs, and external collaborators
- Easier access to big data
- Better GUI support and interactive tools.

New Needs for AI

Unsurprisingly, our survey revealed many faculty who are eager to ramp up research in AI but are held back due to insufficient resources and expertise, which forms part of the basis for the present proposal. The training and testing of AI models requires storage and manipulation of large datasets and extensive specialty (e.g., GPU) hardware, which survey respondents did not presently have access to, but were seeking for their future research. Similarly, respondents are excited about the potential for AI in their research, but they (or their students/postdocs) need training and support to engage in this relatively new research area. Specific requests or suggestions include:

- Significant increases to storage – which needs to be accessible on campus as well as collaborative partners outside Columbia – and needs to be secure (for data with personally identifiable information), but flexible sharing, backed up
- Better data stewardship (e.g., the FAIR principles: findable, accessible, interoperable, and reusable) and lifecycle management

- Advanced training in specific areas
- Basic training (e.g., general AI and cloud computing)
- Some expert staff to assist researchers
- Cloud integration – how and when to move, and manage cloud resources

The research computing “pyramid of needs”

For clarity, we have sorted the needs emerging from surveys, interviews, and discussions into five levels, based on a scale of ascending requirements. We model these five levels after Maslow’s classic Hierarchy of Human Needs. The base level shows foundational computing needs, without which the upper levels cannot be achieved, followed by the second tier, which represents the end-user service infrastructure required to pursue academic research across all disciplines. While outside the purview of this committee, we believe these **basic first and second level needs must be fully addressed**. It is a requirement that the university (CUIT) address and maintain all evolving infrastructure requirements and basic services required by all research faculty staff and students in a timely manner.

The recommendations of this committee directly address the third, fourth and fifth levels of the hierarchy – enabling skilling, supporting guided exploration, and facilitating self-driven research, all intellectually-driven by disciplines themselves. While centralized units such as CUIT may be directed to tactically implement and support many of these higher-level needs, the **scope of the needs must originate in the disciplines themselves and the disciplines must retain agency in their application and growth.** For example, it is unreasonable to expect that a central administrative unit will decide and take responsibility for core decisions such as what language models best suit research in the Journalism school, or what suite of software tools is best for the Business School or the Medical School. The schools themselves must drive these decisions, subject to financial and logistical constraints.

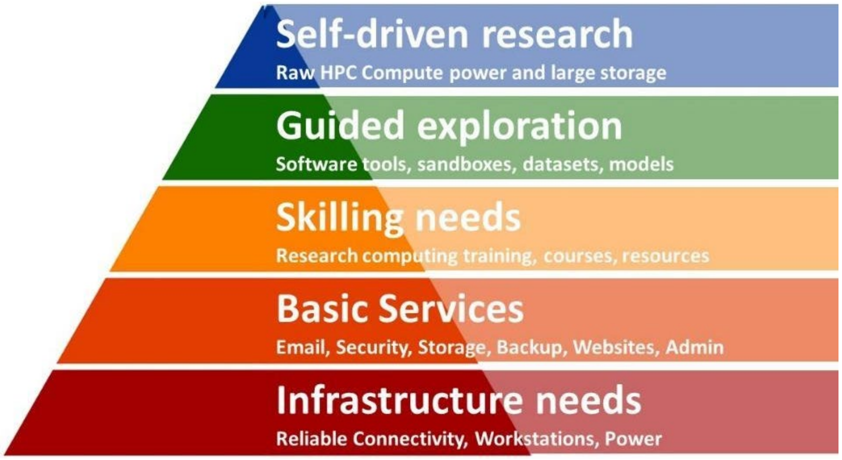

Self-driven research
Raw HPC Compute power and large storage
Guided exploration
Software tools, sandboxes, datasets, models
Skilling needs
Research computing training, courses, resources
Basic Services
Email, Security, Storage, Backup, Websites, Admin
Infrastructure needs
Reliable Connectivity, Workstations, Power

Fig. 4. The research computing pyramid of needs

Summary of emerging needs

1. **Computing power is like a utility**. Basic and even advanced research computing is increasingly assumed to be available much like electrical and communication infrastructure, and physical space. Modern research relies very heavily on the physical instantiation of "Virtual Infrastructure" – the compute, storage, and networks needed to support it. The rise of HPC-AI has elevated the minimal level of infrastructure necessary for research significantly, and the university must make commensurate investments to match. This pressing need is like that of construction, renovation, and maintenance of buildings or laboratories, but serves the entire University.. A non-trivial level of research computing has to be available by default within the University without individual planning.

2. **Sharing resources is essential.** No single researcher can acquire, operate, and maintain the scale and amount of compute and storage that will be needed to conduct cutting edge research in the future. This applies equally to much of the expert staff needed to operate and enable high-impact use. Many resources must be shared. Sharing also encourages high utilization and effective use of the resources, while also facilitating collaboration.

3. **AI is key to creativity**. Deep learning and especially generative algorithms, applied across all domains, such as genomics, material discovery, law, music, arts, journalism, architecture, medical diagnostics, are fast becoming a means to generate new ideas and discoveries. As a university whose entire reputation is based on the ability to discover and

innovate, access to generative AI tools and the ability to modify existing AI tools and develop new ones are key to remaining competitive and even relevant.

4. **Data is the asset**. AI algorithms are increasingly open source, and computing resources, though expensive and in limited supply, are commoditized. A remaining differentiating asset is **data**. For example, end users can get many services for “free” from Google -- but not access to much of Google’s collected data. Data is also crucial to creating unique generative AI capabilities[51–55]. This is certainly true for domain specific foundation models, but also for diverse training data that includes the margins and rare, but important, phenomena. Effectively managing this data throughout its lifecycle is crucial. Robust storage solutions are needed to preserve and share valuable research findings and ensure regulatory compliance. However, simply storing data isn't enough. Lifecycle management practices ensure data is properly categorized, secured, and archived or purged according to its relevance and legal requirements. Therefore, universities must enable researchers to collect, curate, share, and manage data assets reliably, safely and ethically. Cost-effective scalable solutions should be available by the University, either on-site, or a mix of on-site and the cloud resources. This approach should include support and expertise in data lifecycle management, standards, FAIR principles[56,57], storage, curation, and dissemination. Increasingly, data curation and dissemination is also required by Federal research sponsors[58], providing even further incentives for efficient management. Here, CUIT and the University Library can play a key role, much as they have in the past.

5. **Small projects and departments need big resources too**. While it is possible to get government funding and computational resources for mature projects in science and engineering, academic innovation often emerges from many small projects that have no dedicated federal or state funding, and no central administration. Increasingly, these “small” projects rely on sophisticated compute and data resources. This is also true in domains outside of STEM that traditionally were not typical users of HPC, and whose overall funding models may differ across all stages of research (e.g. often supported through internal or private as opposed to Gov’t sources[59].) Many of these projects, as small as they may start, will progressively require large computing power to germinate. Without access to sufficient resources, the innovation stream will slow.

6. **Commercial cloud resources do not match the nature of academic research**. Cloud computing is not a solution to baseline research computing needs. It is expensive, has

limited availability especially at peak times, and is risky in terms of runaway costs and planning. Most importantly, dependency on cloud computing (commercial or government provided) makes academia subservient to industry and government priorities. Cloud services can serve as an important add-on to handle peak loads, temporary fluctuations, or commoditized jobs, and will certainly be a component of Columbia's plan going forward, but it ***cannot*** be the core of our research and data infrastructure.

7. **Cost structure of computing matters.** Enterprise-class computing hardware can be purchased and supported according to a specific cost and renewal schedule that is aligned for typical STEM funding cycles and can be incorporated into shared-facilities that allow for high overall utilization. This shared resource can also be flexibly allocated and proportioned to support access tiers, yet still allow researchers and departments access to modern and powerful compute and data resources. Also, many of the costs of the virtual infrastructure are also shared across (and benefit) the entire university. In contrast, cloud computing is a tightly metered service that incurs significant overhead, and is only available during the performance period of a grant, and provides no future residual value. Ad hoc, or local assets are often the lowest cost, and remain available indefinitely, at low marginal cost to the investigator, though offer lower opportunity for broader sharing and utilization to others vs enterprise-class resources. Faculty consider these cost structures when deciding which form of research computing to use and invest in. The University must support continued access to resources to encourage and promote trust in these shared centralized assets.

8. **Training and support is essential.** Training and support staff are critically needed for training students, faculty, staff and new users. However skilled professional personnel are costly, especially with increased industrial demand. Central coordination of support and expertise at the hardware, data, and runtime engineering are essential to broadly support the university, but cannot scale to provide hands-on domain specific support to every unit or department. Therefore, we expect the disciplines themselves to drive and organize much of the domain specific training (with some central support) rather than being wholly centralized. Further, a substantial collection of training material and resources (eg. free tutorial videos) exists online to cover many introductory and even advanced topics.

9. **Data security and privacy needs to be considered from the start.** Data privacy and security are becoming increasingly important for analyses of granular geospatial data, personal identifying information (PII), and protected health information (PHI), among social,

behavioral, and health scientists across the University. Some can only be accessed on security-certified platforms (e.g. HIPAA certified for PHI). Data breaches are costly due to loss of reputation, fines, and remedial actions. Privacy must be maintained for research study subjects, patients, and employees. University intellectual property must be protected. Therefore, the university must protect the digital storage environment and compute facility against external access to private data. Service providers must enter contractual agreements to protect private data, and infrastructure must be put in place to authenticate and authorize users for accessing private data. The system must support the minimum-necessary paradigm, where users are given access to only those aspects of private data that they need to carry out their current task (e.g., hide patient identities if they are not needed for the current study). A formal governance structure must coordinate the access to private data and administrative coordination across University units involved in purchasing, contracts, user authentication, and systems access authorization must develop clear and streamlined processes for data acquisition, sharing, and access. Meeting Federal Information Security Modernization Act (FISMA) compliance requirements around governance and processes for privacy and security will enhance privacy and facilitate government contracting.

10. **Solutions must be sustainable financially, procedurally, and environmentally**: Whatever resources are chosen, they must be renewable, i.e. upgradable every five years or less. We need a faculty-led standing committee to manage the renewal process and an endowment to support it. The resources should also be highly energy efficient to reduce energy consumption and run on renewable energy, where possible.

11. **Now is the time to raise funds**. Many potential donors interested in having an impact on research understand that the most long-lasting donor-attributable impact may be had by providing the university researchers access to computational resources and training to catalyze and amplify their innovation ability. Many potential benefactors made their fortunes using AI and are well aware of its power. Funders also get more excited about funding pioneers than asking to support catch-up efforts. Now is the time to raise and the window is short because AI is a “winner takes all” game.

Recommendations

We believe that the two key recommendations below will enable Columbia University to thrive in an era of where research is increasingly enabled by high-performance computing and AI.

1. Create the Discovery Accelerator

The University should raise and invest $25M per year (=$250M over a decade) to create and sustain a facility which we tentatively call the Discovery Accelerator.

1. Create a University resource (on- or off-premises) that will comprise necessary computational resources and staffing to operate a mid-sized computational capacity of approximately *60* fully-loaded computational racks with various evolving combinations of GPUs, CPUs, and high-performance storage (and whatever new computing technologies become available), and have flexible architecture to support both supercomputer-style, and node-style workloads. This unit should be lean and efficiently staffed, while offering "containerized" management (i.e., the ability to run custom virtual machines) and security and sharing controls that allow for facile collaboration, and protection. The Accelerator must also support secure data enclaves for discovery on protected or private data. Users must be able to dock and control their own applications and virtual machines, with technical assistance if needed.
2. The proposed budget is intended to sustain operation for about a decade, including hardware upgrades on an ongoing basis. This investment will ensure that the university enables ANY of its faculty and researchers to embark on an AI-accelerated future sooner rather than later.
3. During this first decade, the university will have the opportunity to raise additional funds and explore ways to systematically allocate operating budget, based on the financial analysis of operating expenditure of the accelerator. These calculations will factor such aspects that are difficult to anticipate at this early time, but will likely include the increased grant competitiveness, generated revenue from new or enhanced research and discovery enabled by the accelerator, benefactor interest in supporting specific research, educational, commercial, and societal impact generated by the accelerator, or AI more broadly. We also anticipate some funds to come from in-kind or cost-sharing contributions as well, for heavy or prioritized users. In other words, **faculty will partner with the**

University, and prove the essential value of University investments in the discovery accelerator over the next decade, and help sustain it beyond.

4. Implementation of the accelerator should not be postponed or linked to new construction or acquisition. The accelerator might initially be housed in existing spaces, such as Uris hall basement, power and space permitting. External spaces including new or leased space with low communications latency to data are also a possibility Considerations for "off-prem" should evaluate proximity, cheap reliable power, and high-bandwidth interconnects and peering, and may offer significant savings on operating costs (e.g. LL97). We have identified the Nevis and Lamont campuses as a prime location for this purpose.
5. Government-provided and industry-provided cloud services (such as Empire AI, Microsoft Azure, Google Cloud and Amazon AWS) should not serve as the core accelerator since Columbia does not have direct control over their deployment and allocation priorities. However, the University must still engage vendors and other partners to negotiate favorable rates and access, for both routine compute and storage, as well as peaking capacity.
6. A faculty committee reporting to the EVPR should be established to govern the operational and technical aspects of this facility and direct its ongoing development (see Governance section).

High level financial analysis

Committee members have discussed financial aspects of this recommendation with numerous stakeholders and knowledgeable persons including faculty and staff at peer institutions, people building and managing data centers at companies and universities (e.g. Apple, Google), sales representatives of large equipment vendors (e.g. Nvidia) who are intimately familiar with activities at peer institutions, as well as CUIT staff at Columbia familiar with existing costs and expenses.

We considered two prototypical "fully loaded racks", one for GPU and one for CPU. Both these systems are state of the art (SOTA) at the time of this writing, and the exact configurations should be selected by the faculty leadership and CUIT at time of purchase. Naturally, as time passes the technology will improve, and we expect the performance to increase exponentially but we also assume that the price per SOTA rack will remain roughly constant. We also factor equipment upgrades every five years on a rolling basis (= 20% of the equipment every year).

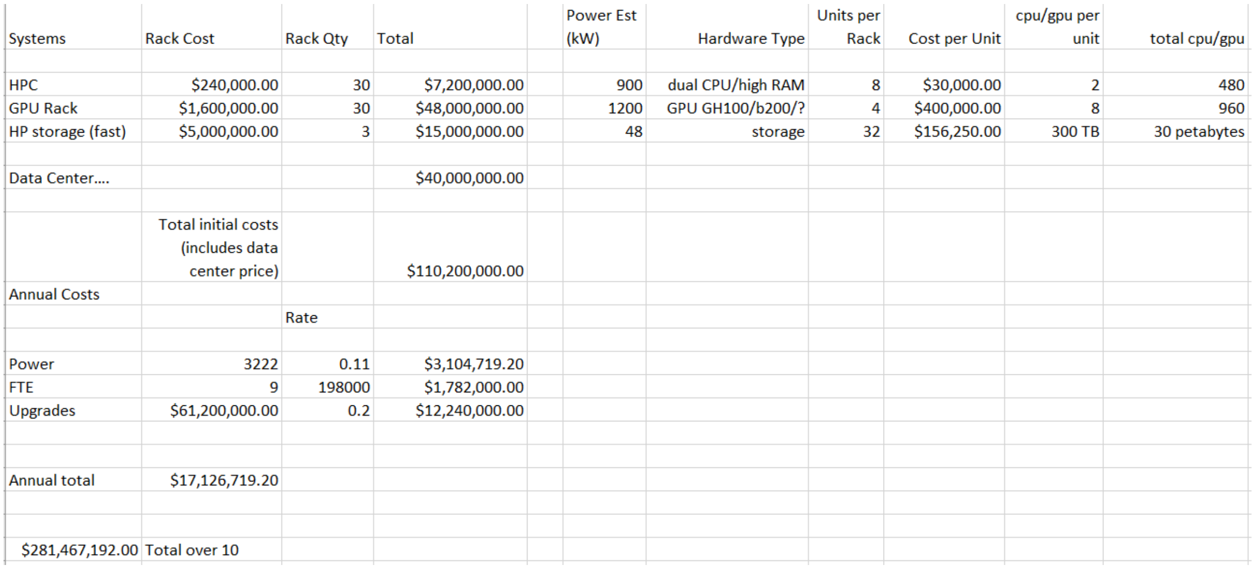
A fully loaded CPU rack contains 8 nodes, each with Dual Intel Xeon (28 cores each) 16 Xeon CPUs per rack. Total rack cost **$240K**

A fully loaded GPU rack contains 4 nodes, each node is an Nvidia DGX style module, each with 8 GPUs, 640 GB GPU memory, 2TB system memory, Dual Intel® Xeon® Platinum 8480C Processors. 32 GPUs per rack, 8 Xeon CPUs per rack. Total rack cost **$1.7M**

A high performance data storage rack is based on ~ 10 PB rack of enterprise storage (Dell Isilon H7000/A3000 mix or equivalent), with an expected 10 year cost of **~5M per rack**.

Power cost. The following calculations are based on $0.11 per kWh, which is a lower rate than in NYC, but is available at some regional data center locations, as well as the area around the Nevis Campus. The area near Columbia's Lamont Campus has even lower average rates, on the order of $0.08. Rates are likely to go down over the years as renewables are introduced into the grid.

Estimated build times: 3-5 months design/engineering, 1-2 months permit/bid, 12-23 months construction, 1-2 months commissioning



Systems	Rack Cost	Rack Qty	Total		Power Est (kW)	Hardware Type	Units per Rack	Cost per Unit	cpu/gpu per unit	total cpu/gpu
HPC	$240,000.00	30	$7,200,000.00		900	dual CPU/high RAM	8	$30,000.00	2	480
GPU Rack	$1,600,000.00	30	$48,000,000.00		1200	GPU GH100/b200/?	4	$400,000.00	8	960
HP storage (fast)	$5,000,000.00	3	$15,000,000.00		48	storage	32	$156,250.00	300 TB	30 petabytes
Data Center....			$40,000,000.00							
	Total initial costs (includes data center price)		$110,200,000.00							
Annual Costs										
		Rate								
Power	3222	0.11	$3,104,719.20							
FTE	9	198000	$1,782,000.00							
Upgrades	$61,200,000.00	0.2	$12,240,000.00							
Annual total	$17,126,719.20									
$281,467,192.00	Total over 10									

Estimated costs for construction and operation of the Discovery Accelerator over a 10-year period. Costs include setup, computers, power, and personnel. Costs do not include real estate.

Possible geographic location of the accelerator: Nevis or Lamont Campuses, or surrounds.

Columbia's NYC campus has several limitations for hosting the Discovery Accelerator hardware. First, power and cooling are limited and existing facilities cannot support this scale of hardware without major renovation and upgrades. Second, the utility rates in NYC are high when compared to surrounding areas. Further, the high-energy requirements for the system could hamper the University's ability to meet Local Law 97 requirements[9], leaving us susceptible to recurring fines. We note that many large companies involved in AI tend to own their own physical data centers rather than using cloud services for reasons of access control, operational cost, and real estate investment. From informal discussions with service providers, we found that large companies (e.g. Microsoft) tend to own about half of their data centers, and rent others from service providers. Whether they rent or own depends generally on whether they have access to the **land and power** in the desired geographic location; they will rent only as a second choice. When companies rent, the datacenter will generally provide only the infrastructure to house the computers: space, network interconnects, power, cooling, and physical security. The computational hardware will be owned and operated by the company.

Luckily, Columbia owns several locations where it has both land and power at the needed scales, including Lamont, Nevis, and other locations upstate. Both the Nevis and Lamont campuses are outside NYC limits, and have lower utility costs. As an example, the Nevis campus has relatively unused physical structures (e.g. the now defunct Columbia Press building) that may be cost-effective to repurpose. The area around Lamont Campus has significantly lower average energy prices than NYC or near Nevis. Moreover, both are close enough to the main campuses to provide low latency, as well as within 30 minutes drive from campus to allow for physical access when needed; Lamont has an hourly weekday shuttle from the Morningside and Manhattanville campuses. At the same time, the campuses are far enough from NYC that they are subject to lower power costs and not subject to power usage fines. Both these sites could be considered possible locations for the Discovery Accelerator Facility.

Cost Comparisons to Cloud Resources

One persistent question that frequently arises before making capital investments in compute resources is the question of on-premises versus cloud-based platforms that are available as a fee-for-service. Columbia-owned resources offer full control of both the environment and availability, and can easily be directed towards University research goals. But it also demands upfront and continuous investment in hardware, software, and IT staff, in addition to space and power. Cloud services, on the other hand, offer a pay-as-you-go model. One only pays for the resources they use, making it more flexible and potentially cheaper for rapidly fluctuating workloads.

However, cloud costs are always metered, and can add up significantly over time, especially if the usage is not carefully managed. (We note that poorly managed jobs are also a local concern, either with real costs from power consumption or administrative management, or opportunity cost for lost cpu/gpu cycles.) Under high-use scenarios, cloud resources are generally considered expensive compared with owned-resources, something that other Universities also report[60]. This can be seen in real time, on one web page given facts about one of Stanford's clusters, "Sherlock", where a running estimate is provided for equivalent costs of running the cluster's compute on on-demand cloud instances instead, which for this current month (June 2024) would be over 3.2 million dollars[61].

Also, for cloud use, some scientific suites or programs may need to be rewritten and ported. Data ingress and egress costs can be significant if large data sets need to be moved, though this is also a concern if the local resource needs to access a cloud based data set. Ultimately, the relative value or expense depends on the specific needs, usage patterns, and overall organizational goals.There may also be some indirect implications on value, relative to the nature of Cap-ex versus Op-ex expenditures, or other costing principles, which are outside the scope of this report..

It is the opinion of this committee that having a significant Columbia-owned resource is worthwhile, especially if it is governed to drive broad and effective adoption across many disciplines, and thus has overall utilization. One expected use of the Discovery Accelerator is the training of domain specific foundation models which require enterprise class GPUs. For reference, we compare the three year cost of an instance on AWS with eight H100 GPUs, assuming 80% utilization. Using the on-demand rate, the cost would be $2.07M[62]. With the 3 year reserved cost, it would be ~$908K. Prices drop further with providers like Lambda[63], which offers similar instances for $587K, By comparison, the University's expected costs over that same period would be ~ $460K. The proposed Discovery Accelerator would have ~120 such machines, resulting >5M per year differences against Lambda, and >18M per year against the "3 year reserved"

Amazon rate. The University cost estimate above includes power and data center costs, but not the operations staff, which is expected to be considerably smaller than this differential.

		AWS Cloud Costs (as of June 2024)			**Lambda Cloud**
Instance type		p3dn.24xlarge	p4de.24xlarge	p5.48xlarge	8x NVIDIA H100 SXM
On Demand per hour		$31.21	$40.96	$98.32	$27.92
3 yr reserved per hour)		$10.42	$14.46	$43.16	
Effective Hours per year	7012.8				
Three year OD		$656,608.46	$861,732.86	**$2,068,495.49**	**$587,392.13**
Three year Res		$219,220.13	$304,215.26	**$908,017.34**	
University Owned					
Estimated Hardware Cost					$400,000
		Power and Cooling (kW)		Over 3 Years	
Power per machine (kW)	10.2	17		$39,341.81	$39,341.81
Spot rate (kW/hr)	$0.11				
NJ datacenter cost[64] ($ per watt over typical life)	$11.40	$19,380.00			
		(cost over 3 years)	University cost over 3 years (not including staffing)		**$458,722**

That said, we expect cloud resources will also be used, and will complement the Discovery Accelerator for peaking capacity, hardware diversity, and particular usage types. We expect the governing body of the Accelerator will periodically assess both the relative costs and expected utility to optimize the overall benefit to the research community and University, just as they will review the hardware selections during refresh and update cycles.

2. Enable all disciplines to engage in Research Computing

The University should invest $5M per year over 10 years (=$50M in total) to facilitate disciplinary research computing engagement by hiring faculty and staff and offering grants for creation of educational materials.

To support the diverse needs of Columbia's research community, the university needs to balance centralized and local, discipline-specific support of the Discovery Accelerator and related research computing tools and services. As is the case today, research support varies greatly by discipline and by the level of funding and expertise available to different research communities. The recommendation seeks to enable all disciplines interested in engaging in research computing to do so, by providing a range of services and options aligned with faculty and researcher needs, including those managing sensitive/restricted data.

A faculty advisory committee will provide guidance and direction to facilitate the best use of centralized support resources, and also work with academic departments and schools to help them drive their own development of support resources for research computing.

The proposed budget of $5 million per year includes researcher-focused staff positions in CUIT and Libraries, funding to assist local/discipline-specific support resources, funding for tools, and funding for a mini-grant program to accelerate adoption across disciplines. This funding will support the following:

1. **Computational Training** - provide comprehensive computational training to support all disciplines to use research computing including but not limited to AI/HPC. This will build upon the Foundations for Research Computing program with a particular focus on fostering local departmental and school training and developing curricula for introductory to intermediate levels of expertise. Centralized training will be provided jointly by CUIT and Libraries staff along with coordinated efforts with local schools and academic departments. Suggested: 4 full-time equivalent (FTE) positions.
2. **Research Data Training** - provide comprehensive research data training to support all disciplines in creating, managing, disseminating and preserving research data and research outputs. Data Curation[65] and data management plans are of particular importance, including but not limited to generative AI and large language models, as well as support for the licensing, copyright and use restrictions that may accompany acquired datasets. Faculty also submitted a set of recommendations in the 2022 Provost Advisory Committee for Libraries Year-End Report, including a coordinated effort by Libraries, CUIT and EVPR to foster a greater understanding and use of data storage and sharing platforms, and promote greater visibility of reliable data storage platforms to the Columbia researcher community. Suggested: 4 FTE positions.
3. **Discipline-Specific Training** - provide funding for specialized support within the academic disciplines, in order to apply the appropriate technologies and infrastructure

needed for particular research and scholarly needs. This distributed support will work closely with CUIT and Libraries to help enable all disciplines to engage in AI and research computing. Suggested: 8 FTE positions.

4. **Develop Training Materials -** develop and/or acquire training materials to support all disciplines in using research computing and AI/HPC. In many disciplines, developing training material will require faculty and researcher engagement to develop specialized educational materials, balanced with a need for more general training and curricular materials developed by Libraries, CUIT and EVPR. Libraries can also support faculty in the creation, discovery, publishing and dissemination of discipline/domain-specific educational materials, via the Academic Commons and published as open educational resources (OER), in order to accelerate knowledge sharing across the university. Suggested: 1 FTE position.
5. **Experimentation and sandbox services and tools -** develop and/or acquire resources to support experimentation in new and emergent research technologies to foster use across all disciplines. In this quickly changing technical environment, it is crucial to provide accessible methods to try out new tools, technologies and computational methods, especially for those faculty, students and researchers who may lack sufficient local capacity. These tools will foster experimentation and testing of large language models and related systems, provide a space for evaluating computational methods and data curation with support from CUIT and Libraries, facilitating training and education, and accelerate knowledge transfer of emergent computational methods via active exploration and experimentation. Suggested: 4 FTE positions + $350,000/year Tools Budget.
6. **Enhance information literacy and competencies** - curate, catalog and provide discovery to collections of computational methods, create incentives to share workflows to accelerate research activities, and address issues of digital rights, algorithmic bias and trust. Enhancing curation and discovery will build upon Libraries existing Academic Commons and CUIT Columbia Data Platform and related programs to support faculty and researcher dissemination of research activity, address scientific reproducibility and accelerate knowledge transfer and use of new computational methods. We advocate for open, documented, transparent methods and approaches to support the mission and values of Columbia University. Suggested: 1 FTE position.
7. **Support long-term sustainable access -** foster discovery, access and re-use of research output, as required by federal funding agencies, by enabling researchers to deposit large

language models, code, datasets and other machine learning tools into approved/appropriate repositories such as the Academic Commons, Columbia Data Platform, and appropriate domain-specific data sharing repositories. Faculty and researchers defined this specific need in the 2022 Provost Advisory Committee for Libraries Year-End Report, to “encourage the migration of Columbia researcher data to reliable and secure storage platforms that meet compliance requirements and align with best practices in data management.” Libraries will enhance current efforts to foster description, deposit, discovery, management and educate faculty and researchers in author rights including license, rights attribution and copyright. CUIT will enhance current efforts in providing research data storage, discovery, analysis and collaboration. Suggested: 1 FTE position

8. **Consultation/Support for Building & Customizing Large Language Models -** provide consultation, training, documentation, awareness and onboarding for students, faculty and researchers in the building, use and application of large language models and machine learning tools. CUIT should expand current efforts to provide general consulting and advisory services to assist faculty in the development and use of large language models and related technologies, connect researchers across disciplines to foster consistency and reduce duplication of efforts, and provide liaison support for domain/discipline-specific departments in developing their own technical expertise. Libraries should expand current acquisition efforts to acquire and manage data sets in support of large language model development, retrieval-augmented generation, and vector databases, and emergent techniques, as well as provide guidance on copyright and use restrictions. Suggested: 1 FTE position.
9. **Researcher Training Support -** provide one-time funding to support training and foster adoption of AI/HPC, help faculty and researchers experiment and learn how to use AI/HPC, and encourage the adoption and creation of computational methods to support AI/HPC across all disciplines. CUIT, Libraries and EVPR should work with the faculty advisory committee and academic departments to administer one-time mini-grants and provide support as needed to accelerate adoption of AI/HPC in new and emergent areas. Suggested: Approximately $800,000 per year for up to 20 mini-grants per year.

Governance

The committee discussed what governance architecture would be best appropriate to manage the proposed recommendations. For example, should it be under the EVPR, or under CIO, or in some joint responsibility.

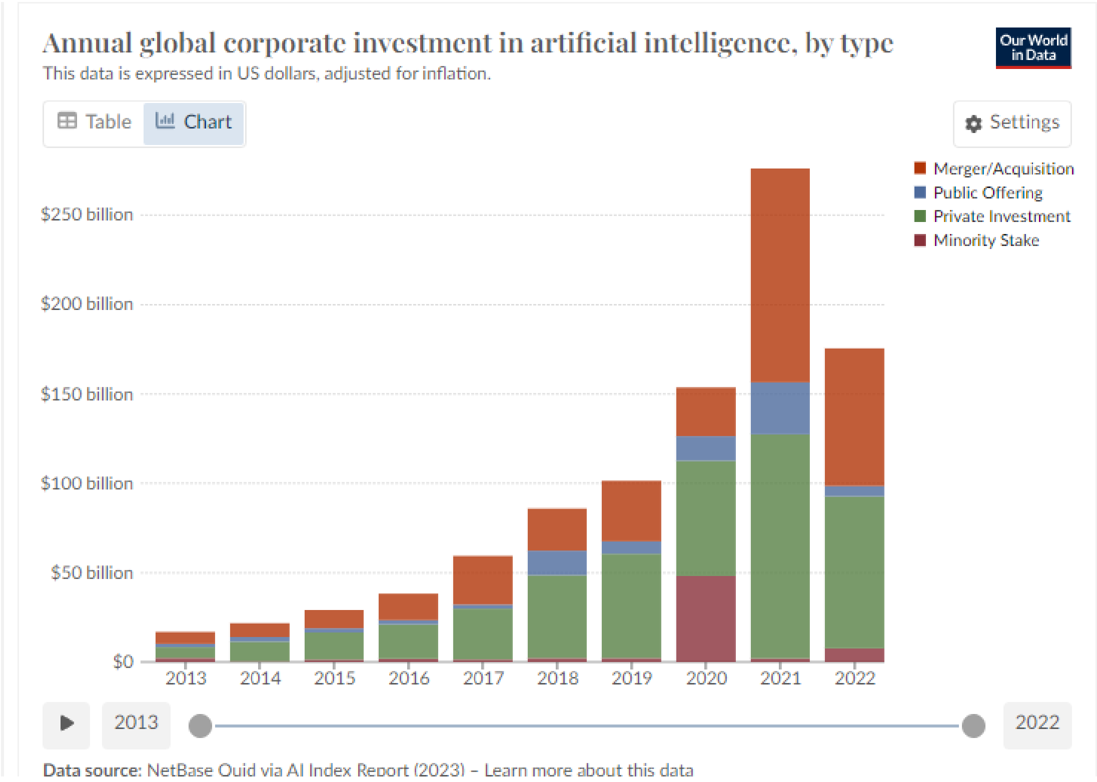
Two key considerations emerged: First, we recognize that those who bear ultimate responsibility for the success of the proposed center over its initial 10-year duration are the research faculty themselves. Since there is no responsibility without authority, decisions pertaining to the investment and operation of the center should ultimately be directed by the research faculty and the EVPR.

Second, it was also recognized that technical capabilities for design, construction, and operation of both the proposed accelerator (recommendation 1) and some of the educational thrust (recommendation 2) would be in the traditional purview of a unit such as CUIT, and therefore under the direction of the CIO.

Since we believe the proposed center will be essential to future research activities, and since the success of all research activities will ultimately determine the viability of the center, we recommend that research faculty must maintain full control (not just provide "input") as to all development and governance aspects of the proposed center, with CUIT or an equivalent unit bearing responsibility for its ongoing implementation and operation. In other words, the persons operating the center would report to a research faculty director(s).

Conclusion: A bifurcation of futures

The next decades will see a bifurcation of growth in the academic sector, much like we are seeing a bifurcation of growth in the commercial sector. Examination of the evolution of the commercial landscape paints a picture of diverging futures: A substantial part of the entire economic growth in the past few years has been dominated by companies that have embraced high performance computing, machine learning and Generative AI. Those companies that fail to make similar investments may be at a competitive disadvantage with potentially adverse outcomes to their existing business models.


Annual global corporate investment in artificial intelligence, by type

This data is expressed in US dollars, adjusted for inflation.

Data source: NetBase Quid via AI Index Report (2023) – Learn more about this data

Fig 5. Annual corporate investments in AI are growing exponentially (Netbase)

Similarly, we believe that universities that embrace new forms of augmented and accelerated discovery, creativity, innovation, and scholarship will thrive and lead, and those that do not will be relegated to a second tier. We want to ensure that Columbia University belongs to the leading tier, and that its students are equipped to become leaders of similarly advancing institutions in industry, government and academia. Our recommendations for the Discovery Accelerator revitalize the relationships between faculty, infrastructure, CUIT, and the Libraries, catalyzing synergistic benefits for research and the whole University enterprise. It is important to emphasize that the field of AI has historically tended to evolve in a winner-takes-all fashion. Therefore, delaying our embarkation on this journey by a few years “to see what other universities are doing” may be a grave mistake. We must act quickly and decisively.

References

1. The Institute launches a college. *MIT Technology Review* https://www.technologyreview.com/2018/12/19/138376/the-institute-launches-a-college/.
2. Yabut, R. USC president launches $1B initiative for computing including AI, advanced computation and quantum computing. *USC Today* https://pressroom.usc.edu/university-of-southern-california-launches-1b-plus-initiative-for-computing-including-ai-advanced-computation-and-quantum-computing/ (2023).
3. Aug 3, H. staff report / P. & 2023. Johns Hopkins makes major investment in the power, promise of data science and artificial intelligence. *The Hub* https://hub.jhu.edu/2023/08/03/johns-hopkins-data-science-artificial-intelligence-institute/ (2023).
4. Princeton invests in new 300-GPU cluster for academic AI research. *AI at Princeton* https://ai.princeton.edu/news/2024/princeton-invests-new-300-gpu-cluster-academic-ai-research.
5. AI at Princeton: Pushing limits, accelerating discovery and serving humanity. *Office of the Dean for Research* https://research.princeton.edu/news/ai-princeton-pushing-limits-accelerating-discovery-and-serving-humanity.
6. Harvard’s Kempner Institute Expands Academic Computing Cluster, Adds Nearly 400 GPUs | News | The Harvard Crimson. https://www.thecrimson.com/article/2023/11/2/kempner-institute-new-gpus/ (2023).
7. NVIDIA Celebrates Groundbreaking of New $213M Research Complex at Oregon State University. *HPCwire* https://www.hpcwire.com/off-the-wire/nvidia-celebrates-groundbreaking-of-new-213m-research-complex-at-oregon-state-university/.
8. A New Social Contract: Sharing our Strategic Priorities | Office of the President.

https://president.columbia.edu/news/new-social-contract-sharing-our-strategic-priorities.

9. Local Law 97 - Sustainable Buildings. https://www.nyc.gov/site/sustainablebuildings/ll97/local-law-97.page.
10. Local Law 97 | NYC Accelerator. https://accelerator.nyc/ll97.
11. Liu, J. & Jagadish, H. V. Institutional Efforts to Help Academic Researchers Implement Generative AI in Research. *Harv. Data Sci. Rev.* (2024) doi:10.1162/99608f92.2c8e7e81.
12. Casado, G. A., Matt Bornstein, Martin. Navigating the High Cost of AI Compute. *Andreessen Horowitz* https://a16z.com/navigating-the-high-cost-of-ai-compute/ (2023).
13. Al-Antari, M. A. Artificial Intelligence for Medical Diagnostics—Existing and Future AI Technology! *Diagnostics* **13**, 688 (2023).
14. Artificial intelligence has long been improving diagnoses. *The Economist*.
15. Kumar, Y., Koul, A., Singla, R. & Ijaz, M. F. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *J. Ambient Intell. Humaniz. Comput.* **14**, 8459–8486 (2023).
16. Simon, F. M. Artificial Intelligence in the News: How AI Retools, Rationalizes, and Reshapes Journalism and the Public Arena. *Columbia Journalism Review* https://www.cjr.org/tow_center_reports/artificial-intelligence-in-the-news.php/.
17. Östling, A. *et al.* The Cambridge Law Corpus: A Corpus for Legal AI Research. *Adv. Neural Inf. Process. Syst.* **36**, 41355–41385 (2023).
18. School, S. L. The Use of Artificial Intelligence in International Human Rights Law. *Stanford Law School* https://law.stanford.edu/publications/the-use-of-artificial-intelligence-in-international-human-rights-law/ (2023).
19. Werner, J. Three Big Use Cases For AI in Law, and Infinite Context Work. *Forbes* https://www.forbes.com/sites/johnwerner/2024/05/20/three-big-use-cases-for-ai-in-law-and-infinite-context-work/.

20. How AI will revolutionize the practice of law. *Brookings* https://www.brookings.edu/articles/how-ai-will-revolutionize-the-practice-of-law/.

21. Law Bots: How AI Is Reshaping the Legal Profession. *Business Law Today from ABA* https://businesslawtoday.org/2022/02/how-ai-is-reshaping-legal-profession/ (2022).

22. The Time is Now to Incorporate AI into Your Legal Services Functions. https://www.americanbar.org/groups/law_practice/resources/law-technology-today/2024/the-time-is-now-to-incorporate-ai-into-your-legal-services-functions/.

23. Machover, T. Media Lab Perspectives: Why Do We Want Our Computers to Improvise? with George Lewis. *MIT Media Lab* https://www.media.mit.edu/events/ml-perspectives-why-do-we-want-computers-to-improvise-with-george-lewis/.

24. ART-IFICIAL INTELLIGENCE. *Columbia Science Review* http://www.thecolumbiasciencereview.com/1/post/2022/11/art-and-ai.html.

25. How AI Is Changing Artistic Creation. https://zuckermaninstitute.columbia.edu/how-ai-changing-artistic-creation, https://zuckermaninstitute.columbia.edu/how-ai-changing-artistic-creation (2022).

26. Henkin, D. Orchestrating The Future—AI In The Music Industry. *Forbes* https://www.forbes.com/sites/davidhenkin/2023/12/05/orchestrating-the-future-ai-in-the-music-industry/.

27. Microsoft AI discovers 18 new battery materials in two weeks. https://www.freethink.com/energy/battery-materials.

28. Lv, C. *et al.* Machine Learning: An Advanced Platform for Materials Development and State Prediction in Lithium-Ion Batteries. *Adv. Mater. Deerfield Beach Fla* **34**, e2101474 (2022).

29. Materials-predicting AI from DeepMind could revolutionize electronics, batteries, and solar cells. https://www.science.org/content/article/materials-predicting-ai-deepmind-could-revolutionize-electronics-batteries-and-solar.

30. Allen, D. L., Abrahamsson, S., Murphy, M. F. & Roberts, D. J. Human platelet antigen 1a epitopes are dependent on the cation-regulated conformation of integrin alpha(IIb)beta(3) (GPIIb/IIIa). *J. Immunol. Methods* **375**, 166–175 (2012).

31. McBride, J. M. *et al.* AlphaFold2 Can Predict Single-Mutation Effects. *Phys. Rev. Lett.* **131**, 218401 (2023).

32. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

33. Toews, R. AlphaFold Is The Most Important Achievement In AI—Ever. *Forbes* https://www.forbes.com/sites/robtoews/2021/10/03/alphafold-is-the-most-important-achievement-in-ai-ever/.

34. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 1–3 (2024) doi:10.1038/s41586-024-07487-w.

35. Lewis, T. AI can read your emotions. Should it? *The Observer* (2019).

36. Cuadra, A. *et al.* The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction. in *Proceedings of the CHI Conference on Human Factors in Computing Systems* 1–18 (Association for Computing Machinery, New York, NY, USA, 2024). doi:10.1145/3613904.3642336.

37. AI's Emotional Revolution is Here. *CMSWire.com* https://www.cmswire.com/digital-experience/ais-next-big-step-detecting-human-emotion-and-expression/.

38. Monteith, S., Glenn, T., Geddes, J., Whybrow, P. C. & Bauer, M. Commercial Use of Emotion Artificial Intelligence (AI): Implications for Psychiatry. *Curr. Psychiatry Rep.* **24**, 203–211 (2022).

39. Khare, S. K., Blanes-Vidal, V., Nadimi, E. S. & Acharya, U. R. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Inf. Fusion* **102**, 102019 (2024).

40. admin. Emotion Detection and Recognition (EDR) Market by Services, Application Areas, Technology, Software Tools, End Users and by Geography - Global Industry Trends and Forecast 2021 - 2026. *Marketresearch* https://www.marketresearchengine.com/reportdetails/emotion-detection-and-recognition-edr-market (2017).
41. {Epoch AI. Data on Notable AI Models. *Epoch AI* https://epochai.org/data/epochdb/visualization (2024).
42. Computation used to train notable artificial intelligence systems. *Our World in Data* https://ourworldindata.org/grapher/artificial-intelligence-training-computation.
43. Computation used to train notable AI systems, by affiliation of researchers. *Our World in Data* https://ourworldindata.org/grapher/artificial-intelligence-training-computation-by-researcher-affiliation.
44. (3) The Great GPU Shortage and the GPU Rich/Poor | LinkedIn. https://www.linkedin.com/pulse/great-gpu-shortage-richpoor-chris-zeoli-5cs5c/.
45. Experiencing Decreased Performance with ChatGPT-4 - ChatGPT. *OpenAI Developer Forum* https://community.openai.com/t/experiencing-decreased-performance-with-chatgpt-4/234269?page=5 (2023).
46. AI Developers Stymied by Server Shortage at AWS, Microsoft, Google. *The Information* https://www.theinformation.com/articles/ai-developers-stymied-by-server-shortage-at-aws-microsoft-google (2023).
47. Kolodny, L. Elon Musk ordered Nvidia to ship thousands of AI chips reserved for Tesla to X and xAI. *CNBC* https://www.cnbc.com/2024/06/04/elon-musk-told-nvidia-to-ship-ai-chips-reserved-for-tesla-to-x-xai.html (2024).
48. Levy, A. Tesla, Meta, Microsoft, and Alphabet All Just Shared Magnificent News for Nvidia Investors. *The Motley Fool* https://www.fool.com/investing/2024/04/30/tesla-meta-microsoft-

and-alphabet-all-just-shared/ (2024).

49. Top-ranked analyst declares JPMorgan ‘the Nvidia of banking’ after it spends $17 billion on tech in a single year. *Yahoo Finance* https://finance.yahoo.com/news/top-ranked-analyst-declares-jpmorgan-211359484.html (2024).

50. Bousquette, I. & Rosenbush, S. Corporate AI Investment Is Surging, to Nvidia’s Benefit. *Wall Street Journal* (2024).

51. Data-Centric AI: AI Models Are Only as Good as Their Data Pipeline. https://hai.stanford.edu/news/data-centric-ai-ai-models-are-only-good-their-data-pipeline (2022).

52. Madan, S. *et al.* When and how CNNs generalize to out-of-distribution category-viewpoint combinations. Preprint at https://doi.org/10.48550/arXiv.2007.08032 (2021).

53. Team, S. Out of distribution blindness: why to fix it and how energy can help. *Snorkel AI* https://snorkel.ai/edited-transcript-detecting-data-distributional-shift-challenges-and-opportunities/ (2023).

54. Parra-Moyano, J., Schmedders, K. & Pentland, A. “Sandy”. How Data Collaboration Platforms Can Help Companies Build Better AI. *Harvard Business Review* (2024).

55. Rajeeva, A. AI datasets need to get smaller—and better. *InfoWorld* https://www.infoworld.com/article/3712223/ai-datasets-need-to-get-smallerand-better.html (2024).

56. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

57. FAIR Principles. *GO FAIR* https://www.go-fair.org/fair-principles/.

58. Data Management & Sharing Policy Overview | Data Sharing. https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policies/data-management-and-sharing-policy-overview#after.

59. Research and Development Expenditures at Colleges and Universities | American Academy of Arts and Sciences. https://www.amacad.org/humanities-indicators/funding-and-research/research-and-development-expenditures-colleges-and.

60. Deloitte. *Research Computing Options Analysis - Benchmark Study. Commissioned by Johns Hopkins University*. (2024).

61. Cavalotti, K. Facts - Sherlock. https://www.sherlock.stanford.edu/docs/tech/facts/.

62. Compute Savings Plans – Amazon Web Services. *Amazon Web Services, Inc.* https://aws.amazon.com/savingsplans/compute-pricing/.

63. GPU Cloud - VMs for Deep Learning | Lambda. https://lambdalabs.com/service/gpu-cloud.

64. Data center construction costs by location 2023. *Statista* https://www.statista.com/statistics/1106621/global-data-center-markets-ranked-by-cost-of-construction/.

65. shanepeckham. Understanding data curation and management for AI projects. https://learn.microsoft.com/en-us/ai/playbook/capabilities/data-curation/ (2024).

66. UF Announces $70 Million Artificial Intelligence Partnership with NVIDIA. https://news.ufl.edu/2020/07/nvidia-partnership/.

67. Waddell, K. MIT is investing $1 billion into AI research. *Axios* https://www.axios.com/2018/10/18/mit-ai-research-1-billion (2018).

68. Ross, D. Transformational Investment in Data Science and AI. *JHU Engineering Magazine* https://engineering.jhu.edu/magazine/2023/12/transformational-investment-in-data-science-and-ai/ (2023).

69. Fisher, P. & Peterka, D. MIT, Discussion with the Office of Research Computing and Data.

70. Szalay, A. & Peterka, D. JHU, Discussion with Alex Szalay. (2024).

44. Liu, J., & Jagadish, H. V. (2024). Institutional Efforts to Help Academic Researchers Implement Generative AI in Research. Harvard Data Science Review.

45. Sastry G, Heim L, Belfield H, Anderljung M, Brundage M, Hazell J, O'Keefe C, Hadfield GK, Ngo R, Pilz K, Gor G. Computing Power and the Governance of Artificial Intelligence. arXiv preprint arXiv:2402.08797. 2024 Feb 13.

Appendix

The following pages include accessory information, and summaries reflecting the needs of individual schools and units at Columbia. These summaries have been authored independently by representatives of those schools.

Acknowledgements

The committee would like to thank the many faculty, researchers, administrators, and staff at Columbia University that participated in the surveys, and subsequent discussions about current and emerging research compute needs. Further, we thank the many individuals and institutions that offered candid and generous discussions and feedback, particularly Peter Fisher (MIT), Alex Szalay (JHU), Curt Hilligas (Princeton), Columbia SRCPAC, and Jeffrey Lancaster (NVIDIA).

Committee members

Chairs		**Members**	
Darcy Petarka *ZMBBI*	X	George Hripcsak *Biomedical Informatics*	X
Hod Lipson *Mechanical Engineering, SEAS*	X	Julien Teitler *SSW*	X
		Laura Kurgan *GSAPP*	X
Ex-officio		Robert Pincus *LDEO*	
Jeannette Wing *EVPR*	X	Roisin Commane *Earth & Environmental Sciences*	X
Robert Cartolano *Libraries*	X	Seth Cluett *Music*	X
Alexander Urban *Chemical Engineering*	X	Siddhartha Dalal *SPS*	X
Frantz Merine *Law School*		Timothy Berkelbach *Chemistry*	X
Gaspare LoDuca *CUIT*	X	Wojciech Kopczuk *Economics*	X
		Ciamac Moallemi *Business School*	X
		Siddhartha Dalal SPS and Dept of Statistics-CAS	
Staff			
Maneesha Aggarwal *CUIT*		Victoria Hamilton *EVPR*	X

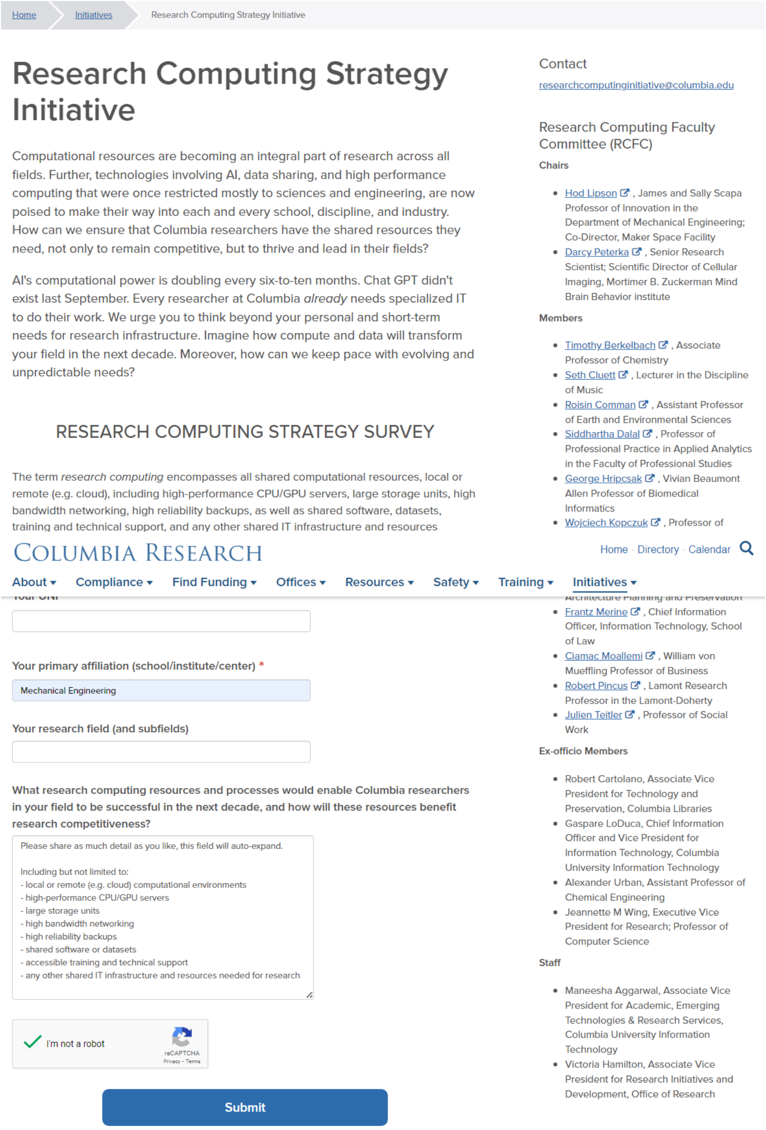
Jessica Eaton *CUIT*	X	[redacted]	

Peer interest and investment in HPC/Compute and AI

Columbia is not alone in recognizing the potential impacts of HPC and AI on universities[2,3,5,6,66–68]. There have been large investments by other institutions across the entire domain of the research and academic enterprise, which presents some challenges disambiguating the specific details of their investments, which range from specific hardware purchases, faculty hires, department and center creation, and building construction. We note that all of these, however, should be seen as recognition that HPC and AI will have profound impact on the generation and interpretation of knowledge, and requires significant attention, strategy, and investment for the modern research university. Recently, Johns Hopkins commissioned Deloitte to perform a benchmark study of JHU and public and private peers with regards to research computing and AI. Their report identifies many of the same issues and opportunities that our panel deliberated to make our recommendations. These include the need for large-scale investment in compute infrastructure and personnel, at a size that was unprecedented before AI and big data, and one that requires continual investments to evolve with changing needs. Other takeaways included the need for centrally coordinated, but faculty-guided investments, personnel, and strategy. This allowed for efficient sharing and joint central and faculty investments in the compute infrastructure for mutual gain. The larger compute needs also drive much higher power and cooling requirements, shifting compute and datacenters away from the main campuses. Further, while cloud resources were considered, most sites reported a lifetime expense per compute unit as significantly higher for cloud-based vs university-owned resources. Beyond the above listed report, members of the committee have also had broad discussions with both academic peer institutions[69,70] and industry about proposed or ongoing investments in HPC/AI/Storage hardware by Universities that are of similar scale to what is proposed in this report.

Survey

A survey was issued broadly to all faculty and postdoctoral researchers soliciting information about future research computing needs. The survey was also available on the open web page where any Columbia-affiliated person could provide input or reach out to committee members: https://research.columbia.edu/research-computing-strategy#!/%23cu_webform-21235



Home > Initiatives > Research Computing Strategy Initiative

Research Computing Strategy Initiative

Computational resources are becoming an integral part of research across all fields. Further, technologies involving AI, data sharing, and high performance computing that were once restricted mostly to sciences and engineering, are now poised to make their way into each and every school, discipline, and industry. How can we ensure that Columbia researchers have the shared resources they need, not only to remain competitive, but to thrive and lead in their fields?

AI's computational power is doubling every six-to-ten months. Chat GPT didn't exist last September. Every researcher at Columbia *already* needs specialized IT to do their work. We urge you to think beyond your personal and short-term needs for research infrastructure. Imagine how compute and data will transform your field in the next decade. Moreover, how can we keep pace with evolving and unpredictable needs?

RESEARCH COMPUTING STRATEGY SURVEY

The term *research computing* encompasses all shared computational resources, local or remote (e.g. cloud), including high-performance CPU/GPU servers, large storage units, high bandwidth networking, high reliability backups, as well as shared software, datasets, training and technical support, and any other shared IT infrastructure and resources

COLUMBIA RESEARCH
Home · Directory · Calendar
About ▾ Compliance ▾ Find Funding ▾ Offices ▾ Resources ▾ Safety ▾ Training ▾ Initiatives ▾

Your UNI

Your primary affiliation (school/institute/center) *

Mechanical Engineering

Your research field (and subfields)

What research computing resources and processes would enable Columbia researchers in your field to be successful in the next decade, and how will these resources benefit research competitiveness?

Please share as much detail as you like, this field will auto-expand.

Including but not limited to:
- local or remote (e.g. cloud) computational environments
- high-performance CPU/GPU servers
- large storage units
- high bandwidth networking
- high reliability backups
- shared software or datasets
- accessible training and technical support
- any other shared IT infrastructure and resources needed for research

✓ I'm not a robot

Contact

researchcomputinginitiative@columbia.edu

Research Computing Faculty Committee (RCFC)

Chairs

- Hod Lipson, James and Sally Scapa Professor of Innovation in the Department of Mechanical Engineering; Co-Director, Maker Space Facility
- Darcy Peterka, Senior Research Scientist; Scientific Director of Cellular Imaging, Mortimer B. Zuckerman Mind Brain Behavior institute

Members

- Timothy Berkelbach, Associate Professor of Chemistry
- Seth Cluett, Lecturer in the Discipline of Music
- Roisin Comman, Assistant Professor of Earth and Environmental Sciences
- Siddhartha Dalal, Professor of Professional Practice in Applied Analytics in the Faculty of Professional Studies
- George Hripcsak, Vivian Beaumont Allen Professor of Biomedical Informatics
- Wojciech Kopczuk, Professor of

Architecture Planning and Preservation

- Frantz Merine, Chief Information Officer, Information Technology, School of Law
- Ciamac Moallemi, William von Mueffling Professor of Business
- Robert Pincus, Lamont Research Professor in the Lamont-Doherty
- Julien Teitler, Professor of Social Work

Ex-officio Members

- Robert Cartolano, Associate Vice President for Technology and Preservation, Columbia Libraries
- Gaspare LoDuca, Chief Information Officer and Vice President for Information Technology, Columbia University Information Technology
- Alexander Urban, Assistant Professor of Chemical Engineering
- Jeannette M Wing, Executive Vice President for Research; Professor of Computer Science

Staff

- Maneesha Aggarwal, Associate Vice President for Academic, Emerging Technologies & Research Services, Columbia University Information Technology
- Victoria Hamilton, Associate Vice President for Research Initiatives and Development, Office of Research

School of Engineering and Applied Sciences (SEAS)

Hod Lipson

The School of Engineering and Applied Science (SEAS) is home to many departments whose research activity is computationally intensive by nature, ranging from physics-dominated academic disciplines such as mechanics, electrical, chemistry and materials, to medical and human-centered areas such as biomedical engineering, and more computationally focused areas such as computer science, applied math and industrial & operations research. Most practical application areas, such as Robotics, span multiple academic disciplines.

Research in SEAS has traditionally consumed research computing in three primary ways: (a) High performance simulations, (b) Data-driven modeling and machine-learning processes, and (c) Design automation, search and optimization. These three application areas are often intertwined, and are used for modeling, insight, forecasting, control, creativity and decision-making applications.

In particular, while automation in design and optimization have long been a consumer of research computing, the recent advent and success of Generative AI tools, has highlighted two key notions: One, that generative design tools will rapidly accelerate scientific discovery and engineering design and are therefore key to any science and engineering entity tasked with discovery and innovation, and second that the tools for generative design require very large amounts of computing power and data, and therefore access to computing power and datasets could become a bottleneck to innovation. Certainly, this challenge is evident in the frenzy of leading industries attempting to build generative models and acquire computational resources to run these generative models and own the datasets to train them.

In addition to computational power, it has become clear that unique datasets have become essential assets. As creativity is increasingly commoditized through generative AI, it is the unique datasets that are used to train and fine-tune these models that hold the key to unique innovation. As academia has traditionally struggled to collect, manage, and secure large datasets, this gap is becoming a problem and making academic research increasingly dependent on external resources.

While faculty, postdocs, staff and graduate students in SEAS have been involved with computational research for a long time, the growing importance (if not dominance) of computational aspects combined with limited resources is straining innovation. Many if not most projects begin as small exploratory endeavors, and if these initial explorations do not have ready access to germinate, the innovation pipeline will dry up. Similarly, as projects begin to grow, if they are stifled by lack of access to computational resources and data storage, their growth too will be throttled. Finally, as researchers turn to industry to fund or fulfill these growing needs, they open themselves to become subservient to short-term industry goals, constraints, and priorities.

Thus, in an attempt to maintain the competitiveness and independence of SEAS research, access to research computing is becoming an essential utility as important and basic as power, space, and communication.

Key observations

1. **The data is the asset**. As AI algorithms are open source, computing power is being commoditized, and programming talent is ubiquitous, the only remaining asset is DATA. For example, you can get almost everything for free from Google -- except their data. Data is also the bottleneck to creating unique generative AI capabilities. Therefore, universities must enable researchers to collect, curate, and manage data assets.
2. **Generative AI is key to engineering creativity**. Generative algorithms, applied in various engineering domains, such as genomics, robotics, material design, antennas, etc, are the new way to generate new ideas and discoveries. As a school whose entire reputation is based in the ability to innovate, access to generative AI tools and the ability to modify existing AI tools and develop new ones are key to remaining competitive and even relevant.
3. **Sharing resources is essential**. No single researcher can acquire, operate, and maintain the amount of computing needed in the future. Resources must be shared.
4. **Small projects need big resources too**. While it is possible to get government funding and computational resources for mature projects, academic innovation begins with many small projects that have no funding, but these small projects still need large computing power to germinate. Without access to fertile grounds, the innovation pipeline will dry.

5. **Computer power is a utility**. Basic research computing is increasingly assumed to be available much like electricity, communication, and real estate space. It has to be available without planning.
6. **Cloud does not match the nature of academic research**. Cloud computing is not a solution to baseline computing needs. It is very expensive, has limited availability especially at peak times, and is risky in terms of runaway costs and planning. It makes academia subservient to industry priorities.
7. **Cost structure of computing matters.** Computer equipment is free of overhead and lasts indefinitely (like a utility). In contrast, cloud computing is a service that incurs overhead and is only available during the performance period of a grant. This cost structure is therefore part of the calculation made by many faculty.
8. **Training and support desired but difficult to scale.** Training and support staff are needed for training students and new users, but dedicated personnel are expensive, difficult to keep updated with changing technologies, and difficult to scale to many people. Also a substantial amount of training material and resources (eg. videos) are already available online. Therefore, investing in training personnel may not be cost effective (compared to using the same funds for cycles).
9. **Need for a local solution**. While it is tempting to put datacenters remotely, placing them "in someone else's back yard" may not be ethical. If we consume computing power we need to learn to deal with it locally.
10. **Solutions must be sustainable financially, procedurally, and environmentally**: Whatever resources are chosen, they must be renewable, i.e. upgradable every five years or less. We need a standing committee to manage the renewal process and an endowment to support it. They must also be sustainable from an environmental point of view.

Graduate School of Architecture, Planning and Preservation (GSAPP)

Laura Kurgan

INTRODUCTION

GSAPP uses computation in studio settings and in research environments. Each student in design studios has a computer at their desk loaded with cloud-based software with a range of computational design tools as well the more familiar suites of software. GSAPP has a number of Centers and Labs which engage in research, which often do require high performance computing. However, most of our faculty do not know what they have access to, paid or unpaid, through the larger university network. I would like to use this opportunity to make sure our faculty know what is available to them, and also, how to make educated choices based on what is available.

The survey was sent out in the summer and was ignored by our faculty. However, computation is used at quite a high level on a daily basis by faculty and students across the school. The reasons for non-response were likely threefold: a.no-one was around, b.the questions were not geared towards a creative/design approach to computation which our faculty need and c. many of our faculty are practitioners with offices outside the University. For this last group the use of computation is likely decided by IT people in their offices, and it is important for this committee to capture the needs of these faculty to understand research and training needs within the school in terms of emerging technologies and rapid change in our field outside of the university.

Because of the low response rate, this document summarizes conversations with 5 GSAPP faculty across 4 out of 8 departments at our school including, Architecture, Urban Planning and Design, Historic Preservation, Computational Design Practices. We still need input from our Real Estate Program. I will try to add more to this document once I have contacted a few more faculty.

As prompted by Hod Lipson, I asked our faculty to think in a Bluesky way.
I spoke to David Benjamin, Anthony Vanky, Jorge Otero Paillos and Leah Meisterlin.

BLUESKY SUGGESTIONS:

1. **Can GSAPP work be used to fine tune Generative AI Models?** GSAPP students spend many hours producing speculative design – building new worlds as well as new ideas for technology in building. The work is archived every year for an online website as well as for the end of year show. Would it be possible to fine tune DALL-E or Midjourney for example (taking into account that copyright issues are fixed first) to generate an ongoing GSAPP ‘language model’ with existing student work. How might students learn from this to move our field forward.

2. **From an Urban Planning/Computational Design Perspective:** Students produce thesis and capstone projects. Could produce a searchable database of past research projects with common themes so that students don't start from ground zero each year and build on knowledge that is created by our own community.
3. **Could an API be leveraged to create a specialized chatbot tailored to the field of historic preservation**. This chatbot would serve the purpose of aiding with municipal regulations and/or specimen identification
4. **Can we host a GSAPP Datalake?** This way, we could capture data that Faculty and Students have produced with sensors they have designed, or data that they have analyzed and collected so that work can be built upon each year. Some of this work is done by Research Data Services in the Library, and we have hosted some amazing datasets that are now looked after and curated by the Library(Historical New York City Project) but is there more we can do based on faculty and student work.
5. **What about Maker Spaces in Multiple Schools across the Morningside Campus?** Many of us use the same equipment across schools. Is there a way to form a network out of these spaces and have support staff that relates to the whole network. That way some departments that receive less grants might benefit from others that receive more?? I'm responding here to Seth's equity question.
6. **Virtual Reality.** The same can be said for VR infrastructure. We have many uses for VR and always a shortage of VR equipment.
7. **Bootcamps.** Can we host a series of interdisciplinary AI bootcamps – especially geared to the formation of customizing AI models for specific fields.
8. **Columbia Large Language Model.** Is there a benefit in creating our own large language model to work outside of the commercial models? I know that more data means better models, but is there something to learn from a non-commercial model – a transparent, open source, foundational model - that might have an impact as well as solve some of the bias problems endemic to the current models.
9. **Centralized GIS Data.** Although Research Data Services goes a long way in solving this, I have been on numerous committees over my 20 years at Columbia asking for Centralized GIS resources. We still have multiple centers and institutes who would like to have access to similar data and no way to find it.
10. **Storage and user access to Lidar Models:** HP Students and others at GSAPP do a lot of work with LiDAR scanning. So far, they have not been able to come up with a way to provide storage and access in the cloud in a way that could make their models more useful to other students, and more importantly to the general campus.

BRIDGE REQUEST BETWEEN BLUESKY AND PRAGMATIC

Some faculty expressed a need that seemed easy to solve if they knew the right people to talk to. When our faculty start a project with Spatial Data, for example, they can go to Research Data Services in the Libraries to talk through their methodology and ask the research librarians (Jeremaiah Trinidad and Eric Glass are particularly helpful) to help them with methodology or with access to other data. But when they are working with large datasets, they often get stuck. A common question for these researchers is: should they work in a distributed cloud for data processing or purchase a large tower to store under their desk? The latter is the most common model at GSAPP and we all know that is not the right answer. Faculty want to know who to talk to for this kind of engineering advice to come up with the correct and least expensive version of what they might need to carry out a project. Again, Research Data Services is great for Faculty and the Empirical Reasoning Center at Barnard is a great resource for Barnard Students on software and methods. It would be great to start and ERC on the Morningside Campus on engineering/hardware questions related to their research. Is CUIT the place for these faculty to go? They are often too expensive for us, as is commercial cloud distributed computation. If not, we need a service like this.

PRAGMATIC REQUESTS:

From Historic Preservation – this might be too fine grain for this doc. Please delete if so – but I wanted to show you that most faculty request unique equipment for their own spaces rather than consider cloud computing.

1. What do you need a computing infrastructure that you don't have?

The HP technology lab needs an in-lab storage solution for digital scanning data sets (estimated to be about 1TB every two semesters, at peak, in order to maintain student scanning work in perpetuity). Transferring large data sets to the cloud would take several days, whereas using a hardwired SSD would save time and increase productivity within a semester.

Additionally, APC units are required to protect against power surges and ensure the computers remain operational, especially during critical data processing. Since photogrammetry projects can run for days or weeks, safeguarding against outages is crucial.

Storage

We have 4T storage and each computer has strong video cards. Depending on the size of the project, processing the data can happen between 1 hour to overnight.

Regarding online cloud storage, I think that is necessary.

Non-user specific / lab data storage

We produce many 3D documentation scans of buildings that take up a lot of data storage.

We need a general email/user to the lab, if we had such a thing, with a Google Drive that would be connected to a general email, then storage that could be continuous regardless of who is working a model at any time. So far there has been no resolution to that matter.

2. Does HP need high-performance computing?
Acquiring an additional tower with the following specifications for the Preservation Technology Lab dedicated to VRI:

- Minimum Requirements for Dedicated Graphics Card: Open GL 4.3, at least 8 GB memory, NVIDIA 1080GTX or equivalent.
- For Stereo Rendering: NVIDIA Quadro.
- Processing: 8 physical cores, such as Intel Core i7, Core i9, or Xeon processors.

Additional Accessories :
3D Connexion Space Mouse with the latest drivers for Laser Faro SCENE software.

Columbia Population Research Center

Julien Teitler

This document summarizes the survey responses of the Columbia Population Research Center (CPRC) affiliates. CPRC has 160 affiliates mostly in A&S, Public Health, and Social Work, but also in CIESIN, Nursing, Architecture, TC, Barnard, and the Medical Center (Pediatrics, Psychiatry, Neurology, Emergency Medicine). CPRC supports NIH research and has been funded by a P2C grant from the National Institute of Child Health and Human Development since 2006. It also receives supplemental resources from Columbia University.

CPRC regularly surveys affiliates about their computing and methodology needs. This document summarizes themes that have emerged from both the UCC survey and ongoing surveys and meetings with CPRC affiliates.

Computing and data access issues raised by CPRC faculty affiliates:

#1 Data sharing within CU and with outside collaborators is very difficult

This is due to a combination of factors: Incompatibility of platforms on the uptown and downtown campuses, the need for UNI authentication to access highly secure computing platforms (and difficulty obtaining UNIs for outside collaborators), the absence of easily accessible medium security platforms (e.g. that allow PII but not HIPAA data).

#2 High performance computing (high cost of entry of HPC)

For researchers outside units that buy into HPC, the cost of membership is very high. The costs of alternatives outside Columbia (e.g. AWS can also be high).

#3 Highly constrained resources are inflexible (particularly uptown)

Required to use Microsoft solutions when there are much better ones out there. This creates uptown-downtown collaboration hurdles (see #1).

#4 Procuring data

There is no resource that assists researchers in obtaining, managing, and distributing data. Most recently, faculty have been asking for help in obtaining L2 voting data and Medicaid claims data. These data require upfront cost to purchase access and ongoing support for managing and distributing the data.

Additionally, Columbia's resources for reviewing and negotiating DUAs and purchase agreements with data vendors seem to be stretched thin and the process takes an unreasonable amount of time, especially in the context of grant or contract funded work. We need additional staff training in DUA and purchasing processes for data products. A lot of the standard forms Columbia provides to potential data vendors are oriented to Human Subjects Research, health data, PHI and PII, but a lot of the data faculty need to license do not have human subjects components. This leads to many rounds of discussion between vendors and Columbia to amend forms and text in the Columbia agreements.

#5 No high-performance computing for sensitive data (the SDE is very limited)

The HPC is good for analysis of non PII data and the SDE is good for less computationally intensive analyses of PII and PHI data, but there is no platform for computationally intensive analysis of PII and PHI data.

#6 Slow speed of adopting collaborative research solutions that meet PII/PHI/HIPAA requirements. E.g. it took years to several years to decide on and roll out Box.

Between the selection of platforms, establishing contracts with vendors, testing, modifying Rascal and IRB platforms to accommodate new platforms, and roll-out, it can take years for CU to come up with solutions to serious research analysis bottlenecks.

#7 The cost structure of most computing platforms (e.g. HPC, SDE, Box) requires that individuals pay for access to secure computing or high-performance computing. This potentially locks out faculty who are doing unfunded research or are between grants. Accessibility could be greatly improved if a basic access tier would be provided free of cost to all faculty and students, with larger, grant funded projects, contributing for higher volume usage.

Recommendations:

1. Most of the research computing bottlenecks experienced by CPRC faculty are related to a lack of coordination across Columbia units, including between uptown and downtown IT departments; between the Provost's office, HR, and CUIT; the procurement office, the IRB, and between CUIT (or RCS) and faculty researchers. Our recommendation is that decisions about new computing platforms, access to platforms, and cost structures of platforms are preceded with consultations across the units listed above and an impact assessment on faculty research.

2. The pricing structure for some of the platforms, particularly for storage and analysis of sensitive data (e.g. Box and SDE), creates disincentives to properly securing data. Columbia should provide centralized funding for anything related to data security to minimize risk to human subjects, or at least a free access tier.

3. Columbia should find a way to support high performance computing with sensitive data, for example, blending the attributes of SDE and HPC.

Lamont campus perspective on research computing survey

Robert Pincus

Survey responses

This summary focuses on the 15-20 responses to the survey from the Lamont Campus and collaborators (e.g. researchers in DEES/A&S and APAM/SEAS who also maintain a presence at Lamont).

Similar to other academic units, respondents note the need for expanded access to computing resources, calling out “reduced barriers to HPC and GPU use” and specifically “HPC (CPU and GPU nodes)” Several mention cloud computing as a means of making large data sets available for analysis. Researchers crave actively-maintained but customizable software environments on computing and analysis platforms.

Barriers to internal and external collaboration came up frequently. With respect to internal collaborations, many of the respondents come from climate science in which it’s common to rely on large (petabyte) archives of data produced and archived remotely. Several respondents pointed out the need for technology and especially people to more effectively use such archives: “data[set] storage needs to be organized centrally so that we don't end up having multiple copies”, and “data[set] storage needs to be organized centrally so that we don't end up having multiple copies.”

With respect to external collaborations, respondents expressed needs for easier access to Columbia computing resources, but more frequently expressed the ability to share large amounts of data with specific collaborators and more broadly. “We need a platform to host large climate data that is also capable of interactive visualization that helps convey the story behind the data to the public”

Consistent with responses from other units, researchers expressed a desire and expectation to have an elevated level of baseline user-level services from the University including backup, file sharing, and technical training (the latter need expressed at all career stages). “We need a system [with] reliable IT support [for] local and remote environment, this cannot be the responsibility of each individual PIs.”

There was dissatisfaction with current funding models in which computing and support are drawn from grants: “For ongoing HPC support, to have a fixed charge that one puts on all grants would be much easier to accommodate than to have to come up with large chunks of money every 5 years. It’s a barrier to keeping up a research program in this area.”

Recommendations

1. Expand access to high-performance computing including GPUs as much as possible through a combination of additional resources, different funding models, and different approaches to provisioning resources
2. Provide diverse ways to store, distribute, and analyze petabyte-scale data stores. This should include e.g. the ability for external users to remotely-access data, perhaps with access restrictions, and flexible computing environments proximate to the data.
3. Develop capabilities to more efficiently share large data sets, especially internally. This should include some combination of human coordination and automated tools to reduce data redundancy.

Arts & Sciences - Sciences response

Róisín Commane and Timothy Berkelbach

This response focuses on the response from the Science faculty (Astronomy, Biological Sciences, Chemistry, DEES, E3B, and Physics) to the survey sent in Summer 2023, discussions with faculty once they had returned to campus in Fall 2023, a meeting between the A&S committee members and A&S admin and an additional survey sent to postdoctoral researchers in Fall 2023.

Summary:

Overall, there was an acknowledgement of a need to move to more energy efficient computing options but many did not see how that was possible with the current infrastructure. Many Faculty and researchers require HPC or cloud computing (both CPU and GPU) and requested better guidance on access and payment, more flexible purchasing options (e.g., lower barrier to entry and/or more frequent purchasing periods), software that is readily accessible to researchers, increased IT support, improved documentation of available resources and easy-to-use research storage and backup. Continued access to computing resources for a short time after people leave was also mentioned. The practice of cutting off IT access the day a researcher leaves seems to be very unusual across academia and especially other peer institutions (e.g. Harvard has a grace period).

Survey responses

1. Three different types of accessibility were mentioned as essential for competitive research and for the successful uptake of HPC or cloud computing and transition away from departmental or group-specific computing resources:

   a. Funding accessibility. The need for a coordinated computing support model that is more integrated with the typical 3 year funding cycle of many in A&S.

      i. *DEES: For ongoing HPC support, to have a fixed charge that one puts on all grants would be much easier to accommodate than to have to come up with large* *chunks of money every 5 years. It's a barrier to keeping up a research program in this area.*

   ii. *Ecology: Columbia really needs to offer HPC computing for less money than having to buy nodes. I realize there are free allotments, but there should be more free options for students and postdocs, as well as faculty. We are way behind our peers and even non-peers. For example, most of my postdocs work on clusters at their PhD institutions because they are always better than what we offer.*

b. Easy access to remote servers. *HPC servers with up-to-date software and interfaces such as jupyterhub, rather than being limited to SSH tunneling, would be a significant improvement over the current infrastructure, and likely encourage more non-specialists to use the facility.*

c. Continued access to complete projects. Other Universities (Harvard, etc) allow a grace period for graduate student researchers to finish computing projects after they have moved on to other positions. This has become a big issue for junior faculty who are encouraging their students to finish on time with papers not fully published. They soon discover that the student loses all computing access, with no grace period and finishing final paper review responses becomes really difficult. All access to CUIT facilities comes from UNI access and a change to the UNI grace period would make research much easier for faculty and students alike. *This is not a problem at Lamont where IT access does not rely on a UNI so projects are not impacted but it has made researchers reluctant to engage with CUIT facilities.

2. Two different types of education were mentioned as essential to encourage uptake.

a. Training for students and postdocs. Most other institutes have extensive online documentation provided for their clusters, which make it easy for a first-timer to use. In general there was concern about training and lack of easy to get students and postdocs from non-computing fields involved in using the clusters/cloud computing.

b. Current users asked for faster response time from Research Computing Services on software installation requests and problems encountered. Suggestions such as a wiki of FAQs or a slack channel where people could ask others for help with technical issues within the Columbia HPC environment, rather than having to wait a week for RCS to respond to a ticket. This is severely impacting research science.

Other comments and suggestions

Physics: *At University of Michigan, the IT office shared between multiple departments set up and managed my Linux workstation and responded to software installation requests and technical issues within a day. In comparison, at Columbia, my department only has a single person who provides IT support, and thus is only able to respond to requests for assistance after multiple days have passed. In addition, University of Michigan provided access to 5 TB of storage on Dropbox and backup of my workstation on Crashplan without direct charge to the researcher. This helped me make sure my data was preserved and saved me the trouble of having to find data storage solutions.*

Astronomy: *observational data streams expected to reach 10 petabytes per year for the Rubin Observatory and 100 petabytes per year for the Square Kilometer Array, so that even if a small fraction is brought local, enormous amounts of storage space will be needed for analysis.*

Arts & Sciences Humanities and the School of the Arts

Seth Cluett

Summary:

There is a significant need to ensure network and computing equity for fields and departments that lack the financial resources to stay current. There is a desire for expansion, clarification and onboarding for the free tier at CUIT to begin to approach the supported access to entry tier cloud computing provided by peer institutions that scales to paid plans with increased use. Cloud-based GPU access to support AV rendering for film, music, and visual art users and site-license access to JupyterHub or equivalent would be desirable to encourage adoption. Faculty requested transition training for software, AI end-user tools, and digital humanities, especially for absolute beginners but also for current graduate students to prepare for job market changes. Energy-use and a desire to ensure a sustainable, green solution is of utmost importance.

Details:

The response rate in the humanities and arts was notably low. Efforts continue in collaboration with administration to gather additional responses. Follow up conversations with colleagues, many felt the survey was not for them or weren't sure how to speculate their future needs in the face of Large Language Model AI discussions in particular. Almost everyone who didn't submit knew that they needed to think about this but expressed that the inevitable impact of recent and future technologies still felt too unknown.

Due in large part to the nature of media rich content creation, the Arts and Music require lab or personal machines capable of end-user AI, ML, graphic rendering, haptics, VR/AR/xR work, video, and audio processing. In conversation, users cited the successful use of cloud-based render farms by peer institutions to support animation, video/film compositing, post-production, and format transfer. The School of the Arts IT group has been exploring this route independently from research computing on campus with some initial success. The Computer Music Center does work in deep & machine learning, AI for creative applications, data sonification and visualization, digital signal processing, music information retrieval (think Spotify-style recommendation algorithms), all of this work is currently done on desktop machines in end-user labs but as enrollments grow and research collaborations expand, these facilities will not sustain research.

Many complained of poor network connections. An informal survey of the wired internet speed in Dodge and Schermerhorn Halls averages 7 Mbps down and 3 Mbps up and wifi ranges from 35Mbps to 120Mbps in rooms where a door interrupts line-of-site to a wifi node. Needless to say, these speeds are insufficient for all academics but are prohibitively slow for any of the kind of computing resources we are discussing. Since the response to this observation, CUIT has imitated a survey of internet speed and developed a plan for modernization. Many faculty cited the need for CUIT to scale the cost for the installation of ethernet or fast wifi with the bandwidth necessary for cloud storage and research computing. While this may seem small, with budgets in the $10k to $20k range, upgrading or installing internet in a facility can consume a year's budget.

For departments or researchers working in fields that lack major grant infrastructure (NIH, NSF, DARPA etc), the need for an effective centralized framework for keeping local or personal computing resources up to date was raised by a number of faculty. Some complained about the difficulty of taking advantage of FRAP funds to acquire computers with capabilities necessary for productive faculty research for which shared computing resources are inappropriate. FRAP is reasonable for consumer computing, however, you cannot add to FRAP or roll it over, so anyone requiring high CPU/GPU, memory, or storage needs have to find other methods to cover their personal research computing needs, often out of pocket.

A number of faculty cited the need for the university to commit to the acquisition or creation of digitized research archives or data sets. Alongside this however, there was a zero-sum-game concern that monies dedicated to high-end computing resources will in turn shift funds away from standard technologies for scholarly activity (subscriptions to archives etc). Many faculty cited the general need for increased local and cloud storage capacity and systematized backup for scholarly work and data.

Lastly, there was a universal desire for expanded training and support for the addition of data-based scholarship and digital humanities for faculty. Moreover, there is an immediate need for resources to train current grad students in fields where the job market will demand computational fluency in some form. While outside the scope of this committee, a computational-humanities cluster-hire was mentioned more than once.

Social Sciences (mainly economics and political science)

Wojciech Kopczuk, Economics A&S and SIPA

I summarize and contextualize responses from social science departments within Arts & Sciences and SIPA — these responses are primarily from economics and political science faculty, with some connections to other fields. There were also closely related comments by social scientists elsewhere (business school, public health, social work) that this summary reflects. It also reflects some in-person and less systematic conversations.

The first thing to note is that there are diverse needs that come up and that it reflects a variety of approaches and types of social science research. The typical social science model does not involve larger labs (although some of those exist and there's been a bit of a trend in that direction), it is usually a collaboration between a small number of researchers who are often not in the same institution (and sometimes solo), it might but does not have to rely on research assistants that would usually be graduate students and sometimes undergrads. Computational needs span the whole spectrum from almost none (some of theory work) to statistical analysis of terabytes of data (textual, geocoded, administrative, health, financial, scanner data, etc.) and/or computational models that require HPCs. Responses reflect that and I'll try to distill the most important themes

1. **Data.** The lack of infrastructure for storing data, bringing in sensitive/human subjects/confidential data, sharing data of any type (including those with restrictions), and integrating all of it with sufficient computational resources is one of the common themes. One of the comments states "Columbia really needs a university-wide established and transparent procedure for procuring, maintaining and providing access to sensitive big data," another one (from an economist in the School of Public Health, but reflecting broader needs I'm aware of) "I've had a lot of difficulty finding a place to store semi-sensitive data, e.g. deidentified claims data that doesn't meet the full HIPAA deidentification standard, like limited data sets. Ideally, there would be a secure computing cluster that was easy to access. I've used the secure data enclave, but it's actually *too* secure for my purposes - I need to be able to move data on and off the enclave."
2. **Computing power.** A theme that's been running through responses is the lack of cost-effective access to modest (by science standards) HPC resources. The option of purchasing a node is viewed as expensive (for social science budgets) and inflexible

(more than needed, set up costs, a longer commitment). Some of it may be unawareness of the free tier, but also the free tier was viewed as inadequate when the only step us is inflexible and expensive.

3. **Technical support and information about what is out there.** The common complaint is lack of convenient access to expertise of many different types — help/information about cloud computing; getting started with parallelizing and GPUs; about data storage; sharing and security options; about specific computational problems (one of the responses was about struggling to analyze large amounts of data using Stata - potentially the wrong tool here, but also, possibly, reflecting the lack of access to expertise about how they could take advantage of HPC)
4. **Sharing resources.** Infrastructure for sharing data with collaborators and students is one theme here; university-level data and software licenses is another (that includes things like Overleaf, Dropbox, ChatGPT, but also statistical software licenses). Some of it is really a call for a well-organized and centrally managed clearinghouse of resources that are already at the university

Zuckerman Institute

Darcy Peterka

The Zuckerman Institute was created to understand the complexities of the brain and mind, and is very broad and interdisciplinary by design. We currently have 54 Principle Investigators spanning 19 departments, including faculty from A&S, SEAS, and CUIMC. We are also home to the Center of Theoretical Neuroscience, and are a key partner with the newly founded SNF Center for Precision Psychiatry. Further, our labs range from purely experimental labs, to purely theoretical labs, and every mix in between, including labs that develop technical methods and others that create, refine, and develop algorithms and pipelines.

The Institute is housed in the Jerome L. Greene Science Center, a relatively new (open < 10 years) building, which appears to give us considerably better baseline network infrastructure, both wired and wireless, than other areas on campus. We also have an on-prem modern data center with a mix of high and medium energy density racks, and currently have ~4 PB of high performance (Isilon) storage coupled to a modest CPU and GPU clusters. The CPU cluster is fully configured for virtualization, and resources can be "rented" at reduced (subsidized) cost. Further, we have a dedicated IT/compute team that interfaces with faculty and staff and works closely with CUIT.

A large fraction of our investigators use at least one of these Institute centralized resources, but to varying degrees and some do not use any. Nearly every lab still has some local storage, with ~20% having their own medium-to-large NAS devices (10s of TB to ~1PB), and many others still relying on ad hoc local storage on desktop computers, or stacks of consumer hard drives. Similarly with compute – nearly every lab has at least one, some many, dedicated high-memory, high core workstations for processing and analysis. Commodity cloud services are also used for data sharing, such as Dropbox, or Google Drive, in addition to more formal tools such as Globus. As an Institute, and some individual faculty, have frequently participated in buy-in rounds for Columba HPC (SRCPAC).

The formal survey response rate was low, though many who did not respond to the survey shared thoughts in person, during subsequent canvassing. Many of these comments echoed points touched upon in some survey responses, but added additional color, and sometimes-new topics entirely.

- People underestimate the costs of storage and compute infrastructure, and thus fail to budget actual expected costs in proposals. Data lifecycle management is a serious issue. While there is a strong push within the neuroscience community for well-formatted and annotated data, there are real and perceived barriers to execute.

- Data that is not discoverable is “cheap data”. While it may have been expensive to acquire, this data has limited value, especially to the broader community, but is still expensive to store.

- Modern methods can generate 100s of GB to 100s of TB per day, per instrument, and many labs have multiple instruments. Rapidly storing, pre-processing, analyzing, exploring and interacting with this data is a problem, and limits better experiments (e.g. closed-loop experiments)

- Collaboration with large data-sets is hard. Network is too slow, and compute methods are often tied to a particular hardware instantiation.

- Real need for software engineers, to move algorithms from proof-of-principle to active deployment – not a priority/motivation for the researchers, even though there is clear understanding this has large value, for reproducibility, adoption, and overall impact.

- Difficulty of hiring, or recruiting researchers. Incoming people have increased expectations for resources.

- Big need for expertise in bioinformatics/big data/interpretation.

- Much of the data is currently resistant to hands-off analysis. Still many manual steps, and parameter tuning, and requires visual inspection. Makes clould work difficult.

- Lack of tools for exploration of very large imaging data.

- Limited access to data center space for individually owned hardware.

- Slow network for large data moves.

- Cloud seems ok for some things, but because specialized expertise is required, people become much more dependent on others for tool development, for ingest, processing, and interaction.

- Data rates are increasing very heavily - we need real-time processing, and potentially industry partnerships for managing streams from proprietary instruments.

- Better collaboration – sharing resources and compute.

- Better ways to understand use, and “touching” of data.

- Better responsiveness from local IT/RC staff. Some complaints about apparent expertise/engagement levels, or ability to solve lab specific compute issues quickly.

- Lack of modern GPU-hardware – complaints of outgoing, or current trainees that they are not getting trained on ML/AI at the speed that is necessary to stay “current”

Survey themes:

Cheap and easy (but good!) storage. Multi-TB to PB scale (raw) datasets are becoming much more common. We need a cost effective way to store these data sets, including backup, while keeping them very accessible to compute infrastructure.

Data and compute (local) need easy remote access, and high-speed transfers.

Access to GPU clusters with high availability (fast to spin up). Queues too long for big jobs on Ginsberg (CU HPC) Better containerization – want access to create an environment beyond home directory installs. Quality was considered generally good, but it is too little, and too rigid.

Higher level of baseline institutional resources for compute (GPU/CPU) and storage at no or low cost.

Access to expert staff across the range of possible functions – people to facilitate running large jobs, set-up/etc. Staff that can help advise on which resource to use, from local workstations, to local HPC to Cloud providers. People that can guide and support local workstation/compute (individually owned)

Better structures for sharing data between campuses (esp. uptown) and collaborators, including protected/HIPAA data

Better access to high performance clusters with managed environments, where all management, and optimizations are essentially abstracted away from the end users.

Needs to be cheap, and not go away during times of resource scarcity. Under desk = pay once, use forever.

Concern on who should administer/archive data long term. Most do not want to be responsible for that, and do not have the background or resources to plan effectively.

Research Computing Needs at Columbia Business School

Ciamac Moallemi

Background

Columbia Business School (CBS) operates its own research cluster and storage infrastructure independent of CUIT and has dedicated research computing administrators. In general, users are very happy with the setup. A key aspect of this is the fact our administrators provide very high touch support. Our users are not always sophisticated and dedicated/personalized support has been extremely important. Part of this is the ability to provide customized hardware and software environments. In the past we have had issues with university offerings for research computing such as SRCPAC, because they are restrictive and inflexible especially in their software administration. T**he central ingredient in the success of research computing at Columbia Business School has been the fact that those responsible for its design and administration are directly accountable to faculty.** This is done by having the research computing function report to the Senior Vice Dean for Research and the Faculty Computing Committee at the Business School. We strongly believe that future research computing efforts will only be successful if research computing administrators report to and are accountable to the school and faculty as opposed to CUIT.

Future Needs

I am summarizing future needs based on 9 responses received to the survey, interviews with users, as well as my own experience chairing the CBS Faculty Computing Committee for the past 5+ years:

- **GPUs.** One area of current shortfall is access to GPUs. Users would like to fine-tune and perform inference on state-of-the-art large language models as well as other generative AI models. This hardware to do so is very expensive (e.g., 8xNVIDIA H100-80GB server is $300K+) also in short supply thus very hard to purchase. Researchers have been constrained to using smaller models due to lack of access to high memory GPUs (especially currently access to NVIDIA A100/H100 GPUs is extremely challenging).

- **HIPPA.** Some of our researchers work with personal health information and need HIPPA compliant servers. CUIT offers a “Secure Data Enclave”, but this is so limited (max 25GB of collaborative workspace, only 1-2 users can access data at the same time, etc.) that is not useful. Our researchers have had success leveraging HIPPA compliant servers administered by CUIMC, but this requires collaboration with and sponsorship by CUIMC researchers. It would be great to have a modern university wide HIPPA compliant offering available outside of CUIMC.
- **Research engineering support.** As research computing gets more sophisticated, users need an increasing amount of engineering support. For now, our dedicated support model has worked well, but we may be beyond our abilities along more challenging technical dimensions, for example for deep learning models trained on GPUs.
- **Cloud computing.** We have very little expertise in cloud computing. This will become increasingly important, especially as it is an important avenue to obtain access to more exotic hardware like high memory GPUs. It would be helpful to leverage university-wide cloud computing efforts.
- **Physical facilities.** We rely on the university for rack space, power, and networking for our cluster, and hope to continue to do so.
- **Hardware.** We have continuing needs for CPUs and storage. In general, we have developed our own solutions independent of the university and these have worked well. It would be great to leverage university offerings, but the restrictive SRCPAC model does not work for us. If the university could offer bare metal access to hardware, or offer VM-level access (like cloud providers such as AWS or GCP), that might be something we could potentially use.

Faculty Input on Computing Needs - Law School

Frantz Merine

Feedback was gathered through discussions with research-active faculty members. The following insights are from discussions with faculty who recently made technology purchases through the IT Helpdesk using their research budgets/accounts.

- Autonomy - The ability to have full administrative control over the computing environment
- Division of Labor - Clear roles and responsibilities over the tech-stack
- Remote Access and Virtualization - Enabling access to resources fromanywhere
- Support and Training - Providing assistance and education in using computing resources effectively.
- High-Performance Computing (HPC) - Harnessing powerful computational capabilities for complex tasks.
- Dedicated Software and Hardware Resources - Access to specialized tools and equipment tailored to research needs.
- Data Storage and Management - Efficient organization and retention of research data.
- Data Analytics and Machine Learning - Use computational power for advanced data analysis and model training.
- Visualization and Graphics - Rendering and interpreting complex data through graphical representation.
- Security and Compliance - Ensuring data protection and adherence to regulatory standards

Faculty Input on Computing Needs - School of Professional Studies (SPS):

Siddhartha Dalal

Feedback was gathered through discussion with faculty and through the survey.

1. The survey responses were few since the survey was conducted during the summer.

2. **Diversity of Needs:** The faculty in SPS is very diverse with 18 departments including negotiations, Bioethics, Applied Analytics, Information & Knowledge Strategy, Insurance management, Enterprise Risk Management, etc. They all have very different needs including those who are satisfied with their laptops to those who require GPUs. The following observations are based on a discussion with several faculty members who are working on relevant topics requiring serious computing.

a. **Minimal Cloud Account:** There is a great interest in using GPU and Cloud for research and teaching. However, accessibility and usability were two serious problems. Given the prevalence of cloud, the faculty felt that there was a need for all faculty to be provided with a minimal cloud account.This will encourage more faculty to become conversant with it and experiment with other associated technologies.

b. **Research Engineers to Support Local Needs:** Given the diversity of faculty in SPS, it was critical to have a research engineer affiliated with CUIT to help SPS faculty in configuring components to create a custom workable system for each faculty/class need.

c. **Shareable Data Repositories and Storage:** There is a need for data repositories which can be shared across the university which will encourage collaboration across departments and schools. Currently it has been hard to move data around different schools, or keep track of other relevant research.

d. **Access and Licensing of External Data:** At times, where there is a need for accessing data from external resources. There is a lack of information about whether Columbia has already an access, or if not, then having help in negotiating data access. After all the data is the fuel which powers all the intellectual activities across the campus.

e. **Configuration Management System:** There is a need for some sort of easily usable configuration management system to integrate components and libraries. For example, though there is Singularity virtual environment in HPC clusters, it is rather complex to get any new components added since the users do not have SU privileges and getting any new component requires hand-holding by CUIT.

f. **Chatbot for Columbia Use:** For commonly available information, it would be worthwhile for CUIT to develop a chat-gpt kind of interface for novice users.

G. **Support for New Emerging Technologies:** Currently we do not have much support for the computers which are other than MAC and Windows outside of the College of Engineering. We need to develop this for other OSs, e.g., Ubuntu machines, which are commonly used in Machine

Learning. As the new technologies and computing environments become available, it would be important for CUIT to develop expertise and support faculty in the use and trouble-shooting of components using these new technologies.

CUIMC Faculty Response on Research Computing

In 2022, the Columbia University Irving Medical Center (CUIMC) created a short-term task force on Research Information Technology and Computing (RITC) to advise on the status and future of RITC services at CUIMC, completing its report in February 2023. Based on that report and on the 2023 university Research Computing Faculty Committee survey, we present these findings.

CUIMC Main Findings

Across all CUIMC schools, research progress is often impeded by ***inadequate information technology (IT) infrastructure and research services*** requiring workarounds and external subcontracts resulting in lost scientific opportunities and revenue.

Financial, networking, security, access, and reliability challenges exist for researchers using the available on-premises ("on-prem") and cloud computing services. These challenges are compounded by our ***aging IT infrastructure and lack of adequate coordination*** between operational units (e.g., CUIMC IT, Facilities, and the Center for Computational Biology and Bioinformatics (C2B2) in the Department of Systems Biology).

To be a leader in biomedical research and research computing, CUIMC must ***develop and implement a research strategic vision*** and partner with the rest of Columbia University as well as NewYork-Presbyterian Hospital, and participate in regional and national consortia to seize opportunities and address threats.

The task force evaluation suggests that for the foreseeable future, from both a financial and an operational perspective, CUIMC research information technology and computing needs will be best served by a ***hybrid model***. This model should offer a transparent, navigable, coordinated mix of on-prem and cloud resources for high-performance computing (HPC), data storage and management, and central, fundamental IT services that support local innovation. This model should undergo a regular review.

We emphasize the importance of ***data privacy*** at CUIMC given the volume of protected health information (HIPAA), which must be taken into account for any shared infrastructure. Sharing data generally requires several appropriate levels of permission and protections.

CUIMC Vision

CUIMC needs a single service-oriented “front door” for central support of research HPC and IT. This will facilitate equitable and streamlined access to HPC, data storage, data management, and support services throughout the research lifecycle. CUIMC IT must be guided by faculty leadership and faculty needs assessments to prioritize RITC central services while partnering with expertise in the schools and department to maximize efficiencies, enable research innovations and drive new funding. CUIMC faculty leadership will evaluate needs to make strategic decisions about future investments and partnerships with CU and NYP, as well as industry and other academic partners, to support the research vision and mission.

CUIMC Recommendations

1. We need leadership and governance reorganization around RITC support services at CUIMC. We propose (a) a new CUIMC role, Chief Research Information Officer to oversee strategic planning and direction of resources devoted to RITC across the four schools at CUIMC; (b) a CUIMC Faculty Advisory Committee on RITC; (c) hire a CUIMC Director of RITC, reporting to the CUIMC CIO, to operationalize and implement the vision developed by CRIO and Advisory Committee; and (d) encourage school-level RITC leadership.

2. We must engage in short and long-term strategic planning for HPC and data storage and management (on-prem and cloud). This must be done in collaboration with the rest of the university.

3. We must coordinate and invest in existing CUIMC, school, and departmental IT resources to support the research mission. We must gather data from all user groups, upgrade the network, establish a phased plan to support a hybrid HPC environment, and increase RITC services. Collaboration with the university will likely result in economies of scale and stability.

4. We must create a financial model and invest in infrastructure to support integrated RITC operations.

CUIMC Additional Materials

CUIMC Landscape Analysis

Research at the medical center and the university involves exponentially larger and more complex data necessitating ***access to reliable high-capacity computing power and data storage systems***.

CUIMC does not have ***faculty-level leadership and governance nor a singular strategic plan*** for RITC. There are no guiding principles or needs evaluations in prioritizing investments and infrastructure.

Networking conditions within and between campuses and externally are a significant barrier to scientific progress.

CUIMC ***HPC resources*** have out-of-date infrastructure and inadequate central support from Facilities and IT and are viewed as unreliable by some researchers and groups. The medical center is not well-positioned to support the expanding HPC environment (HPCE) for the research community. Prioritization, coordination and efficiency, and financial investment are all required. Central support from IT and Facilities teams will be essential for a data center's reliability, sustainability, security, and integrity.

"On-prem" HPC (through C2B2) and "off-prem" cloud services (through CUIMC IT & Departmental CITGs) are provided. However, ***messaging about options, services, and pricing lacks coordination and can conflict***.

Cloud computing presents new challenges (and opportunities) for academic medical centers, including: (1) an inability to estimate and monitor costs accurately; (2) complex and evolving cloud pricing models; (3) rapidly evolving sets of cloud computing software services; and (4) training and support for cloud computing usage.

To plan and budget appropriately, grant-funded investigators need more resources and services to estimate costs for the use of cloud services vs. on-prem HPC and data storage. ***Upfront guidance and transparent fee structures to compare all options are necessary.***

Many critical research services and trainings (through CUIMC IT, CUIT and at schools/departments) are ***poorly accessible, unknown or simply lacking***. RITC support is uneven, with some departments investing in support (CITGs) and others relying on central support that needs to be more adequately staffed or deployed.

A review of several leading academic medical centers (e.g., Vanderbilt, Stanford and others) underscore the importance of a ***robust academic and research-focused governance structure*** involving faculty for strategic planning, needs assessment, prioritization, transparency, and communication. This requires collaboration with the university and health system to leverage investments and vendor partnerships to maximize service and impact.

CUIMC Responses to the 2023 Research Computing Faculty Committee Survey

This is a summary of the CUIMC responses to the Columbia research computing survey. There were 96 CUIMC responses as of 9/8/2023 with two duplicates. It is not intended to replace the original comments, which are worthy of individual review. The purpose here is to put those comments within the whole context. That something was mentioned frequently probably means that it is important, but that something was mentioned only once does not mean that it is unimportant. Some responses were clearly frequent because they were mentioned in the survey prompt. A comment about a single area like data storage could really be about infrastructure, funding, training, privacy, or other, and the attempt here was to separate them into the underlying intents to the extent possible.

Comment	Number
Infrastructure	133
Storage	32
Huge databases (terabytes)	6
Supporting inter-institutional work	4
High-performance computing	25
GPU	11
Backup	14
Designed for compliance	13
HIPAA	9
FERPA	1
Semi-secure data	1
Tools (overlaps services)	10
RedCap	5
Other (R, SPSS, Digital signing, Shiny apps)	4
High-speed network (mostly for data transfer)	7
Access to health record data (mostly Epic)	7
Shared datasets (mostly internal for this row)	7
Remote access for research	5
WIFI	4
General refence to cloud solutions	3
Classroom IT	2
Provision of external data sets	2

Services	36
IT support in various forms	16
Examples (HPC, Linux, R, GPU, software update)	5
Collaborative research computing	7
Examples (Ronin, Colab, code repository)	3
Statistical services	5
Bioinformatics services	2
Other (language processing, machine learning, anonymization, ...)	6
Funding (thrust was mostly how to pay)	14
Free or cheap computing cluster	5
Software licenses	3
Training	9
Examples (coding, language processing, AI, statistics)	4
Policy	3
Allow tools beyond Microsoft	1
No more phishing emails	1
Less strict control of workstations and servers	1

Data storage, high-performance computing, and backups were predominant. There were requests both for local and cloud versions of data storage and computing, depending on the context. GPUs came up repeatedly. Access to health record data was frequently requested, along with a request for access to the Epic system itself (e.g., to provide decision support). Compliance with privacy regulations like HIPAA also came up frequently. Inter-institutional work came up in several contexts including data sharing and external collaborative research.

Many comments were requests for tools and services. The most common was for IT support in various forms, but there were requests for research-specific tools like RedCap and collaboration software and services like bioinformatics and statistics.

Several generic IT requests surfaced, such as getting the WIFI to work, remote access from home for research, and classroom provisioning. Training came up as did a few policy issues. A number of comments were really requesting funding rather than a specific change in infrastructure or services.

CUIT Point of View

The committee endeavors to recommend a set of centralized research computing resources that are broadly useful enough to provide the means for every researcher to be competitive technically and financially on every grant. There are two major issues that must be tackled in order to create something useful enough to draw researchers towards the resources (as opposed to mandating they use them):

1. The resources must be relevant to the needs of a variety of researchers and not become antiquated in a short amount of time
2. The resources must be significantly more cost effective than "going it alone"

When it comes to relevant resources, one must consider the variety of research, the diversity of researchers, and constant evolution of available technology. Those factors demonstrate definitively that making any single resource (or type of resource) available will not suffice to meet the committee's charge. This implies that there must be a portfolio of technologies provided, or even more to the point, a variety of services available for researchers to choose from. Services can range from datacenter floor space to managed GPU clusters to cybersecurity responses on grant proposals – the only common thread being that they are useful enough to the research community to warrant their cost and upkeep.

To define and maintain a relevant portfolio, there must be faculty governance. Not unlike SRCPAC, but perhaps with a wider charge, a faculty governance committee to continuously monitor the portfolio, adding services and convalescing them as needed will be essential to the success of anything that is built and maintained centrally. Importantly, the governance must be representative of the diversity of schools and researchers; anything built centrally should not be just for the few technically skilled researchers, rather these resources should be useful across the spectrum of researchers regardless of their technology acumen. Certainly there are obvious resources that are identifiable enough at the outset, such as datacenter space, HPC/GPU cluster access, access to AI tools (such as LLM's and machine learning platforms), cloud environments, etc. – but all of those items that we can list need additional detail to turn them into an actual service offerings, and faculty input is required to do just that. The Executive Vice President for Research office should play a role in this governance, ensuring that faculty participation is consistent and representative.

Defining a portfolio and providing the rules around it will only be valuable if there are people to carry out those directions from the governance committee. This means that beyond the portfolio software, hardware, space requirements and other physical elements there will need to be an appropriately skilled central team to maintain the entire portfolio. A professionally managed team that provides the structure to maintain the necessary talent that supports the research resources will be essential to keeping the entire portfolio of offerings relevant. Beyond just having the resources available, it will manage on-boarding, career-pathing, leave management, performance reviews, training/skilling up and the other myriad of management activities required to keep a team performing well and serving the researchers. The central team would also be charged with communication and awareness campaigns, to make sure researchers know what is available and how to access it. Communication materials would not be limited to simply "advertising" internally to the university, but there could be elements of communications materials that can help researchers when they are developing proposals. Finally, An essential feature of the centralized team would be to continue and develop the embedded model in place today (or "data/research navigator" model at CUIMC). This model provides schools and departments with dedicated skilled resources to assist with bringing technology to their research and maintaining it throughout the grant lifecycle.

Having a portfolio defined and a supporting group of human resources will only be useful if it's financially attractive to the researchers, and that will only be true if there is enough central funding to subsidize these resources to a point whereby the financial differences are impossible to ignore or argue with. That is why it is important that we get buy-in from school deans and central leadership to provide on-going financial support so that we can "lift all boats" at Columbia. It is therefore recommended that the committee act as the first iteration of the faculty governance committee and settle on a list of initial services. CUIT emerging technology team and RCS will then act in the place of the future state support model and provide the necessary costs, both capital and operating, to present to university leadership. Assuming approval, we would then move forward with building the centralized team under CUIT and bootstrapping the initial services.

None of these services would be compulsory, rather the opposite – it should be that researchers seek to use these services. This effort is not meant to supplant resources that researchers already have and find sufficient or superior to any centralized resource. However, we believe that the most cost-effective way to get the vast majority of researchers access to the technology they need while staying competitive on their grant proposals will be to build the portfolio and supporting structures centrally. We seek to provide more access to more people, thereby making it worth the investment to the deans and university leadership.