AI at the Edge of IoT

Dr. L S Jayashree
Professor, Dept of CSE
PSG College of Technology
Coimbatore

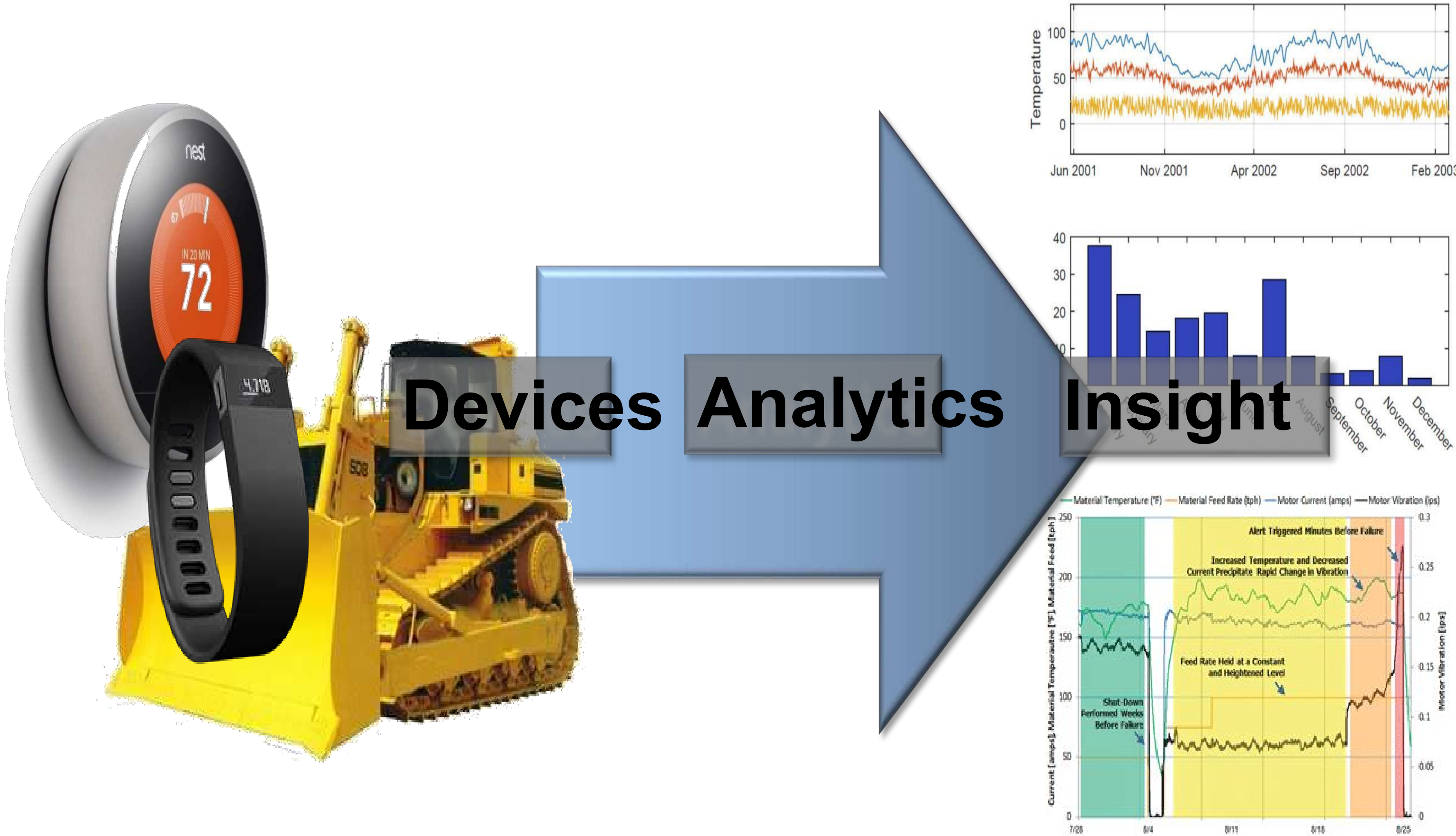
The agenda

- The IoT Data Processing Architecture
- Types of IoT Data Analytics
- Cloud Analytics vs Edge Analytics
- Why Edge Computing ?
- The Hardware and Software Enablers of Edge Computing

- differe
- networking of physical devices –
  - Ex: traffic light in smart city

Physical things + computing + connectivity + sensors + actuators = IoT

IoT Analytics- A Broad Picture



Why IoT ?

Connect → Monitor → Analyze → Control → Create Value

Physical things can sense, communicate and collaborate creating a greater intelligent outcome to bring more value to the users of those things

Users

- Monitor & control devices remotely
- Use efficiently
- Service life of things can extend
- Improved experience through collaboration of things
- Autonomy of things

Manufacturers / Administrators

- Real-time insights into device location, condition, usage and performance
- Open new opportunity to monetize value added services around product usage
- Ability to improve future products
- Infield product upgrades for extended lifecycle

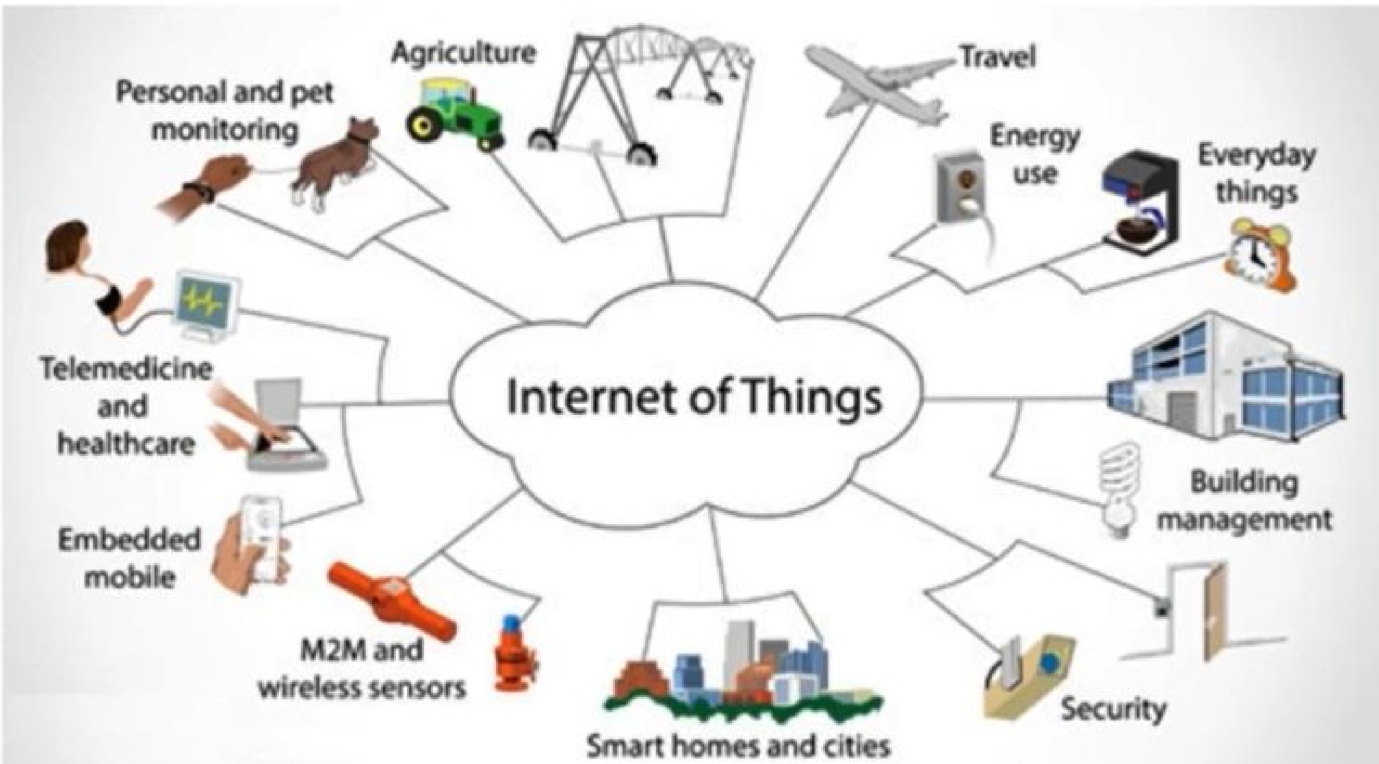
Broad Categories

- **Consumer IoT** – Smart home, Smart buildings

- **Industrial IoT** – machines in a particular industry gets connected Ex: Medical equipments can be controlled and monitored

- **Civic IoT** – smart public services like transportation , smart grid etc..

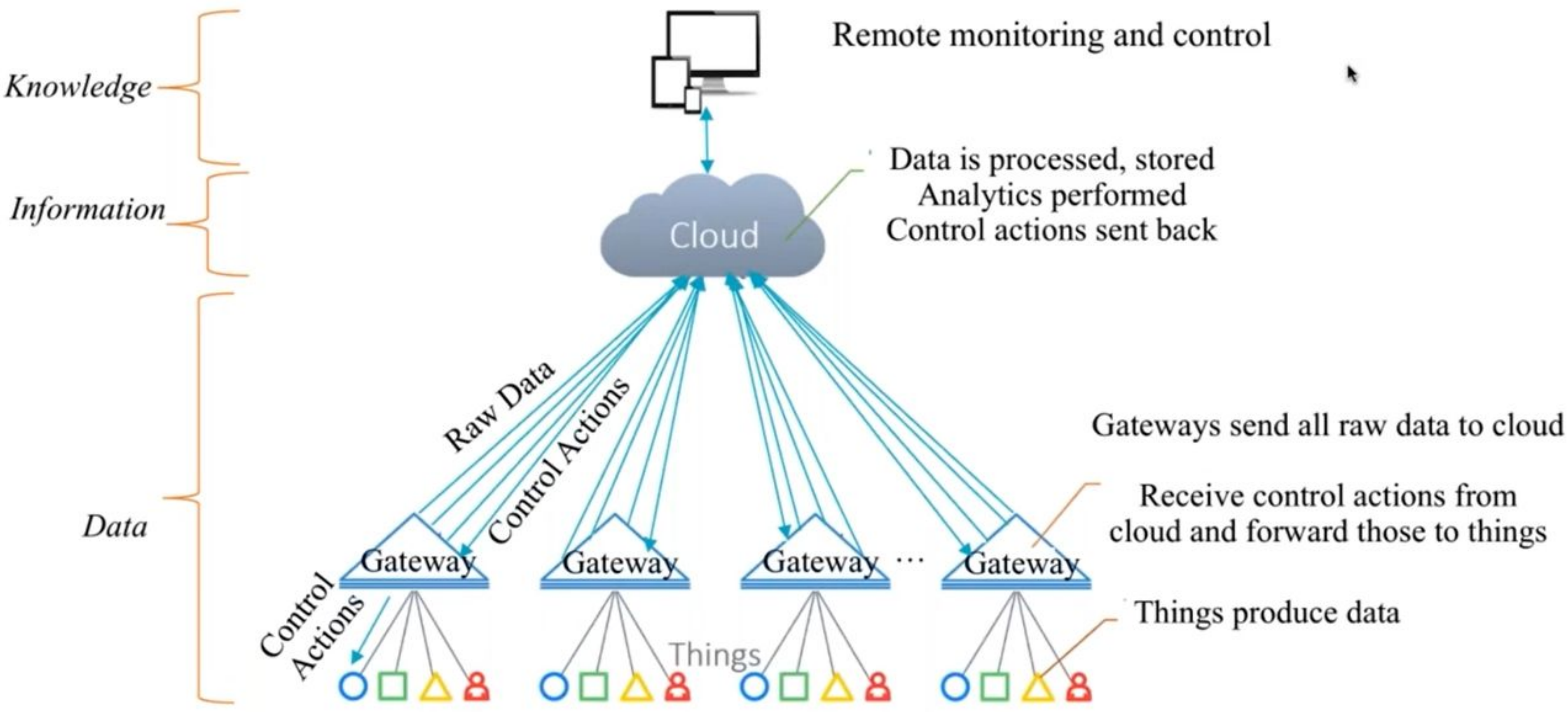
Consumer IoT ***Industrial IoT*** ***Civic IoT***



What does it take?

- Sensors / Actuators
- Single Board Computers, SoCs
- Embedded programming
- Intelligent Devices (mobile)
- Internet
- Cloud computing
- Big Data
- Analytics
- Artificial Intelligence

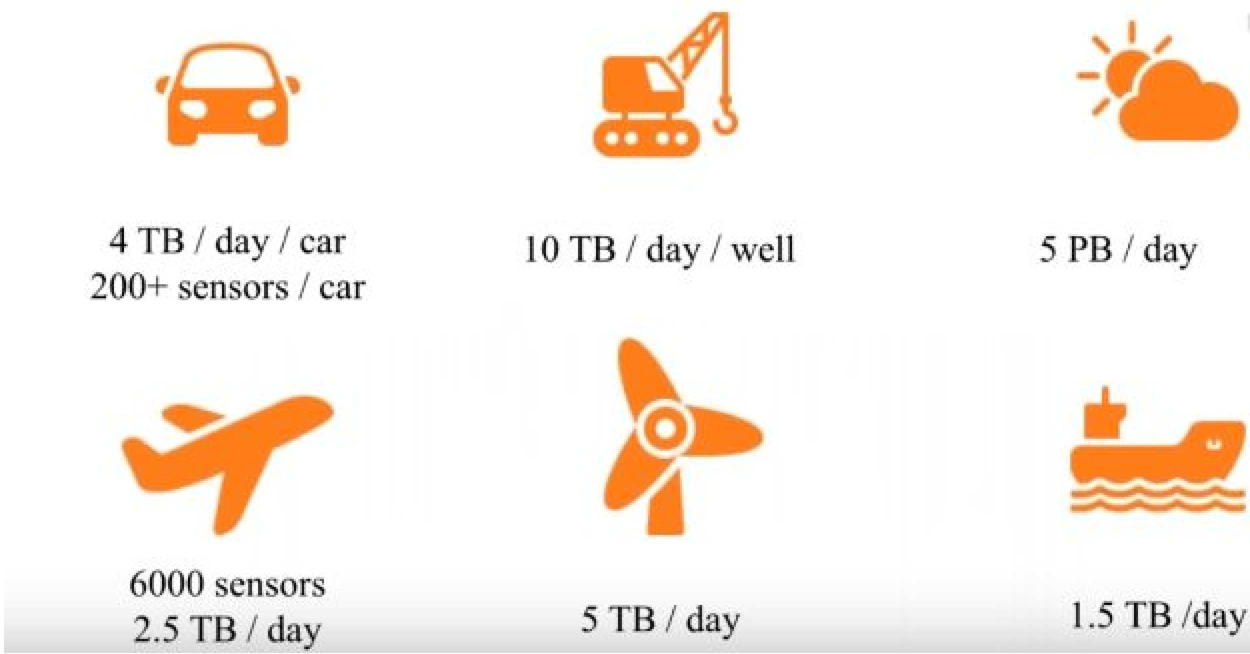
ARCHITECTURE



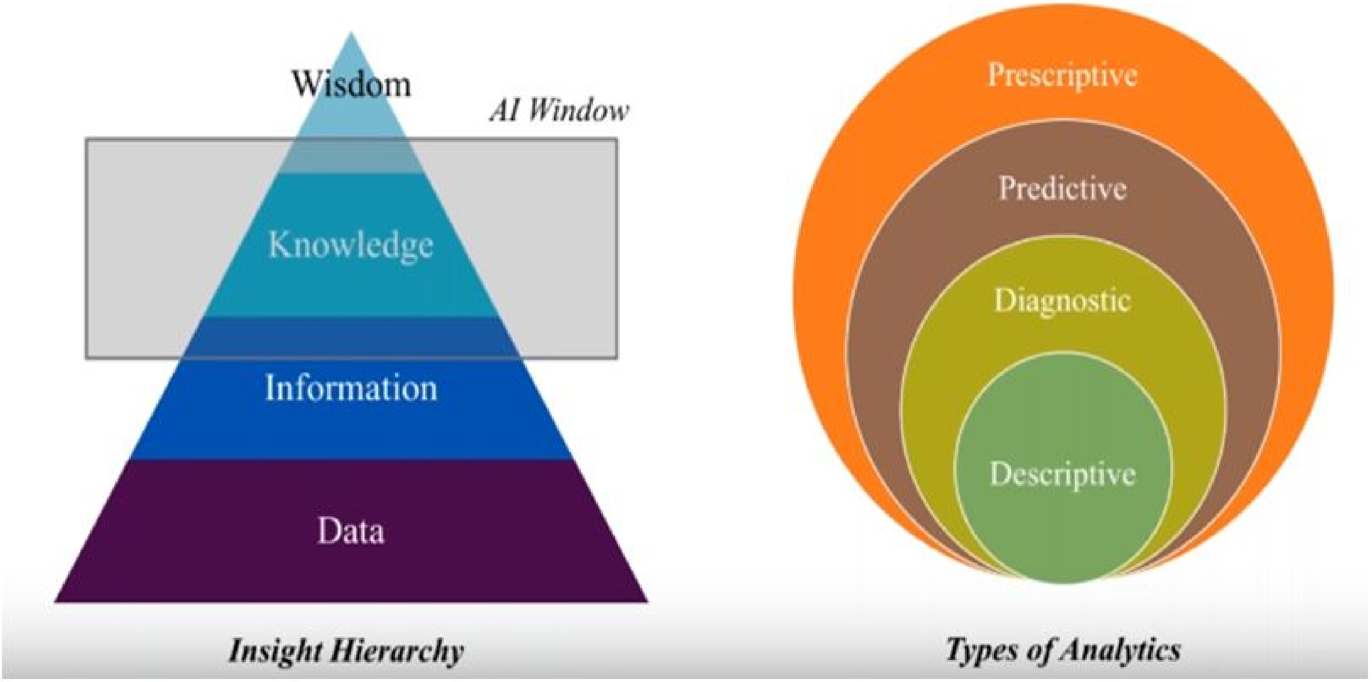
Challenges

- Security
- Privacy
- Interoperability
- Over-the-air upgrades
- Huge data volumes (data intensive)
- Real-time actionable insights
- Complex event processing
- Standardization has not kept pace with growth

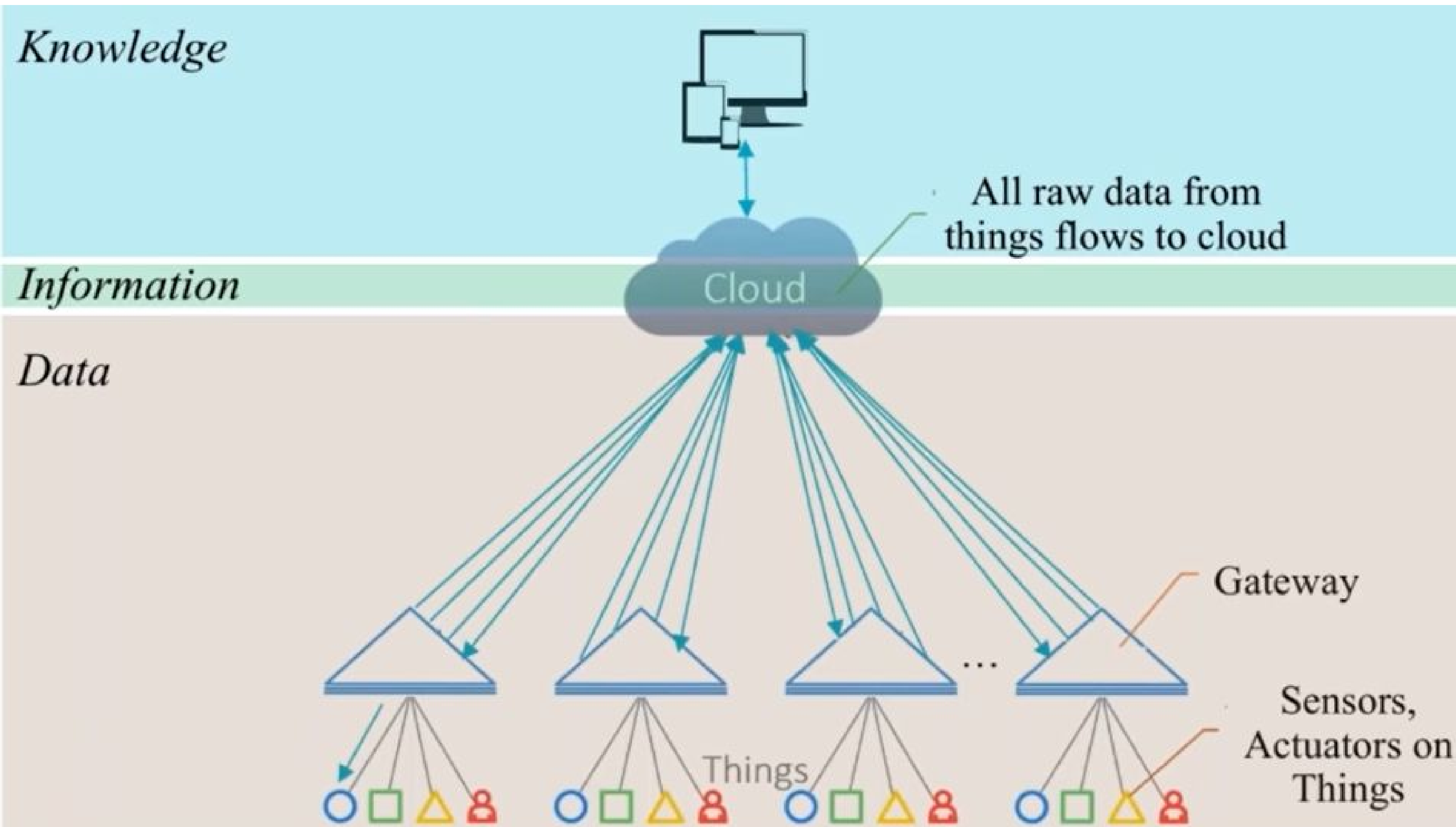
Data Volumes



How to draw insights??



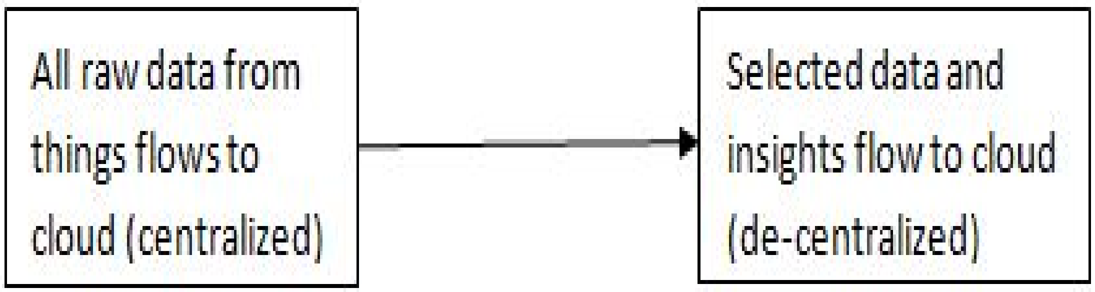
Insights in Centralized IoT



Data Processing challenges

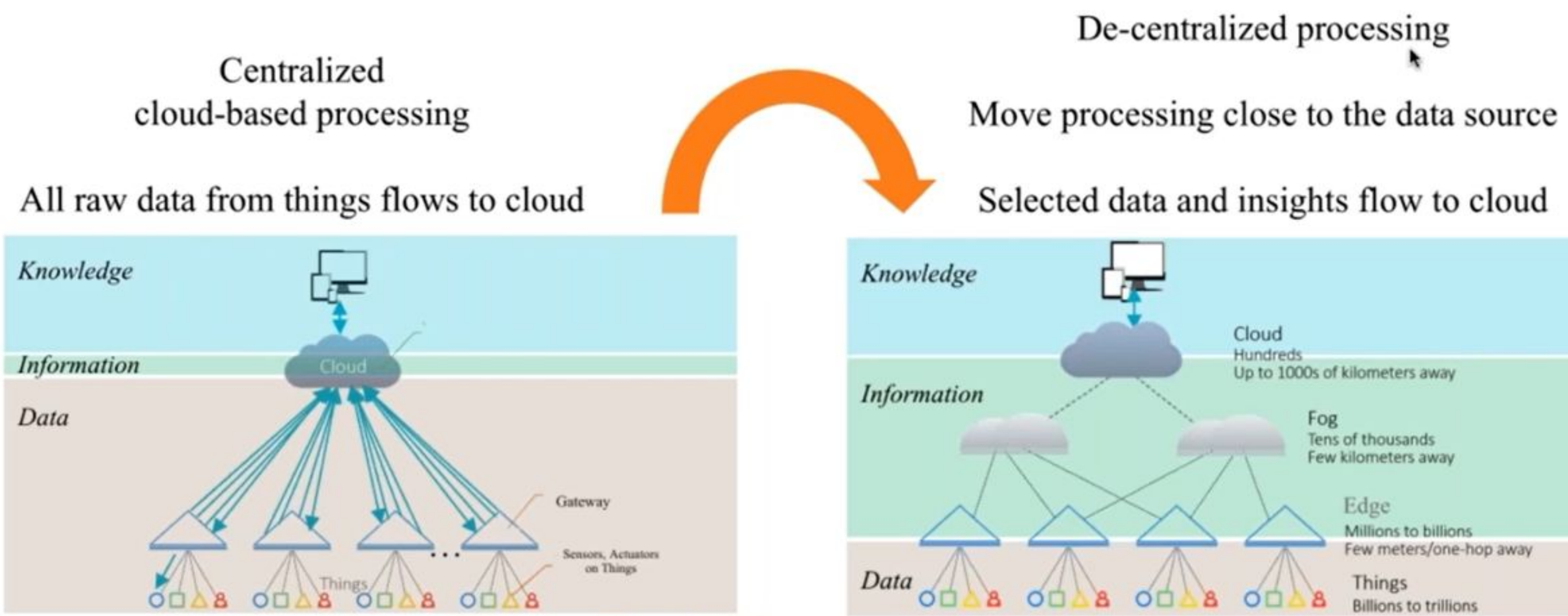
- Intermittent Connection
- Insufficient Bandwidth
- Delayed (No) actions
- No real-time insights
- Security & Compliance
- High Latency
- Perishable data

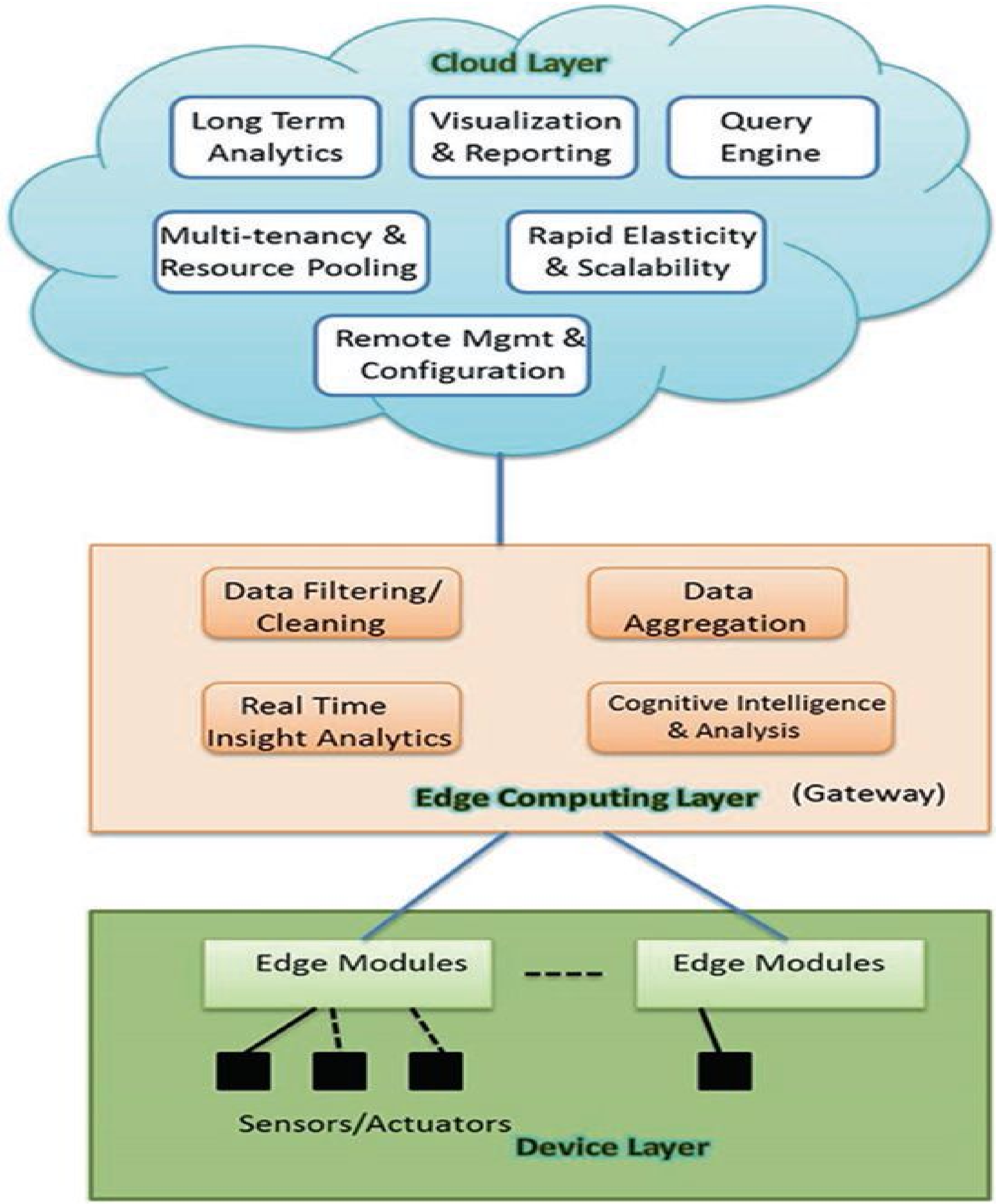
  (data loses its value when not used appropriately)



The solution to the problem is the **de-centralized processing with edge/fog computing.**

Solution – Edge / Fog Computing

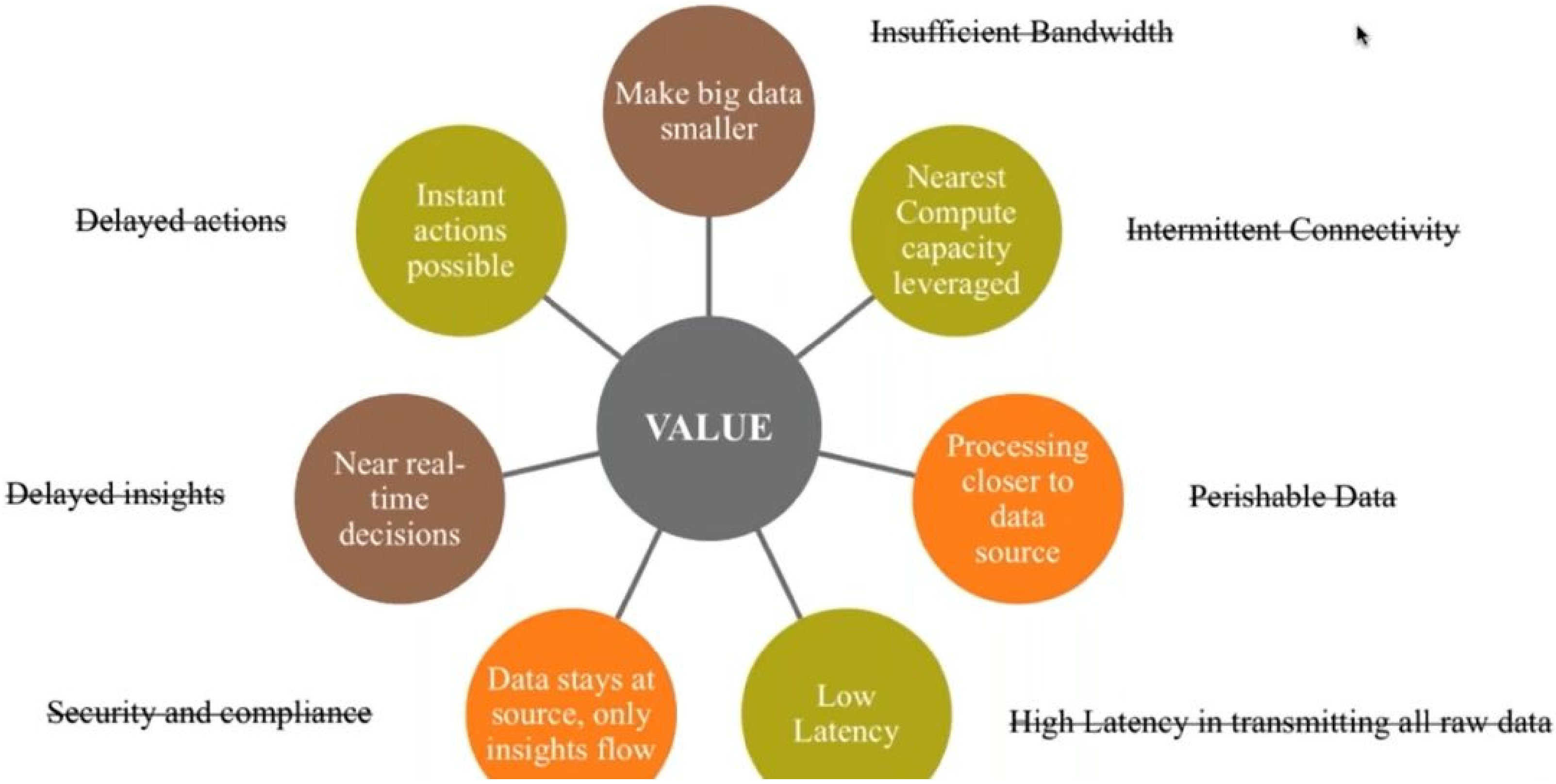



Cloud Layer
Long Term Analytics
Visualization & Reporting
Query Engine
Multi-tenancy & Resource Pooling
Rapid Elasticity & Scalability
Remote Mgmt & Configuration
Data Filtering/ Cleaning
Data Aggregation
Real Time Insight Analytics
Cognitive Intelligence & Analysis
Edge Computing Layer
(Gateway)
Edge Modules

Edge Modules
Sensors/Actuators
Device Layer

- Availability of higher computing power and storage for lesser cost
- Cost of development and form factor of the embedded devices continue to drop.
- Ever growing volume of data from the world around us
- Availability of advanced machine learning and data analytics techniques

Advantages of de-centralized approach



Capabilities needed at the edge

- Compactness
- Ruggedness for out-door, harsh environmental conditions (shock, temperature ..)
- Security
- Remote administration, monitoring and control
  - Software updates
- Modularity for easy subsystem replacement
  - Device upgrades

The HARDWARE ENABLERS OF EDGE COMPUTING

Processors & Accelerators

Micro Data Centers

Converged Systems
Compute + Storage + Network

Advanced SoCs powerful enough to run full-fledged OS + complex algorithms

5G, SDN, NFV

Complete Edge Software Stack with
Analytics / Machine Learning Libraries

PROCESSORS & System on chips

E5 – 8 cores, **3.8 GHz, 135 W**

E3 – 4 cores, 3.5 GHz, 80 W

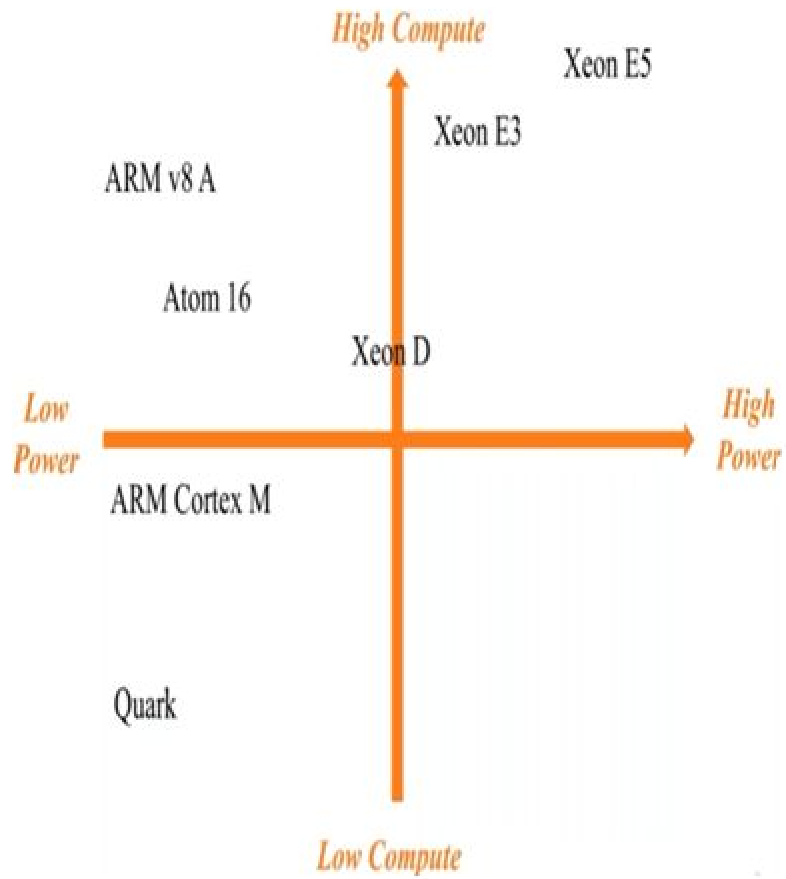
D – 16 cores, 1.3 GHz, 45 W

1-8 cores, 1.9 GHz, 2 - 20 W

32 - 400 **MHz, 0.025 W** to 2.2 W

M – 1 to 2 cores, 400 - 800Hz, 2+ W

A V8 – 4 cores, **2.5 GHz, 2+ W**



ACCELERATORS

FPGA

- Reprogrammable for new problems of ever-shifting technologies and business models
- Up to 1500 GFLOPs still less energy-hungry
- Used in Intel's driver-less cars platform
- Neural network execution for Bing and in Azure

Advantages of FPGA over GPUs

(1) provides a consistent throughput, invariant to the size of application workload

(2) offer both spatial and temporal parallelism

(3) FPGAs feature 3 - 4 times lower power consumption and up to 30 times better energy efficiency

Some use cases

ASIC (Application Specific Integrated Circuits)

- **Google's Tensor Processing Unit (TPU)**

- 15 to 30 times faster neural networks in less power
- Saved Google cost of 15 new data centres
- Costly, permanent in nature, not reprogrammable
- Google, Facebook need AI chip on personal devices to run neural networks

Converged Infrastructure for Edge Devices

- Compute, storage, network and virtualization components grouped together in a package
- Compact in size
- Rugged for harsh edge conditions
- Optimized for low energy consumption
- Engineered to optimize performance and cost
- Ready to commission

Converged Servers for Gateway

- Intel Xeon D / E3 @ 1.3 to 3.5 GHz
- 4 – 32 cores
- 64 – 512 GB RAM
- Workstation – class GPU
- IU form factor
- Upto 4 TB storage
- 10 – 40 Gbps network
- 45 – 100 W

Compute for Things-to-Cloud

Microcontrollers	SoC	Gateway	Server	Micro Data Center
				
Small size, low power low cost High performance apps like DSP (Digital Signal Processing), sensor fusion, motor control	Server class SoC with ARMv8 64 bit 54 cores 3.0 GHz Upto 1 TB memory 100 Gbps I/O bandwidth 10 to 100 GbE Integrated hardware accelerators for security, storage, networking and virtualization	Intel Atom / i5 dual-core 1.46 / 1.9 GHz On-board GPU 4 - 8 GB RAM 32 - 64 GB SSD 1U / 2U form factor Rugged for harsh edge	Intel E5-26xx V4 family 64 to 176 cores 32GB to 2TB RAM 9.6 - 460 TB storage 12M+ IOPS, 60GB/s transfer rate 2U (H 3.5” x W 17.25” x L 36.5”) Rugged for harsh edge	10s of processors in a single 19in 15U to 30U rack Rapidly deployable indoor, outdoor Designed to withstand in rugged, high risk zones

Data Centers @ Edge

Micro data center

Standalone rack-level systems

Contain all of 'traditional' data center in one box

Designed for specific set of workloads

Modular data center

A modular data center connected to the power grid at a utility substation

Modules can be shipped to be added, integrated or retrofitted as required

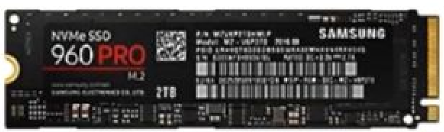
Portable data center

Portable modular datacenter

20 to 40 feet container size

Complete IT infrastructure in a shipping container.

Storage



1 TB SSD
M.2 U form factor

Intel 'Ruler' SSD
1 PetaByte Storage
1U server rack

IBM Tape
330 TB
Palm size

Transfer Rates

HDD – 120-150 MB/s
SSD – 500-600 MB/s

Interface Comparison

SATA

150 – 600 MB/s
Queues : 1
Q Depth: 32

SAS

750 MB/s
Queues : 1
Q Depth: 254

PCI EXPRESS
nvm EXPRESS

1 to 3GB/s
CPU cycles: 50+% less
Queues : 65000
Q Depth: 65000

Software @ Edge

IBM Watson IoT.

AWS Greengrass

Azure IoT Edge

PREDIX

Apache EDGENT

EDGE X FOUNDRY™

OpenWhisk

Fog IOx

thingworx®

OpenFog®

Cloud Stack


Business specific IoT applications
Microservices
OpenWhisk
Spark
InfluxDB
Rasin.io
Kubernetes
Master
thingworx
openstack
intel
Xeon
inside
GPU
FPGA
ASIC

Edge Stack

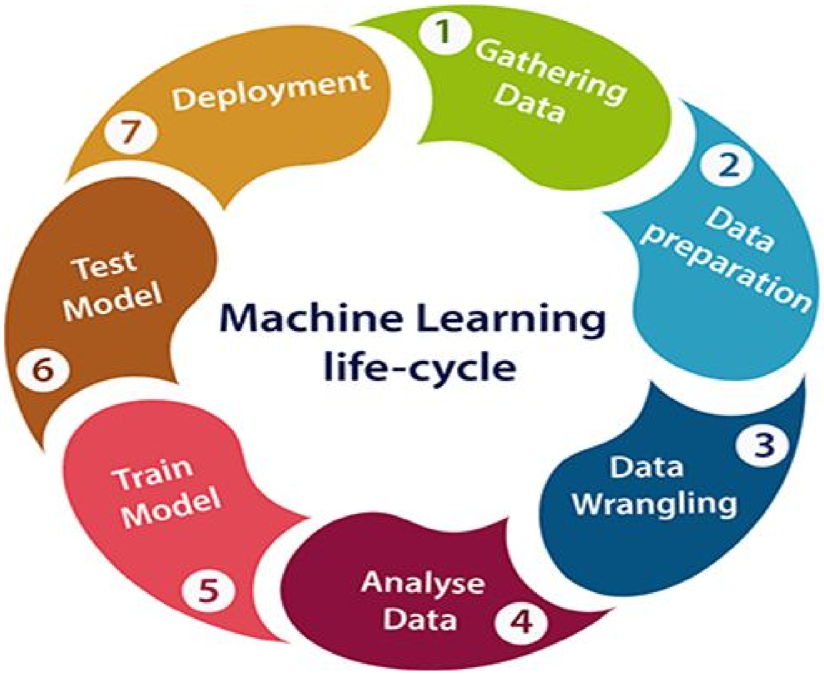

Business specific IoT applications
Microservices
OpenWhisk
Rasin.io
Kubernetes Node
InfluxDB
thingworx
SYSTEM STUDIO
mbed SDK
sdk
snappy
IoT
IOx
Raspbian
ARM mbed
intel Xeon
intel Atom
arm
ARTIK

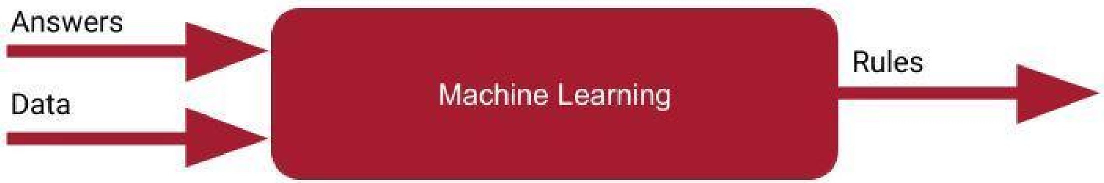
Things Stack

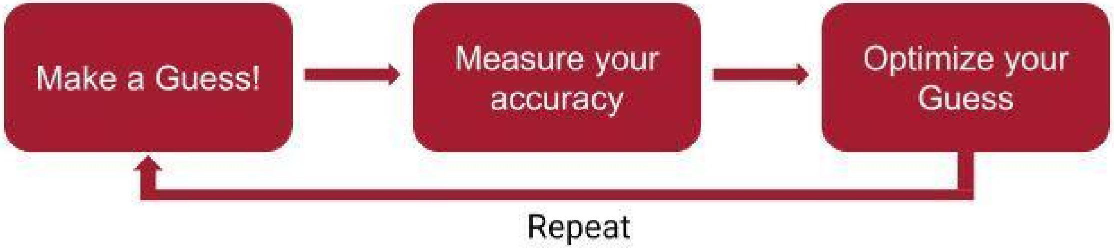
Challenges at Edge

- Replication of data – when, what, how
- Security at resource constrained edge devices
- Storage limitations
- Creating analytical model in one place and executing it in multiple places
- Creation and exchange of ML model within nodes
- Complexity Management
- Peer-to-peer communication requires mature standards
- Dependent in progress in communication network and related infrastructure

An overview of Tiny ML

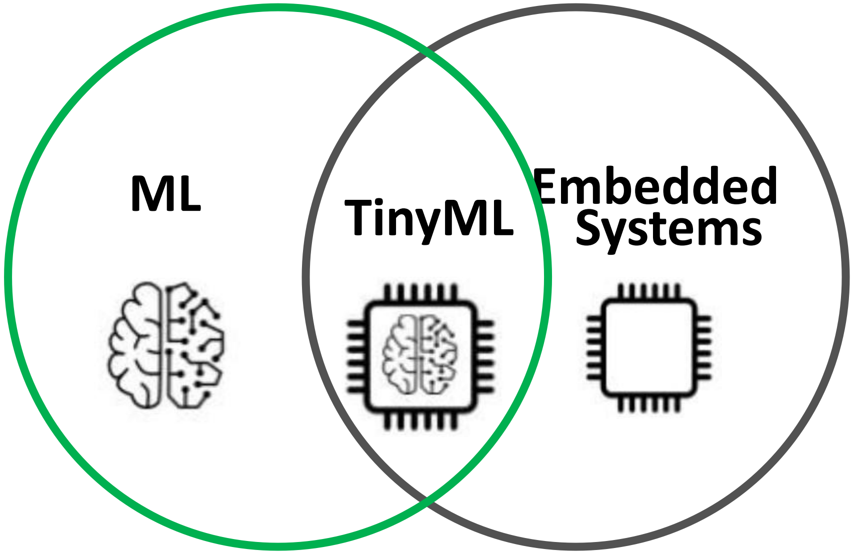

1 Gathering Data
2 Data preparation
3 Data Wrangling
4 Analyse Data
5 Train Model
6 Test Model
7 Deployment
Machine Learning life-cycle

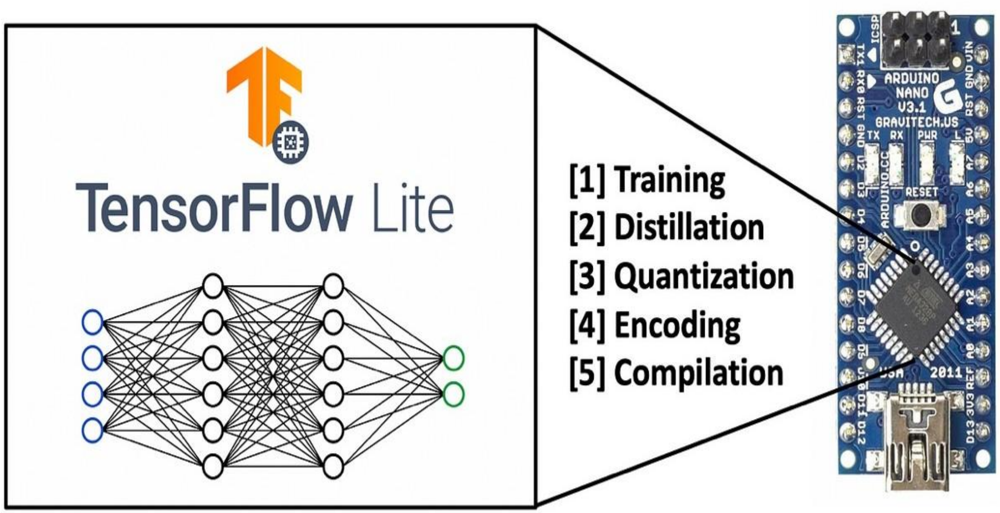

Answers
Data
Machine Learning
Rules


Make a Guess!
Measure your accuracy
Optimize your Guess
Repeat

Foundations and Applications of TinyML

- **TinyML:** Emerging area where **ultra large powerful ML models are converted into executables for embedded systems** that are battery operated and mostly well beyond the operation capacity of the smart phones (e.g., microcontrollers)

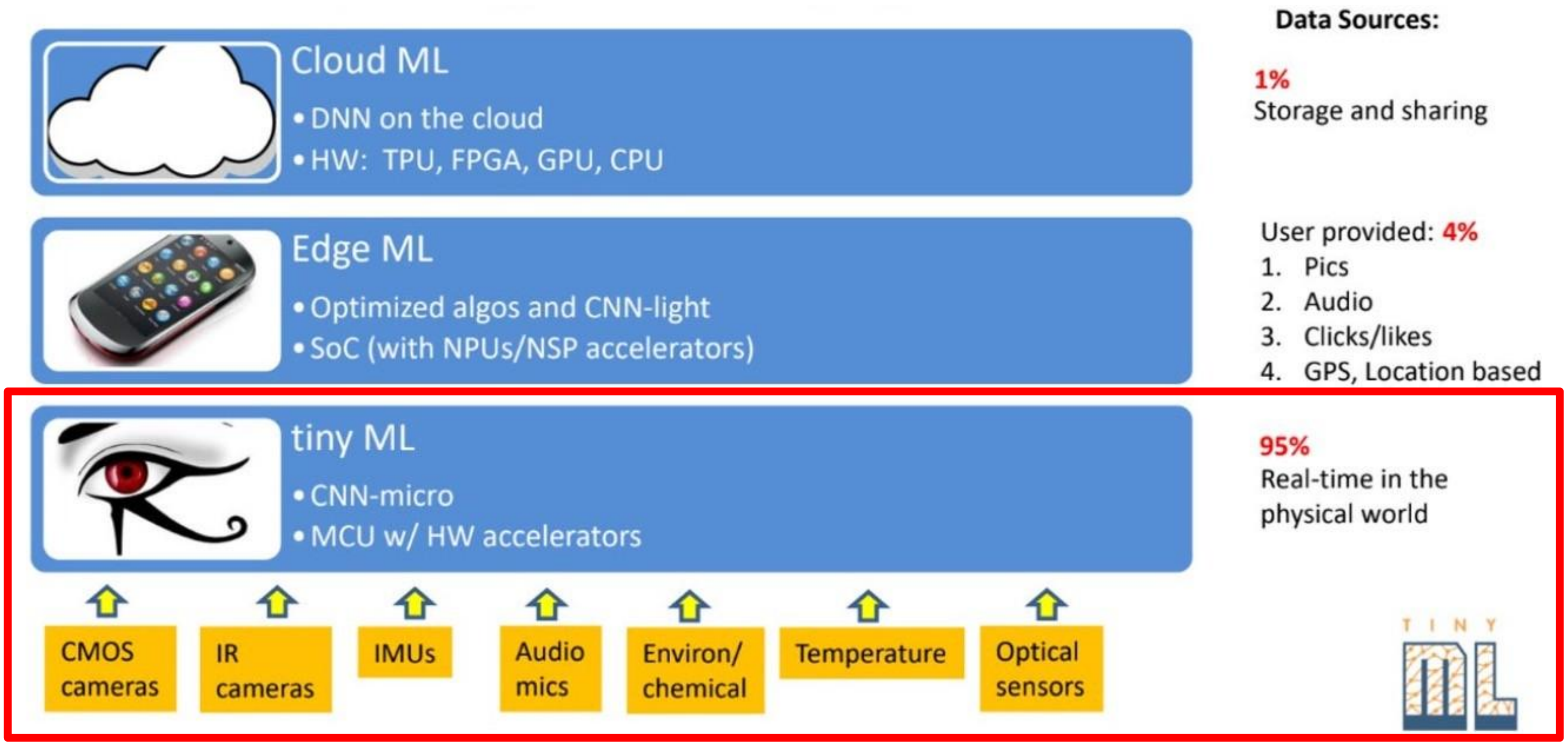




Source: https://towardsdatascience.com/tiny-machine-learning-the-next-ai-revolution-495c26463868

Foundations and Applications of TinyML

- TinyML is **real-time processing of time-series data that comes directly from sensors**



Source: https://www.tinyml.org/about/

Foundations and Applications of TinyML

- TinyML has **applications in agriculture, health, retail, energy industry**, and more…

Plant disease classification with TensorFlow Lite on Android

Source: https://yannicksergeobam.medium.com/plant-disease-classification-with-tensorflow-lite-on-android-part-2-c2d47371cea3

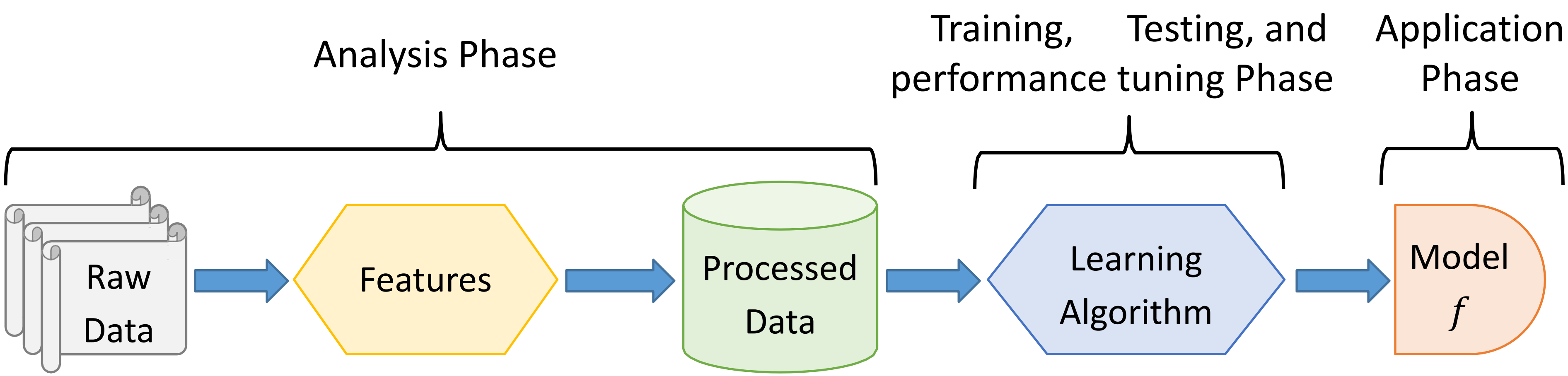
Solar Scare Mosquito: A solar-operated device that sits on stagnant water to create air bubbles at regular intervals to avoid the breeding of mosquitoes

Source: https://theindexproject.org/award/nominees/6558

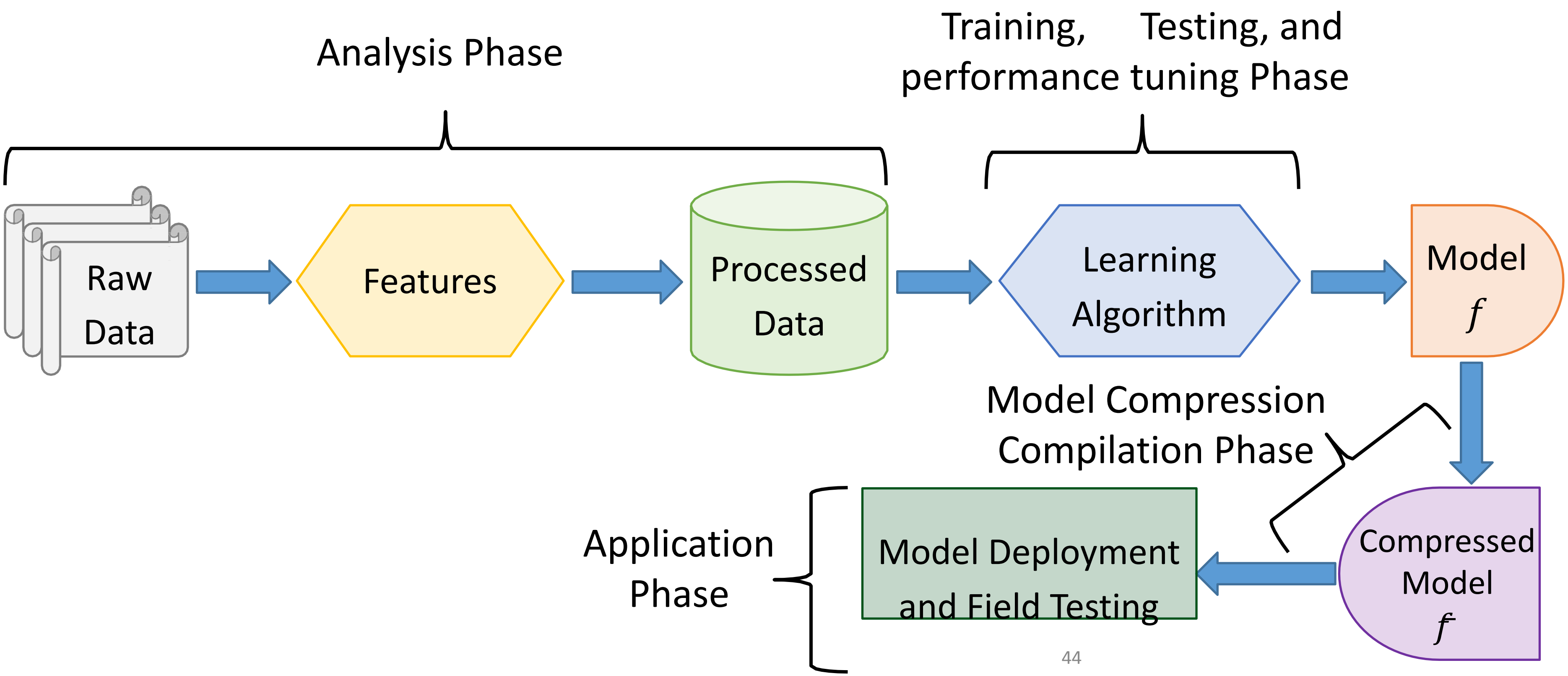
TinyML for keeping an eye on the inventory of goods on the shelf in retail establishments and sending out warnings when it runs low

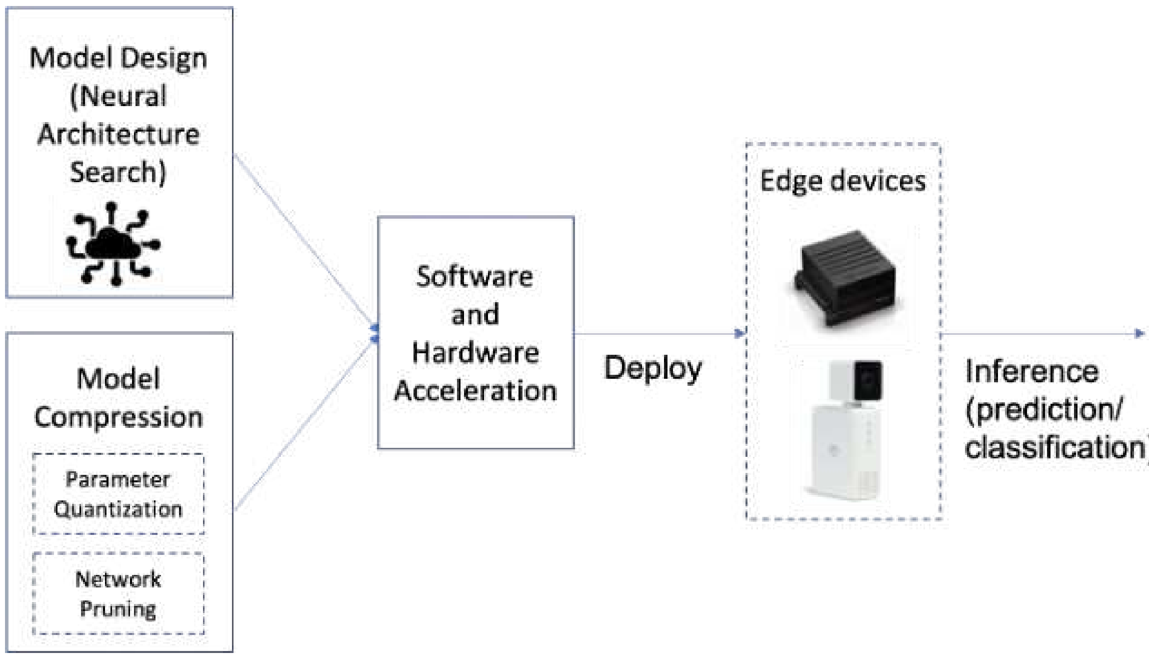
Source: https://www.supermarketnews.com/store-design-construction/amazon-go-goes-smaller

A Schematic View of ML and Its Phases

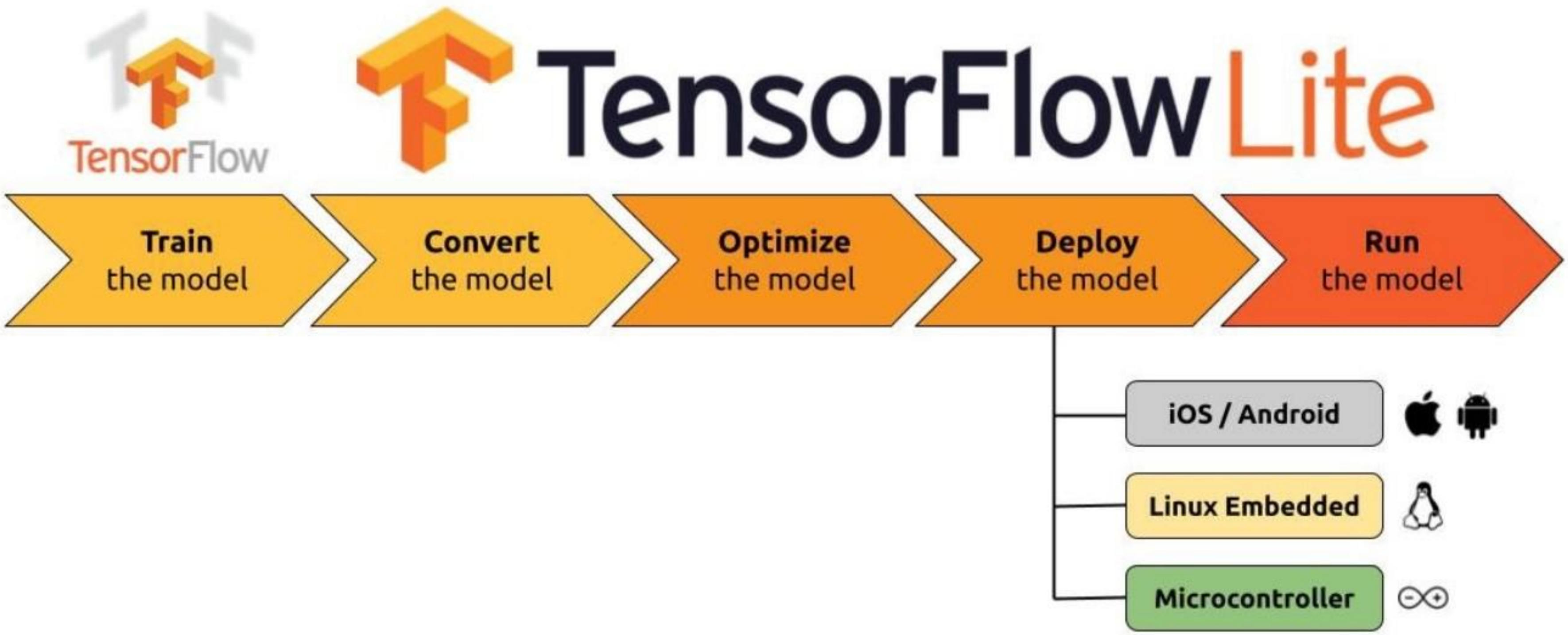


A Schematic View of TinyML and Its Phases

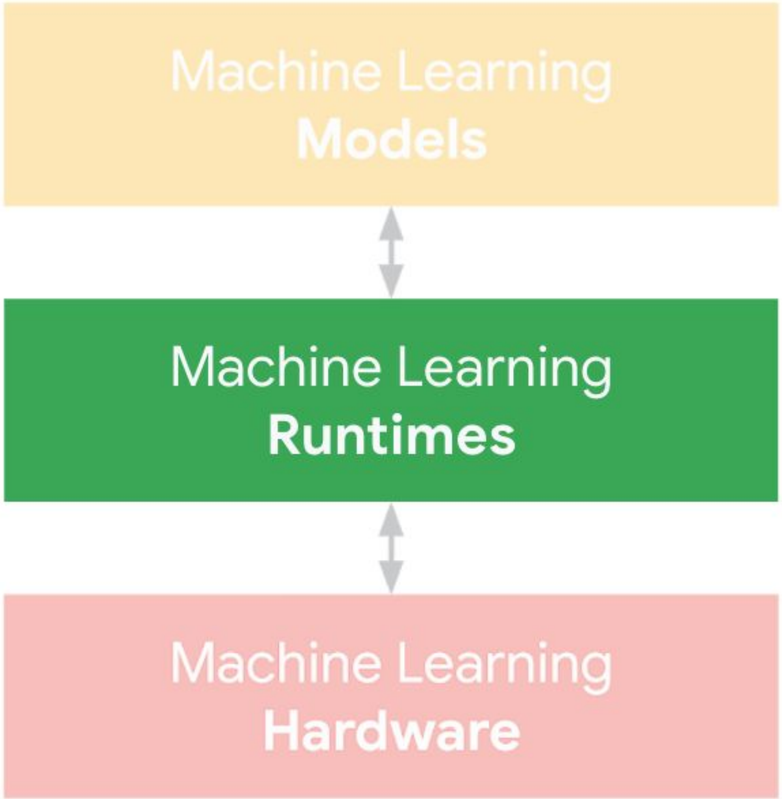


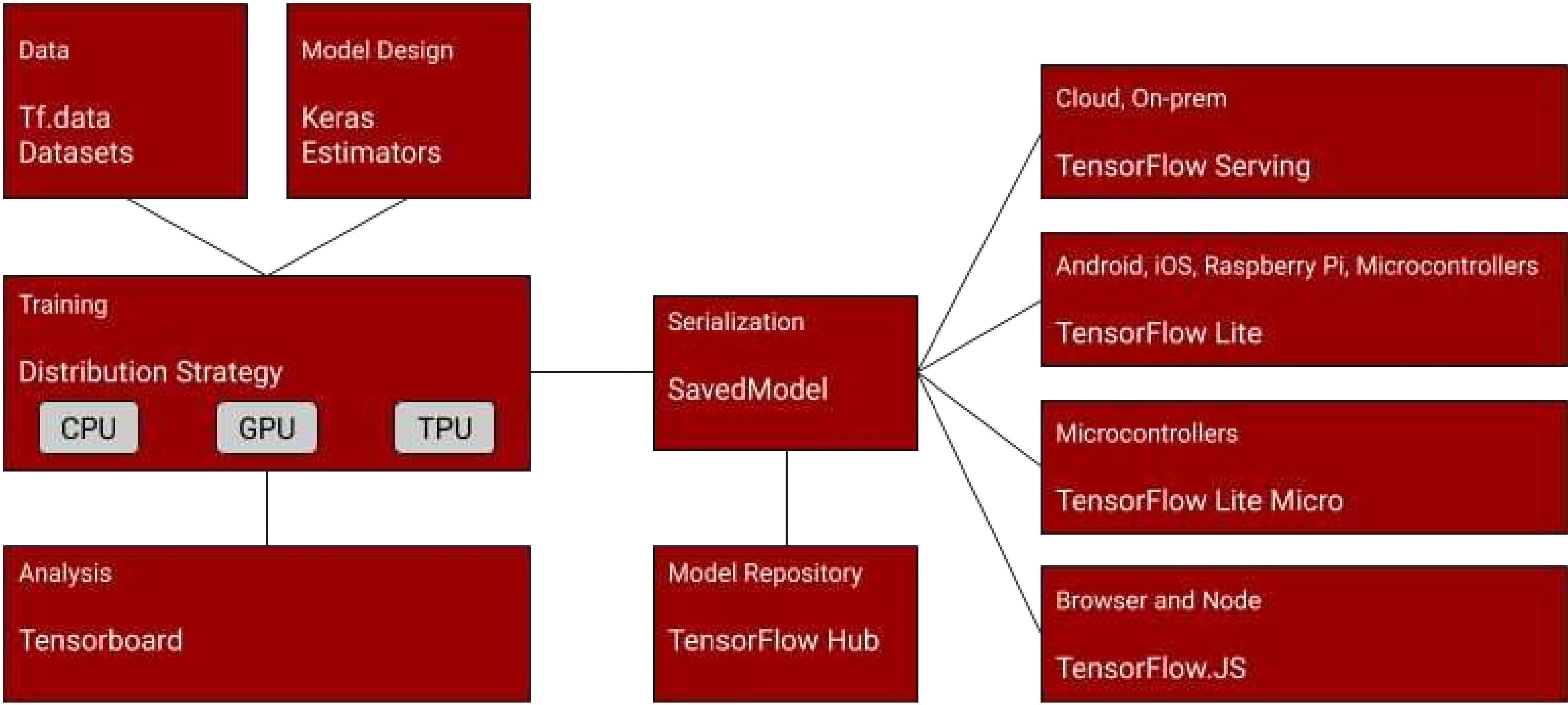

Model Design (Neural Architecture Search)
Model Compression
Parameter Quantization
Network Pruning
Software and Hardware Acceleration
Deploy
Edge devices
Inference (prediction/ classification)

TensorFlow (TF) and TFLite Workflow for TinyML



Source: https://leonardocavagnis.medium.com/tinyml-machine-learning-for-embedded-system-part-i-92a34529e899


Machine Learning
Models
Machine Learning
Runtimes
Machine Learning
Hardware


Data
Tf.data
Datasets
Model Design
Keras
Estimators
Training
Distribution Strategy
CPU
GPU
TPU
Analysis
Tensorboard
Serialization
SavedModel
Model Repository
TensorFlow Hub
Cloud, On-prem
TensorFlow Serving
Android, iOS, Raspberry Pi, Microcontrollers
TensorFlow Lite
Microcontrollers
TensorFlow Lite Micro
Browser and Node
TensorFlow.JS