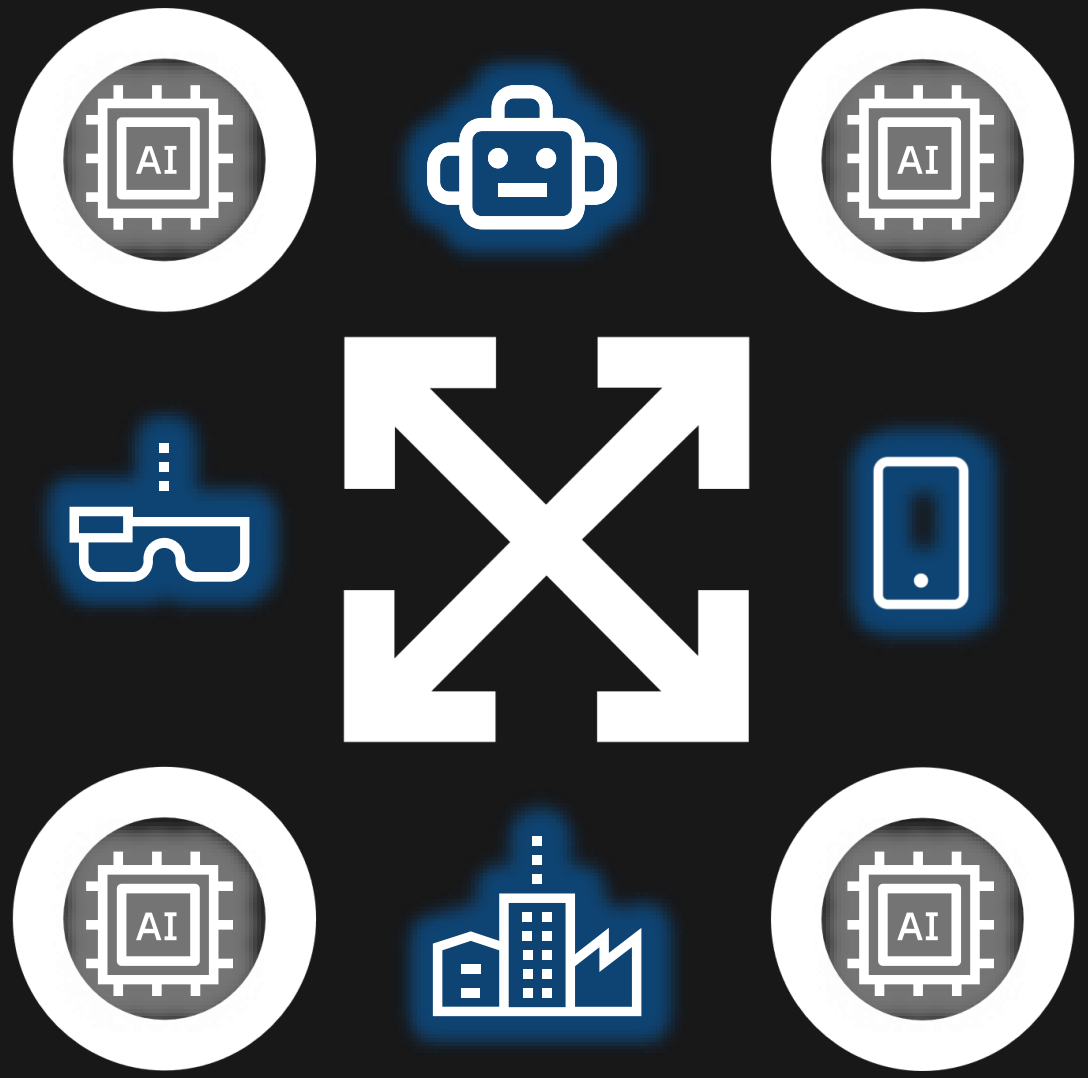
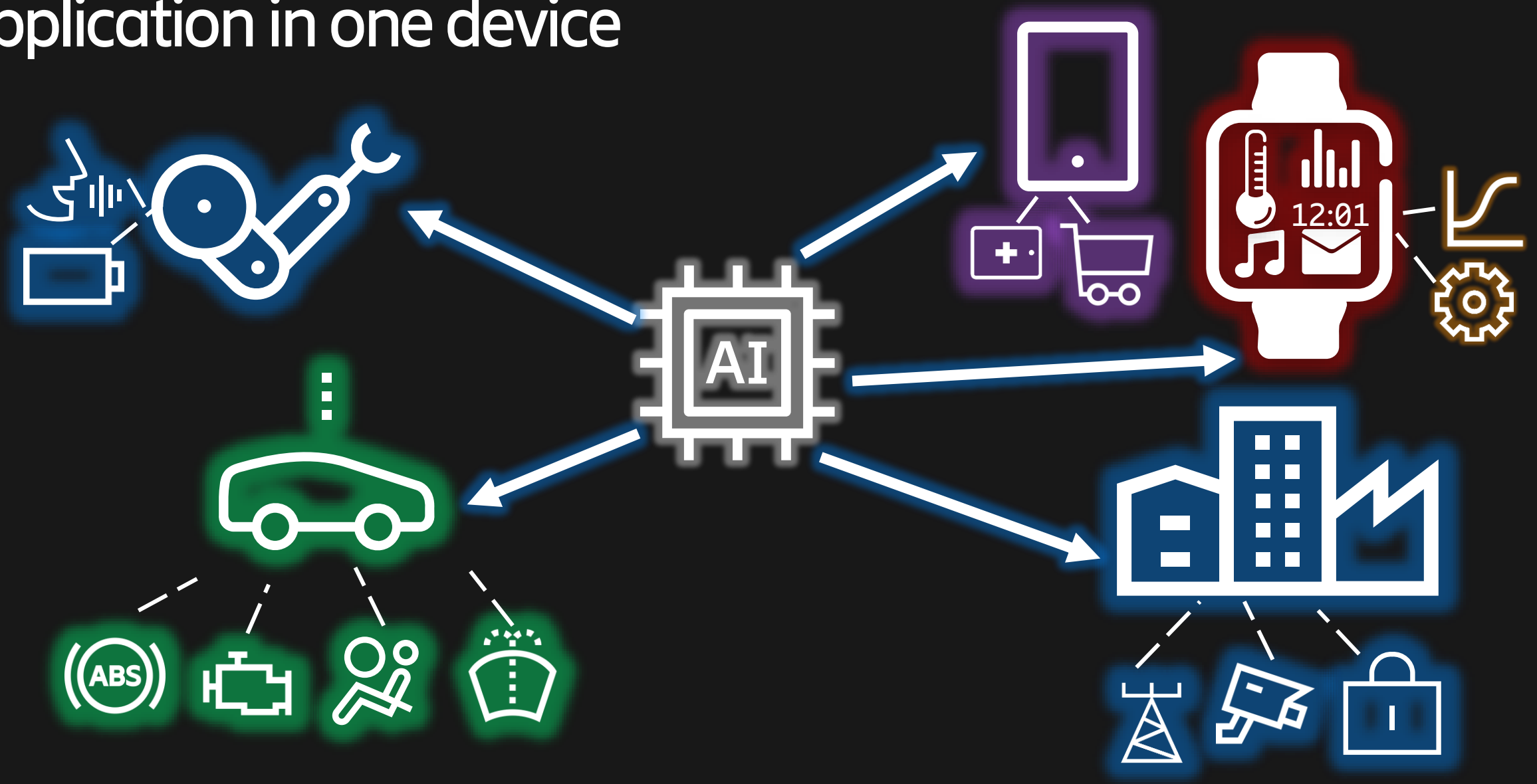
Intelligence Services Provisioning and Distribution

To the edge and beyond...

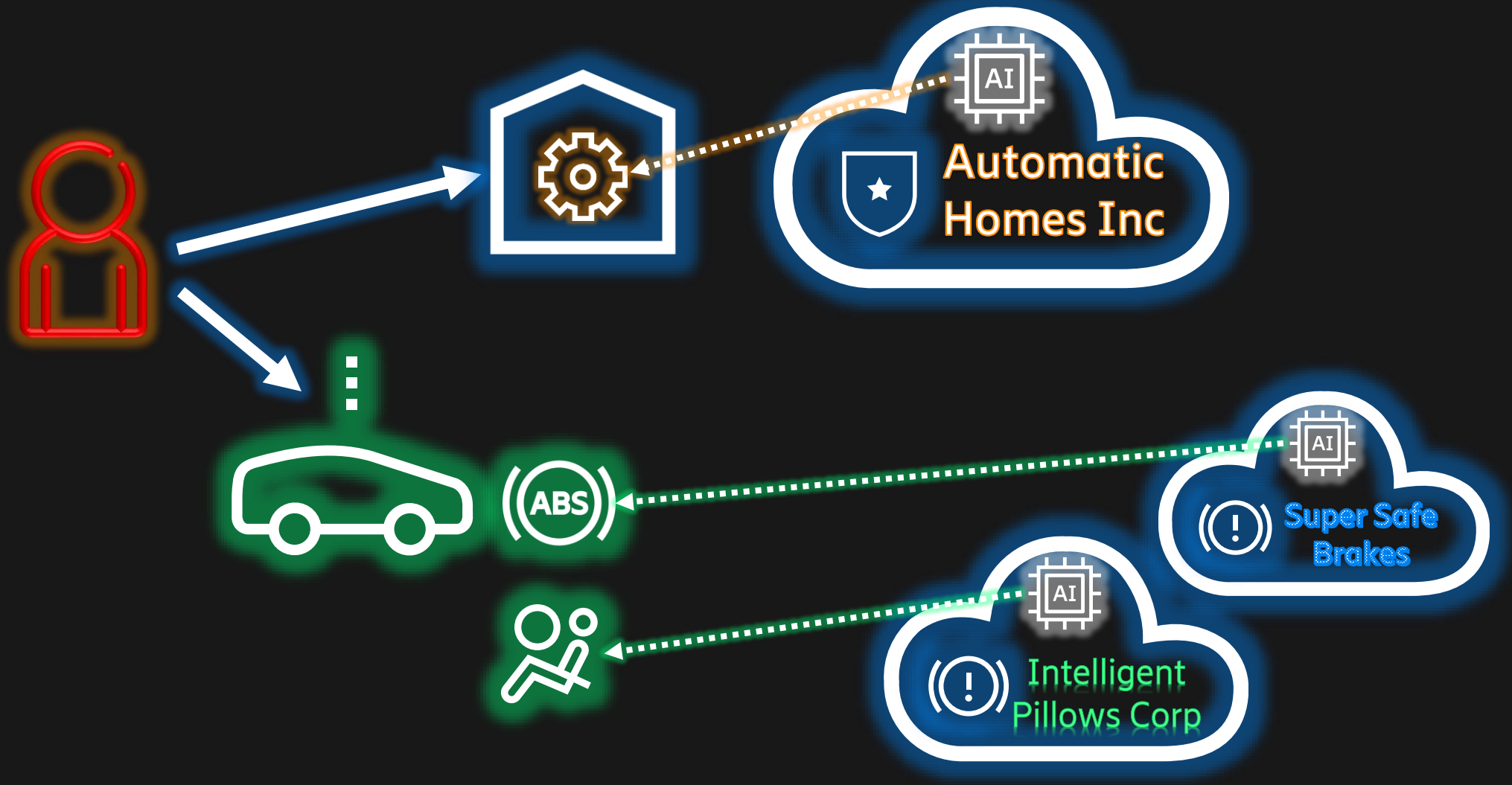
Edgar Ramos



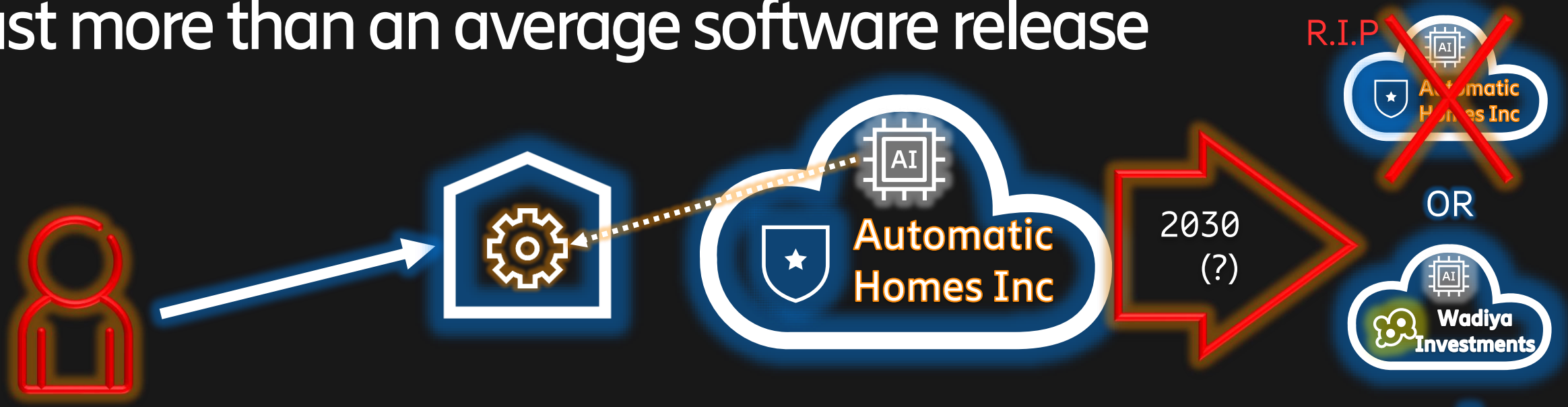
Intelligence is much more than one application in one device

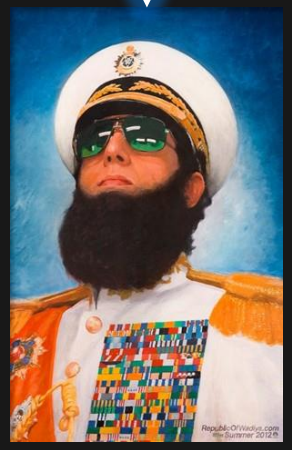


Intelligence is in devices and systems are expected to last more than an average software release

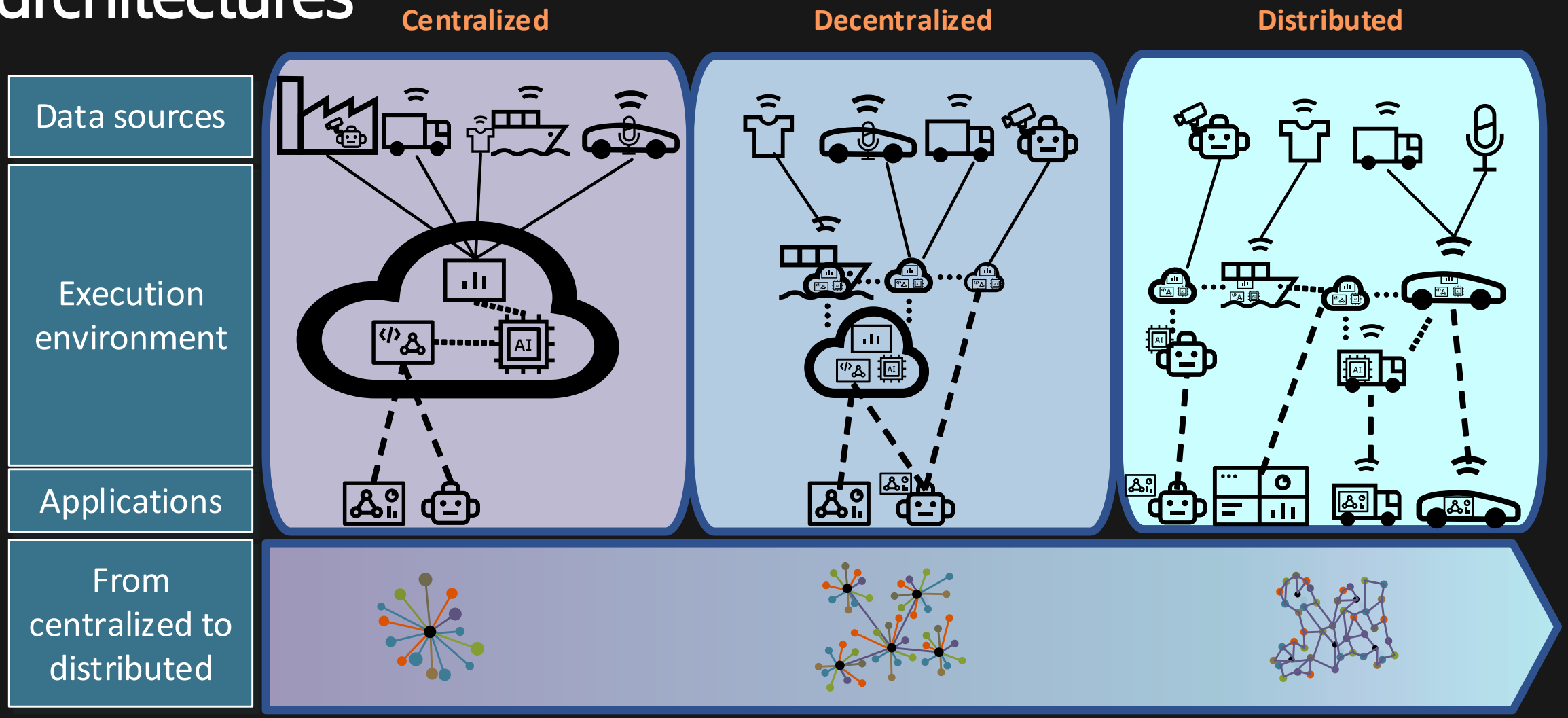


Intelligence is in devices and systems are expected to last more than an average software release



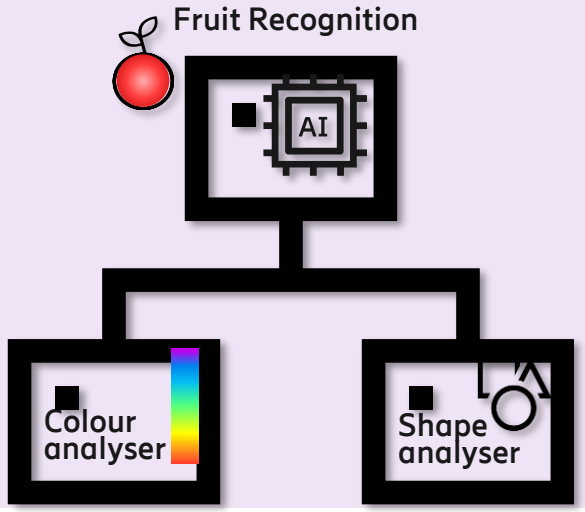


MI Centralized vs. Distributed architectures

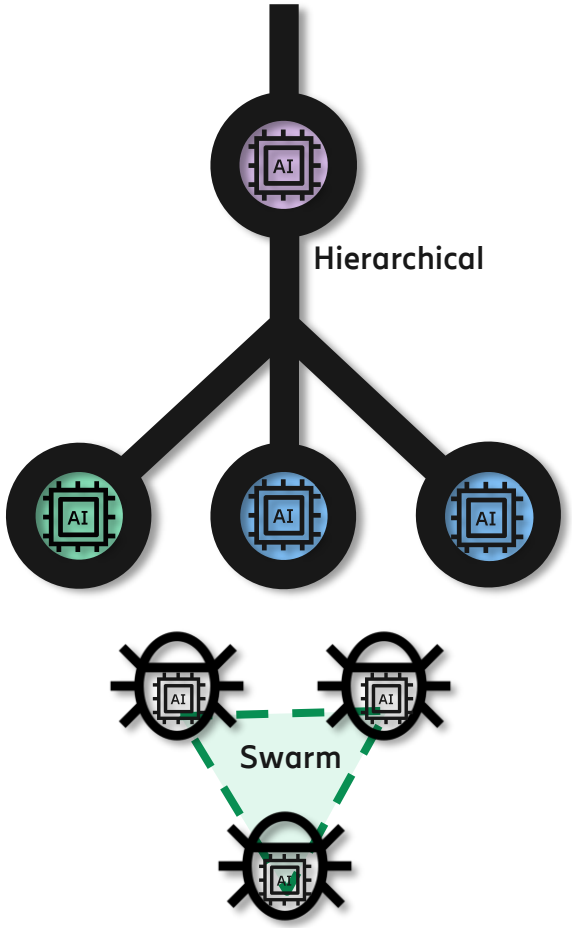


Machine Intelligence Distribution Aspects

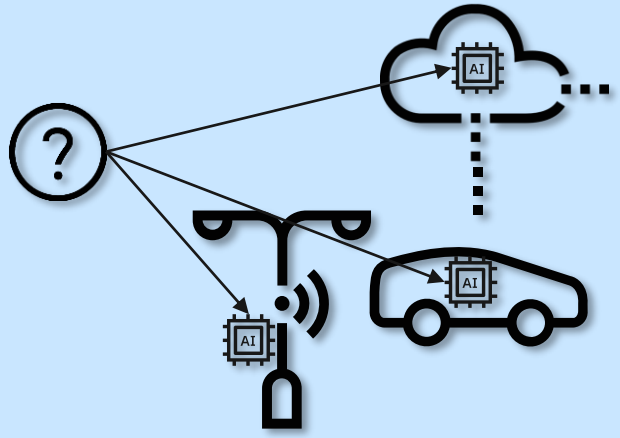
AI Service Functional Distribution



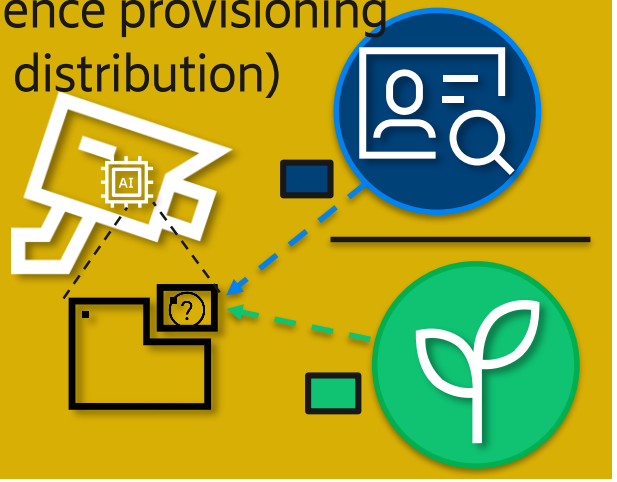
Agent Functional Distribution



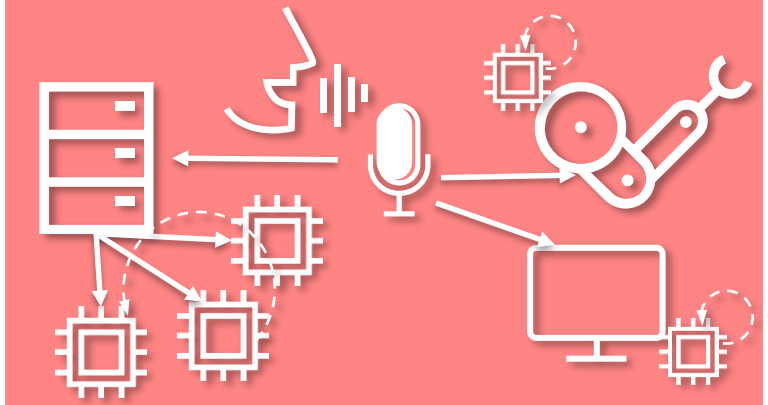
Inference Execution Distribution



Intelligence provisioning (Model distribution)

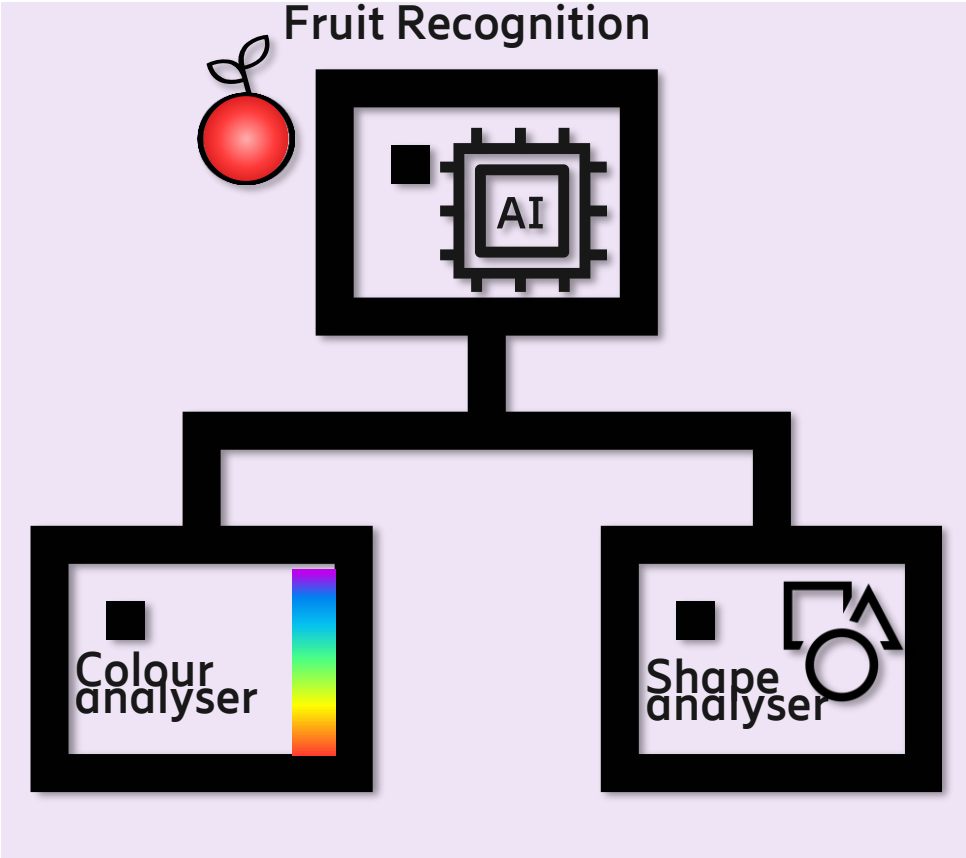


Training Execution Distribution



Machine Intelligence Distribution Aspects (1/5)

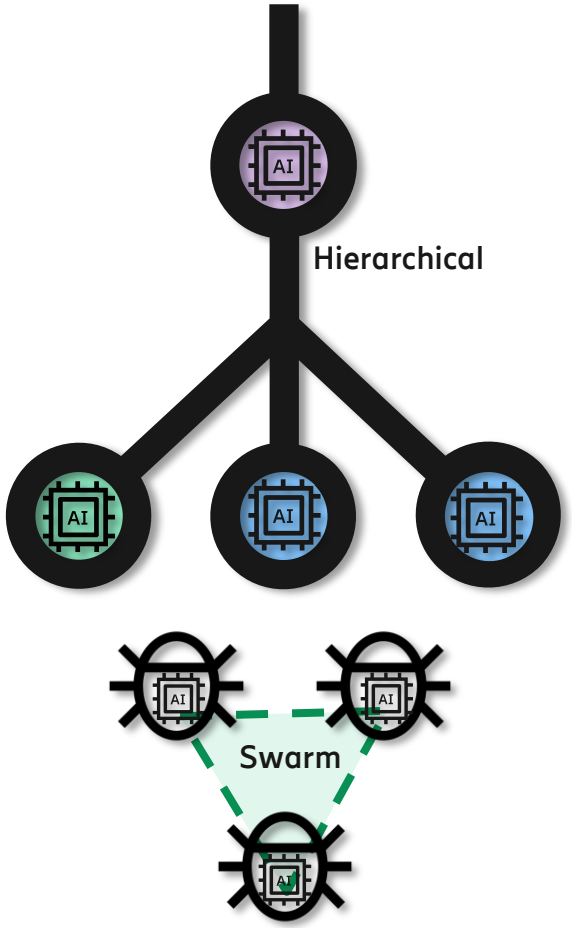
AI Service Functional Distribution



- AI Models **modularity**
- Combination (**composition**) of multiple domains and techniques to achieve a concrete task
- A concrete example:
  Generative Adversarial Network (GAN)

Machine Intelligence Distribution Aspects (2/5)

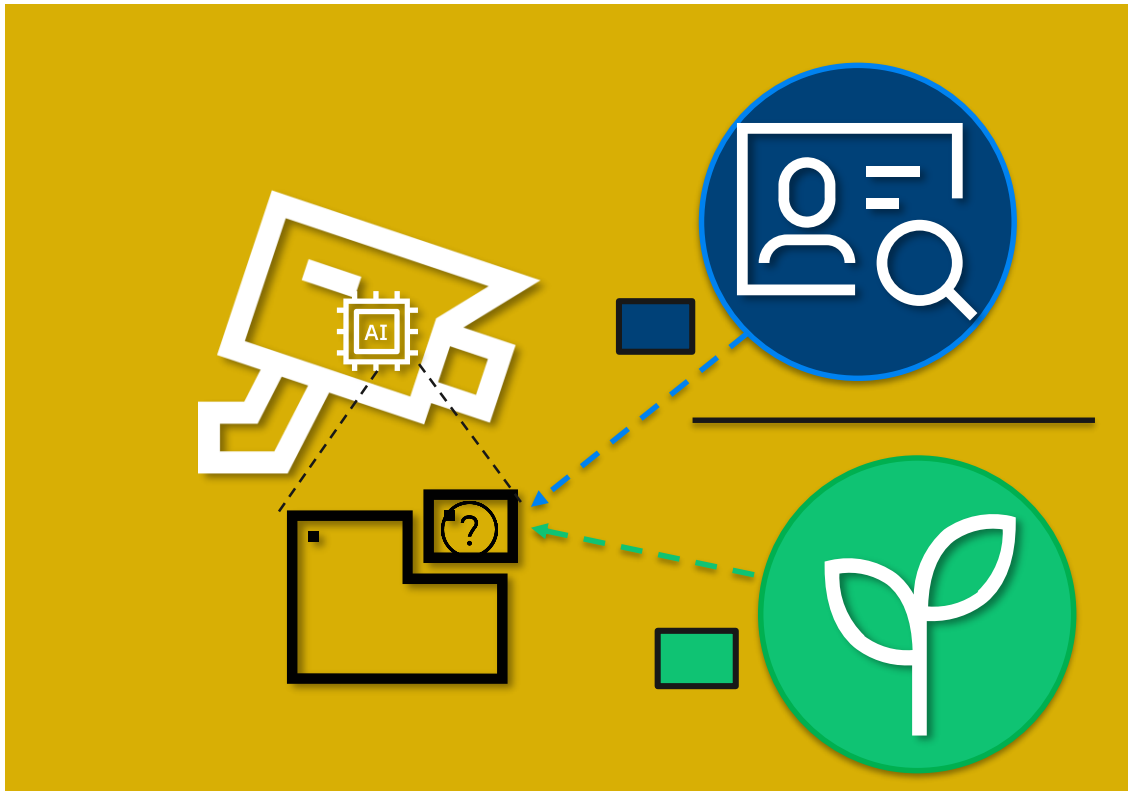
Agent Functional Distribution



- Intelligence **interaction** with other intelligences
- The **rational agents** strategies and conducts involved
- The degree of **communication** between agents
- The perception of intelligence **organization**
- Interesting example:
  Neural networks that design other neural networks

Machine Intelligence Distribution Aspects (3/5)

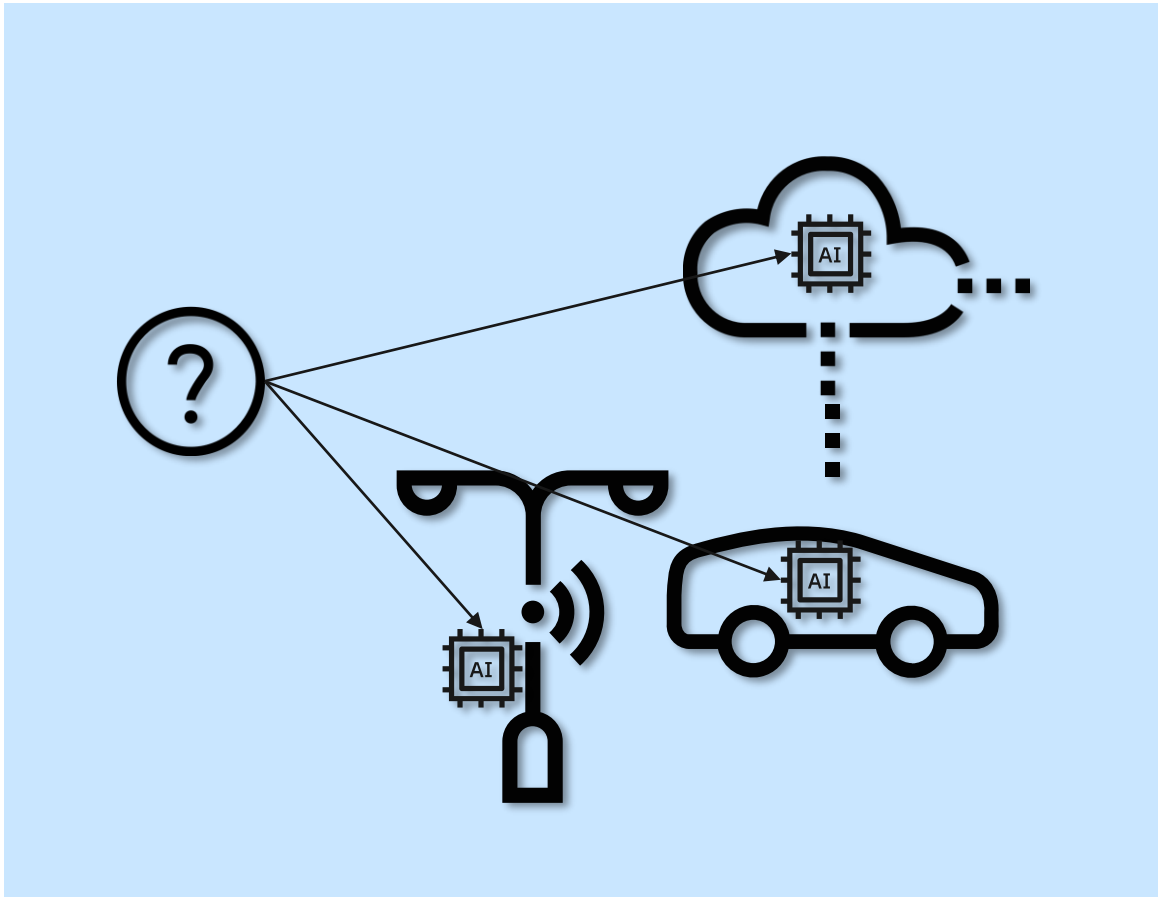
Intelligence provisioning (Model distribution)



- **Onboard** or **update** an intelligence model
- Exploiting **general hardware**, cloud facilities and local acceleration from AI
- Mass marketing and **standardizing** AI models
- Example:
  ONNX (Open Neural Network Exchange)

Machine Intelligence Distribution Aspects (4/5)

Inference Execution Distribution



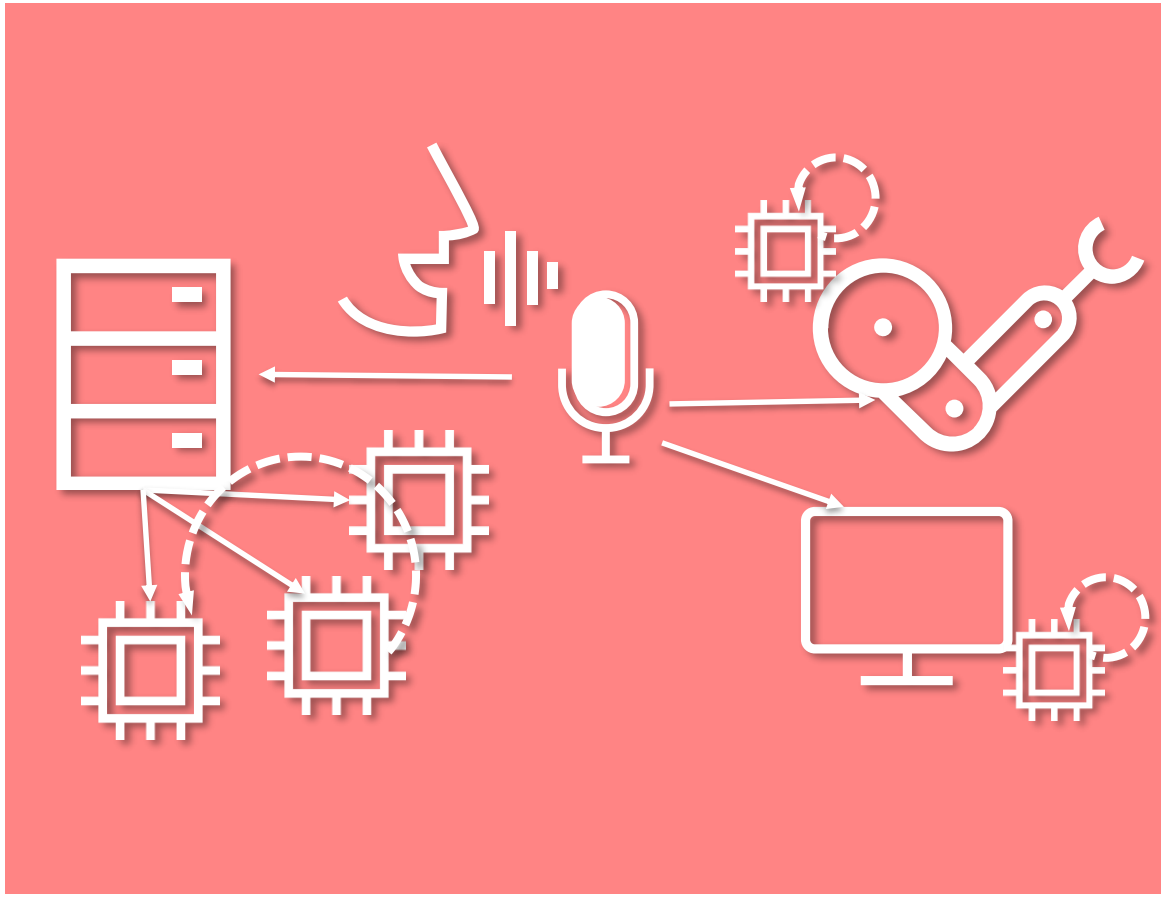
- **Where and how** the intelligence inferencing is executed
- Privacy, hardware capabilities, connectivity, latency and control-loop response times as main discriminator **requirements**
- May include different intelligence solutions for different **domains**
  - Execution domain may also switch according to availability, **environment** changes or variable requirements

Interesting example:

Actor Based microservices frameworks (e.g. Calvin)

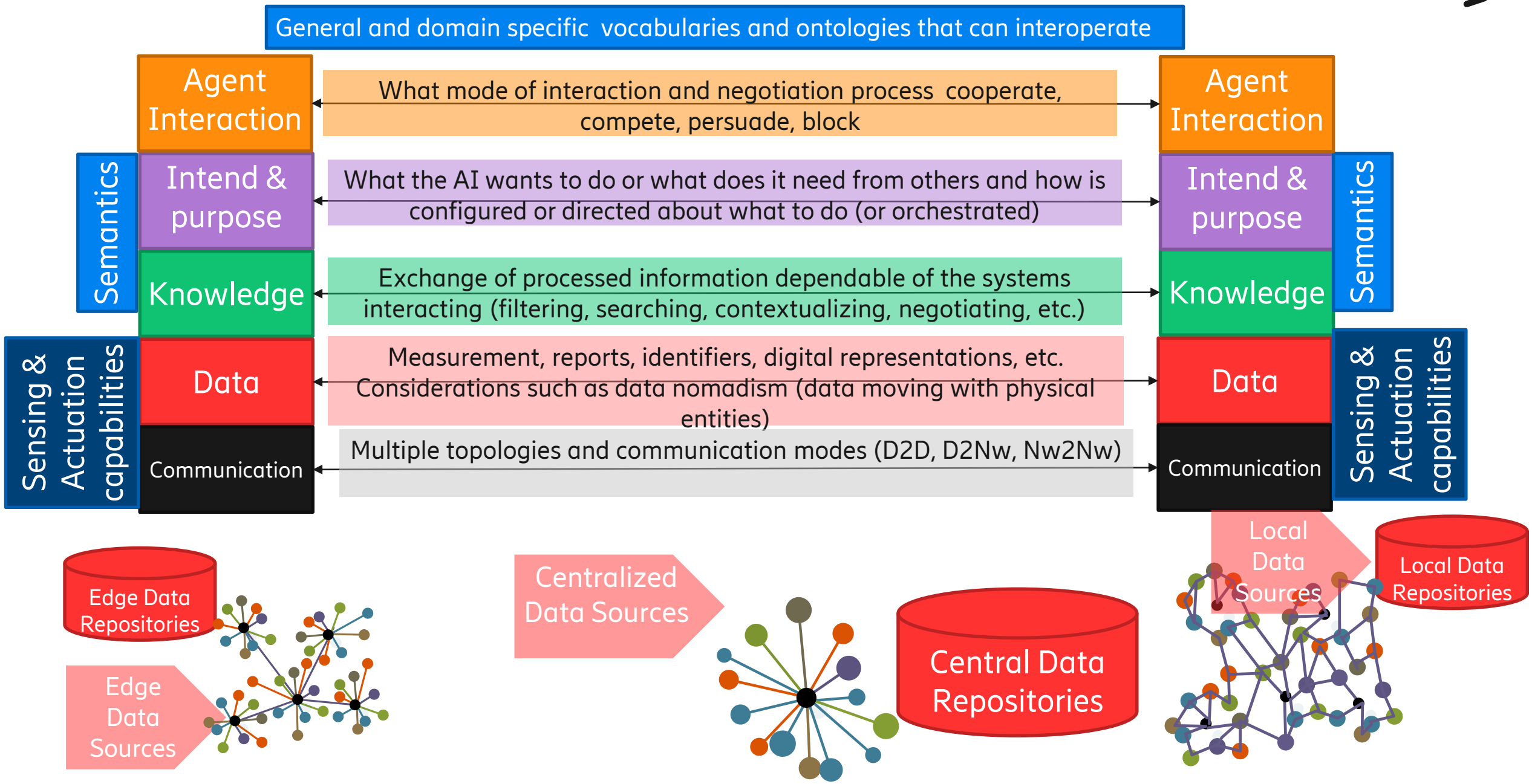
Machine Intelligence Distribution Aspects (5/5)

Training Execution Distribution

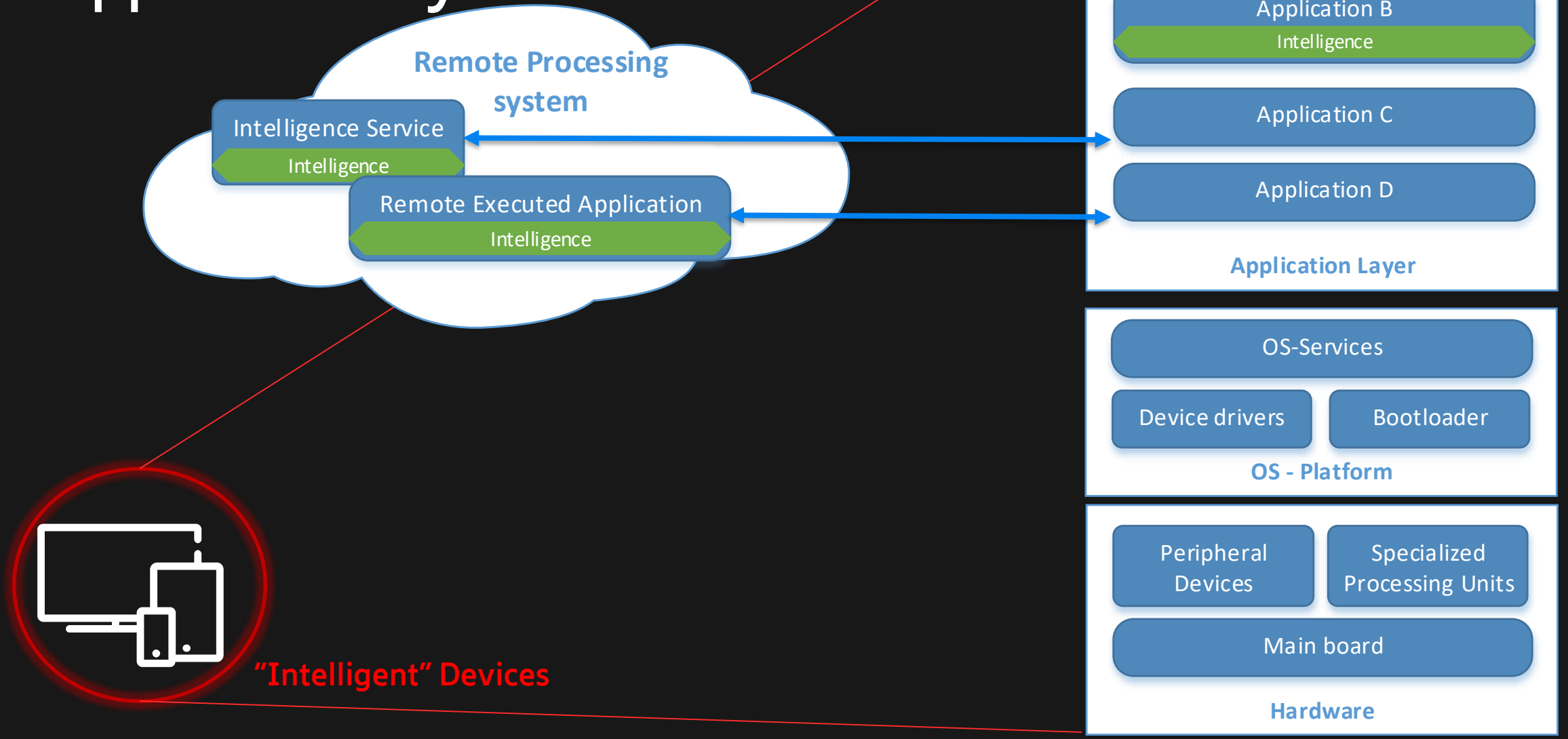


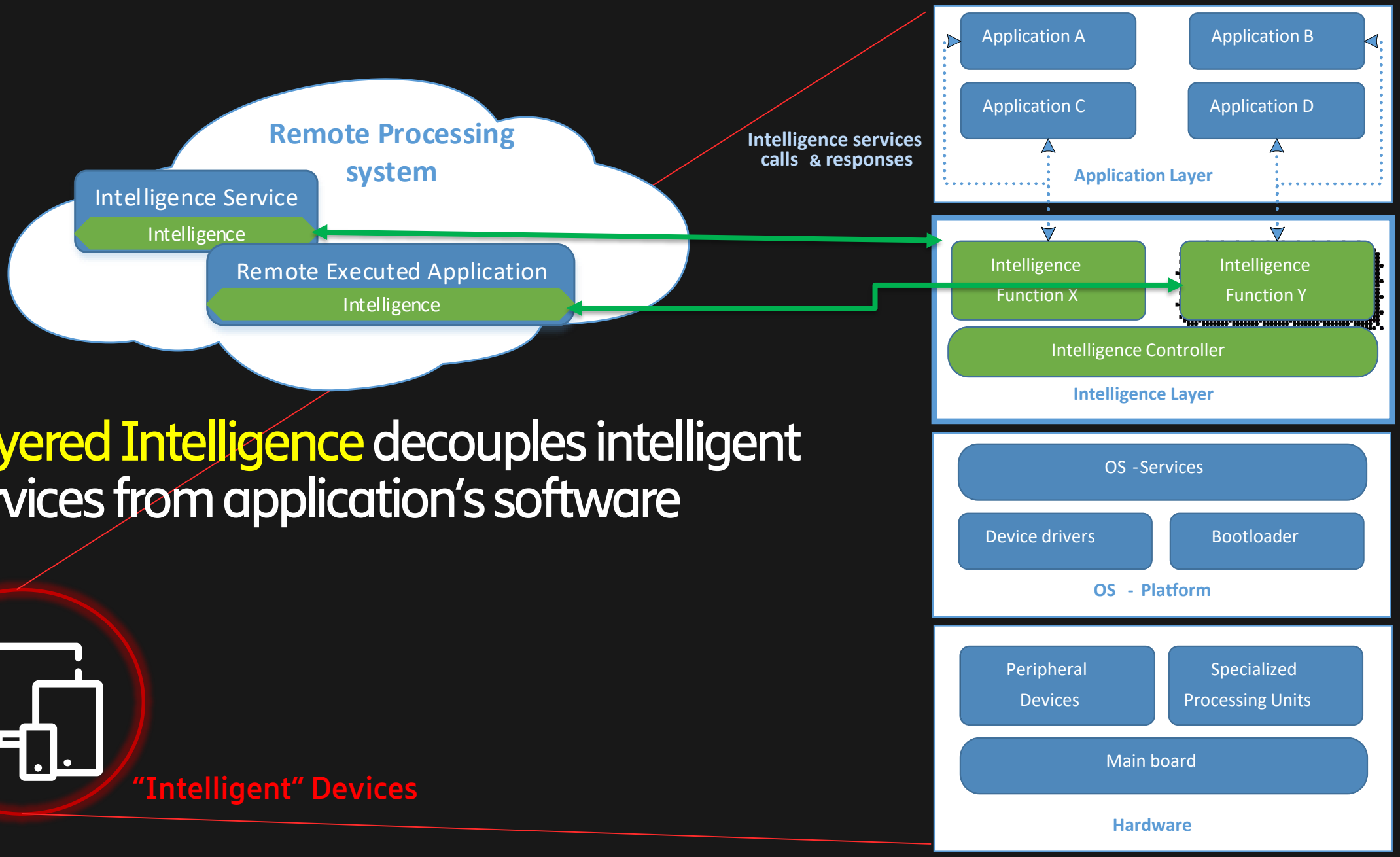
- Where and how the intelligence **training is executed**
- Similar than inference execution distribution
- Additionally to hardware constrains and particularities of the inference hardware, the **location and volumes of training data** and the sensitivity plays a big role in the decision of distribution
- **Continuously generated data** and consolidation is another consideration
- Interesting example:
  Federated learning

High-level AI-2-AI Network stack

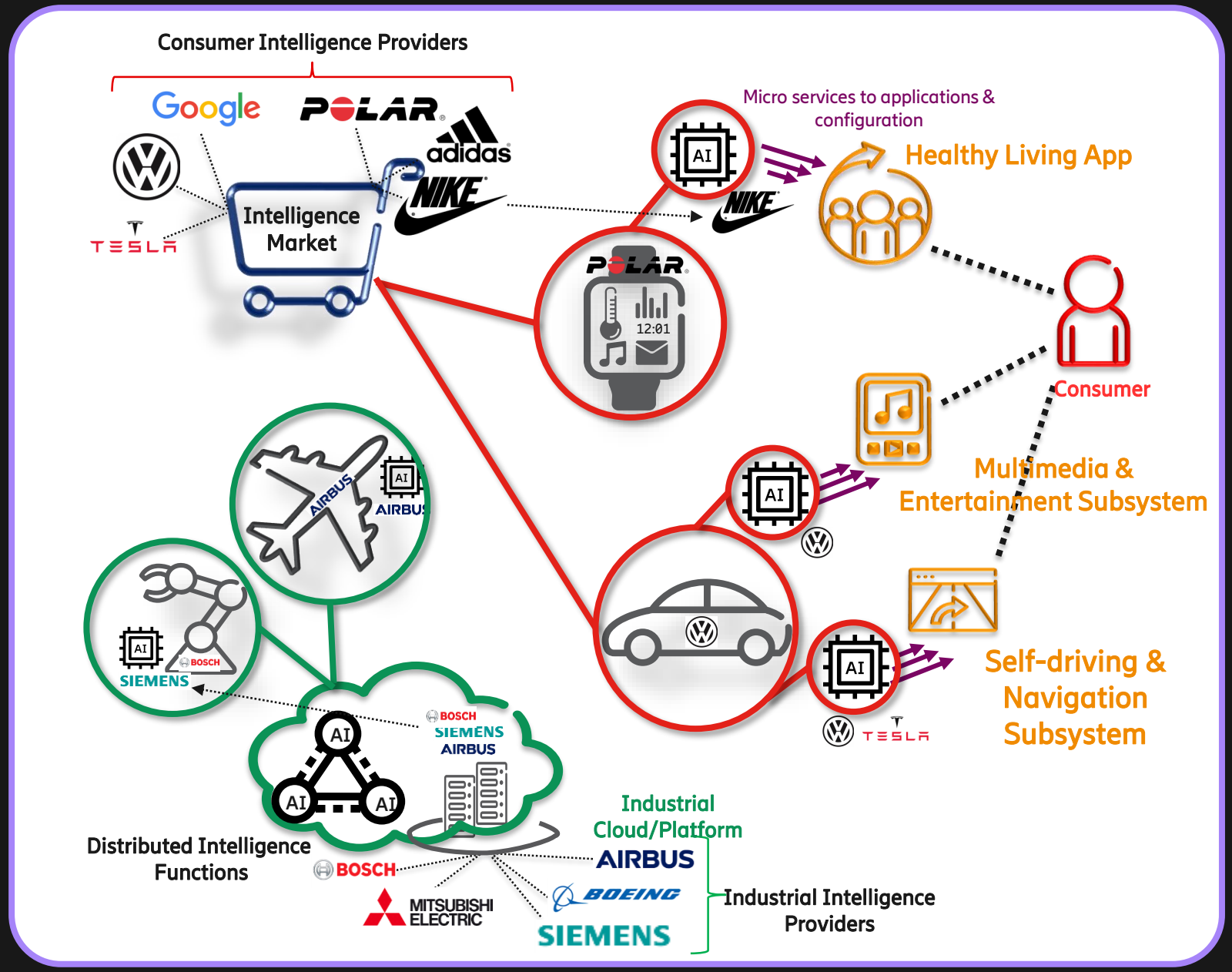


Legacy Intelligence belongs to the application layer in the Stack



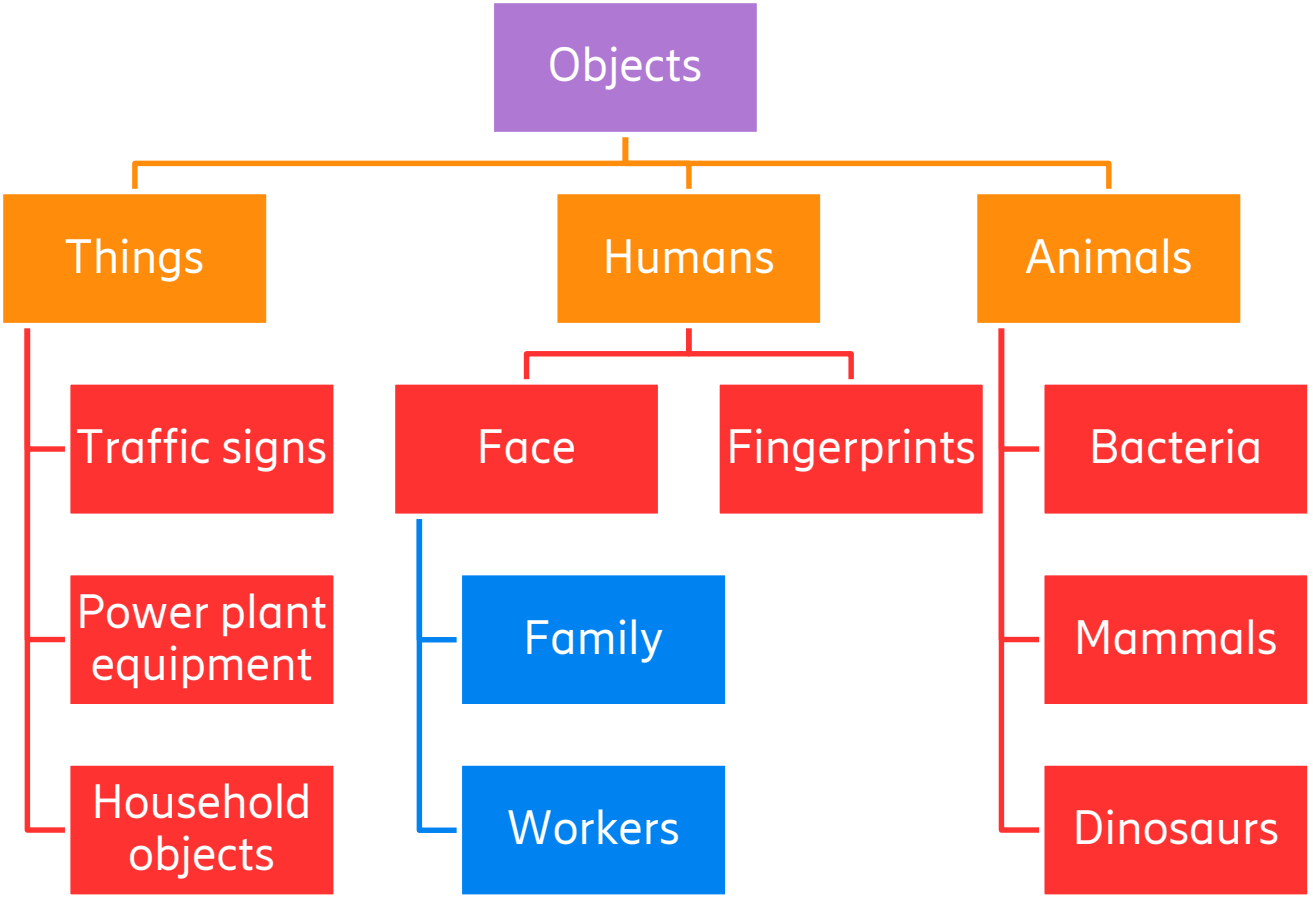


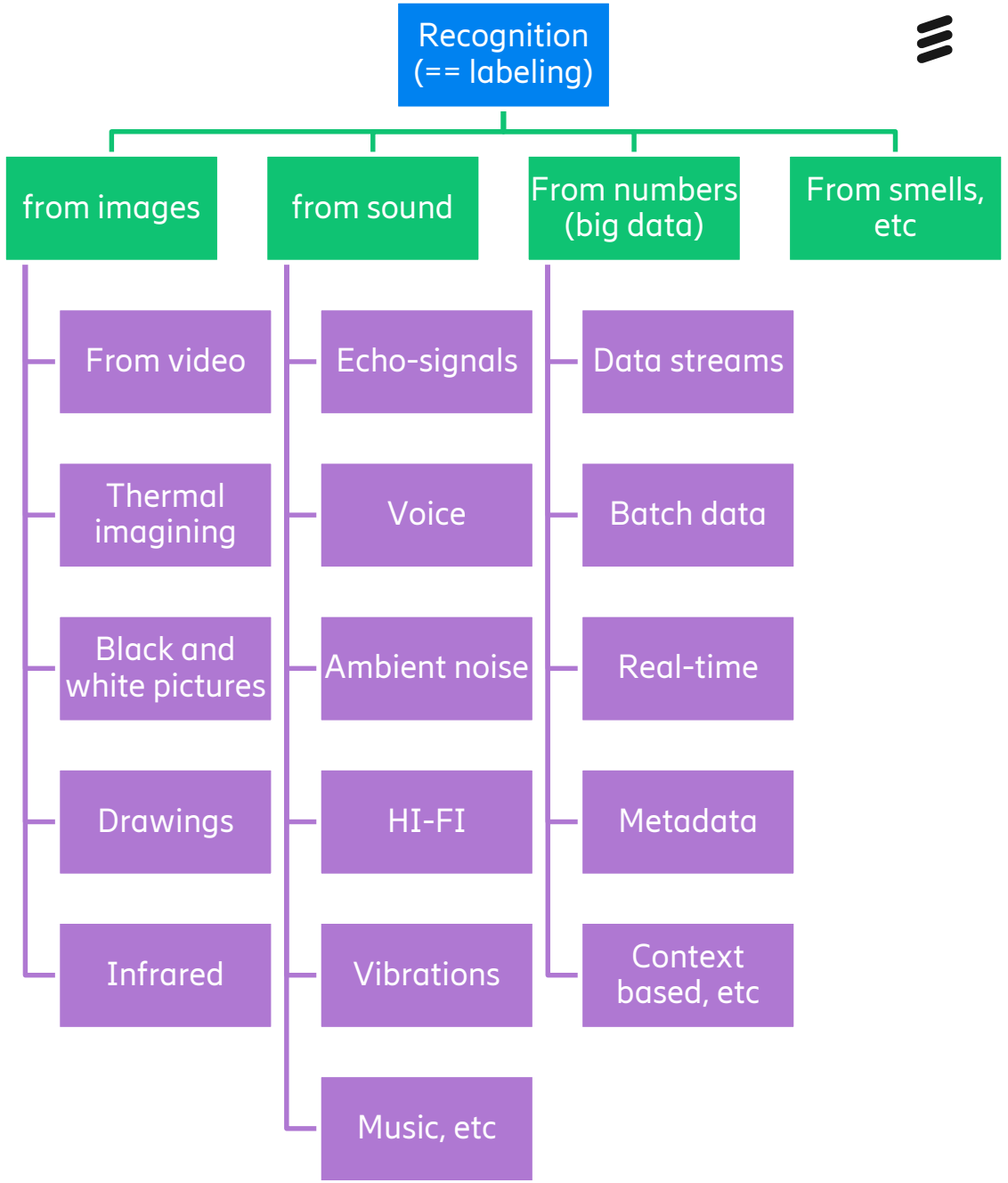
Layered Intelligence decouples intelligent services from application's software



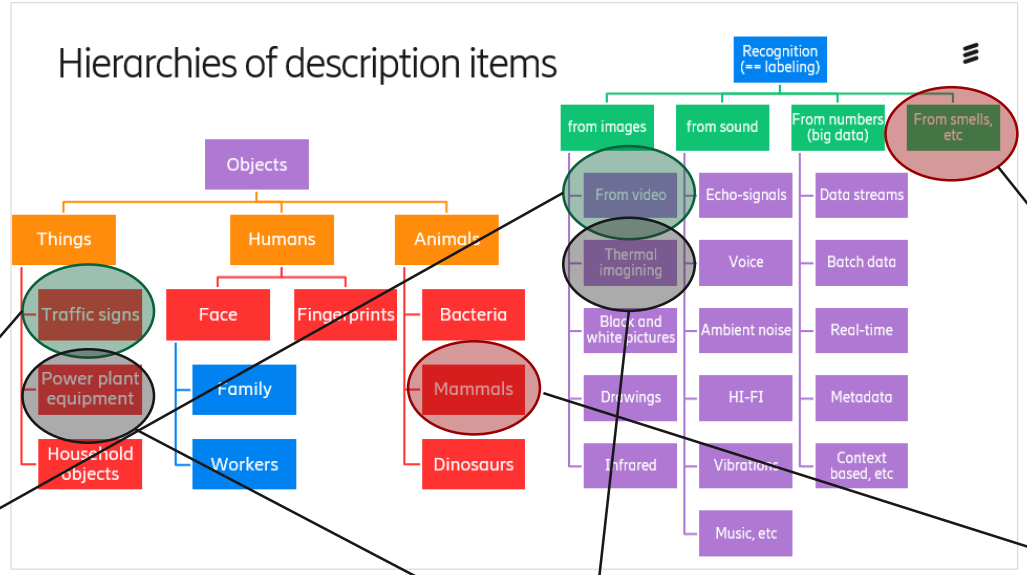
An ecosystem emerges when intelligence is distributed all the way to devices

Hierarchies of description items





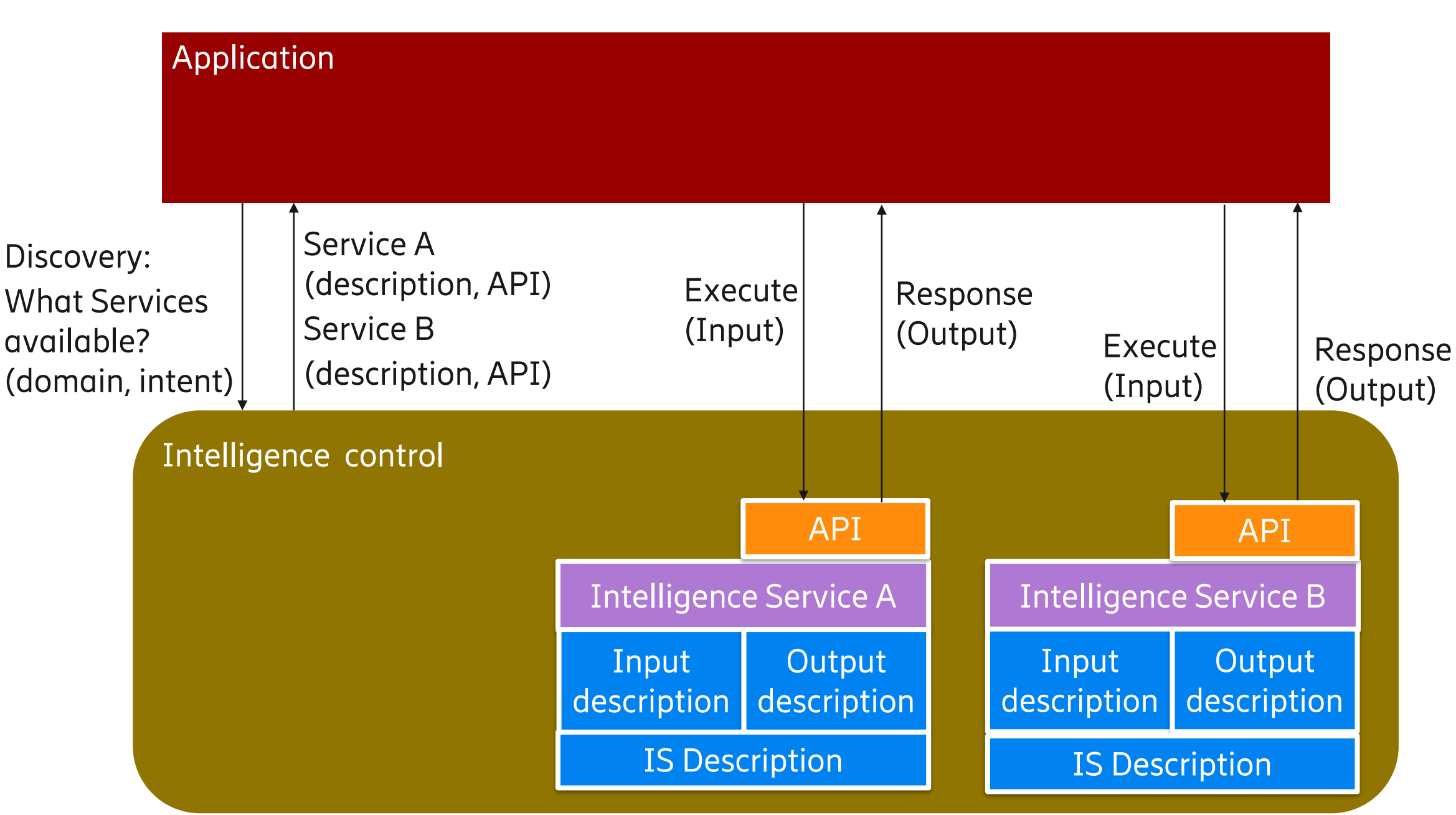
IS – simple description example



-Traffic sign
- Recognition from video

-Power plant equipment
- Recognition from thermal images

- Mammals
- Recognition from smell

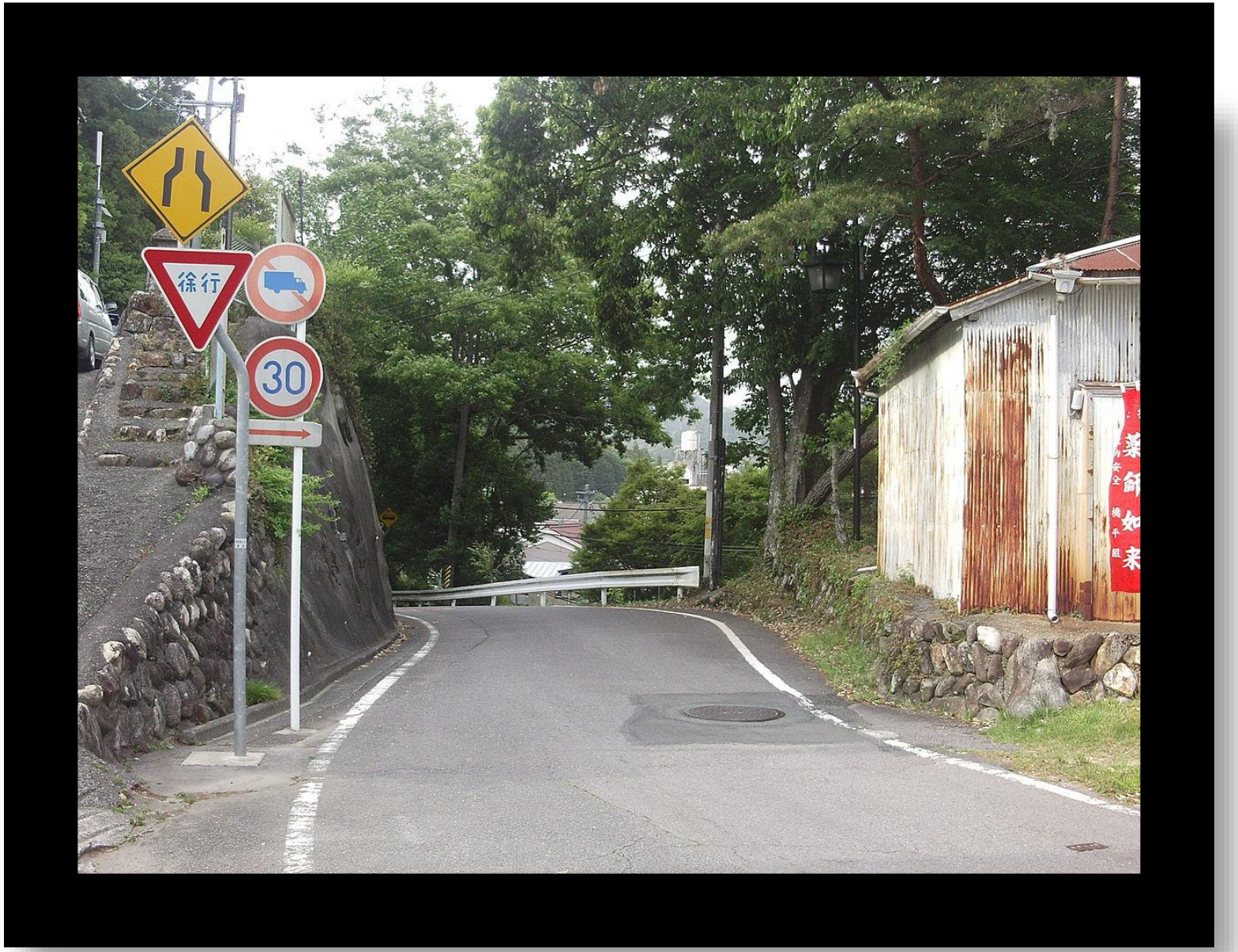

Application
Discovery:
What Services
available?
(domain, intent)
Service A
(description, API)
Service B
(description, API)
Execute
(Input)
Response
(Output)
Execute
(Input)
Response
(Output)
Intelligence control
API
Intelligence Service A
Input
description
Output
description
IS Description
API
Intelligence Service B
Input
description
Output
description
IS Description

API

Intelligence Service

Picture	Traffic Sign category

Traffic sign recognition



API

Intelligence Service A

Picture	Traffic Sign category

Traffic sign recognition

API

Intelligence Service B

Picture	Traffic Sign category, Coordinates in image

Traffic sign recognition

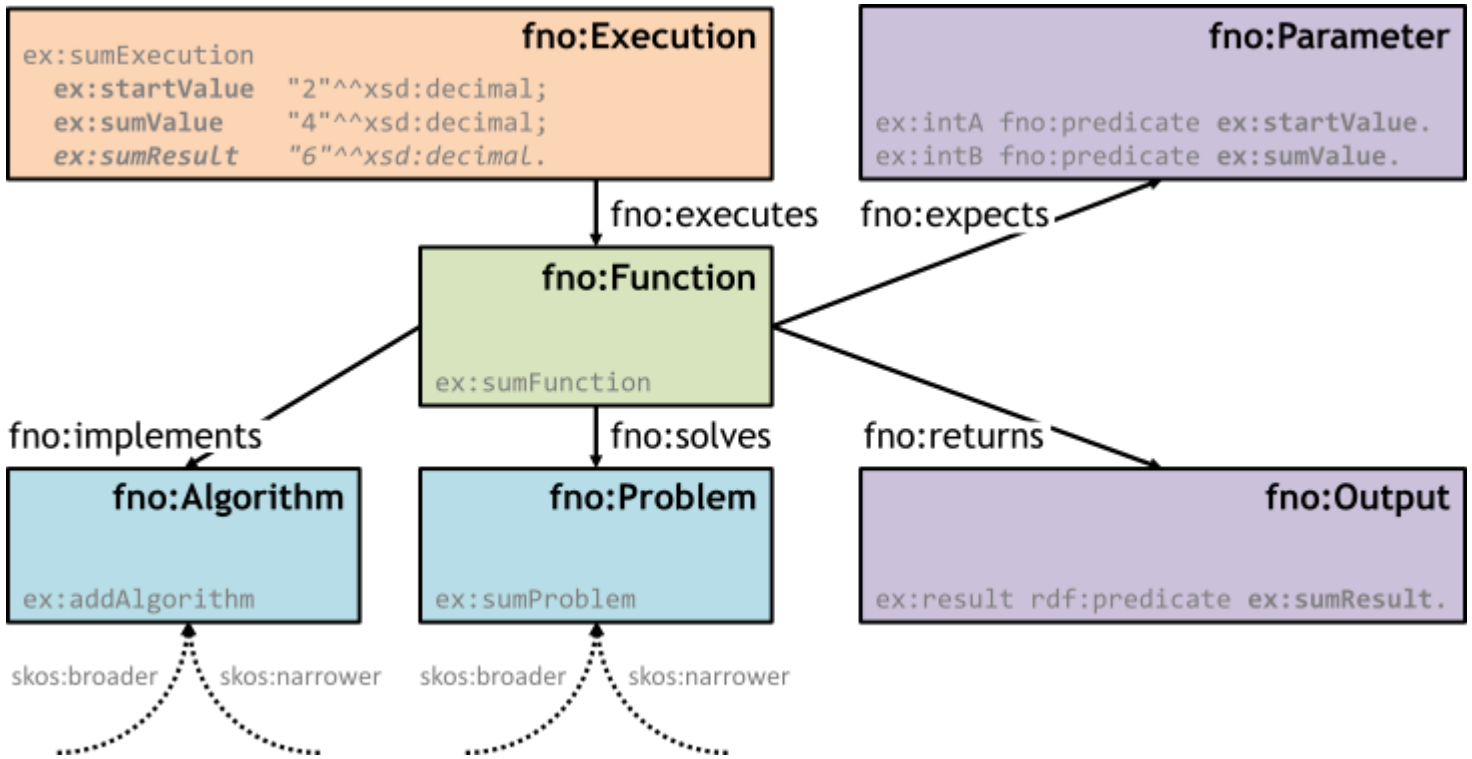
API

Intelligence Service C

Picture, Language to translate	Traffic Sign category, Coordinates in image, Translation

Traffic sign recognition and translation

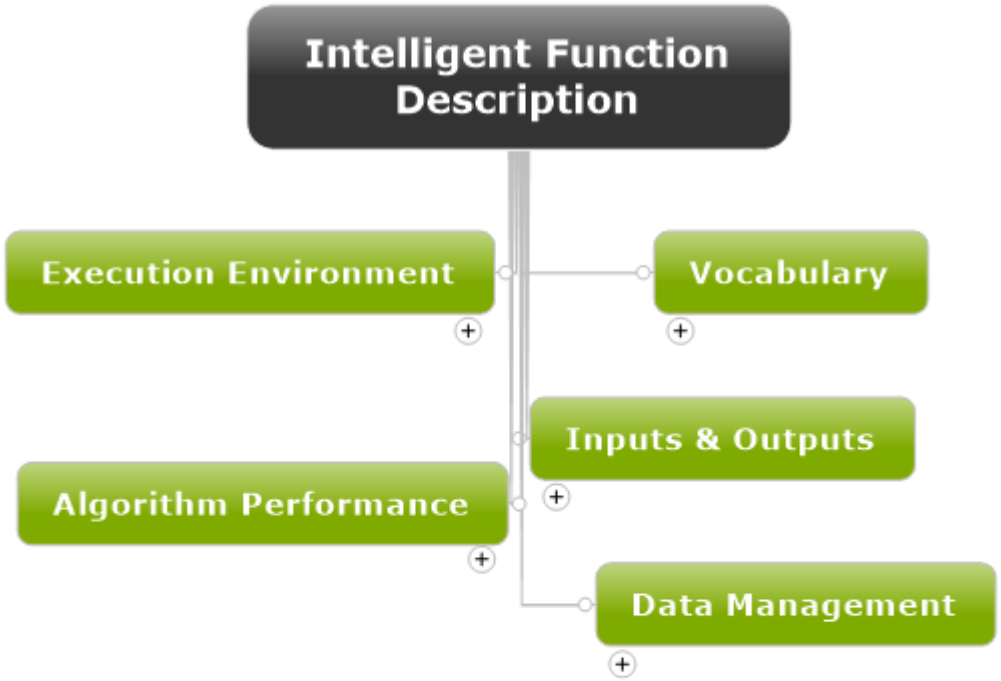
The function ontology: Good starting



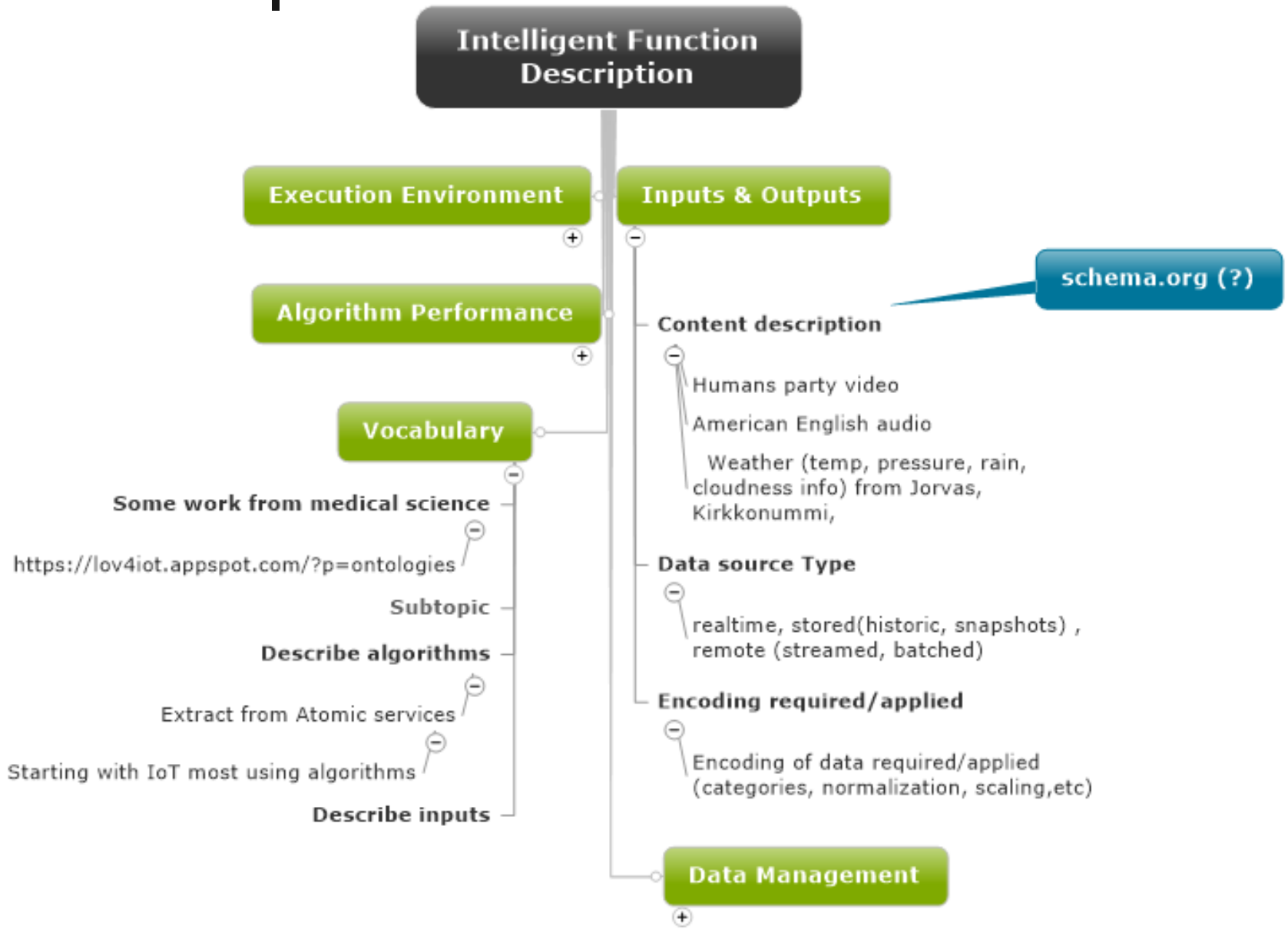
https://fno.io/
https://fno.io/spec/
https://fno.io/ontology/index-en.html

What else?

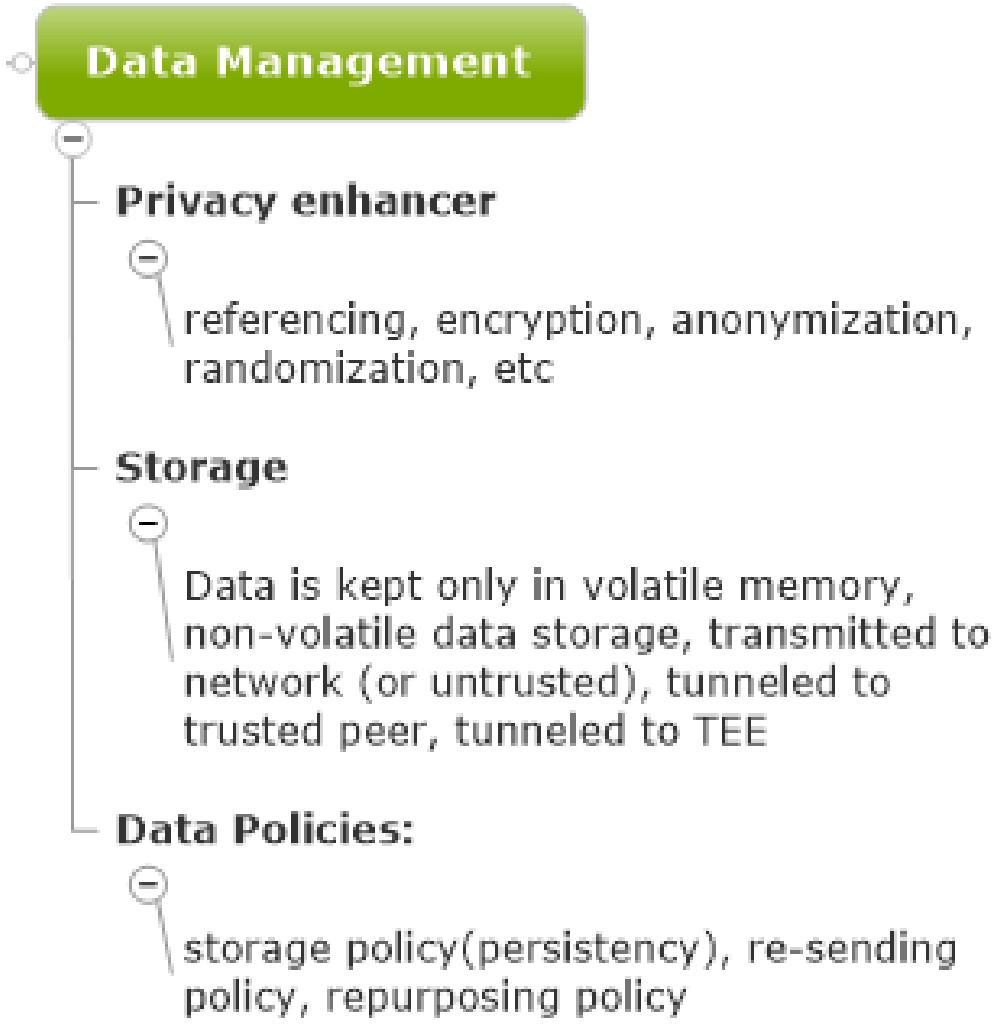
- Vocabulary of description
- More semantically meaningful Inputs -& output
- Execution environment
- Algorithm performance
- Hardware performance
- Data management



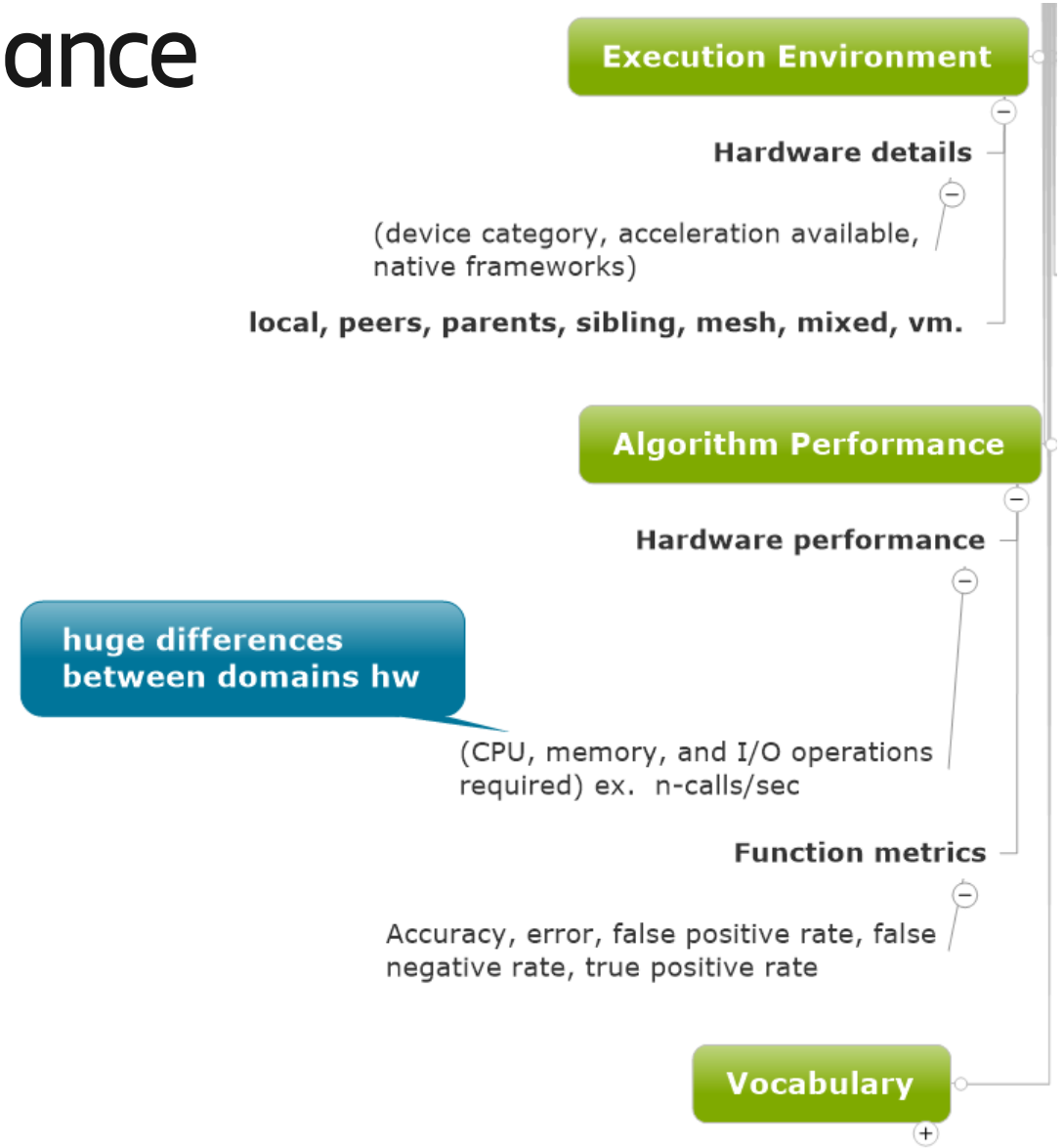
Input & Output descriptions

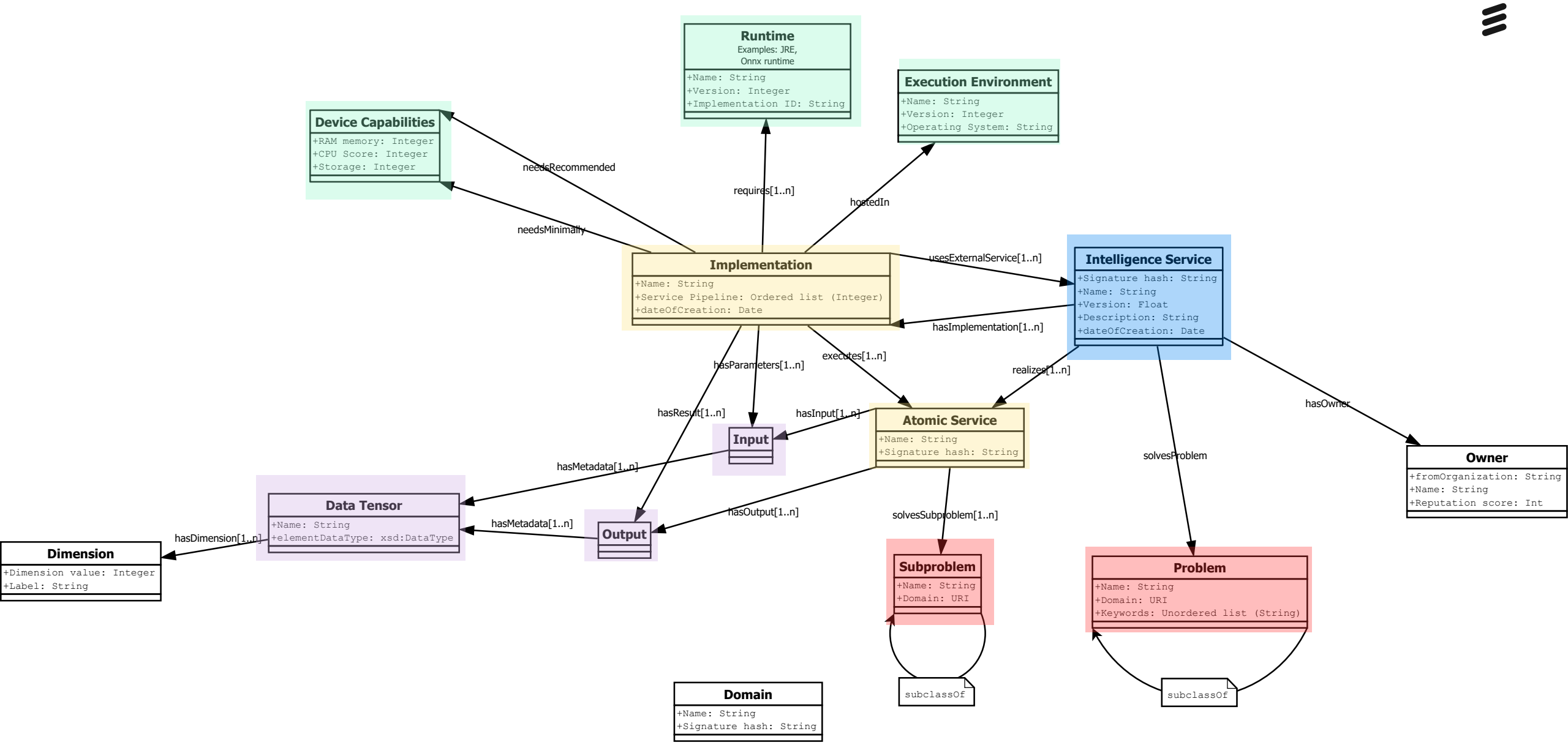


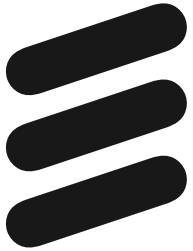
Data management policies



Execution environment & Algorithm Performance




Runtime
Examples: JRE,
Onnx runtime
+Name: String
+Version: Integer
+Implementation ID: String
Execution Environment
+Name: String
+Version: Integer
+Operating System: String
Device Capabilities
+RAM memory: Integer
+CPU Score: Integer
+Storage: Integer
needsRecommended
needsMinimally
requires[1..n]
hostedIn
Implementation
+Name: String
+Service Pipeline: Ordered list (Integer)
+dateOfCreation: Date
usesExternalService[1..n]
Intelligence Service
+Signature hash: String
+Name: String
+Version: Float
+Description: String
+dateOfCreation: Date
hasImplementation[1..n]
hasParameters[1..n]
executes[1..n]
realizes[1..n]
hasResult[1..n]
hasInput[1..n]
Input
Atomic Service
+Name: String
+Signature hash: String
hasOwner
Owner
+fromOrganization: String
+Name: String
+Reputation score: Int
hasMetadata[1..n]
solvesProblem
Data Tensor
+Name: String
+elementDataType: xsd:DataType
hasMetadata[1..n]
Output
hasOutput[1..n]
solvesSubproblem[1..n]
hasDimension[1..n]
Dimension
+Dimension value: Integer
+Label: String
Subproblem
+Name: String
+Domain: URI
Problem
+Name: String
+Domain: URI
+Keywords: Unordered list (String)
subclassOf
subclassOf
Domain
+Name: String
+Signature hash: String



Ericsson.com/IoT