What is Edge AI?

advian

CONTENT

What is Edge AI?

Simply put, Edge AI is a combination of Edge Computing and Artificial Intelligence.

AI algorithms are processed locally, either directly on the device or on the server near the device. The algorithms utilize the data generated by the devices themselves. Devices can make independent decisions in a matter of milliseconds without having to connect to the Internet nor the cloud. Edge AI has almost no limits when it comes to potential use cases. Edge AI solutions and applications vary from smartwatches to production lines and from logistics to smart buildings and cities.

How does Edge AI work, what type of business benefits does it bring and how can you get started with Edge AI? Read this page and find out - let's begin our journey to the edge!

👉 Edge Computing

Edge computing consists of multiple techniques that bring data collection, analysis, and processing to the edge of the network. This means that the computing power and data storage are located where the actual data collection happens. What is meant by the network edge? Well, depends on the use case - it could be a mobile phone, IoT-device, self-driving car or even a cell tower. Read more about edge computing from Samu's blog. **Read more about edge computing from Samu's blog**.

👉 Artificial Intelligence

Broadly speaking, in Artificial Intelligence a machine mimics human reasoning: such as understanding languages and problem solving. Artificial intelligence can be seen as advanced analytics, (often based on machine learning) combined with automation. This pragmatic definition covers all current AI applications.

You can think of Edge AI as analytics that takes place locally and utilizes advanced analytics methods (such as machine learning and artificial intelligence), edge computing techniques (such as machine vision, video analytics, and sensor fusion) and requires suitable hardware and electronics (which enable edge computing). In addition, location intelligence methods are often required to make Edge AI happen.

DOWNLOAD AN INTRODUCTION TO INTELLIGENT SAFETY EBOOK

Edge AI devices include smart speakers, smart phones, laptops, robots, self-driven cars, drones, and surveillance cameras that use video analytics.

Although most of the analysis and decisions made by machines already happen on the edge, the greatest benefits are obtained when the findings produced by the devices are linked to business processes. Therefore, **modern data platforms**, capable of handling large amounts of location and streaming data, are also needed to enable real-time computing.

> **Tip for the busy ones: there is a handy navigation pane on the left side. Use it to move straight to the section that interests you.** 👀

How Edge AI helps to generate better business

When edge computing is combined with artificial intelligence, we get an unbeatable combo.

Edge AI speeds up decision-making, makes data processing more secure, improves user experience with hyper-personalization, and lowers costs — by speeding up processes and making devices more energy efficient.

An example of this could be a hand-held tool used in a factory. The tool is embedded with a microprocessor that utilizes Edge AI software. The tool's battery lasts longer, when data doesn't have to be sent to the cloud. The tool collects, processes, and analyses data in real-time, and after the work day, the tool sends the data to the cloud for later analysis. A tool embedded with AI could for example turn itself off in the event of an emergency. The manufacturer receives valuable information about how their products are working and can utilize this information in further product development.

Latency

Data transfer to cloud and back takes time. This time, latency, is usually about 100 milliseconds. Often this is not a problem, but sometimes the response time requirement is so high, that even latency is too much. For example, new Porsches are equipped with hundreds of sensors that continuously produce massive amounts of data on how the car is operating. Porsches are embedded with NVIDIA's GPU processor and Kinetica's analytics software. The automation takes the wheel if needed.

If the car's speed is 200 kilometers per hour, latency of even milliseconds is too much. The decision to brake comes too late when the car is already in the ditch.

Real-time analytics

With edge computing, it is possible to reach near real-time analytics. Analysis takes place in a fraction of a second - which is crucial in time critical situations. Let's think about machines on a factory assembly line. If a robot on the assembly line is activated at a wrong time or too late, it may result in a damaged product or the product may move further on the assembly line unprocessed and untouched. If the mistake goes unnoticed, the faulty product may end up in the market or cause damage in later phases of the production.

Scalability

Research organisation IDC predicted that there will be 41.6 billion connected **IoT devices generating 79.4 zettabytes of data** in 2025. As volumes grow, new innovative ways for efficient analysis and data processing are needed.

When most of the data processing is done locally, on the edge, centralized service or data transfer won't become a bottleneck. Edge AI use cases typically involve large amounts of data. If you have to process video image data from hundreds or thousands of different sources simultaneously, transferring the data to a cloud service is not a viable solution.

Information security and privacy

Less data in the cloud means less opportunities for online attacks. Edge often operates in a closed network, which makes stealing information harder. Also, it is harder to bring down a network consisting of multiple devices.

Generally you can say that anything that has a security element, must be done on the edge. As an example, we can think about the intelligent safety monitoring systems in a factory. When machines are not working as they should or when people are moving in a prohibited area, the alarm should go off before the accident has even happened.

As already mentioned, when data processing happens locally, there is no need to send data to a cloud environment. Because of this, it becomes pretty hard to access data without permission. Also, sensitive data that is processed in real-time, such as video data, might only exist for a blink of an eye before it disappears. In these type of situations, it is easier to ensure data privacy and security, because the intruder should gain direct access to the physical device, where the data is being processed.

LEARN MORE ABOUT INTELLIGENT SAFETY SOLUTIONS

Automated decision-making

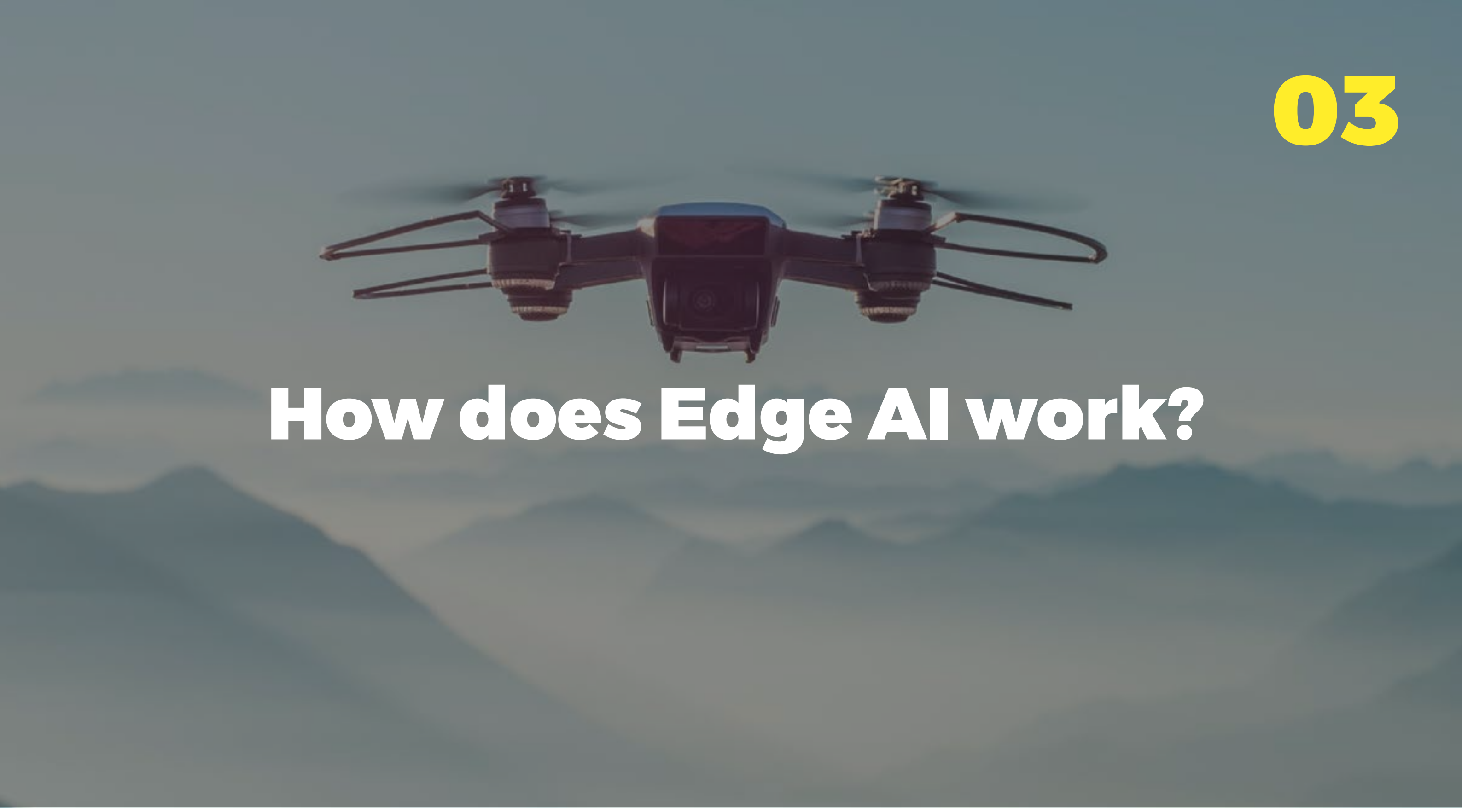
There are hundreds and hundreds of sensors in a self-driving car that constantly measure e.g. the position of the vehicle and the speed of tire rotation. The driving computer can make the necessary decisions regarding steering, braking and the use of throttle based on the collected data from the sensors - automatically.

Reduced costs

Due to scalability of analytics and reduced latency in making critical decisions, edge can bring significant cost reductions for your organization. In addition to time, **edge can save bandwidth** - the need for data transfer is reduced. This also makes devices more energy efficient.

Processing and analyzing large amounts of data in cloud is not cheap. If you want really fast response times when analyzing constant data streams or large amounts of historical data, you will have to buy a lot of capacity from a cloud service, such as GPU computing. Sometimes this turns out to be so expensive that it ruins business case calculations.

It goes without saying that Edge AI requires local computing power and investing in hardware, but even so, Edge AI is often the most cost-efficient solution.



03

How does Edge AI work?

What are the fundamentals and basic principles behind Edge AI? How does it actually work? Keywords here are machine learning, AI and edge computing.

In a typical machine learning setting, we start by training a model for a specific task on a suitable dataset. Training the model basically means that it is programmed to find patterns in the training dataset and then evaluated on a test dataset to validate its performance on other unseen datasets, which should have similar properties **to the ones that the model is trained on**.

Once the model is trained, it is deployed or in other words ”put to production”.

It can now be used for inference in a specific context, for example as a **microservice**. Inference refers to the process of using a trained machine learning algorithm to make predictions.

Once the model works as wanted, predictions produced by the model can be utilized in improving business processes. Typically, the model works via an **API**. The model output is then either communicated to another software component, or in some cases, visualized on the application front-end for the end user.

Cloud is not enough

The rise of cheap computing and data storage resources with cloud infrastructure has given new opportunities to **leverage machine learning at scale**. However, this comes at the cost of latency and data transfer challenges due to bandwidth limitations.

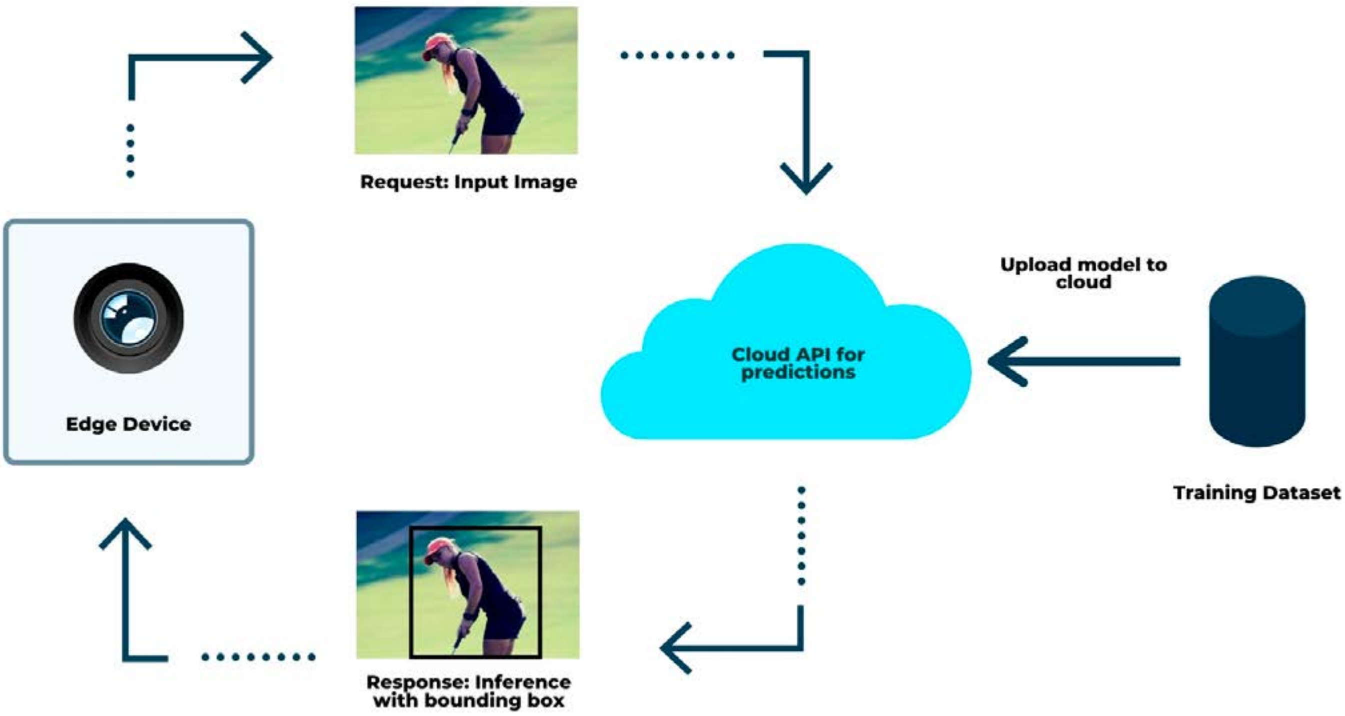
Training a machine learning model is a computationally expensive task well suited for cloud-based environment, whereas **inference requires relatively low computing resources**.

Given the new trends of **Industry 4.0**, autonomous systems and intelligent IoT devices, the old paradigm of executing inference in cloud is starting to get increasingly less suitable as needs for real-time predictions grow.

DOWNLOAD OUR FREE SMART FACTORY EBOOK HERE

If a machine learning model lives in the cloud, we first need to transfer the required data (inputs) from the end-device, which it then uses to predict the outputs. This requires a reliable connection and if we assume that the amount of data is large, the transfer can be slow or in some cases impossible. If the data transfer fails, the model is useless.

In the case of successful data transfer, we still need to deal with latency. The model naturally has some inference time, but the predictions also need to be communicated back to the end-device. It’s not hard to imagine, that in mission-critical applications, where low latency is essential, this type of approach fails.



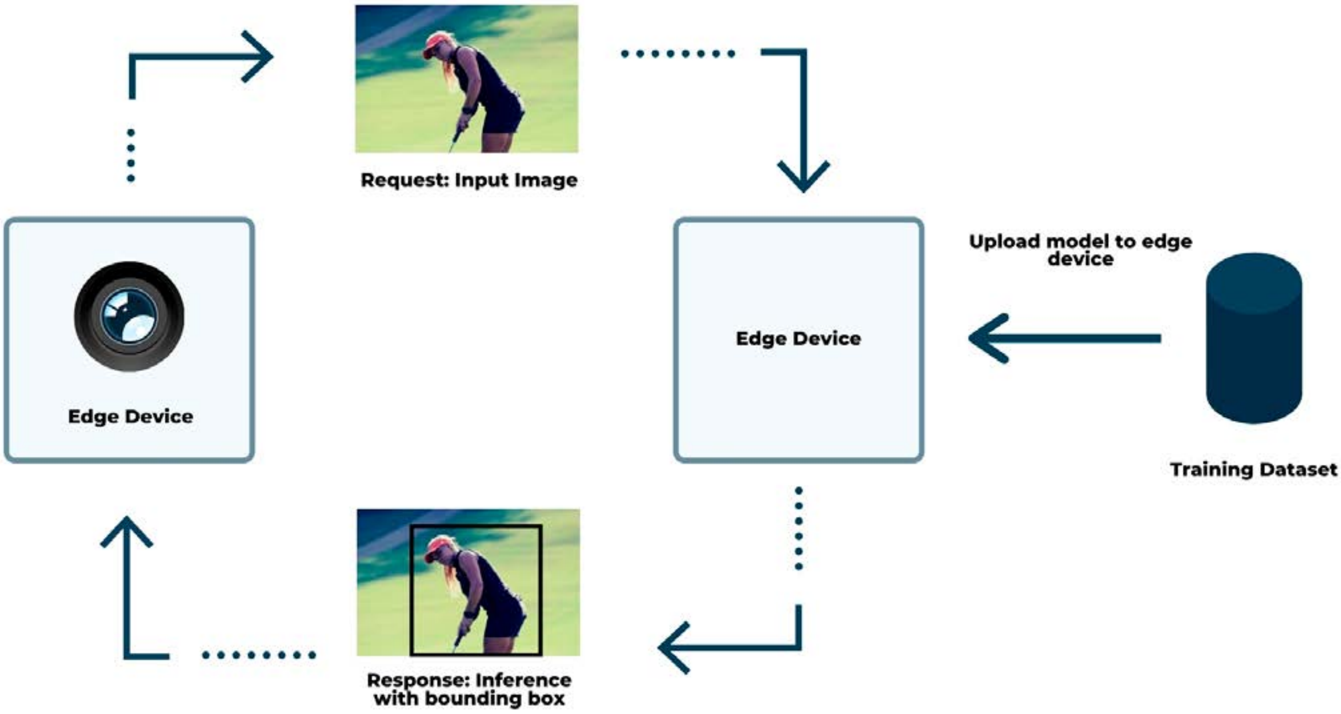
Computationally more powerful edge devices have enabled a new way of performing machine learning and artificial intelligence - Edge AI.

Mikä mahdollistaa Edge AI:n?

In the traditional setting the inference is executed in a cloud computing platform.

With Edge AI, the model works in the edge device without requiring connection to the outside world at all times. The process of training a model on a consolidated dataset and then deploying it to production is still similar to cloud computing though. This approach can be problematic for multiple reasons.

First, it requires building a dataset by transferring the data from the devices to a cloud database. This is problematic due to bandwidth limitations. Second, data from one device can not be used to predict outcomes from other devices reliably.

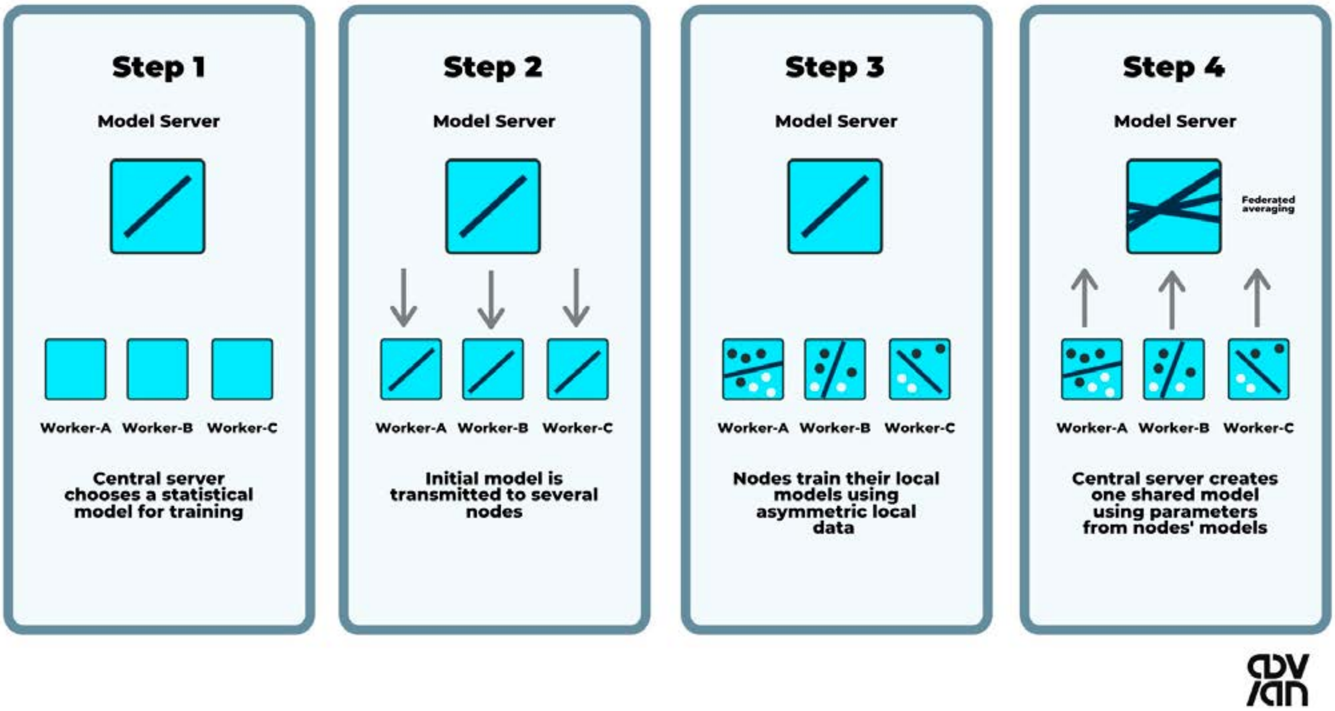

Request: Input Image
Edge Device
Edge Device
Upload model to edge device
Training Dataset
Response: Inference with bounding box

Finally, collecting and storing a centralized dataset is tricky from a privacy perspective. Legislative limitations such as GDPR are creating significant barriers to training machine learning models. Moreover, the centralized database is a lucrative target for attackers. Therefore, the popular statement that edge computing alone answers to privacy concerns is false.

For tackling the above problems, federated learning is a viable solution.

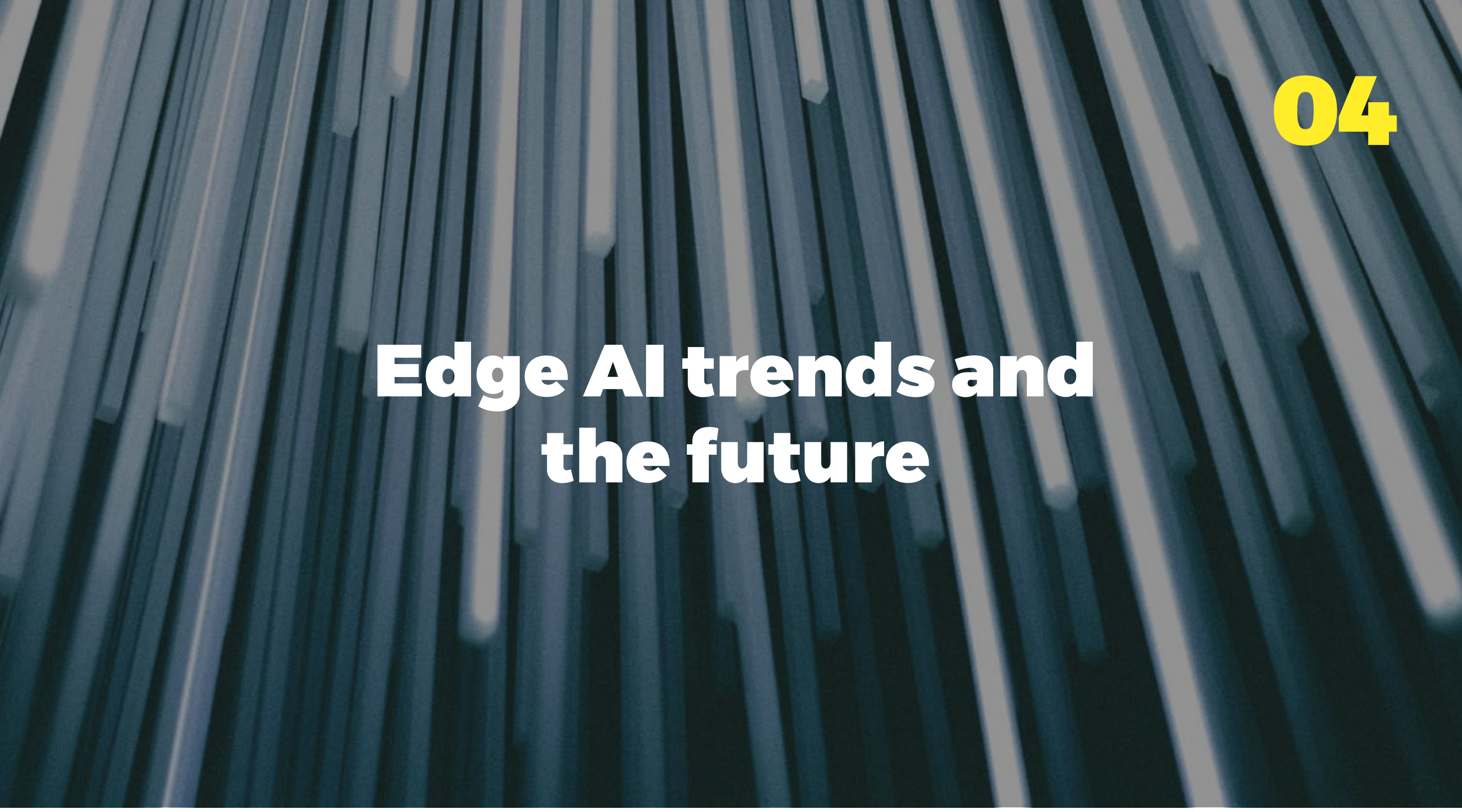
Federated learning is a method for training a machine learning model on multiple client devices without having access to the data itself.

The models are trained locally on the devices and only the model updates are sent back to the central server, which then aggregates the updates and sends the updated model back to the client devices. This allows for hyper-personalization - while preserving privacy.


Step 1
Model Server
Worker-A
Worker-B
Worker-C
Central server chooses a statistical model for training
Step 2
Model Server
Worker-A
Worker-B
Worker-C
Initial model is transmitted to several nodes
Step 3
Model Server
Worker-A
Worker-B
Worker-C
Nodes train their local models using asymmetric local data
Step 4
Model Server
Federated averaging
Worker-A
Worker-B
Worker-C
Central server creates one shared model using parameters from nodes' models

Edge computing is not going to completely replace cloud computing, rather it's going to work in conjunction with it.

There are still multiple applications, where cloud-based machine learning performs better, and with basic Edge AI the models still need to be trained in cloud-based environments. In general, if the applications can tolerate cloud-based latencies or if the inference can be executed directly in the cloud, cloud computing is a better option.



04

Edge AI trends and the future

There is always a lot of hype associated with new technology, but there are several concrete reasons for the growth of the Edge AI market.

According to the Global Edge AI Software Market Growth **report**, the Edge AI software market alone will grow from $ 346.5 million to about $ 1.1 billion by 2024. Edge AI hardware and consulting market will grow at the same pace. **Grand View Research** estimates that the total global Edge computing market will grow 37.4 percent per year and will be worth $ 43.4 billion by 2027.

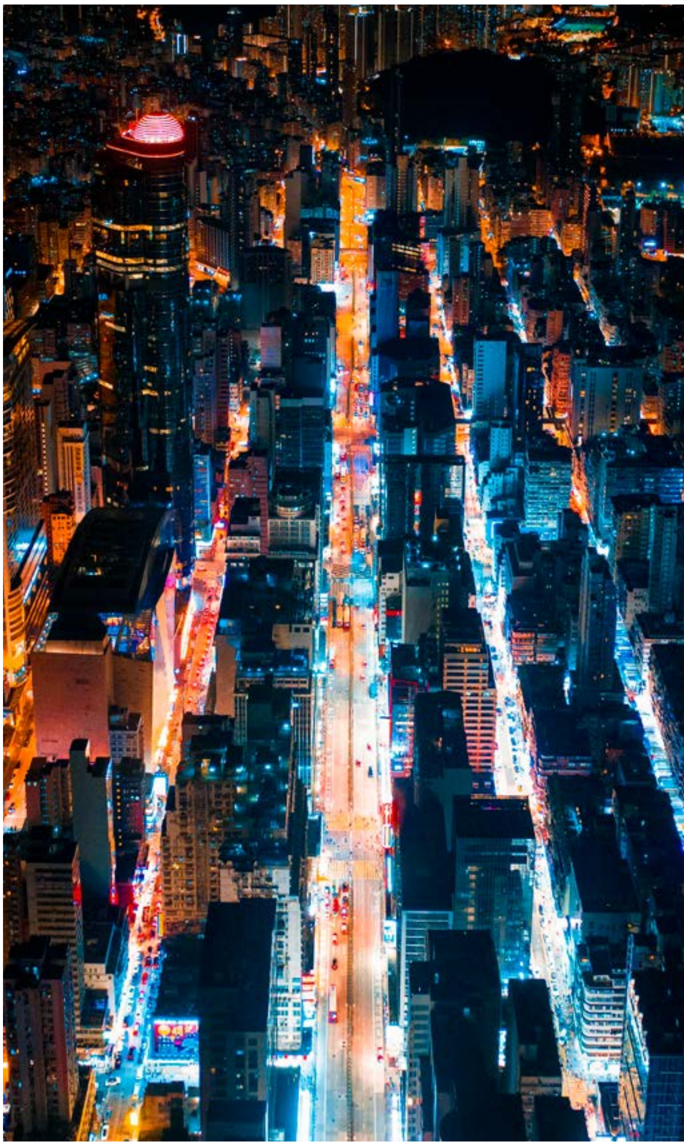
5G

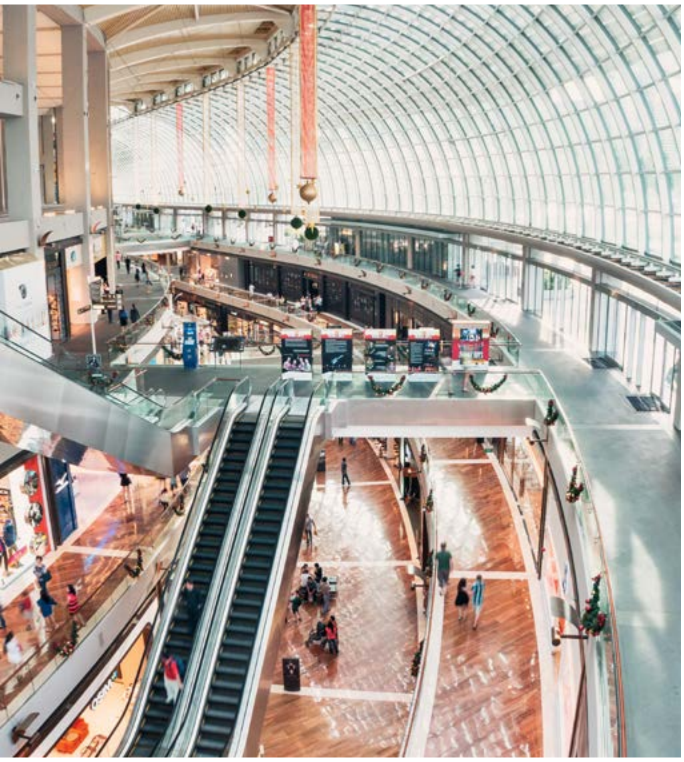
5G networks enable the collection of large and fast data streams. The construction of 5G networks begins gradually, and initially they will be set up very locally and in densely populated areas. The value of Edge AI technology increases when the utilization and analysis of these data streams are done as close as possible to devices connected to the 5G network.

Massive amounts of IoT generated data

IoT and sensor technology produce such large amounts of data that even collecting the data is often tricky and sometimes even impossible in practice. For example, the latest Airbus A350 air crafts have 50,000 sensors that collect 2.5 terabytes of data every day. In comparison, this is more data than the whole Wal-Mart's massive Teradata data warehouse had in 1992.

Data tells us nothing if it is detached from its place of origin and does not have metadata describing the meaning of the data. Therefore, simply retrieving data is not enough. To summarize, you could say that only Edge AI makes it possible to fully utilize the much-hyped IoT data. A massive amount of sensor data can be analysed locally, and operational decisions can be automated. Only the most essential data is stored in a data warehouse located in the cloud or in a data center.





Customer experience

People expect a smooth and seamless experience from services. Nowadays, a delay of just a few seconds could easily ruin the customer experience. Edge computing responds to this need by eliminating the delay caused by data transfer.

In addition, sensors, cameras, GPU processors and other hardware are constantly becoming cheaper, so both customized and highly productized Edge AI solutions are becoming available to more and more people.

05

Examples of Edge AI use cases

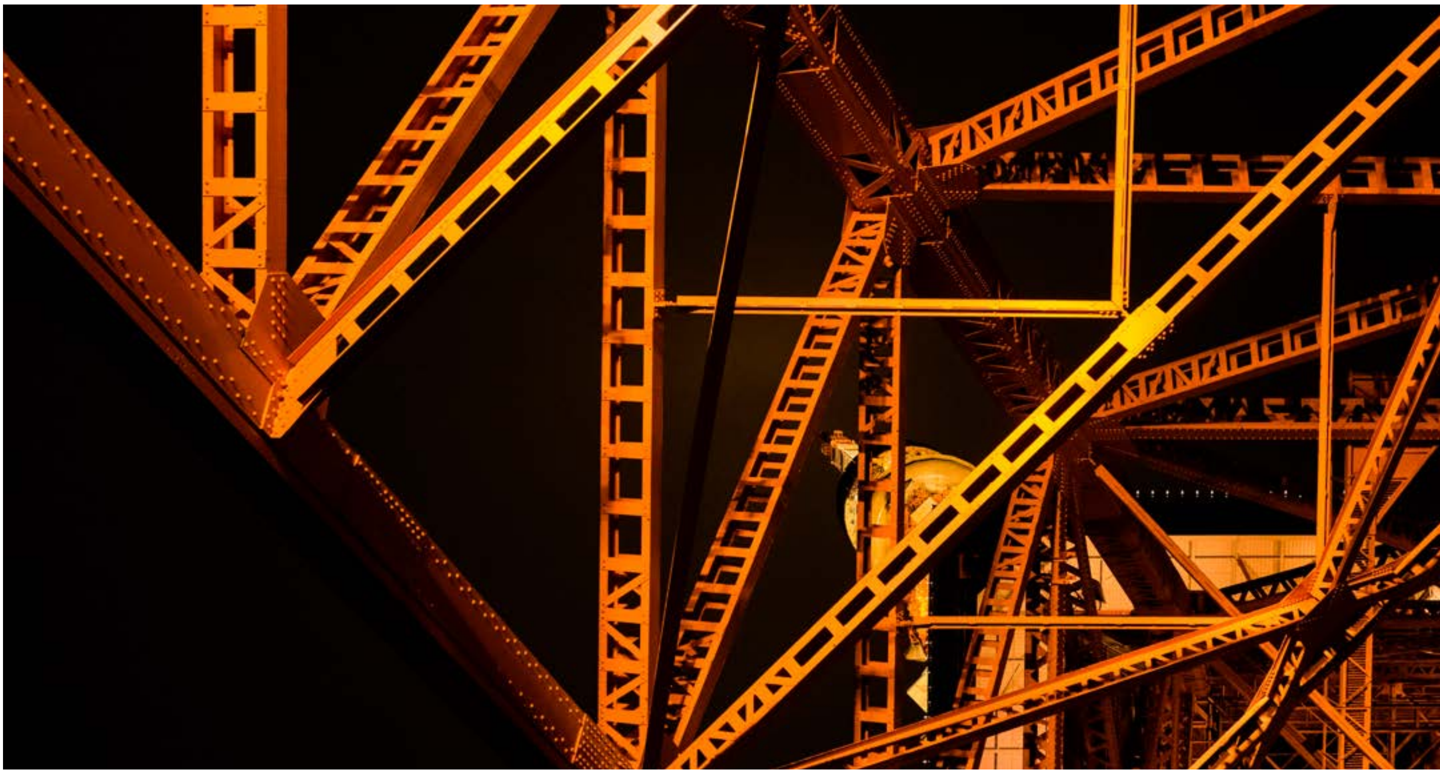
Edge AI is particularly beneficial in the manufacturing sector (possible use cases include proactive maintenance, quality control, production line automation, and safety monitoring through video analytics) and in the traffic and transportation sectors (including autonomous vehicles and machinery). Other growing industries in Edge AI are retail and energy industries.

Manufacturing

One of the most promising Edge AI use cases is manufacturing quality control. Advanced machine vision (video analytics), an example of Edge AI, can monitor product quality tirelessly, reliably and with great precision.

Video analytics can detect even the smallest quality deviations that are almost impossible to notice with the human eye.

Production automation requires advanced analytics, for example in the prediction of equipment failures. Analyzing the data from the sensors and detecting abnormalities in near real-time makes it possible to shut the device off before it breaks. This can save you from significant hardware damages or even injuries. Automatic analysis of material flows by video analysis, for example, is also a promising use case.



Transportation and traffic

Passenger air crafts have been highly automated for a long time. Real-time analysis of data collected from sensors can further improve flight safety.

While fully **autonomous and fully unmanned ships** may not become a reality until years from now, modern ships already have a lot of advanced data analytics.

Edge AI technology can also be used, for example, to calculate passenger numbers and to locate fast vehicles with extreme accuracy. In train traffic, more accurate positioning is the first step and a prerequisite towards autonomous rail traffic.

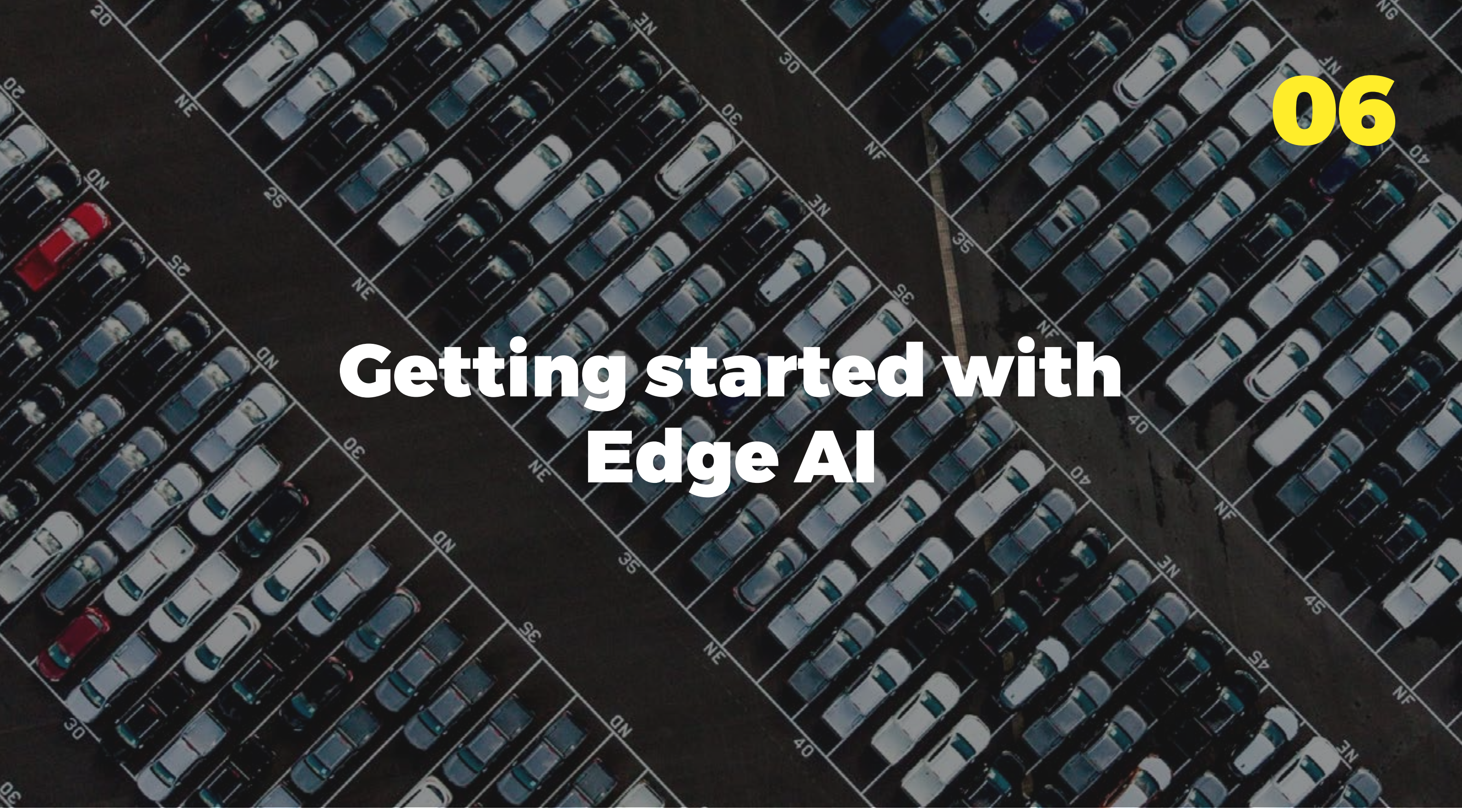
Energy

A smart grid produces a huge amount of data. A truly smart grid enables demand elasticity, consumption monitoring and forecasting, renewable energy utilization and decentralized energy production. However, a smart grid requires communication between devices, and therefore transferring data through a traditional cloud service might not be the best alternative.

Retail

Large retail chains have been doing customer analytics for a long time. The analytics is currently largely based on an analysis of completed purchases, i.e. receipt data. Although good results can be achieved with this method, the receipt data does not tell you everything. It doesn't tell you how people move around the store, how happy they are, what they stop to watch, etc. Video analytics analyses fully anonymized data extracted from a video image and provides an understanding of people's purchasing behaviour that can improve customer service and the overall shopping experience.

DOWNLOAD OUR RETAIL ANALYTICS EBOOK



06

Getting started with Edge AI

If you read this far, you probably already wonder how Edge AI solutions could work as a part of your business.

As in modern knowledge management and business intelligence, success in Edge AI depends on having courage and an open-minded attitude: be agile, dare to fail, learn from your mistakes, and scale your success into new business processes and service development.

Book a free half-hour telephone conversation with us, so we can brainstorm together how Edge AI solutions could benefit your business, what kind of use cases are feasible in your situation and what kind of challenges different implementations might have.

Book a free consultation today

You risk nothing by having a chat with us to explore if Edge AI could help your business.

So, please, book a call with Mr. Janne Honkonen, our Chief Executive Officer, at your convenience.

Also, don't forget to subscribe to our blog in case you haven't yet done so.

Thank you for your attention!

BOOK A CALL

Subscribe to our blog

www.advian.fi