Edge Computing for beyond 5G

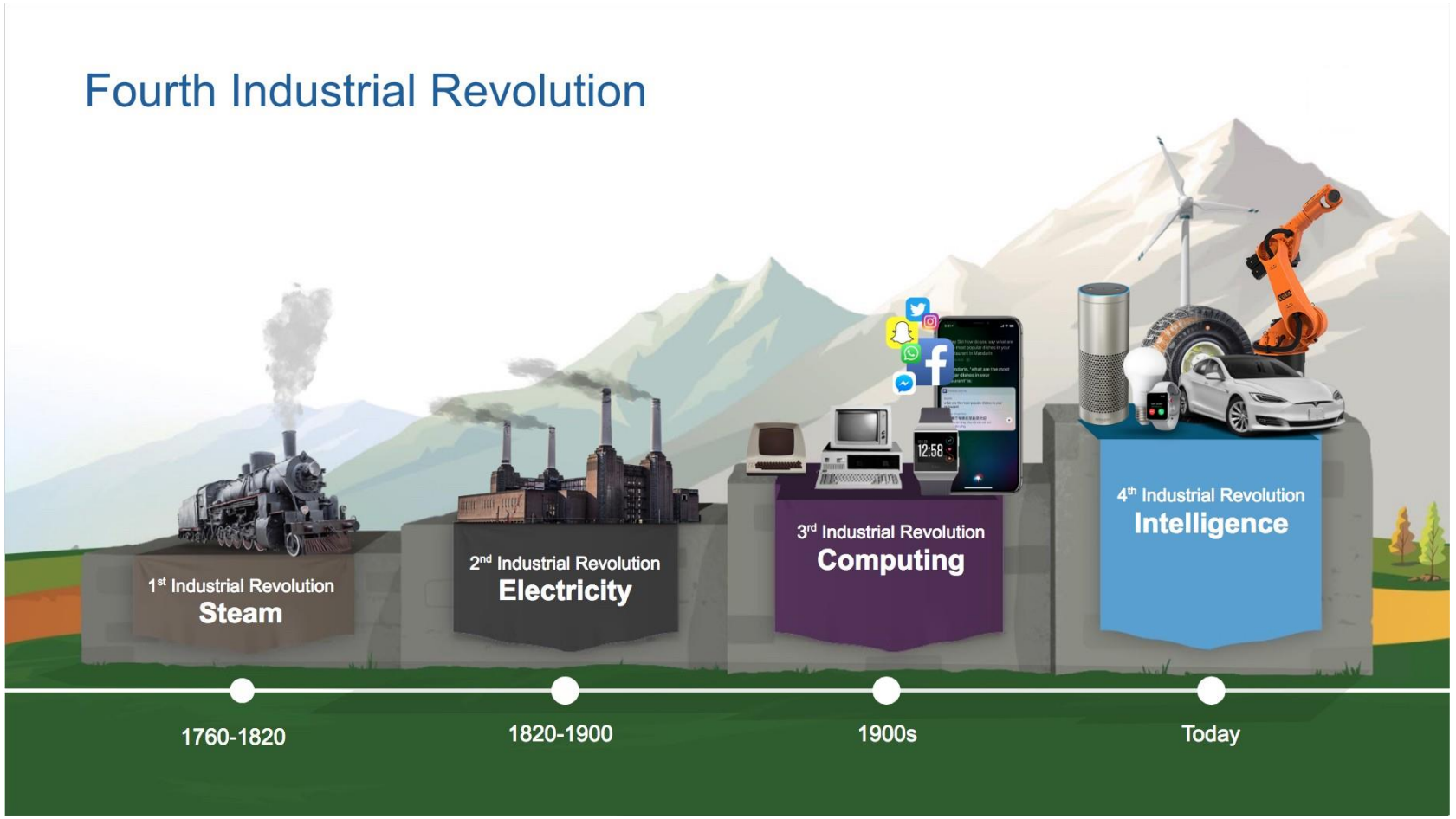
Choong Seon HONG

Professor, Department of Computer Science
and Engineering, Kyung Hee University,

KHU

INTELLIGENT NETWORKING LAB

Introduction-Industry 4.0

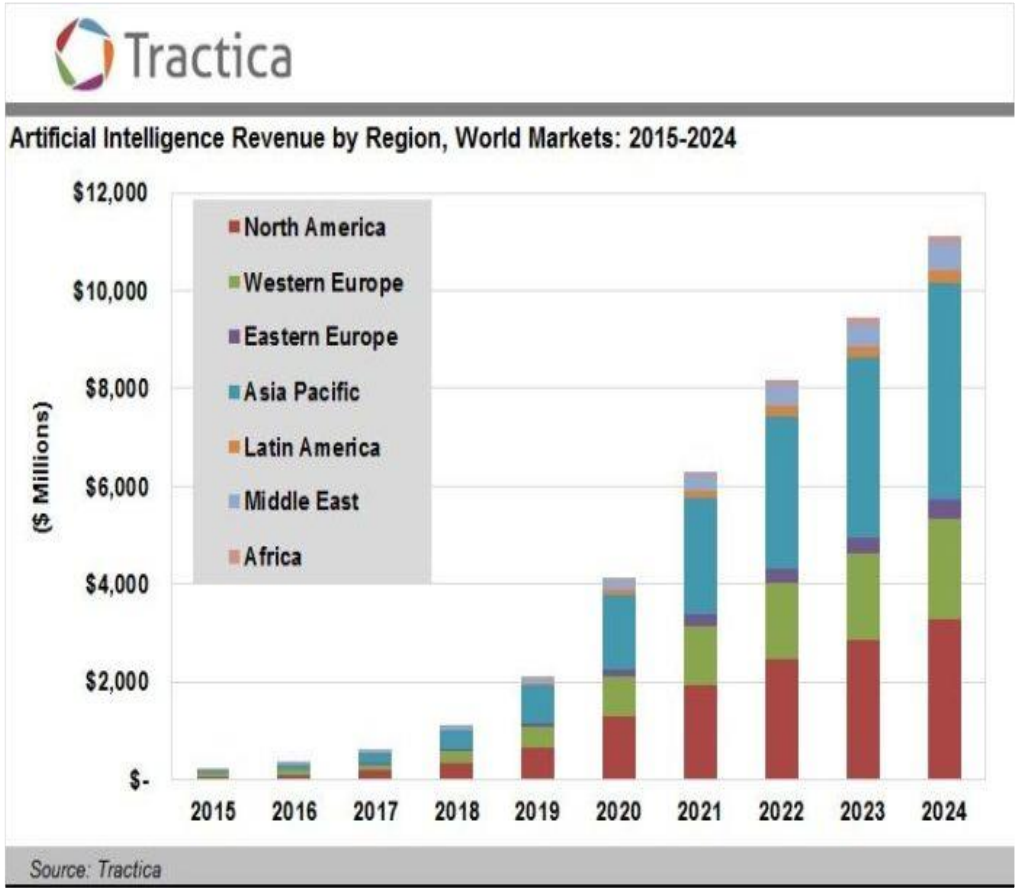


Introduction – Edge Application

- Shopping with Augmented Reality

[1] https://www.youtube.com/watch?v=UQcJSZPpNhA&feature=youtu.be

AI Applications



AI technology is currently being applied to solutions for various use cases such as Agriculture, Financial Services, Retail, and Energy, and it is expected that sales from AI technology will continue to increase globally.

Agriculture

- To evaluate the best crop choices against various parameters such as soli quality and the needs of the farmers

Financial Services

- Many investment firms are using deep learning algorithms to improve speed, optimize and increase the efficiency of recognizing trends across vast data sets to obtain competitive advantages.

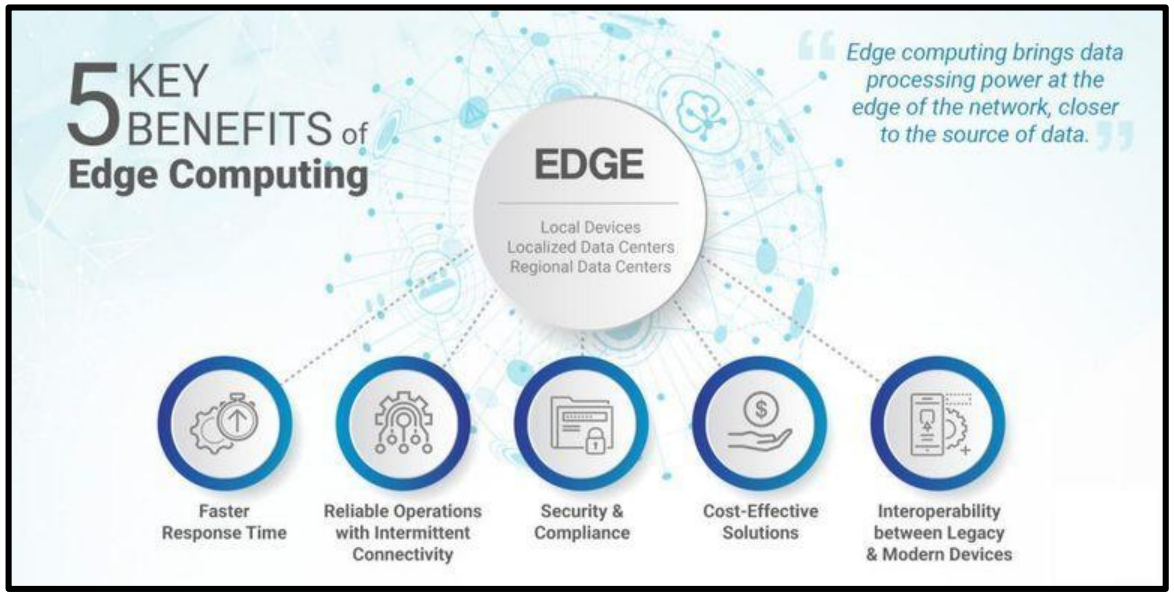
Retail

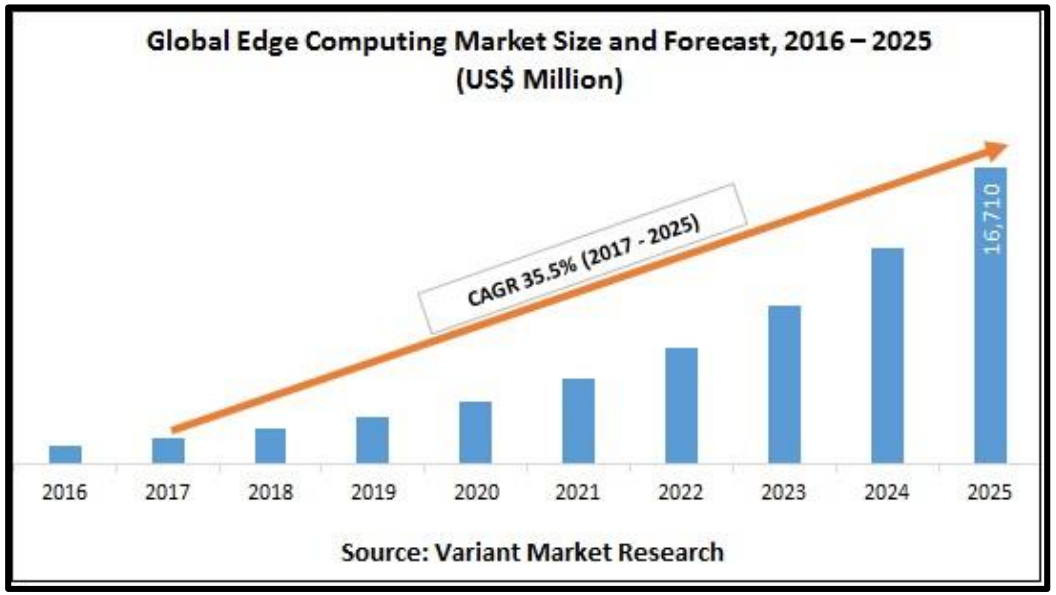
- Optimization of pricing based on the product demand.
- To offer seamless experience to customers and deliver greater customer satisfaction.

Energy

- To improve and increase efficiency and awareness of gas plants and power grid systems.

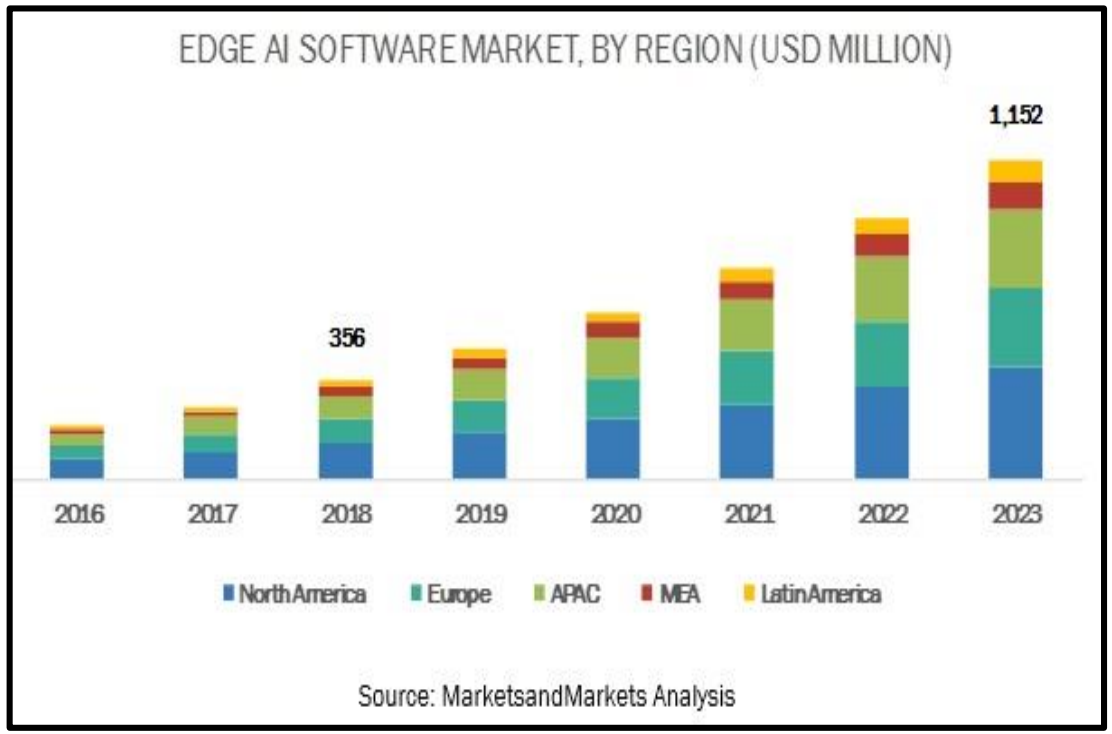
Edge Computing Market Trends

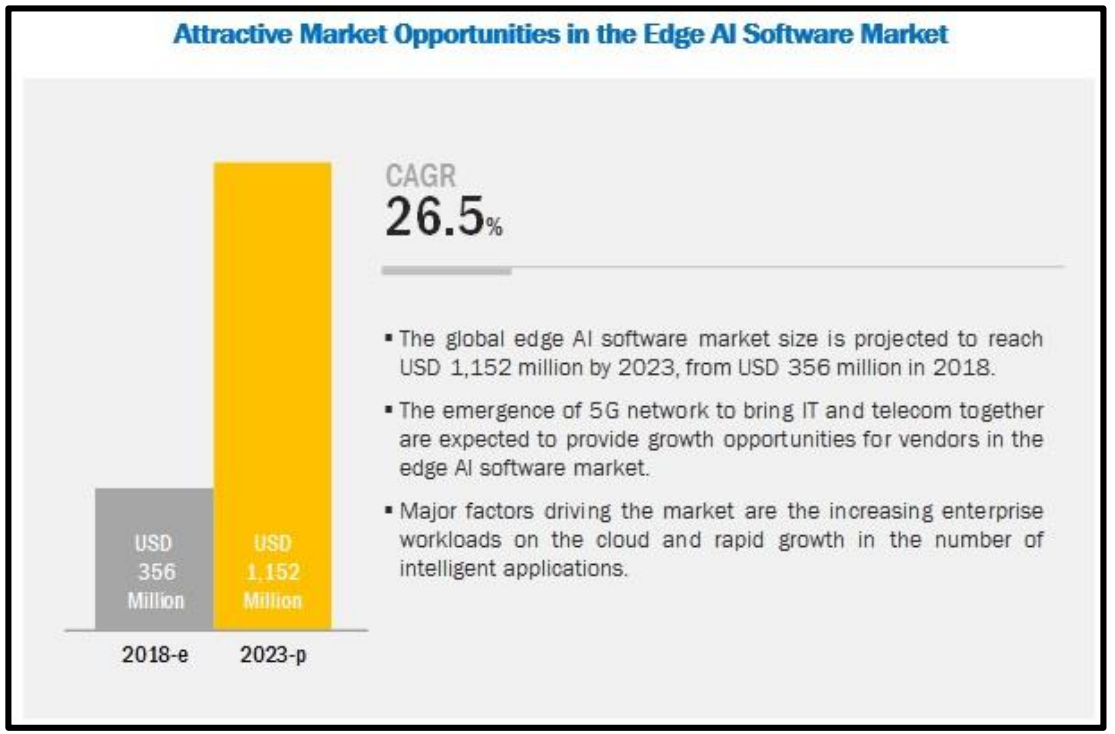




- The global edge computing market is expected to reach about $ 16.71 billion by 2025, (2017~2025) Year average growth rate of 35.5%
- Among the many fields that utilize Edge Computing in addition to 5G, it is expected to show the highest growth rate in areas where fast service provision is important (Smart City, Smart Factory, etc.)

Edge AI Software Market





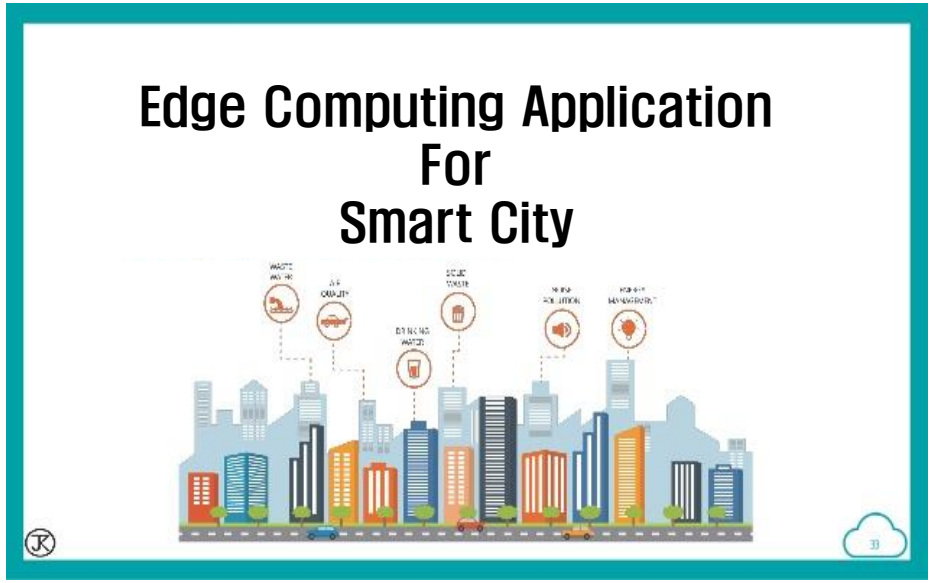
- The Edge AI software market is expected to grow at a CAGR of approximately 26.5% from $ 356 million in 2018 to $ 1,122 million in 2023, driven by increased cloud loads and accelerated development of various AI applications.

Introduction – Edge AI

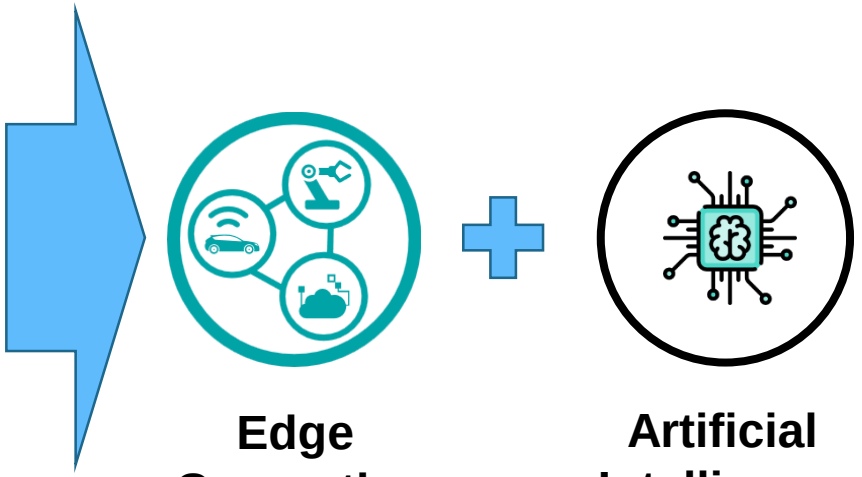
✓ Changes in the computing paradigm (Centralized Cloud Computing -> Distributed Edge Computing)
- Real-time, high-capacity, low-latency service increases
- The emergence of smart city and the necessity of utilization of artificial intelligence due to the 4th industrial revolution

→ **Increase in role of artificial intelligence algorithm in distributed edge computing environment**

4th Industrial Revolution



Smart City and Edge Computing



※ 출처: 2019년 국내 ICT 시장 10대 전망, 한국 IDC

Introduction –Edge AI

The current premise in classical ML is based on a single node in a centralized and remote data center with full access to a global dataset and a massive amount of storage and computing power, sifting through this data for inference.

Nevertheless the advent of a new breed of intelligent devices and high-stake applications ranging from drones to augmented/virtual reality (AR/VR) applications, and self driving vehicles, makes cloud-based ML inadequate. These applications are real-time, cannot afford latency, and must operate under high reliability, even when network connectivity is lost.

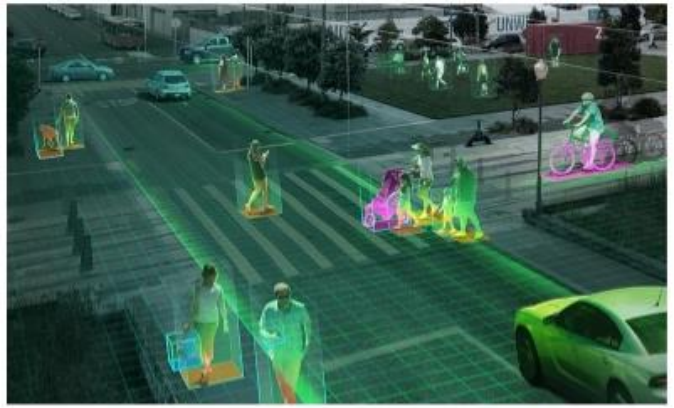
WHY AI AT THE EDGE MATTERS

Bandwidth	Latency	Availability
		
1 billion cameras WW (2020)	30 images per second	50% of world at less than 8mbps
30B Inference/Second	200ms latency	Only 73% 3G/4G availability WW

"Billions of intelligent devices will take advantage of DNNs to provide personalization and localization as GPUs become faster and faster over the next several years." – Tractica

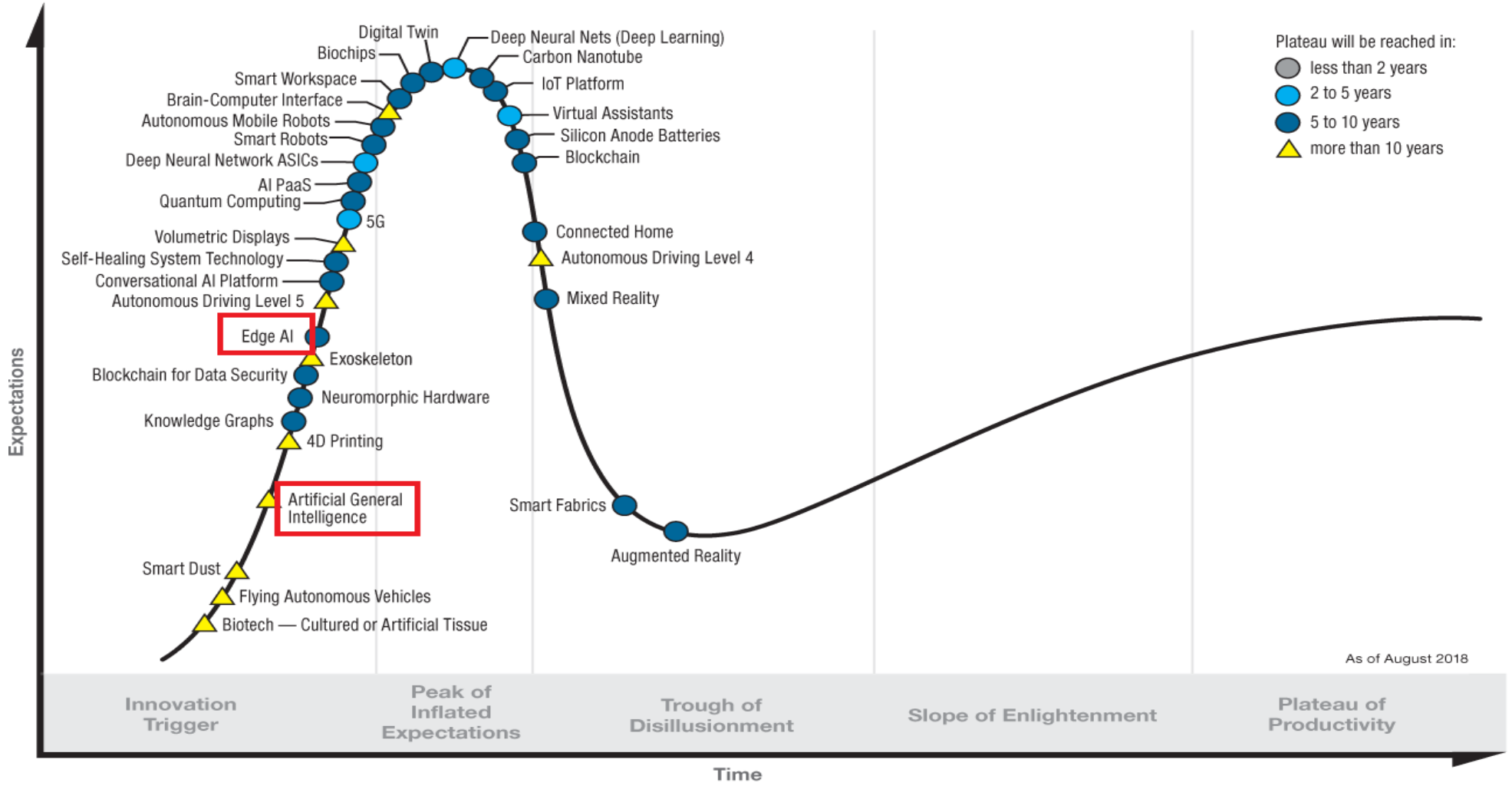
Introduction – Edge AI

- Indeed, an autonomous vehicle that needs to apply its brakes, cannot allow even a millisecond of latency that might result from cloud processing, requiring split second decisions for safe operation [2], [3].
- A user enjoying visual-haptic perceptions requires not only minimal individual perception delays but also minimal delay variance, to avoid motion sickness [4], [5].
- A remotely controlled drone or a robotic assembler in a smart factory should always be operational even when network connection is temporarily unavailable [6]–[8], by sensing and reacting rapidly to local (and possibly hazardous) environments.



[2] S.-C. Lin, Y. Zhang, C.-H. Hsu, M. Skach, M. E. Haque, L. Tang, and J. Mars, "The architectural implications of autonomous driving: Constraints and acceleration," in Proc. of the 23rd ACM ASPLOS, ASPLOS '18, (Williamsburg, VA, USA), pp. 751–766, ACM, Mar. 2018.
[3] M. K. Abdel-Aziz, C.-F. Liu, S. Samarakoon, M. Bennis, and W. Saad, "Ultra-reliable low-latency vehicular networks: Taming the age of information tail," in Proc. of GLOBECOM [accepted], (Abu Dhabi, UAE), Dec. 2018.
[4] J. Park and M. Bennis, "URLLC-eMBB slicing to support VR multimodal perceptions over wireless cellular systems," ArXiv preprint, vol. abs/1805.00142, May 2018.
[5] ABI Research and Qualcomm, "Augmented and virtual reality: The first wave of 5g killer apps," white paper, Feb. 2017.
[6] T. Kagawa, F. Ono, L. Shan, K. Takizawa, R. Miura, H. Li, F. Kojima, and S. Kato, "A study on latency-guaranteed multihop wireless communication system for control of robots and drones," in Proc. of 20th WPMC, (Yogyakarta, Indonesia), pp. 417– 421, Dec. 2017.
[7] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs," IEEE Transactions on Wireless Communications, vol. 15, pp. 3949–3963, June 2016.
[8] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. Galati Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," ArXiv preprint, vol. abs/1809.01752, Sept. 2018.

Technology Trends

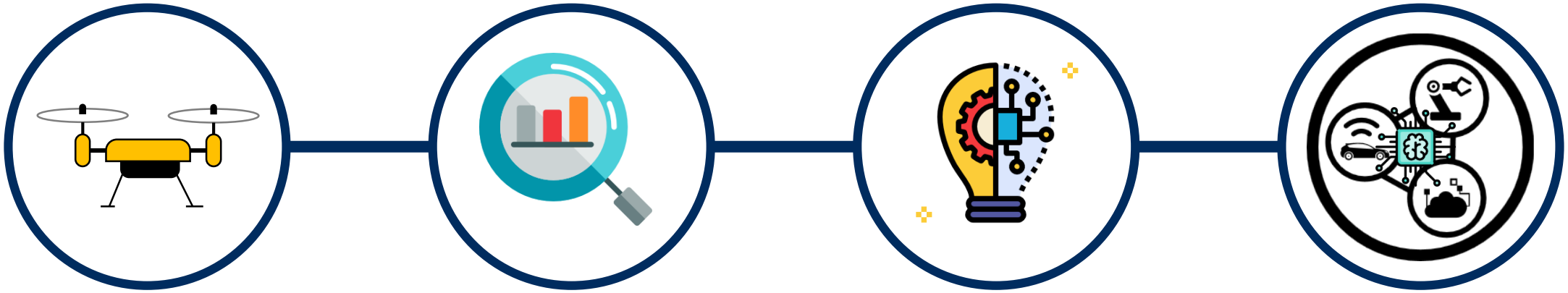


※ Source : Gartner Hype Cycle for Emerging Technologies 2018

Gartner 2019 Ten Trends in the Future

① Autonomous Things, ②Augmented Analytics, ③ AI-Driven Development, ④ Digital Twin, ⑤ Empowered Edge, ⑥ Immersive Experience, ⑦ Block chain, ⑧ Smart Spaces, ⑨ Digital Ethics and Privacy, ⑩ Quantum Computing

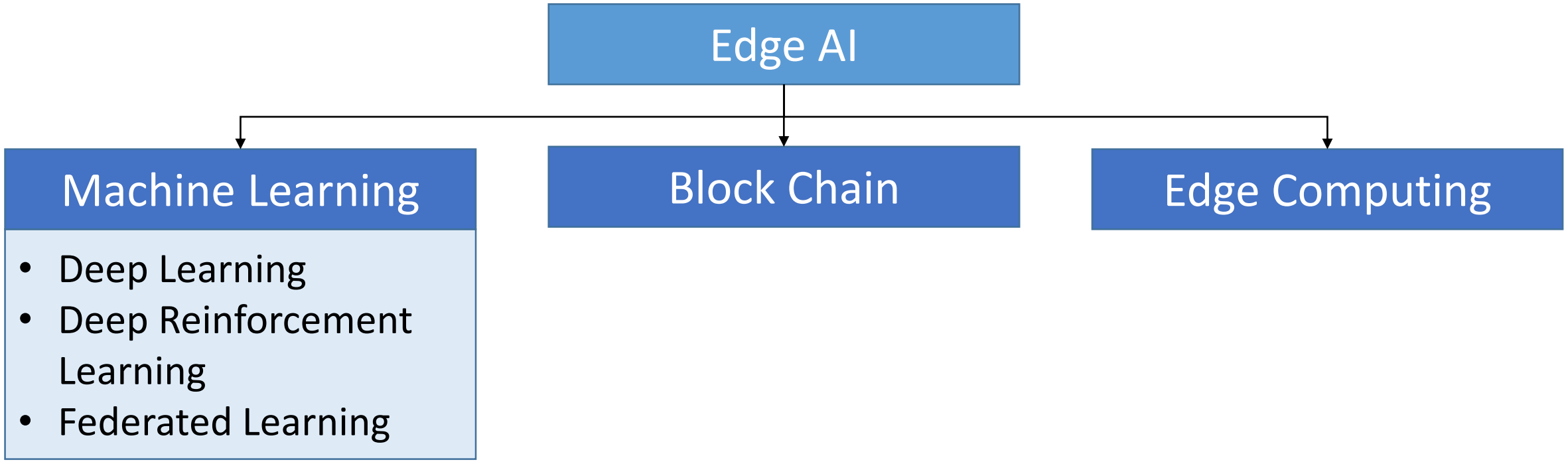
Future technology change



Autonomous Things | Augmented Analytics | AI-Driven Development | Empowered Edge

※ 출처: Top 10 Strategic Technology Trends 2019, Gartner

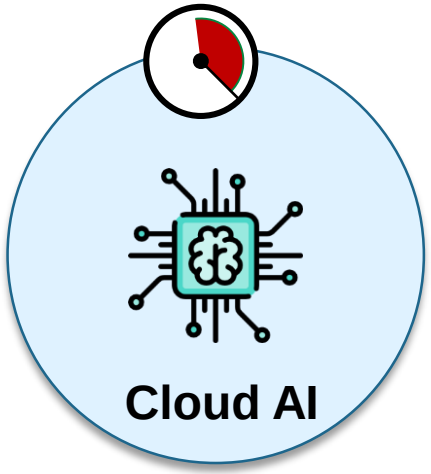
Enabling Technologies to Implement Edge AI



Cloud VS. Edge AI

Pros

- ❑ Can train large neural network model
- ❑ High computation resources
- ❑ Big Data processing
- ❑ Easy to scale
- ❑ Low cost storage

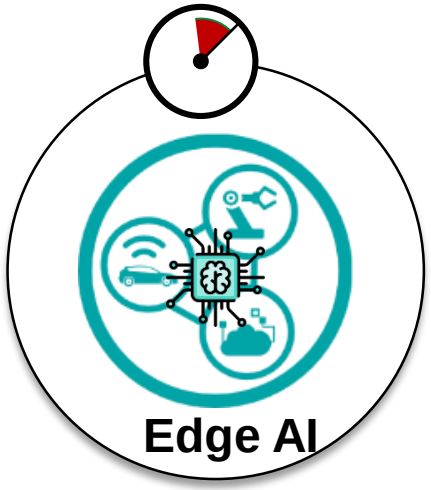


Cons

- ❑ High service delay
- ❑ High bandwidth cost
- ❑ Sending raw data over the Internet to the cloud have privacy, security and legal issues

Pros

- ❑ Real-time predictions for mission-critical applications
- ❑ Efficient use of network bandwidth
- ❑ Process data closest to the source
- ❑ Low latency response
- ❑ Support mobility



Cons

- ❑ Low computation power than the cloud
- ❑ Computation resources are Less scalable than cloud

Supervised Learning

By feeding an input data sample, the goal of supervised learning is to predict a target quantity, e.g., regression, or classification of the category within predefined labels. This ability can be obtained by optimizing the NN parameters by feeding training data samples, referred to as a training process. In supervised learning, the input training samples are paired with the ground-truth output training samples. These output samples 'supervise' the NN to infer the correct outputs for the actual input samples after the training process completes.

Unsupervised Learning

The training process of unsupervised learning is performed using only the input training samples. In contrast to supervised learning, unsupervised learning has no target to predict, yet aims at inferring a model that may have generated the training samples. Clustering of un-grouped data samples and generating new data samples by learning the true data distribution, i.e., a generative model, belong to this category

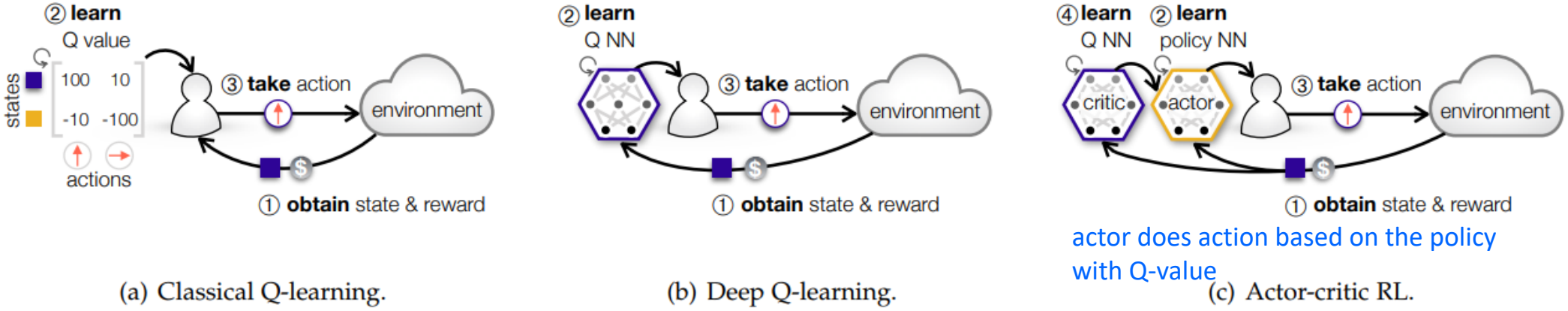
[9] Park, J., Samarakoon, S., Bennis, M. and Debbah, M., 2018. Wireless network intelligence at the edge. *arXiv preprint arXiv:1812.02858*.

Type of Machine Learning Schemes for Cloud and Edge AI

Reinforcement Learning (RL) [9]

The goal of RL is make an agent in an environment take an optimal action at a given current state, where the interaction between the agent's action and state through the environment is modeled as a Markov decision process (MDP).

- When each action is associated with a return, the agent takes an action that maximizes its predicted cumulative return, e.g., Q-learning that maximizes the Q value for each state, as illustrated in Figure(a).
- In Q-learning, the larger state dimension, the more computation. This problem is resolved by deep Q-learning as shown in Figure (b), where a NN approximates the Q function and produces the Q values by feeding a state. These value-based RL can take actions only through Q values that are not necessarily required. Instead, one can directly learn a policy that maps each state into the optimal action, which is known as policy-based RL whose variance may become too large [10].
- Actor-critic RL is a viable solution to both problems, comprising a Neural Network (NN) that trains a policy (actor NN) and another NN that evaluates the corresponding Q value (critic NN), as visualized in Figure (c).



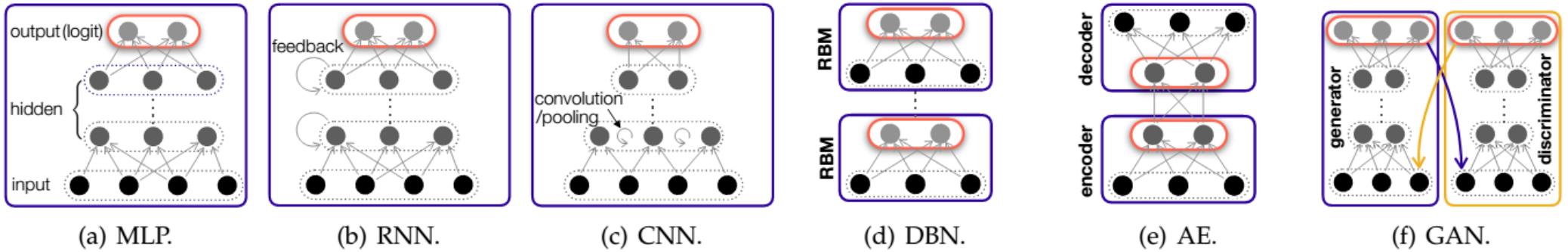
(a) Classical Q-learning. (b) Deep Q-learning. (c) Actor-critic RL.

Examples of RL: (a) classical Q-learning without any Neural Network; (b) deep Q-learning with a Neural Network, and (c) actor – critic RL with actor and critic Neural Networks.

[9] Park, J., Samarakoon, S., Bennis, M. and Debbah, M., 2018. Wireless network intelligence at the edge. *arXiv preprint arXiv:1812.02858*.
[10] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function ap proximation," in Proc. of the 12th NIPS, NIPS'99, (Colorado, USA), pp. 1057–1063, MIT Press, Dec. 1999

Challenges of utilizing Deep learning in Cloud and Edge AI

- It requires high computation resources to process the big data (high-dimensional data) to train the prediction models.
- It is difficult to find a suitable prediction model among the various types of deep learning models, such as Multilayer Perceptron (MLPs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Convolutional Recurrent Neural Networks (CRNNs), etc. [11].
- It is difficult to tune parameters such as the number of layers (i.e., the depth of the network), types of layers (e.g., Convolutional, Recurrent and Fully Connected layers), and learning rate to improve the accuracy of the prediction model.

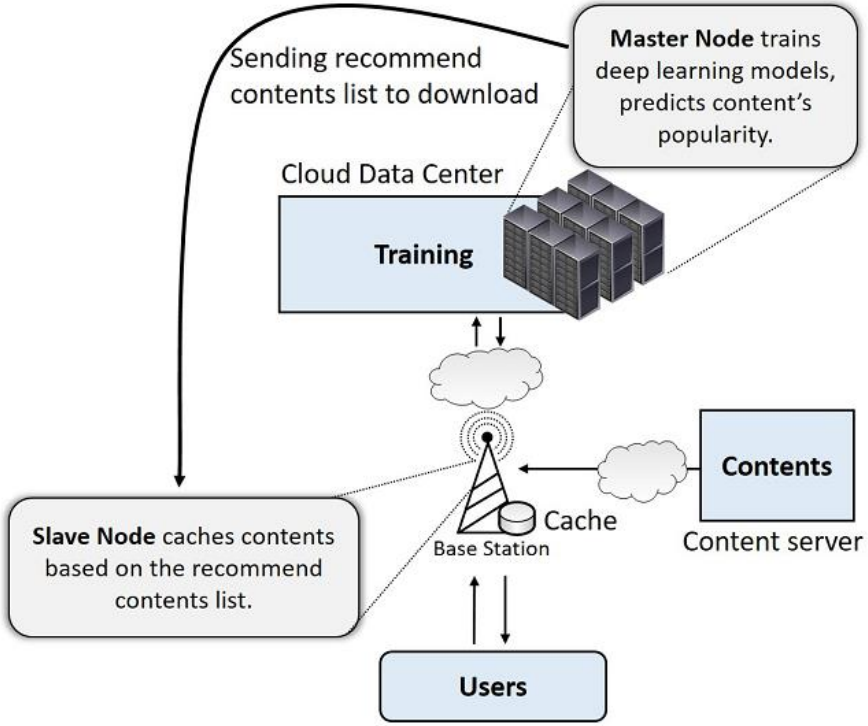


Types of NN architectures: (a) multilayer Perceptron (MLP); (b) recurrent neural network (RNN); (c) convolutional neural network (CNN); (d) deep belief network (DBN) with restricted Boltzmann machines (RBMs); (e) auto encoder; and (f) generative adversarial network (CAN) [3]

[9] Park, J., Samarakoon, S., Bennis, M. and Debbah, M., 2018. Wireless network intelligence at the edge. *arXiv preprint arXiv:1812.02858.*
[11] I. Goodfellow, Y. Bengio and A. Courville, "Deep Learning", in MIT Press, 2016.

Content Caching at Edge (Centralized)

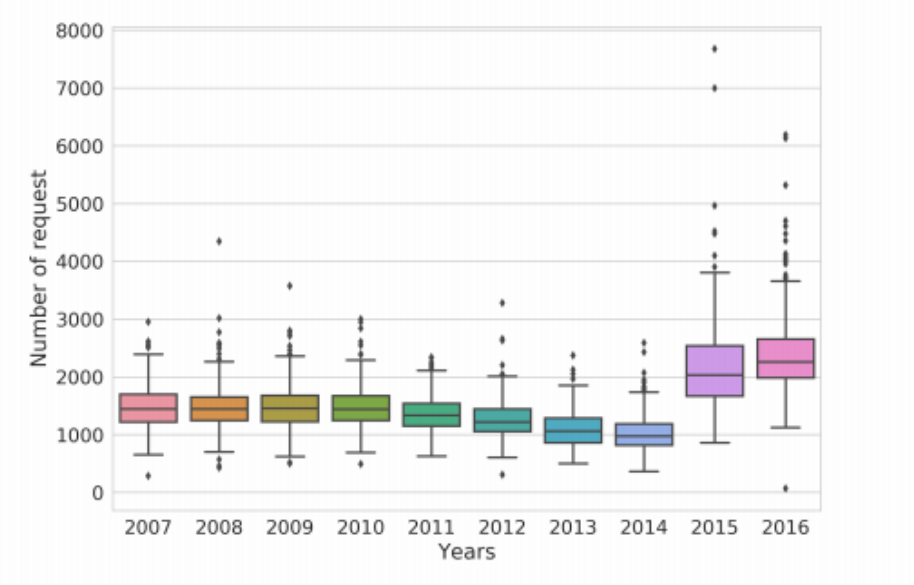
- Caching popular contents at edge nodes such as base stations is a crucial solution for improving users' quality of services in next generation networks.
- However, it is very challenging to correctly predict the future popularity of contents and decide which contents should be stored in the base station cache.
- Recently, with the advances in big data and high computing power, deep learning models have achieved high prediction accuracy.



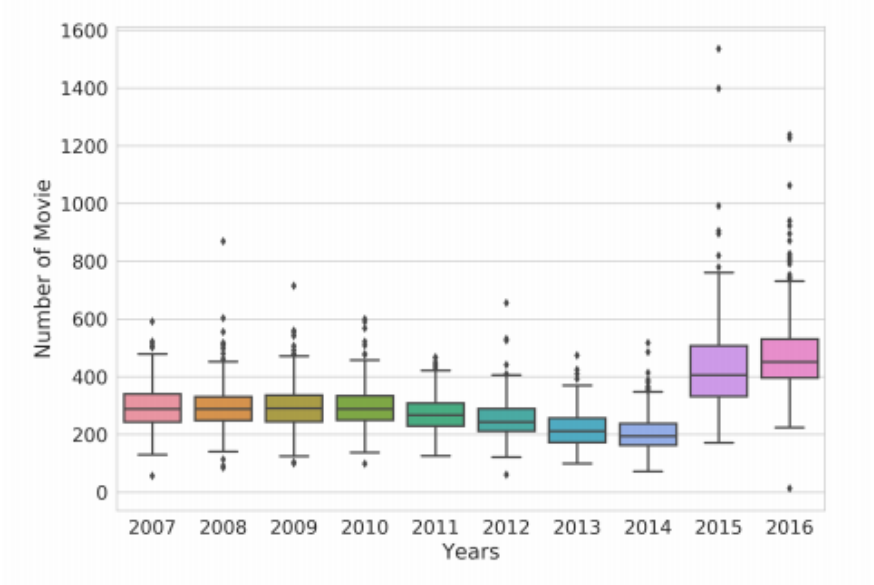
System model of learning-based caching at the edge.

[12] Kyi Thar, Nguyen H. Tran, Thant Zin Oo, Choong Seon Hong, "DeepMEC: Mobile Edge Caching Using Deep Learning," IEEE Access, Vol.6, Issue 1, pp.78260-78275, December 2018

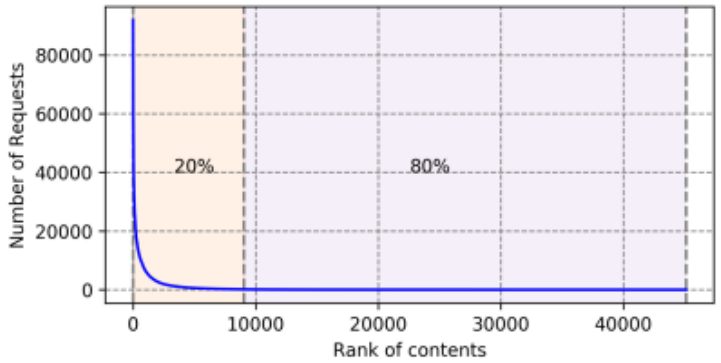
Content Caching at Edge (centralized)



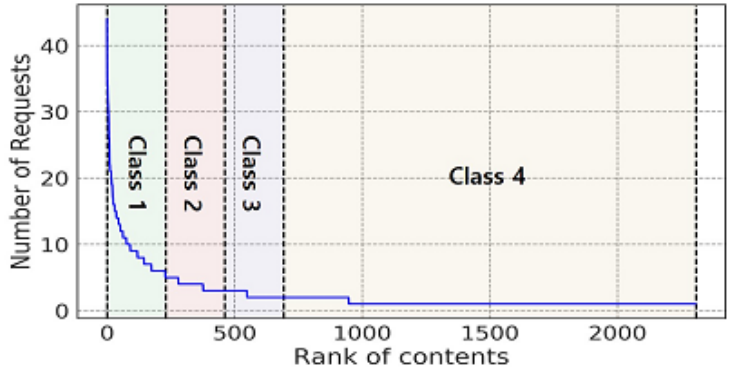
Daily request count of all contents (movies) for each year.



Daily request count of top 20% contents (movies) for each year.



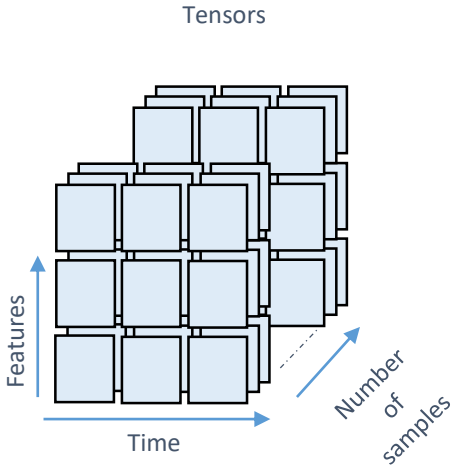
Ranking of all contents (movies) based on total request count received.

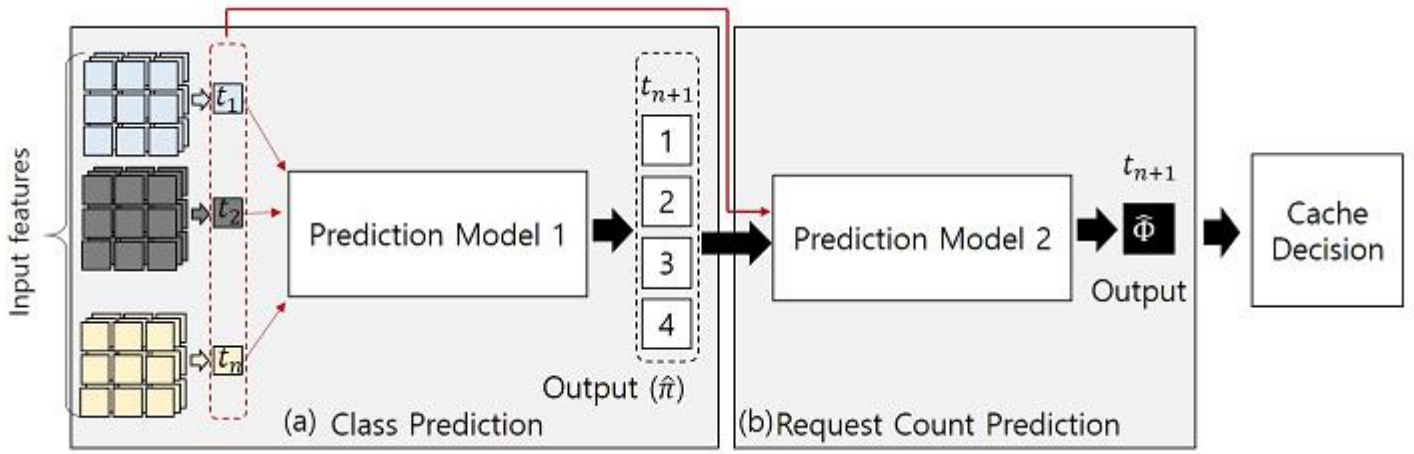


Ranking of all contents (movies) based on daily received content requests and classification labels.

[12] Kyi Thar, Nguyen H. Tran, Thant Zin Oo, Choong Seon Hong, "DeepMEC: Mobile Edge Caching Using Deep Learning," IEEE Access, Vol.6, Issue 1, pp.78260-78275, December 2018

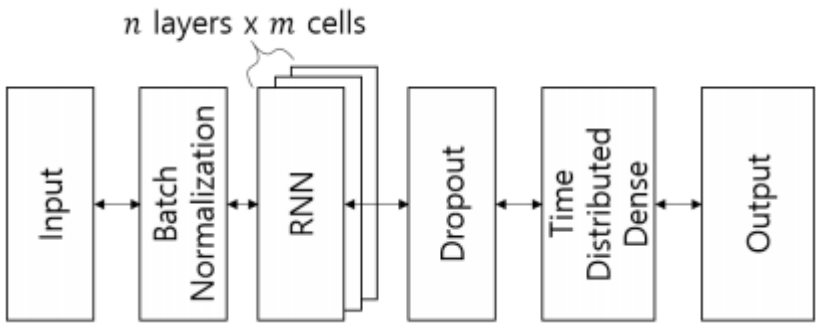
Content Caching at Edge (centralized)



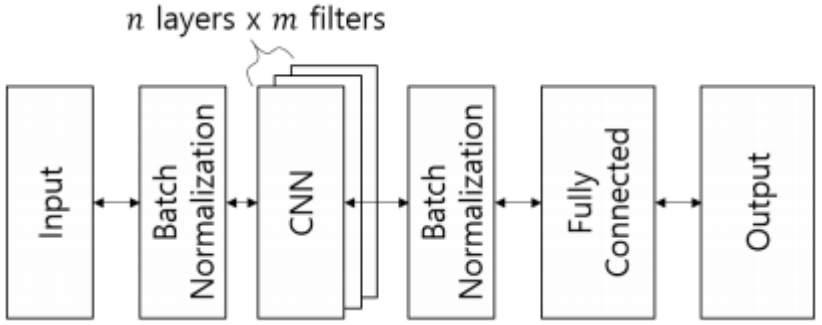


The overview prediction model used in the proposed scheme: (a) Class prediction, (b) Request count prediction.

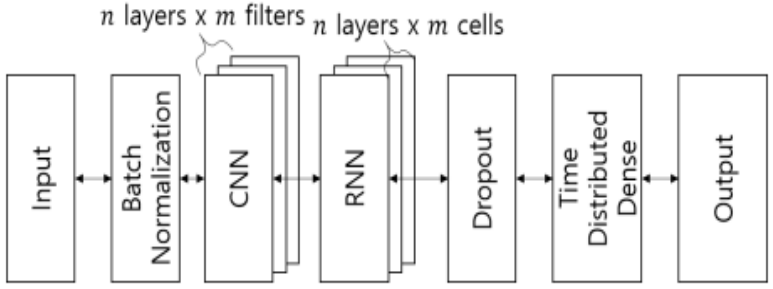
Randomized Model Search



Recurrent neural network model.



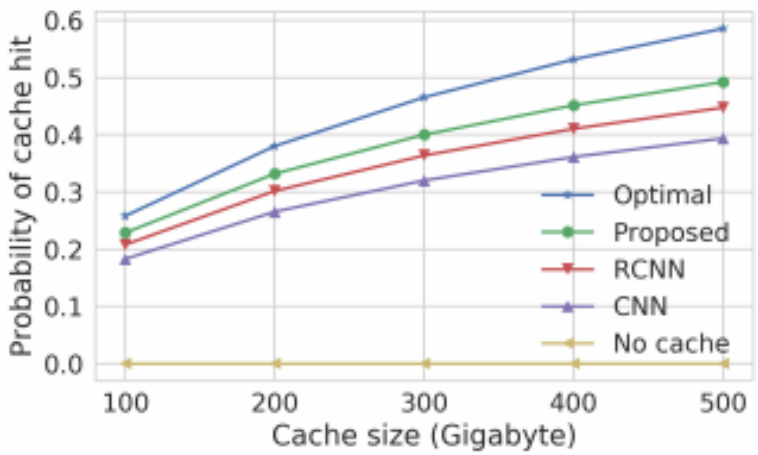
Convolutional neural network model.



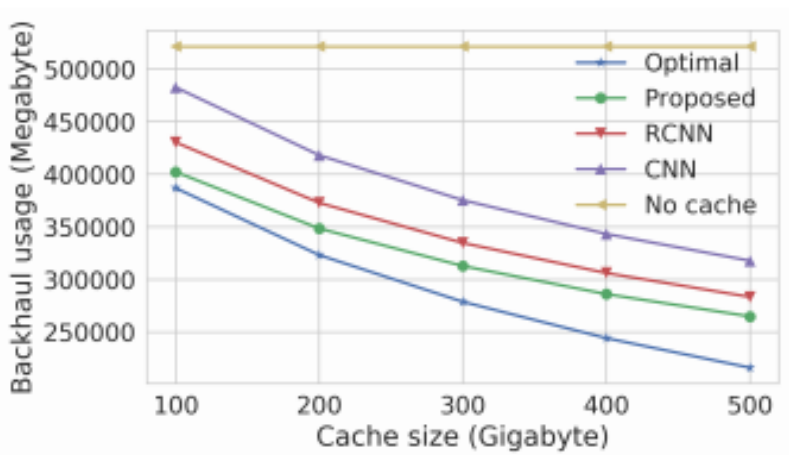
Convolutional recurrent neural network model.

[12] Kyi Thar, Nguyen H. Tran, Thant Zin Oo, Choong Seon Hong, "DeepMEC: Mobile Edge Caching Using Deep Learning," IEEE Access, Vol.6, Issue 1, pp.78260-78275, December 2018

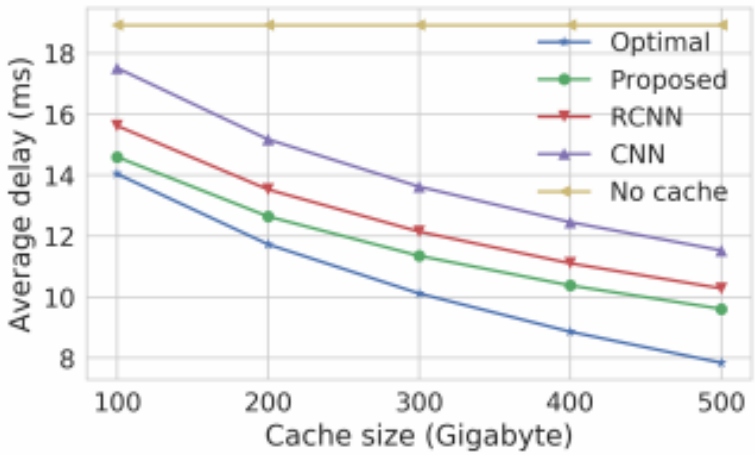
Content Caching at Edge (centralized)



(a)



(b)

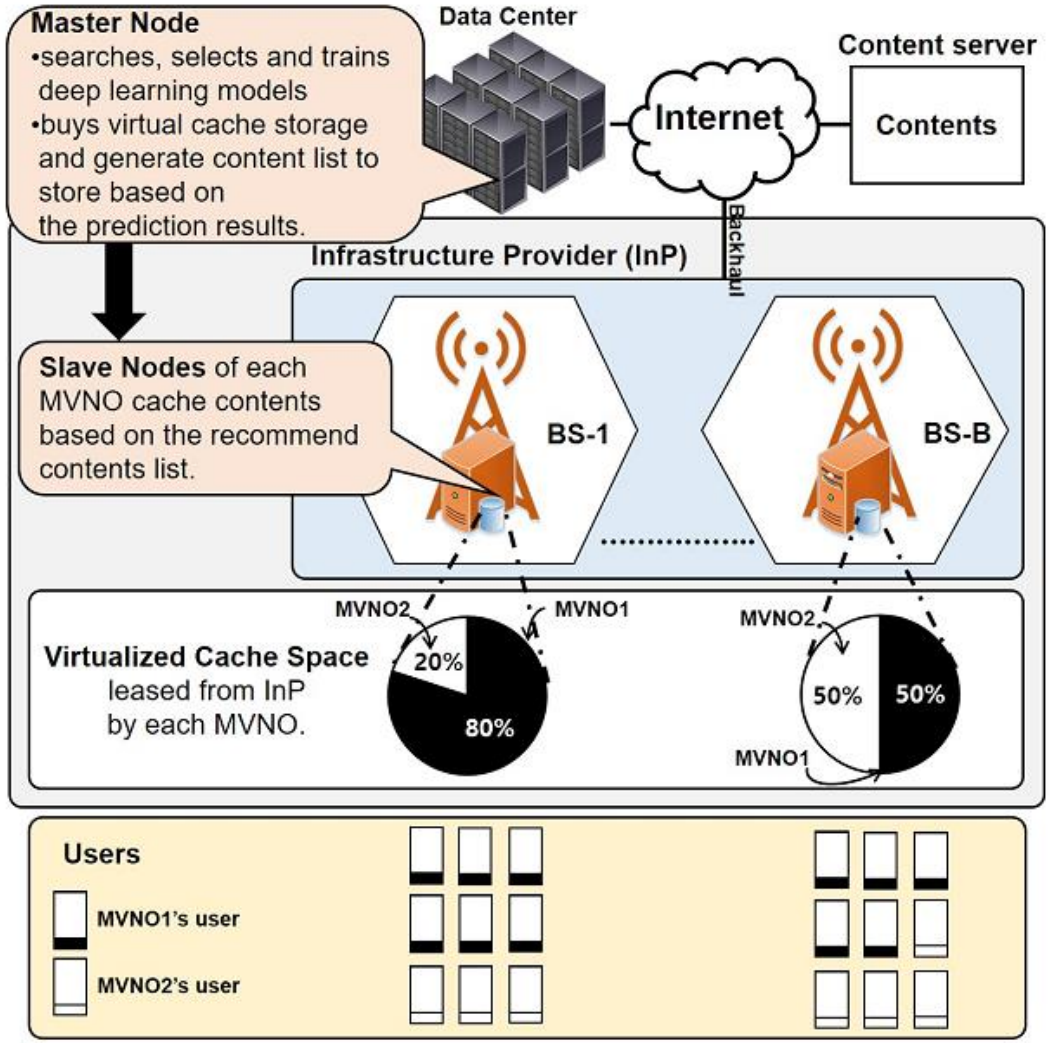


(c)

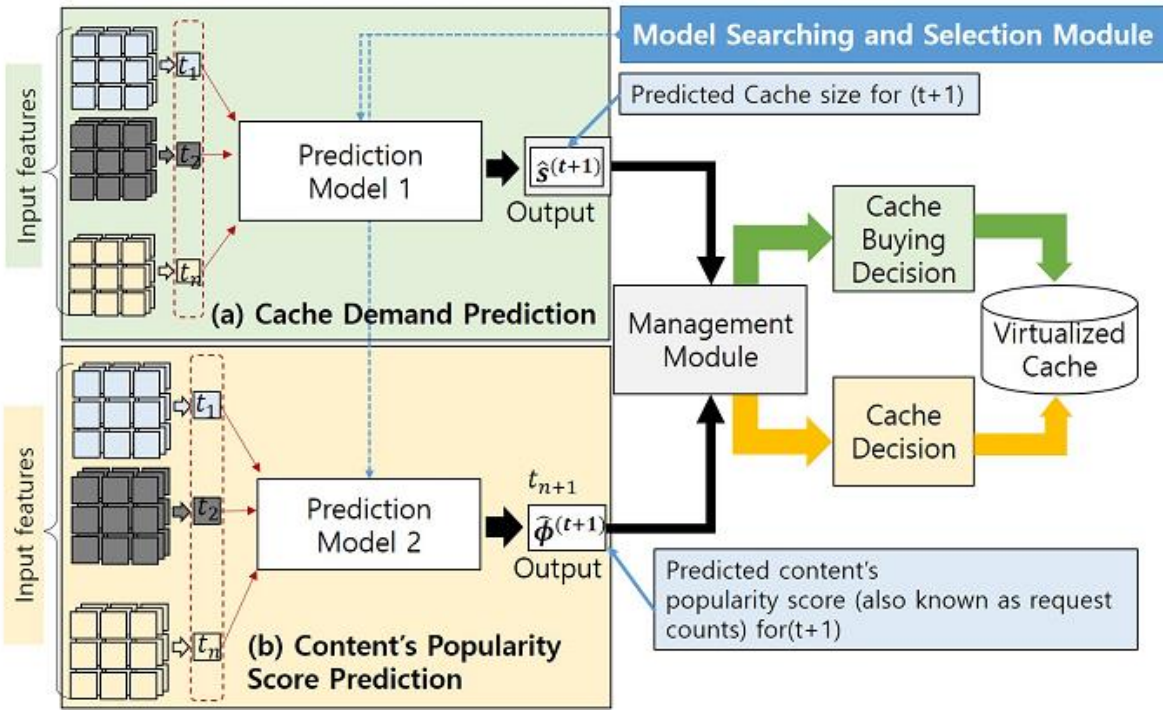
Caching related performance comparisons: (a) Cache hit, (b) Backhaul usage comparison, and (c) Access delay comparison.

[12] Kyi Thar, Nguyen H. Tran, Thant Zin Oo, Choong Seon Hong, "DeepMEC: Mobile Edge Caching Using Deep Learning," IEEE Access, Vol.6, Issue 1, pp.78260-78275, December 2018

Content Caching at Edge (centralized)



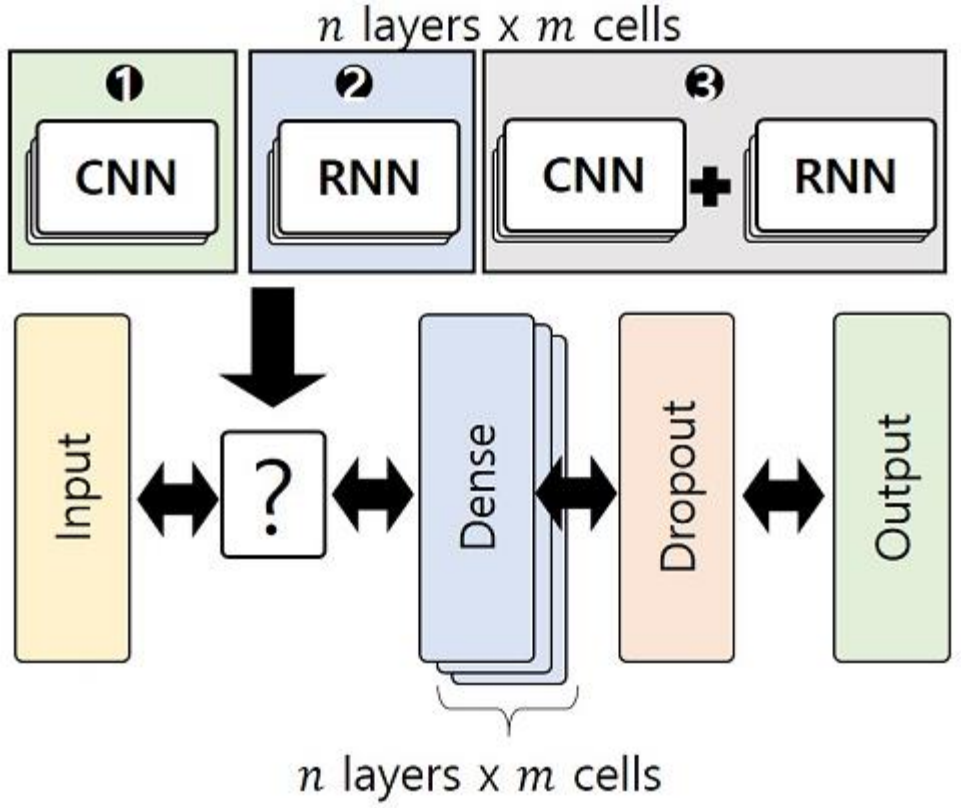
Virtualized cache management for MVNO.



The overview prediction model used in the proposed scheme: (a) Cache demand prediction, (b) Content's popularity scores prediction.

[13] Kyi Thar, Thant Zin Oo, Yan Kyaw Tun, Do Hyeon Kim, Ki Tae Kim, and Choong Seon Hong, "A Deep Learning Model Generation Framework for Virtualized Multi-access Edge Cache Management,"

Content Caching at Edge (centralized)



Deep Learning Models Framework.

Summary of output layer activation function.

Activation	Equation	Range
Softmax	$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$	[0,1]
Sigmoid	$\text{Sigmoid}(z_i) = \frac{1}{1+e_i^{-z}}$	[0,1]
Tanh	$\text{Tanh}(z_i) = \frac{e_i^{z} - e_{[i]}^{-z}}{e_i^{z} + e_i^{-z}}$	[-1,1]
Relu	$\text{Relu}(z_i) = \begin{cases} 0, & \text{for } z_i < 0. \\ z, & \text{for } z_i \geqslant 0. \end{cases}$	[0,1]

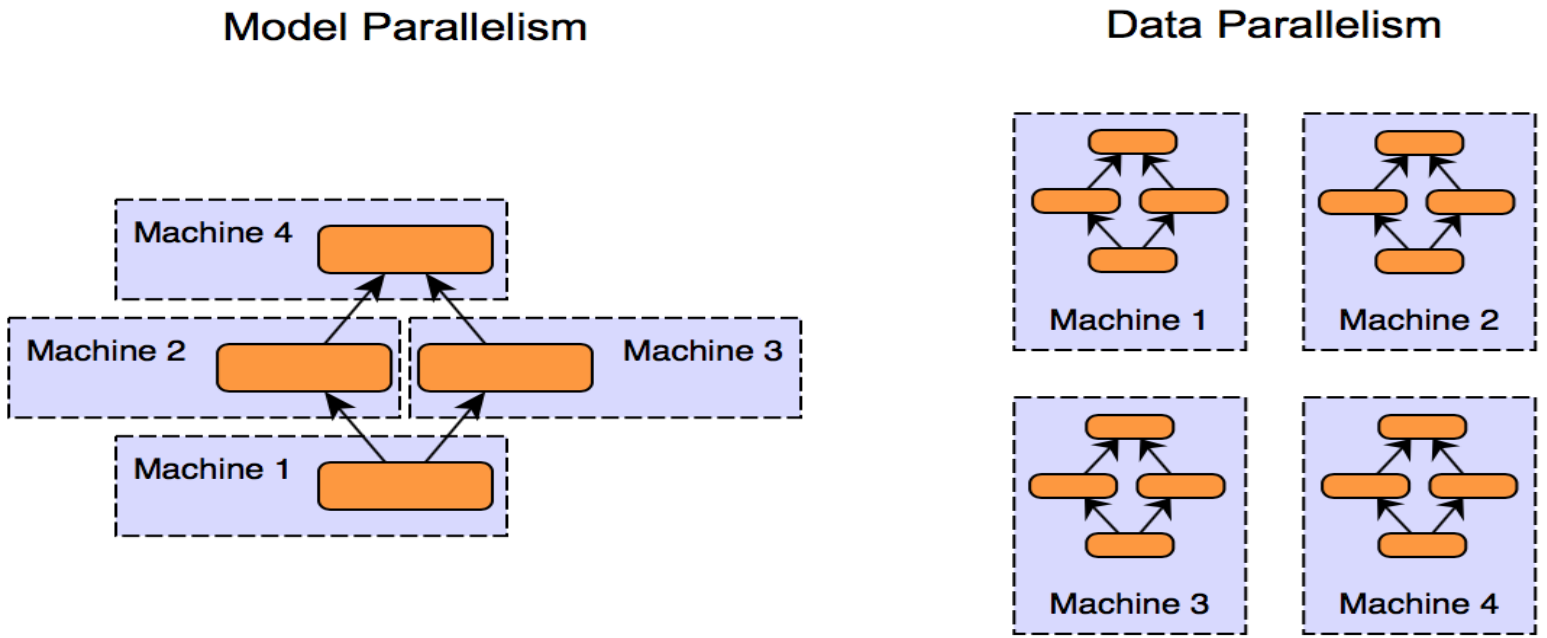
[13] Kyi Thar, Thant Zin Oo, Yan Kyaw Tun, Do Hyeon Kim, Ki Tae Kim, and Choong Seon Hong, "A Deep Learning Model Generation Framework for Virtualized Multi-access Edge Cache Management,"

Cloud to Edge AI

- Classical ML exerts severe demands in terms of energy, memory and computing resources, limiting their adoption for resource constrained edge devices.
- The new breed of intelligent devices and high-stake applications (drones, augmented/virtual reality, autonomous systems, etc.), requires a novel paradigm change calling for distributed, low-latency and reliable ML at the wireless network edge (referred to as edge ML).

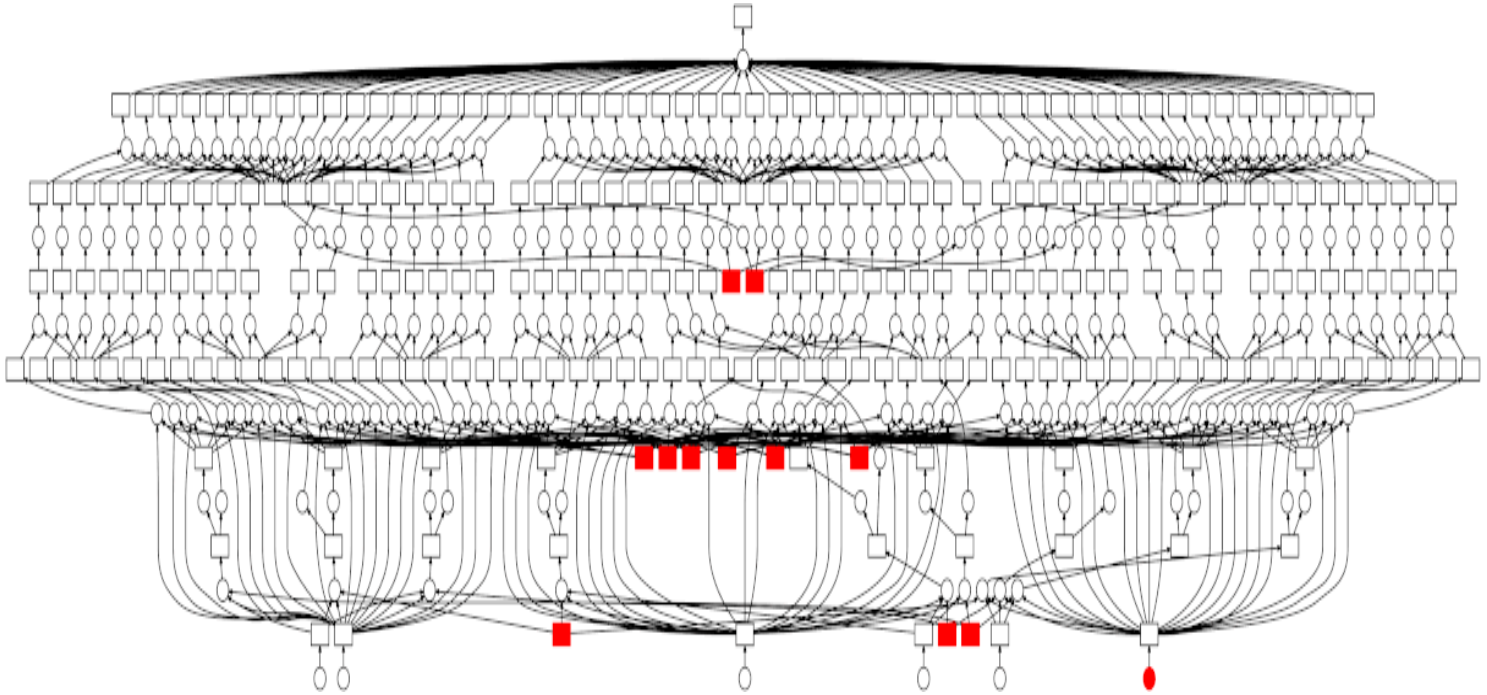
Distributed Training

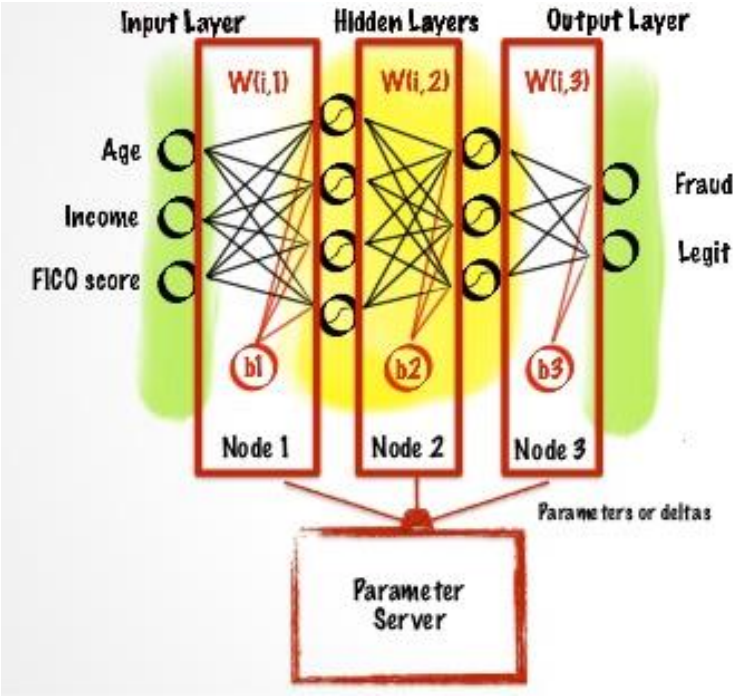
- **Model Parallelism:** When an Deep Learning model size is too large, a single Deep Learning structure can be split into multiple segments that are distributed over multiple devices, i.e., model parallelism or split.
- **Data Parallelism:** An Deep Learning training process can be split by parallelizing the training data samples to multiple devices that have an identical deep learning model structure, referred to as data parallelism or split.



[14] https://xiandong79.github.io/Intro-Distributed-Deep-Learning

Model Parallelism



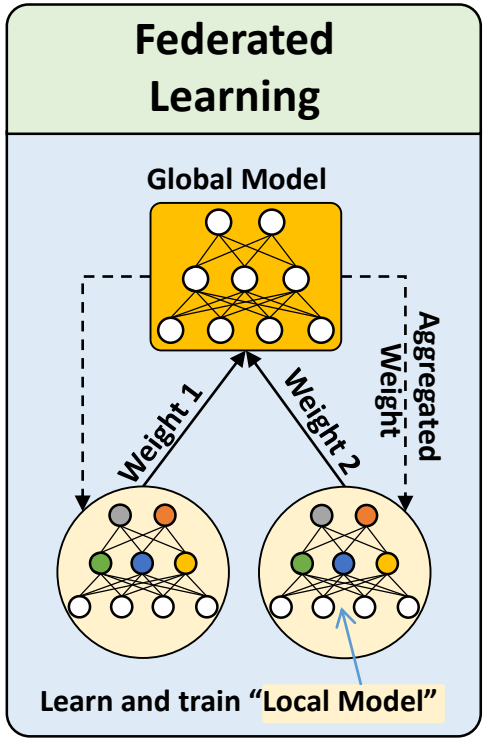


[15] https://jcrist.github.io/dask-sklearn-part-1.html

[16] https://www.slideshare.net/MateuszDymczyk1/deep-learning-at-scale-70024644

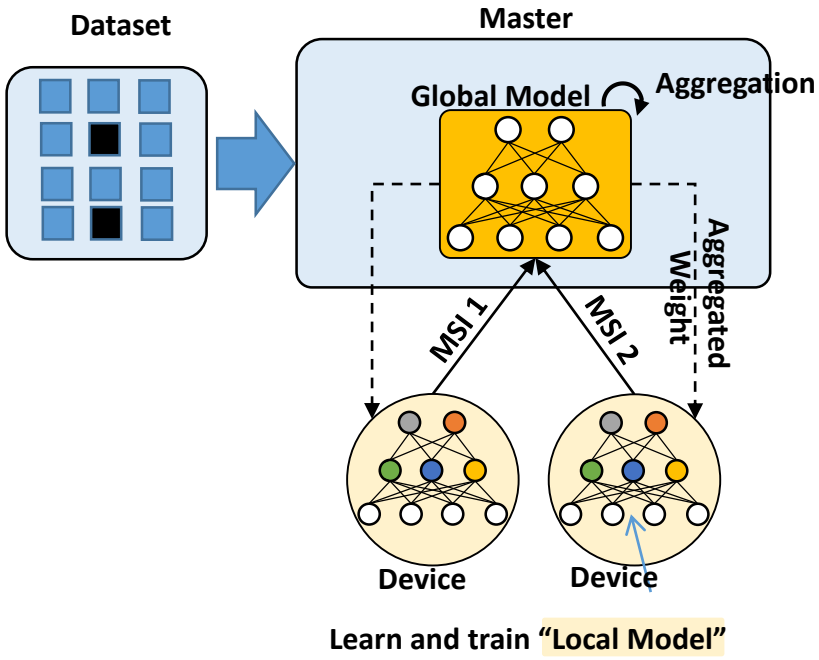
Data Parallelism to Federated Learning (FL)

- Standard machine learning approaches require centralizing the training data on one machine or in a datacenter.
- Federated Learning is a machine learning setting where the goal is to train a high-quality centralized model with training data distributed over a large number of clients each with unreliable and relatively slow network connections.
- Learning algorithms for this setting where on each round, each client independently computes an update (model state information (MSI))to the current model based on its local data, and communicates this update to a central server, where the client-side updates are aggregated to compute a new global model.

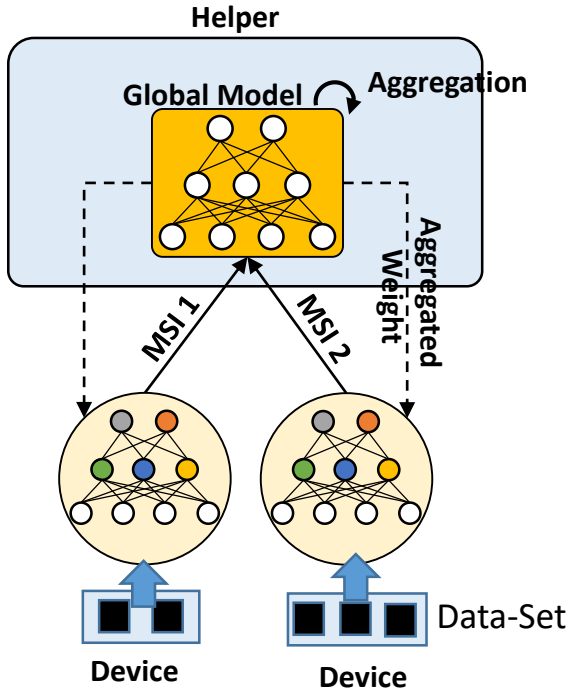


[17] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

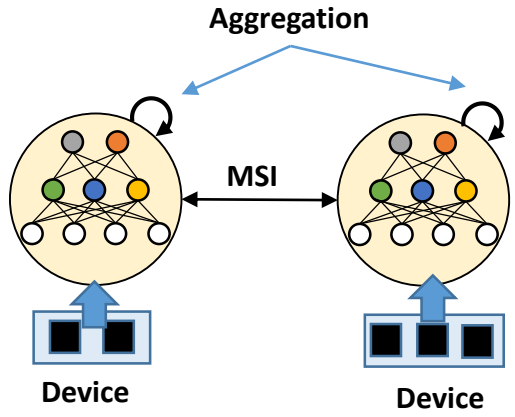
Data Parallelism



(a) Master-Devices



(b) Helper-Devices

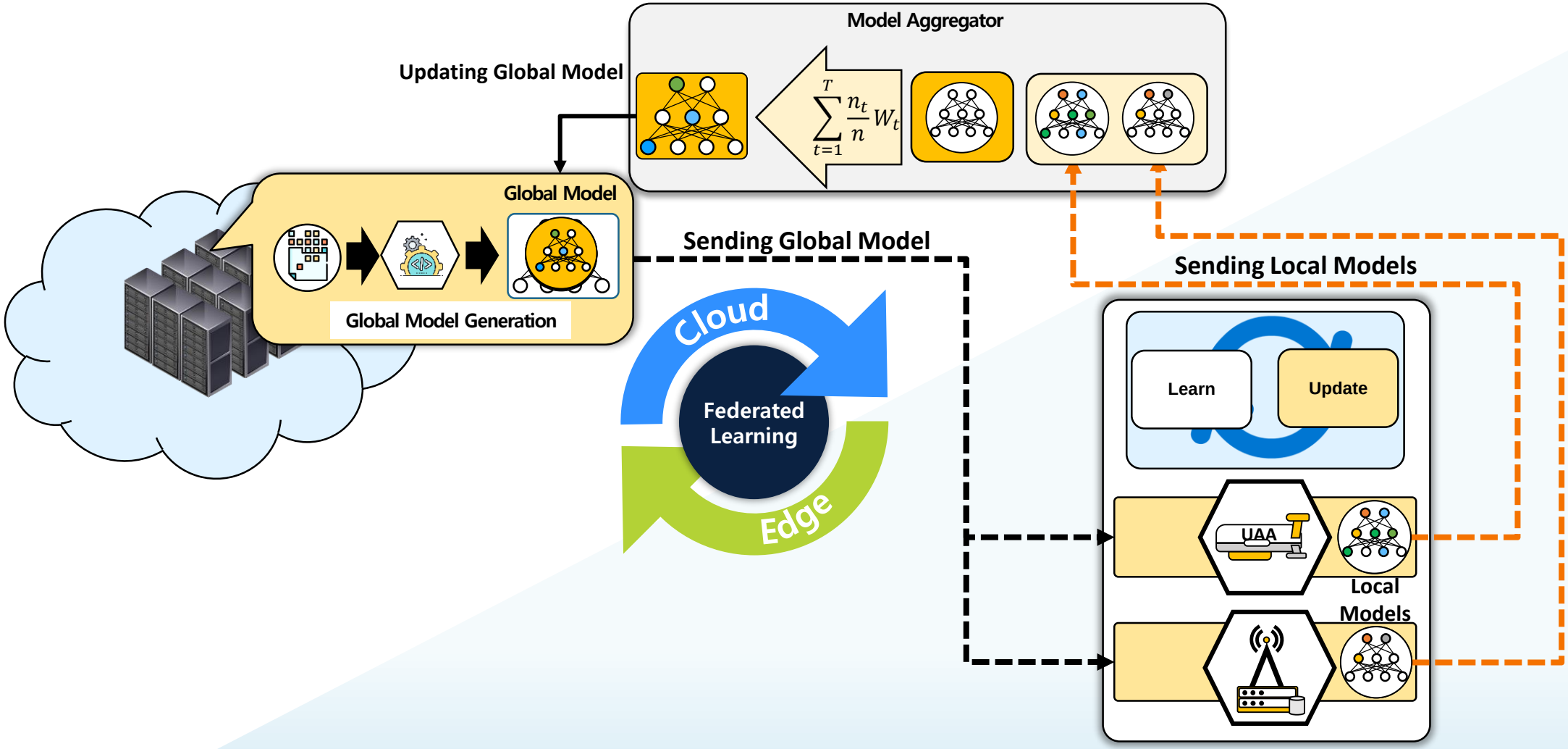


(c) Device-Device

model state information (MSI)

[9] Park, J., Samarakoon, S., Bennis, M. and Debbah, M., 2018. Wireless network intelligence at the edge. *arXiv preprint arXiv:1812.02858*.

Federated Learning (Key technology to enable edge AI)



What impact might FL bring to the AI community?

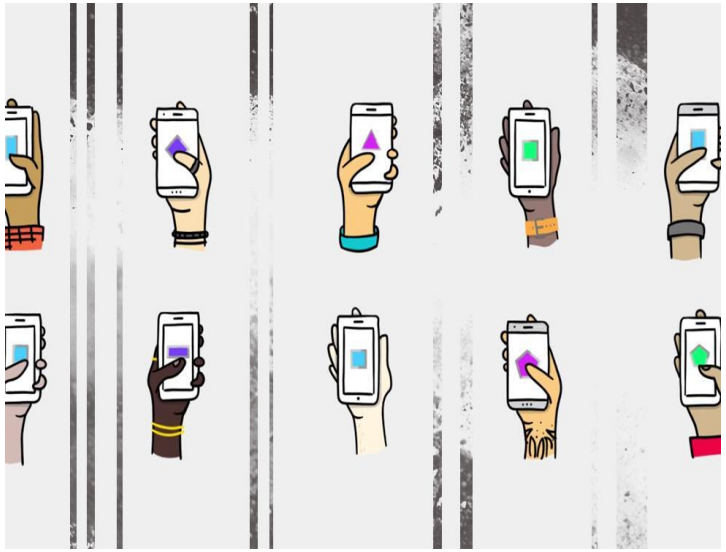
Federated learning opens up a brand new research field in AI.

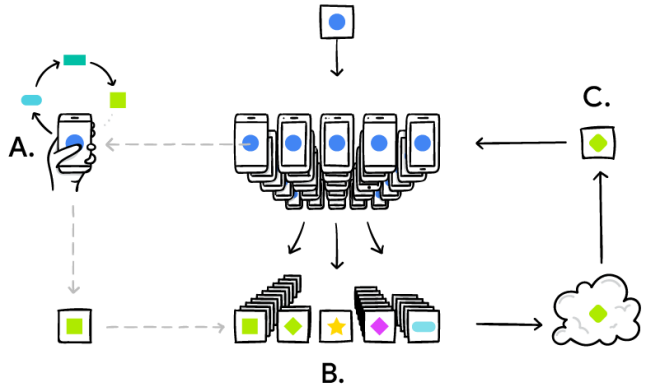
Today, gigantic amounts of data are generated by consumer devices such as mobile phones on a daily basis.

These data contain valuable information about users and their personal preferences: what websites they mostly visited, what social media apps they mostly used, what types of videos they mostly watched, etc.

With such valuable information, these data become the key to building better and personalized machine learning models to deliver personalized services to maximally enhance user experiences.

Federated learning provides a unique way to build such personalized models without intruding users' privacy.





[18] https://medium.com/syncedreview/federated-learning-the-future-of-distributed-machine-learning-eec95242d897

Unique Characteristics of Federated Learning

Federated Learning [19] might similar to distributed machine learning on a technical level, there are some major differences to applications in data centers where the training data is distributed among many machines [20].

- ❑ Huge number of clients
- ❑ Non-identical distributions
- ❑ Unbalanced number of samples
- ❑ Slow and unstable communication

[19] https://florian.github.io/federated-learning/

[20] Konečný, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T. and Bacon, D., 2016. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

How is Federated Learning and improvement to AI?

- Personal data never leaves the user's device, only updates made to the model are transferred. This data is encrypted making it impossible for anyone to intercept the data and retro engineer it.
- The updates are lighter than the original user's data. Consequently the overall workload needed is lower in Federated Learning than in cloud based architectures or in edge computing, which makes it cheaper and more convenient.
- The model is located in the user's device, allowing for real time inferences with no latency problems.

	Centralized Learning	Edge Computing	Federated Learning
Privacy	X	X	✓
Bandwidth	X	✓	✓
Latency	X	✓	✓
Cost/ Feasibility	✓	X	✓

[21] https://medium.com/frstvc/otium-neural-newsletter-1-federated-learning-a-step-closer-towards-confidential-ai-efe28832006f

Potential future trends related to FL?

- Federated learning is revolutionizing how machine learning models are trained.
- Google has just released their first production-level federated learning platform, which will spawn many federated learning-based applications such as on-device item ranking, next-word prediction, and content suggestion.
- In the future, machine learning models can be trained without counting on the compute resources owned by giant AI companies.
- And users will not need to trade their privacy for better services.

[22] https://medium.com/syncedreview/federated-learning-the-future-of-distributed-machine-learning-eec95242d897

Limitations of Federated Learning

- Maintaining massively distributed systems
- Limited connectivity to all the devices at all the time
- Unbalanced data in terms of bias or feedback. This problem however can be reduced to a certain extent by smartly selecting devices from which to get a feedback at a given moment.
- Developing an infrastructure or models which can keep up with the pace of the dynamic and continuous learning involved in the approach
- Running optimization algorithms across highly distributed data sets

[23] Ammad-ud-din, M., Ivannikova, E., Khan, S.A., Oyomno, W., Fu, Q., Tan, K.E. and Flanagan, A., 2019. Federated Collaborative Filtering for Privacy-Preserving Personalized Recommendation System. *arXiv preprint arXiv:1901.09888*.

Model Aggregation Optimization

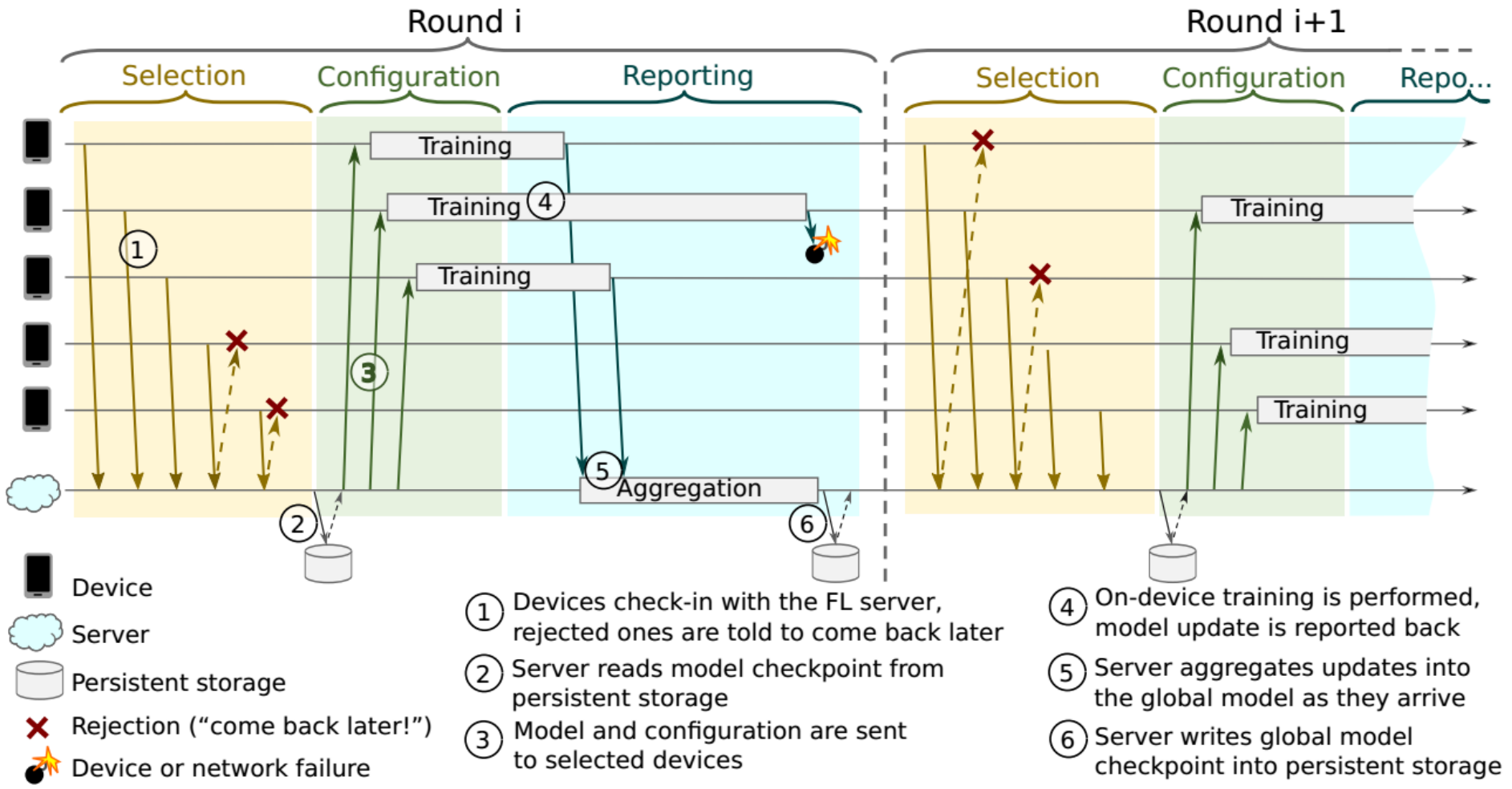


Figure 1: Federated Learning Protocol

$$\sum_{t=1}^{T} \frac{n_t}{n} W_t$$

Federated averaging

[24] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konecny, J., Mazzocchi, S., McMahan, H.B. and Van Overveldt, T., 2019. Towards Federated Learning at Scale: System Design. *arXiv preprint arXiv:1902.01046*.

Model Aggregation Optimization

Algorithm 1 `FederatedAveraging` targeting updates from K clients per round.

Server executes:

initialize w_0

for each round $t = 1, 2, \ldots$ **do**

Select $1.3K$ eligible clients to compute updates

Wait for updates from K clients (indexed $1, \ldots, K$)

$(\Delta^k, n^k) = \text{ClientUpdate}(w)$ from client $k \in [K]$.

$\bar{w}_t = \sum_k \Delta^k$ *// Sum of weighted updates*

$\bar{n}_t = \sum_k n^k$ *// Sum of weights*

$\Delta_t = \Delta_t^k / \bar{n}_t$ *// Average update*

$w_{t+1} \leftarrow w_t + \Delta_t$

ClientUpdate(w)**:**

$\mathcal{B} \leftarrow$ (local data divided into minibatches)

$n \leftarrow |\mathcal{B}|$ *// Update weight*

$w_{\text{init}} \leftarrow w$

for batch $b \in \mathcal{B}$ **do**

$w \leftarrow w - \eta \nabla \ell(w; b)$

$\Delta \leftarrow n \cdot (w - w_{\text{init}})$ *// Weighted update*

// Note Δ is more amenable to compression than w

return (Δ, n) to server

[24] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konecny, J., Mazzocchi, S., McMahan, H.B. and Van Overveldt, 2019. Towards Federated Learning at Scale: System Design. *arXiv preprint arXiv:1902.01046*.

Model Aggregation Optimization

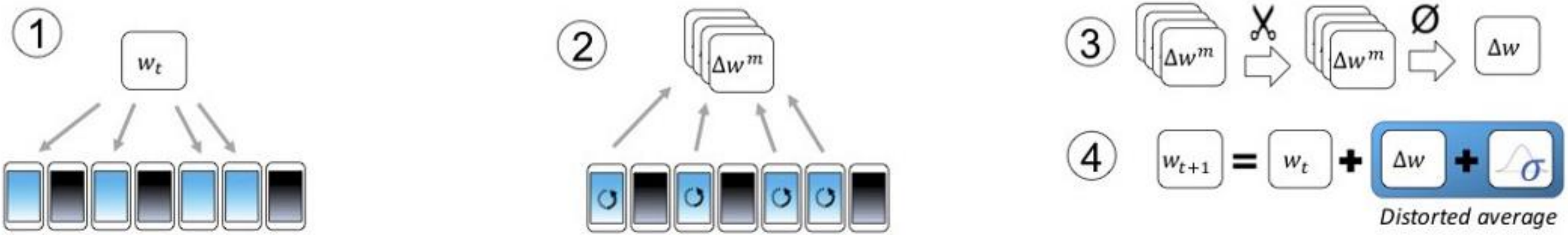
Differentially Private Federated Learning: A Client Level Perspective

1. Incorporating a randomized mechanism into a communication round of Federated Learning

- Sampling m_t out of K clients at round t
- Distributing central model w_t

- Clients optimize w_t on their data, leading to w^k
- Updates $\Delta w^k = w^k - w_t$ are centralized

- Update with clipped averaged & distorted models



2. Privacy accountant; called before each communication round

- σ_t *and* m_t determine increase of δ during a communication round
- If δ_{t+1} stays below threshold $\hat{\delta}$ for a certain σ_t *and* m_t, a new round may start

④ $$w_{t+1} = w_t + \frac{1}{m}\sum_{k=1}^{m} \Delta w^k + N\left(0, \frac{S^2 {\sigma_t}^2}{m_t}\right)$$ ➡ ${\sigma_t}^2 / m_t$

How should we choose ${\sigma_t}^2/m_t$ for all t, such that model performance is maximized throughout the course of training, while δ stays smaller than $\hat{\delta}$.

t	: communication round index
m_t	: number of subsampled clients at round t
σ_t	: noise parameter at round t
δ_t	: probability that ϵ-diff. privacy is broken at round t
w_t	: central model
w^k	: model optimized by client k

[25] Geyer, R.C., Klein, T. and Nabi, M., 2017. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*.

Federated Learning with Blockchain

- Machine learning models trained on data from blockchain-based marketplaces have the potential to create the world's most powerful artificial intelligences.
- They combine two potent primitives: private machine learning, which allows for training to be done on sensitive private data without revealing it, and blockchain-based incentives, which allow these systems to attract the best data and models to make them smarter.
- The result is open marketplaces where anyone can sell their data and keep their data private, while developers can use incentives to attract the best data for their algorithms to them.

[26] https://medium.com/@FEhrsam/blockchain-based-machine-learning-marketplaces-cb2d4dae2c17?fbclid=IwAR2acQ2t143gp4chJTPqcJMClTVpqPaLGv8xuxc8rFJwdFqQIn5reK5O7PM

Federated Learning with Blockchain

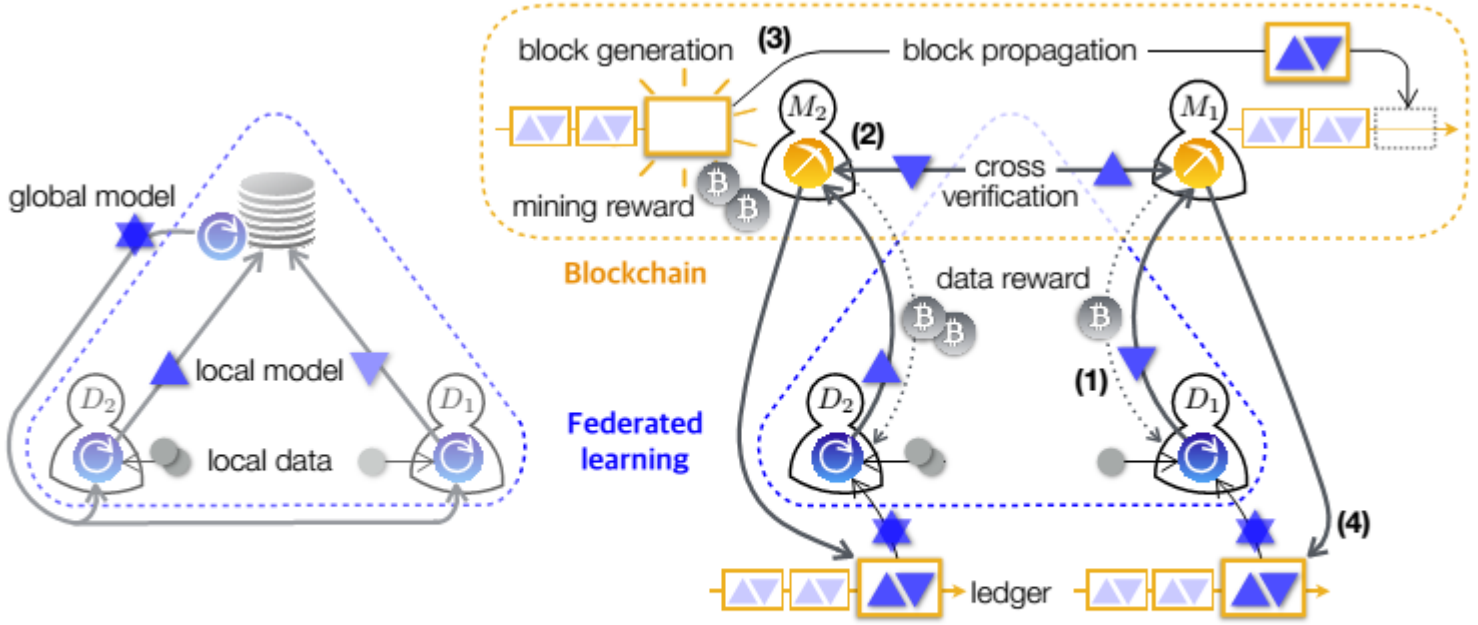
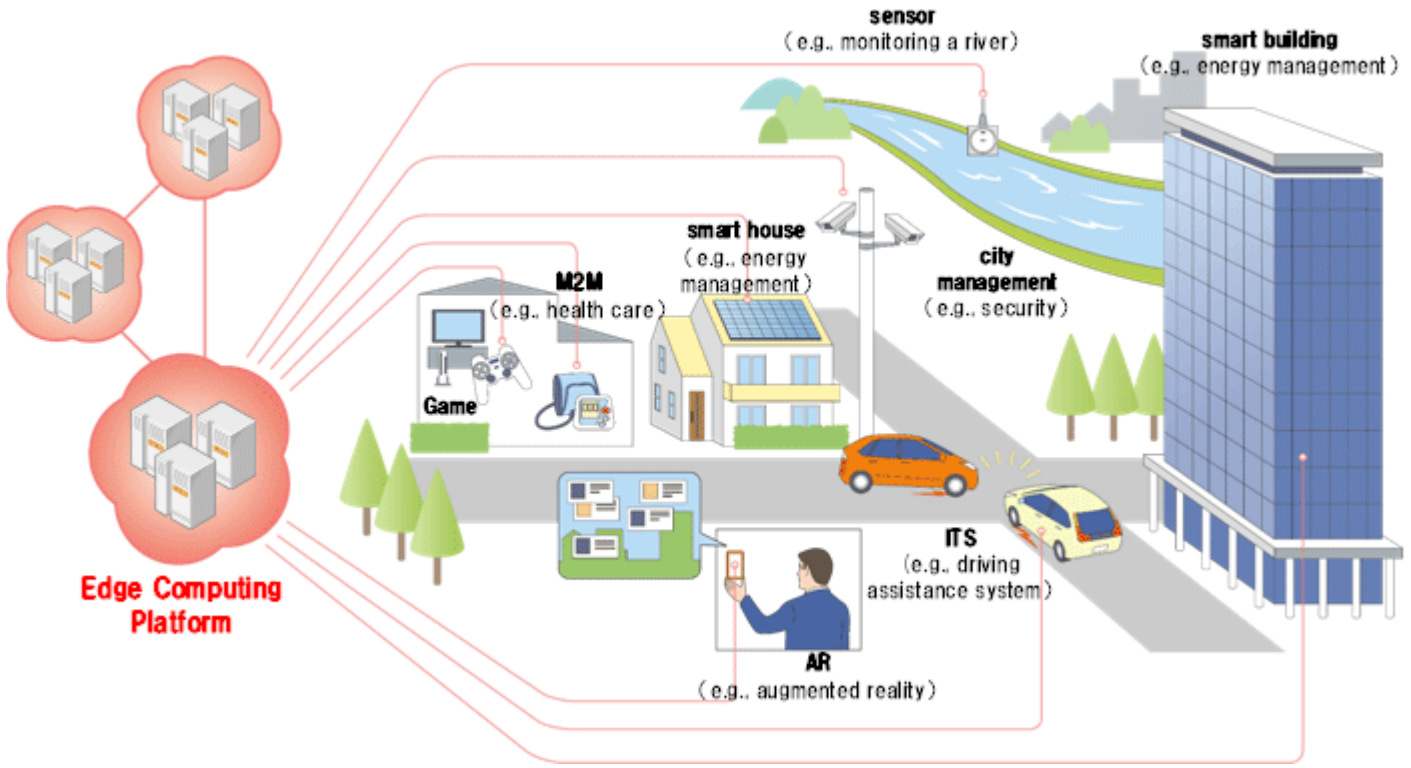


Fig. 1. An illustration of (a) the vanilla federated learning (FL) [4], [5] and (b) the proposed block-chained FL (BlockFL) architectures.

Receiving the reward proportional to the number of its data samples to locally update the model.

[27] Kim, H., Park, J., Bennis, M. and Kim, S.L., 2018. On-device federated learning via blockchain and its latency analysis. arXiv preprint arXiv:1808.03949.

Edge AI Use Cases in Edge Analytics Application



What is analytics?

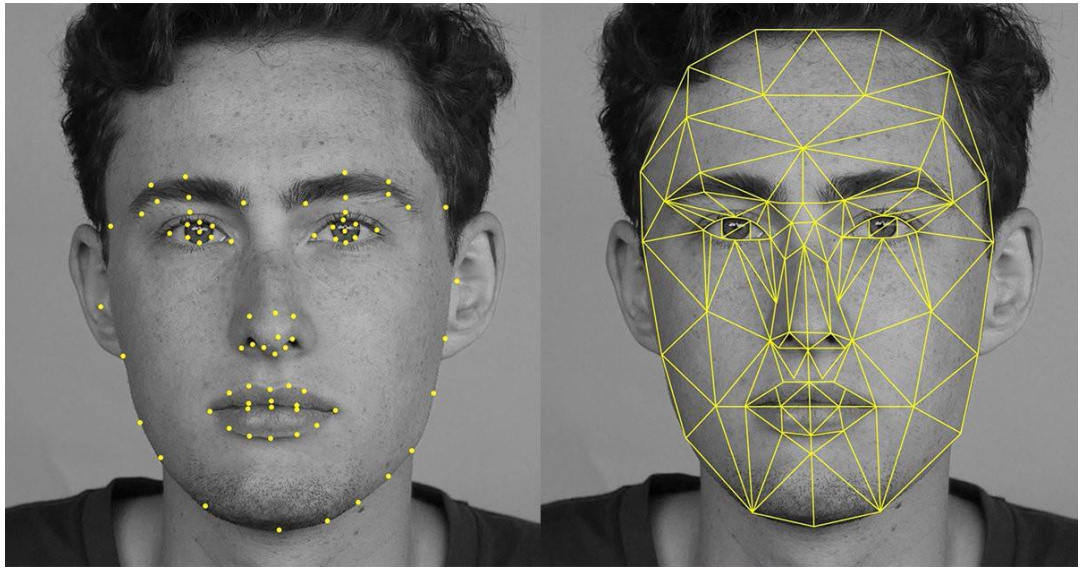
"Analytics" refers to the systematic use of technologies, methods, and data to derive insights and enable fact-based decision-making for planning, management, operations, measurement, and learning.

[28] https://www.talend.com/resources/edge-analytics-pros-cons-immediate-local-insight/

Edge AI Use Cases in Edge Analytics Application

Image analytics :

- Image analytics is a classic AI application area. The availability of huge numbers of images on the web and of pre-classified data sets has recognition of various object types.
- For example, real-time recognition of a constantly changing scene based on video streaming requires high data bandwidth if performed in the cloud. Alternatively, AI on the Edge enables local analysis of the visual scene in various flavors, such as understanding the scene for context analysis, simultaneous multi-object detection and recognition for obstacle avoidance, people identification for secure access, and more.

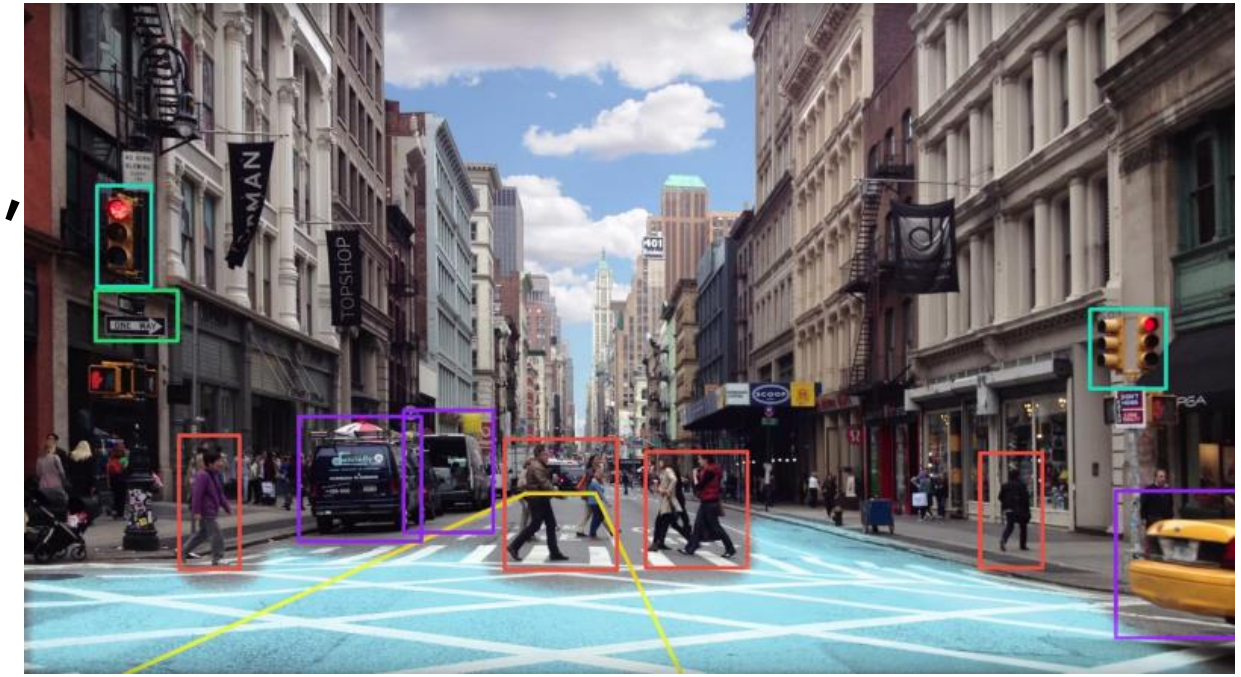


Edge AI Use Cases in Edge Analytics Application

- Surveillance and Monitoring: Deep Learning-enabled smart cameras could locally process captured images to identify and track multiple objects and people, detecting suspicious activities directly on the edge node.
- Smart cameras minimize communication with the remote servers by only sending data on a triggering event, also reducing remote processing and memory requirements.
- Intruder monitoring for secure homes and monitoring of elderly people are typical applications.

Edge AI Use Cases in Edge Analytics Application

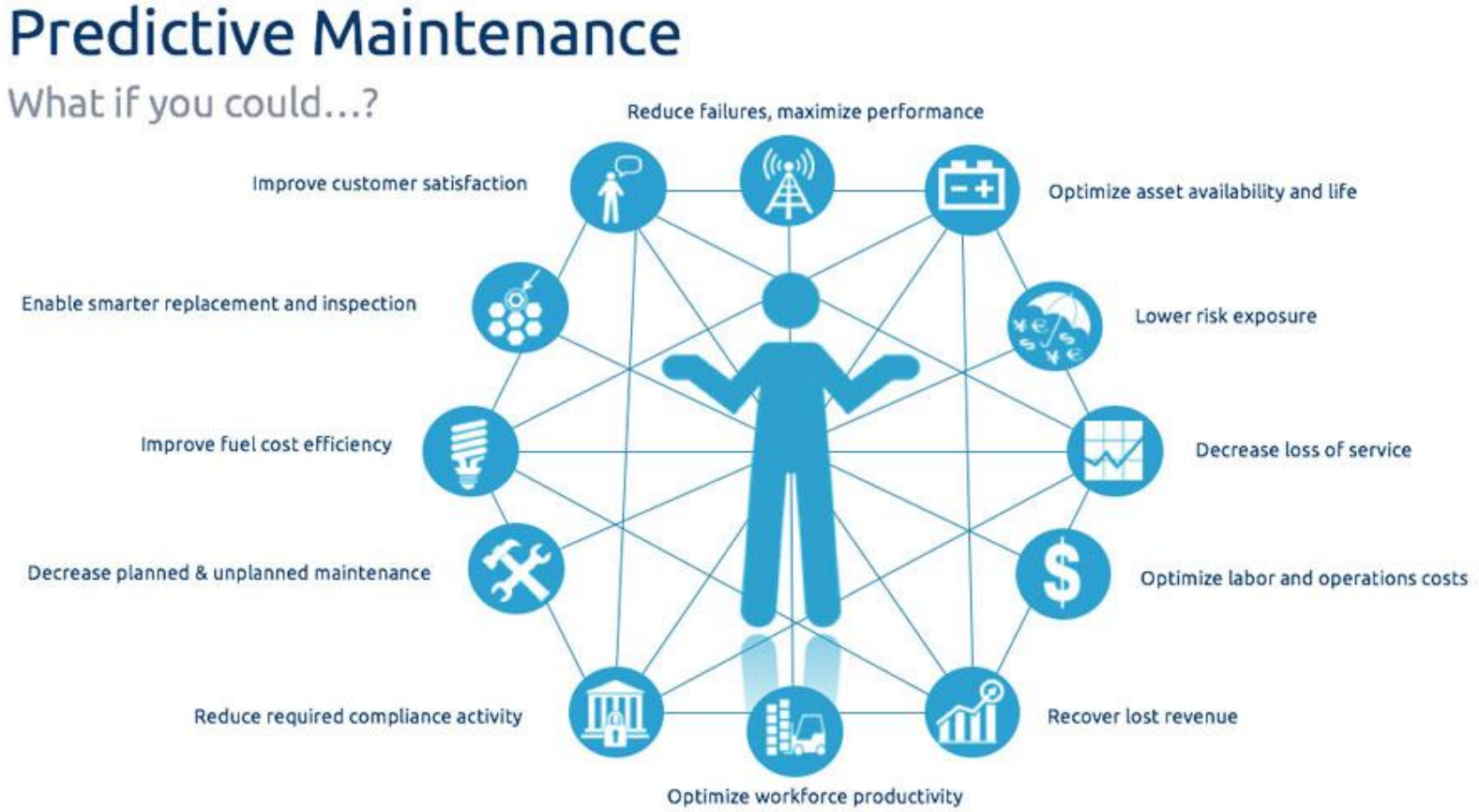
- Autonomous Vehicles: A smart automotive camera can recognize vehicles, traffic signs, pedestrian, road, and objects locally, sending only information needed to perform autonomous driving to the main controller.
- A similar concept can be applied to robots and drones.



Edge AI Use Cases in Edge Analytics Application

Predictive Maintenance in Factories:

- Sensors attached to a machine can measure vibration, temperature, and noise levels and AI performed locally can infer the state of the equipment, potential anomalies, and early indications of failure. In this case, local Deep Learning could also communicate with cloud-based services to deliver data for specific analyses and corrective actions.



Edge AI Use Cases in Edge Analytics Application

Table 1. Trade-offs of the different types of maintenance

	Benefits	Challenges
Reactive	• Maximum utilization of tooling or machine components	• Potentially greater damage to machine beyond failed part • Unplanned downtime • Higher maintenance costs
Planned	• Less likelihood of broken machinery • Less unplanned downtime • More cost-effective than reactive	• Increased replacement costs over time • Need for additional spare parts inventory • Increased planned downtime
Proactive	• Longer lifespan of equipment • Decreased downtime, planned and unplanned • More cost-effective than run-to-failure or planned maintenance • Lower spare parts inventory	• Ongoing maintenance and monitoring • Need for organizational changes • Increased training

Source: Deloitte analysis.

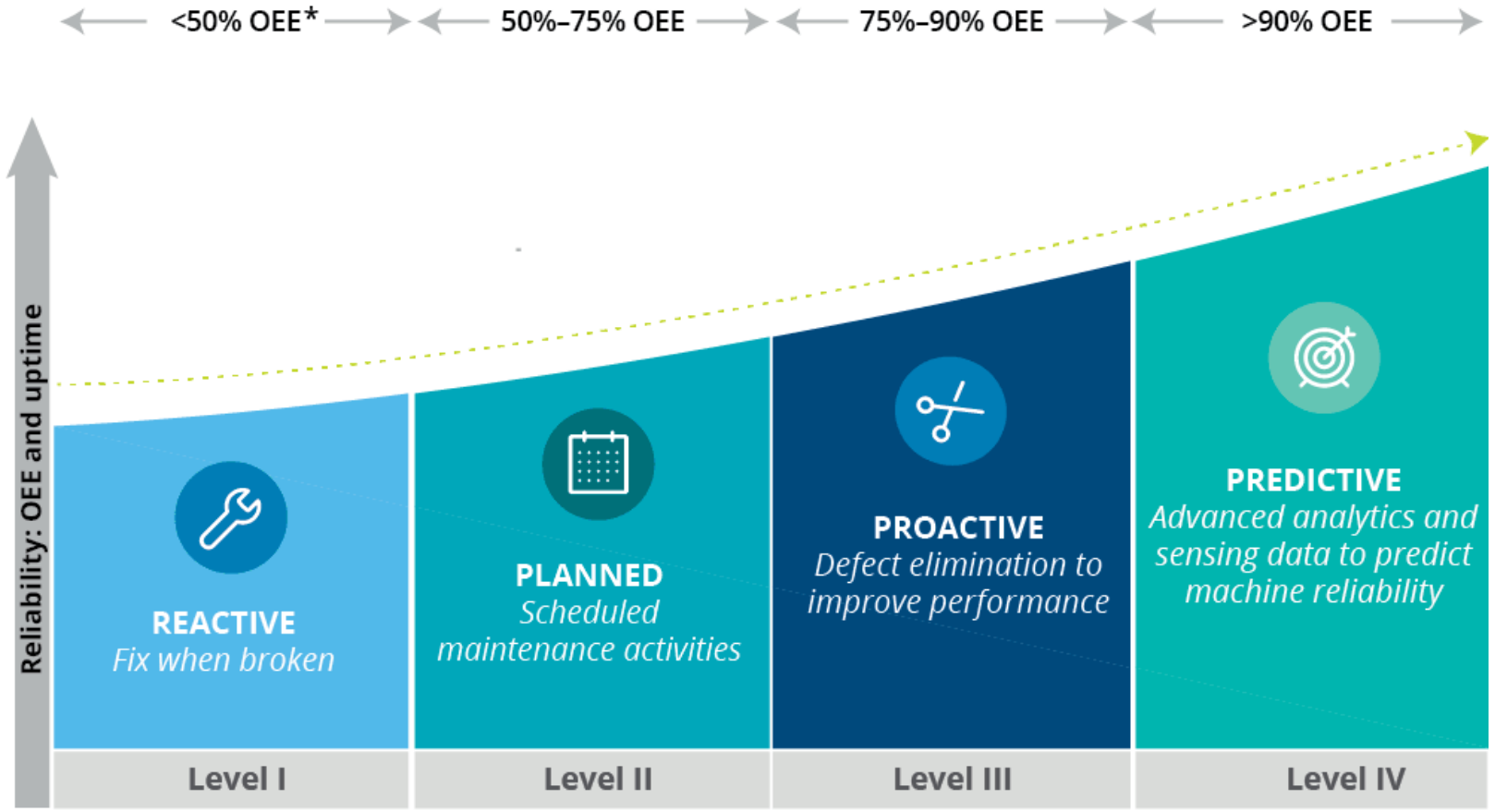
Deloitte University Press | dupress.deloitte.com

[29] https://www2.deloitte.com/insights/us/en/focus/industry-4-0/using-predictive-technologies-for-asset-maintenance.html

Edge AI Use Cases in Edge Analytics Application

Edge Analysis in Smart Factory

Figure 1. Maintenance strategy continuum



* Original equipment effectiveness (OEE)

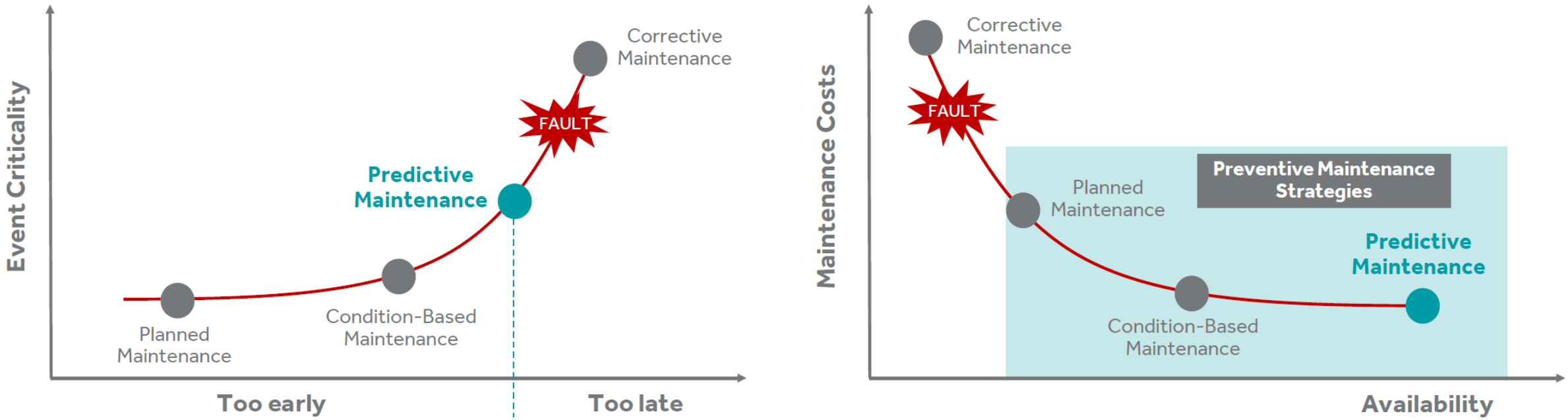
Source: Deloitte analysis.

Deloitte University Press | dupress.deloitte.com

[29] https://www2.deloitte.com/insights/us/en/focus/industry-4-0/using-predictive-technologies-for-asset-maintenance.html

Edge AI Use Cases in Edge Analytics Application

Edge Analysis in Smart Factory



This data gives valuable insights regarding engine life, service history, recommended versus actual fuel levels, etc. which are constantly monitored in real time to ensure efficient operations. Predictive maintenance enhances coordination between maintenance team and supervisors, thus assisting them with effective decision making, backed by data. The decision makers can then take a call whether to expedite a service or look for more information. These solutions not only help cut maintenance costs but more importantly avert a crisis leading to enhanced customer satisfaction and stickiness.

[30] https://www.gslab.com/blog-post/predictive-maintenance/

Conclusion

- Edge AI can be deployed by combining Federated Learning, Block Chain, and Edge Computing
- Previously, powerful AI apps required large, expensive data center-class systems to operate. But edge computing devices can reside anywhere, as demonstrated in the above use cases.
- AI at the edge offers endless opportunities that a can help society in ways never before imagined.

References

- [1] https://www.youtube.com/watch?v=UQcJSZPpNhA&feature=youtu.be
- [2] S.-C. Lin, Y. Zhang, C.-H. Hsu, M. Skach, M. E. Haque, L. Tang, and J. Mars, "The architectural implications of autonomous driving: Constraints and acceleration," in Proc. of the 23rd ACM ASPLOS, ASPLOS '18, (Williamsburg, VA, USA), pp. 751–766, ACM, Mar. 2018.
- [3] M. K. Abdel-Aziz, C.-F. Liu, S. Samarakoon, M. Bennis, and W. Saad, "Ultra-reliable low-latency vehicular networks: Taming the age of information tail," in Proc. of GLOBECOM [accepted], (Abu Dhabi, UAE), Dec. 2018.
- [4] J. Park and M. Bennis, "URLLC-eMBB slicing to support VR multimodal perceptions over wireless cellular systems," ArXiv preprint, vol. abs/1805.00142, May 2018.
- [5] ABI Research and Qualcomm, "Augmented and virtual reality: The first wave of 5g killer apps," white paper, Feb. 2017.
- [6] T. Kagawa, F. Ono, L. Shan, K. Takizawa, R. Miura, H. Li, F. Kojima, and S. Kato, "A study on latency-guaranteed multihop wireless communication system for control of robots and drones," in Proc. of 20th WPMC, (Yogyakarta, Indonesia), pp. 417– 421, Dec. 2017.
- [7] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs," IEEE Transactions on Wireless Communications, vol. 15, pp. 3949–3963, June 2016.
- [8] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. Galati Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," ArXiv preprint, vol. abs/1809.01752, Sept. 2018.
- [9] Park, J., Samarakoon, S., Bennis, M. and Debbah, M., 2018. Wireless network intelligence at the edge. *arXiv preprint arXiv:1812.02858.*
- [10] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function ap proximation," in Proc. of the 12th NIPS, NIPS'99, (Colorado, USA), pp. 1057–1063, MIT Press, Dec. 1999
- [11] I. Goodfellow, Y. Bengio and A. Courville, "Deep Learning", in MIT Press, 2016.
- [12] Kyi Thar, Nguyen H. Tran, Thant Zin Oo, Choong Seon Hong, "DeepMEC: Mobile Edge Caching Using Deep Learning," IEEE Access, Vol.6, Issue 1, pp.78260-78275, December 2018
- [13] Kyi Thar, Thant Zin Oo, Yan Kyaw Tun, Do Hyeon Kim, Ki Tae Kim, and Choong Seon Hong, "A Deep Learning Model Generation Framework for Virtualized Multi-access Edge Cache Management,"
- [14] https://xiandong79.github.io/Intro-Distributed-Deep-Learning
- [15] https://jcrist.github.io/dask-sklearn-part-1.html

References

- [16] https://www.slideshare.net/MateuszDymczyk1/deep-learning-at-scale-70024644
- [17] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
- [18] https://medium.com/syncedreview/federated-learning-the-future-of-distributed-machine-learning-eec95242d897
- [19] https://florian.github.io/federated-learning/
- [20] Konečný, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T. and Bacon, D., 2016. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.
- [21] https://medium.com/frstvc/otium-neural-newsletter-1-federated-learning-a-step-closer-towards-confidential-ai-efe28832006f
- [22] https://medium.com/syncedreview/federated-learning-the-future-of-distributed-machine-learning-eec95242d897
- [23] Ammad-ud-din, M., Ivannikova, E., Khan, S.A., Oyomno, W., Fu, Q., Tan, K.E. and Flanagan, A., 2019. Federated Collaborative Filtering for Privacy-Preserving Personalized Recommendation System. *arXiv preprint arXiv:1901.09888*.
- [24] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konecny, J., Mazzocchi, S., McMahan, H.B. and Van Overveldt, T., 2019. Towards Federated Learning at Scale: System Design. *arXiv preprint arXiv:1902.01046*.
- [25] Geyer, R.C., Klein, T. and Nabi, M., 2017. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*.
- [26] https://medium.com/@FEhrsam/blockchain-based-machine-learning-marketplaces-cb2d4dae2c17?fbclid=IwAR2acQ2t143gp4chJTPqcJMClTVpqPaLGv8xuxc8rFJwdFqQln5reK5O7PM
- [27] Kim, H., Park, J., Bennis, M. and Kim, S.L., 2018. On-device federated learning via blockchain and its latency analysis. arXiv preprint arXiv:1808.03949.
- [28] https://www.talend.com/resources/edge-analytics-pros-cons-immediate-local-insight/
- [29] https://www2.deloitte.com/insights/us/en/focus/industry-4-0/using-predictive-technologies-for-asset-maintenance.html
- [30] https://www.gslab.com/blog-post/predictive-maintenance/

Thank You!

Q & A