Getting to the core of Edge ML and TinyML

Filipa Peleja
Lead Data Scientist, Levi Strauss & Co.

Around 10% of enterprise-generated data is created and processed outside a traditional centralized data center or cloud.
By 2025, Gartner predicts this figure will reach 75%

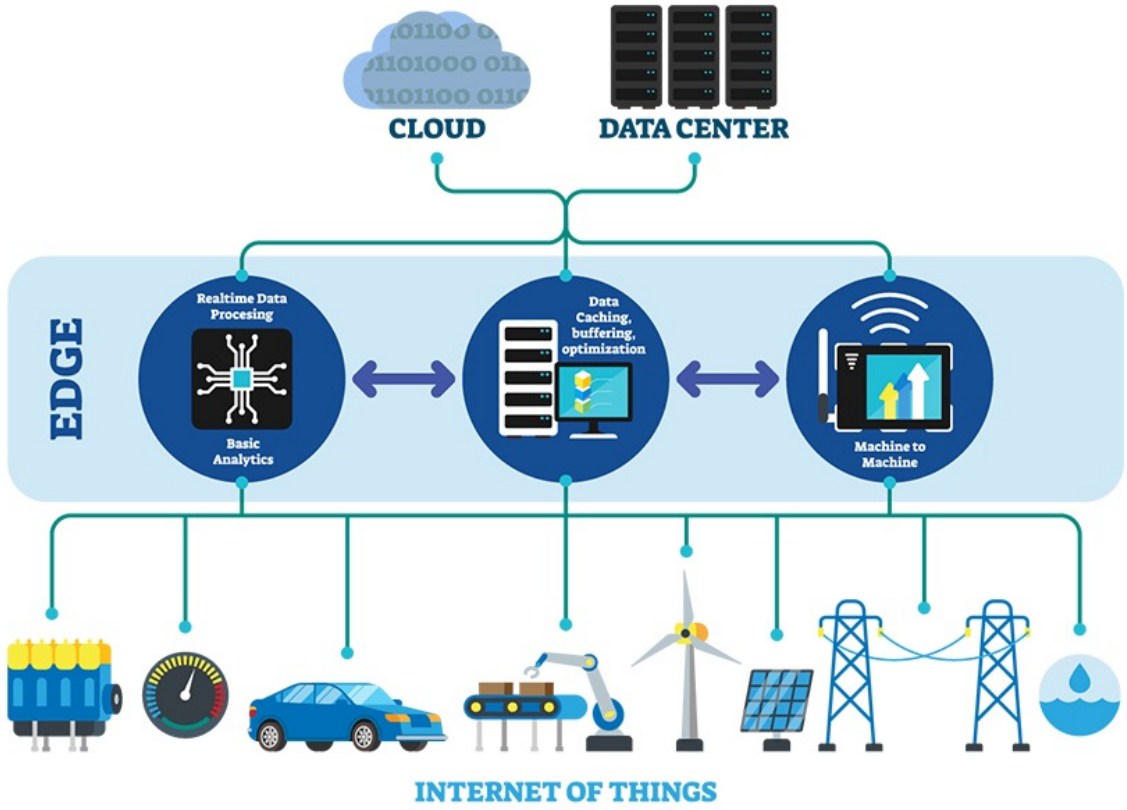
Gartner, 2018
Rob van der Meulen

https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders

AI on the edge is not new
but ***not popular either***

* *What is edge computing?*
* *How edge computing relates to AI?*
* *Why aren't we using it?*

Edge Computing

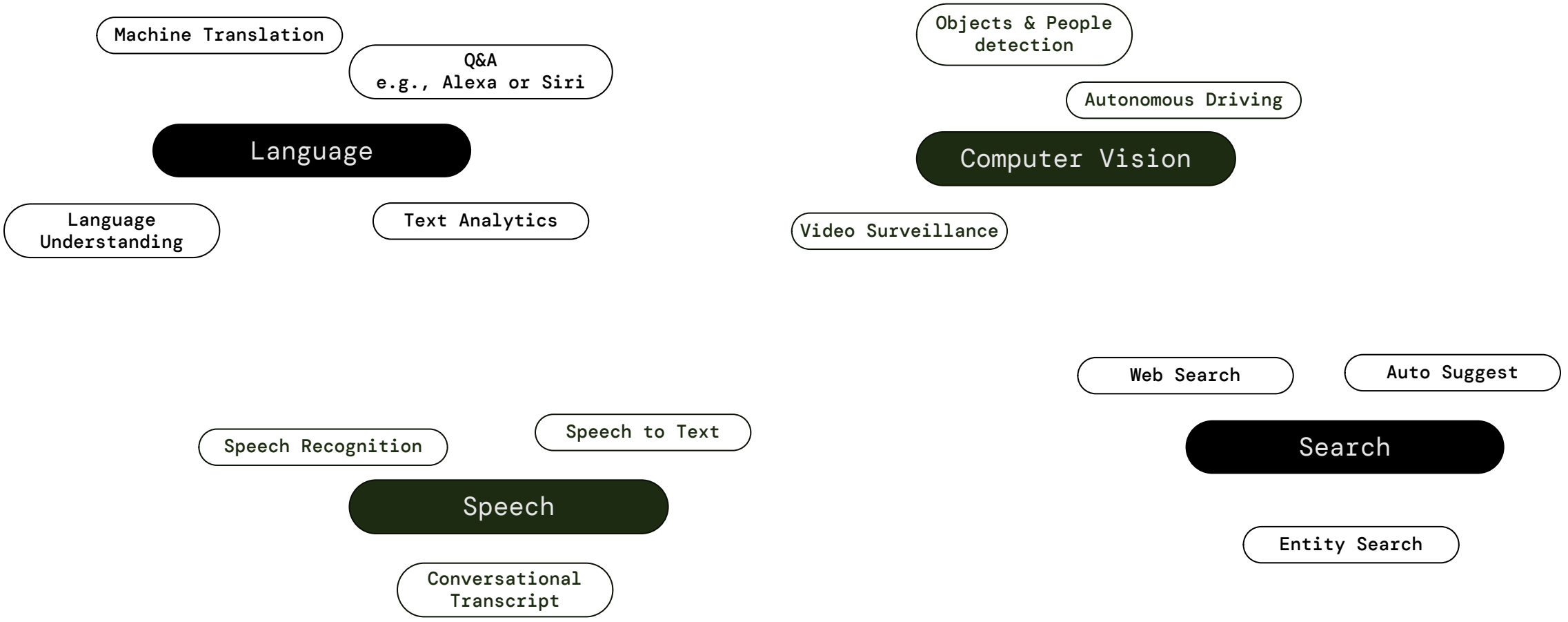


https://innovationatwork.ieee.org/real-life-edge-computing-use-cases/

- Connectivity
- IP Security
- Offline Support
- Bandwidth
- Resource Utilization
- Latency

AI & ML

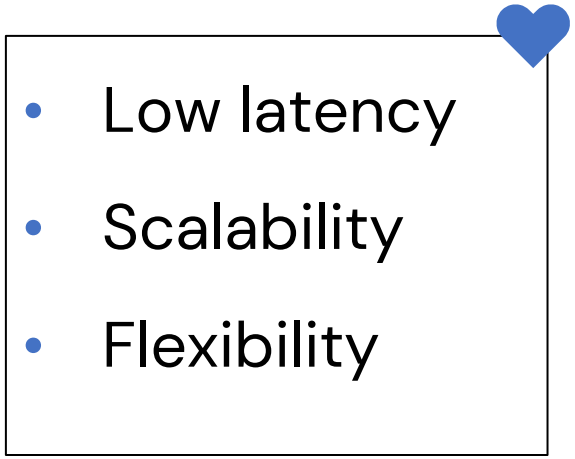
Artificial Intelligence & Machine Learning



AI solves hard problems

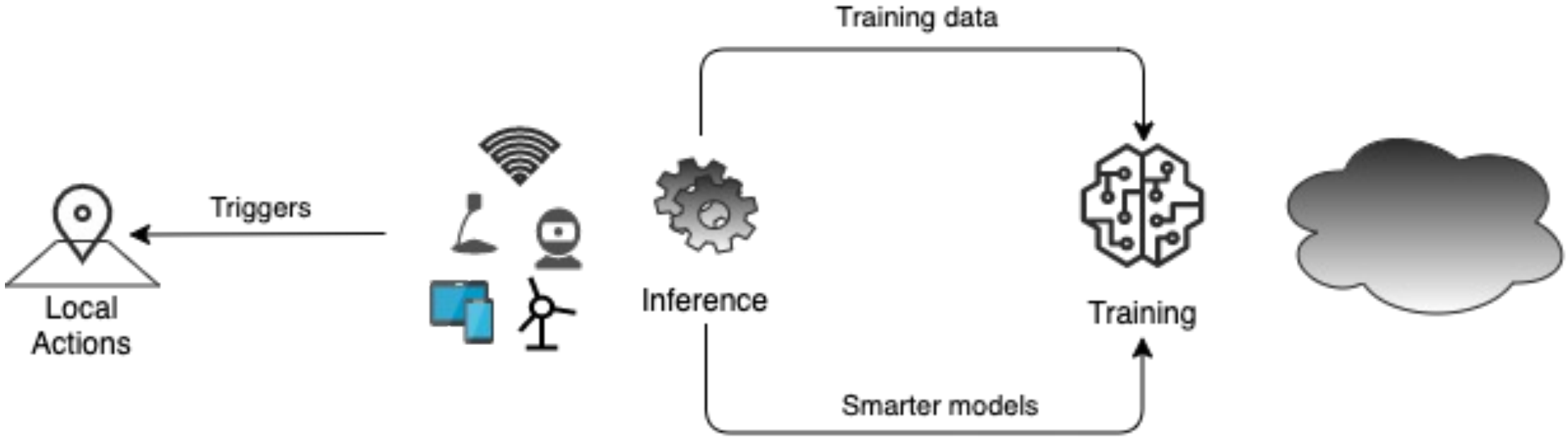
Edge computing & AI

- Autonomous Vehicles
- Healthcare Devices
- Security Solutions
- Retail Advertising
- Smart Speakers
- Smart Agriculture
- More..



Edge computing & AI

Extend Intelligence to the edge



Edge computing & AI

Extend Intelligence to the edge

- Devices act locally based on the data they generate
  - **Continues to benefit from cloud**, but overcomes part of its limitations by acting locally

- HOW?
  - Build & train Machine Learning models in the cloud
  - Deploy optimized models in the target device
  - Accelerate inference applications **on the edge**
  - Devices take action quickly even when disconnected

Edge computing & AI

Extend Intelligence to the edge

- Train in the cloud
  - Computing power
    - E.g., access to GPU from lambda functions to speed-up inference
  - Large volumes of data

- Inference at the edge[1]
  - Low latency
  - Bandwidth saving

[1] Regulation and privacy can and should be secured at the edge

Top Edge Computing Companies

Edge computing

Some of the top Edge Computing Companies[1]

- AWS Greengrass ML
- Google TensorFlow Lite
- EdgeX Foundry
- Azure Stack Edge
- ClearBlade

Company	IoT, ML, and AI at the edge	Security
AWS	✔	✔ AWS cloud security
Azure	✔	✔ Activation keys, passwords, certificates, double encryption, and restricted access
ClearBlade	✔	✔ API access encryption, authentication, and authorization
Dell	✔	✔
EdgeConneX	✔	✔
Section	✔	✔Web application firewall (WAF), bot management, certificates, and IP restrictions and blocking

[1]Top 6 Edge Computing Companies 2022, April 12, 2022
https://www.enterprisenetworkingplanet.com/management/edge-computing-companies/

Edge computing

Some of the top Edge Computing Companies[1]

- AWS Greengrass ML
- Google TensorFlow Lite
- EdgeX Foundry
- Azure Stack Edge
- ClearBlade

Company	IoT, ML, and AI at the edge	Security
AWS	✔	✔ AWS cloud security
Azure	✔	✔ Activation keys, passwords, certificates, double encryption, and restricted access
ClearBlade	✔	✔ API access encryption, authentication, and authorization
Dell	✔	✔
EdgeConneX	✔	✔
Section	✔	✔Web application firewall (WAF), bot management, certificates, and IP restrictions and blocking

[1]Top 6 Edge Computing Companies 2022, April 12, 2022
https://www.enterprisenetworkingplanet.com/management/edge-computing-companies/

Greengrass

AWS IoT Greengrass ML Inference

Open-source edge runtime and cloud service provided by Amazon that helps build, deploy and manage device software.

- Pre-build MXNet and Tensorflow packages
- Lambda actions
- Deploy trained models in the cloud
- GPU access
- Allows to bring our own framework e.g., Caffe2 or CNTK

https://aws.amazon.com/greengrass/ml/

TensorFlow Lite

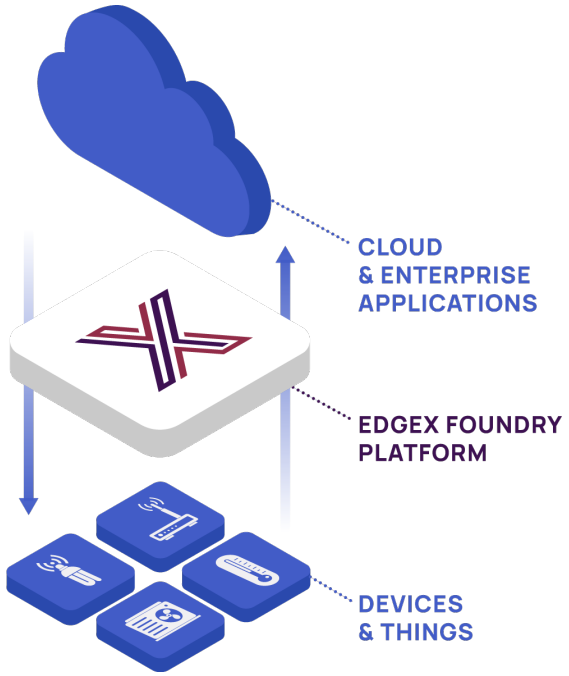
Developed by Google internal use, and now open-source. It is a collection of tools to convert and optimize TensorFlow models to run on mobile and edge devices.

- Open-source deep learning framework designed for on-device inference
- Allows to run trained models on mobile, embedded and IoT devices
  - Supports different platforms e.g., Linux, Android and iOS
- Offers an user-friendly way to build applications using TF ML models in iOS and Android devices
- Offline inference does not rely (at all) in internet connection. This allows to deploy solutions in remote cases where internet is expensive or scarce

https://www.tensorflow.org/lite
https://viso.ai/edge-ai/tensorflow-lite/

EdgeX Foundry

Open-source, vendor neutral hosted by Linux Foundation, that provides a common framework for IoT edge computing



- Edge Xpert is a recent a the addition that allows computer vision and edge AI support via an add-on built around Intel's OpenVINO AI toolkit.

https://www.edgexfoundry.org/
https://www.edgexfoundry.org/why-edgex/
https://linuxgizmos.com/edgex-foundry-based-edge-software-adds-ai-support/

Challenges

Challenges

Edge computing & AI

- Security at the edge requires broad preparation
  - Limited resources limit the capability to ensure security
- Stakeholders alignment to migrate workloads to the edge
  - Involves changes across multiple teams
- Moving from an architecture with e.g., a few servers & locations to an environment of thousands of individual smaller locations
  - Impacts architecture design, strategy and points-of-failure e.g., see the first point related to security
  - Edge envs are heterogenous by nature, therefore requires a thinking for an hybrid solution
- For many ML models running models on "limited hardware" impacts the quality of the models

https://enterprisersproject.com/article/2022/2/edge-computing-strategy-5-potential-gaps-watch

A closer lens on TinyML

TinyML

Tiny Machine Learning

Intersection of Embedded Systems and AI that involves developing systems that run ML models on ultra-low-power microcontrollers. The idea is to push the implementation to where the information source is.

How to implement TinyML?

- Machine learning frameworks that support TinyML applications
  - TensorFlow Lite
  - PyTorch Mobile
  - Edge Impulse

TinyML also focus on ML solutions for good[1]

[1] https://www.tinyml.org/event/tinyml-for-good/

TinyML aims to improve the products on four fronts

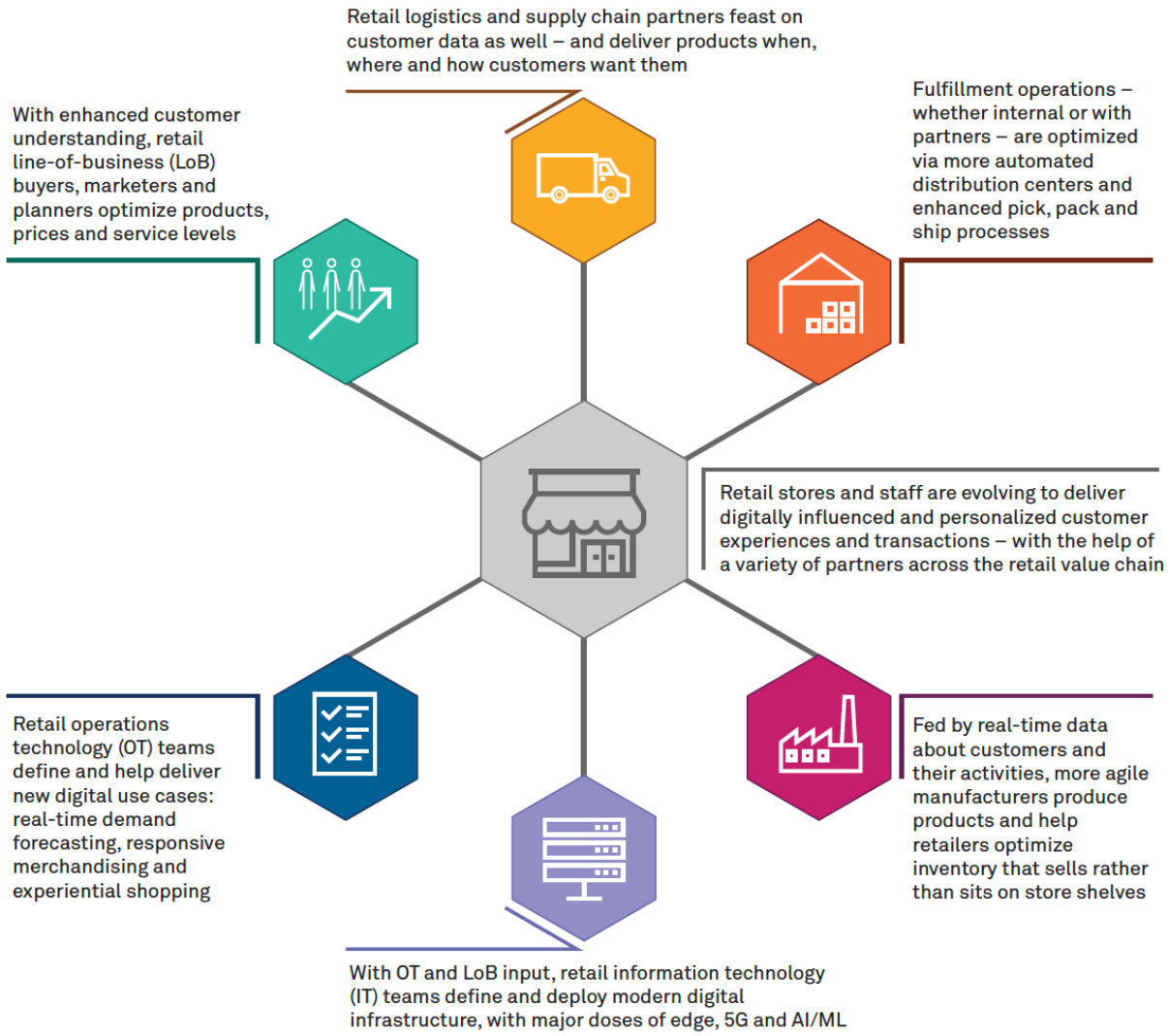
- **Privacy –** by running ML programs on the edge, it is possible to help in the risk of such a privacy incident[1]
  - E.g., data analysis will be kept on the actual device which can positively impact a user's data privacy,
- **Power –** processing data on the device cuts off energy consumed in the data communication layer
- **Cost –** by not transmitting data, the cost of setting up servers and the on-device radio is cut
- **Reliability –** Not sending data anywhere means that the response time is reduced. Server maintenance or outage is no longer a problem.

1 https://internetinnovation.org/general/creation-of-innovation-tinyml-can-enhance-privacy-efficiency-of-iot-devices/
2 https://research.aimultiple.com/tinyml/

Edge & Retail Digital Transformation

Edge + retail digital transformation

- 38% retail enterprise infrastructure is already edge-based
- 77% retailers say they will increase their edge infra in the upcoming 2 years, and 27% claim it will be a significant increase
- 63% of retailers plan to augment their enterprise edge compute infra with private 5G network



Pathfinder, Transforming Retail with Edge Compute, 5G, IoT and AI – Dell, 2021

Join us!

levistrauss.com/work-with-us/data-science/

- Product Recommendations
- Search & Browse
- Loyalty
- Consumer Experience

Computer Vision

NLP & Information Retrieval

Recommendation Systems

and more

The future of fashion, apparel and retail is digital — powered by data, computer vision and machine learning.

*We are harnessing petabytes of images and a treasure trove of data from the last 167 years to **solve exciting problems in fashion, design, manufacturing, supply chain, business and society** — all of it while leading with our values. Join us to make your mark on our digital transformation.*

Katia Walsh, Ph.D.

Senior Vice President & Chief Strategy and AI Officer

Thank you!

Filipa Peleja

Lead Data Scientist, Levi Strauss & Co.