

IT Roadmap for enabling AI at Universities

A practical roadmap for universities to build AI-ready infrastructure, align culture and workflows, and enable researchers and leadership to realize lasting value from AI

Med finansiering från:

VINNOVA



Medfinansieras av
Europeiska unionen

Table of contents

Executive Summary..... 3

Introduction 4

The Academia Context: A Fragmented Operational Landscape 5

Prerequisites for a Successful AI Transformation 9

A Practical Roadmap for an AI-Ready IT Environment in Academia..... 10

Strategic Alternatives: The Build-Versus-Buy Decision..... 16

Key Takeaways 18

Conclusion 19

Authors and acknowledgements..... 20

The project “Data-driven organizations – Best practices for operationalization of AI in Sweden” 21

Executive Summary

Universities are under increasing pressure to support AI across administration, student services, education, and research. While expectations are rising quickly, most institutions lack the infrastructure, workflows, and organizational readiness required to deliver AI at scale. This white paper presents a practical, experience-based roadmap, developed through the Vinnova-funded DDO project and dialogues with Swedish universities, that guides IT departments in building an AI-ready environment.

The roadmap emphasizes staged progress over large, speculative investments. It begins by rapidly expanding access to GPU resources and introducing container-based workflows, enabling researchers to adopt modern practices while lowering operational friction. It then transitions into establishing a Kubernetes-native, hybrid architecture that supports both predictable local workloads and temporary high-volume demands. Finally, it consolidates culture, governance, and long-term sustainability.

Taken together, these steps provide a realistic path for universities to modernize their IT foundations, align technical capabilities with educational and research needs, and ensure that AI becomes an integrated, reliable part of everyday academic operations.

Introduction

Artificial Intelligence (AI) is transforming research and education. Researchers rely on it to accelerate discovery, and leadership views it as a driver of competitiveness. Yet, many institutions currently lack the infrastructure and practices needed to support this shift across the university, from internal administration and student-facing services to educational activities (both applied and development-oriented) and research.

Responsibility for "AI transformation" often defaults to the IT department, but IT cannot deliver it alone. The department provides the technical backbone; realized value comes when researchers adopt new workflows. Without organizational support for this cultural change, infrastructure remains just hardware.

This white paper presents a practical, experience-based roadmap for university IT departments. Building on insights from the Data-Driven Organizations (DDO) project, it outlines a structured path to AI readiness.

The strategy advocates for staged progress rather than immediate, costly leaps to "state-of-the-art" solutions. This synchronizes technical upgrades with user adoption, ensuring infrastructure scales efficiently as researchers learn to utilize it.

This guide is written for IT departments, yet the insights are equally relevant for university leadership and researchers. Successful AI adoption relies on aligning technical upgrades with cultural shifts, and a shared perspective will smooth the path forward.

The conclusions and recommendations originates from discussions between Malmö University IT department, Linköping University IT department, and the DDO project participants Red Hat, Stormgrid, Proact, IBM, Aixia, Netapp, and AI Sweden. Together we explored their current setups and goals. We decided to leverage the combined experience from the group and the DDO learnings to design a high level roadmap for the IT departments.

The Academia Context: A Fragmented Operational Landscape

Understanding the starting point for a typical Swedish university is essential before outlining any roadmap for AI readiness. Through dialogues with several institutions, recurring organizational and technical patterns have emerged that we wish to highlight.

Universities operate as a duality: they are unified institutions yet act as collections of semi-independent units. Because research funding is typically individual, decisions regarding hardware and tools are made locally, resulting in a landscape that is cohesive in purpose but structurally fragmented, adding to this technical universities have a long history of seeing internally built tools used for various purposes and it is to be expected that this will continue in the AI-area, in fact given so called “vibe-coding/engineering” now available it will probably increase. Such solutions are used both university wide and at research group or individual levels, adding to the fragmentation.

Many university IT departments successfully manage traditional on-premises environments (typically virtualization-based) providing shared services and administrative systems. Some have integration with external hyperscalers for short-term needs. However, few offer specialized centralized resources for researchers, such as GPU clusters or Kubernetes environments, with the same ease as traditional services.

Without centralized GPU access, AI work creates a polarization of resources, leaving a critical gap in the middle:

- **The Small Scale (Workstations):** Lacking institutional alternatives, researchers purchase individual workstations with consumer-grade GPUs and the same type of setup is implemented in computer labs for teaching, albeit through central IT. While this provides immediate access, it is inefficient at an institutional level on the research side and on the allocation to students on the teaching side (since computer labs tend to be shared with courses not doing AI, hindering training jobs to run outside the specific timeslot for the AI course). It duplicates costs, creates inconsistent reliability, and makes security or compliance nearly impossible. Furthermore, it

isolates researchers from modern frameworks designed for shared, distributed environments.

- **The Large Scale (HPC):** On the other end, researchers utilize national or european supercomputing resources (like NAISS or EuroHPC). These offer immense power but are heavily utilized and time allocations on them are handled similarly to any other research grant, that is under intense competition and with typical lead times of 4–8 months for approvals. This is out of the scope of IT-departments at universities.

There is currently no bridge between the extremes, there is little to no “medium” resources.

AI Requirements Across the Four Domains in Academia

To understand what is implicitly expected from the IT department, it is helpful to examine the university’s four main areas of need: internal administration, student-facing services, education, and research. Each of these domains interacts with AI in different ways and therefore places distinct requirements on IT.

- **Internal (central) administration** can in many cases rely on cloud-based solutions, as a large portion of administrative data is non-sensitive. However, certain use cases will still require internally hosted systems or governance structures aligned with GDPR and the AI Act. Administrative workloads therefore span a spectrum: many can be offloaded to external platforms, while others demand the same level of protection as traditional IT systems.
- **Student services** typically rely on open university information combined with personal student data. These services, such as chatbots, recommendation engines, or individualized study support, are generally suitable for cloud deployment, provided that personal data handling remains compliant with GDPR and the AI Act.
- **Education** presents a dual challenge. Today, most AI course requirements can still be met through dedicated “edu-clusters” that support longer-running training jobs in AI courses than GPU-equipped computer labs can. However, the landscape is quickly becoming more complex. As AI education matures, development-

oriented courses will need experiment-tracking tools (e.g., Weights & Biases, MLflow) and full MLOps flow setups to train engineers for AI-mature organizations. At the same time, applied disciplines, such as medicine and the humanities, will increasingly request operationalized AI solutions based on research outputs and adapted for teaching scenarios.

- **Research** requires the most clearly defined set of self-hosted systems. Researchers today either work on personal workstations, inefficient and difficult to govern, or wait months for access to national supercomputers. A medium-scale centrally managed environment is essential for scaling the capability to support AI for research as demand for it grows across disciplines. This includes: training smaller models, preparing optimized container workflows before moving to supercomputing resources, or running inference workloads that are typically not allowed on HPC systems due to low GPU utilization. It should be stressed that supercomputing resources have tough demands on optimization of code, and in the AI space this usually means that your training code needs to be optimized to push data into the GPU fast enough so that GPU utilization stays above the defined threshold where jobs are killed and pushed out of the scheduler queue. In addition, there is a strong need for centrally hosted experiment trackers and related support tools, both for on-premises use and for consistency across external systems. Since HPC allocations are limited and interruptions are common, researchers often need to move workloads mid-experiment (e.g., from BerzeLiUs to Alvis). This creates a practical need for seamless interconnectivity between university-hosted support systems and external supercomputing resources. Finally, we foresee that very soon many researchers will start requesting internal AI solutions for working with their own research publications before they are submitted for review, requiring the same level of protection as restricted IP at private companies. All these mentioned points drive a need for internally hosted solutions for research.

This whitepaper focuses mostly on the research part of the setup since it is the area with the most clear cut requirement on self-hosted systems (which was the main scope of the DDO project). Any implemented

solutions introduced for research will however form the foundation for emerging AI needs in other areas.

In essence, the pain points we wish to address in this paper will focus on a roadmap that should improve availability and efficiency by providing a path to centralized GPU resources, whether on-premises or in the cloud, with a particular focus on small- to medium-scale needs.

Prerequisites for a Successful AI Transformation

AI transformation is not a background task; it requires a clear mandate and strong executive sponsorship anchored at the top of the organization. Leadership must clearly communicate goals and provide the operational support necessary to achieve them. Without this explicit prioritization, the initiative will stall against the demands of daily operations.

This transformation demands an intentional shift from maintaining existing systems to building new capabilities. Leadership must explicitly allow for a temporary reduction in general productivity. Managing and serving GPU resources through Kubernetes is a specialized challenge distinct from standard IT operations. The necessary skills must be obtained through training, experimentation, and strategic hiring. Expecting IT staff to master these new tools without this investment is unrealistic.

Finally, cultural change is driven by doing, not just discussing. The most effective way to drive adoption is to provide technical "sandboxes" for hands-on experience and to actively identify and promote local champions. These early adopters bridge the gap between IT and research, proving the value of the new environment while the broader organization catches up.

A Practical Roadmap for an AI-Ready IT Environment in Academia

Building an AI-ready environment at a university is not a single procurement decision nor a purely technical upgrade. It is a staged transformation that balances researcher needs with a long-term architectural vision, while also managing organizational change. The following three-step roadmap outlines a pragmatic and achievable path for IT departments in academia.

It spans approximately three years, which is both quick and slow depending on one's perspective. The transformation, changes in processes, and redirection of financial resources will take considerable time. Experience shows that this will also create a knowledge gap that will need to be filled on both the IT side and the user side.

The technical challenge will be focused more on doing the right things at the right time rather than engaging in extensive R&D. The key is to avoid reinventing the wheel and instead identify mature, ready-made solutions that can be integrated and adapted to fit the university's needs.

Year 1: Initial Enablement and Early Engagement

The most urgent challenge is to provide researchers with reliable and usable acceleration capacity within existing budget constraints. This first phase establishes momentum: researchers gain tangible benefits quickly, and IT secures credibility as an enabler of research ambitions.

- **Scale GPU access quickly:** Equip existing rack servers with low- to mid-range GPUs and begin procurement of dedicated GPU nodes. At this stage, availability and ease of access are more important than peak performance. The goal is to ensure that AI workloads no longer depend solely on personal hardware. As an example, there are small form factor GPUs (e.g. Nvidia L4 24GB) that fit in virtually any rack server PCIe slot, that can run most mid-sized LLMs. Even if you do not use existing servers, the cost per GPU can be kept low which can ease the sourcing process. On the opposite end of the spectrum, high end GPU (e.g. 8 x Nvidia B200) servers are also competitive if paired with a suitable

framework that maximizes utilization. The key takeaway is that both options enable progress; the only wrong choice is inaction.

- **Introduce a lightweight container scheduler:** Researchers and doctoral students are often accustomed to VMs or local execution. A gentle shift can be initiated by offering container-based workflows where the “reward” is access to scarce GPUs. The scheduler acts both as a technical enabler and a behavioral nudge toward modern practices. By utilizing a lightweight, Docker-based platform like **AiQu**, IT can deliver immediate value and resource efficiency while avoiding the complexity of a full-scale platform deployment at this early stage.
- **Make containers the standard entry point:** Containers are the recommended industry practice and pair well with DevOps/MLOps workflows, but most researchers are not using them today. To encourage adoption, access to both GPUs and preconfigured, validated frameworks should be tied to using containers for the user workloads. This creates a clear incentive: researchers gain more capability with less overhead when adopting containers.

Further enable this by providing a robust support ecosystem. This includes source code management and automated pipelines (using platforms like GitLab or Gitea) alongside secure container registries (such as Harbor or JFrog Artifactory). These tools allow researchers to easily build, store, and deploy their workloads to Podman or Kubernetes environments without managing the underlying plumbing.

- **Engage the community early:** Success in this phase hinges on framing the new environment not just as infrastructure but as a shared, evolving resource. Introductory workshops, clear documentation, and peer-to-peer champions can smooth the transition. Foster cross-functional collaboration with researchers and embed a culture of continuous learning. This ties directly back to the AI transition being a whole-organization transformation rather than solely an IT responsibility.
- **Aim for IT infrastructure that natively supports AI operations:** A common mistake is building AI capabilities as isolated add-ons

rather than incorporating them into regular operations. Build technical foundations, from GPU architecture to deployment and operations, that integrate AI as a standard, supported capability within the base IT infrastructure. In other words, users, data, and storage management should function consistently regardless of whether the task involves AI.

- **Leverage supplier and partner relationships:** This transition presents opportunities to create shared value with existing and future suppliers. Both the university and its partners can benefit from deeper collaboration and co-development, particularly during the early stages of capability building.

The same container-based workflows and GPU resources introduced for research will also form the foundation for emerging AI needs in education, student-facing services, and certain administrative applications, ensuring that early technical investments benefit the entire university. This is enabled by providing a “medium” solution as discussed above.

Year 2: Establishing the Kubernetes-Native Core

With a base of GPU access and container familiarity established, the next step is to construct a scalable and future-proof architecture. This stage is less about acquiring hardware and more about operationalizing scale, developing processes, governance structures, and ways of working around the new platform.

- **Establish cloud-ready ways of working:** Introduce Kubernetes-based platforms with integrated portals and schedulers. Once workloads run in containers, transitioning to Kubernetes becomes a natural next step. Kubernetes makes it possible to scale consistently and tap into external resources such as IaaS (Infrastructure-as-a-Service), PaaS (Platform-as-a-Service), and other public cloud services.

A helpful resource is Red Hat Developer Sandbox, a free environment full of examples that closely mirrors what users can expect from a Kubernetes platform.

- **Migrate workloads gradually:** Avoid “big bang” transitions. Allow researchers to experiment in Kubernetes sandboxes, then move

partial workloads, and only later migrate core projects. Maintaining parallel VM environments during the transition ensures continuity and reduces resistance.

Universities generally have access to significant academic discounts on enterprise software solutions, which is a valuable resource to leverage. For example, this makes exploring Red Hat OpenShift as a base platform for Kubernetes, and OpenShift AI as a framework for AI workloads, particularly compelling.

- **Expand GPU capacity using a hybrid strategy:** While self-hosted AI tools and daily prototyping constitute a predictable base load perfect for on-premises hardware, student courses and funded research projects often create massive, temporary spikes. For example, a student lab may require hundreds of GPUs for two weeks, or a research grant might fund a massive one-time training run. A hybrid model allows these irregular or high-volume workloads to "burst" to the cloud, preventing the institution from over-investing in local hardware that would sit idle the rest of the year.

Furthermore, this resolves a common financial mismatch between central planning and research funding. On-premises infrastructure represents Capital Expenditure (Capex), best suited for the predictable baseline the university must guarantee. Conversely, cloud resources are an Operational Expenditure (Opex), allowing heavy, temporary workloads to be billed directly to external grants or course budgets. This distinction enables researchers to fund their own peak capacity without forcing central IT to absorb the long-term depreciation costs of hardware needed only for short-term projects.

- **Establish clear pathways to external resources:** Cloud platforms offer cost and time efficiency for specific workloads, but complexity often hinders adoption. By integrating external resources directly into the Kubernetes platform, IT can unify authentication, data handling, and access management. This makes the "official" route the path of least resistance, ensuring researchers choose supported tools over shadow IT. Crucially, centralized integration allows IT to

automate governance regarding data processing locations, preventing accidental compliance violations.

- **Model-as-a-Service internally:** Introduce an internal Model-as-a-Service layer to reduce the widespread duplication that occurs when individual researchers or departments run their own instances of the same models. Many baseline models (text classifiers, image models, embedding generators) tend to be used by many groups, and by using frameworks such as vLLM, total resource consumption can be reduced significantly.

Navigating the External Resource Landscape

Global Hyperscalers, best for massive scale and broad service catalogs. Examples: AWS, Google Cloud, Azure.

Global & specialized clouds, are often simpler to navigate with more transparent cost management. Example: IBM Cloud

Regional & Compliant Clouds, best for workloads requiring strict EU data residency or lower latency within the Nordics. Examples: Hetzner, GleSYS, PROACT Hybrid Cloud.

Specialized AI Infrastructure, best for dedicated, high-performance GPU compute without the overhead of general-purpose clouds. Example: Airon, Aixia.

Model-as-a-Service, best for accessing specific models via API without managing the underlying infrastructure. Example: Berget AI.

Year 3: Cultural Consolidation and Operational Steady State

The final stage is not technical but cultural. By this point, containerized and Kubernetes-native workflows should have become the norm, and the role of IT shifts toward refinement, stability, and long-term sustainability.

- **Reduce reliance on legacy environments:** Gradually phase down VM-heavy setups and personal workstations as container-first workflows become standard. This frees up resources and reduces fragmentation, enabling a more coherent and maintainable platform strategy.
- **Phase out distributed resources:** Existing consumer-grade GPUs are difficult to integrate properly into Kubernetes. It is often more efficient to let these distinct nodes reach end-of-life naturally while expanding centrally hosted resources in parallel.
- **Embed new ways of working:** Adoption becomes sustainable only when researchers and doctoral students clearly see value in the model. Ongoing training, peer champions, and integration into course curricula help reinforce the shift. The aim is to make these practices part of the expected academic workflow, not an optional add-on.
- **Reach a steady-state balance:** Establish a resilient equilibrium between on-premises and cloud resources, centralized and local GPU availability, and flexibility versus governance. At this stage, infrastructure is no longer a bottleneck but a catalyst for research and innovation.

The roadmap demonstrates that universities do not need to leap directly to “state-of-the-art.” Instead, they can sequence their journey: first provide acceleration, then modernize platforms, and finally consolidate culture and governance. This ensures that researchers see benefits quickly, adoption progresses smoothly, and the institution reaches a sustainable and AI-ready state within three years.

It’s worth acknowledging that year 3 is rather thin, this is intentional. We believe that if one manages to reach a state of continuous learning and development the need for drastic changes will be reduced. The goal with this roadmap and mindset is to bring AI readiness into the “mundane” and simply something that is a part of regular IT operations.

Strategic Alternatives: The Build-Versus-Buy Decision

Many of the activities in the roadmap result in an infrastructure assembled from individually selected components. While this offers maximum modularity, it places a significant burden on the university IT department. You become responsible not only for day-to-day operations but also for the "integration tax", ensuring interoperability between storage, compute, schedulers, and diverse AI frameworks.

This burden can be reduced by strategically choosing frameworks that cover more areas at the same time, such as Red Hat OpenShift AI or AiQu, but you can go one step further. This paper would not be complete without presenting an alternative path: choosing a vertically integrated supplier, such as IBM. The essence of this approach is that the entire stack, from hardware to the orchestration layer, is engineered as a cohesive unit. These solutions are purpose-built for reliability and predictable performance, often including service-level agreements (SLAs) that cover the full infrastructure.

In general, this route is the fastest path to value. It effectively outsources the architectural complexity, converting what would be a multi-year internal R&D project into a single procurement decision.

A common hesitation regarding integrated solutions is the cost of licensing or support contracts compared to "free" open-source alternatives. However, this comparison is often flawed as it ignores the Total Cost of Ownership (TCO).

- **The Hidden Costs of "Free":** Self-hosting open-source tools requires significant internal effort. The cost is not paid in licenses but in staff hours, time spent on integration, security patching, and maintaining "glue code" rather than supporting research.
- **The Value of Support:** A support contract is effectively an insurance policy. It guarantees access to specialized expertise when things break and ensures that critical infrastructure does not depend on the tacit knowledge of a few key employees. For many, the cost of a

support contract is lower than the cost of hiring the dedicated DevOps engineers required to maintain a home-grown stack.

Another common concern is vendor lock-in. In our experience, the fear of lock-in is often overstated, especially when weighed against the benefits of operational simplicity. Every choice, even an open-source one, introduces some level of dependency (lock-in to a specific technology or staff skill set).

When a vendor solution offers predictable delivery, integrated compliance controls, and reduced internal overhead, the trade-off is often highly favorable. It allows the university to trade "theoretical flexibility" for "practical capability," delivering reliable AI resources to researchers today rather than a perfect architecture tomorrow.

However, vertical integration is not a silver bullet. These platforms prioritize stability over bleeding-edge flexibility. For researchers needing to modify the underlying OS, swap out schedulers, or test pre-release hardware, an enterprise appliance can feel like a straitjacket. Additionally, the higher upfront transparency of costs comes with less elasticity; you pay for the premium experience even for low-priority workloads that could have run cheaply on commodity hardware.

Key Takeaways

The roadmap illustrates that technical progression goes hand in hand with organizational transformation. For IT departments navigating this shift, five principles stand out:

- **Momentum beats perfection:** Do not wait for a perfect architectural master plan. Immediate access to simple GPU resources builds trust and secures the "quick wins" necessary to justify larger investments later.
- **Align education with industry practice:** The path forward is built on containers. By adopting industry standards, the university "surfs" on global innovation rather than fighting against it. This ensures that the tools researchers use are robust, portable, and aligned with the skills demanded by the labor market.
- **Hybrid is the default:** Acknowledging that workloads will span both on-premises and cloud environments allows for a flexible architecture. This approach handles the academic fluctuation between steady-state research and massive, grant-funded spikes without financial waste.
- **Don't reinvent the wheel:** The market for AI infrastructure is mature. Leveraging existing enterprise frameworks and academic discounts is almost always faster and more sustainable than maintaining custom, home-grown integration code.
- **Don't overcentralize:** Facilitating the "medium" means providing frameworks, kubernetes clusters and virtualization environments where the users are free to operate with a high degree of freedom. This will balance efficiency with the freedom that research and education requires.
- **Culture is the bottleneck:** The best infrastructure is useless if researchers cannot easily adopt it. Investment in training, "champions," and user support is not an optional add-on, it is the primary vehicle for ROI.

Conclusion

Building an AI-ready university is a balance of urgent needs and long-term vision. The roadmap demonstrates that universities do not need to overhaul their entire environment overnight. By sequencing the journey, first providing acceleration, then modernizing platforms, and finally consolidating governance, institutions can achieve stability without stalling innovation.

Critically, this journey reveals that infrastructure is no longer a purely technical concern. Whether navigating the financial trade-offs of hybrid cloud or choosing between building vs. buying platforms, every IT decision directly impacts research competitiveness.

For the IT department, this represents a fundamental shift from operational support to strategic leadership. You are no longer just keeping the lights on; by removing technical friction and democratizing access to resources, you are providing the engine for the next generation of academic discovery.

Authors and acknowledgements

Main authors

Kim Henriksson, AI Sweden, kim@ai.se

Fredrik Viksten, Linköping University, Department of Electrical Engineering, fredrik.viksten@liu.se, and AI Sweden, fredrik.viksten@ai.se

Co-authors

Håkan Stensby, Linköping University IT, hakan.stensby@liu.se ,

Charles Morrall, Proact, charles.morrall@proact.se

John Magnusson IBM john.magnusson@se.ibm.com

Acknowledgements

Malmö University IT department: Staffan Krook staffan.krook@mau.se who initialized this exercise and Magnus Wikfors magnus.wikforss@mau.se for sharing their big picture and acknowledging the small, medium and large challenge.

Daniel Hemberg, ATEA, daniel.hemberg@atea.se, for walking us through the art of creating a white paper and being a sounding board.

The project “Data-driven organizations – Best practices for operationalization of AI in Sweden”

This material has been produced as part of the Vinnova-funded project *Data-driven organizations – Best practices for operationalization of AI in Sweden* (DDO), a project lasting just under two years with twenty participants from private sector, public sector, and academia. Together, they tackled issues concerning large- and small-scale operation of AI solutions and how to enable and use AI broadly across an organization.

The work focused on three specific use cases: Local sustainable operation of AI, legal and technical prerequisites for effective infrastructure, and how to create the best conditions for keeping thousands of AI models in operation.

A compilation of all material produced within the framework of DDO is available on [AI Sweden's website](#).

The organizations that participated in DDO were [Aixia](#), [Hewlett Packard Enterprise](#), [Hopworks](#), [IBM](#), [Linköping University](#), [NetApp](#), [Predi](#), [Proact](#), [RISE](#), [Red Hat](#), [Region Halland](#), [Sahlgrenska University Hospital](#), [Statistics Sweden](#) (Statistiska Centralbyrån), [The Swedish Tax Agency](#) (Skatteverket), [Stormgrid](#), [The Swedish Transport Administration](#) (Trafikverket), [Volvo Parts](#), [Region Västra Götaland](#), [Santa Anna](#), and [AI Sweden](#).

The project was funded by the participating organizations and Vinnova. AI Sweden is in part financed by the EU.