ELLIIT 2030 TECHNOLOGY FORESIGHT

New directions for strategic research in IT and mobile communications

ELLIIT

Excellence Center at Linköping – Lund in Information Technology

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

ELLIIT is one of two strategic research environments created by the Swedish government in 2010 in the area of IT and mobile communications.

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

This document summarizes ELLIIT's general long-term vision, and outlines specific, new research directions that urgently need investments.

Contributors: Patrick Doherty, Inger Erlander Klein, Michael Felsberg, Görel Hedin, Martin Hell, Ingrid Hotz, Jörn Janneck, Charlotta Johnsson, Christian Kowalkowski, Liang Liu, Anders Rantzer, Anders Robertsson, Cory Robinson, Per Runeson, Isaac Skog, Cristian Sminchisescu, Walid Taha, Fredrik Tufvesson, Anders Ynnerman, Karl-Erik Årzén, Kalle Åström

Contact: Erik G. Larsson, director, ELLIIT, erik.g.larsson@liu.se

More information: www.liu.se/elliit

Images: (Shutterstock.com) **1, 28** Maksim Samusenko 4–5 Nomad_Soul 6–7 Mopic 8–9 whiteMocca 10–11 metamorworks 12–13 PopTika 14–15 GarryKillian 16–17 Alones 18–19 Phonlamai Photo 20–21 TippaPatt 22–23 NPFire 24–25 Yurchanka Siarhei 26–27 Funtap

Design and layout: Linnkonsult

Print: Tryckeriet i E-huset, Lund 2019

FOCUS THEMES

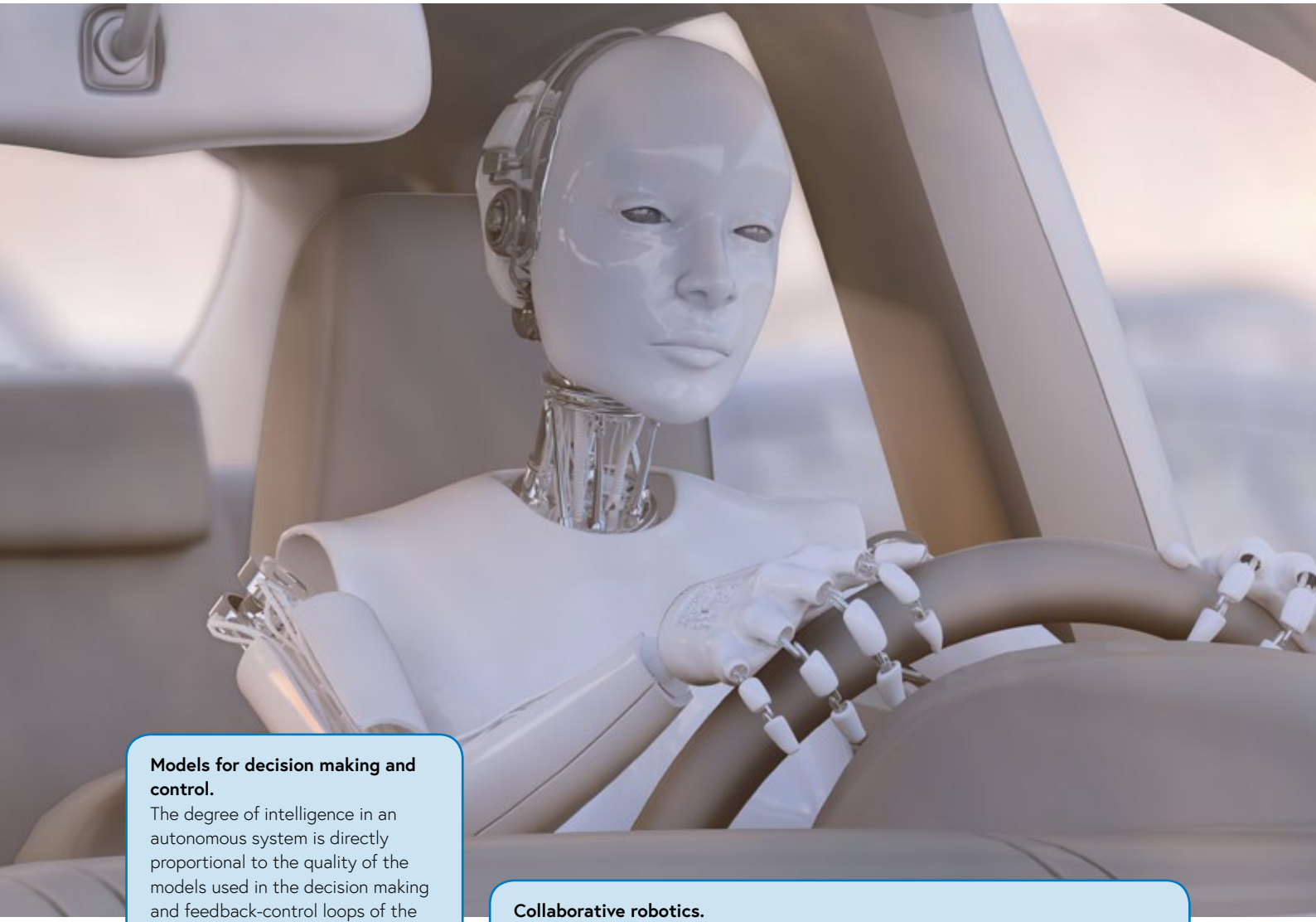
EMERGING RESEARCH THRUSTS, TECHNOLOGIES, AND CHALLENGES

FOCUS THEMES

Autonomous vehicles and robots.

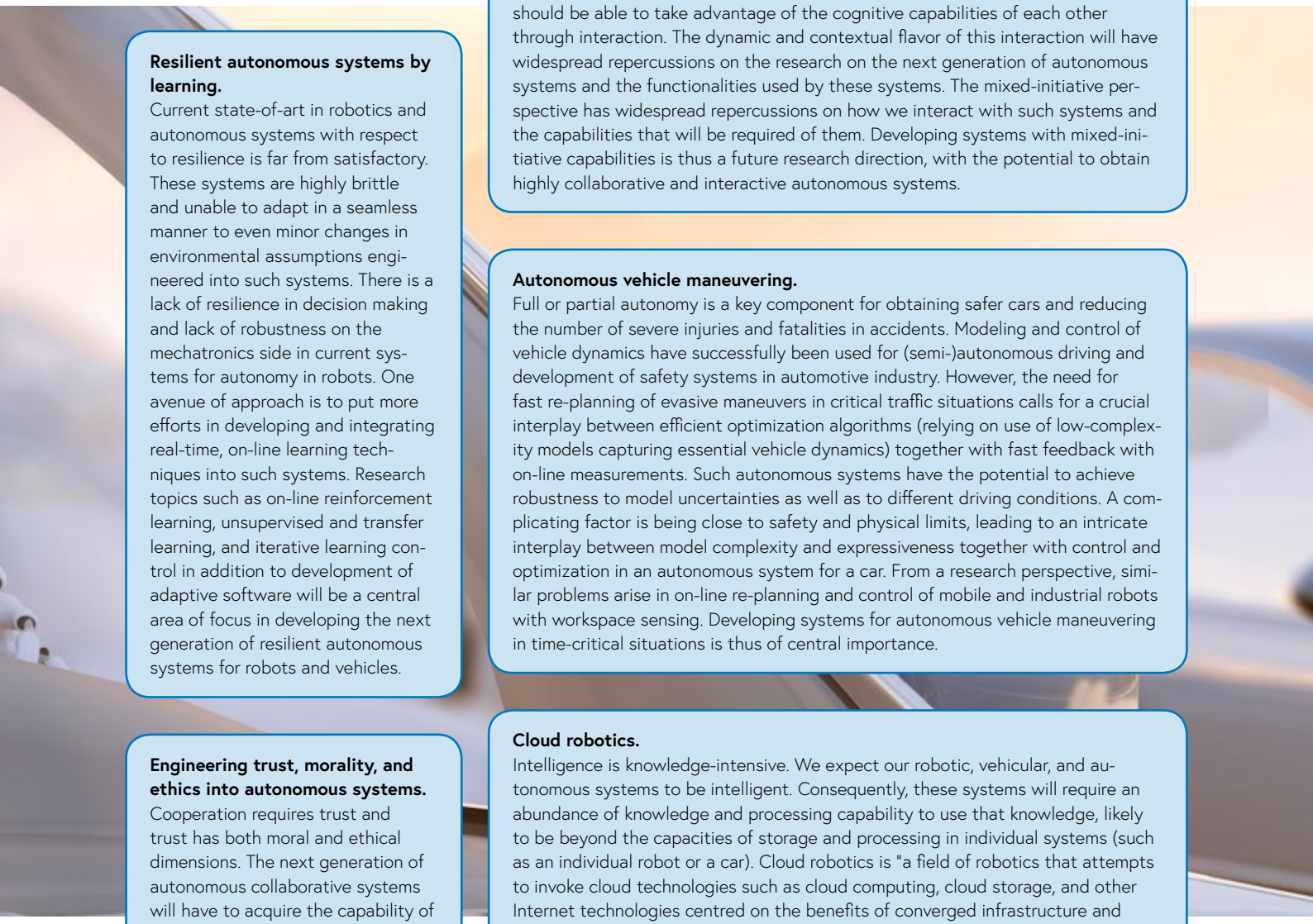
Traditionally, autonomy has been characterized in an individual machine- or system-centric manner, where different levels of autonomy have been described relative to a specific robot, vehicle, or autonomous system. The modern view of autonomy is broader and more encompassing with a shift from a system-centric point of view, to a relational point of view between the autonomous systems themselves and their users. In this context, autonomy is characterized as a bounded, contextual, dynamically controllable capability shared and constrained through interaction with human users. Issues surrounding mission autonomy become paramount and its interaction with traditional control autonomy introduces new ways to think about architectures, functionalities, and interaction with humans. The convergence of artificial intelligence (AI) and robotics into integrated systems rather than being viewed as separate disciplines is central. This melding of disciplines will drive research and technology development in this area for generations to come.

Future research in the field should focus on the following topics.



Models for decision making and control.
The degree of intelligence in an autonomous system is directly proportional to the quality of the models used in the decision making and feedback-control loops of the system, their timeliness in usage, in addition to the veracity of those models to those aspects of the system itself and the embedding environment they are intended to model. Models, and algorithms that use models for decision making and control, are central components in autonomous systems for robots and vehicles, and much research is being directed into this area.

Collaborative robotics.
While the traditional setup of collaboration between a human user and an autonomous system has been restricted to pairs — one human, one robot —, the new generation of autonomous systems will be required to have capabilities allowing them to interact and collaborate in a larger system of systems to achieve shared objectives. This trend requires development of new models for interaction, collaboration, sensing, control, decision making, and problem solving involving distributed teams of human, autonomous, and robotic systems. Major emphasis will be placed on research for achieving distributed sensing and control, distributed planning and decision making, distributed use of models, and dynamic fusion of models. Here, both centralized and decentralized approaches to collaboration, together with their combinations, will be additional areas of research interest.



Mixed-initiative interaction.
The next generation of autonomous systems are not intended to replace humans, but rather to extend human capabilities through collaboration and interaction. Such systems will be required to interact in a more natural and deeper manner with human users, which implies new research directions. Human users and robots should be able to take advantage of the cognitive capabilities of each other through interaction. The dynamic and contextual flavor of this interaction will have widespread repercussions on the research on the next generation of autonomous systems and the functionalities used by these systems. The mixed-initiative perspective has widespread repercussions on how we interact with such systems and the capabilities that will be required of them. Developing systems with mixed-initiative capabilities is thus a future research direction, with the potential to obtain highly collaborative and interactive autonomous systems.

Resilient autonomous systems by learning.
Current state-of-art in robotics and autonomous systems with respect to resilience is far from satisfactory. These systems are highly brittle and unable to adapt in a seamless manner to even minor changes in environmental assumptions engineered into such systems. There is a lack of resilience in decision making and lack of robustness on the mechatronics side in current systems for autonomy in robots. One avenue of approach is to put more efforts in developing and integrating real-time, on-line learning techniques into such systems. Research topics such as on-line reinforcement learning, unsupervised and transfer learning, and iterative learning control in addition to development of adaptive software will be a central area of focus in developing the next generation of resilient autonomous systems for robots and vehicles.

Autonomous vehicle maneuvering.
Full or partial autonomy is a key component for obtaining safer cars and reducing the number of severe injuries and fatalities in accidents. Modeling and control of vehicle dynamics have successfully been used for (semi-)autonomous driving and development of safety systems in automotive industry. However, the need for fast re-planning of evasive maneuvers in critical traffic situations calls for a crucial interplay between efficient optimization algorithms (relying on use of low-complexity models capturing essential vehicle dynamics) together with fast feedback with on-line measurements. Such autonomous systems have the potential to achieve robustness to model uncertainties as well as to different driving conditions. A complicating factor is being close to safety and physical limits, leading to an intricate interplay between model complexity and expressiveness together with control and optimization in an autonomous system for a car. From a research perspective, similar problems arise in on-line re-planning and control of mobile and industrial robots with workspace sensing. Developing systems for autonomous vehicle maneuvering in time-critical situations is thus of central importance.

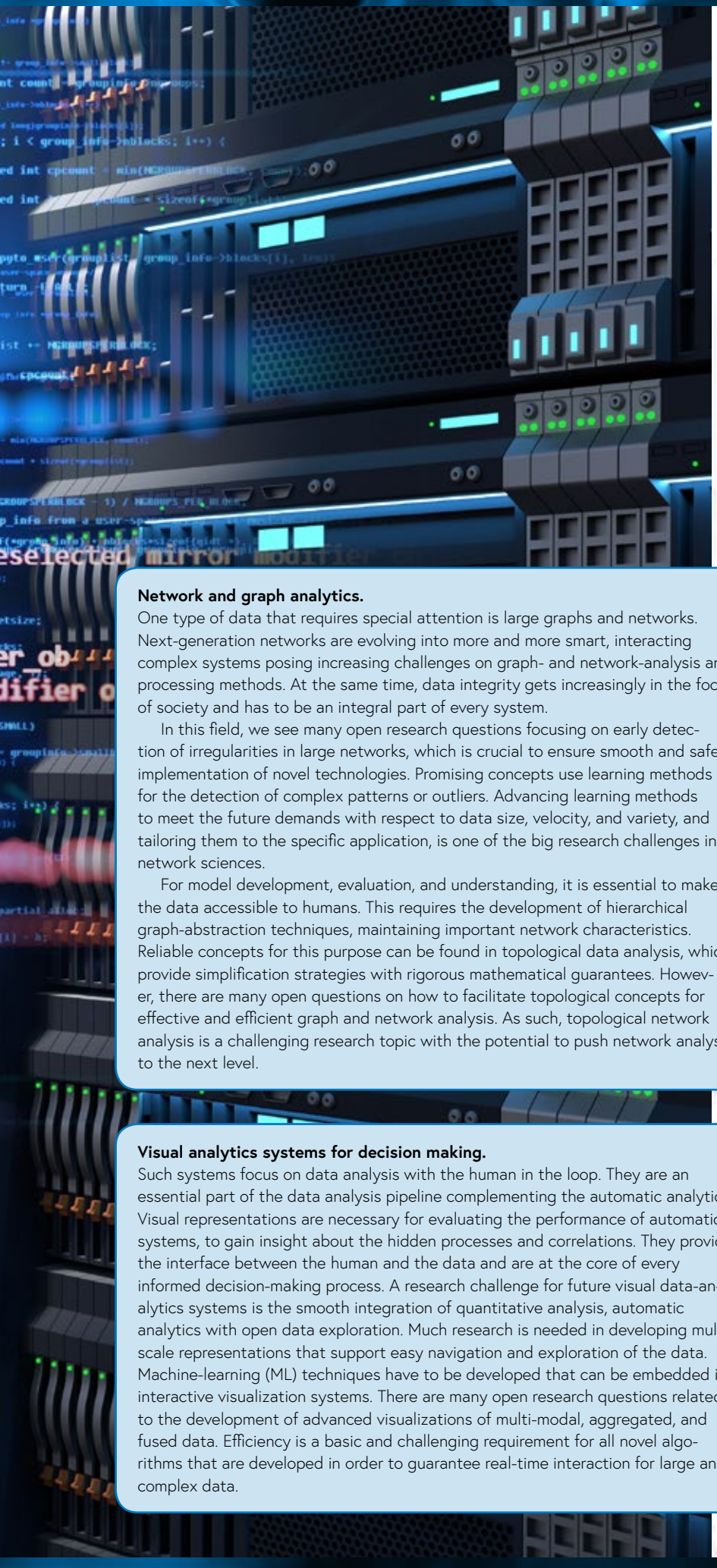
Engineering trust, morality, and ethics into autonomous systems.
Cooperation requires trust and trust has both moral and ethical dimensions. The next generation of autonomous collaborative systems will have to acquire the capability of making not just decisions, but decisions constrained by moral and ethical considerations just as humans do. This topic is an emerging area of research and of pressing importance in achieving co-habitation of this technology with humans and perhaps one of the most exciting future topics in autonomous systems with very few concrete results related to engineered mechanisms.

Cloud robotics.
Intelligence is knowledge-intensive. We expect our robotic, vehicular, and autonomous systems to be intelligent. Consequently, these systems will require an abundance of knowledge and processing capability to use that knowledge, likely to be beyond the capacities of storage and processing in individual systems (such as an individual robot or a car). Cloud robotics is "a field of robotics that attempts to invoke cloud technologies such as cloud computing, cloud storage, and other Internet technologies centred on the benefits of converged infrastructure and shared services for robotics" (Wikipedia). There are many research challenges associated with cloud robotics and its counterpart for autonomous vehicles. For instance, the cloud offers an ideal infrastructure for capacity-intensive optimization of vehicle maneuvers, as well as for collective robot learning. Large collections of robots could share experiences in the form of trajectories, initial and desired conditions, control policies, as well as data on resulting performance and effects of task execution. Robot learning would be greatly accelerated by this approach as would distributed reasoning. How to best utilize cloud resources for autonomous systems is one of many fascinating and challenging research questions that will be considered, with close connection and relevance to, for instance, smart manufacturing and Industry 4.0.

Big data and network science.

Data in all its facets have had far-reaching societal and commercial impact in recent years. This implies visions about unforeseen new possibilities for industry, society, and sciences. At the same time, big data is one of the biggest challenges for the society of the future. As a consequence, data-science centers are being founded all over the world. A special significance have data-analytics aspects, which likewise demand for novel automatic learning-based methods as well as novel visual data-analytics methods with the human in the loop. This entails the need for reliable and efficient analytics systems for monitoring and steering, for understanding and predicting the behavior of dynamic systems, and for interaction with large-scale and multilayered data. Such systems must be able to deal with missing or compromised data and inform the user about data uncertainty. Meeting all these demands requires large research efforts in all related disciplines.

In-situ analysis — moving the analytics to an early stage in the pipeline.
In many cases, the sheer size of the data generated in simulations makes it infeasible to save all the data and overwhelms post-processing analytics tools. In-situ analytics is an emerging processing paradigm, which performs analysis as the data is generated and has the potential to reduce the amount of data drastically. However, many of the traditional algorithms are not suited for the use on large-scale parallel machines and there are many open research topics. For instance, we see a special need in research on in-situ workflows and methodologies, in-situ data reduction and compression, derivation of measures for effectiveness and efficiency of in-situ algorithms, and predictive models for in-situ-analysis costs. A further challenge requiring research is to use the paradigm when it comes to exploration-oriented use cases.



Network and graph analytics.
One type of data that requires special attention is large graphs and networks. Next-generation networks are evolving into more and more smart, interacting complex systems posing increasing challenges on graph- and network-analysis and processing methods. At the same time, data integrity gets increasingly in the focus of society and has to be an integral part of every system.

In this field, we see many open research questions focusing on early detection of irregularities in large networks, which is crucial to ensure smooth and safe implementation of novel technologies. Promising concepts use learning methods for the detection of complex patterns or outliers. Advancing learning methods to meet the future demands with respect to data size, velocity, and variety, and tailoring them to the specific application, is one of the big research challenges in network sciences.

For model development, evaluation, and understanding, it is essential to make the data accessible to humans. This requires the development of hierarchical graph-abstraction techniques, maintaining important network characteristics. Reliable concepts for this purpose can be found in topological data analysis, which provide simplification strategies with rigorous mathematical guarantees. However, there are many open questions on how to facilitate topological concepts for effective and efficient graph and network analysis. As such, topological network analysis is a challenging research topic with the potential to push network analysis to the next level.

Visual analytics systems for decision making.
Such systems focus on data analysis with the human in the loop. They are an essential part of the data analysis pipeline complementing the automatic analytics. Visual representations are necessary for evaluating the performance of automatic systems, to gain insight about the hidden processes and correlations. They provide the interface between the human and the data and are at the core of every informed decision-making process. A research challenge for future visual data-analytics systems is the smooth integration of quantitative analysis, automatic analytics with open data exploration. Much research is needed in developing multi-scale representations that support easy navigation and exploration of the data. Machine-learning (ML) techniques have to be developed that can be embedded in interactive visualization systems. There are many open research questions related to the development of advanced visualizations of multi-modal, aggregated, and fused data. Efficiency is a basic and challenging requirement for all novel algorithms that are developed in order to guarantee real-time interaction for large and complex data.

Graph learning and signal processing.
Graph learning refers to the "reverse engineering" problem of inferring the topology of a complex network from observations of data, and to the learning of node attributes from partial observations. This class of problems in turn is relevant in a number of applications, ranging from biology (example: reconstructing a gene-regulatory network from gene-expression profiling), neuro-imaging (example: revealing the structural organization of the brain), engineering (example: localizing failures in power grids and computer networks), and analysis of social networks and networks of autonomous agents (example: inferring the strength of connection between individuals or agents based on observed actions or affiliation to the same services, inference of opinion dynamics and interactions among agents, and detection of communities based on this information). Future Internet-of-things (IoT) and 5G scenarios are based on principles such as high-density deployment of sensing, interfacing and computing devices. These smart devices are expected to produce large amounts of data, useful also to investigate the interaction graphs among agents involved in common tasks.

Recent algorithmic progress in graph learning has been facilitated by the availability of large-scale information ("big data") and availability of processing power, similarly to what is driving the deep-learning (DL) revolution. Urgent research is needed for the development of, for example, semi-supervised methods for node-attribute and graph-topology learning, as well as estimation of trust/mistrust among proximal agents, information that can be used to establish reliable interactions between them.

Communications and networks beyond 5G: sensors, IoT, and cloud.

Networks beyond 5G must handle entirely new requirements in terms of power and spectral efficiency, latency and reliability. Emerging future applications prompt for new physical-layer breakthroughs: large-scale adoption of augmented and virtual reality, massive machine-type and sensor communications, and large-scale deployments of Internet-of-things. In addition, the data deluge creates new opportunities and challenges for processing, analysis and exploitation of emergence of big data in networks.

Wireless access.
In terms of wireless access, massive multiple-input-multiple-output (MIMO) technology started as a wild academic idea about ten years ago and has since then become the main component of the 5G new radio (NR) physical layer. ELLIIT teams have had an internationally recognized leading role in its development and demonstrations.

The next revolution is likely to be the development of intelligent electromagnetic surfaces (also known as "cell-free" systems), that coherently interact with devices and sensors in its vicinity — an entirely new concept for wireless infrastructure. Interaction with these surfaces will enable communications and positioning with orders-of-magnitude better performance as compared to current technology, natural human-computer interaction and remote monitoring of life processes.

In this context, centralized, cloud-based radio access networks, with highly distributed antenna systems enabled by digital fronthaul interfaces supporting coherent large-scale baseband processing, must be developed.

Machine learning and AI techniques facilitating new applications.
The machine-learning revolution combined with the massive amounts of data enabled by 5G physical-layer technology (especially, but not limited to massive MIMO) will open up new possibilities for large-scale environmental sensing. Use cases include mobility prediction based on 5G access-point baseband data, detection and tracking of moving objects, short- and long-term prediction of user mobility patterns, estimation of traffic flows, counting the number of persons in a room, guarding against intrusion in protected spaces, remote health monitoring, and through-the-wall imaging.

Joint radar, communication and positioning methods and combination of RF data and point clouds obtained from images or video analytics may also be used to support positioning and large-scale sensing algorithms. Another application that may benefit from the technology is gesture recognition. Entirely new user experiences will be facilitated when for example gesture recognition is combined with tactile feedback (for instance via ultrasonic haptics) — with applications both in professional disciplines (for example health care, machine operation) and in entertainment and gaming.

Future IoT and machine-to-machine communications.
Future IoT solutions will rely on ultra-low-power sensor nodes, relying on back-scattering communications using radio-frequency (RF) energy harvesting, wireless power transfer and charging, and ultra-low-power wakeup radios. Array transceivers will be a basic enabling technology here.

Concurrently, cable-based solutions will be replaced by ultra-reliable large-scale wireless systems for real-time control of collaborative machines, vehicles, and swarms of autonomous robots and vehicles.

New frequency bands.
Another direction for beyond-5G systems that needs to be considered is the further increase of carrier frequencies for the wireless links to the lower THz and upper GHz range. While this sounds trivial and as an already solved problem, the ten-to-hundredfold increase in Doppler spreads and hardware limitations of energy-efficient electronics at those frequencies call for reconsideration of the physical layer techniques. For these reasons, we should study new approaches like orthogonal time frequency space (OTFS) where the data is transmitted/coded in the delay-Doppler domain, where the channel shows a more stationary behavior and the system is less sensitive to phase noise.

Ultra-robust connectivity and machine-learning algorithms — security aspects.
Machine- and deep-learning technologies can facilitate new algorithmic solutions for cases where accurate physical models are unavailable. But an important challenge is the brittleness of learning networks to adversarial attacks. Generally, adversarial examples are generated by finding small perturbations that can be applied to the input, in order to cause misclassifications. While mostly known in the computer vision field, this vulnerability affects all applications of machine learning. When ultra-critical connectivity (> 99.99999% reliability) is of concern, reliance on learning algorithms opens up a new security concern — especially because of the open (broadcast) nature of wireless channels, which facilitates easy insertion of adversarial signals. The study of adversarial attacks and the possible defense policies is a fundamental step towards the application of robust and resilient learning-based signal processing and wireless-communication algorithms.

Hardware-near signal processing and compensation for impairments.
The anticipated extreme proliferation of small (even disposable), ultra-low-power IoT hardware units will require hardware costs and power consumption to be driven down by an order of magnitude. As a consequence, low-end hardware will have to be used, which implies high levels of distortion (for instance from non-linearities) and noise. New algorithmic techniques are required to optimally achieve high-end performance with large numbers of low-end devices of this sort. Since the physical modeling of the devices is rather complicated (or potentially impossible at all) and the manufacturing variability among them is foreseen to be large, machine learning techniques may be a fundamental enabler for this emerging algorithmic technology.

Security, privacy, and trust.
These aspects must be integrated into the design of new systems, rather than added as add-ons. Security threats concern eavesdropping, jamming and spoofing; these problems have escalated in the last decade and are likely to further increase in magnitude. The increased reliance on machine-learning solutions itself constitutes a new class of security threats, as especially deep-learning networks are known to be extremely brittle and susceptible to adversary attacks, a threat that is currently only vaguely understood. Privacy is another major and spiraling concern, especially with the "surveillance as business model" that drives much of the Internet.

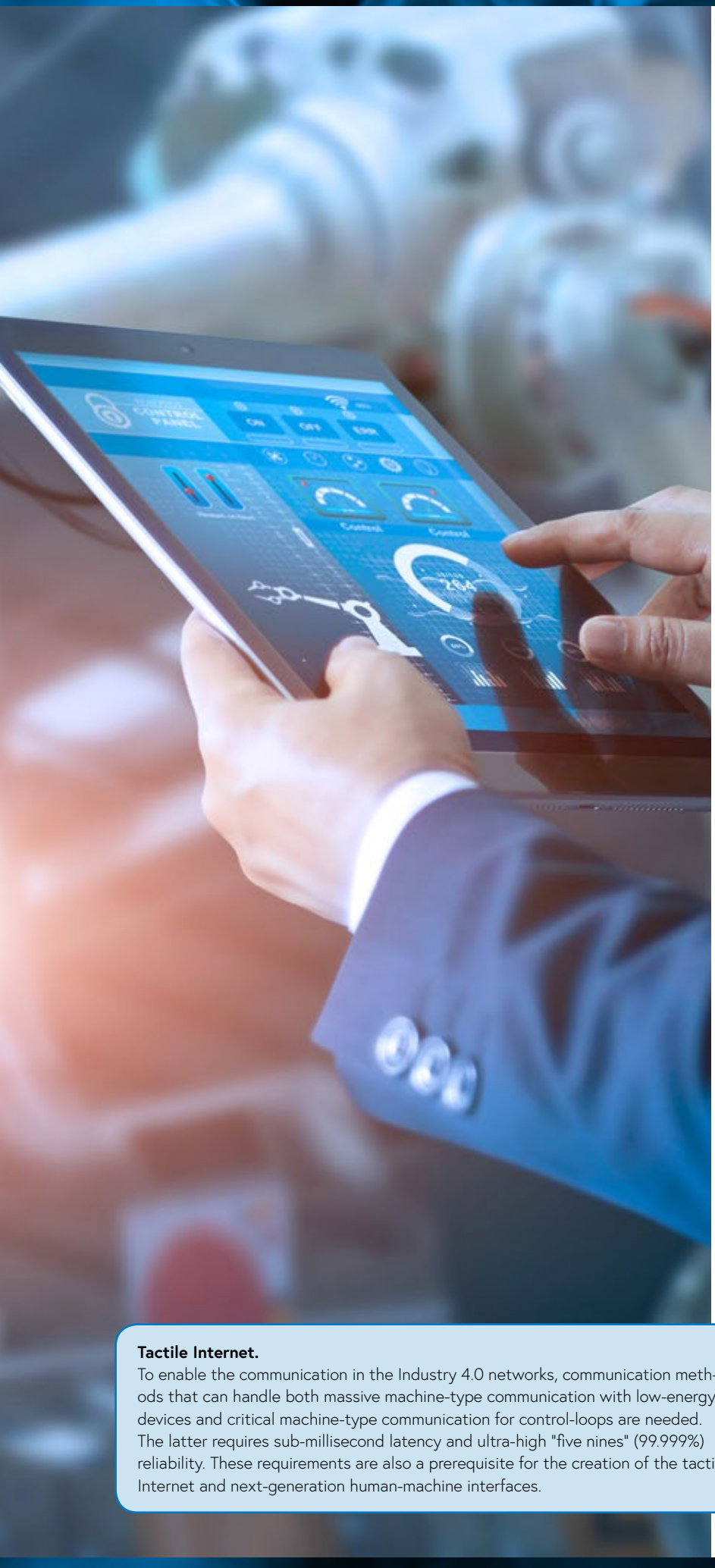
Distributed processing in energy-constrained wireless networks.
For networks with nodes having limited resources in terms of computational capabilities or battery power it is essential to pre-process the data in a distributed manner at a node level to determine if there is information of interest at a central level, and if so, to extract the essential data for transmission. This calls for new joint processing-communication approaches enabling extreme low-power nodes with sensing, computing, storage, and communication capabilities.

Future cloud infrastructures.
The networked society, empowered by hundreds of billions of connected devices, fundamentally changes the way we communicate and compute. Currently, the limited computing and storage capacity of end-user and IoT devices are complemented by remote data centers. However, there are limitations on what kind of interactive and real-time applications that can be deployed on today's cloud due to its inability to provide guaranteed end-to-end performance with low and jitter-free latency. Distributed cloud and fog architectures in combination with 5G radio access is a proposed remedy for this. A fog-computing infrastructure is hyper-distributed and resource-heterogeneous, ranging from user-proximal data centers or edge nodes at the network edge to traditional distant data centers. The edge data centers are typically small data centers associated with a base station, a production plant, some transportation system infrastructure, or an office building. In the fog, software applications can be dynamically deployed in all types of network nodes to meet their individual performance goals. Dynamic resource management, orchestration and network-function virtualization are areas to be studied in this context. Other research topics of interest are programming and communication models as well as software-defined infrastructures for both data-center hardware and networks.

Industry 4.0.

The term Industry 4.0, also known as smart manufacturing, refers to a further development stage in the organization and management of the entire value chain in the manufacturing industry. The base for this further development is the last decade's evolution of the information and communication technology (ICT) that has paved the ground for creation of smart factories. This means factories where equipment, machines, and robots collaborate seamlessly and respond autonomously to unscheduled disruptions in the production process or changes in business opportunities. Beyond that, the smart factories will have the capability to interface with each other as well as with other infrastructure in the society, and share knowledge and insights about how products are used and operate throughout their full life cycle. This will enable new value chains, disruptive business models, more rapid product development, customized production, more efficient supply chains, better use of production resources, holistic life-cycle management, and similar.

There are many challenges, both technical and non-technical, in realizing the ideas of Industry 4.0 and smart manufacturing. Non-technical challenges exist in the areas of economics and management (new business models), human resources (lack of specialists), law (clarification of rights and obligations), education (training courses and renewed curricula), and similar. The technical challenges span across a broad spectrum of areas such as standardization, data management, data security, network dynamics, artificial intelligence and machine learning, and similar. Five of the most important technical challenges are found within the following areas.



Distributed machine intelligence.
To enable scalability and the development of low-cost robots and machines with a high level of intelligence, methods for distributed machine intelligence and artificial intelligence across devices, systems, and the cloud are needed.

Distributed data analytics and machine learning.
To handle the exponential growth of sensorized and digitalized devices, distributed and hierarchical data analytics and machine-learning methods to transform large quantities of heterogeneous and spatially and temporally distributed data to actionable and concentrated high-level system information are needed.

Secure data sharing and privacy.
Tomorrow's businesses will need to cooperate with partners and suppliers to collaboratively design new products, or to optimize supply chains by sharing production lines and facilities. Therefore, methods for securely sharing data among trusted parties and other classes of users, that can perform certain operations on the encrypted data and view results using homomorphic encryption technologies, are needed.

Standardization.
Standardization is crucial part of the development of Industry 4.0, as it involves a high number of stakeholders and cannot be realized by proprietary solutions from one part alone. Hence, international and industrial collaboration is required to develop technical standards allowing stakeholders to share information and data amongst each other, without removing individual stakeholders' competitive advantages or revealing their internal technical implementation. Such standards are imperative for the success of industrial wide-scale realization of the Industry 4.0 and smart manufacturing concept.

Tactile Internet.
To enable the communication in the Industry 4.0 networks, communication methods that can handle both massive machine-type communication with low-energy devices and critical machine-type communication for control-loops are needed. The latter requires sub-millisecond latency and ultra-high "five nines" (99.999%) reliability. These requirements are also a prerequisite for the creation of the tactile Internet and next-generation human-machine interfaces.

Intelligent assistants and tools.

The digital era requires the automatic collection of knowledge and automated decision support in the form of intelligent assistants and tools that can analyze and visualize the massive amount of information available and constantly produced, in databases, in social media, and on the Internet. This precise information harvested through data analysis consists of terabits per second of streaming data that needs to be reduced to megabits per second for human interaction and decision making. In order to solve this technical and societal challenge, it is of the highest importance to develop decision-support systems, which adaptively reduce the cognitive load caused by large and rapid information flows while ensuring application-dependent mission-critical decision time-scales. Intelligent assistants and tools will thus be needed in several scenarios involving humans such as centralized decision-support arenas frequently manifested as screenscape environments containing novel multisensory interfaces. Another scenario is personalized and contextualized decision support in the wild/field, based on hand-held and wearable pervasive technology. A challenge is to enable full collaborative decision support, which interlinks different flavors of decision support involving human and cyber-physical agents in the field and operational centers for analysis and operational steering on one or several locations.

The challenges involved in enabling the scenarios described above span across several disciplines. Below we give examples of such project areas in which ELLIIT has unique possibilities to spearhead development of new technologies.

Supporting human-level interaction — cognitive companions.
In many applications, remote autonomous systems will be embodied as avatars. In training, learning, or personalized health care, the avatar could be manifested as a photo-real digital human. In other applications, such as big-data analysis or UAV-fleet management, the avatar, or cognitive companion, represents visualizations of multi-source aggregated data, or even advanced instrumentation in a driver environment. There has been renewed interest in the use of immersive technologies for data exploration. This has been spawned by the renewed interest in VR/AR and the improved possibilities offered by modern GPUs and improved display and tracking.

Upcoming research will have to deal with new software architectures that can form dynamic application-defined networks with transparent use of underlying highly heterogeneous and changing physical networks. These architectures further need to support high-level context-aware composition of autonomous services, evolvability of the software, and flexible security, supporting secure communication in the presence of dynamically changing networks and components.

New protocols for interaction with collaborative multi-agent autonomous systems at variable levels of autonomy will also be a challenge. This will require formal specifications and techniques for interaction processes in terms of delegation, contracts, speech acts, and other protocols as well as formal models for shared tasks associated with joint planning and decision making.

In addition, research will have to focus on new visual representations that are tailored to the needs of the application at hand and that make use of the immersive dimension to add clear value. For data analysis and decision support, immersion needs to be stable, controlled and reproducible. There is thus a clear need to develop perceptual metrics for assessment of visualization and interaction quality in immersive environments.

Also, investigations of how explorative visualization affects human cognitive structure and experience, and the limits of the massive information that is necessary to integrate in order to understand complex data, will have to be performed.

New visual representations for streaming heterogenous data.
The range and heterogeneity of data will require new autonomous methods for choosing visual metaphors based on ontologies of visualization methods, and new data-dependent, adaptive, and suggestive visualization protocols. As mentioned above, there is a lack of visualization targeted towards representing dynamic data.

Research will have to focus on new representations that effectively make use of the strengths of human perception in 4D and are tailored to the spatio-temporal resolution of human perception. An interesting notion that can be exploited is the strong spatio-temporal correlation that most data sets exhibit. This is an inherent feature as the characteristic time scale at which interesting events occur are correlated to the speed of information propagation. Thus, regions of interest and level of detail in space and time are strongly correlated.

New adaptive data representations that are trained (on-line or off-line) on the data under constraints enforcing sparsity are also of great importance. Sparse representations of visual data not only lead to high compression ratios and efficient processing, but also enables the application of novel data analysis and machine-learning approaches. Another aspect of sparse representations is that we can apply theory from compressive sensing to significantly improve the efficiency in the image synthesis.

In addition, research will have to deal with protocols and instantiations of ontologies for visualization methods and data representations, which given the range and heterogeneity of data select visual metaphors to be used in visualizations.

Supporting semantic-level communication.
A very large part of our knowledge lies in text, calling for research in natural-language processing (NLP). This research area is rapidly advancing and has pioneered many of the techniques in deep learning as well as in software tools and techniques for parallelizing computations. High-tech American and Chinese companies like Google, Facebook, Tencent, and Baidu are dominating the current development, and it is very important from a national perspective to have strong players in this research, from many perspectives, including economy, democracy, and societal security.

A challenge for future research in this area lies within multilingual techniques to support low-resource languages, like Swedish, for example, using multi-language sources like Wikipedia. Another challenge concerns semantic analysis of text (entity resolution and extraction of relations) to build universal knowledge graphs from very large multilingual corpora: Wikipedia, encyclopedia, newspapers, and similar, potentially including all the written works produced to date.

Multilingual question-answering systems to use these graphs and answer any question is also a focus area for research, as well as handling social-media content with intricate interconnections, non-standard spelling, code switching (alternating between different natural languages in a single conversation), and similar. Further, research will need to focus on new NLP techniques for specific analyses like sentiment analysis and context-aware analysis, together with much improved parallel software pipelines for data acquisition as well as for data processing, promoting easy experimentation.

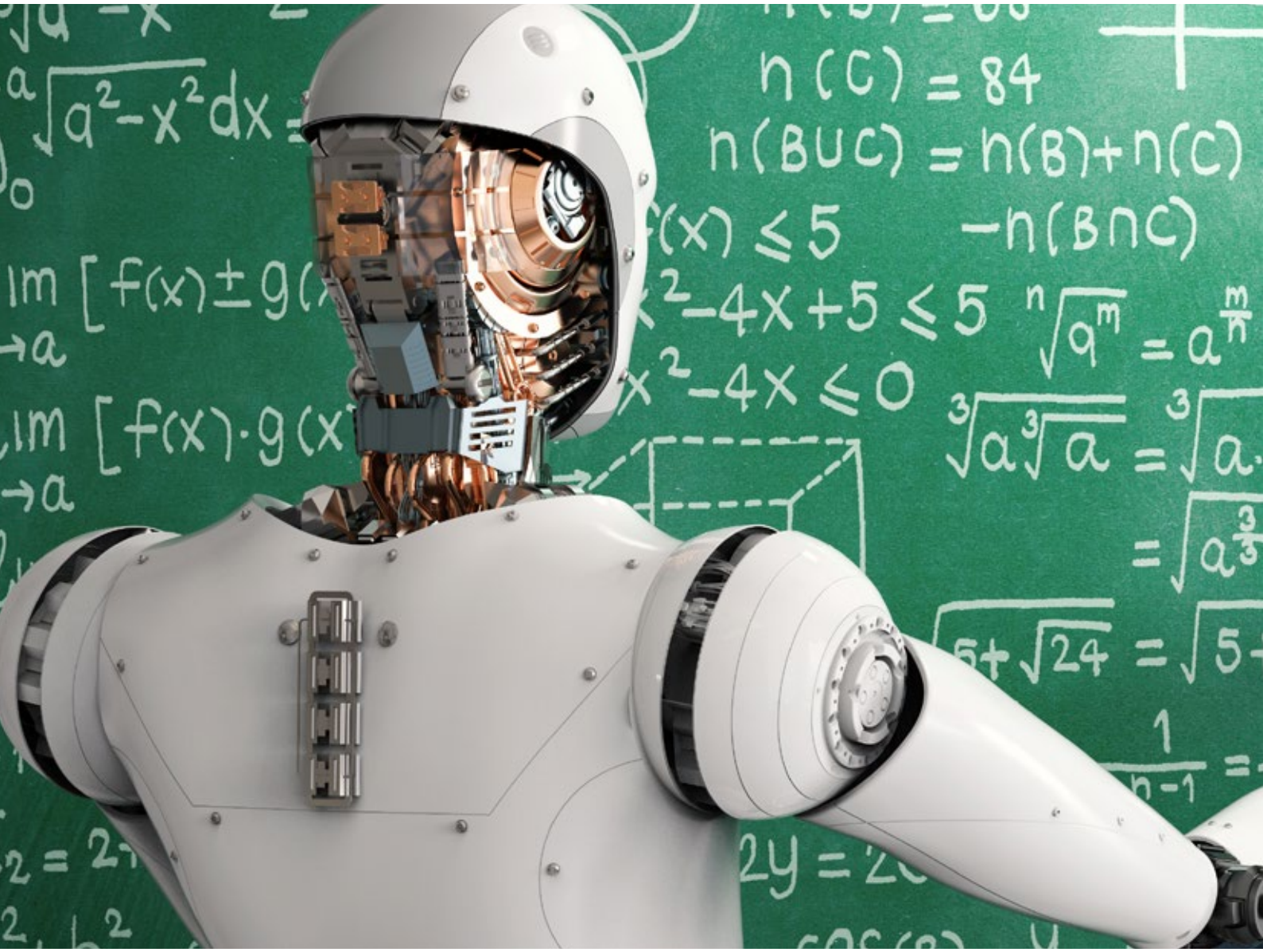
EMERGING RESEARCH THRUSTS, TECHNOLOGIES, AND CHALLENGES

AI, large-scale algorithms, machine learning, deep learning, and XAI.

Several areas within data science such as natural-language processing and computer vision have recently seen major progress due to novel methods within machine learning and deep learning. Specifically, computer vision has seen a scientific revolution since 2012, where deep learning has enabled major progress. At the same time, computer vision has been one of the most important success stories to deep learning. In general, most success stories of machine learning have occurred in areas where humans are naturally strong, such as reading sign language or playing Go. Progress has been more moderate in areas where computers were already doing well, such as playing chess and statistical inference for isolated problems.

It is sometimes wrongly assumed that progress within deep learning has been triggered solely by the availability of GPUs and big data. Although core elements such as back-propagation and convolutional layers existed before, many recent mathematical and algorithmic concepts, as well as novel topologies, have been essential for the success of deep learning. In fact, classical research fields such as statistics and optimization theory have during recent years developed in strong synergy with applications in learning and data processing.

The rapid growth of the research area is expected to continue and will be of strategic importance for Sweden during the next decade and even further. ELLIIT researchers are well positioned to contribute to theory and methodology for AI and machine learning in several directions.



Perception-action learning.
The approach of perception-action learning has been around in machine perception for more than 20 years. It has strong ties to research on adaptive control, which has been a stronghold of Sweden since the 1960s. Its main idea is inspired from biology and makes systems observe the effect of their own actions to their environment, integrating ideas from reinforcement or feedback learning, cognitive bootstrapping, embodied learning, affordances, predictive coding, and random exploration. Due to recent progress in machine learning and to tighter collaboration across research communities, the approach of perception-action learning will likely see major progress in cyber-physical settings as well as purely virtual settings. Integration with control engineering will enable new applications with hard real-time constraints.

Verification, confidence levels, security, integrity, and originality.
These aspects of deep learning address criticism to machine learning raised in the past and also during the revolution of deep learning: How can we know how reliable a deep network is (verification)? What can the network tell us about the its confidence in the output (confidence levels)? How can we be sure that the output has not been modified in some unintended way (security)? How can we limit the amount of personal data in the internal models (integrity)? How can we differentiate between real data and generated/adversarial data (originality)? The first two questions are important to many practical applications and the latter three are at a meta-level, having a major influence on the acceptance of AI/ML systems by society.

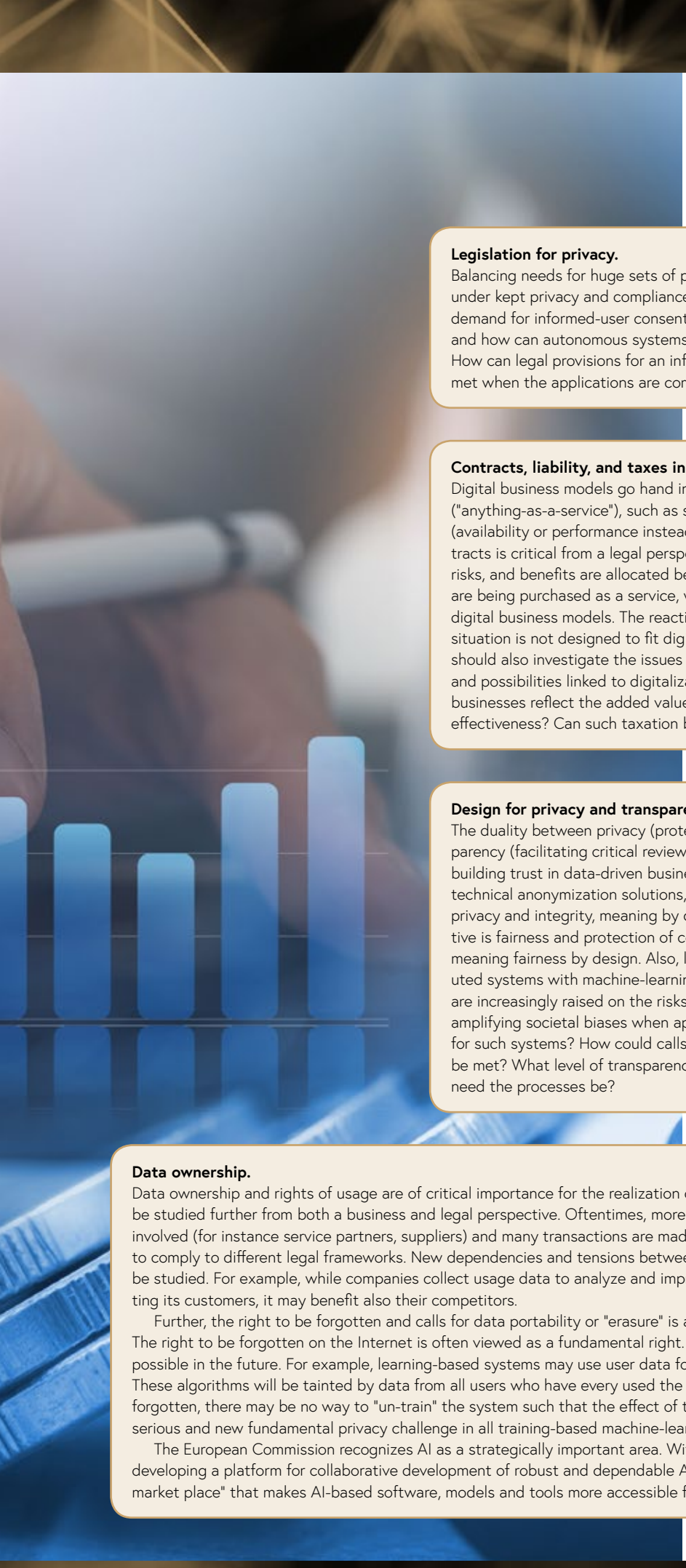
Computational aspects of ubiquitous AI/ML.
AI/ML will be omnipresent, meaning ubiquitous, both as inference machinery but also for incremental and distributed learning. Current hardware demand does not scale well with this development, neither economically, nor from an energy perspective, nor from the perspective of raw materials. Current research focuses on reducing the bit depth of calculations and variables, but in the long run, the overall network optimization must be broken down into learning atoms and alternative paradigms, for example spiking networks. The atoms enable not only distributed optimization, but also re-use of partial results across networks and possibly domains. However, the recombination of these atoms poses substantial problems regarding the applied techniques for partial result representation and its algorithmic processing. Also, new hardware needs to be developed, for instance generally available TPUs and NVRAM-based systems, that consume less power and achieve at least GPU-level performance.

General optimization algorithms for large-scale data.
The rapid development of machine learning algorithms during recent years has to a large extent been driven by experimental research. Mostly, however, the mathematical foundations of the success are still poorly understood, which is a major limiting factor for most of the research directions discussed above. A challenge for current research is therefore to distill the essential aspects of working algorithms and analyze their fundamental limitations. More often than not, this can be done using well established mathematics such as gradient steps, projections, randomization, and fixed-point iterations, but in a new context. For example, the theory for mass-transportation problems dating back to Gaspard Monge in the 1780s has recently become an efficient tool for analysis of statistical learning algorithms. Similarly, fixed-point iterations of deep learning are being analyzed and improved using calculus of variations. An open problem is also the computational integration of learnt and physically-inspired layers in deep networks (for instance resulting in backpropagating gradients through relaxations or entire optimization layers) so that problem-domain constraints are supported and generalization is achieved with considerably less training data and under weaker forms of supervision.

Digital business models and legal aspects.

The digital transformation has disrupted several business domains, such as the music and tourism industries, and is making traditional industry boundaries increasingly blurred. By harnessing digital and data-driven technologies, companies can create novel business models or reconfigure existing ones, thereby enabling new mechanism for revenue generation and value creation. Implications include new ways of doing business at a new scale, for example new business ecosystems, modes of collaboration, and customer experiences (for example Spotify, Uber, Airbnb, and other "platforms") — and challenged legislation (for instance copyright, the labor and tax regulation of the taxi sector, and hotel regulations). As a counter reaction, new legislation (currently GDPR) forces businesses to change their digital and data-handling practices to preserve user privacy and enable more trust in data-driven European markets.

There is an interplay between the technical solutions, business models, and legislation that must be explored in combination as a system, rather than addressed one by one. Technical solutions, such as blockchains and digital contracts, may be business enablers but never solve the business challenges alone. Further, technical solutions may be designed to align with legislation, community values and social norms, in contrast to adding regulation as an afterthought. Increasingly, the digital platforms find themselves mitigating both community norms and regulatory jurisdictions as a design feature, demanding them to take normative stances and train algorithms to implement policies (for example Facebook, Google Search, Twitter). Research and practice are in their infancy in addressing this interplay, and there is a need for both observational studies of practice and design science to address how to mitigate these challenges. Extending digitalization into new societal and industrial sectors, research challenges to be addressed in cross-disciplinary constellations include the following aspects.



Legislation for privacy.
Balancing needs for huge sets of personal data to train machine-learning systems under kept privacy and compliance with, for instance, GDPR, and particularly the demand for informed-user consent. How does the legal practice of GDPR evolve, and how can autonomous systems be designed to adhere to the legal practice? How can legal provisions for an informed consent-based use of personal data be met when the applications are complex and autonomous?

Contracts, liability, and taxes in digital business.
Digital business models go hand in hand with new service-based business models ("anything-as-a-service"), such as subscriptions and outcome-based contracts (availability or performance instead of physical goods). The design of such contracts is critical from a legal perspective since the contract stipulates how costs, risks, and benefits are allocated between the parties. As more and more offerings are being purchased as a service, we should study the legal implications of such digital business models. The reactive nature of law implies that the current legal situation is not designed to fit digital business models. For these reasons, research should also investigate the issues of accounting, taxation, and insurance needs and possibilities linked to digitalization. For example, how can taxation of digital businesses reflect the added value under kept incentives to improve business effectiveness? Can such taxation be technology-independent?

Design for privacy and transparency.
The duality between privacy (protection of individuals' information) and transparency (facilitating critical review of how decisions are made) is at the core of building trust in data-driven businesses and autonomous systems. This requires technical anonymization solutions, designed to enable data utilization under kept privacy and integrity, meaning by design and not by regulation. Another perspective is fairness and protection of consumers' rights under reduced net neutrality, meaning fairness by design. Also, liability for autonomous, geographically distributed systems with machine-learning algorithms requires transparency. Concerns are increasingly raised on the risks of AI and ML possibly reproducing or even amplifying societal biases when applied. What are the legal and moral implications for such systems? How could calls for algorithmic transparency or accountability be met? What level of transparency is called for, and where? How "explainable" need the processes be?

Data ownership.
Data ownership and rights of usage are of critical importance for the realization of digital business models and should therefore be studied further from both a business and legal perspective. Oftentimes, more parties than only the buyer and the seller are involved (for instance service partners, suppliers) and many transactions are made between international parties that may have to comply to different legal frameworks. New dependencies and tensions between parties is another consequence that should be studied. For example, while companies collect usage data to analyze and improve products and processes, thereby benefitting its customers, it may benefit also their competitors.

Further, the right to be forgotten and calls for data portability or "erasure" is a significant design challenge to be explored. The right to be forgotten on the Internet is often viewed as a fundamental right. But it is not clear if this is even going to be possible in the future. For example, learning-based systems may use user data for training, and for build-up of trained algorithms. These algorithms will be tainted by data from all users who have every used the system. In case a particular user requests to be forgotten, there may be no way to "un-train" the system such that the effect of this user is removed from the system. This is a serious and new fundamental privacy challenge in all training-based machine-learning systems on-line.

The European Commission recognizes AI as a strategically important area. Within Horizon 2020, for example, researchers are developing a platform for collaborative development of robust and dependable AI systems. A long-term goal is to establish an "AI market place" that makes AI-based software, models and tools more accessible for industry.

Next-generation software technology.

Networked and mobile software applications are central to future digitalization, both for industry and society. Today's cloud-centric solutions are gradually starting to be complemented with edge-centric computing as edge devices become more powerful and programmable. These applications call for a new generation of software technology that can handle a number of challenges that current research and practice have only started to address — and it needs to provide means to deal with the increasing complexity of systems.

Robustness and resilience.
Next-generation software technology must support robust and resilient applications in the presence of dynamically changing networks and mobile devices and subsystems, unreliable networks, and communication failures due to both benign and malicious influences. Support for high-integrity and zero-downtime software deployment, upgrade, and modification is required, and it should be possible to initiate and validate such operations remotely, as devices will be too numerous for human intervention and may be physically unreachable.

Safety and security.
It will be of critical importance to eliminate security and safety concerns by design to as large extent as possible, minimizing possibilities for both attacks and bugs. At the same time, next-generation software must be able to interoperate across highly heterogeneous networks, devices, operating systems, programming languages, and APIs. It will also be required to provide foundations for modeling, reasoning, and analyzing software systems operating within probabilistic and continuous spaces, so as to enable rigorous integration with communication and mobility.

Control and testing.
The increasing use of AI will pose significant challenges to software quality control and testing, as AI systems are expected to learn and to evolve with higher degrees of freedom than previous software technologies. Next-generation software technology must include novel methods for specifying, testing, securing, verifying, and ensuring the accountability of software that employs AI, as well as techniques for employing AI in the construction of new software.

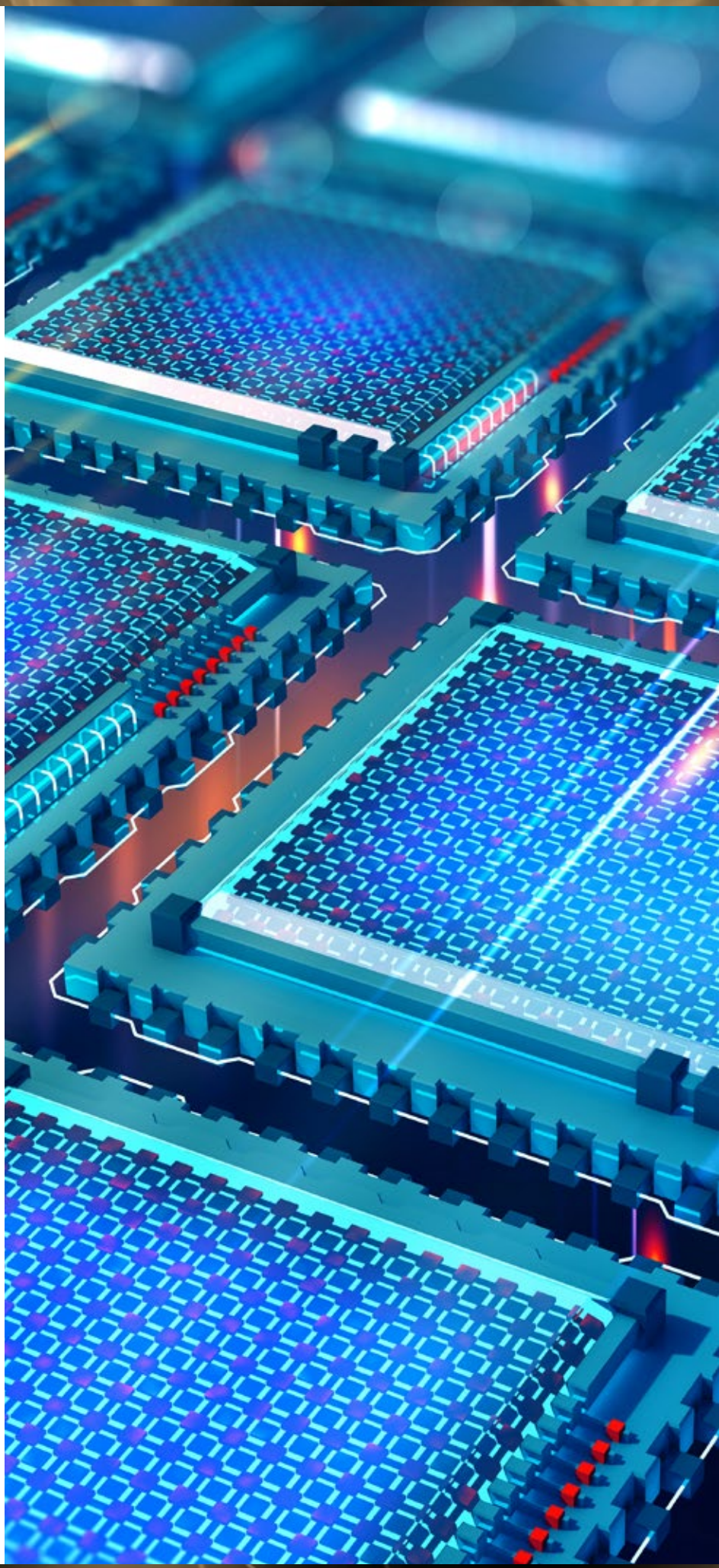
Languages, tools, methods, and architecture.
Next-generation software technology will need to meet future expectations, demands and challenges through the development of new languages, tools, methods, and software architectures that support the safe, easy, and flexible programming of such mobile intelligent applications. Examples include agile methods, distributed immutable data structures (like blockchains), reactive programming constructs, container technology, and static and dynamic analysis, to just name a few.

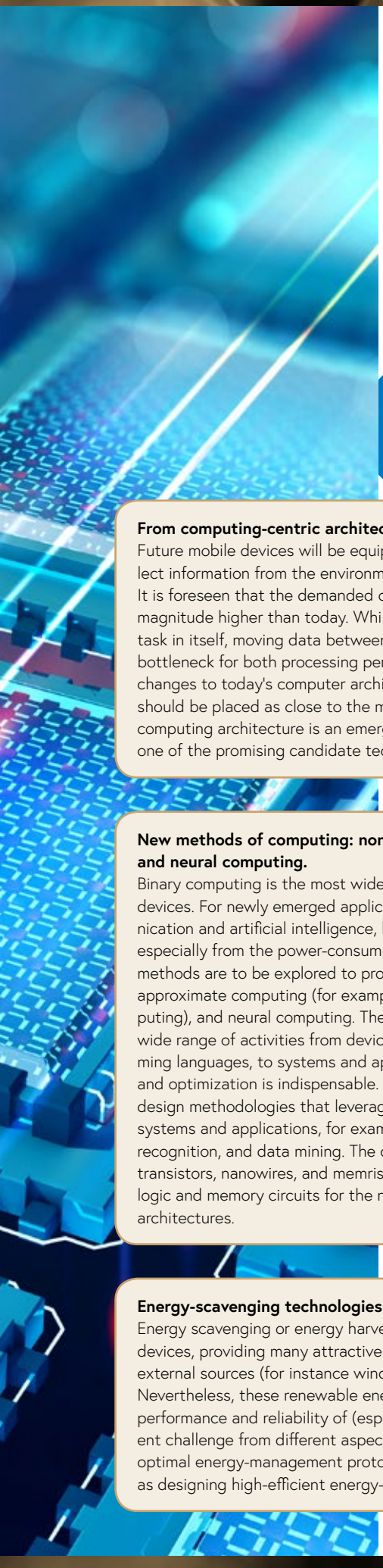
Short development cycles.
The ability to experiment with systems that are already deployed and in operation requires short development cycles.

Mobile processing architectures and devices.

The proliferation of communicating and computing technologies that is anticipated will pose several challenges. For the number of devices anticipated in the future, deploying a system that is deemed "low-power" by today's standards would entail unreasonable levels of energy consumption. Increasing connectivity between devices not only multiplies possible failure modes but also opens the door to emergent operational behaviors at scales never seen before. The longevity of systems poses further challenges not only for standardization processes but also for upward compatibility, security, and hardware/software co-design. Finally, if unmanaged, devices that reach obsolescence can entail a significant environmental liability.

To meet these emerging challenges and to enable the envisioned Internet of everything, intelligent devices, mobile virtual reality, and true autonomous systems in the future, we need significant advances in mobile processing architectures and devices. Reducing power consumption while keeping high performance will continue to be of paramount importance both in hardware and in transmission. New methods will be needed to rigorously verify the security, reliability, and predictability of (all levels of) communication architectures in a variety of physical contexts. New standards, software architectures, security protocols, hardware, and compilation methods will be needed. Sustainability must be addressed through technologies for planned obsolescence such as bio-degradable ICT components or reliable methods for sequestering and recycling ICT components embedded in obsolete products. More specifically, research efforts and innovations are needed from (but not limited to) the following areas.





From computing-centric architecture towards storage-centric architecture.
Future mobile devices will be equipped with a massive number of sensors to collect information from the environment and to extract intelligence via data mining. It is foreseen that the demanded on-chip memory capacity will be several orders of magnitude higher than today. While integrating large-size memory is a challenging task in itself, moving data between memory and computing units is becoming the bottleneck for both processing performance and power consumption. Fundamental changes to today's computer architecture are necessary, and the data processing should be placed as close to the memory as possible. This new storage-centric computing architecture is an emerging research area and in-memory computing is one of the promising candidate techniques.

New methods of computing: non-binary computing, approximate computing, and neural computing.
Binary computing is the most widely used computing method for today's mobile devices. For newly emerged applications like new generations of wireless communication and artificial intelligence, binary computing starts to show its inefficiency, especially from the power-consumption perspective. New types of computing methods are to be explored to provide application-specific optimization, including approximate computing (for example voltage overscaling and stochastic computing), and neural computing. The research on new computing methods spans a wide range of activities from devices, circuits, computing architectures, programming languages, to systems and applications, where very tight cross-level design and optimization is indispensable. For instance, approximate computing employs design methodologies that leverage the unique feature of error resilience in many systems and applications, for example wireless communication, machine learning, recognition, and data mining. The development of emerging devices, such as new transistors, nanowires, and memristors, should also be exploited to form efficient logic and memory circuits for the new computing and processing methods and architectures.

Computing architectures and programming models for high-dimensional data structures.
New applications constantly ask for computing devices to provide a higher degree of parallelism to support operations such as matrix computation, 3D filtering, and similar. The computing-architecture evolution from the current vector machines to higher-dimensional processors requires joint research efforts in the area of processing elements, storage units, and interconnection technology. At the same time, the model of the sequential-instruction-set computer has been a very powerful abstraction. Yet there is no common computing model tying different parallel architectures together, and consequently there are no common software tools and no common programming languages. There is an urgent need to bridge the gap between massively parallel hardware architectures and programming models, and thus tools for software development.

Energy-scavenging technologies.
Energy scavenging or energy harvesting can potentially reduce the dependence on the supply of battery energy for mobile devices, providing many attractive benefits from the deployment and maintenance perspectives. Energy can be harvested from external sources (for instance wind, sound, vibration, solar, thermal, and RF energy) or recycled from internal energy leakage. Nevertheless, these renewable energy sources are intermittent and unpredictable, making it very challenging to maintain the performance and reliability of (especially battery-free) mobile devices. Further research is indispensable to address this inherent challenge from different aspects and levels, including understanding variable energy sources and their features, developing optimal energy-management protocols, scheduling energy harvesting and energy consumption in a cooperative manner, as well as designing high-efficient energy-reception, harvesting, and translation circuits.

Design for security, privacy, and trust.

The ongoing, and in some domains upcoming, digitalization will result in a significant increase in digital data. This data has to be generated, transmitted, stored and analyzed, and through all these steps there are several security-related challenges, but also opportunities. In many cases, the data can be used to improve, for instance, production processes, transportation systems, and our health, and contribute to a more sustainable society. These positive effects are driving the development, but the rapid development has caused severe security problems. Malicious data can cause erroneous decisions in processes and systems and will have severe consequences when the data is used to supervise and control critical functions, infrastructure, and factories. IoT devices are expected to significantly grow in numbers in the near future, and protection of these devices, for instance against jamming, eavesdropping or spoofing attacks, will be essential to the trust we can afford to put into data and functionality provided by IoT devices. This section presents specific areas, both related to IoT and focusing on other aspects as a consequence of digitalization, where we believe that research needs to pay particular attention in the next decade.

Lightweight cryptography.
NFC- and RFID-enabled devices, and also devices anticipated to run for an extended period of time on battery, must be very energy-efficient. Though there are several well-known and secure encryption algorithms, the need to focus on energy consumption is increasing. More data is being encrypted, we see a higher focus on sustainable solutions, and many devices operate without a continuous power source. It is thus important to develop, analyze, and better understand cryptographic algorithms that are suitable in environments that are very constrained. These algorithms must not only encrypt the data, but must simultaneously provide authentication functionality as well. Further, it must be possible to only authenticate a subset of the data, while the rest is both encrypted and authenticated, known as authenticated encryption with associated data (AEAD).

Post-quantum cryptography.
The encryption that we often use today has been analyzed and undergone scrutiny for a long time. How to build secure encryption algorithms is to a large extent well understood. A future problem is the development of quantum computers, which can be used to break many of the encryption systems that we use on a daily basis. The development of new algorithms that are secure also from attacks by quantum computers has been ongoing for about a decade, but our knowledge about these algorithms is still very limited in relation to what we know about currently used algorithms. We are relatively unprepared for the day when powerful quantum computers are reality, and we need to put more effort into understanding the algorithms that can resist attacks carried out using quantum computers.

Side-channel attacks.
In all computations, optimizations focusing on speed and efficiency have always been a primary concern. Sharing of hardware resources is necessary in systems running several concurrent processes, while also supporting several users with different privileges. Encryption and authentication must keep a secret key, but devices are often still in the hands of potential adversaries. Side-channel leakage is information leakage that is not directly related to an algorithmic or implementation-related flaw, but instead uses properties of the implementation to deduce information from, for example, power, timing or electromagnetic information. Thus, even mathematically proven algorithms, and implementations free from classical software bugs, will be at risk if the implementation or hardware design allows for side-channel leakage. With more systems and devices, and with more sensitive data handled by these, we need to further focus on how to build hardware that can ensure separation between contexts of different privilege, and that do not allow for information leakage through power usage or execution time.

Spoofing and jamming attacks on the physical layer.
A wireless network may be attacked at its physical layer by jamming (intentional disruption of a link by transmission of man-made interference) or spoofing (insertion of false or misleading information into the data stream). This class of threats, traditionally considered a "military problem", is increasingly relevant to civilian systems. The principally important context is 5G and beyond. Wireless services other than communications are also subject to these threats, most notably navigation systems (GPS, GALILEO, Baidu, GLONASS). Spoofing and meaconing (interception and rebroadcast of navigation signals) attacks are becoming increasingly cost-effective; even if security services are in place, replay and physical-layer attacks are still possible. Array processing, statistical signal processing and machine learning are fundamental enablers for the development of low-cost, easy-to-deploy, and effective countermeasures.

Distributed anonymous authentication.
Ledger and blockchain technologies are fundamental enablers for distributed authentication. While the concept has existed for some time, real implementations have been limited by practical factors such as high power consumption, and there are serious unsolved challenges. The use of these technologies may also require entirely new types of connectivity, that conventional wireless systems or 5G do not support.

Anonymity.
Newer, more-connected, and more pervasive generations of Internet-connected devices will enable services and functionality not previously possible, but also creating "the most effective mass-surveillance infrastructure" ever created. Due to their ability to store, process, and analyze data in the cloud, IoT devices will enable the collection and cross-referencing of large amounts of distinct data items. Consequently, new privacy threats will result. The eventual ubiquity of IoT sensors allows privacy violations on a new scale: sensors at work, home, and in public could lead to wide-scale losses of personal privacy. As 95% of individuals can be identified with merely four spatiotemporal points, it is critical to secure IoT systems and the information they collect. Arising from the creation and use of these systems, there exists opportunity for developing systems ensuring true anonymization of data. By fully anonymizing data, financial gain by fraudulent actors through obtaining and reselling sensitive data is lessened. Societal stakeholders should identify, create, and maintain systems where individual data is able to fuel complex research questions while anonymous enough to reduce the allure for fraudulent actors. Ethical considerations for ensuring true anonymity must be ensured by engineers and programmers while developing these new technologies and protocols.

Protection of personal data.
In future scenarios where technology will allow more immersive interactions with surrounding environments (for instance VR, AR) there will come the capability to acquire large amounts of data in order to identify anyone we walk past or interact with. Identification of private individuals will become ubiquitous. As consumers and citizens continue to rely on for-profit companies collecting and selling personal data in exchange for free or discounted services, the availability of sensitive data will only increase. Further, new types of data that can be considered sensitive personal data will emerge. As individuals' on-line and public presence continues to grow, citizens must be protected from social sorting systems based on undesirable sociodemographic identifiers. Already, some countries analyze an individual's aggregated social-media presence to develop financial credit systems where individuals' financial information is not available. As personal data continues to be a valuable, tradable commodity, stakeholders need to ensure citizens that their valuable personal data is sovereignly protected. Better, more robust systems are needed to ensure that full-control personal data is protected, regulated, and guaranteed in digitally connected systems where trust is paramount. Research informing citizens and consumers about digital privacy literacy will be of critical importance for consumers to be informed of how to lessen the availability of sensitive personal data, and therefore to enable individuals the decision to remain private or public.


ELLIIT
Excellence Center at Linköping – Lund
in Information Technology