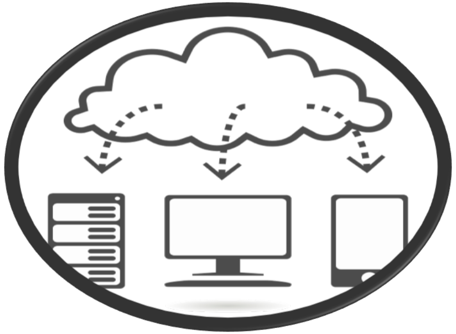
Distributed AI for Intelligence at Edge

Presenter: Dinesh Verma
IBM Fellow, Distributed AI
IBM T. J. Watson Research Center
Email: dverma@us.ibm.com

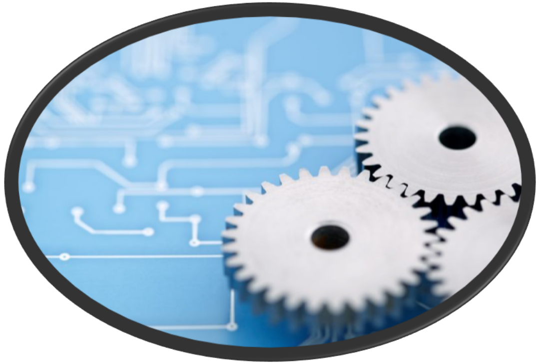
Many promising technologies are converging together to transform the world within the next decade

Cloud Computing

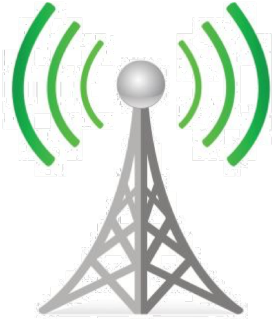
Edge



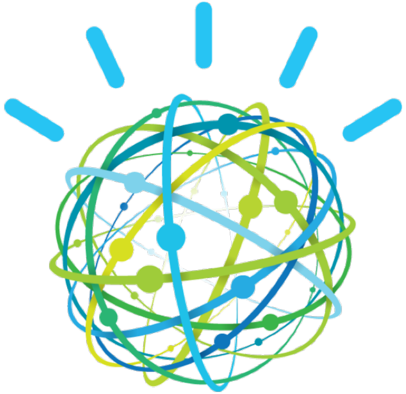
Internet of Things



5G



Artificial Intelligence



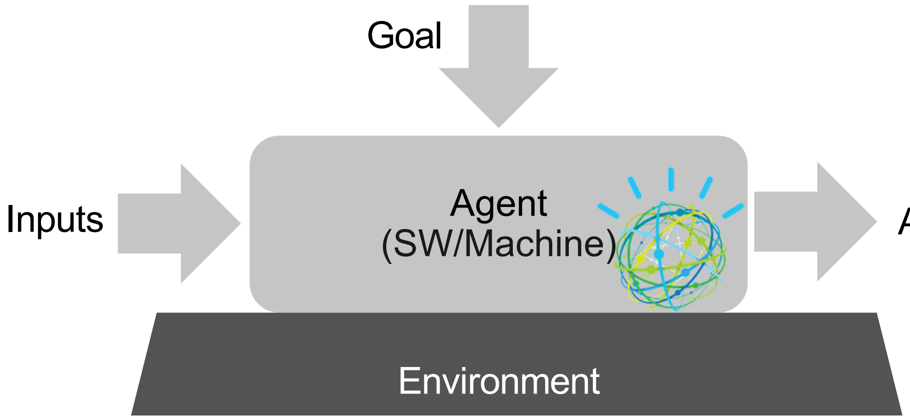
What is AI (really)

Definitions of Intelligence

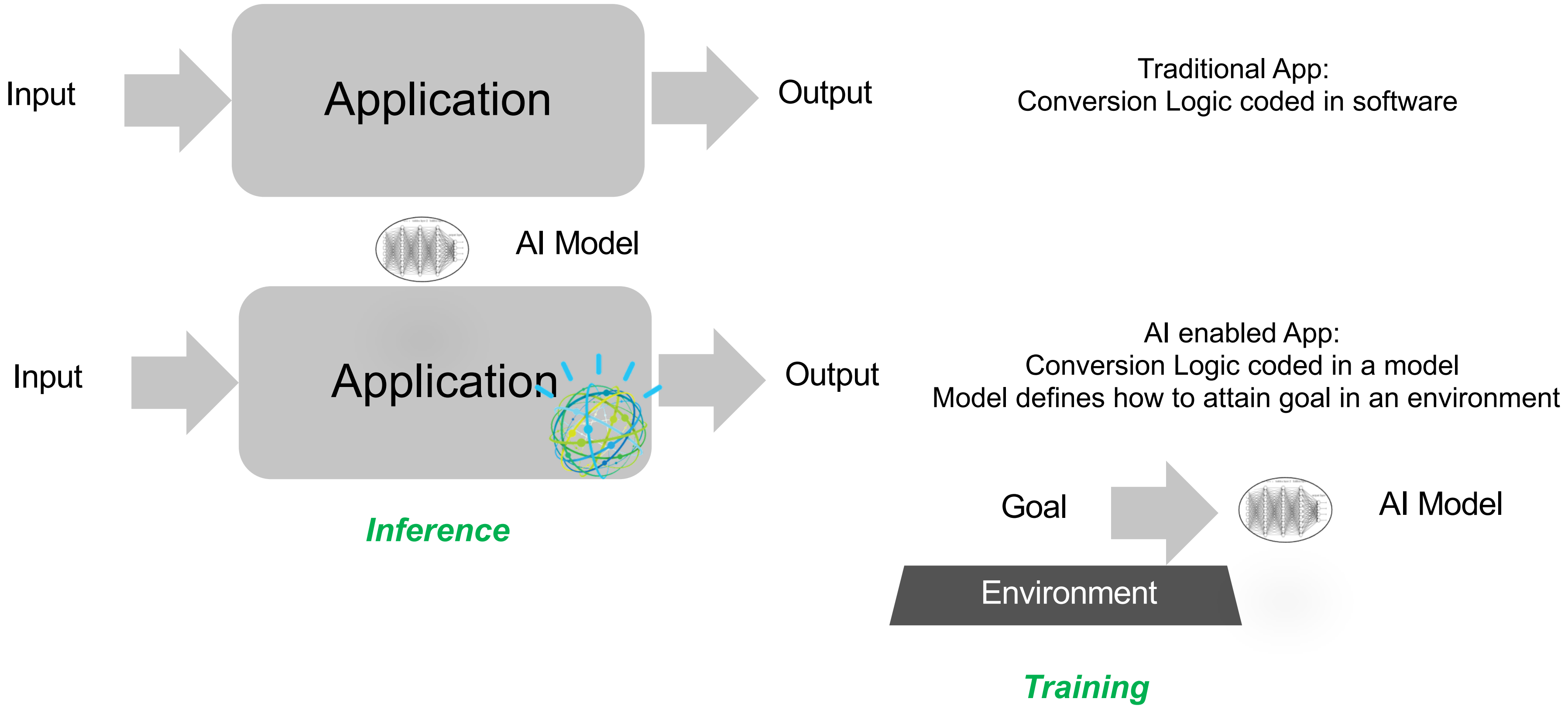
- Too many – For a collection of definitions, see Legg, Shane, and Marcus Hutter. "A collection of definitions of intelligence." Frontiers in Artificial Intelligence and applications157 (2007): 17

Informally

- “Intelligence is a measure of an agent’s ability to achieve goals in a wide range of environments.
- Artificial Intelligence is the ability of a machine to display intelligence, i.e. an ability to achieve its goal in a variety of environments



Any Computer Application converts an input to an output



AI versus not AI

From the conversion of input to output, an AI enabled model and a non-AI model are equivalent

– Every Software Program can be converted to a equivalent Turing Machine Program

– Every AI model can be converted into a equivalent “Lookup table”

AI does not enable us to do anything we could not already do using a software encoded logic

What AI provides is non-functional attributes

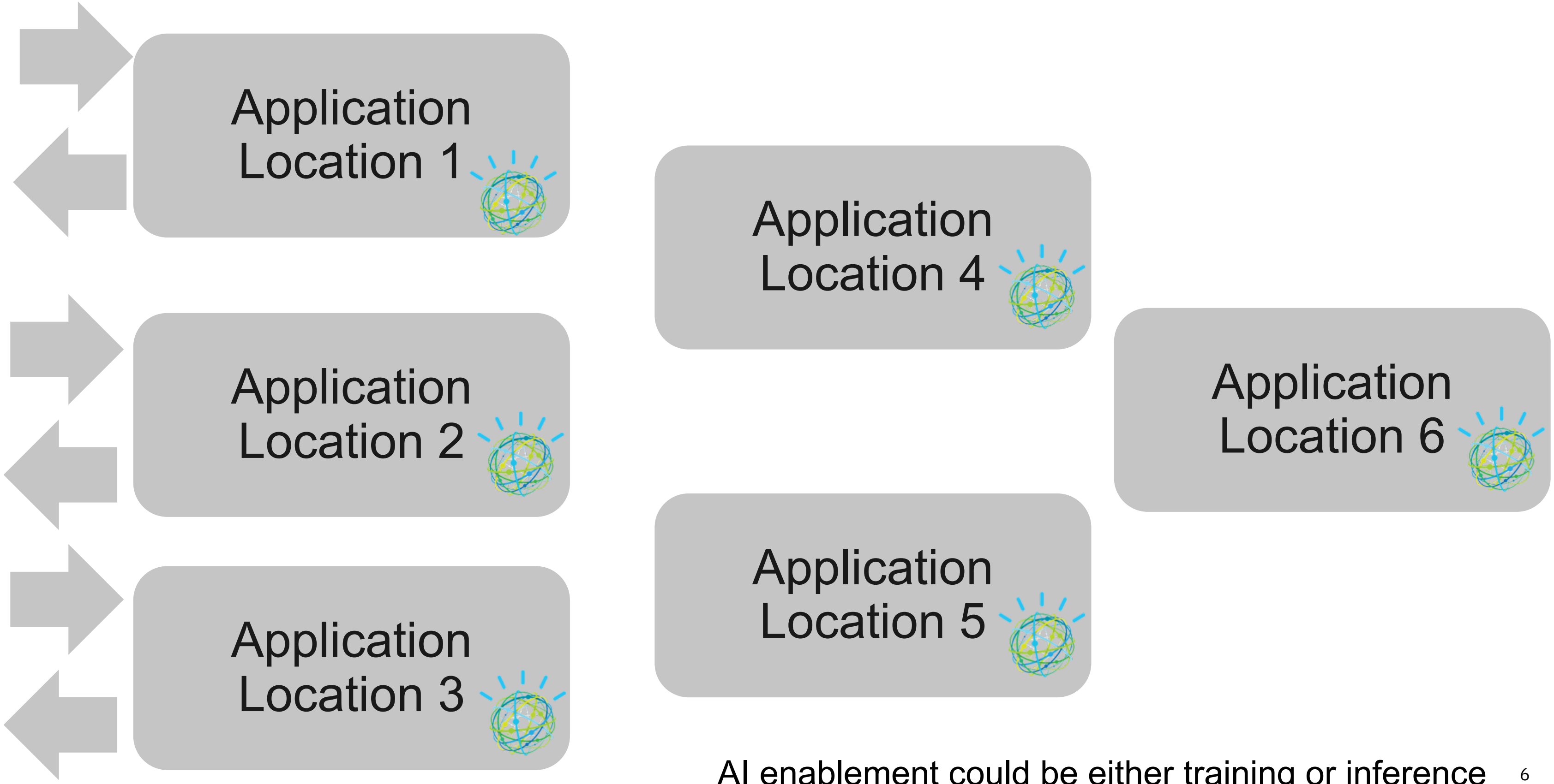
Cheaper

- Use different skills, e.g. data scientists instead of programmers
- Uses less of an expensive resource (network/storage/processor)

Better

- More flexible and adaptable
- Faster time to decision making
- More maintainable

Distributed AI – application where AI enablement happens at many different locations

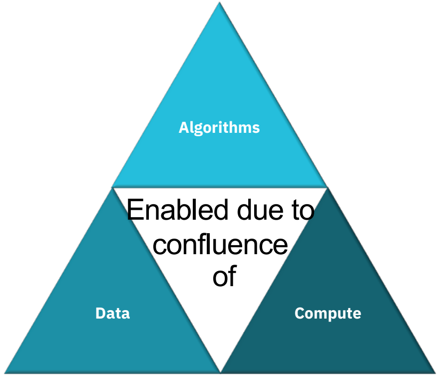


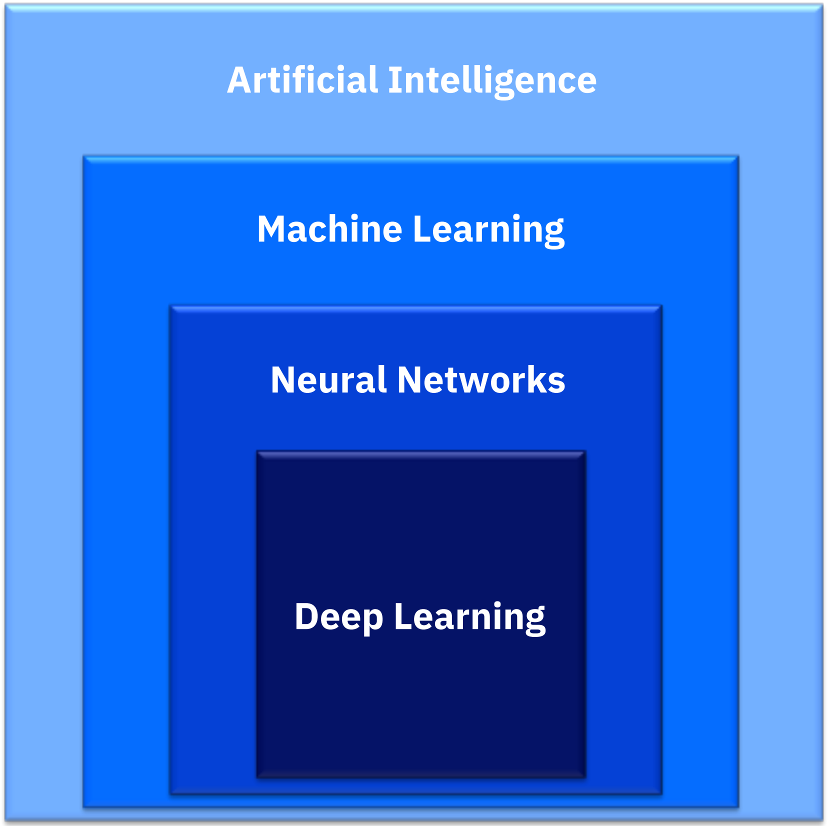
AI enablement could be either training or inference

Creating an AI model

AI model can be created in a variety of ways

- Symbolic AI: A human being provides the model
- Machine Learning: A machine extracts the model from training data
- Neural Networks: A specific type of AI model
- Deep Learning: AI model using neural networks with many layers





AI model encodes a spectrum of capabilities

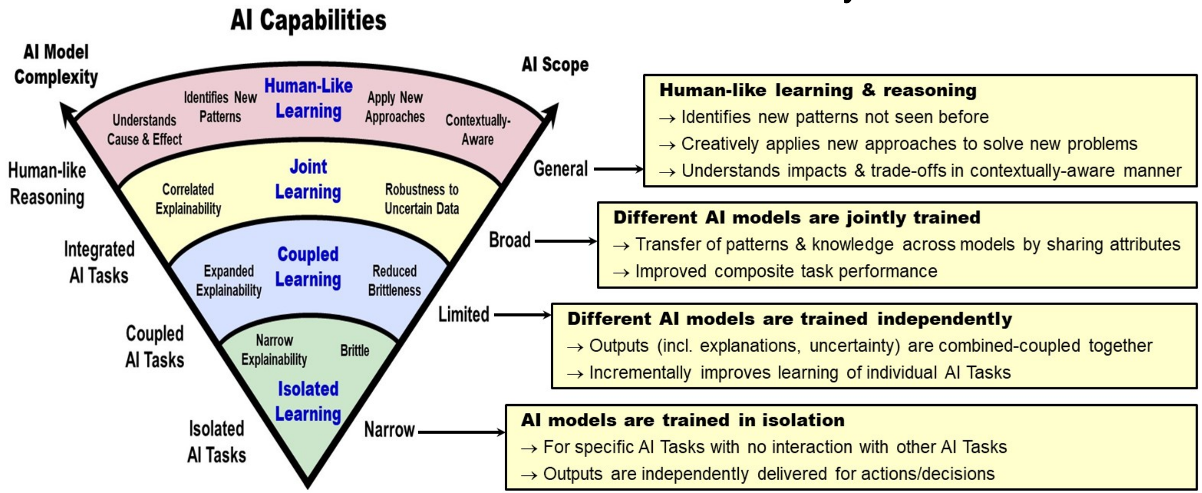
The capabilities inherent in AI are related to:

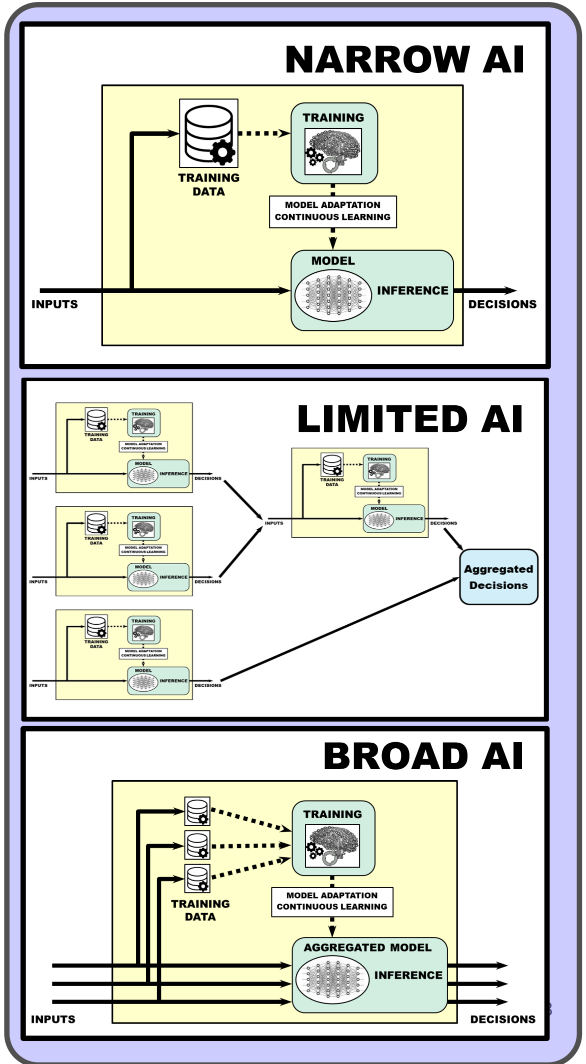
→ Scope of the learning

→ Complexity of the AI learning model

→ Potential level of explainablity

→ Potential for robustness under uncertainty



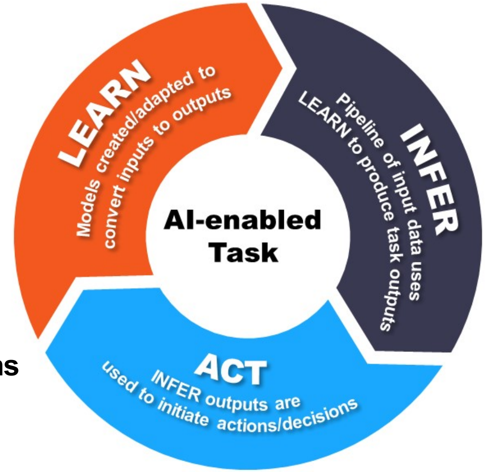


Source: Cirincione, G., & Verma, D. Federated machine learning for multi-domain operations at the tactical edge.
In SPIE DSS - *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (May 2019).

The process for creating an AI-enabled application

Systematic approach to create AI:

1. **Design AI/ML to meet goals**
2. **Obtain/curate relevant training/ validation data**
3. **Train models under a variety of conditions → dinky, dirty, dynamic, deceptive**
4. **Validate AI/ML in realistic conditions**
5. **Analyze to understand & predict its behavior/performance for safety**
6. **Determine allowable use cases including learning during deployment & autonomy**

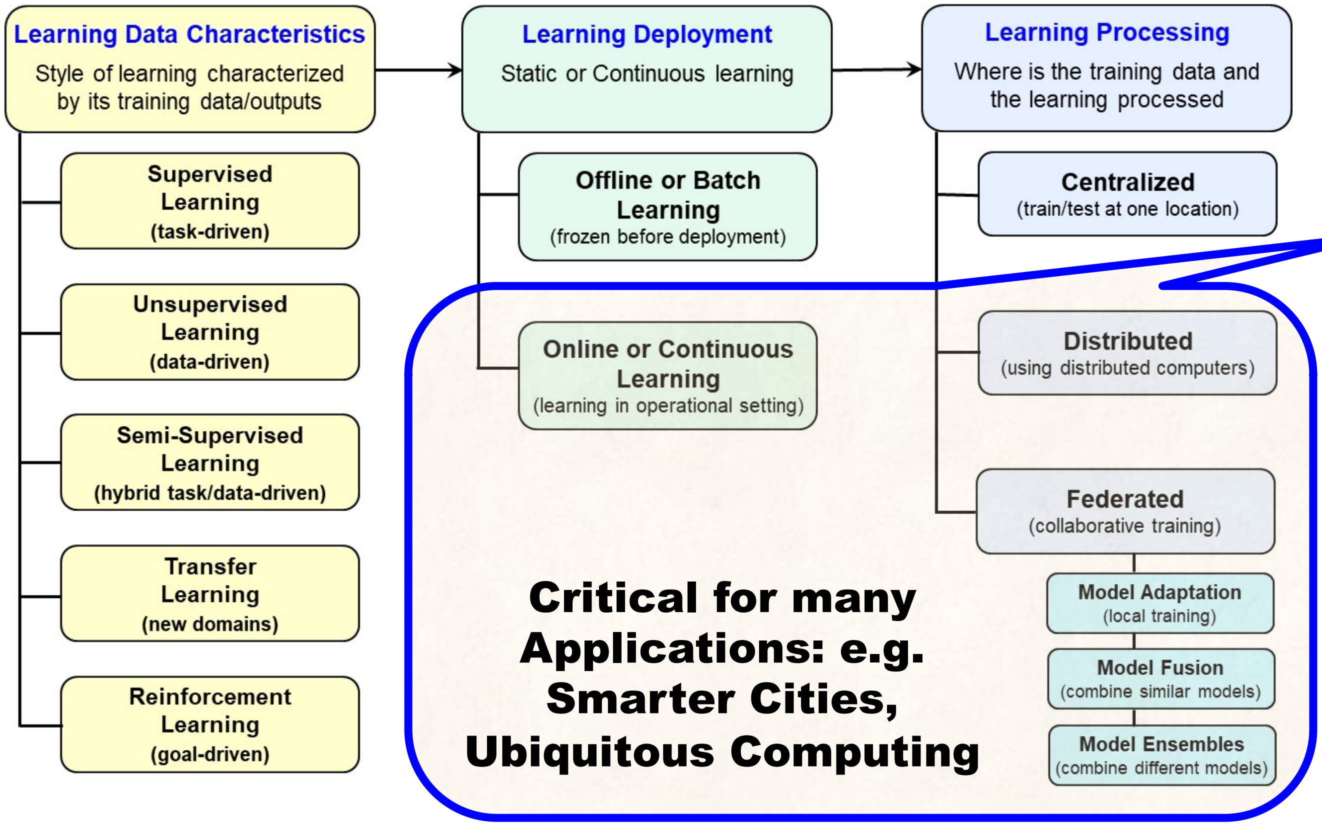


Design Inputs for Learn → Infer → Act Processes

1. **Where does learning, inferring, acting occur? (not necessarily co-located)**
2. **What are the performance & resiliency requirements of the action?**
3. **What level of autonomy is required?**
4. **What constraints will exist in the operational setting?**
5. **What is the availability & location of the training & input data?**
6. **How large & complex is the model to be used?**
7. **Where is computational capacity to support learning & inferencing?**

Source: Cirincione, G., & Verma, D. Federated machine learning for multi-domain operations at the tactical edge. In SPIE DSS - *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (May 2019).

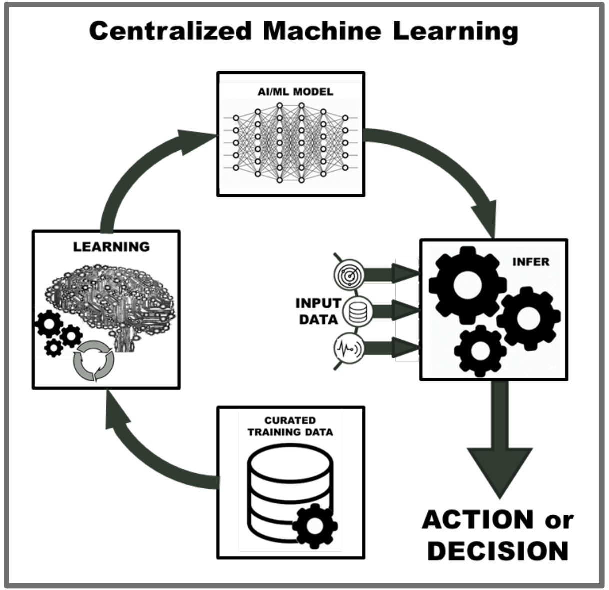
Factors involved in enabling an AI-based application



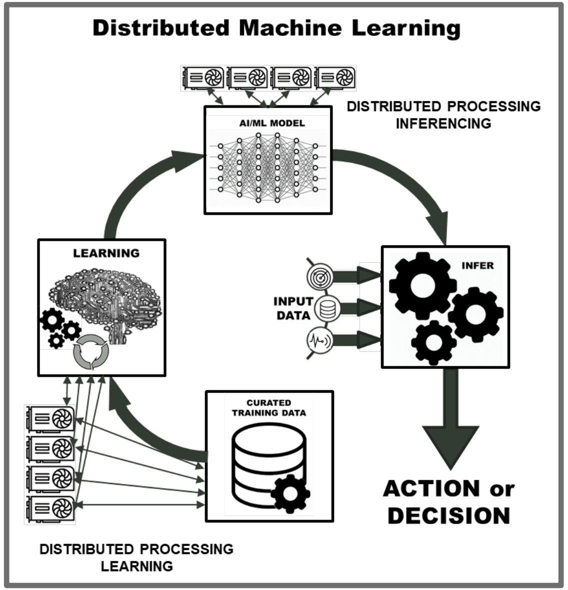
- **Operational adaptability**
- **Limited computational resources at Edge**
- **Limited networks**
- **Sharing restrictions**
  - → **Regulations**
  - → **Policies**
  - → **Security**

Source: Cirincione, G., & Verma, D. Federated machine learning for multi-domain operations at the tactical edge. In SPIE DSS - *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (May 2019).

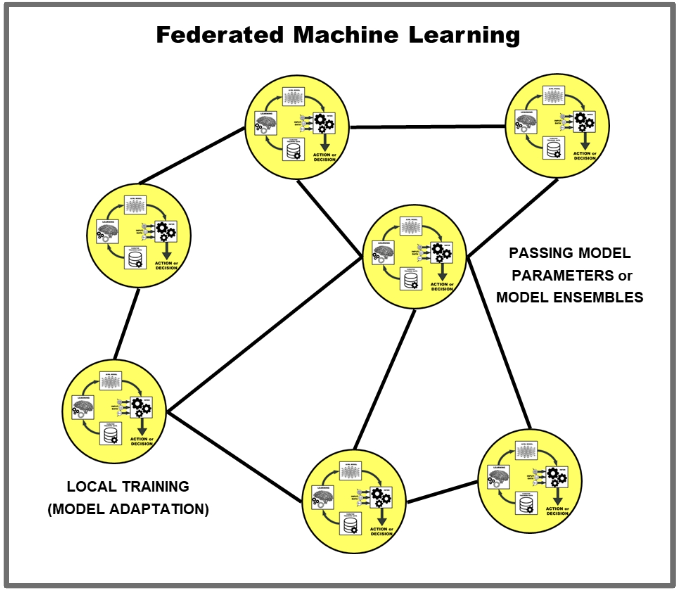
Different options for Learning Processing



When the Network is Always Reliably Available



When the Network is Reliable but a single processor is not adequate

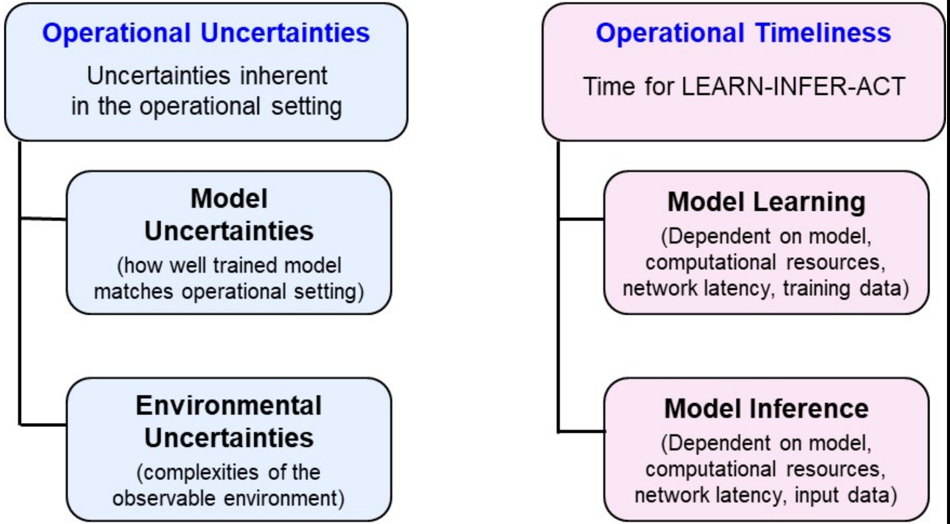


When the Network is unavailable or crosses a regulatory domain

Source: Cirincione, G., & Verma, D. Federated machine learning for multi-domain operations at the tactical edge.
In SPIE DSS - *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (May 2019).

Any resilient application has to deal with the inherent uncertainties in the environment

Operational factors impacting AI/ML effectiveness



- → **Partial Observability**: Distributed operations without full state of the environment
- → **Dynamics**: Unpredictability due to conditions that are changing:, fast mobile vehicles, rapidly changing situations, complex environments such as urban settings, manufacturing floors; unexpected and adversarial behavior of people
- → **Noise**: Introduced in wireless communications networks, sensors inputs, & information sources
- → **Conflicts**: Mismatches in inputs from distributed sensor & information sources
- → **Deception**: bad actors manipulating the inputs into the system to disrupt & deceive, kids trying to play a prank on the system, participants trying to game the system, scammers trying to see if they can get an unfair advantage

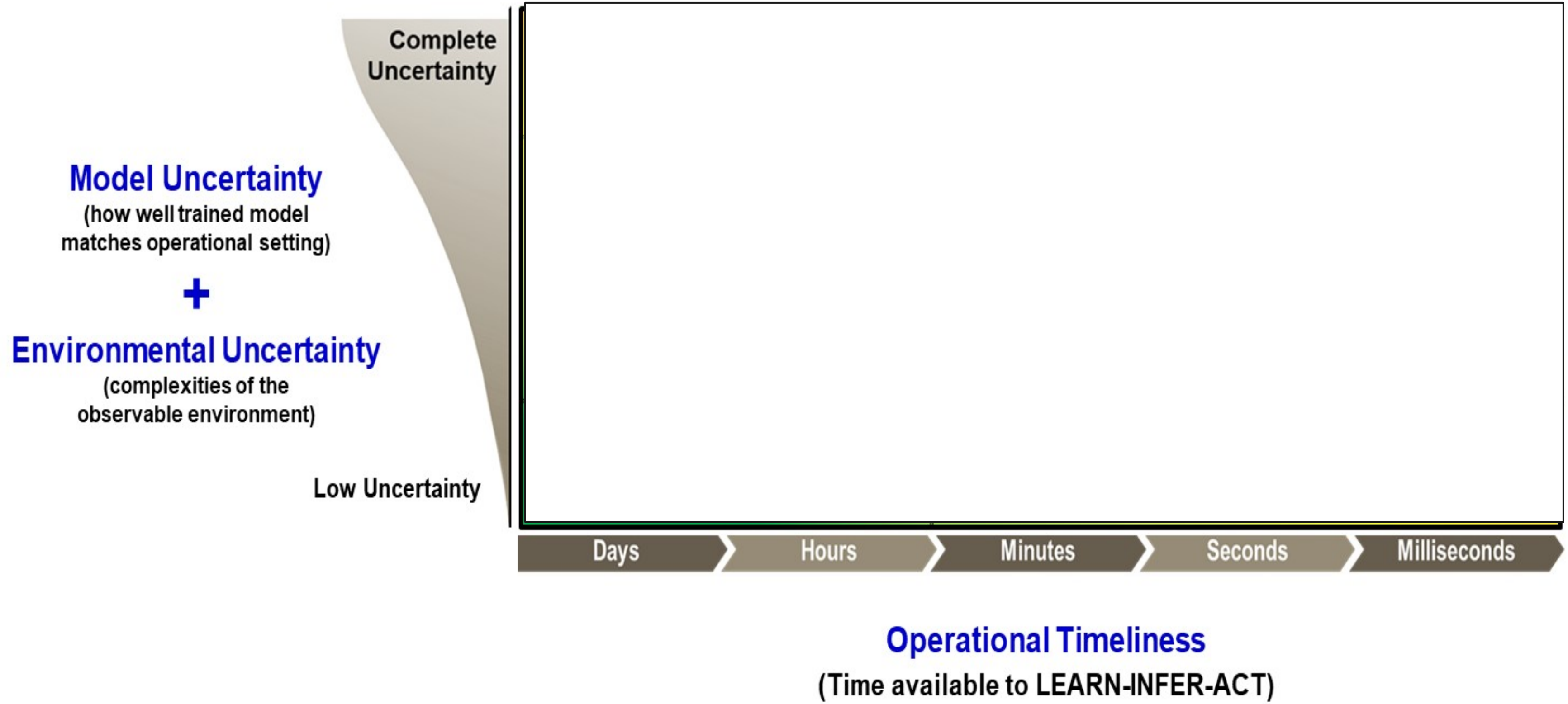
Source: Cirincione, G., & Verma, D. Federated machine learning for multi-domain operations at the tactical edge. In SPIE DSS - *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (May 2019).

Applications need to meet operational timeliness requirements

Attributes of Operational Timeliness requirements in some smart computing environments

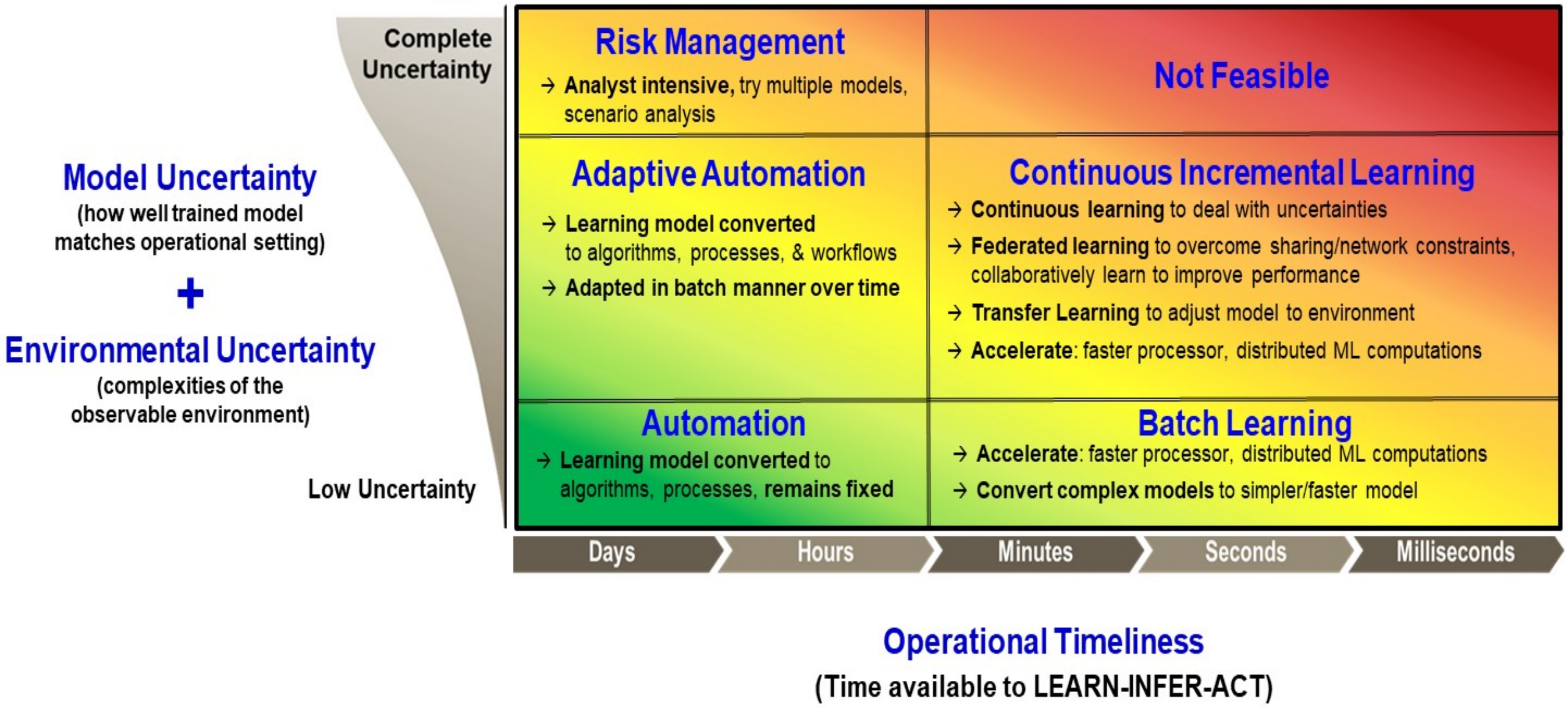
	Attributes	Operational Timeliness	Examples
Non Real-Time	• Complex Analytics from heterogenous and structured/unstructured sources • Strategic decision making	Hours to Days	• Supply Chain • Asset Maintenance • Business Process Optimization • Customer Journey Analysis • Customer Retention Analysis
Near Real-Time	• Speed is important, but some delays are acceptable • Quick response with soft or hard deadlines	Seconds to Minutes	• Voice Assistants • Safety of Crowds and Cities • Premises Access with Biometrics • Industrial Worker Safety • Bank Fraud Detection
Real-Time	• Constant input with a steady data output requirement • Frequently hard deadlines	Seconds to milliseconds	• Manufacturing Quality Control • Autonomous Vehicle Control • Production Robotic Control • Industrial Safety and Controls

Any Solution addresses a point in uncertainty-timeliness space



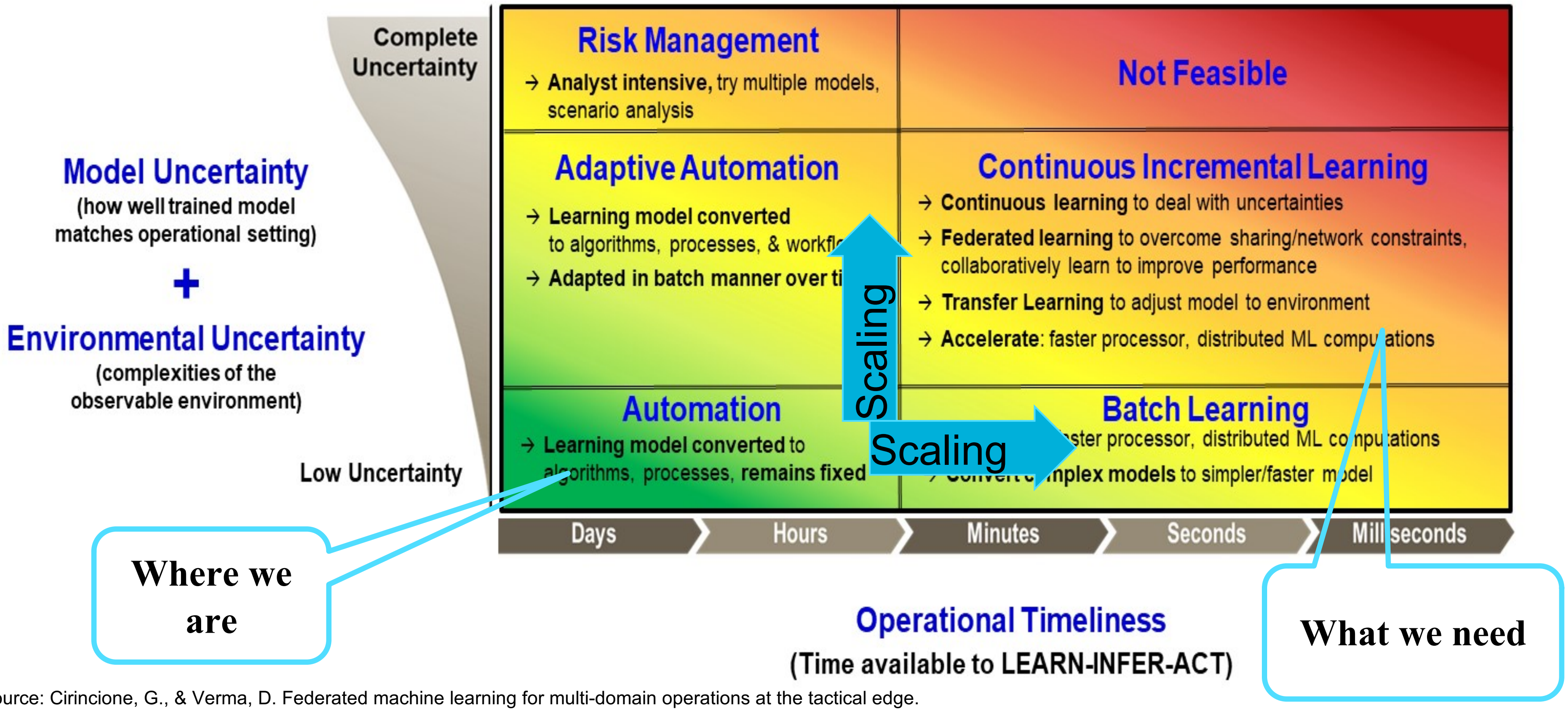
Source: Cirincione, G., & Verma, D. Federated machine learning for multi-domain operations at the tactical edge.
In SPIE DSS - *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (May 2019).

Different types of approaches required at different regions



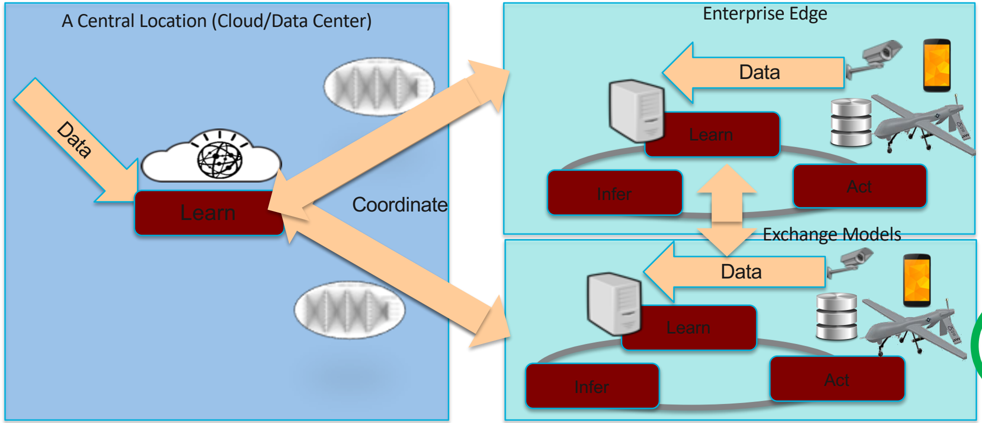
Source: Cirincione, G., & Verma, D. Federated machine learning for multi-domain operations at the tactical edge.
In SPIE DSS - *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (May 2019).

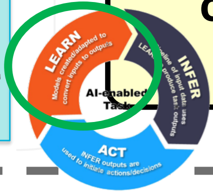
Distributed AI is required to address the operational timeliness requirements



Source: Cirincione, G., & Verma, D. Federated machine learning for multi-domain operations at the tactical edge.
In SPIE DSS - *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (May 2019).

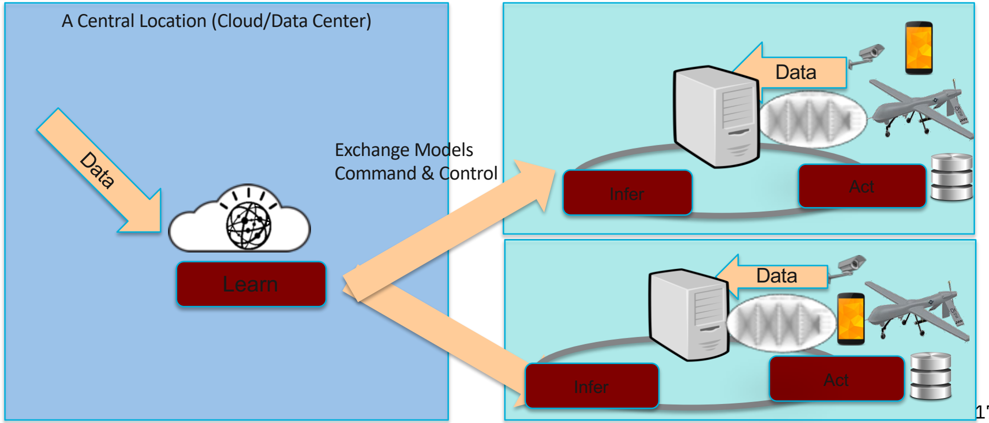
Two Approaches for Distributed AI during model training phase



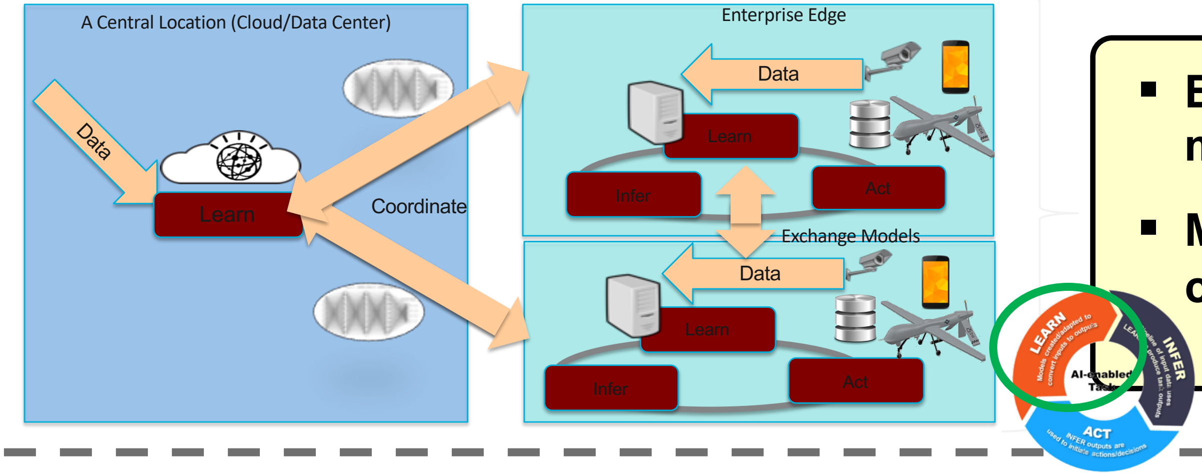


- **Each edge site learns the AI model from their local data**
- **Models are exchanged and combined into a composite.**

- **Cloud learns the model and sends to each edge site**
- **Each edge site adapts the model to match its environment.**



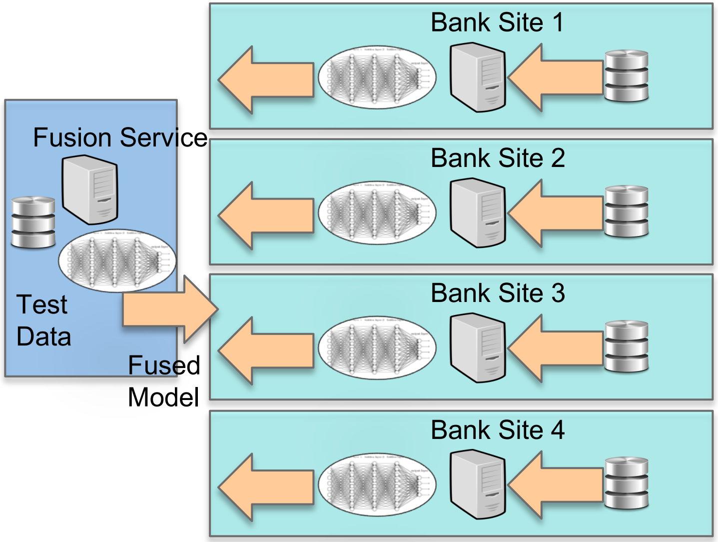
Two Approaches for Distributed AI during model training phase



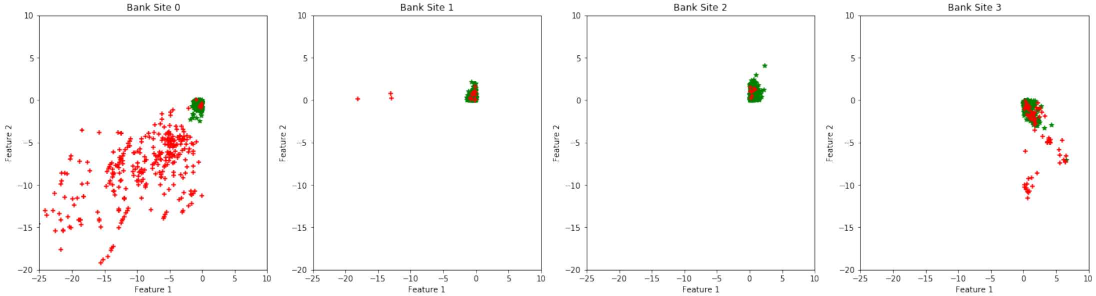
- **Each edge site learns the AI model from their local data**
- **Models are exchanged and combined into a composite.**

Improving credit card fraud models using Federated Learning

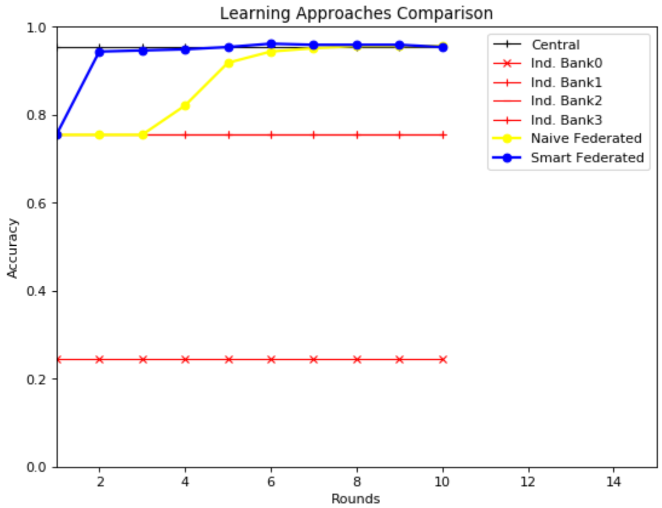
Scenario



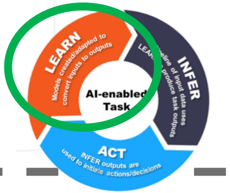
Data



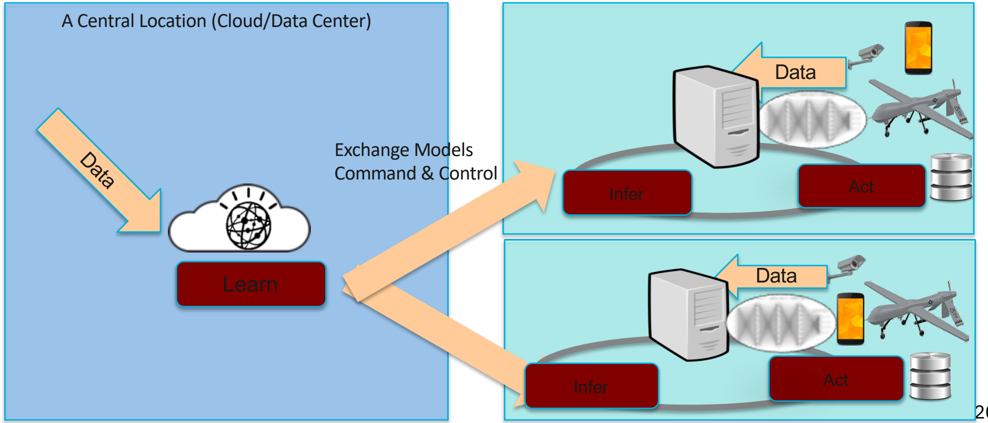
Results



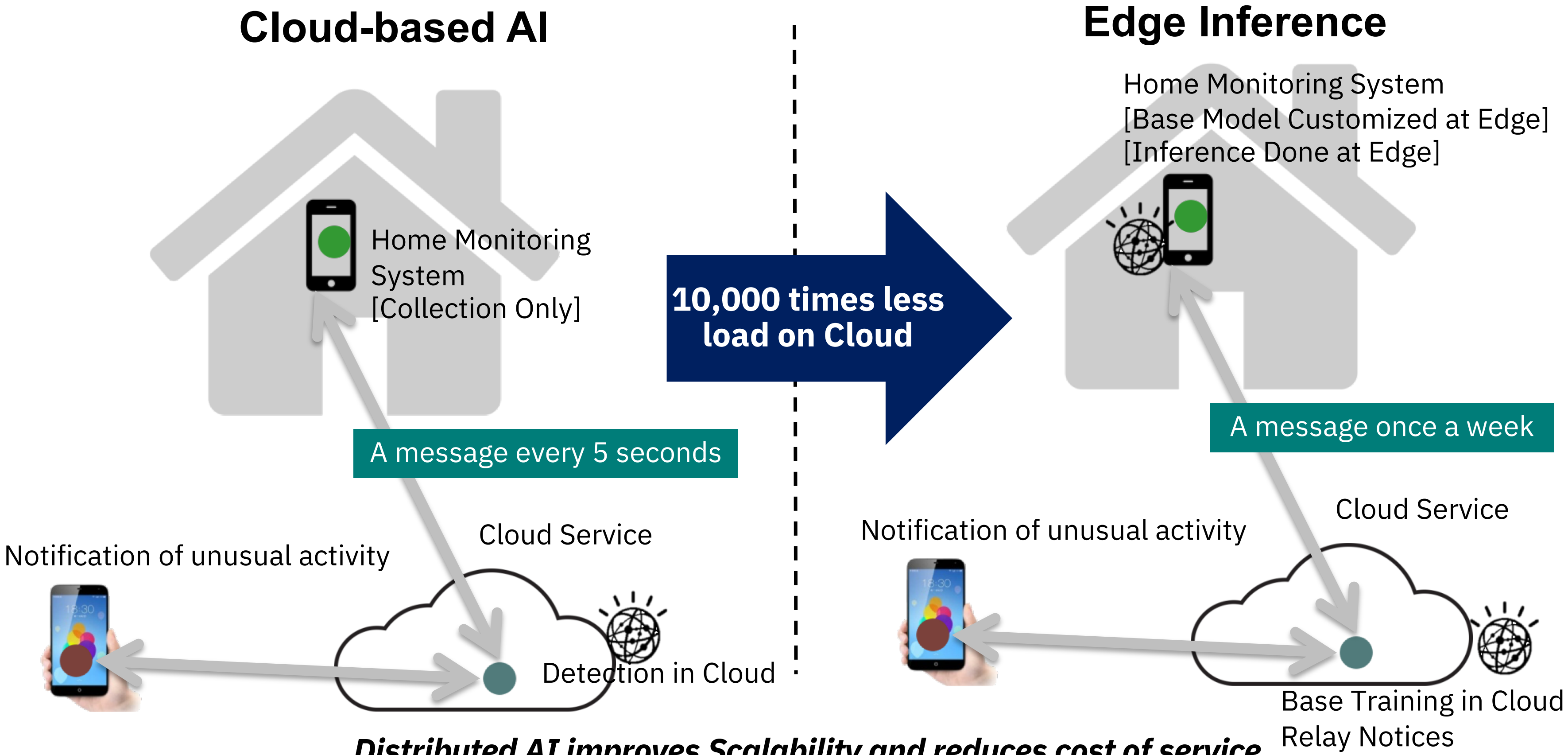
Two Approaches for Distributed AI during model training phase



- **Cloud learns the model and sends to each edge site**
- **Each edge site adapts the model to match its environment.**

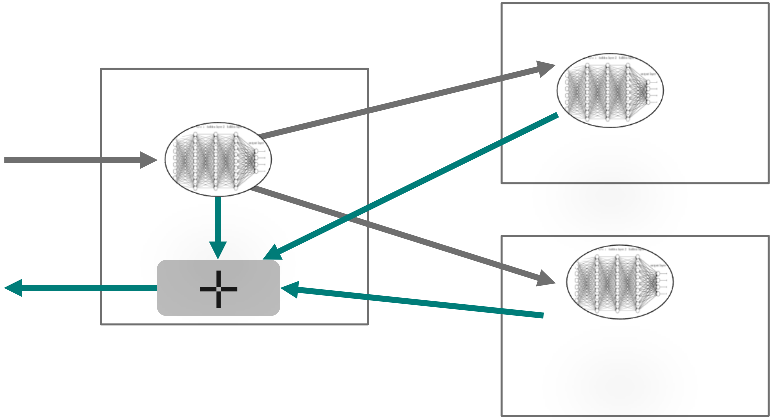


A Specific Example: Comparing two Smart Computing Implementations

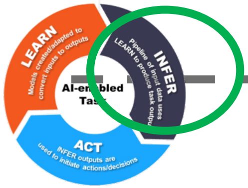


Distributed AI improves Scalability and reduces cost of service

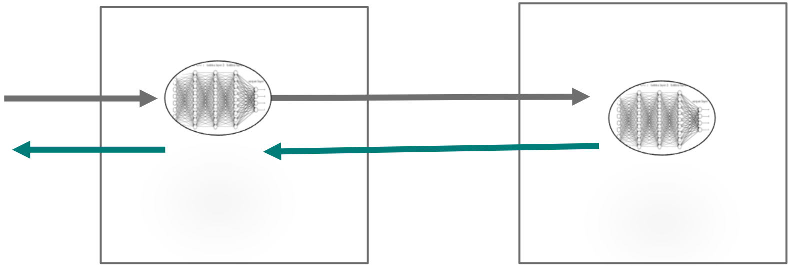
Two Approaches for Distributed AI during the infer/act phase



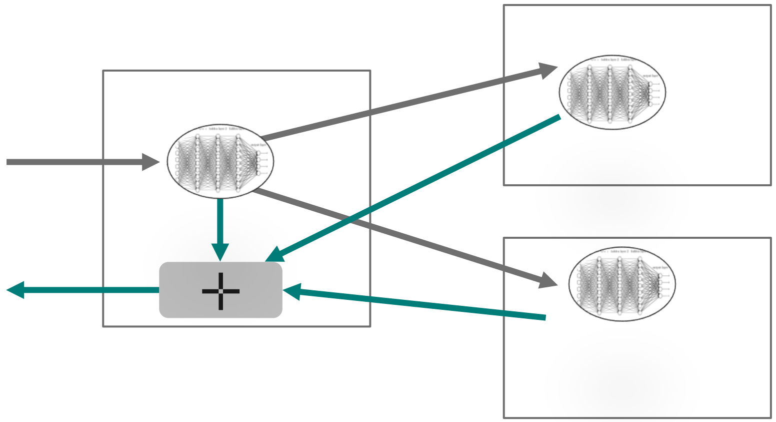
- **Each site sends the inference request to one or more peer sites**
- **Results are combined into a aggregate answer by requesting site.**



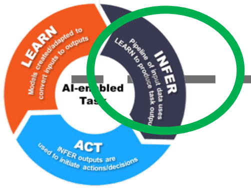
- **Edge Site gets the model from Cloud Site and adapts it locally**
- **It checks its ability to perform the inference and if not, consults the cloud site**



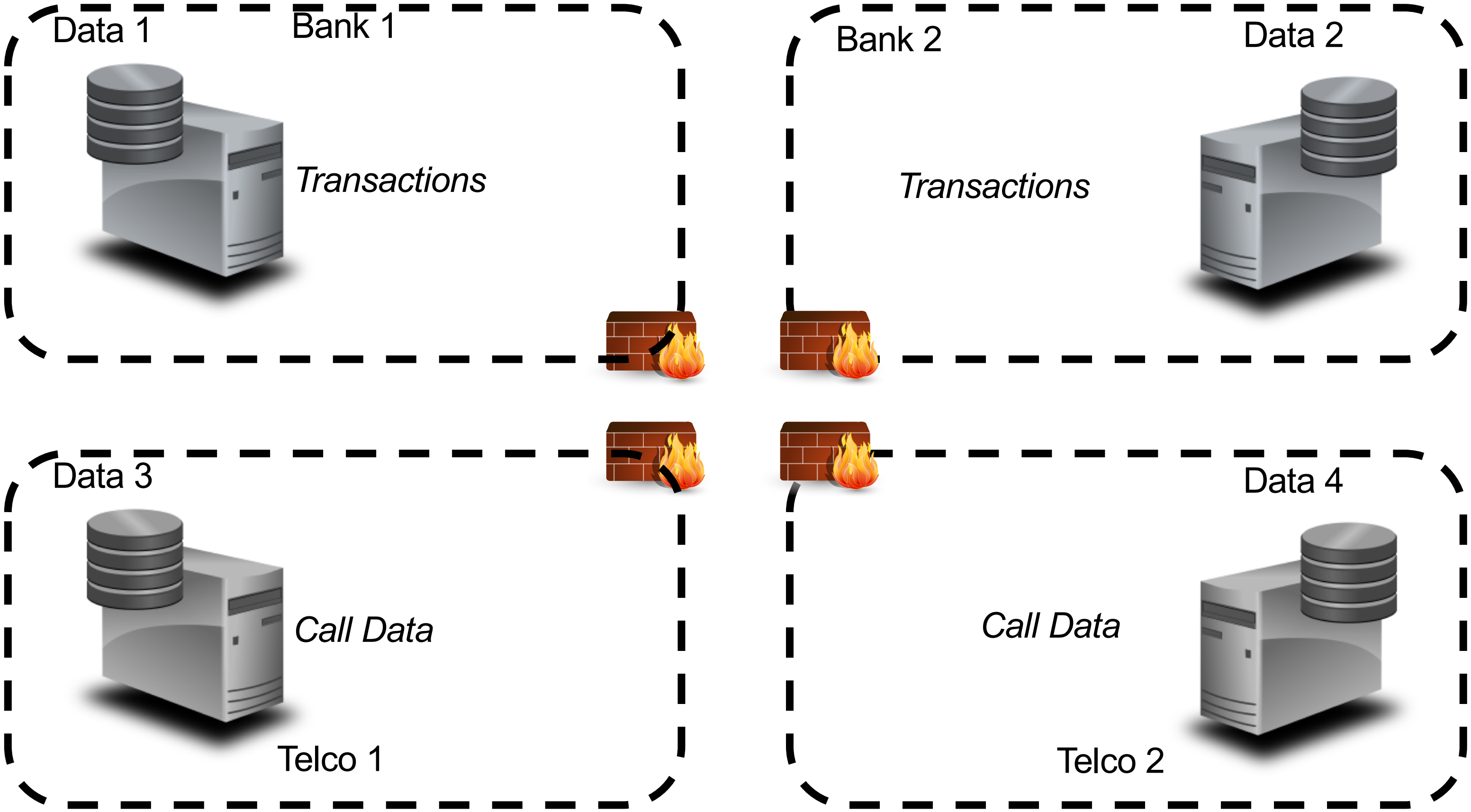
Two Approaches for Distributed AI during the infer/act phase



- **Each site sends the inference request to one or more peer sites**
- **Results are combined into a aggregate answer by requesting site.**



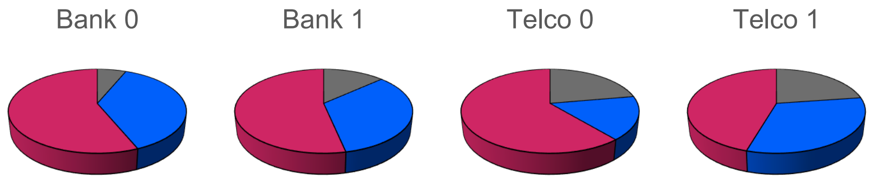
Example: The World Environment



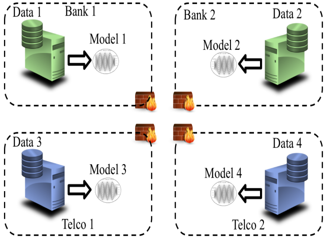
2 Banks and 2 Telcos want to detect scammers

Typical Results

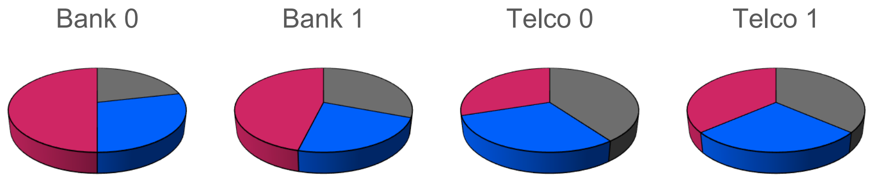
True Positives False Positives False Negatives



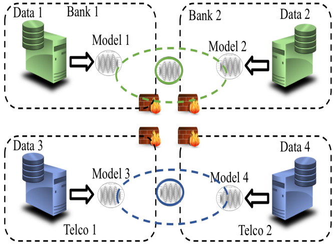
Independent Detection



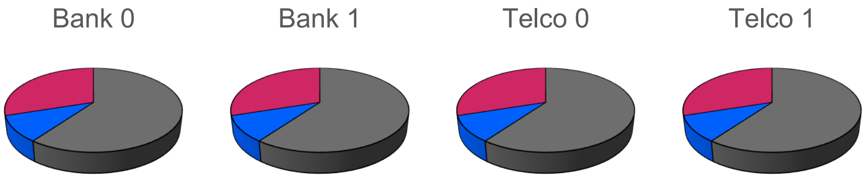
Each organization
Works independently
to detect scammers



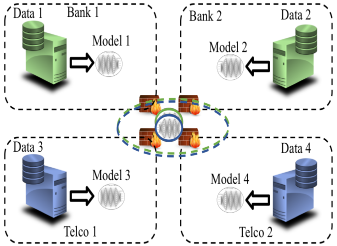
Federated Learning



Banks and Telcos
use Federated Learning
to detect scammers

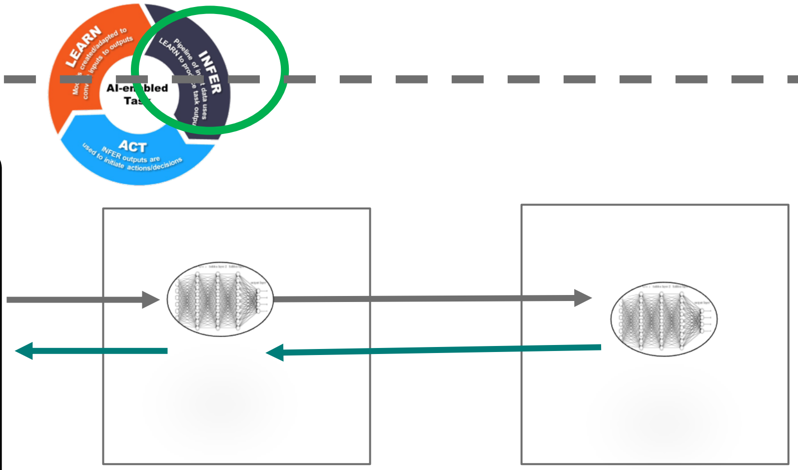


Federated Learning + Federated Inference



Banks and Telcos
use Federated Learning
and Federated Inference
to detect scammers

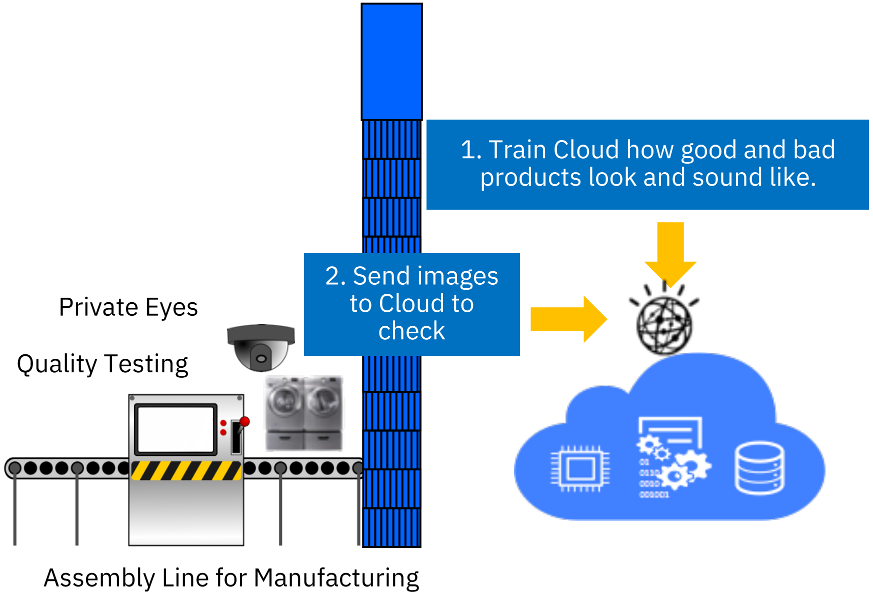
Two Approaches for Distributed AI during the infer/act phase



- **Edge Site gets the model from Cloud Site and adapts it locally**
- **It checks its ability to perform the inference and if not, consults the cloud site**

A Specific Example: Meeting Requirements of Manufacturing

Cloud-based AI

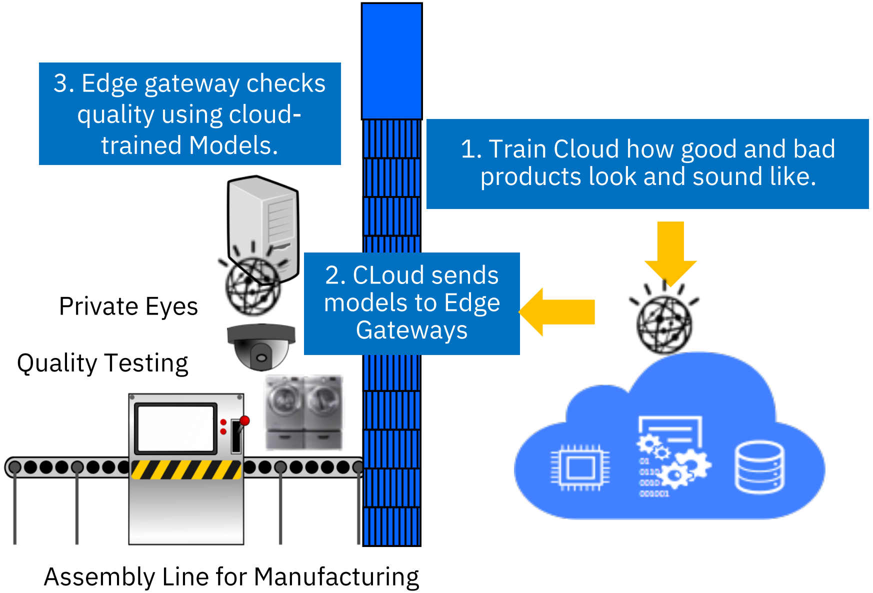


Time per check = 1000 ms:

- Network (10*100ms) + Image Time (2ms) + Inference Time (1ms)

Operational data goes to the cloud

Edge Inference



Time per check = 50 ms:

- Network (10*5ms) + Image Time (2ms) + Inference Time (2ms)

Plus No Operational data leaves premises

Distributed AI improves operational timeliness and privacy risks