ELLIIT TECHNOLOGY FORESIGHT

2023 Companion

IT and mobile communications are transforming our lives and constitute a backbone of Swedish industry.

ELLIIT is one of two strategic research environments in Sweden in the area of IT and mobile communications. It has been funded by the Swedish government since 2010, and expanded significantly in 2020 to meet the needs of technology development for the digitalization of society.

The overarching goals of ELLIIT are to supply Swedish industry with competence, to educate the next generation of engineers, and to generate research results that can be directly exploited.

Partners are Linköping University, Lund University, Blekinge Institute of Technology and Halmstad University. Linköping University acts as coordinator.

In 2019, we published the ELLIIT 2030 Technology Foresight. The ELLIIT Technology Foresight 2023 Companion describes ten additional, emerging trends and challenges that we have identified in our strategic work. The Technology Foresight together with the Foresight Companion will drive ELLIIT's long-term development.

Contributors: Atila Alvandpour, Amir Aminifar, Bo Bernhardsson, Zheng Chen, Ove Edfors, Johan Eker, Michael Felsberg, Anders Hansson, Fredrik Heintz, Martin Höst, Maria Kihl, Volker Krueger, Marco Kuhlmann, Jan-Åke Larsson, Fredrik Lindsten, Liang Liu, Yurii Malitskyi, Jose Peña, Anders Rantzer, Joachim Rodrigues, Per Runeson, Mattias Tiger, Jonas Unger, Lars-Erik Wernersson, Farnaz Adib Yaghmaie, Anders Ynnerman, Karl-Erik Årzén, Kalle Åström

Contact: Erik G. Larsson, director, ELLIIT, erik.g.larsson@liu.se,
Fredrik Tufvesson, co-director, ELLIIT, fredrik.tufvesson@eit.lth.se

More information: elliit.se

CONTENTS

1. New methodology in machine learning 5
2. Learning for dynamics and control 6
3. AI-driven decision support 7
4. Swarm robotics 8
5. DevOps and MLOps 9
6. Future mobile communications 10
7. Distributed wireless intelligence. 11
8. The network as a compute fabric 12
9. New developments in semiconductor technology . 13
10. Quantum technology and computing 14

1

NEW METHODOLOGY IN MACHINE LEARNING

The field of machine learning continues to advance. In the areas of natural language processing (NLP) and computer vision (CV), a steady **increase in the sizes of both models and data sets, together with novel model architectures,** have continuously improved performance. For example: five years ago, a large language model comprised around 100 million parameters; today, models have more than 1 trillion parameters.

Attention-based models such as transformers are becoming increasingly popular, especially for NLP. Such architectural improvements are, however, largely based on human innovation. Systematic ways of optimizing model structures, for example through neural architecture searches, will likely become increasingly important as the models themselves become more complex.

Parallel to this development, a new learning paradigm, termed **self-supervised learning,** is on the rise. Classical approaches based on supervised, or reinforcement, learning are inevitably limited by the availability of supervisory signals, coming either from manually annotated training data, or the number of trial-and-error experiments carried out by a reinforcement learning agent. Self-supervised learning, on the other hand, enables the AI system to learn a generic world model in an unsupervised way, simply by observing the world.

Self-supervised learning has been hugely successful in natural language processing, and to some extent for computer vision. Generalizing these methods to **efficient joint handling of video and multiple data modalities** could result in the next major leap in machine intelligence. Indeed, having access to a generic world model does not only allow the AI system to excel at a specific task; it also opens up for continual learning where the system more easily can adapt to new situations from limited (or even zero) experience. By extending self-supervised learning to causal world models, by which the system can reason about the effect of its own actions, this development could reach yet another level.

In addition to CV and NLP, there are many other areas where machine learning is becoming a disruptive enabling technology: biochemistry, materials science, medicine, and climate science, to name a few. In many of these applications, data are comparatively scarce, making it more challenging to learn models of highly complex relationships and processes. To tackle this, one promising approach is to build prior knowledge in the form of **simulators and mechanistic models** into otherwise flexible data-driven models. Furthermore, **integrating semantic representations and ontologies** into machine learning, for instance through neural-symbolic computing, can lead to improved reasoning and explainability.

2 LEARNING FOR DYNAMICS AND CONTROL

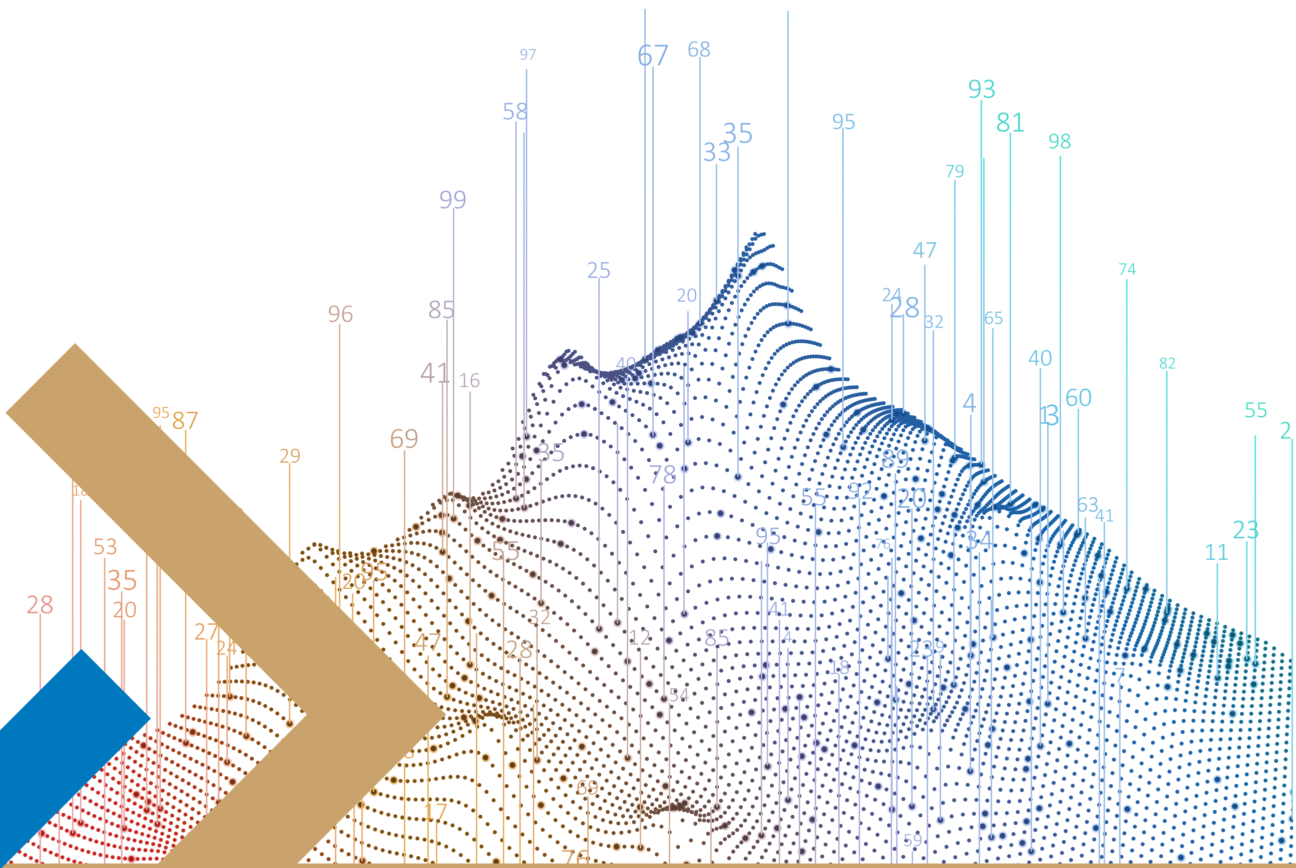
Devices that sense and control the physical world are expected to soon be the largest generator of data. The availability of a huge amount of such real-time data emerging from the physical world encourages a combined effort by experts in fields such as control theory, machine learning, statistics, and optimization. While control theory traditionally has been based on a model-based design paradigm there is an increasing interest in **data-driven approaches** to control and optimization of dynamical processes.

Recent progress in **reinforcement learning** has been rapid and delivered several impressive demonstrations of learning intelligent behavior. However, this has typically been in situations where data can be collected safely, for a long time, and with low cost for possible failure, such as in games and simulated systems.

Learning from existing data sets, such as those coming from normal operation of a large-scale production facility, can however be fundamentally limited by a lack of information in available signals.

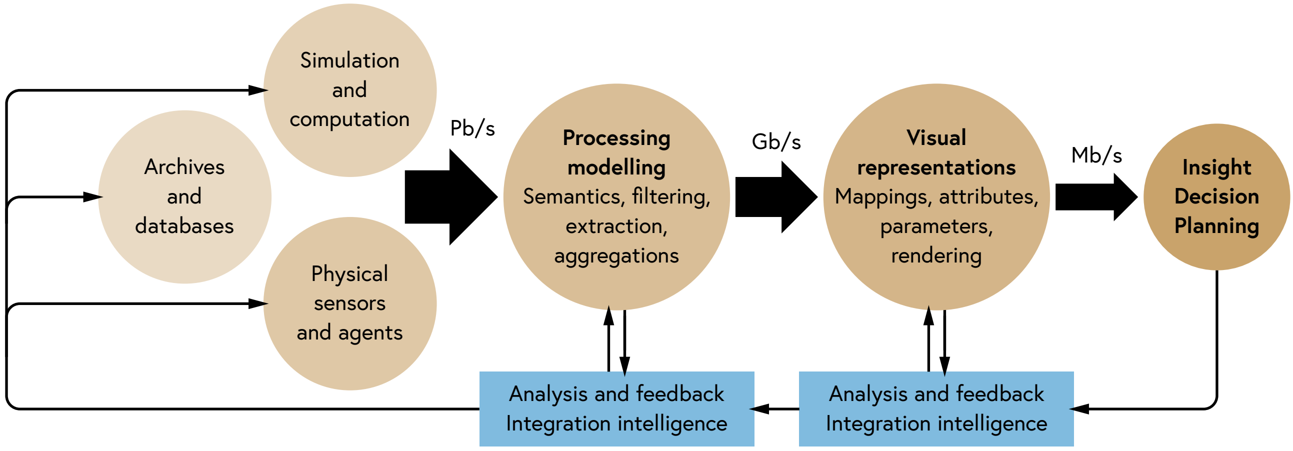
For the prediction model to be able to generalize into unseen situations it needs help from **prior information.** A major challenge is to formulate reasonable prior information that remains valid in uncommon situations, and to obtain efficient learning while maintaining safe operation. The resulting prediction model should also provide **system understanding** so that possible unsafe future behavior is identified and avoided and any opportunities for improved performance, such as redesign of the physical system or parameter adjustment, can be identified.

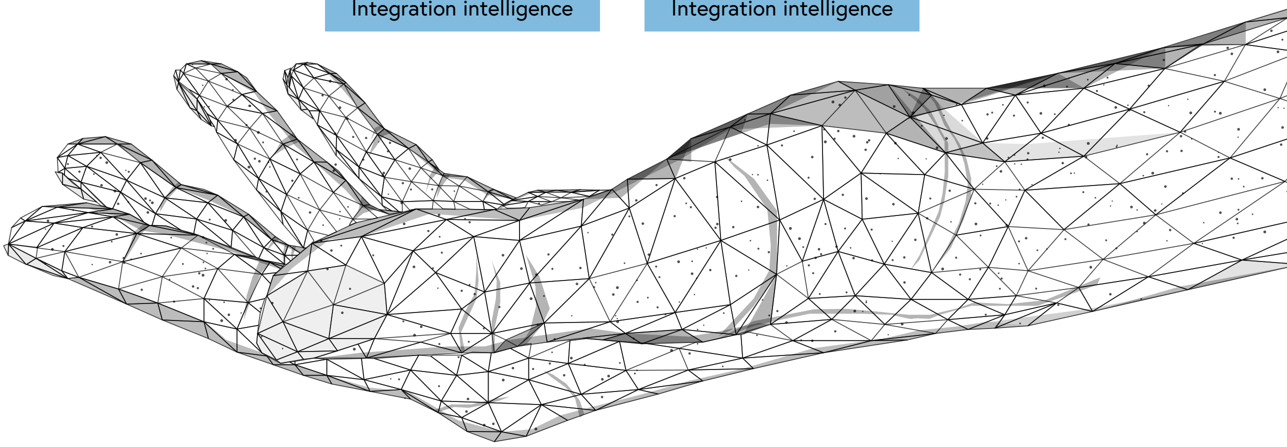
Applications where learning for dynamics and control will play a major role include robotics, transportation, health care, manufacturing, and agriculture.



3

AI-DRIVEN DECISION SUPPORT





Improved support for human decision making is of crucial importance in a variety of contexts and environments in many areas of society. **Seamless integration of AI** at several stages in the decision pipeline opens new opportunities for research and applications. A fundamental challenge is the **rapidly increasing amount and variety** of data available, for example, through sensors, simulation, databases, and from the Internet. This data holds invaluable information describing complex situations needed for human experts to make fast assessments and take informed decisions. The information gathered through data collection and analysis often consists of terabytes per second of streaming data that needs to be reduced to megabytes per second for human interaction and decision making.

The starting point is the increasingly sophisticated software and hardware systems that utilize decision support in the collection, analysis, and presentation of data in aggregated forms. To take decision-support systems and environments to the next stage, two central challenges need to be addressed.

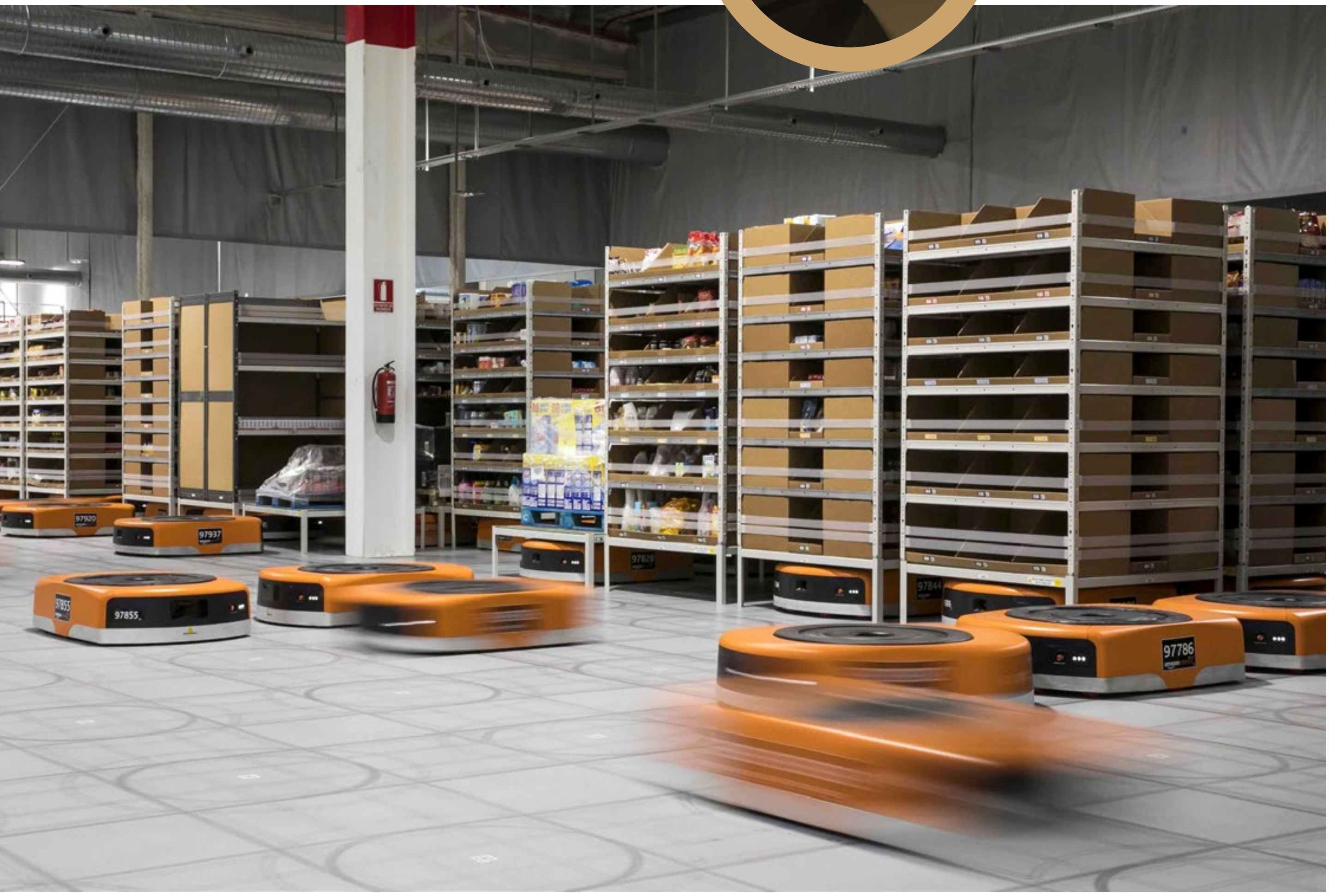
The first challenge is **seamless integration of simulation** as a tool for exploration of different future scenarios, with the goal of answering the “What if?” question and to predict how different interventions affect the outcome of a scenario.

The second challenge is to use machine learning and AI systems to **close the loop between the human user(s) of the system and the data sampling and scenario-simulation** systems. This scenario-integration intelligence can serve to filter the data, impose semantics, and augment the visual representations to support the analysis and planning performed by the users.

Another important aspect of systems intelligence is to steer simulation parameters based on data analysis to **further improve the data augmentation.**

To solve these challenges, research questions directed towards tangible interfaces, display technologies, immersive visualization, multimodal interaction, data mining and reduction, machine learning, visual representations, immersive technologies (VR/AR), and image analysis and synthesis need to be addressed.

SWARM ROBOTICS 4

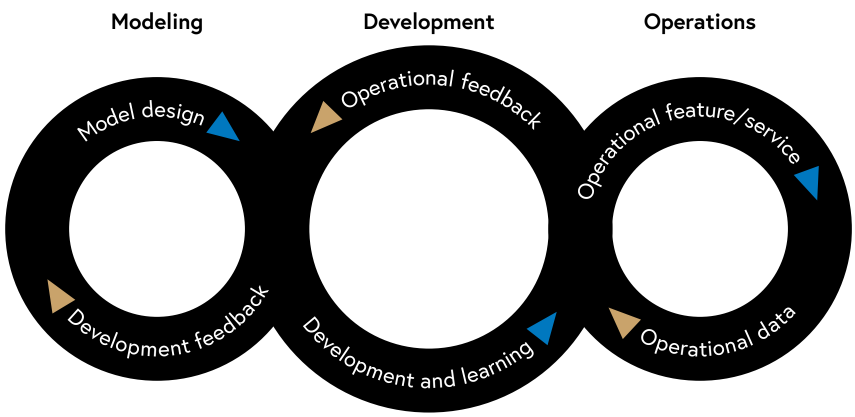


In swarm robotics, many simple robots perform tasks through collaboration, inspired by how swarms of insects, such as ants and bees, behave. Large robots usually have sufficient computational resources for deliberate reasoning, learning, sensing, and actions. Swarm robots, on the other hand, are comprised of simple individual robots; it is the swarm that performs a complex action, meaning that **the intelligence is on the system level** rather than with the individuals. The individual robots are controlled through simple reactive behaviors or reflexes, which in turn directly map sensory stimuli into an actuator output.

Swarm robotics has recently gained momentum within the study of Internet-of-Things (IoT) and distributed systems. Especially when **centrally coordinated and controlled through the cloud,** this technology has great potential to enable novel applications. An example is the Amazon KIVA system. Therein, the goal of each robot is centrally coordinated but the individual robots operate autonomously. Sensing and navigation are kept simple, relying on visual tags on the warehouse floor.

Scientific challenges include **self-healing by exploiting redundancy, collective decision making, collective fault detection, collective perception and knowledge accumulation, cooperative localization for improved navigation, and swarm planning of missions.** A key difficulty is to predict the swarm behavior emerging from local interactions among individual robots: especially, it is challenging to **prove the efficiency of a swarm solution** for a given application goal and context.

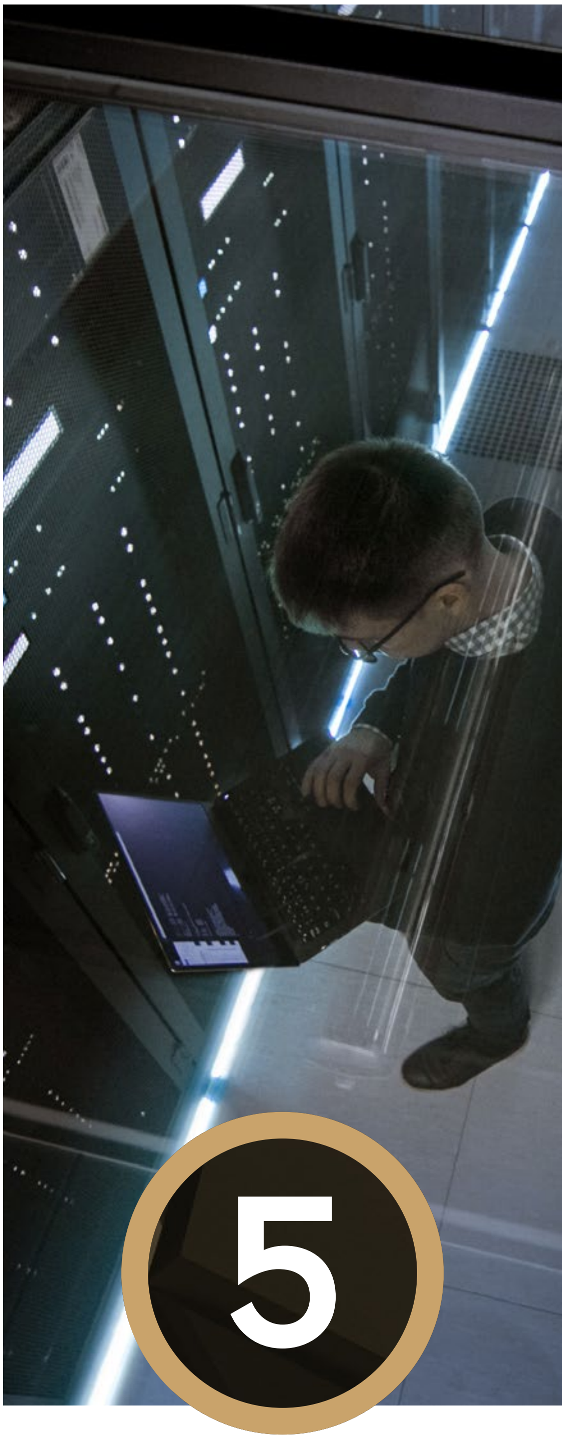
DEVOPS AND MLOPS



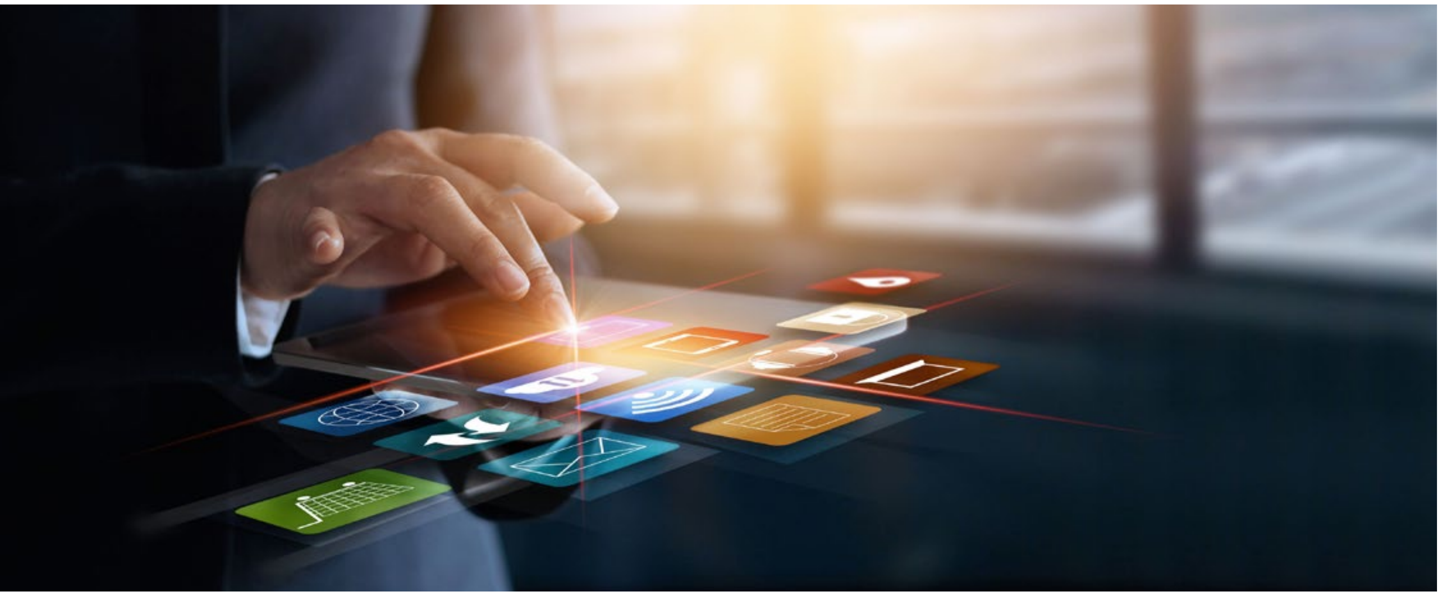
Software-intensive systems underpin most societal functions and constitute the differentiating value in ever more products and services. The development towards agile software-engineering practices has swept over industry domains during the last two decades. In its extremes, continuous integration and deployment (CI/CD) of software updates is done several times an hour, which forces the **development and operations of software-intensive systems to be integrated,** often referred to as DevOps. This enables **continuous feedback from operations** to development and **continuous experimentation approaches,** such as A/B testing, to guide development towards data-driven and gradually automated design decisions.

With the introduction of machine learning (ML) components in products and services, new challenges for software engineering arise. The continuous **remodeling, retraining, and deployment of trained ML models** also take place at a high pace, and are referred to as MLOps. Data pipelines feeding MLOps raise new challenges of **data curation, annotation, versioning, and quality assurance.**

While these trends are driven by industrial practice, there are several research challenges to address, in both foundational and applied research. For example, an understanding of **prescriptive knowledge** of DevOps and MLOps practices, along with required tool capabilities, and **limits to their application,** for example, in safety-critical systems, must be developed. Methods and tools to ensure **traceability and authenticity of data, as well as reproducibility of results,** including versioning of data, models, and tools, are needed. Methods for **automation of decisions** in DevOps and MLOps, for both feature and model selection, reducing manual engineering effort and optimizing the value of the resulting products and services must be developed. Finally, the application of DevOps and MLOps to embedded and cyber-physical systems bring challenges related to the ability to **upgrade systems during operation, security, memory and computational constraints,** and risks related to **physical damage.**



FUTURE MOBILE COMMUNICATIONS



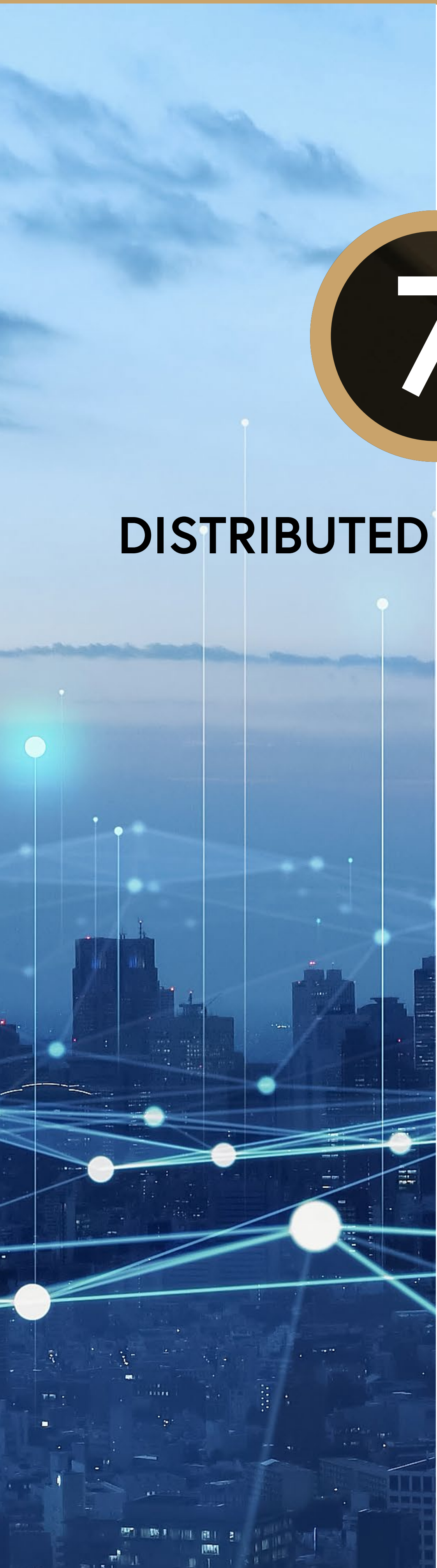
With 5G wireless systems deployed, the race towards the sixth generation – 6G – is on. 6G will enable entirely new applications, made possible by its exceptional, envisioned performance: peak data rates exceeding 100 Gbps, wide area coverage at multi-Gbps levels, extreme energy efficiency, low cost, imperceptible (sub-ms) latency, seven-nines reliability, and device densities an order of magnitude higher than with 5G. As some of these targets touch the theoretical limits, entirely new wireless communication concepts need to be developed.

Moving to sub-THz frequencies, where spectrum resources are abundant, will facilitate grossly increased peak data rates but also brings inherent **limitations in terms of coverage and reliability.** These limitations can be mitigated by dense deployments, but this raises another important challenge: **economy of scale.** Existing technical solutions for sub-THz communication use expensive and exotic semiconductor technology. The development of new energy-efficient, low-cost solutions is critical.

The use of large antenna arrays for wireless transceivers has met with great success, ultimately resulting in Massive MIMO technology that was adopted in the 5G standard. In addition to **increasing the spectral and energy efficiency** of the communications enormously, Massive MIMO also enables entirely new possibilities for **high-resolution sensing of the wireless propagation environment.** The 6G development will take Massive MIMO technology further to the use of large, *distributed* antenna arrays and intelligent surfaces for coordinated transmission and reception over physically large regions. These arrays need to be interconnected with high-capacity links and the antenna signals must be co-processed in an energy-efficient manner.

Another direction is **low and zero-energy ("passive") wireless devices** that will substantially rely on energy harvested from radiofrequency signals or ambient sources such as solar cells. Such devices are envisioned to constitute the backbone of massive low-maintenance sensor deployments and future Internet-of-Things (IoT). In low/zero-energy devices, extreme energy efficiency must be achieved on all levels, from the circuit solutions to the physical-layer signal design to the protocols used to access and wake up sleeping devices.

Artificial intelligence (AI) and **machine learning (ML)** have developed tremendously recently and provide means to autonomously optimize, tune, control and configure increasingly complex systems. These technologies are already used in 5G for, by way of example, beam selection, but are far from exploited to their full potential. In 6G, AI will be a critical component in achieving flexible, energy-efficient, and reliable high-performance services, with low deployment and operational costs. Specific problems where AI/ML has potential include **mitigation of non-linear distortion and intermodulation problems, traffic scheduling, assignment of infrastructure/computational resources, and positioning services.**



7

DISTRIBUTED WIRELESS INTELLIGENCE

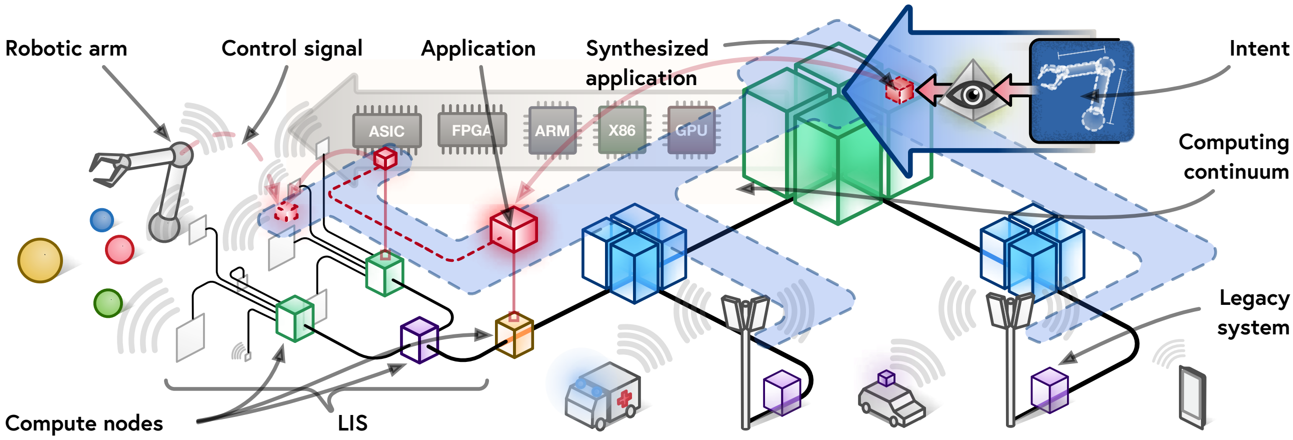
As we enter the era of connected intelligence and massive Internet-of-Things (IoT), new challenges appear in the intersection between wireless communications, machine learning, and distributed algorithms. The multidisciplinary nature of the IoT technologies requires **new perspectives on the joint design of communications, computation, and control.** A prime example is collaborative machine learning at the wireless network edge, where machine learning parameters and models are exchanged among edge devices, either in a fully distributed manner or coordinated by a central server. The **limitation of wireless communication resources** (frequency, time, space) is the bottleneck that determines how much information data can be reliably transmitted and received, which in turn affects the data compression loss, communication latency and convergence of the learning models.

The design principles of wireless resource allocation in distributed intelligent systems must adapt to the specific tasks performed by machines instead of human perception. This naturally means that next-generation wireless networks will take **a major shift from traditional rate-driven design to task-aware information processing, communication, and data aggregation** over large-scale distributed systems. In a broad sense, this is related to the concept of goal-oriented semantic communication that recently emerged as one component of post-Shannon information theory that goes beyond the principle of transmitting error-free bits.

This fundamental shift gives rise to many new questions and challenges, including communication design and signal processing, energy efficiency, and cross-layer resource management. In addition, new algorithm designs are needed that can **combine communication efficiency with data privacy, security protection against malicious threats, and robustness in model training.** Each of these aspects pose many new challenges, as there can be no definitive one-fits-all solution for all practical scenarios.

8

THE NETWORK AS A COMPUTE FABRIC



The evolution of 6G communications requires a fundamentally redesigned software infrastructure, a network compute fabric, including massive computing resources to make distributed, complex, and coordinated decisions throughout the whole infrastructure. The compute infrastructure is required to provide the foundation for hosting both the **massive computing capacity** required for the wireless network infrastructure itself, and an **elastic computing platform** for hosting demanding user-centric third-party applications.

The paradigm shift from cell-based Radio Access Network (RAN) technologies to cell-free technologies, such as Large Intelligent Surfaces (LIS), will pose totally novel orchestration and infrastructure challenges, where **massive hyper-local computing** is required. Adding energy constraints and sustainability goals make these challenges even harder to overcome.

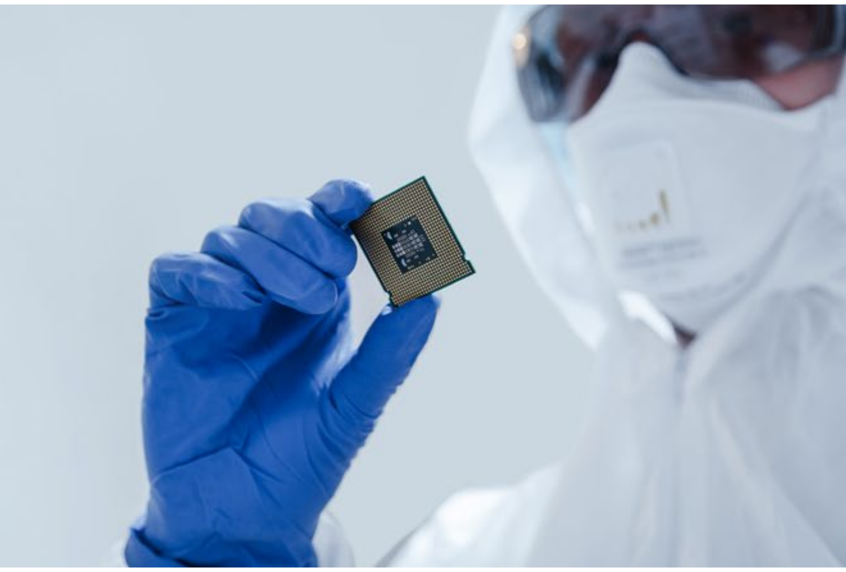
6G will incorporate millions (potentially billions) of connected devices. Traditional monitoring and management approaches in such systems will incur enormous energy and performance overheads. Therefore, it is essential that new techniques are developed to allow for **energy-efficient collection and analysis** of key metrics on a massive scale. Further, the complex architecture and the requirements of the applications require **anomaly detection** that can offer high precision and on-demand network management, service isolation, and multitenancy.

Applications will require extreme availability and reliability, despite running on nodes that are potentially unprotected from external impact. Consequently, the compute infrastructure must meet **extreme availability and reliability** demands. However, the infrastructure will also provide an abundance of resources on the global scale, which enables opportunities for performance improvements, by allowing the system to trade performance with availability and reliability.

Additionally, the application software itself needs to be re-imagined, due to the complexity of the system. Applications need to be both **extremely resilient and quality-elastic,** meaning that they must dynamically adapt to prevailing resource availability, wherever they happen to reside in the infrastructure, without failures. Therefore, applications should be synthesized from an intent, such as an objective to control a robot in a certain manner given a set of constraints. The system would synthesize the intent into a quality-elastic implementation that can be deployed and function where it is deemed needed by the system orchestrator.

NEW DEVELOPMENTS IN SEMICONDUCTOR TECHNOLOGY

9



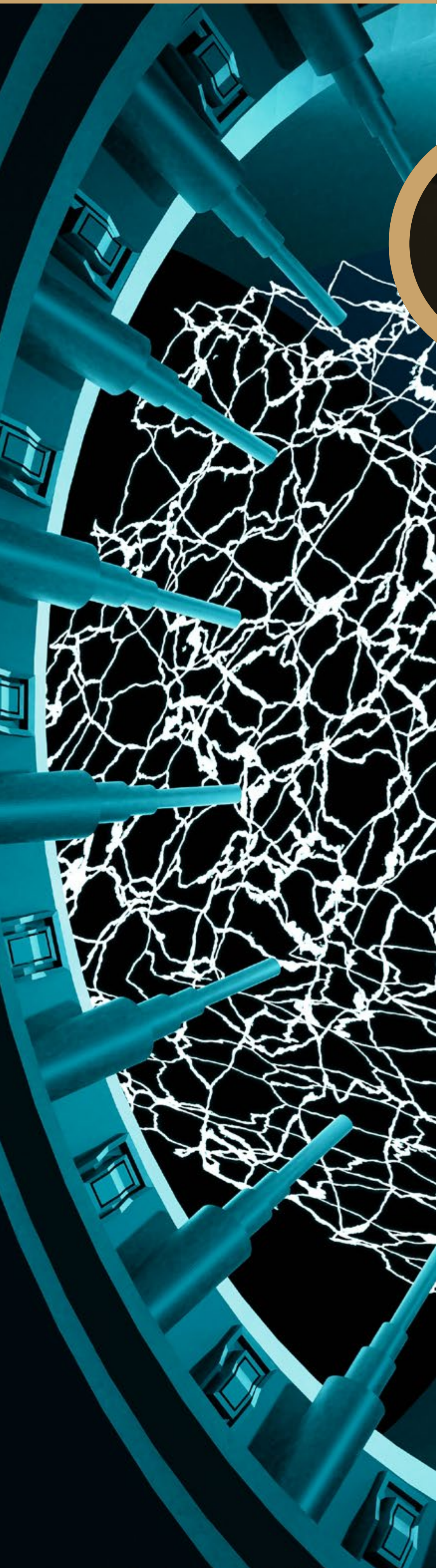
Semiconductor technology is a key enabler for digitalization and the green transformation. With increasing numbers of electronic components, it is essential to **optimize chip energy efficiency** within, for instance, IoT, connectivity, electrification, and computing. The EU has announced the **European "Chips Act"** with strong efforts on **research, support to the ecosystem** in areas of chip design, **testing,** and **advanced nodes.** In addition, efforts are targeting **resilience in European chip production.** The focus of the Chips Act will be on FD-SOI scaled down to 10 nm and below, leading-edge process technology at 2 nm and below, and advanced heterogeneous system integration (2.5D and 3D integration). Silicon-based CMOS, competitive in terms of cost and maturity, will be utilized for design of future advanced systems on chips, with nodes selected by application functionality, cost, and performance.

Efficient implementation of AI technologies including machine learning will require design of systems at the latest technology nodes. It is hence critical to get access to technology and to master the chip design at these nodes. Pure physical downscaling of CMOS transistors will be challenged by the performance benefits in power and intrinsic delay that can be accessed by **better adaptations of computational structures and system architectures, and 3D technology design and functional integration.**

Next-generation communications and networks will require intelligence and communication capabilities in devices and objects to enable people and things to work together intelligently – Internet of Things (IoT). This in turn requires highly tailored application-specific smart wireless **energy-efficient sensors,** which are often **self-powered** utilizing energy harvesting. These should support an increasingly wide range of application areas such as medical devices and healthcare, safety/security/surveillance of important infrastructures, automotive and electrical cars, and industrial automation.

Future communication systems (6G and beyond) will include operation in the sub-terahertz bands (100 GHz and above). Current roadmaps suggest that future critical circuit blocks will be based on **heterogeneous technologies integration** combining both advanced Si-CMOS and BiCMOS technology, but also GaN (PAs) and III-Vs (LNAs). **Innovative circuit design** of RF and **mixed-mode building blocks** as well as **supporting digital system design** need to be researched together with system-level and domain-specific knowledge to support key foundational applications.

The **merger of memory and logic technology** opens up for new concepts including in-memory computing and neuromorphics. The combination of materials innovation and system knowledge is essential to reduce the overall power consumption. Down the road, quantum technology will emerge as a viable candidate for computing and data analysis.



10

QUANTUM TECHNOLOGY AND COMPUTING

Quantum mechanics is just over a century old, and has given us technology such as the transistor, the laser, and the atomic clock. These are all enabled by collective effects of quantum systems and have led to the 20th-century technological revolution that included computers, optical-fiber communication, and the global positioning system, all of which are vital to the world economy.

We have now reached a point where we can build technology that uses **single-system quantum effects,** enabling highly sensitive measurements and efficient protocols. This is known as the **second quantum revolution** and the term "quantum technology" used in its modern meaning refers to technology from this second revolution.

An example of applications of this second quantum-revolution technology is **Quantum computers** that use single-quantum systems to perform calculations much more efficiently than is possible in classical systems, to solve problems ranging from the design of new chemical catalysts, protein folding, complex optimization problems and crypto cracking, to the answering of fundamental scientific questions in nuclear physics, high-energy physics, and other fields.

Another example is **Quantum communication** that promises secure communication even in the presence of the crypto-cracking tools from quantum computers and enabling distributed quantum computing, and in the long run, the quantum Internet.

A third example is **Quantum sensing** that provides high-sensitivity sensors for ultra-high-precision microscopy, positioning systems, clocks, and gravitational, electric and magnetic-field sensors, and also allows us to reach optical resolution beyond the wavelength limit.

The emerging field of quantum technology is developing rapidly with many new applications expected in a near future.

Images: **1–16** Patuk Bohdan/Shutterstock.com **2–3** Rawpixel.com/Shutterstock.com
5 Phonlamai Photo/Shutterstock.com **6** majcot/Shutterstock.com **7** Butusova Elena/Shutterstock.com
8 Amazon **9** Gorodenkoff/Shutterstock.com **10** PopTika/Shutterstock.com **11** metamorworks/Shutterstock.com
13 Vladeep/Shutterstock.com **14** adison pangchai/Shutterstock.com
Design and layout: Linnkonsult
Print: Tryckeriet i E-huset, Lund 2023

ELLIIT
Excellence Center at Linköping – Lund
in Information Technology