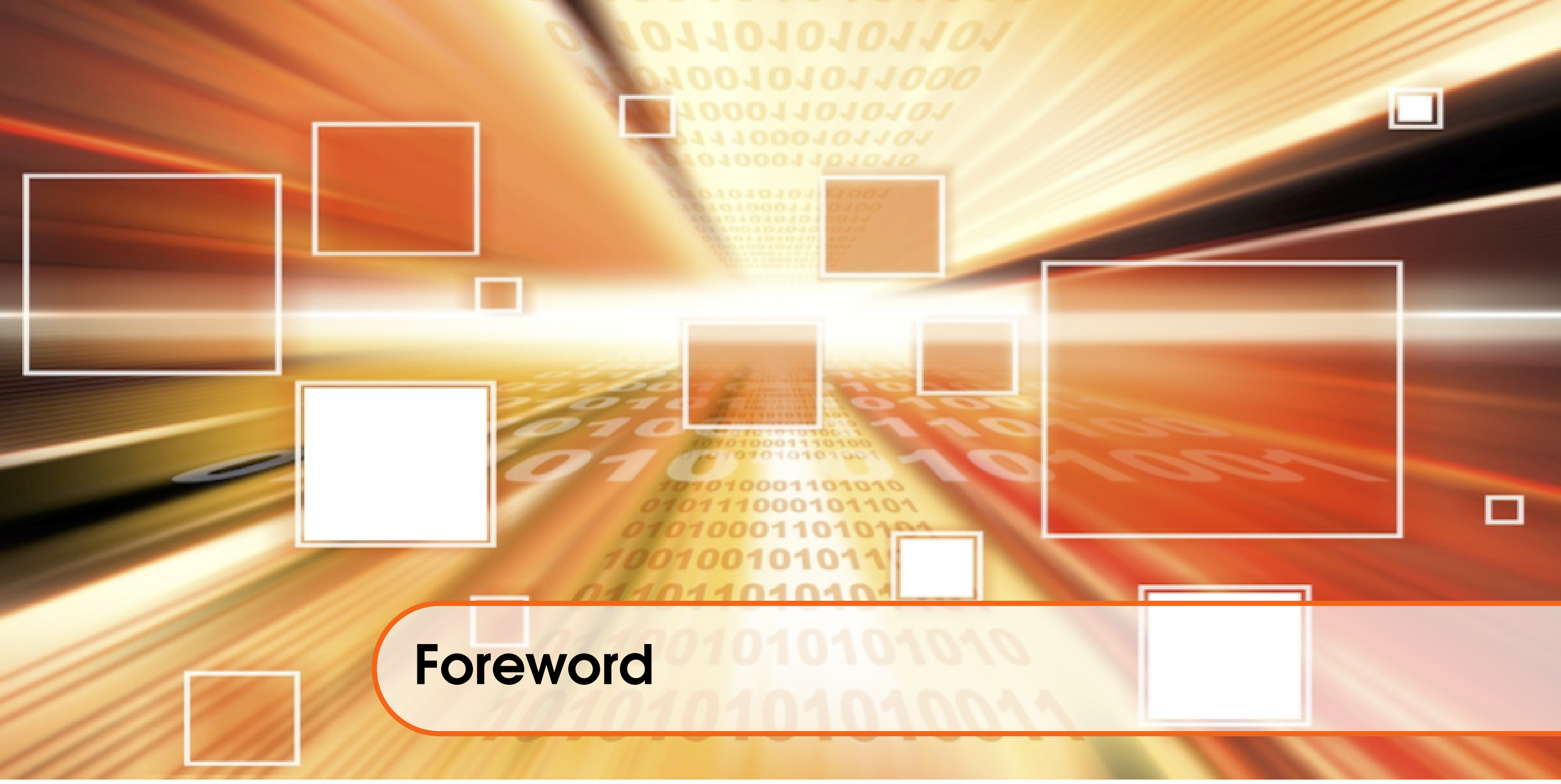
Swedish Infrastructure for AI

Comissioned by The Wallenberg AI, Autonomous Systems and Software Program & The Swedish National Infrastructure for Computing

Editor: Anders Ynnerman

Copyright © 2018-12-05 WASP & SNIC

Contents



Foreword

Today almost all industrial countries are investing heavily in the development of AI, and in particular, Deep Learning. In Sweden, government, industry, and academia all recognize the need to act quickly to strengthen our position in the field. The Knut and Alice Wallenberg Foundation (KAW) has donated an impressive 1 billion Swedish Crowns to increase AI development and use within the Wallenberg AI, Autonomous System and Software Program (WASP). The program will use the funding to support the best AI researchers in Sweden, but a substantial part will also be used for international recruitment of both young and established AI-researchers. Already today, Swedish AI researchers report that the limited availability of adequate computing infrastructure is hampering productivity. With the increasing volume of Swedish AI research over the coming years, resulting from the investments from KAW and others, adequate computing resources will be a crucial factor for success. Instead of forcing each research group or university to find the funding to build these resources, it is much more cost effective to build a common platform that can be used by all, preferably operated within SNIC, which has the expertise to organize the use of national HPC resources. It is therefore with great foresight that KAW has donated not only funding for research but also an additional 70 million Swedish Crowns for investment in computing resources dedicated to AI research. We are most grateful for this. I would also express gratitude to Professor Anders Ynnerman and colleagues for taking time out of their busy schedules to provide this report.

Professor Mille Millnert
Chairman SNIC & WASP

1 — Executive Summary

This report is comissioned by the Wallenberg AI, Autonomous Systems and Software Program (WASP) and the Swedish National Infrastructure for Computing (SNIC). The report forms the foundation for a planned investment in dedicated nationally available computational resources for research using AI methodology, such as Machine Learning.

The report gives an overview of the evolving hardware and software landscape for AI/ML, analyses the scientific workflow and identifies the stages in which large-scale resources are needed. The core of the report is an account for eight selected academic uses cases, each outlining challenges, potential breakthroughs and technical requirements. The report also presents industrial use cases, which are used to emphasize differences and similarities between industrial and academic workflows. Based on the analysis of use cases, demands, and future workflows, 10 major findings are presented:

1. GPUs are the current processing engine of choice in AI work flows, but in the longer perspective other more efficient technologies may be on the horizon.
2. Investment in processing cycles needs to be augmented with a tightly coupled peripheral infrastructure, containing elements such as hierarchical storage (memory, SSD and disk), interconnection infrastructure, and network connectivity.
3. Internationally, substantial investments in infrastructure for AI are currently taking place in many countries.
4. HPC and AI have different access and resource provisioning methods. Convergence towards a Cloud-like delivery model is desired.
5. As the use of AI approaches is emerging in many new fields, advanced user support and training is of high priority to make efficient use of the infrastructure.
6. Access to large scale GPU clusters is an essential and costly stage in the AI workflow, but it also contains several other stages with varying needs for resources, data and software.
7. Efficient, robust and secure data handling is mission critical in most applications.
8. There are currently islands of advanced AI users in the Swedish academic community, and a challenging large number of rapidly emerging applications spanning across the full range of disciplines.

9. The AI software landscape is changing rapidly with a continuous flow of new tools and updated technologies.
10. Users should be presented with one national entry point regardless of the organization of the physical infrastructure.

The overall main conclusion from analysis of the needs expressed in the uses cases is that:

> *There is an urgent and rapidly growing need for computational resources supporting AI workflows. Unless investments are made rapidly and with long term plans for continuous build-up, the competitiveness of Swedish research in AI/ML is jeopardized.*

The reference group for the investigation performed an analysis of the findings and possible modes of provision of resources, such as AIaaS (AI as a Service), infrastructure grants to individual research groups, a coordinated national investment, and possible combinations thereof. The conclusion is summarized in the following recommendations:

- Investment in one large scale GPU-based, state-of-the-art system, which is integrated into the Swedish National Infrastructure for Computing.
- Investment in tightly integrated, hierarchical storage solutions to provide extreme bandwidth access to data and leverage such existing resources in SNIC.
- As operational costs are significant they must be included or covered by other sources.

To ensure the provision of high quality services to the community it is furthermore necessary that:

- A new delivery model based on convergence of HPC and Cloud paradigms is developed.
- Advanced user support for effective use of the AI resource is established.
- Outreach and training efforts are initiated at the national level.
- Coordination with Swedish industrial resources should take place and industrial collaborations must be supported.

The reference group is very grateful for the many contributions from the vibrant and emerging AI/ML community towards this report. The AI infrastructure initiative is very timely and will promote Swedish research as well as make it possible for Sweden to align with the international build-up of infrastructures in this strategic domain.

2 — Preliminaries

2.1 Motivation

This report takes its starting point in the rapid, on-going development of Artificial Intelligence and in particular Machine Learning (ML) approaches, and their deployment in essentially all sectors of society. The Wallenberg AI, Autonomous Systems and Software Program (WASP) has launched strategic efforts to promote both research on AI itself and the application of AI based approaches in the WASP domains and beyond. At the same time other areas of research are aggressively, and with the support of funding agencies, building up research based on deployment of ML, and the use of ML is thus spreading through many areas of academia, industry and society.

State-of-the-art research, development and use of ML require access to a range of computational and storage resources. In view of this need, WASP and the Swedish National Infrastructure for Computing (SNIC) have expressed the intention to support provision of AI/ML resources to the Swedish academic community. This report has been commissioned with the objective of providing guidance on how to best invest in, and provide such support.

2.2 Mandate and Scope

The report should contain a clear recommendation for potential investment in hardware during 2019, as well as needed investment in operations and support. The analysis and scope of the report should cover present needs and estimated development until 2025. The report should only deal with nationally available infrastructure and services, and thus does not include specific needs expressed by research groups. Nor does it cover general support for use of AIaaS[1] resources. The report should focus on academic needs, but should also address potential collaborations with non-academic partners. The report should be presented to the WASP and SNIC boards no later than November 2018.

[1] AI-as-a-Service

2.3 Method

A reference group with representatives from academic users and developers, HPC centres, and industrial partners was appointed to lead the investigation. The group has the following members:

- Prof. Anders Ynnerman, Linköping University, Chairman
- Dr. Johan Eker, Lund University and Ericsson AB
- Prof. Erik Elmroth, Umeå University
- Prof. Michael Felsberg, Linköping University
- Prof. Matts Karlsson, Linköping University
- Prof. Erwin Laure, Royal Institute of Technology

The chairman of the reference group carries the overall responsibility for the investigation and is the main editor of the report. The reference group also includes advisory panels covering the following topics:

- Hardware and software solutions for ML
- International development
- Swedish leading edge research in need of ML resources
- Industrial needs and collaborations
- User support and training

The panels presented their initial findings at an open hearing held in Stockholm on June 1^{st} 2018. The results from the hearing formed the basis for the list of findings presented in the report. The chairman of the reference group is the main author of those sections of the report not provided by the specific panels.

3 — Introduction

3.1 Hardware and Software Solutions

The hardware and software landscape for AI is developing rapidly, with the current situation being significantly improved over the past five years through development of such approaches as accelerator technology and deep learning (DL) software. This rapid development is expected to continue, possibly driven by other software trends, making any long-term predictions rather uncertain.

3.1.1 AI in the Scientific Workflow

AI methods and software are not applied in isolation but are part of larger scientific or industrial workflows involving pre-and post-processing, data management, and other simulation or analysis steps. Figure 3.1 (taken from [1]) shows the typical ecosystem in which AI methods (in this specific case a machine learning system) are embedded.

This ecosystem is not specific to AI applications but applies to any (HPC) workflow, although there might be important differences in what is expected from certain components. This is particularly true when it comes to service models and access to resources. While computing

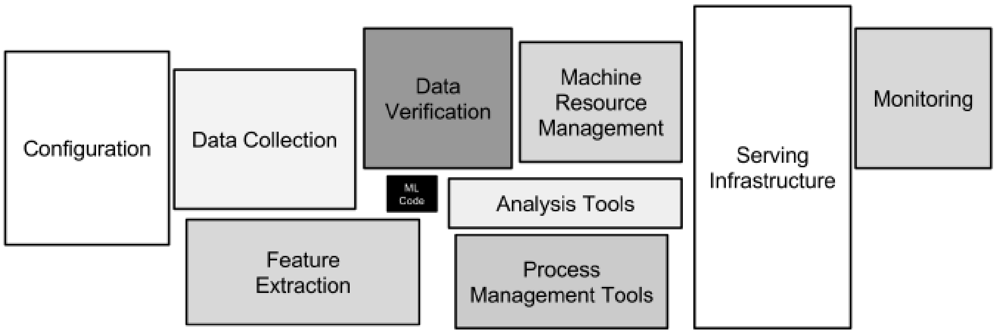


Figure 3.1: An illustrative AI ecosystem.

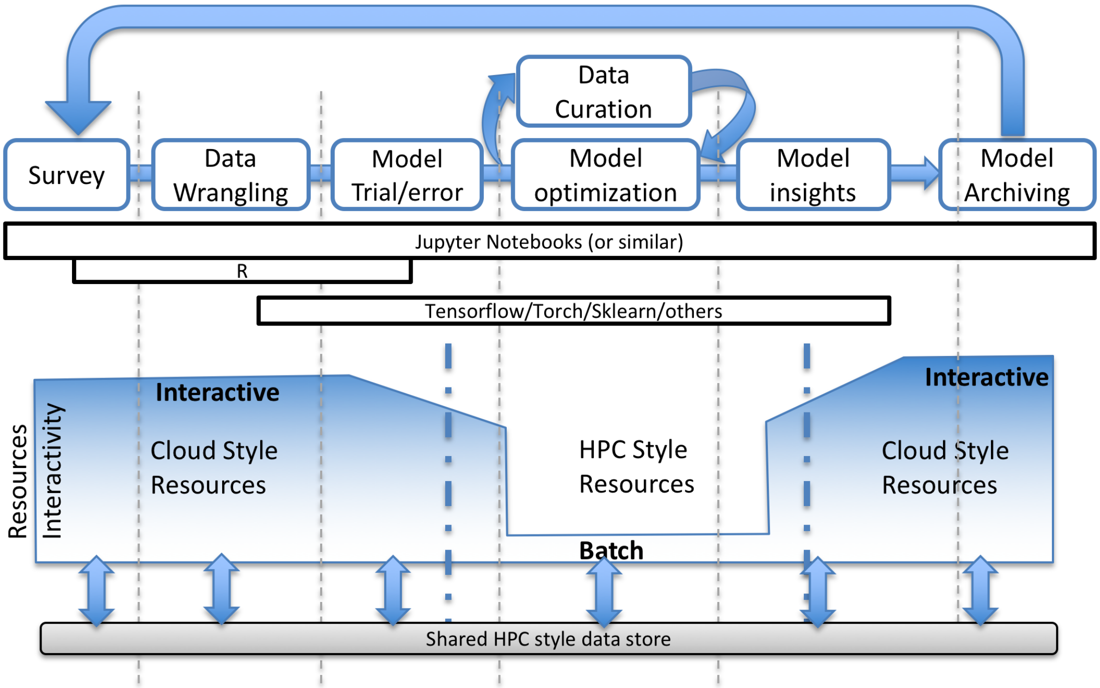


Figure 3.2: An illustrative AI workflow.

resources are delivered to classical HPC applications in a rather static, batch-like model, AI applications are also in need of more interactive modes, including dynamic resource allocation and streaming of data. As a consequence several service models are required to support these different use cases. Ideally, such systems should ease the uploading of data, downloading of results, and job submission and application control.

Another important component is data handling, since AI is dependent on access to large amounts of data. Many aspects of data handling, including fast data access and efficient large-scale storage are essential, but aspects such as data quality assurance, data provenance, format conversion, and data privacy are also important components of the AI ecosystem, as they can be for other applications. Following the principles of Open Data and of the European Open Science Cloud (EOSC), an AI infrastructure needs to provide appropriate data management subsystems, following the FAIR principles: Findable, Accessible, Interoperable, and Reusable.

3.1.2 AI Workflows

There are different kinds of workflows for AI problems, but typically for machine learning there are large datasets and the challenge is to create a system to learn a pattern present in one dataset (the training set) and later recognize the presence of that pattern in other, related datasets.

To solve problems of this type, data and AI scientists go through a workflow that requires different techniques, software, and hardware at each stage, as illustrated in Figure 3.2. Different problems will have specific workflows, but in general the workflow may be: intention definition (survey), data wrangling, model trial/error, model optimization, and model insights and archiving.

The stage of *Intention definition and survey* includes formulating a hypothesis that certain generalizable information can be built from the dataset. For example to detect the presence of a cat in a photograph, from a set of photographs that have been human-tagged with “contains cat"/"does not contain cat". At this point a survey is required to determine the existence of models that solve this or similar problems. Those "similar problem models" are key: most new AI problems are solved by modifying an existing model that is sufficiently close in nature. For example, it is possible that a deep learning model to detect dogs in photos can be altered slightly to detect cats. Among the requirements for this is access to a library of models for past AI problems to support the creation of new ones. Such a library must include model descriptions, training datasets, and usable code. Users must subsequently be able to add new models and datasets to the library.

In the *Data wrangling* stage the scientists process the data to increase the capacity of the learning systems or to satisfy particular requirements, for example by filtering, averaging, withering, anonymization, or any other kind of data processing step. As part of the workflow management it is important that datasets are handled using proper version control techniques in order to allow reproducibility. The requirements for this phase include the capacity to upload and store the initial dataset, as well as any intermediate versions of the dataset together with metadata that describes its changes (provenance). The requirements also include the need for interactive computing resources of intermediate scale to test data transformations and run quick experiments to evaluate models. The final large dataset transformations require larger-scale resources, ideally accessible via a batch system.

At the *Model trial* stage the scientist defines a set of models and runs training processes over the dataset to identify which one best solves the problem. Users usually perform small training tests with quick adjustments and re-runs before running larger training processes. The requirements during this phase include both interactive resources of intermediate capacity for the quick model trials and more powerful batch resources for the training processes to select the best models for the next stage.

The *Model optimization and final training* stage is usually performed for models that are parameterized for, for example, deep learning structure, learning rate, and model constant. The space of parameter values is explored by running multiple training processes on the dataset to determine the best settings for the model. Powerful batch compute resources are required to perform multiple training steps in parallel, each with with different parameters. In addition, orchestration tools are required to setup and perform the multiple training steps and to automate model selection by analyzing training results. The system requirements include low latency synchronous interconnects to optimize parallel synchronous learning.

In the final stage, *Model insights and archiving*, models are often analyzed to understand why they perform as they do. This usually requires visualization tools to analyze intermediate data steps in the model and is fundamental to provide knowledge that will help to use all or part of the model for other AI problems. This brings requirements of interactive resources of intermediate capacity to run the model as well as mechanisms to store the model, training dataset, and insights into the model, for future use. [2]

3.1.3 The Evolving Software Landscape

The software landscape includes both software for different particular AI methods and for use at different stages of the workflow. Before detailing this landscape it should be pointed out that this is a software landscape subject to very rapid development with new software packages, providing new features or better performance, quickly replacing others. Below, we characterize software based on functionality and illustrate with some currently available packages:

Software for *General Machine Learning*, including software mainly for supervised and unsupervised learning, methods for data analysis, experimental setup, and validation. Examples include SciKit-learn, Spark MLLib, and H20.ai.

Software for *Deep Learning*, including specific tools for creating neural net architectures and possessing capabilities for hardware acceleration and distributed computation. Software is typically general purpose but has many applications in language and speech recognition, and in computer vision. Examples include Tensorflow, Keras, Torch, Microsoft Cognitive Toolkit, and Theano.

Software for *Reinforcement Learning* is software aimed to act in a particular environment and is often simulator-based. Examples include OpenAI Gym, DeepMind Lab, and Berkeley RISE Lab stack.

Software for *classical AI including Reasoning and Planning* is often designed in the form of libraries, and targets problems such as semantic reasoning and planning. Examples include SOAR, Metric-FF, and CST.

Programming Tools and Management Software is software for setting up and performing computations, for example programming environments and tools for managing projects, data, and models. Many of the applications of AI and machine learning will need massive amounts of training data. Storing this data and calculating features from them is a big data problem and thus the system should have access to very large-scale storage and the capabilities to run big data platforms such as Spark and Hadoop. Examples of other tools and management software include Jupyter, GitLab, and Hops.

In addition to the packages described above, the research community uses a lot of *specialized software* tailored to small user groups and with functionality very much dependent on research area and interests.

3.1.4 The Evolving Hardware Landscape

The computing hardware used for AI is a mix of general-purpose CPUs and GPUs and specialized hardware such as FPGAs and accelerators, such as ASICs/TPUs, as they are well-suited for different parts of the workflow. The CPUs spend most of their area on deep caches, making them ideal for providing low latency data access for applications with random or non-uniform memory accesses. GPUs, on the other hand, are optimized for compute intensive workloads and streaming memory models and so provide high throughput. GPUs are sometimes claimed to have much higher memory bandwidth, by as much as a factor of ten, but this comparison very much depends on what memory level comparison is under consideration. Regarding specialized hardware, the FPGAs provide programmable logic whereas the different accelerators are fixed.

In practice, GPUs and accelerators are primarily used for large-scale deep learning (and other machine learning) processing whereas CPUs are the primary choice for most other parts of the workflow. Notably, sometimes the choice of CPUs is based on a lack of libraries supporting GPUs and accelerators, which may be the case, for example, for researchers working on their own machine learning codes.

In addition to computational resources, the storage infrastructure is important to allow for the storage and high-speed delivery of large amounts of data, which can only be expected to increase in both capacity and demand over time.

3.2 International Outlook

AI and its related IT infrastructure is currently a hot topic internationally, with many countries developing plans to provide dedicated support for AI, on both the research and infrastructure sides. This includes significant investment in AI infrastructure, together with programs to improve the system and application software. A recent trend is to see AI infrastructure not as a separate entity but as an integrated component of the main national HPC infrastructure. Thus full scientific or industrial workflows, combining AI methods with other simulation or analysis components, is facilitated. In addition, synergies within the IT ecosystem (processing, memory, storage and networking) can be utilized.

In this section, we summarize some of the major international developments.

At the end of March 2018 the French President presented his vision to make **France** a world leader in AI [3] based on the extensive report compiled by Cèdric Villani [4]. This report addresses various issues from education/training, research, infrastructure, access to open databases, fairness and ethics. Among these recommendations a major one is related to the provision of AI infrastructure for the French AI research community. This recommendation is currently being addressed by the French National Infrastructure for Supercomputing (GENCI) by means of systems with converged HPC/AI nodes, with additional nodes specific to AI. While access modes to AI nodes will be different from those used for classical HPC, one topic of interest for GENCI is the convergence between HPC and AI, when HPC needs AI (such as for in situ post processing of data, smart computational steering, fast convergence of numerical methods, or for maintenance or smart scheduling of resources) and when AI needs HPC (such as for scaling up (new) AI methods using massive datasets).

A particular focus is put on supporting the needs of the French AI community (in addition to the HPC one) by supporting a large number of tools and environments, rapid provision of such environments, flexibility and ease of use (providing similar experiences as using cloud resources, portals, GUIs), containerization, security, and the inclusion of elastic access modes in the existing peer-review based allocation mechanism. Currently, these joint systems are used by some 40% by the French AI community; research and industry alike.

A similar approach is being followed in **Finland** where the national supercomputing centre, CSC, is planning a computational platform that combines traditional HPC with data intensive computing and AI based on requirements put forward in "*The Scientific Case and User Requirements for High-Performance and Data-Intensive Computing in Finland 2017–2021*" [5]. Finland is investing over 270 MSEK in renewal of their national HPC infrastructure, of which some 50–100 MSEK will be invested in the converged HPC/AI platform, excluding storage.

The **UK** has recently published a policy paper entitled "*AI Sector Deal*" [6] that is expected to influence future e-infrastructure roadmaps towards AI.

The GCS partners in **Germany** operate a variety of different evaluation systems and the Jülich Supercomputing Centre (JSC) has plans to install a major system optimized for data-intense science and data-analytics applications, such as non-volatile memory (NVM) at node level and high-performance storage tiers, by 2020.

Many other countries, including Norway, Ireland and the Netherlands, are putting in place smaller systems to gain experience and build up competence in this area.

In the Nordic countries a letter of intent has been submitted to the Nordic e-Infrastructure Collaboration (NeIC) for increased collaboration and exchange of knowledge within the region.

Outside of Europe, the major **US** (e.g. the currently most powerful supercomputer in the world, *Summit*) and **Japanese** systems (*Post-K*) have been designed to support both, classical HPC and AI applications, although using different technologies. As part of their Exascale initiative the US have also recently launched a *Machine Learning Center* that aims at the development of highly scalable and efficient ML software for the Exascale[1]. This centre receives some 4 MUSD per year.

References

[1] D Scully, G. Holt, D. Golovin, et al. “Hidden Technical Debt in Machine Learning Systems”. In: *NIPS’15 Proceedings of the 28th International Conference on Neural Information Processing Systems* (2015) (cited on page 11).

[2] J Wood et al. “Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data-European Commission”. In: *European Union* (2010) (cited on page 13).

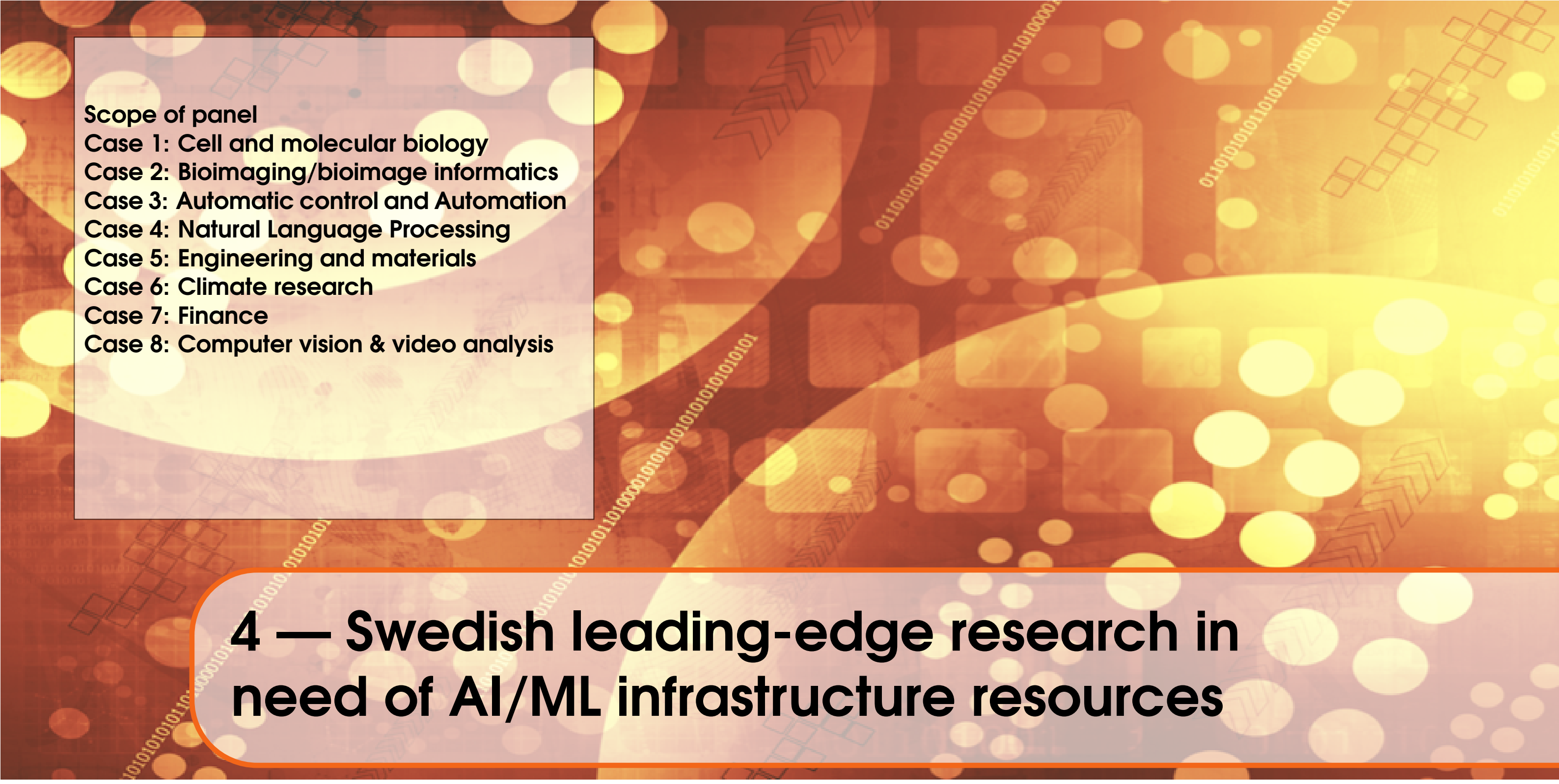
[3] URL: `https://www.aiforhumanity.fr/en/` (cited on page 15).

[4] URL: `https://www.aiforhumanity.fr/pdfs/MissionVillani%5C_Report%5C_ENG-VF.pdf` (cited on page 15).

[5] URL: `https://www.csc.fi/-/sciencecase2017-2021` (cited on page 15).

[6] URL: `https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal` (cited on page 15).

[1] https://www.exascaleproject.org/ecp-announces-new-co-design-center-to-focus-on-exascale-machine-learning-technologies/



4 — Swedish leading-edge research in need of AI/ML infrastructure resources

4.1 Scope of panel

This chapter presents the findings of the panel with respect to the mission to identify and present use cases for a future AI/ML infrastructure. The selection of areas was made to represent potential user groups and to highlight the present and future needs in a variety of disciplines and different usage patterns. The cases do not cover all the current and potential future users and they should not be seen as any guidance to which disciplines and cases are of high priority, but can serve as clear examples giving insights into the potential needs and the impact that might be generated. The following cases were selected:

1. *Cell and molecular biology*
2. *Bioimaging/bioimage informatics*
3. *Automatic control and Automation*
4. *Natural Language Processing*
5. *Engineering and Materials*
6. *Climate Research*
7. *Finance*
8. *Computer vision & video analysis*

For each of these cases the panel was asked to delineate the scientific challenges and their potential impact in terms of specific breakthroughs that could be made using AI/ML. An estimate of required resources to execute the research program, as well as the corresponding technical challenges, concludes each case. It should be noted that the scope and level of details varies between the cases, reflecting the varying degree of experiences of AI/ML. In some cases the case is based on on-going research projects using significant resources, whereas other cases are based on extrapolated needs for future projects.

4.2 Case 1: Cell and molecular biology

Cell biology (cytology) and molecular biology are branches of biology that study the structure and function of the cell, and the molecular basis of biological activity in the systems of a cell, respectively. Both branches are fundamental to the understanding of how cells work and thus form the basis to many other areas, such as research into cancer.

4.2.1 Scientific challenges and potential impact

In structural biology studies different modalities, such as cryo-electron microscopy (cryo-EM), X-ray crystallography and flash X-ray imaging, generate tremendous amounts of data with megapixel images generated at kHz rates at the most advanced facilities. Since these images reflect the scattering of individual photons or electrons, with each image possibly reflecting a unique sample, traditional image and video compression approaches are applicable to only a limited extent. How to compress and make sense of these data flows is an area of active interdisciplinary research.

So far, processing approaches have generally focused on modelling individual shots, and only then compiling conclusions. The interest in time-resolved studies, as well as the direct modelling of the heterogeneity of samples, is increasing, however, with a commensurate increase in sizes of the required data. Since proteins are frequently molecular machines, understanding what parts can move and what parts are flexible require a holistic modelling of a full dataset, especially when the signal to noise ratio for individual shots is very low. In such conditions one would want to move beyond the current traditional modelling based on a simplified view of the physical processes, and benefit from advances in machine learning.

During the development of these techniques, repeated training (not only inference), running over hours of recorded data, can be necessary. Even with some amount of prefiltering the data volumes can be vast. Beyond what is common in any image-processing Deep Learning (DL) application, the physics of photon diffraction also suggests that Fourier transform layers should be included in the nets. The global nature of the Fourier operator introduces some challenges regarding the efficient implementation of these models.

The complete understanding of the energy landscape of a protein, as characterized by the probability of observing it in different shapes in a flash X-ray imaging dataset, would be directly applicable in pharmaceutical applications since most drugs work by binding to proteins with high specificity, inhibiting or enhancing their function. Currently, structural models from experiments are static and the only way to understand the mechanics is by molecular dynamics simulations.

A desirable outcome would be to create a hybrid learning model, that unifies an ML framework with experimental structure data.

Possible breakthroughs on the horizon which might be enabled by use and development of DL/ML in this domain are:

- **Atomic resolution 3D reconstruction of a membrane protein of unknown structure** (the 2012 Nobel prize in Chemistry was awarded for the elucidation of the structure of one family of challenging proteins using current techniques).
- **An experimentally founded description of the 4D mechanics** (3D + reaction dynamics) of a central macro-molecular machine in the human cell, such as the ribosome or RNA polymerase II. Previous work on the static ribosome structure was awarded the Nobel prize in 2009.

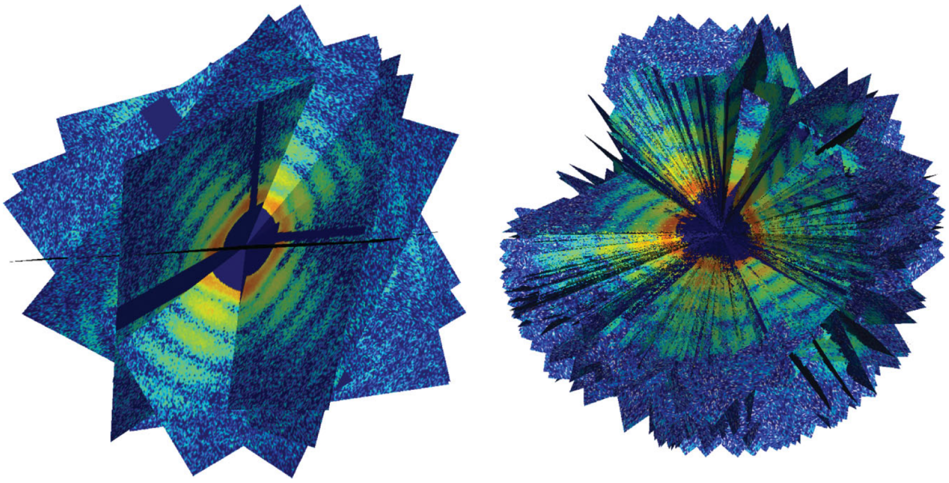


Figure 4.1: 3D reconstruction of a protein. Left: a few separate collected 2D slices merged into a 3D volume; right: a few hundred slices, creating an (almost) dense volume.

- **Real-time feedback, including tentative reconstructions**, while data collection is ongoing. Beam time at these facilities is extremely precious and competitive (for example 60 hours per year can be awarded to a whole consortium), so immediate feedback on the quality of data in order to decide whether to improve the experimental conditions or move on to the next sample, is crucial.

4.2.2 Technical challenges and infrastructure requirements

In order to address the above mentioned challenges and achieve the full impact of the field, specific requirements need to be addressed by local and national infrastructure. Data of the order of hundreds of terabytes generated at local, national, or international infrastructure centres need to be streamed to computational nodes and this process needs to be repeated hundreds of times when training new models. The data itself also needs to be archived for at least 15 years for the purpose of reproducibility of results.

Thus cell and molecular biology presents enormous challenges for the bandwidth between local nodes and national centres as well as in the storage capacity of local infrastructure. High bandwidth backbones are required, which connect directly to scalable computational resources, such that incoming data need not be buffered. This is especially true for real-time use cases. Computations are typically SIMD (Single Instruction Multiple Data), but with a high level of dependency, thus requiring joint, wide bandwidth internal buses.

Among existing alternatives for computational infrastructure, classical CPU-based nodes suffer from a need for heavy inter-node communication whereas their flexibility in control structures remains unexploited. In contrast, GPU-based nodes are quite flexible regarding control structures and particularly with regard to shared memory resources. However, their availability is limited, both in gross numbers and also with respect to on-demand scalability. The inherent physics of diffraction experiments necessitates whole-image Fourier transforms. These are efficient on GPUs but some other forms of suggested deep learning accelerators, those with a focus on the convolutional operations used in more traditional image processing deep learning, might not be able to handle this task. The extreme dynamic range in the recorded signal also makes the issue of double/mixed/half precision models, in this specific area, something in need of thorough investigation.

4.3 Case 2: Bioimaging/bioimage informatics

Automation of microscopy, including sample handling and microscope control, has enabled the rapid collection of digital image data from biomedical samples, transforming imaging cytometry into one of the most data-rich scientific disciplines. Bioimage informatics aims at the development of methods for efficient analysis of such biomedical data. Application areas include digital pathology and large-scale image-based drug screening, as well as advanced ultrastructural imaging such as cryo-EM imaging and reconstruction, also requiring development of novel computational approaches to go from image data to scientific discoveries.

Development of such computational approaches is an important part of the activities at the national Science for Life Laboratory (SciLifeLab), and within SciLifeLab, the BioImageInformatics facility provides support for image-based research `https://www.scilifelab.se/facilities/bioimage-informatics/`.

4.3.1 Scientific challenges and potential impact

Both in traditional compute-intensive fields and new AI-focused ones it is obvious that Deep Learning (DL) technologies will continue to rise in importance. There is some lower-hanging fruit where researchers simply use ML to interpret or classify results of calculations, but we also see more and more examples where DL is replacing traditional compute techniques such as quantum chemistry, raytracing, and optimization – and this too can be expected to grow rapidly.

As part of the human protein atlas (`https://www.proteinatlas.org/`) researchers within SciLifeLab are working with AI in a range of microscopy data applications including cryo-EM imaging and investigating methods to improve reconstruction both of single particles and tomograms, digital pathology, and investigation of protein localization.

In all these fields, the primary AI/ML research is on DL. The interest from the community for applying DL to large-scale image data is growing quickly, and it has been decided to prioritize internal technology development in this area, focusing on making novel techniques available to a broader community. *'Imaging' was voted the number one future area during the 2018 SciLifeLab summit in April 2018.* The interest from industry is also growing. Examples include the use of DL for image-based large-scale drug discovery projects in collaboration with AstraZeneca (funded by SSF) and evaluation of antibiotic effects in collaboration with the SMEs Q-linea and Astrego diagnostics.

Possible breakthroughs on the horizon enabled by use and development of DL/ML in this domain are:

- deep learning to progress **digital pathology** and provide guides for pathologists,
- learning approaches to utilize the full potential of **novel microscopy imaging modalities**, and
- **learning of complex models** for image-based basic research such as the Human Protein Atlas as well as industry-relevant drug-screening efforts.

4.3.2 Technical challenges and infrastructure requirements

Locally, GPU clusters with Nvidia Titan X Pascal GPUs (12Gb memory) are being used. As a complement, the Analytic Imaging Diagnostics Arena, AIDA, at CMIV in Linköping (`https://medtech4health.se/aida/projekt/`) is being used. In the absence of suitable national GPU clusters, 'Narvi' in Tampere, Finland, is being used for a large scale (>700 patients,

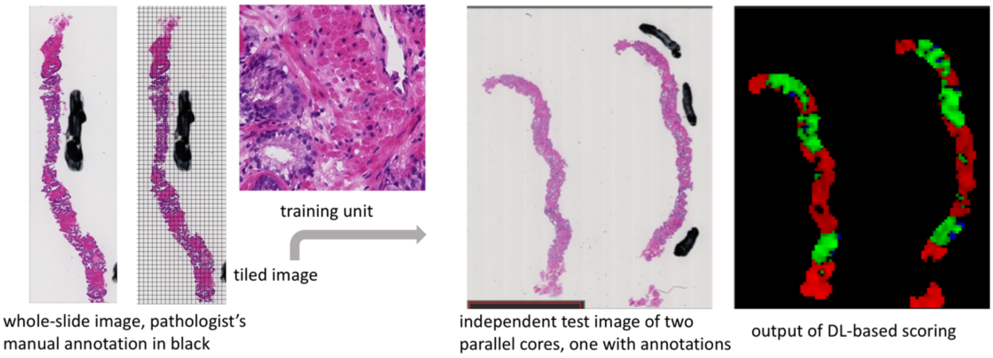

training unit
tiled image
whole-slide image, pathologist's manual annotation in black
independent test image of two parallel cores, one with annotations
output of DL-based scoring

Figure 4.2: Prostate cancer grading (an ongoing collaboration between SciLifeLab Uppsala University and the Karolinska Institute). Working with a collection of 80,000 core needle biopsies with clinical data on patient level and approximate cancer regions manually marked at pixel-level in 6000 samples, scoring based on deep learning is approaching 99% accuracy.

80 000 biopsies) digital pathology/prostate cancer project. Here, disk bandwidth is a major problem. Each GPU node has a local 540 GB SSD for temporary data. Copying the dataset (750 GB) from the slow shared disk to the node's local SSD before the run and deleting it afterwards is a substantial bottleneck during training. A further resource, although limited to projects involving AstraZeneca, is their infrastructure in Sköndal.

In industry the trend is towards specialized hardware, such as Google's TPUs, and the challenges there is likely to be focused on highly energy-efficient inference hardware. It is unlikely that any users will be happy with a national infrastructure for that. Rather than focusing on "one infrastructure to rule them all", the important question is what parts might not be served well by the current infrastructure. For many users the greatest challenge is simply that they need a lot of scripts, new development and storage – and that can be handled quite well on current SNIC resources.

The part that is not served well is the work focused on Deep Learning training. Presently (and within the next 5 years) there is simply no competition to Nvidia for general purpose usage of this type, partly because of the number of tools that are now written to use CUDA and partly because of very efficient tensor units. The other key difference compared with traditional hardware is that these processors have very high memory bandwidth, but this is limited to the high-end professional cards, not the cheaper consumer devices.

The required setup is a smaller copy of a machine like Nvidia's in-house "Saturn V": a group of nodes like DGX-1 or DGX-2. A handful of such nodes are currently many times more powerful than the largest SNIC machines for AI. They should be scheduled entirely with container-systems such as docker and kubernetes – both because that is where AI approaches are going, and because it will train all the users in the tools that currently dominate industry. Similarly, storage should likely be handled with object storage and/or volumes attached dynamically to containers.

Envisaged infrastructure: GPU computer clusters, provided via SNIC, with good bandwidth and close-to-cluster temporary data storage would be highly valuable. SciLifeLab has also identified such infrastructure as a prioritized need for the future development.

4.4 Case 3: Automatic control and Automation

Automatic control is sometimes referred to as the the hidden technology, since it is everywhere around without us ever noticing it. For example, the cruise control in a car or the heating system in a building. Automatic control is about making physical systems behave in a desired way. Traditionally, automatic control has been model-based rather than data driven, but machine learning is becoming an increasingly important ingredient. In particular, in areas related to automation and autonomous systems.

4.4.1 Scientific challenges and potential impact

Automatic control and machine learning have many topics in common. One example is Partially-Observable Markov Decision Processes (POMDP) which were first introduced in the control community. Another example is reinforcement learning that has strong connections both to dynamic programming, dual control, and iterative learning control. Reinforcement learning has been successfully deployed to robotics systems, where the system autonomously learns an optimal behaviour via trial-and-error. The control system designer provides an objective function that is used as feedback to a learning system. It has successfully been applied to a wide range of systems, such as walking robots and autonomous helicopters.

So far, most machine learning applications have been based on static models, while in most control applications dynamics (often physics based) plays an important role. There is a huge potential for future applications where machine learning is used for control of physics-based dynamical systems. The ability of deep neural networks to approximate nonlinear mappings with high-dimensional input spaces and their potential for offline and online learning makes them interesting candidates for use in, for example, model-based control as well as in adaptive control where neural network models can be used as inverse models. The connection between adaptive control and machine learning is currently being re-explored, driven partly by the abundance of data and computing resources.

Adaptive control systems are designed with the ability to adjust to changes in the operating conditions. An outer loop adapts the inner control loop by learning the environment and updating the process model. Artificial neural networks have been used to guide the outer loop in selecting different initial models by classification of the physical process. This can be an important step towards more autonomous systems.

Machine learning has successfully been applied in the process industry for data mining and analytics. Extraction of information from process data, and transferring it to effective knowledge for decision support. To effectively carry this out, machine learning algorithms have long played an important role. Smart manufacturing and factory automation are related domains where machine learning currently is making major inroads. Predictive maintenance systems detect forthcoming faults and schedule repairs to prevent unwanted and costly downtime. Similarly, anomaly detection can be applied to monitor the state of a manufacturing site to reduces operational cost and improve reliability.

Product life cycle management based on machine learning analyses data from all the different production steps, as well as the end-user behavior to improve quality throughout the complete life span of the product series.

Machine learning has been applied to control of very large scale and complex systems that are typically difficult to model, such as whole data centers or production plants. For example, several successful deployments of highly autonomous data centers have been reported with significant energy savings.

Autonomous vehicles is an area of much attention and promise. It has not only the potential of drastically decreasing the number of accidents (roughly 1.2 million fatalities each year), but also the potential of redefining the way we think about automotive transportation. However, the size of training sets is prohibitive and a challenge.

At its heart automatic control is model based, while machine learning is data driven. The former brings proofs and guarantees and the latter adds support for very large scale and highly complex systems of systems. Possible breakthroughs in the area of automation and control from access to large scale AI-infrastructure are:

- **Robust and intelligent control systems** that can be reasoned about by combining machine learning and classical control theory.
- Smart **manufacturing**, which will reduce operational cost and improve product quality. This has the potential to result in huge savings and improvements.
- Evolution of highly **autonomous systems**, such as self driving vehicles and self-governing data centers, which has the potential of saving lives and reducing energy consumption.
- Application of reinforcement learning for **robotics** systems has the potential to greatly simplify programming.
- Modeling and control of **high-dimensional non-linear systems**, which are unfeasible using traditional methods.

A large-scale infrastructure with support for AI would greatly benefit the development of control theory and control systems to address new applications areas and more complex systems.

4.4.2 Technical challenges and infrastructure requirements

- **Data** Lack of access to real world data sets and a secure data management systems is currently hindering progress in the development of ML system for large scale systems, e.g. smart manufacturing, autonomous cloud, etc. Data coming from production environments is typically sensitive and cannot be openly shared. A secure data storage system, with support for data lineage and data provenance is needed.
- **Service delivery** Typical tools for control engineers are Matlab, Modelica, Julia or Python. A service access model to machine learning facilities that can be combined with established workflows would greatly reduce barriers and promote uptake. Furthermore, many of the established software packets runs very efficiently on modern CPU with vector support (e.g. Intel AVX) and need not GPU support.
- **Compute** Traditional control system design would benefit from a large scale data parallel service allowing extensive simulations of parameterized models.

Computational infrastructure such as access to high performance data centers will be critical for research to address all the above challenges.

4.5 Case 4: Natural Language Processing

Natural Language Processing (NLP) is one of the high impact application areas of AI/machine learning (together with computer vision). It is a multi-disciplinary area drawing from both classical and computational linguistics. The goal is to extract semantic information automatically from text and one of the most important applications is automatic translation from one natural language to another (for example from English to Chinese). The field is, today, totally dominated by machine learning methods with all state of the art systems being based on ML. The results find application in everyday tools, from Google translate to dialogue systems in digital assistants such as Siri.

4.5.1 Scientific challenges and potential impact

A typical example of an NLP task is automatic machine translation from one natural language to another. An idea of the multiple challenges involved in this task can be appreciated by realizing how natural language utterances are often ambiguous and dependent on context for interpretation. State-of-the-art systems today, such as Google Translate, are based on Deep Learning and training the networks requires enormous amounts of training data, such as large parallel corpora of data such as Bibles in multiple languages or European Union documents. Fundamental NLP tasks that are subsystems within translation and other tasks include part-of-speech tagging (POS), parsing, semantic role labelling, disambiguation and sentiment classification. Other related NLP tasks which are becoming increasingly important include Question-Answering systems and even more general Dialogue Systems (see Google Duplex for an impressive example: `https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html`)

The predominant architecture used in these tasks is that of recurrent neural networks, such as Long Short Term Memory networks (LSTMs) and Gated recurrent units (GRUs), which typically are harder to train than Convolutional Neural Networks (CNNs) and are also increasingly used in computer vision. Recent advances have involved new architectures such as the Transformer network (`https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html`), which builds upon the Attention mechanism. Hyperparameter optimization is carried out with thousands of runs, evaluating many factors including different layer structures, cost functions, optimization methods, learning rates and initializations. Even for relatively simple networks training from scratch often takes days, even using using multiple GPUs, while fine-tuning then requires several additional hours. A recent trend in application development has been to make them multimodal: combining text with image and video data. These applications are even more computationally demanding.

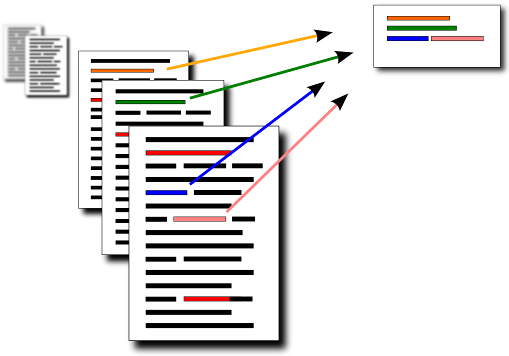


Figure 4.3: Automatic summarization of multiple discrete text documents.

The availability of computational resources is one of the major bottlenecks for further progress in research. Without suitable additional resources research on NLP advances will be seriously jeopardized. Commercially available resources are also not an option as the costs for access are, according to trials with Azure, of the order of a 50% additinal cost on top of each PhD student.

Successful NLP research will have a major positive effect on a wide variety of

applications going beyond classical ma-
chine translation systems. Natural language technologies are expected to become much more widespread over the coming years, and an infrastructure investment could lead to breakthroughs in commercially important technologies such as:

- **Information extraction systems** which convert the information expressed in text into a structured format.
- **Automatic summarization systems** which condense one or more large documents into a short summary containing the most important information.
- **Interactive spoken-language systems**, including dialogue systems and spoken-language interfaces. Siri and Alexa are only the very beginning of the possibilities.

4.5.2 Technical challenges and infrastructure requirements

Deep Learning for NLP and related applications requires primarily GPU resources rather than the CPU resources which have, historically, been the focus of SNIC. Although SNIC already provides some GPU resources these constitute less than 1% of the nodes in place. The demand is so much larger that, of the projects awarded in 2018, only some 20% of the needed resources can be provided) and so a shift in focus from CPU to GPU provision is called for.

Local resources are also necessary to support initial trials and to avoid waiting periods caused by the scheduling of shared resources. A major bottleneck for local systems is, again, the available GPU resources. Usually groups have a few GPU cards but server-grade hardware is usually not affordable for individual research groups without additional funding, such as through WASP.

Server hardware is more common at the national level, but the research described here requires interactive and exclusive access to GPU nodes for those resources. Interactive access is necessary in order to efficiently monitor and re-launch parameter trials and for debugging. Exclusive access is necessary due to the large amounts of data, which also require large scratch spaces of the order of terabytes.

4.6 Case 5: Engineering and materials

Engineering is a very broad topic, ranging from the prediction of small-scale phenomena like the propagation of cracks or the motion of turbulence in a mixing chamber, to very large complete systems such as power plants, bridges or ships. It is therefore difficult to provide a general view as to exactly what resources are needed. In the following, however, attempts to give some guidance as to future needs based on the example of fluid mechanics and turbulence, which on its own spans a very wide field including different application areas such as biomedical flows, atmospheric flows and flows in engineering devices).

4.6.1 Scientific challenges and potential impact

Computing has a long tradition in fluid mechanics, which goes back to the first steps by the meteorologist Richardson in the 1920s. Since then turbulent flow simulation, in particular, has proven to be one of the largest users of computer time on the world's largest clusters. This is related to the inherent nonlinearity of the governing Navier-Stokes equations, which tend to excite smaller and smaller scales as the so-called Reynolds number (comparable to the speed or the size of the problem) is increased. In the past, large-scale computations in fluid dynamics have mainly been used to reproduce experimental setups and consequently to calculate statistics of the turbulent flow. Once validated against experiments, these statistics could then be used to tune physics-based engineering models (for instance the coefficients for the widely-used eddy-viscosity-based k-epsilon turbulence model). Considerable effort has gone into the development of more complicated models, including wall function, large-eddy and detached-eddy simulation (LES and DES, respectively) techniques, although without a major breakthrough being made.

Similarly, during recent years, with the advent of the notion of so-called coherent structures, three- or even four-dimensional education techniques were developed, in which features in the turbulent flow were identified, and potentially tracked in time. This is done in an effort to better understand the nonlinear dynamics of the flow and thus to arrive at a lower-order description of the relevant ingredients that need to be modelled. At the same time optimization methods for flows (for example using shape optimization to reduce drag, or topology optimization to increase mixing) have been developed, mainly using classical adjoint-based gradient methods.

So far, machine learning and artificial intelligence has only rarely been used in fluid mechanics. Traditional calculations and postprocessing methods were based on physics-based assumptions, and thus required little case-by-case tuning. This is, however, gradually changing and ML-based approaches have been used, for example to train turbulence models for specific non-standard situations (such as separation prediction, which is notoriously difficult using eddy-visocity models), or to devise low-order models to predict simple recurrent motions in transitional and turbulent flows. At the same time it is believed that extracting coherent turbulence structures from vast training data will boost the understanding of the flow dynamics. This has so far been limited by the size of the underlying training data sets, for which millions of degrees of freedom need to be tracked at each time step.

Possible breakthroughs due to exploitation of ML in turbulence research could be:

- improvement of control techniques based on prior learning, for example to reduce drag or delay separation,
- prediction of wall models to speed up simulations,
- extraction of coherent structures and recurrent phenomena in turbulence,
- generation of realistic inflow conditions based on training data sets.

4.6.2 Technical challenges and infrastructure requirements

Traditional computations in fluid mechanics rely on supercomputing resources, mainly CPU based systems, but are gradually moving to exploit accelerators such as GPUs. Any large-scale computation in turbulence requires on the order of months to complete, and produces terabytes of data. The limiting factor has, historically, been the computing time, but the available memory and storage are becoming significant limitations, in particular if one considers generating a learning database. Most applications of ML in fluid mechanics are based on standard software frameworks, such as TensorFlow, and Python packages. Depending on the type of analysis, GPU-based hardware is advised, with dedicated high-speed storage for fast access to the data.

At the same time it should be mentioned that there is currently no group with expertise in fluid mechanics and ML in Sweden. There are a few projects being started right now, which will certainly lead to relevant results within a comparably short time.

4.7 Case 6: Climate research

Climate and weather research involves the analysis of large amounts of data that can be either observational or generated through simulations. These datasets are increasing in size incredibly rapidly as Earth-observing satellites become more plentiful and exploit higher quality measurement devices, and as climate models become more powerful. A high-resolution climate model can easily produce petabytes of data and so new techniques to efficiently extract key features from this wealth of data are needed.

4.7.1 Scientific challenges and potential impact

Observational data from remote sensing from space or ground contains a wealth of information and new ways of analyzing them can reveal more details of the various phenomena which can be identified. For example the three-dimensional structure of clouds might be identified, or detailed information might be obtained through first principle modeling of the flows in simulations containing clouds. These types of information could then be used to train models to incorporate the evolution of clouds and to enable exploration of the effect they have on the flow and how they interact with the solar and terrestrial radiation leading to significantly improved climate models. Domain-knowledge-driven machine learning techniques can thus be applied to these detailed data to construct submodels to be used within coarser-resolution models of the global climate system. In this way, we can generate more reliable and faster climate simulations that will allow for more simulations and so larger ensembles, which in turn can provide more statistically certain climate predictions.

Many other of the non-resolved, so-called ‘parameterized’ processes such as radiation, turbulence, cloud microphysics, and interaction with surfaces, could also make use of techniques of this type and so be be incorporated, both into numerical weather prediction and into climate models. Furthermore, the analysis of a large ensemble of disparate model results targeting specific features such as low pressure systems, extreme winds or extreme precipitation, is also well suited for exploration through ML processes.

Since the amount of data is increasing extremely rapidly in the domain of climate and weather research, it will continue to be necessary to advance the range, efficiency and effectiveness of available techniques for analyzing them. Scientific insights by efficient use of these new techniques has the potential to advance the field significantly and increase understanding of the complexities of the climate system and its sensitivity to external forcing such as by greenhouse gas increase. Such a greater understanding, together with greater certainty in the results and predictions for climate change, will be of enormous benefit to global society.

4.7.2 Technical challenges and infrastructure requirements

Deep neural networks are proven to be the state-of-the-art technique to achieve high accuracy in image recognition and pattern identification, which are essential in many climate science applications, such as climate predictions and extreme event forecasts. In order to conduct high accuracy predictions, observational and/or simulation datasets can easily go up to PBs with high dimensionality. It imposes great challenges on storage and computational resources. Comparing to traditional image recognition problem, whose images are usually represented in 3 RGB channels, climate datasets have much higher dimensions. For example, humidity or air temperature at one geographical location can easily have one hundred vertical layers. Thus, neural networks with many layers and more complicated structures are required to perceive these high dimensional data in order to make a high accuracy prediction. It means more memory is required to store the parameters of the ML models. Considering the memory limits on a single GPU, multiple GPUs with large memories are needed. At the same time, algorithms that can efficiently coordinate the execution of machine learning tasks that are deployed on multiple GPUs are worth investigation. Software stacks for big data storage and batch/stream processing, such as Hadoop, Spark, and Flink, may be also required.

Recently, deep learning for extreme weather event prediction has achieved promising results. These models require the real-time analysis of weather events come in the format of time series. In order to allow neural networks to capture the order dependence between each climate snapshot in a sequence, long short-term memory (LSTM) networks are usually used together with convolutional neural networks (CNN). It is similar to the realm of video analysis, which requires high-end GPUs and powerful computing platforms.

Usually, the amount of computational resources determines the level of resolution for climate simulations, higher resolution simulations would be possible if simulation time could be reduced without compromising simulation fidelity, which could lead to potentially new insights in climate science research. In this research direction, we explore novel neural networks for predicting some parameterized modules in climate simulation. These neural networks will not only focus on improving prediction accuracy in order to provide results with fidelity, but also pay attention to the prediction speed. In other words, we would like to keep the models as small as possible under specific accuracy constraints. Thus, iterative tunings are required in order to achieve the optimal balance between model accuracy and prediction speed. In this case, local exclusive access to GPUs is necessary.

Because of the high dimensionality and volume of climate data, state-of-the-art computing platforms with GPU support is compulsory for climate science research.

4.8 Case 7: Finance

The financial industry is rapidly adapting and developing techniques in applied mathematics, machine learning, and statistics for a range of tasks including pricing, risk management and business intelligence applications. The field is concerned with data-driven decision making based on a wide variety of data sources across time-scales, ranging from slowly varying macroeconomic variables and business cycles to high-frequency electronic markets on that change on a micro-second level.

4.8.1 Scientific challenges and potential impact

Standard approaches in quantitative finance rely on modelling of asset prices as stochastic processes that are calibrated to current market conditions. The stochastic models can then be analysed, either analytically or by stochastic simulation methods, for pricing, hedging and risk management purposes. The stochastic models are often idealized in the sense that they capture certain, but not all, stylized features of the data, which makes them tractable and easy to interpret, but not particularly accurate.

The standard paradigm in mathematical finance is currently being challenged by the development of methods in machine learning, deep learning, and reinforcement learning. In the new framework hedging strategies, trading decisions and other market actions can be formulated by neural networks that enable 'model-free' approaches to be developed. Consequently, the actions produced by the model do not depend specifically on the chosen market dynamics and other idealized assumptions but may include market frictions such as transaction costs and liquidity constraints, as well as exogenous information such as news or business reports. Moreover, generative models can be developed that capture market fluctuations very efficiently, which enables the generation of future scenarios building upon complex relationships between the underlying variables. Such generative models could be implemented, for example, to study high-frequency market fluctuations or more general risk management decisions.

Possible breakthroughs on the horizon enabled by use and development of DL/ML in this domain include:

- methods for pricing and hedging of financial instruments using model-free approaches that incorporate realistic features of the underlying market,
- improved risk management through better understanding and prediction of high-frequency price movements, by combining generative models for generation of future scenarios and by learning of market execution strategies
- reliable trading performance and better customer pricing by designing automated strategies for trading execution through reinforcement learning.

4.8.2 Technical challenges and infrastructure requirements

The ability to perform state-of-the-art research in this area, which relies on the ability to analyse high-frequency data from electronic markets, train large-scale deep neural networks, and run large-scale Monte Carlo simulations on generative models, requires extensive computational resources. Without suitable computational hardware such research will not be competitive on an international level.

With regard to the needs of computational infrastructure significant new demands can be anticipated in the near future: specifically with regard to access to GPU-clusters for researchers.

Current availability of GPU resources is limited: for example KTH hosts a SNIC facility, Tegner, offering 9 Nvidia Tesla K80 cards. Far larger GPU-based systems are called for. Locally, within the research groups in the departments, there are demands for smaller GPU clusters (20-100 GPUs) for development, training and testing. On the larger scale, to seriously compete with state-of-the-art international research in machine learning, there is a need for computational GPU-resources on the scale of Beskow or better, offering hundreds of GPU nodes. Comparable systems and research examples include the following:

- Baidu (China): Deep Learning Scaling is Predictable, Empirically `http://research.baidu.com/deep-learning-scaling-predictable-empirically/` – 50 GPU-years
- Facebook (US): Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour `https://research.fb.com/wp-content/uploads/2017/06/imagenet1kin1h5.pdf` – 256 GPUs in parallel
- DeepMind (UK): Mastering the game of Go with deep neural networks and tree search `https://storage.googleapis.com/deepmind-media/alphago/AlphaGoNaturePaper.pdf` – 50 TPU-years

4.9 Case 8: Computer vision & video analysis

Computer vision is a very broad, multidiscplinary field which has long had a close relationship with AI. This is particularly true for the area of video analysis that addresses the task of extracting semantic information from continuous video streams, for example those originating from cameras in autonomous vehicles such as self-driving cars or drones.

4.9.1 Scientific challenges and potential impact

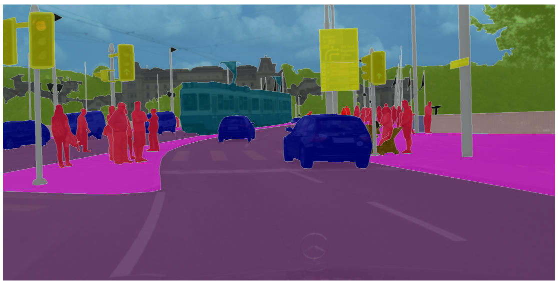


Figure 4.4: Segmentation example Zürich from the CityScape dataset `https://www.cityscapes-dataset.com/examples/`.

A good example to understand the scientific challenges presented by video analysis of streams from a self-driving car is the detection and localization of other actors appearing in the scenario being monitored by the video. This is usually accomplished by semantic segmentation of individual frames of the video into image regions identified as being other cars, pedestrians, cyclists, traffic signs, road surface, curbs, vegetation, buildings and so on. State-of-the-art approaches in this area apply Deep Learning for these tasks and training the networks requires enormous amounts of training data. As an example over 4000 videos were used in the training applied in the case described in `https://competitions.codalab.org/competitions/19544#learn_the_details`.

Successfully conducted research on video analysis will have major positive effect on various applications that assume working visual perception systems. Besides the previously mentioned self-driving cars, also drones and autonomous boats will benefit. Furthermore, video surveillance could also be automated to a large extent, removing the need for human operators in this tedious task. Finally, human-machine interaction could greatly benefit due, for example to improvements in gesture recognition, body language interpretation, and mimics.

These examples generalize well to other applications and other subfields of computer vision. The primary platform to publish highest level results in the field is the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Papers at this conference must contain novel methodological approaches that are convincingly evaluated in large-scale experiments. There is no doubt that we have only seen the beginning of the impact of AI in computer vision and video analysis.

Possible breakthroughs on the horizon which will be enabled by use and development of DL/ML in this domain are:

- on-the-fly learning for real-time video segmentation,
- reinforcement learning for dynamical vision-based control models, and
- learning of complex models for image-based estimation of physical entities.

4.9.2 Technical challenges and infrastructure requirements

During the design and tuning of the network the training of the network needs to be repeated hundreds of times to evaluate such aspects as different layer structures, cost functions, optimization methods, learning rates, and initializations. Effective set-ups need to be verified in ablation studies where leave-out tests need to be run on all components to show their relevance for the result.

Even for relatively simple networks training from scratch often takes as much as six days on a single GPU and fine-tuning will typically require a further 2–4 hours. Advanced topologies such as Resnet-152, which are required for such problems as advanced segmentation methods, need large amounts of GPU memory and as long as three weeks for training on 4 Nvidia Tesla K80 GPUs (See, for example, `https://bit.ly/2Dsu1EG`). One hundred runs would, therefore, require exclusive use of about 6 years on 4 K80s or one year on 24 K80s. The availability of computational resources is therefore the most significant bottleneck for further progress in research in this area.

Deep Learning for video analysis and related areas in computer vision requires primarily GPU resources rather than the CPU resources which have been the primary focus for SNIC in the past. Due to the large training datasets (of the order of terabytes) that need to be traversed hundreds of times, large scratch spaces and extremely high local bandwidth are also required. On the software side, mostly standard frameworks, such as TensorFlow, and container-based solutions are required.

Local resources are necessary to do initial trials and to avoid waiting periods caused by the scheduling of shared resources. A major bottleneck for local systems is the available GPU memory in consumer systems. For example the standard network system 'ResNet-200' cannot be trained on an Nvidia GTX 1080 with 8 GB of memory. Server-grade hardware is usually not affordable for individual research groups without additional funding, such as that which might be provided through WASP. Uniform container-based solutions might be a route to simplify the transfer from local systems to server-based solutions.

Server hardware is more common at the national level but the described research requires interactive and exclusive access to GPU nodes for those resources since interactive access is necessary in order to efficiently re-launch parameter trials, improving the accuracy of estimation methods, and for debugging. Exclusive access is necessary due to the large amounts of data involved.

It can be concluded that without suitable resources, research on video analysis and other Deep-Learning-based areas of computer vision will be seriously jeopardized. Also commercially available resources are not an option as the costs for access are, according to trials with Azure, in the order of an additional 50% on top of each PhD student.

5 — Industrial Use Cases

Machine learning is quickly finding its way into a wide range of industries. In this chapter a number of sample use cases from Swedish industry are presented to illustrate similarities in requirements with those from academia, as well as additional requirements on the compute infrastructure.

5.1 Industrial Automation

AI applications have until now been dominated, but not restricted to, machine learning applications targeting systems such as condition monitoring and predictive maintenance. Here, the Data are traditional measurements like vibrations, temperature, pressure etc. This leads to some substantial volumes of data, but so far these have not been so large as to require larger clusters, either for model training or inference.

In the future, when going into more fleet-based analysis as well as into autonomous systems in areas such as robotics, marine and mining applications larger volumes of data will result. This anticipated increase in information will require orders of magnitude more computation, in particular for model training.

As in most domains, the key to success is access to large amounts of annotated data. Suppliers of automation solutions face an additional challenge since the data is usually owned by the customers. This often makes the data commercially sensitive. Hence data privacy and data provenance will be a key factor for any interaction with academia in the use of external infrastructure. Local, on-site clusters are commonplace and small clusters will usually be sufficient. The biggest bottleneck at the moment seems to be related to software licence costs rather than hardware.

The digitalization of the automation industry is driving the development of AI services, typically placed in external public cloud infrastructure, utilizing built-in AI services and APIs for delivering those solutions to customers. During the research phase it may, however, be more practical and cost efficient to use on-site clusters or similar. Here a collaborative infrastructure could be an interesting alternative.

5.2 Machine Vision

The area of image processing and network attached equipment for the safety and surveillance industry is expanding rapidly. Traditional machine vision is rapidly being supplanted with machine learning based approaches, particularly for computer vision applications, but increasingly also for image compression and rendering. The machine vision market is seeing rapid adoption of these technologies, both as part of products in terms of specialized hardware and as part of product and software service development. By far the most challenging workloads in this domain are associated with machine learning training. Algorithms and frameworks are evolving rapidly and, as a general rule, image data sets are typically in the range of a couple of hundred gigabytes.

On-site computational resources could be augmented with external capacity to cover needs that vary greatly over product release cycles. However, software and dataset licensing makes commodity cloud services difficult and it is non-trivial to address these legal issues. The size of the datasets often prohibits, or at least complicates, collaboration across sites. Efficient and flexible ways to share large data sets is of great importance in the facilitation of collaboration, both across sites and between organizations. Data and algorithms are typically version controlled to allow for reproducibility and traceability.

5.3 Telecommunications

The application areas for machine learning within telecommunications are constantly expanding and range from advanced customer support tools to the realization of zero touch and highly autonomous systems capable of self-healing and adaptation to new operating conditions. By introducing AI/ML-driven automation in network operations it is possible to enable intelligent operations to predict, prevent and handle events without human intervention. On a regular day millions of alarms are raised on base-stations and left to be addressed by site engineers. AI and machine learning can play an important role in supporting humans in dealing with this vast influx of data. For example, anomaly detection based on AI is a promising possibility to reduce the need for human intervention while also improving response times. One further possibility is to go beyond simple fault detection by also providing AI-based root cause analysis. This means the system not only helps by identifying and classifying the fault, but also allows for tracing back to the initial failing component which led to the identified fault. Finally, predictive maintenance opens up the possibility of detecting patterns in equipment behaviour in order to predict when and if systems of components are about to fail or operate with reduced capacity, thus avoiding failures before they happen. The data handled is very diverse and is, by its very nature, decentralized. This type of data would benefit from streaming data analytics, rather than batch processing.

Data is typically owned by customers and is potentially sensitive in a business sense. Some data sources are also sensitive from a privacy point of view, possibly containing end-user identifiers. These aspects place additional constraints on where data can reside and how it can be transferred and processed in external systems.

AI will help service providers to handle the growing complexity in mobile networks. Machine learning can be used to leverage the skills of experienced engineers and technicians. Streaming data analysis and decentralized ML are interesting areas for collaboration, as are the requirements for secure data management and provenance.

5.4 Autonomous Driving

The future car will be heavily reliant on different kinds of AI technologies. Cameras are used to create situational awareness by detecting and tracking different kinds of traffic participants. One key enabler of visual perception is deep learning, which is showing great potential in terms of applicability and performance. It is necessary to have a scalable infrastructure for training, simulation, and prototyping in order to increase development speed (essentially reducing training times) and innovation in terms of capabilities to evaluate different algorithms and architectures.

The workloads are quite substantial since models are trained for image recognition using deep learning with millions of parameters. Distributed training sessions are also possible to further increase parallelization. Training times range from hours to weeks depending on needed functionality and dataset size. Reduction in training times is important in order to enable fast prototyping and so increase development speed. Video sequences are collected together with other sensor input and vehicle signals. The sequences are annotated with ground truth that will be later used in the training.

A typical user in this area makes use of both local computers and servers (GPU-based) to train networks. The former is primarily for debugging and quick tests while the latter is used for full-fledged training sessions. Access to GPUs is currently a limiting factor for development. In the future this problem will be further accentuated due to the need for larger networks and more functionality. Ease of use for developers is also a growing problem.

The number of AI developers and users in this area is expected to increase by a factor of ten in the coming two years. In terms of computational capacity there is a need to ensure the availability of infrastructure, improve ease-of-use, and to provide mechanisms to share capacity among multiple units to improve utilization. As both the user base and the datasets grow, data governance models and policies are becoming increasingly important. In the future additional computational needs can be expected due to, for example, the generation of synthetic data (generation of photo realistic images or video sequences) and semi-automatic annotation using deep learning.

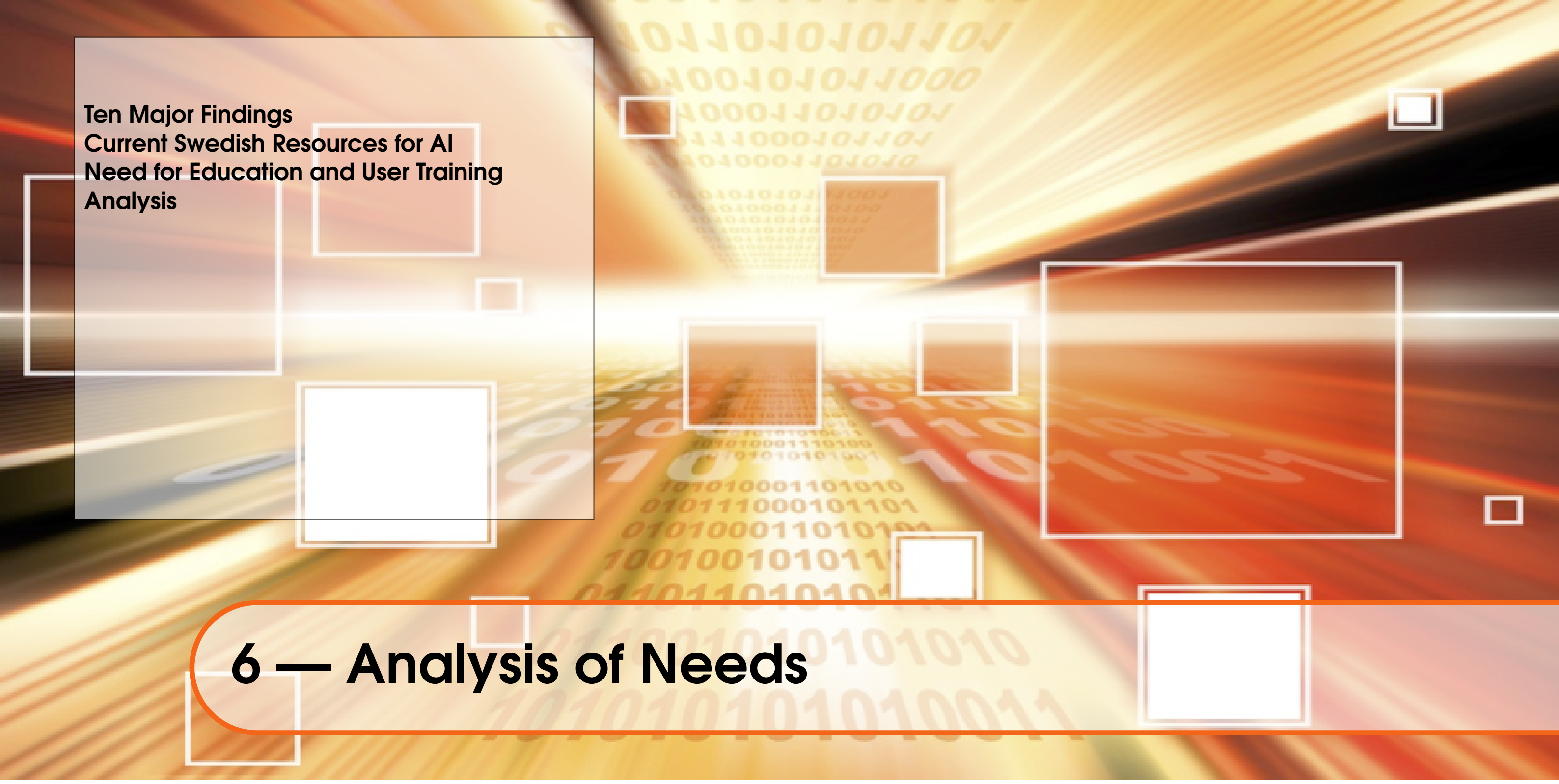
5.5 Medical Imaging

IT systems for data-intensive parts of health care is an expanding area. Today it mainly consists of approaches to the analysis of image data in radiology and pathology (microscopy) but in the future other areas, such as genetics and proteomics, can be expected to present similar data challenges. AI is a crucial part of the toolbox for the development of both current and next generation systems. While the classic AI task of detecting/classifying image findings is, of course, an important part of the process, the scope for what AI can provide is clearly much broader than that.

As is often the case with AI, the key to success is to obtain large amounts of high quality training data and to work effectively with it. This is unusually challenging in medicine because ground truth is determined by specialist physicians who are very expensive to use as an annotator. There is also large variation, both in assessments between individuals and in computer skills between different hospitals. Affordable access to large amounts of computational power is needed, ideally in the form of GPUs, together with a smooth management system that enables flexible switching between fast development cycles on local machines and major training efforts on nodes in centralized cloud services. It is crucial to integrate this with a proper version management system of the entire learning pipeline such as is, for example, found in Pachyderm.

Traditionally AI decision support has been developed by extracting a limited amount of training data from health care. A vision for the future would be to enable easier access to the clinical systems so that AI models can gain access to very large amounts of data and have the opportunity to continuously refine and calibrate the models. In this way methods and platforms that allow research and innovation can work closely with clinical activities and the data sources which are so very important. There are many challenges with this approach, particularly in the protection of these sensitive patient data. Anonymization will be a basic requirement in this context.

A highly relevant AI venture in healthcare is AIDA (https://medtech4health.se/aida), a Swedish national forum for AI and medical imaging which is being driven by CMIV at LiU. The goal is very applied: to develop AI innovations that take the final steps into actual usefulness in health care. In this forum we have an infrastructure for clinic-related research and development which we also plan to expand with a focus on large-scale data collection, a data arena or data factory in line with the idea of opening up data to the health care sector as described above.



6 — Analysis of Needs

6.1 Ten Major Findings

Based on the reports provided by the panels, an overview of existing resources, international outlook, and technology foresight has been created and a number of general noteworthy items have been identified. We list here the ten most significant findings:

1. GPUs are the current processing engine of choice in AI work flows but, in the longer perspective, other, more efficient, technologies may be on the horizon.
2. Investment in processing cycles needs to be augmented with a tightly coupled peripheral infrastructure, containing elements such as hierarchical storage (memory, SSD and disk), interconnection infrastructure, and network connectivity.
3. Internationally, substantial investments in infrastructure for AI are currently taking place in many countries.
4. HPC and AI have different access and resource provisioning methods. Convergence towards a Cloud-like delivery model is desired.
5. As the use of AI approaches is being adopted in many new fields, advanced user support and training is of high priority to make efficient use of the infrastructure.
6. Access to large scale GPU clusters is clearly an essential and costly stage in the AI workflow, but it also contains several other stages with varying needs for resources, data and software.
7. Efficient, robust and secure data handling is mission critical in most applications.
8. There are currently islands of advanced AI users in the Swedish academic community, and a challengingly large number of rapidly emerging applications spanning across the full range of disciplines.
9. The AI software landscape is changing rapidly with a continuous flow of new tools and updated technologies.
10. Users should be presented with one national entry point regardless of the organization of the physical infrastructure.

These findings form the framework for the analysis of current and future needs, as well as the recommendations for investments and supporting actions.

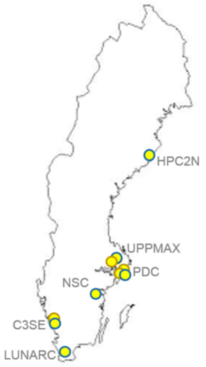


Figure 6.1: The SNIC Centers hosting nationally available HPC resources and providing user support and training to academic users in Sweden.

6.2 Current Swedish Resources for AI

To make recommendations regarding investments in view of needs it is necessary to first have an overview and assess the currently available resources.

Generally available computational resources in Sweden are provided by the Swedish National Infrastructure for Computing (SNIC). SNIC, a collaboration of 10 Swedish universities, offers large-scale computing resources to Swedish researchers on a merit basis via its six operating centers (See Fig. 6.1). The use of resources is free of charge once a proposal has been approved by a national allocation council. SNIC resources can, in principle, also be used by AI researchers but those resources are currently not tuned towards AI needs in either their hardware provision or their access mechanisms.

The SNIC centres have started to explore hardware configurations better suited for AI, particularly through the provision of GPUs, but these are mostly small experimental systems and their configuration is still tuned towards classical simulation workloads and not to AI. It should be noted that SNIC has invested in GPU based HPC systems in the past, but the utilization was low and they have now been decommissioned. The recent large-scale investments of SNIC: Kebnekaise at HPC2N and Tetralith at NSC, provide some more GPU resources but, again, their configuration is tuned towards classical HPC workloads. At the same time, typical AI software frameworks are tuned for massive GPU systems and not optimized for more standard HPC applications.

In this situation, Swedish researchers in need of larger computational resources, and without access to their own resources, often resort to cloud based resources like Amazon Web Services. It is also likely that some researchers have, through close industrial collaborations, access to the rapidly developing landscape of industry-operated resources.

On the commercial side, Vinnova is investing heavily in a new AI centre in Gothenburg (DataFactory) which aims to provide a testbed for companies interested in AI. To what extent this centre will provide IT infrastructure is unknown at the time of writing of this report.

The SICS ICE data centre, owned and operated by RISE SICS North in Luleå, provides an infrastructure and cloud research and test environment. The facility is open for use primarily by European projects, universities and companies. The facility can also be used for AI activities and Tensorflow has recently been added to its service portfolio.

At the base level research groups often have smaller in-house installations, in some cases small evaluation systems provided by a vendor, that are not sufficient for larger runs and are typically not accessible by outside users. Some groups have, however, invested in sizable infrastructure, for example the Laboratory of Molecular Biophysics that operates a 300 GPU cluster at UPPMAX in Uppsala.

6.3 Need for Education and User Training

User support for the AI/ML community is similar to the user support already being provided by several of the current supercomputer centres within SNIC in that it provides for installation of software, necessary and relevant performance tuning and some programming. As described in the previous sections ot this report, however, AI/ML also presents a need for a compute workflow optimization, in order to fully utilize centralized compute resources, as well as for data curation and storage optimization. One particular aspect is the fact that in AI/ML some model optimization processes are repeatedly performed using the same reference database as input which calls for efficient data curation.

AI/ML workflows require not only easy access to optimized computational back-ends but also a much more interactive approach, both in the development and in the deployment and production phases (see Fig. 3.2). In classical HPC, even if there is a development phase with a need for interactivity, most of the production computing is, in reality, run as batch jobs requiring a well-organized queuing system. Thus, one important aspect to keep in mind is that the traditional HPC usage is readily executed with great user satisfaction employing fair-share principles whereas the AI/ML community has a workflow that might be better served using a time-share principle.

With this in mind it is clear that an investment in an infrastructure for AI must be augmented with a significant effort to support users at all levels. The core elements in such an effort should be:

- An extended helpdesk function with rapid turnaround times and several communication channels.
- In-depth support from application experts targeting new users and new domains.
- Introductory courses introducing ML and presenting the resources available.
- Outreach efforts to reach new domains and user groups.

On a general level there is a need for application support for choosing the type of methodology, code and/or hardware platform. In the area of AI/ML there is a plethora of on-line resources for introduction and training. One of the tasks of the user support function will thus be to curate these resources and present them as a guide describing available courses/workshops at Universities as well as how they interface to the Swedish infrastructure. This would be beneficial for enhancing collaboration and serve as a basis for pointing out where new efforts should be directed. Application support also has to include code installation, code tuning and performance optimization and the first level of basic code handling.

It can be concluded that an investment in an AI resource in Sweden has to be accompanied with a level of user support that goes beyond the traditional HPC helpdesk and user support functions to make sure that new users are provided with an easy introduction to the resources and that advanced users can obtain and maintain optimal usage.

6.4 Analysis

The Swedish academic user community has, in a very short time, adopted the new paradigm of AI and is beginning to capitalize on the possibilities offered by deployment of AI approaches in their research. This rapid adoption has been stimulated by strategic initiatives by private and governmental funding agencies. The foundation for this report lies in section 4, where potential scientific breakthroughs are described and the required resources that would enable these are outlined. Even though the needs are hard to gauge in absolute numbers, it is clear that they are already significantly larger than the currently available resources in Sweden can support. Depending on the nature of the research and at different phases of the development, the kind and scale of resource can vary significantly. In some cases access to commercially available AIaaS will be sufficient and cost effective, whereas in other cases the data sizes and processing needs go far beyond that which can be accommodated, and dedicated infrastructure is needed to support the predicted research projects. It is also clear that the needs are rapidly increasing as use of AI is spreading into new domains and new problems are being addressed using machine learning across virtually all scientific disciplines.

It is also noticeable that the user communities have vastly varying levels of maturity when it comes to the knowledge and awareness of the potential and limitations of AI based solutions. This, in itself, calls for significant efforts to raise awareness and to educate the Swedish scientific community in the use of AI based tools in their research.

The items that are currently found to be most critical are:

- Access to large scale GPU based systems.
- Tightly integrated, large scale and high bandwidth hierarchical storage solutions.
- Support for full workflows for training and service.

Unless action is taken on these items the projected deliverables and breakthroughs described will not be achievable. In the following chapter the current infrastructure is described and the additions and changes needed to provide a sufficient infrastructure meeting the needs summarized above are then outlined for hardware, services and other supporting activities.

From the analysis of the use cases it is also clear that meeting the current needs is a matter of extreme urgency as some leading research groups, with advanced users and applications already in place, are in need of large scale resources to remain at the international forefront.

7 — Meeting the needs

7.1 The Landscape of AI services

As is clear from the analaysis, the hardware requirements for AI applications will, in at least the foreseeable future, include sizable clusters of servers with substantial amounts of both CPU and GPU capacity. They will require tightly integrated storage and provision of sufficient memory, connectivity, and interconnectivity. As their usage will differ from traditional HPC usage, they will come with new challenges with respect to delivery models and operation.

The AI workflows place particular requirements on the services to be provided, with some stages requiring many short executions and significant human interaction (for example for configuration, modeling, optimization and parameterization) while other stages require large-scale computations for training of full systems. Although the software landscape is evolving quickly, rapidly producing new packages which become heavily used while others become outdated and deprecated, it appears clear that the modes of operation will require a mix of interactive services for immediate access and quick response and other services for truly large-scale computations. Ideally, these different modes of operation should be provided in an integrated environment together with repositories for data, codes, log files, and other requisites, and with software supporting the workflow management itself.

7.2 Possible Scenarios for Investment

Given the documented clear need for increased support for AI workflows there are several ways to provide access and support. Given the opportunity to address this need through funding made available the alternatives need to be assessed. We thus provide an analysis of the strengths and weaknesses of three different solutions:

New National AI Resource at SNIC

An investment in large scale resource hosted by SNIC and connected to the existing SNIC infrastructure and services.

+ SNIC has established mechanisms for allocating resources to users and provides an existing, functional and efficient eco-system (experienced data centres, HPC and storage resources, connectivity, user support and training programs). By placing an AI resource at SNIC, synergies with the existing SNIC operations can be gained and full scientific workflows can be supported. Furthermore, SNIC has a competitive cost structure.
- SNIC has little experience of operating AI resources and the manner in which SNIC systems are configured and made available is tuned to classical HPC workloads. By investing in a new resource, technical diversity will be limited since only a small number of different configurations will be possible with the current funding levels. The configurations will also tend to be rather static, with limited possibility to change over time.

Procure Services

Funding is made available to research groups after a review of proposals. The funding is earmarked for procurement of services. Alternatively access can be negotiated with a provider at the national level and resources allocated according to the SNIC model.

+ By procuring services on the free market it may be possible to use configurations tailored for specific problems. The funds could also be used in a flexible way over time, allowing exploitation of new solutions as they appear on the market.
- The (in principle) flexibility might be hampered by the need to follow the rather strict public procurement procedures. Procuring services on the open market is typically more expensive than investing in hardware services (at least at the time of writing this report considering published access costs). Integrating AI components in scientific workflows where other components are not executed on the same cloud will be difficult. Although it can be expected that a wide range of different services and configurations will be offered, it is likely to be difficult to make providers change their offerings to meet specific needs.

Provide Funding to Research Groups

Funding is made available to research groups after a review of proposals, and no earmarking is made. The groups will then procure and operate group level servers and infrastructure.

+ Research groups can purchase solutions precisely tuned to their problems and have full control over them.
- Resources or services procured by individual research groups would be unlikely to be available to other groups and synergies with other national services cannot be exploited. The resources might be operated by groups with little operational experience and providing training and support for such customized solutions from the outside would be difficult. The entry bar for new users might be too high since they would have to deal with a wide array of detailed technical questions by themselves.

7.3 Integrating AI resources into SNIC

It is the conclusion of the reference group that the solution that optimizes the services provided through a direct investment by KAW is a tight integration with SNIC resources in the national

HPC infrastructure. This is motivated by the following considerations. SNIC has:

- an existing infrastructure with storage, network access, and the existing HPC systems will leverage an AI resource,
- operational environments for large scale computational resources,
- well developed user support functions such as help desks and applications experts that can be extended to include AI services,
- an application process for granting access to resources.

The fact that AI applications call for integrating interactivity into the HPC environments, rather than providing interactive resources as a stand-alone complement, is a strong driver in HPC to continue the already on-going convergence between the traditional service delivery models in HPC and cloud data centres. This convergence is also being driven from cloud data centres which, although traditionally focused on interactive provisioning of virtual machines, have increasingly been handling task-oriented jobs much like traditional HPC batch jobs but often much shorter and in larger numbers. From a resource management point of view this somewhat converged scenario requires resource management systems that can efficiently use the hardware to run both interactive services with varying load and unknown duration and tasks of a well-defined size and length. The ability to solve this problem and provide the solution needed for AI research has the potential to revolutionize HPC and, in addition, provide new types of services for areas not involving AI. It is recognized that this shift of HPC services will need support in terms of research and development projects.

A SNIC resource alone, however, will not be sufficient to cater for all needs and we will most likely see a mix of the three scenarios above, and with multiple sources of funding. In particular some users will benefit from exploiting commercially available AIaaS on-line services. For large-scale usage they may appear unfeasible from a cost-perspective but they may provide a convenient entry-point for new users interested in exploring the opportunities afforded by machine learning in their research processes. It is therefore important that well defined and smooth migration paths from AIaaS services to the proposed national resource are provided. It is also the conclusion of the group that an immediate investment should be seen as an initial step into an emerging domain of services with rapidly increasing importance.

7.4 Configuration of Resource

It is recognized that a SNIC resource for AI can be manifested in several different ways. As the usage patterns of AI is a rapidly evolving area it is hard to provide long term recommendations for configuration of a resource. There are, however, some observations that may guide the choices:

- Tight integration with memory and storage hierarchies to provide extreme bandwidth to data is key.
- Current usage patterns primarily build on massive utilization of separate GPUs, thus with limited demands on communication between nodes in a cluster. The trend is, however, moving towards scalable solutions which would depend on inter-node communication.
- The most cost-effective operation is likely to be obtained by investment in one large-scale clustered resource in one location. This, however, depends on considerations that have to be made by SNIC as it involves center-specific knowledge.
- Varying specific demands for configuration may call for investments in separate clusters with different hardware profiles.
- Proximity to data sources could be a deciding factor for location of processing capacity.

Based on these observations it is the conclusion of the group that the initial investment in a resource should not be fragmented into several units but should be kept as one unit to maximize cost efficiency in investment and operation, as well as enabling development of scaling methodology to support very large projects. It is also imperative that the allocations committee ensures that highly prioritized research gains access to large fractions of the resource capacity for individual tasks, to avoid fragmentation of that capacity. We also realize that the selection of the best SNIC site for the resource will be competitive and a plan for location and integration into the SNIC national services will have to be presented by the SNIC management.

7.5 Operational Competence and Related Research

The rapid introduction of AI methodology into the HPC infrastructure has some fundamental implications for the providers of the resources. this entails both handling of new hardware architectures and memory hierarchies as well as navigating the evolving software landscape. From a national perspective it is thus imperative that Sweden develops and maintains a high level of competence in this area. This calls for further build-up of highly skilled operational staff and application experts for advanced user support. This should ideally be done in close collaboration with research groups active in hardware and software research in the domain, which in turn will open up new opportunities for techology-oriented research projects.

7.6 Summary of Recommendations

Based on the analysis and conclusions the recommendation of the reference group can be summarized as:

- Investment in one large scale, GPU-based, state-of-the-art system which is integrated into the Swedish National Infrastructure for Computing.
- Investment in tightly integrated, hierarchical storage solutions to provide extreme bandwidth access to data and leverage such existing resources in SNIC.
- As operational costs are significant they must be included or covered by other sources.

To ensure the provision of high quality services to the community it is necessary that:

- A new delivery model based on convergence of HPC and Cloud paradigms is developed.
- Advanced user support for effective use of the AI resource is established.
- Outreach and training efforts are initiated at the national level.
- Coordination with Swedish industrial resources should take place and industrial collaborations must be supported.

It should also be emphasized that the committee sees the initial investment as a first step towards catering for the current urgent needs and with the expected growth of the use of AI based workflows significant further investments will be necessary.

8 — Acknowledgements

This report has been made possible by the contributions of a large group of people who have provided background information, texts and comments on the report. Primarily we would like to thank the members of the supporting advisory panels and all the participants of the open hearing day in June 2018. We have decided not to list the names of contributors as the editorial work have had large impact on provided texts and conclusions in the report are solely based on discussions and conclusions drawn by the reference group. We are also grateful for extensive proofreading by Dr. Matthew Cooper.