Report: Shaping the Future of UK Large-scale Compute

This survey captured the views of 247 researchers and innovators across various universities, organisations, and firms. While physics, engineering, and computer science were the most represented fields, valuable insights were also gathered from the biological sciences, medicine, social sciences, industry representatives and others.

The findings highlight key concerns and recommendations for the UK's Large-Scale Computing (LSC) strategy:

- The respondents have stated their impression that the UK is lagging behind in its overall compute capacity. This takes two forms: lack of access to existing compute systems as well as insufficient compute power to tackle the biggest workloads. There is a difference here between the total capacity required within the system and the way in which that capacity is apportioned into different systems.
- The strategy should address a wider range of factors beyond just compute power. These include data storage, skills and training, software services, and data transfer, to tackle bottlenecks effectively, or avoid creating them when new compute comes online.
- Transferring large data sets, codes, and software for use on new systems presents a significant barrier for adoption. A key aspect of this effort should be the development of standardised data transfer protocols and tools to streamline data movement between systems.
- Industry users have indicated they require less computational power on average than their academic counterparts. Cost is paramount for these users, which is compounded by their reliance on cloud-based computing solutions.
- The UK research compute communities are divided among regions, institutions and available LSC systems that have different resources, cultures, and requirements. Working to ensure federation and interoperability will mean a more integrated and efficient compute ecosystem, leveraging the UK's already significant knowledge resources.
- Some of the skill requirements mentioned by respondents can be met by robust software and software research engineer (RSE) resources.
- There is a strong interest in establishing platforms for collaboration and interaction between research communities. These platforms would facilitate sharing best practices, relevant tools, and information regarding LSC for research and innovation.

Report on large scale computing community engagement

Recent international developments in Large-scale compute, such as the inauguration of the world's first exascale computer in Oak Ridge National Lab in the US and the announcements of other petascale and exascale computers in Europe, Japan, and Korea, among others, have evidenced a world-wide push towards securing large compute capacities to enable high-impact digital research. Transformative research and innovation, in areas such as weather modelling, pandemic preparedness and AI development, among many others, has become increasingly dependent on more powerful Large-Scale Compute systems.

In the case of the UK research compute environment, the Independent Review of the Future of Compute (FoC) has helped to illustrate the pressing need to update the country's compute capacity to meet changing research needs. Increasing compute capacity, as well as promoting and facilitating access to current and future compute resources, is then a crucial requirement in achieving the UK's goals and maintaining the UK's position as a Science and Technology Superpower world leader in breakthrough digital innovation. Promoting this requires a comprehensive, long-term strategy that takes advantage of the country’s considerable research community and resources. It also depends on comprehensive and proactive coordination between researchers, providers, and funders.

Since the publication of the FoC Review, significant efforts and funds have been dedicated to develop new compute resources aimed at supporting researchers and providing advanced data processing capabilities. Among these efforts, the Artificial Intelligence Research Resource (AIRR), and the Pathway to Exascale programme are worth mentioning due both to their scale and to their role in providing state-of-the-art compute facilities.

To make sure that the correct mechanisms and strategies are put in place, and that the compute requirements of current and future users are met in the most effective way possible, UKRI has embarked on a community engagement initiative. This has been done in recognition of the difficulties inherent in reaching the heterogenous communities that make up the UK’s compute environment. As such, efforts have been taken to include the views of UKRI’s constituent councils and other relevant actors through council-led workshops, stakeholder and community engagement activities and technical working groups, among others. To ensure that improvements in the UK compute infrastructures are made to fit the requirements of users, both present and future, UKRI has sought to implement a broad community engagement strategy.

Future of UK Large Scale Compute Survey

The Future of UK Large Scale Compute (LSC) survey constitutes a key component of UKRI's multifaceted stakeholder engagement strategy for the creation of a UK Large Scale Compute Strategy. With 247 individual responses, this survey seeks to complement other initiatives such as town halls, workshops, and external consultancies, and should not be considered a replacement of those spaces but instead add to the discussion. The survey instrument itself was designed to gather broad input from diverse research and innovation communities. The survey framework incorporated provisions for deeper engagement with interested participants through follow-up mechanisms. This combined approach aims to capture crucial insights on user needs, challenges, and current use cases for high-performance computing systems, while at the same time allowing for continued engagement efforts with the community. By incorporating this wealth of data garnered from various engagement methods, UKRI is positioned to develop a compute roadmap that effectively addresses the evolving requirements of the research and innovation landscape.

Figure 1. type of LSC users

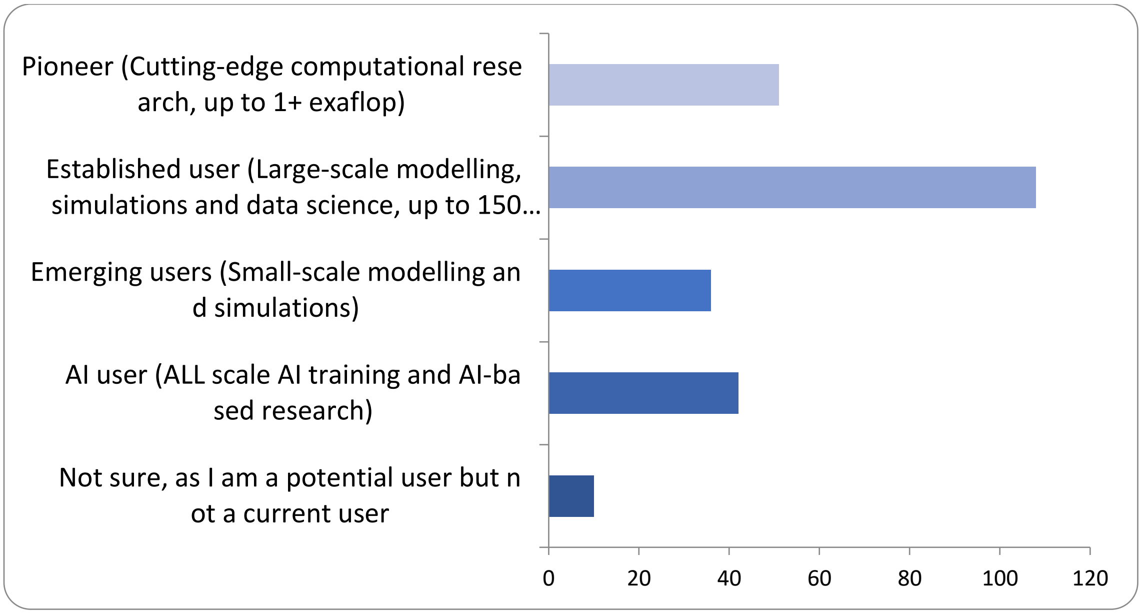


Figure 2. Users by sector

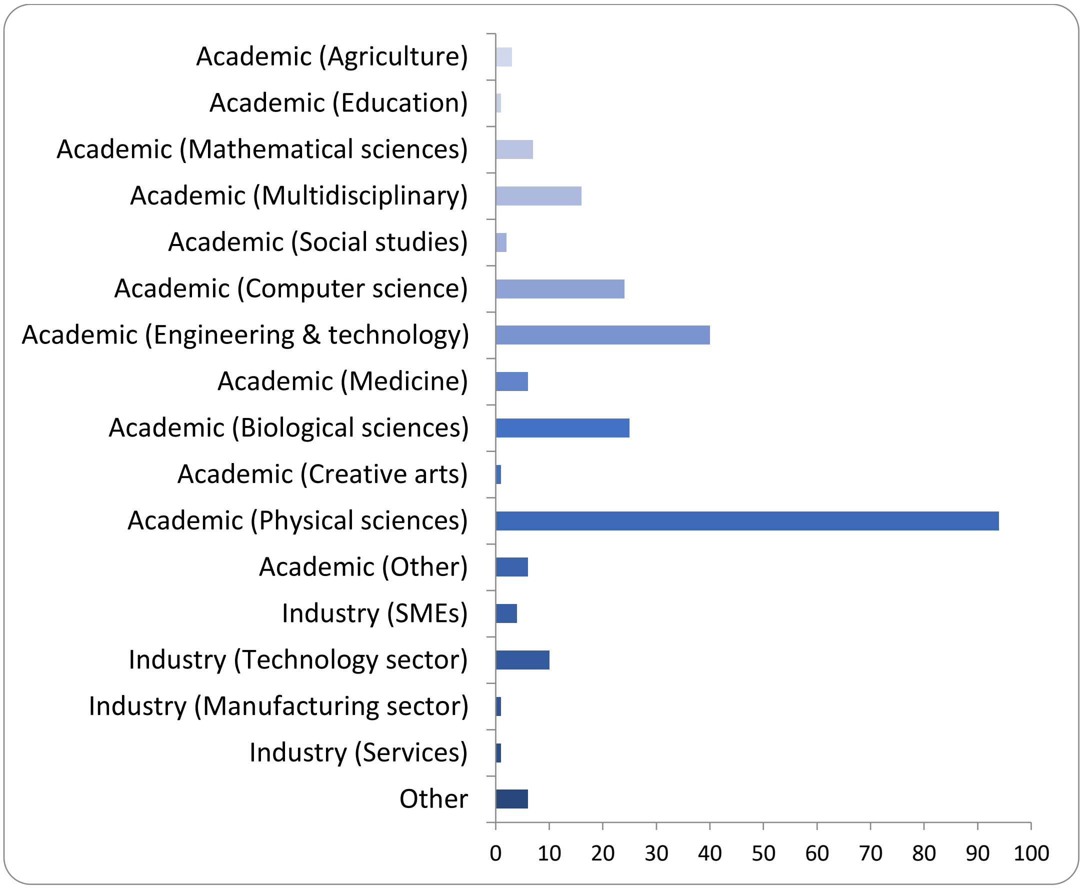


Figure 1 shows the self-identification of users and potential users of HPC resources, with the largest group being established users. However, the survey also captured significant responses from less traditional sources, such as pioneer, emergent and AI-focused users. This, while not necessarily representative of the research community, helps to understand its heterogenous nature. On the other hand, **Figure 2** illustrates the fact that some sectors are more solidly established than others. In this case, the largest user communities come from the Physical Sciences, Computer Science and Engineering sectors, though the others do have a significant presence.

Relationship between industry and LSC resources

A clear distinction emerges when examining responses from industry representatives, particularly small and medium-sized enterprises (SMEs) and the wider technology sector, and their relationship with national LSC resources compared to academia. Industry users have indicated they require less computational power on average than their academic counterparts. **Cost is paramount for these users, which is compounded by their reliance on cloud-based computing solutions**. However, this focus on cost isn't a limitation on future growth. **Industry respondents specifically cited software support and training as critical factors in enabling them to significantly increase their use of LSC** for future research and development (R&D), and have referred to key UKRI resources, particularly those provided by Hartree as necessary in scaling up their research capabilities. This indicates a willingness to embrace more demanding computational needs as innovation continues to push the boundaries of what's achievable. In essence, **industry respondents see national LSC resources as a potential future partner in their R&D efforts.** However, overcoming the cost and access barriers and providing adequate software support and training are key to unlocking this potential.

Furthermore, most of the industry respondents have identified themselves as primarily AI users, with some identifying as pioneer. Since most of the work and research they undertake using LSC is AI-related, this means that **GPU and machine learning-optimized systems are currently seen as a priority for industry-led innovation**. Additionally, most compute resources for the UK industry, cloud computing, hardware, commercial software and even some specialised training programmes, are provided by foreign firms, creating concerns of dependencies, vulnerabilities in supply chains, and other risks.

The heterogenous nature of the UK's LSC environment

The respondents have mentioned using a large variety of LSC resources, both national and international to suit their research needs. A significant portion stated that they regularly use more than one system, normally of different tiers, either because they need to scale up their compute demand for particular processes or because they require specific architectures and support systems for their workloads. There is also mention of the need of better support to facilitate the change to larger and newer machines, such as petascale and exascale LSCs. **This highlights the need to increase interoperability and federation measures to allow for easier transference of data as well as the porting of codes and programmes in order to make a more integrated and efficient compute ecosystem.** Given that the ability of researchers to effectively use state-of-the-art systems is currently ensured, securing interoperable capabilities is seen as a critical step in increasing the country's compute power in research, according to its users.

In terms of usage and preference of LSC systems within UK research communities **the most used type are tier 2 (Regional) supercomputers, with approximately 39% of surveyed researchers using them with certain regularity, followed by 36% and 31.2% for Tier 3 (Institutional) and Tier 1 (National), respectively**. Users also describe significant differences when using different system types, from service provisioning as well as specialised and workload-efficient architectures which points further towards silos of compute in the UK research ecosystem. **There is also reported use of other international and private systems**, such as limited use of cloud compute for certain limited computations as well as employment of larger LSCs, either through international schemes such as PRACE or through specific partnerships with large systems like Oak Ridge national Lab's Frontier system and Finland's LUMI.

Figure 3. Ranking of priorities for the LSC ecosystem

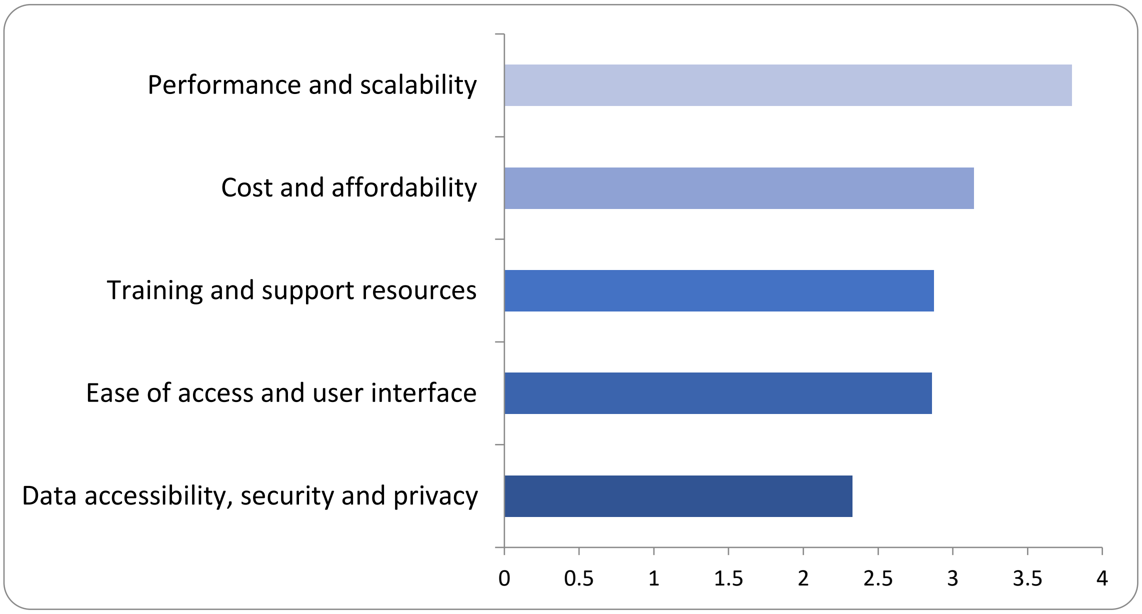
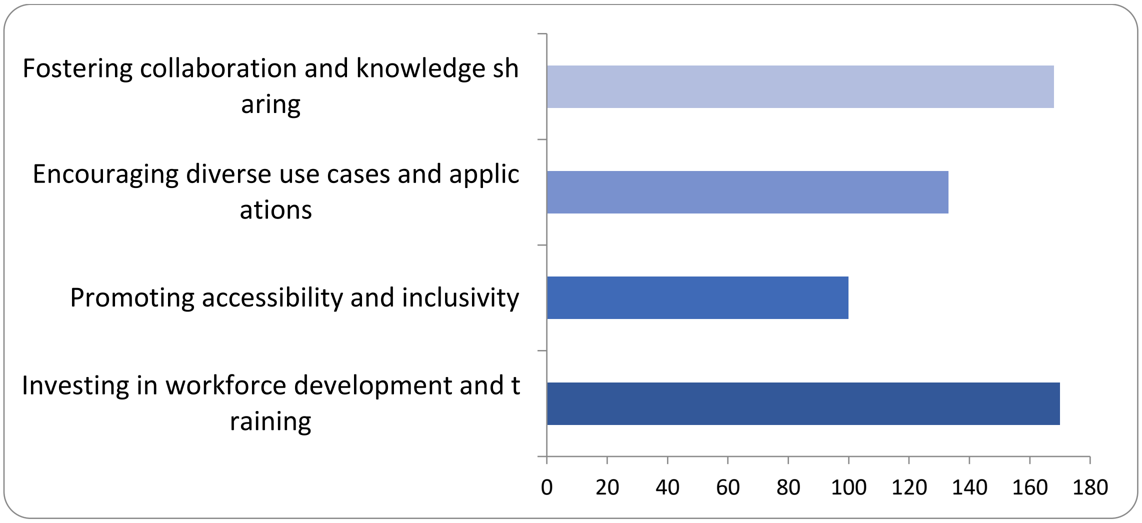


Figure 4. Priorities for UK large-scale compute landscape in the future



Priorities in a national LSC ecosystem

Survey respondents were asked to rank their priorities for the UK LSC ecosystem according to their current and expected needs as well as their experience with existing systems. **Figure 3** shows that while Performance and Scalability is considered the main priority, all five options were prioritised rather evenly among respondents. The answers given in the survey points towards a general consensus of the **need for a holistic strategy for improving UK capabilities, focused on addressing potential bottlenecks**. This also recognises that addressing any country's research compute requirements goes beyond increasing raw processing power.

Requirements and resources also differ significantly within and without individual research communities. Beyond differences between academic and industry sectors, there are also regional differences that have to be addressed. **Access to power, high upload speed for data transfer, access to skilled workforce or training resources can vary greatly across the country and has an impact on the research and innovation outcomes.** This can be addressed and managed by a holistic infrastructure strategy that considers the specific requirements of the communities and works to eliminate inequalities through capacity building and collaboration.

Adapting to new systems

The biggest concern surrounding new LSC resources, like exascale and AI-optimised supercomputers, is adapting to their unique architectures, software, and workflows. **This highlights the critical need for enhanced interoperability between these systems.** To achieve this, increased collaboration and dedicated resources are essential. Research Software Engineers (RSEs), training programs, and upskilling initiatives can equip researchers with the skills to utilise these new systems effectively. This wouldn't just benefit current Tier 1 and Tier 2 users seeking to leverage exascale power, but also researchers with lower or less frequent compute needs. These users could still leverage the training and software infrastructure to enhance their capabilities.

Furthermore, transferring large data sets, codes, and software for use on new systems presents another significant barrier to adoption, impacting both new and established users. This further emphasises the importance of a coordinated effort to build computational capacity. **A key aspect of this effort should be the development of standardised data transfer protocols and tools to streamline data movement between systems.** This approach, combining user upskilling with technical solutions, will ensure smoother adoption of next-generation LSC resources and maximise their impact on scientific discovery.

Ensuring UK's leadership in science through access to compute

The respondents have highlighted a critical challenge for the UK research ecosystem: limited access to high-quality, large-scale compute systems compared to other established technology superpowers. This disparity stems from a lack of coordinated investment and a clearly defined strategic focus. The consequences are potentially severe. Without access to cutting-edge computational resources, researchers may be forced to simplify complex scientific investigations, diminishing their value and impact. **Additionally, the inability to conduct research at the leading edge could drive talent overseas, weakening the UK's overall research competitiveness**. Over 40% of all respondents explicitly mentioned that lack of sufficient LSC is their main concern in regard to digital research.

There's a demonstrably increasing demand for LSC resources, driven by several factors. Firstly, the size and complexity of datasets in fields like genomics and climate modelling are increasing significantly. Analysing these datasets requires immense computational power to extract meaningful insights. Secondly, the development of sophisticated software and tools in areas like machine learning necessitates powerful computing resources to train complex models and perform large-scale simulations.

The current computational landscape demands a balanced approach. While the development of exascale computing capabilities is underway to tackle problems of increasing complexity, **the continued value of smaller and more diverse LSC systems cannot be overlooked.** Researchers across various disciplines have mentioned a clear reliance on these systems for critical tasks like data analysis, high-fidelity simulations, and complex visualisation. Each user has a specific requirement profile, and some rely on particular resources more than others, such as specific technologies and software that aide in visualisation or particular data access and storage agreements for sensitive medical data, as an example. Given the stated scarcity of compute time, even within established tiers (tiers 1 and 2), dedicating petascale and exascale LSC facilities to smaller or exploratory computations would represent an inefficient allocation of resources.

Demands for "more compute power" don't always translate into exascale needs. Many researchers have declared a need for readily available institutional and regional systems for their daily work**. A key focus should be on establishing access models that are both equitable and inclusive**, effectively utilising the existing national computational infrastructure. This approach ensures that researchers from diverse backgrounds have equal opportunities to pursue cutting-edge research, unhindered by bureaucratic hurdles or technical complexities.

Beyond compute

In some cases, researchers may not necessarily need faster machines or vast compute power, or these needs are compounded by other obstacles. Conversely, the use of diverse tools and different approaches might actually lessen the demand and streamline the use of resources. Because of this, respondents were asked to identify other existing and potential barriers to their digital research processes, apart from accessing more compute power. **The two factors that were most commonly cited were the need for more training opportunities and skilled workers and accessible software services**.

Training a next generation of users

A crucial aspect identified by both researchers and industry representatives is the necessity for ongoing skills development. This ensures researchers and their collaborators possess the necessary competencies to effectively and efficiently LSC resources. The current LSC landscape in the UK necessitates a certain level of technical expertise for operation. This requirement extends beyond the lead researcher to encompass their support teams as well. Specifically, this encompasses ensuring staff are equipped with the skills to manage data and its associated metadata appropriately, maintain and update code effectively, and support transitions between different tiers and architectures of computing power.

Furthermore, **fostering the dissemination of information regarding relevant tools, software, and processes among users is critical**. This collaborative approach would facilitate more efficient research processes and pave the way for the adoption of novel computing methodologies.

However, to maximise the return on investment in skills and training resources, a tailored approach is essential. Exhaustive, time-consuming training courses may not be suitable for the fast-paced research environment and available workforce. PhD students and research assistants, for example, are often simultaneously familiarising themselves with their specific research areas and may not have the time or inclination to become broad experts in compute systems. In such cases, **a more targeted**

approach focusing on equipping staff with the specific programming and data analysis skills directly relevant to utilising the LSC systems may be more beneficial.

A comprehensive software ecosystem

Alongside the focus on continuous skills development, respondents identified another crucial factor: the need for better software services. These services bridge the gap between the skills training programs can't fully equip researchers with and the ever-evolving demands of their work. Ideally, these comprehensive support structures should seamlessly integrate into the research process, guiding users and university staff towards achieving their desired research goals efficiently and sustainably.

The data gathered from respondents highlights the importance of such services including access to highly trained personnel like RSEs. Their expertise, combined with well-maintained and comprehensive software libraries and appropriate license agreements, would provide an invaluable foundation for researchers. **This software support system would be particularly beneficial in facilitating the transition to newer, more powerful LSC systems.** Additionally, it would empower researchers to adopt GPU systems optimized for workloads in artificial intelligence and machine learning. The critical role of software services is further emphasized by the fact that a lack of such support was cited by several researchers as a major barrier to upgrading to more powerful or specialized systems. By investing in software services and environments, the UK can empower researchers to fully leverage the potential of LSC resources.

The role of storage and research data management

Alongside the challenges previously mentioned, according to respondents, **data management has become a top concern for researchers and the development of future proof LSC systems**. This stems from a confluence of factors. The explosion of data, with modern scientific experiments generating ever larger and more intricate datasets, necessitates sophisticated digital resources for effective handling and utilization. The growing complexity of digital infrastructure further highlights the need for robust data storage solutions in developing effective LSC systems.

Respondents have emphasized the prevalent lack of adequate data storage. **Long-term remote storage infrastructure ensures data remains accessible over extended periods, critical for long-term research projects.** Additionally, short-term on-site storage facilitates efficient data processing during active research phases, enabling researchers to work with their data swiftly.

Furthermore, current, and potential users have stressed the need for strong data ecosystems. These ecosystems should provide comprehensive support throughout the data lifecycle, encompassing data collection, curation, processing, and storage. Building robust data pipelines is crucial for ensuring the quality, efficiency, and responsible conduct of research. **One specific concern raised by some researchers relates to limitations in data transfer, as well as the lack of skilled data professionals.** Insufficient bandwidth and trained personnel can hinder the movement, access and curation of large datasets for computational analysis, creating a bottleneck in the research workflow.

Across various disciplines, particularly social and medical sciences as well as industry, concerns regarding data governance and ethics have emerged as a significant barrier for collaborators and administrators. These concerns stem from uncertainties surrounding both the origin and nature of large datasets used in research, as well as the current state of global and national data management standards. This lack of clarity leads to hesitation in utilising LSC resources, potentially hindering research progress. To address this, ensuring researchers have access to effective guidance on responsible data usage and establishing clear data management standards are crucial steps. By taking these actions, the UK can promote responsible research innovation and foster a more confident environment for researchers working with large datasets.

Collaboration and sharing of best practices.

A recurring theme identified by researchers and industry representatives is the need for more formalized and regular communication channels between LSC users. These channels encompass dedicated online platforms for knowledge sharing, facilitating discussions and housing resource repositories. At these events, researchers could present their work and methodologies employed on specific supercomputers, offering a glimpse into real-world LSC applications. **Complementing these online and showcase-based interactions, hosting workshops and conferences specifically focused on LSC utilisation would foster deeper collaboration and knowledge exchange**.

Given the intricate nature of research requiring LSC resources, supporting the flow of information and experiences is critical for nurturing a thriving and effective research ecosystem. Implementing such programs would significantly address concerns regarding skills and training, particularly the reported lack of awareness among UK researchers about available tools and methods. By actively sharing best practices and fostering a culture of knowledge exchange, researchers would be better equipped to tackle complex problems and maximise the potential of LSC resources.

Furthermore, respondents have expressed an interest in the creation of further spaces **promoting the participation of UK researchers in international forums specifically focused on LSC,** which would expose them to cutting-edge advancements and facilitate collaboration with leading international researchers and organisations. These forums often feature presentations, panel discussions, and networking opportunities, allowing researchers to stay abreast of the latest developments in the field and forge valuable connections.

Additionally, fostering collaborative spaces for international researchers and organisations to share knowledge and best practices could be facilitated through initiatives such as joint research projects, international research fellowships, and virtual guest lectures. By creating opportunities for international collaboration, the UK research community would benefit from the exchange of expertise and the ability to tackle global challenges by leveraging combined resources.

These combined efforts would create a vibrant and supportive research environment. This environment would enable UK researchers to access the knowledge, expertise, and tools needed to fully utilize LSC resources and contribute to advancements across various scientific disciplines.