جامعة اليرموك
YARMOUK UNIVERSITY

KEYNOTE LECTURE

Resource-Efficient Intelligence

Where Edge AI, TinyML, and Frugal AI Meet Sustainable Innovation

Prof. Dr.
Khalid M.O. Nahar

Professor of
Artificial Intelligence
and Machine Learning

Yarmouk University
Faculty of IT



Bringing Intelligence to Sensors

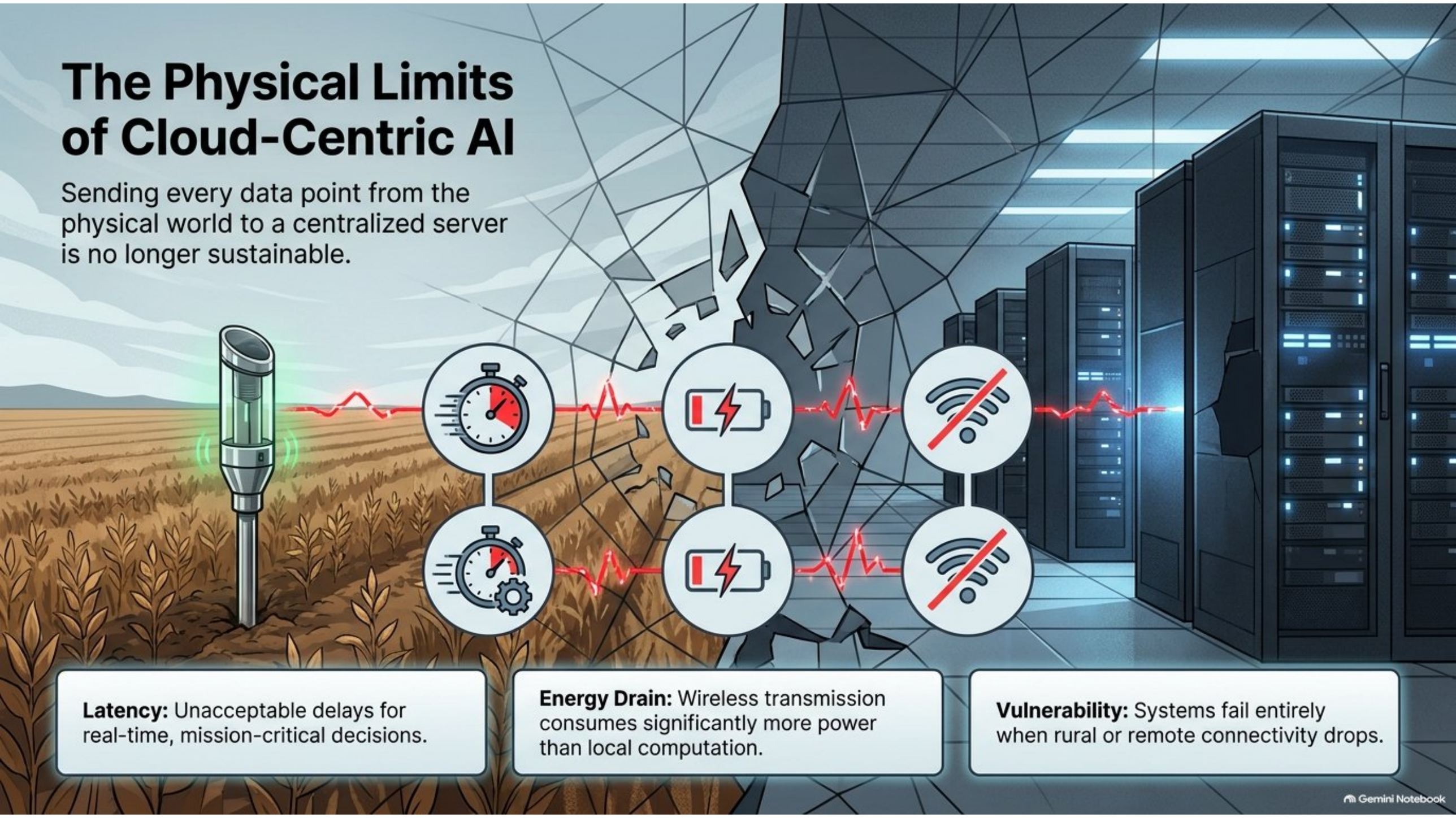
The shift from cloud-centric AI to frugal, edge-based TinyML ecosystems.



Gemini Notebook

The Physical Limits of Cloud-Centric AI

Sending every data point from the physical world to a centralized server is no longer sustainable.



Latency: Unacceptable delays for real-time, mission-critical decisions.

Energy Drain: Wireless transmission consumes significantly more power than local computation.

Vulnerability: Systems fail entirely when rural or remote connectivity drops.

Gemini Notebook

The Sensor That Understands

The Old Paradigm: Dumb Sensor



- Records audio.
- Requires Wi-Fi.
- Streams gigabytes of canopy noise.
- Dead battery in 3 days.

The Breakthrough: Intelligent Node

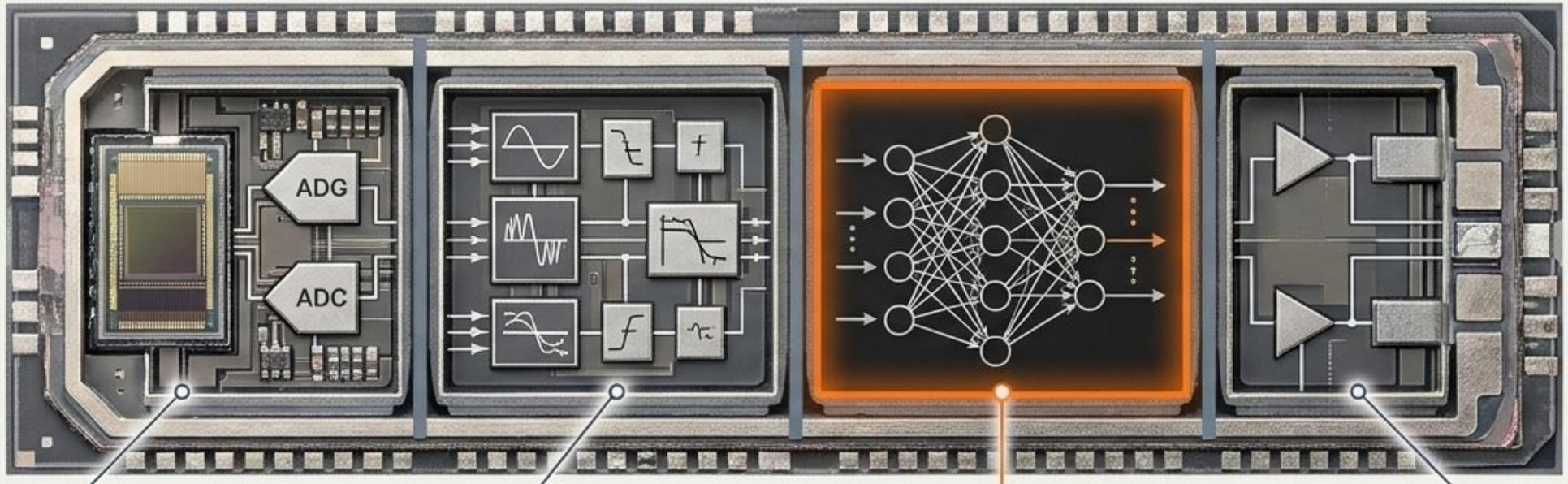


- Listens offline.
- Powered by a palm-sized solar panel.
- Runs a neural network in 20 kilobytes.
- Detects chainsaw acoustic signatures 3km away and transmits a 1-byte alert.

We are moving from a world of sensors that only measure, to a world of sensors that perceive.

Gemini Notebook

Anatomy of an Intelligent Sensor



1. Sense

Front-end hardware converts physical reality (audio, vibration, heat) into raw electrical signals.

2. Pre-Process (DSP)

Digital Signal Processing cleans the signal efficiently (e.g., turning audio into a spectrogram) before AI is even invoked.

3. Infer

A compact neural network (often <100KB) runs locally to classify the pattern or detect anomalies.

4. Act / Report

The chip triggers a local action (stop motor) or transmits a tiny, distilled message.

TinyML is not just AI—it is an engineering discipline fusing signal processing with machine learning.

Gemini Notebook

What is Edge AI?

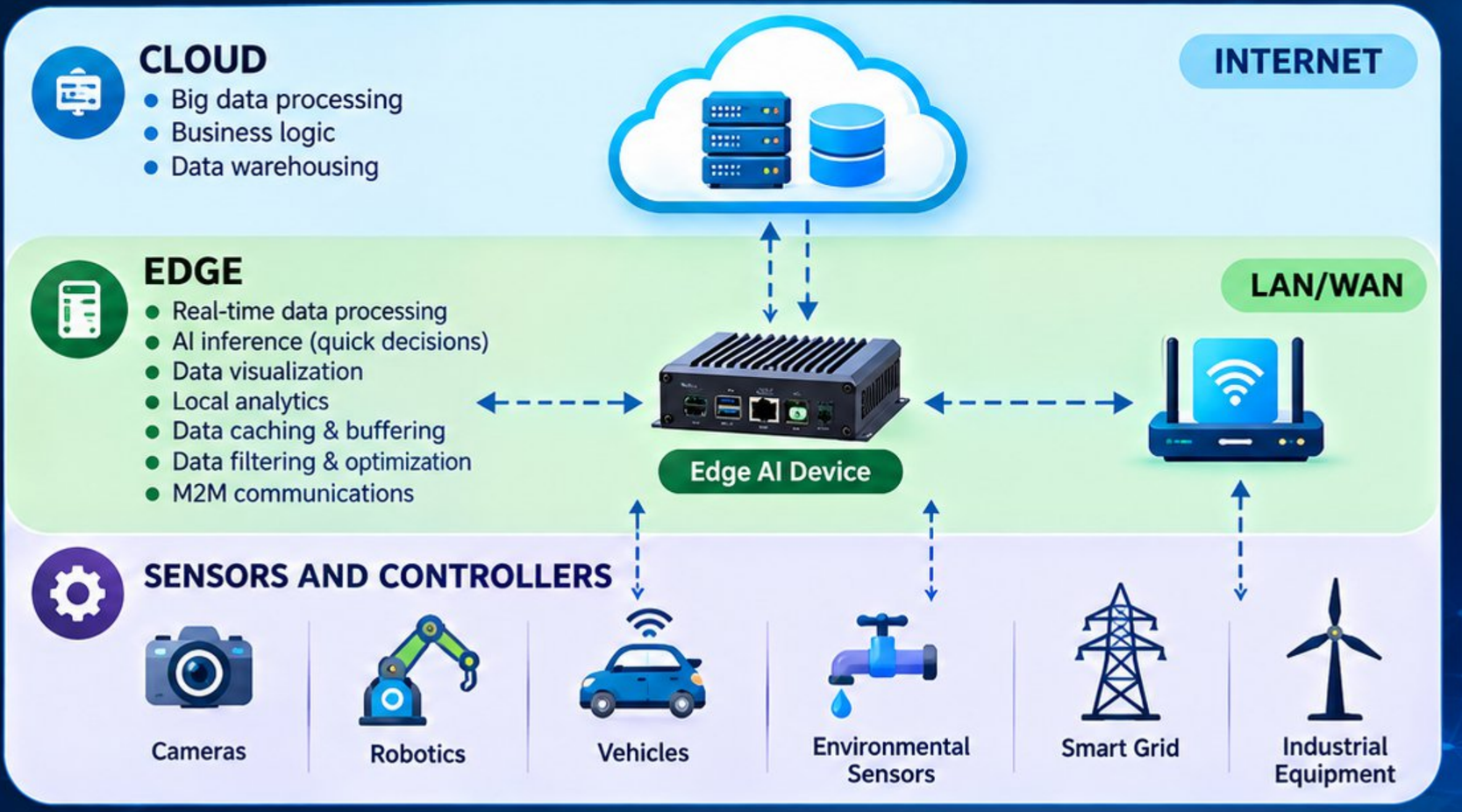
Edge AI refers to running artificial intelligence models locally on a device (the “edge”) rather than sending data to the cloud.

Smart Cameras
Real-time vision and object detection

Autonomous Drones
Onboard decision making and navigation

IoT Sensors
Local data analysis and alerts

Wearables
Personalized health and activity monitoring



Closing the Distance to the Physical World











Maximum Distance

Centralized (1980s)

Mainframes.

Cloud Computing (2000s)

Remote massive data centers.

Mobile Computing (2010s)

Intelligence in user pockets.

Edge AI (Present)

Local gateways, cameras, vehicles.

TinyML (The Frontier)

Intelligence inside the sensor itself.

Zero Distance

Intelligence at the point of origin. Zero distance between the event and the decision.

Gemini Notebook

What is TinyML?

TinyML is a subset of Edge AI focused on running machine-learning models on very small, low-power microcontrollers (MCUs), such as **ARM Cortex-M** boards.

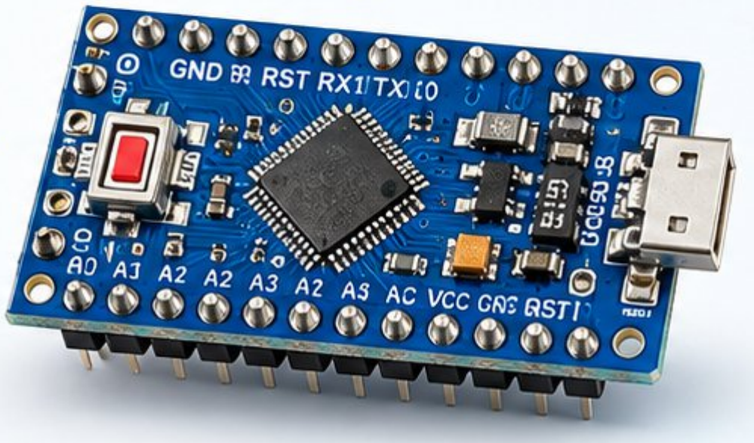
Extremely Small — Kilobytes instead of megabytes.

Low Memory — Typically 256 KB RAM or less.

Low Power — Can run for years on a coin battery.

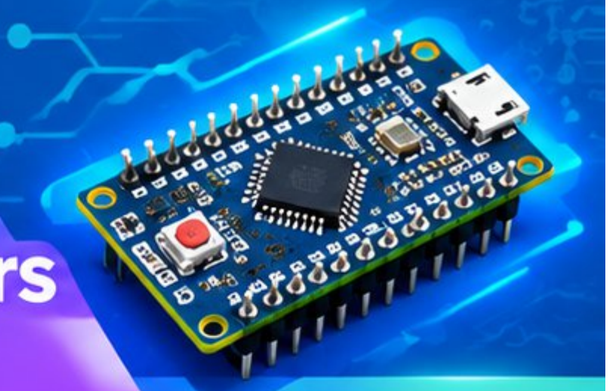
Real-time Processing — Instant responses at the edge.

TinyML:
The Future of Machine Learning on Edge Devices



TinyML =

Machine Learning on Microcontrollers with <1MB RAM



Ultra-Lightweight

Models:
10–500KB

Power Efficient

Inference:
1–5mW

Real-Time

Latency:
<10ms

TinyML Hardware Examples

Small boards... Big intelligence!



Arduino Nano 33

Processor
ARM Cortex-M4

Clock Speed / Memory
64 MHz, 256KB RAM

Price Range
$25–35

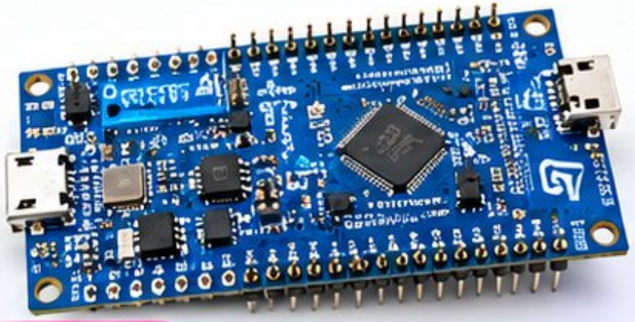


ESP32-S3

Processor
Dual Xtensa LX7

Clock Speed / Memory
240 MHz, 512KB RAM

Price Range
$5–10



STM32F7

Processor
ARM Cortex-M7

Clock Speed / Memory
216 MHz, 512KB RAM

Price Range
$8–15

The Definitive Spectrum of Compute

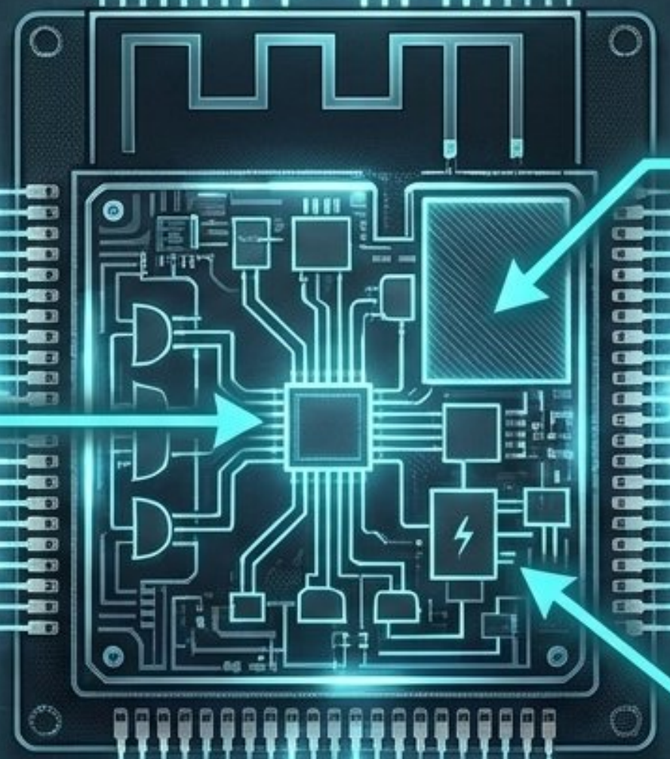
Compute Tier	Power Budget	Memory	Network Dependency	Core Function	Hardware Cost
Cloud AI	Megawatts	Terabytes	100% Dependent	Massive global training & analytics	Millions ($)
Edge AI	Watts to Tens of Watts	Gigabytes	Intermittent	Real-time local inference (gateways, phones)	$10s to $1,000s
TinyML	Milliwatts / Microwatts	Kilobytes	Zero	Always-on embedded sensing	Cents to $5

Billions of microcontrollers ship annually. TinyML awakens the silent silicon already embedded in our world.

Gemini Notebook

The Microcontroller Reality

Embedding intelligence requires operating an AI brain on less power than an LED light bulb.



Flash (Storage)

Typically under 2 MB. Insufficient for standard neural network weights.

SRAM (Working Memory)

Often limited to 256 KB. Cannot hold traditional deep learning activation buffers.

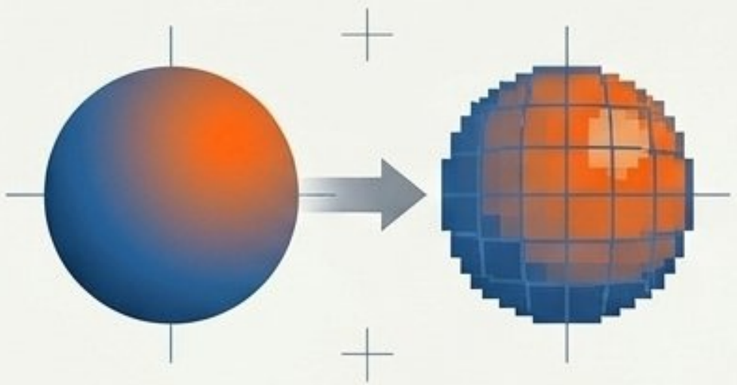
Power Supply

Runs on coin-cell batteries or intermittent solar. Processing must finish in milliseconds to return to deep sleep.

Gemini Notebook

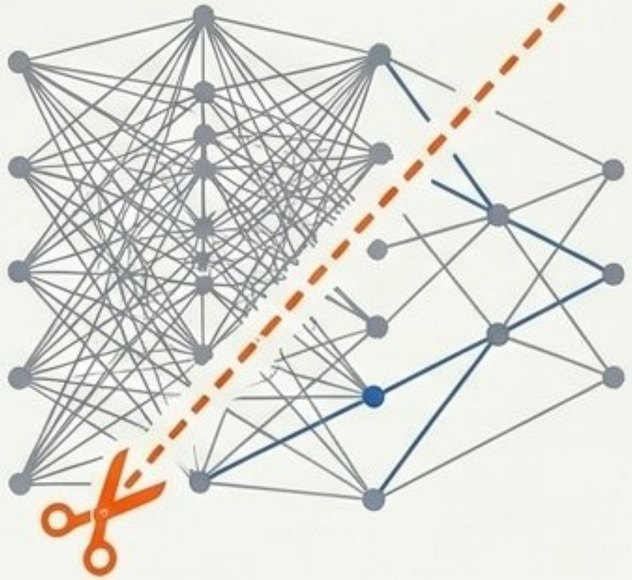
Shrinking the Brain: The Engineering Puzzle

1. Quantization (Resolution reduction)



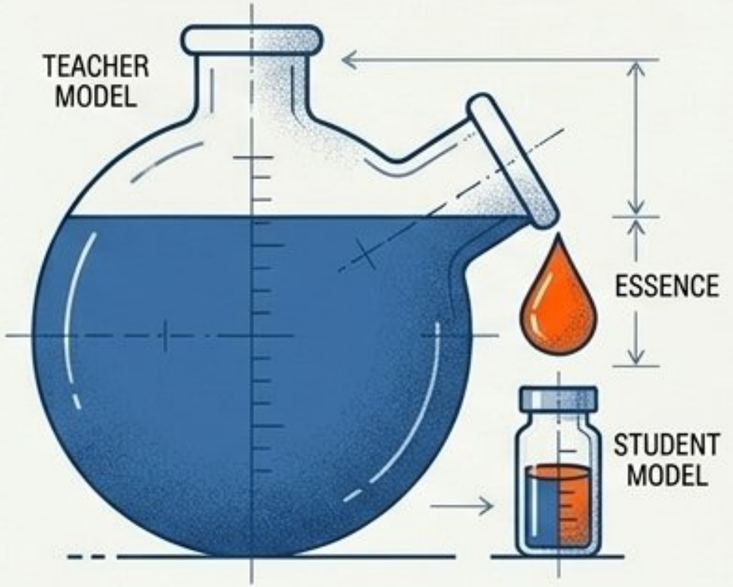
Converting 32-bit floating-point weights to 8-bit integers. Shrinks the model 4x with less than 1% accuracy loss.

2. Pruning (Structural reduction)



Deleting up to 90% of the network's unnecessary connections. The essential architecture survives.

3. Knowledge Distillation (Essence transfer)



A massive cloud 'teacher' model trains a tiny 'student' model to imitate its reasoning, fitting supercomputer intelligence into 40 kilobytes.

Gemini Notebook

From model to embedded device

A four-step path from training to on-device inference.

TRAIN → RUN

01 TRAINING

Train the model

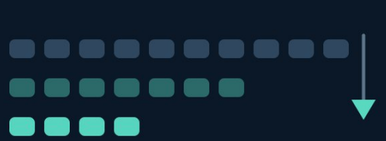


Train on a workstation or in the cloud using TensorFlow or PyTorch.

- TENSORFLOW
- PYTORCH

02 OPTIMIZATION

Optimize the model

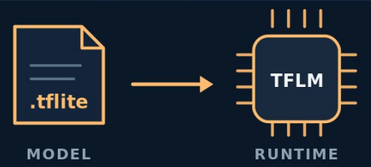


Quantize, prune or compress to reduce the model's memory and compute footprint.

- SMALLER FOOTPRINT

03 CONVERSION

Convert for the edge

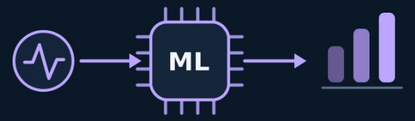


Export a .tflite model. For microcontrollers, use the TFLite Micro runtime (TFLM).

- TFLITE MODEL
- TFLM RUNTIME

04 DEPLOYMENT

Deploy & infer



Load the model and runtime on the target device. Inference then runs locally.

- MOBILE
- LINUX EDGE
- MCU

ON-DEVICE INFERENCE — Sensor data in. Predictions out. The model runs locally on the device.

Small models, capable devices

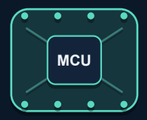
Microcontrollers and edge accelerators bring inference close to sensors and users.

HARDWARE → INFERENCE

01 HARDWARE TARGETS

Embedded boards & accelerators

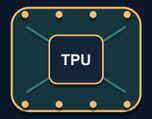
Choose a platform that fits the model and resource budget.



MICROCONTROLLERS

Compact, low-power boards

- Arduino Nano 33 BLE Sense
- ESP32
- STM32
- Raspberry Pi Pico



EDGE ACCELERATOR

Google Coral

Edge TPU

Match the model format and runtime to the target hardware.

02 ON-DEVICE INFERENCE

Tasks handled locally

The device processes inputs and returns predictions on the spot.

Wake-word detection

Recognize “Hey device.”

Gesture recognition

Read movement from a camera or motion sensor.

Predictive maintenance

Spot unusual motor vibration.

Low-power vision

Classify images on compact cameras.

TYPICAL INPUTS: AUDIO • MOVEMENT • VIBRATION • IMAGES

WHY ON-DEVICE?

Local inference can reduce round-trips and keep raw inputs on the device.

Real-time decisions, right on device

On-device inference can respond quickly, work offline, and keep sensitive inputs local.

LOCAL → RESPONSIVE

01 OFFLINE-READY



Works offline

Run the deployed model without an internet connection.

02 LOW LATENCY



Faster decisions

No network round-trip; local responses often take milliseconds.

03 DATA PRIVACY



Keep inputs local

Raw sensor data can stay on-device, rather than being sent to a server.

04 LOW POWER



Battery-aware ML

TinyML is designed for small compute and power budgets.

INFERENCE SPECTRUM

Choose how close inference runs to the sensor

CENTRALIZED → ON DEVICE



CLOUD

Data centers • largest models

Centralized compute • global scale



EDGE

Nearby gateway or device

Smaller models • regional processing



TinyML

Microcontroller • compact model

On-device inference • ultra-low power



PRACTICAL NOTE

Actual latency and battery life depend on the model, hardware, and workload.


Simulate All

Case Study: MobileNetV2 in the Orchard

Detecting Apple and Mango leaf diseases using a pipeline optimized for low-power edge deployment.

Step 1: Input & Augmentation

High-res field images undergo random rotations and shifts to ensure model robustness against varying field conditions.

Step 2: Feature Extraction

Utilizing the pre-trained MobileNetV2 base (unfrozen for fine-tuning) for highly efficient spatial mapping.

Step 3: TinyML Optimization

Global Average Pooling reduces dimensions -> 128-unit Dense Layer (ReLU) -> 50% Dropout to prevent overfitting -> Softmax classification.

Step 4: Deployment

Converted to TensorFlow Lite for ultra-low power execution on the sensor node.

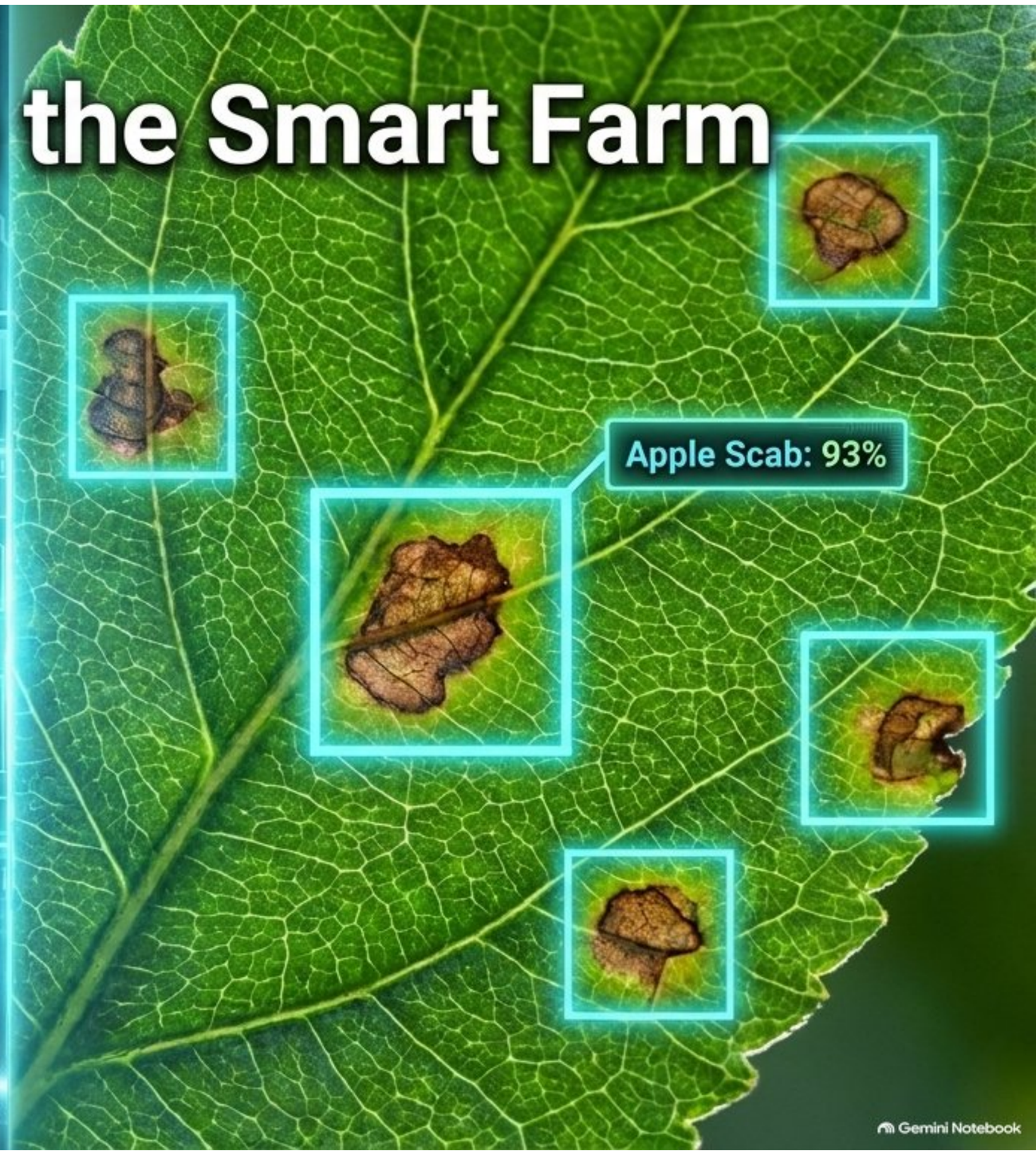
Gemini Notebook

Ground Truth in the Smart Farm

In precision agriculture, immediate disease detection prevents widespread crop failure. However, smallholder farms and remote orchards lack the bandwidth for cloud-based computer vision.

The TinyML Solution:

Moving inference directly to the camera sensor. No cloud connection required, no subscription fees, and real-time execution in the field.



Gemini Notebook

On-device vision for drones

Compact models can interpret camera frames locally, even with limited connectivity.

AERIAL PERCEPTION



01 AERIAL PERCEPTION

Where local vision helps

On-board models can assist field teams in:

Infrastructure inspection
Flag visible cracks or damage for human review.

Search & rescue support
Highlight possible people or hazards for response teams.

Environmental surveys
Map vegetation, shorelines, or wildlife from imagery.

- ON-DEVICE INFERENCE
- HUMAN-REVIEWED

ON-DEVICE BENEFIT

On-board analysis can reduce delay and the need to stream every frame.

IoV Attack Simulator

Exploring Security Threats in the Internet of Vehicles

Realistic Attacks | Smarter Detection | Safer Roads



Types of Attacks

Denial-of-Service (DoS)
Jamming or flooding the communication channel to disrupt vehicle connectivity.

Spoofing Attack
Impersonating a vehicle or sending fake messages to manipulate the system.

False Data Injection
Injecting malicious or incorrect data into the network to mislead vehicles and systems.

Potential Consequences

- Loss of communication
- Wrong decisions
- Accidents / Safety risks
- Service disruption
- Privacy breaches

1. ATTACK
The attacker sends malicious data or signals.

2. COMMUNICATION
The data travels through V2X / network channels.

3. IMPACT
Vehicles receive false data, causing incorrect behavior.

4. DETECTION & DEFENSE
AI/IDS analyzes traffic and identifies the attack.

5. SAFER IoV
Build resilient systems for connected vehicles.

SECURING THE INTERNET OF VEHICLES

Real-world applications

Compact models turn local signals into useful decisions across health, industry, and the home.

SENSE → INFER → ACT

01 HEALTHCARE

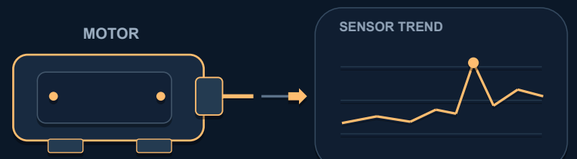


ECG anomaly detection

Flag unusual heart-rhythm patterns in ECG data for clinician review.

- ECG • HEART-RHYTHM SIGNALS

02 INDUSTRIAL



Predictive maintenance

Track vibration or temperature to spot emerging equipment problems.

- VIBRATION • TEMPERATURE

03 SMART HOME



On-device wake words

Recognize “Hey device” locally, without sending continuous audio to the cloud.

- LOCAL WAKE-WORD DETECTION

COMMON PATTERN

Sense locally → infer with a compact model → act where the data is captured.

The Next Hurdle: Scalability vs. Sustainability



TinyML solves the compute problem, but deploying tens of thousands of autonomous sensors introduces a new crisis. Static AI models will drain batteries blindly. How do we manage thousands of devices without creating a massive e-waste graveyard?

Gemini Notebook

Orchestrating the Edge

Computational Demand



Real-time Energy Availability

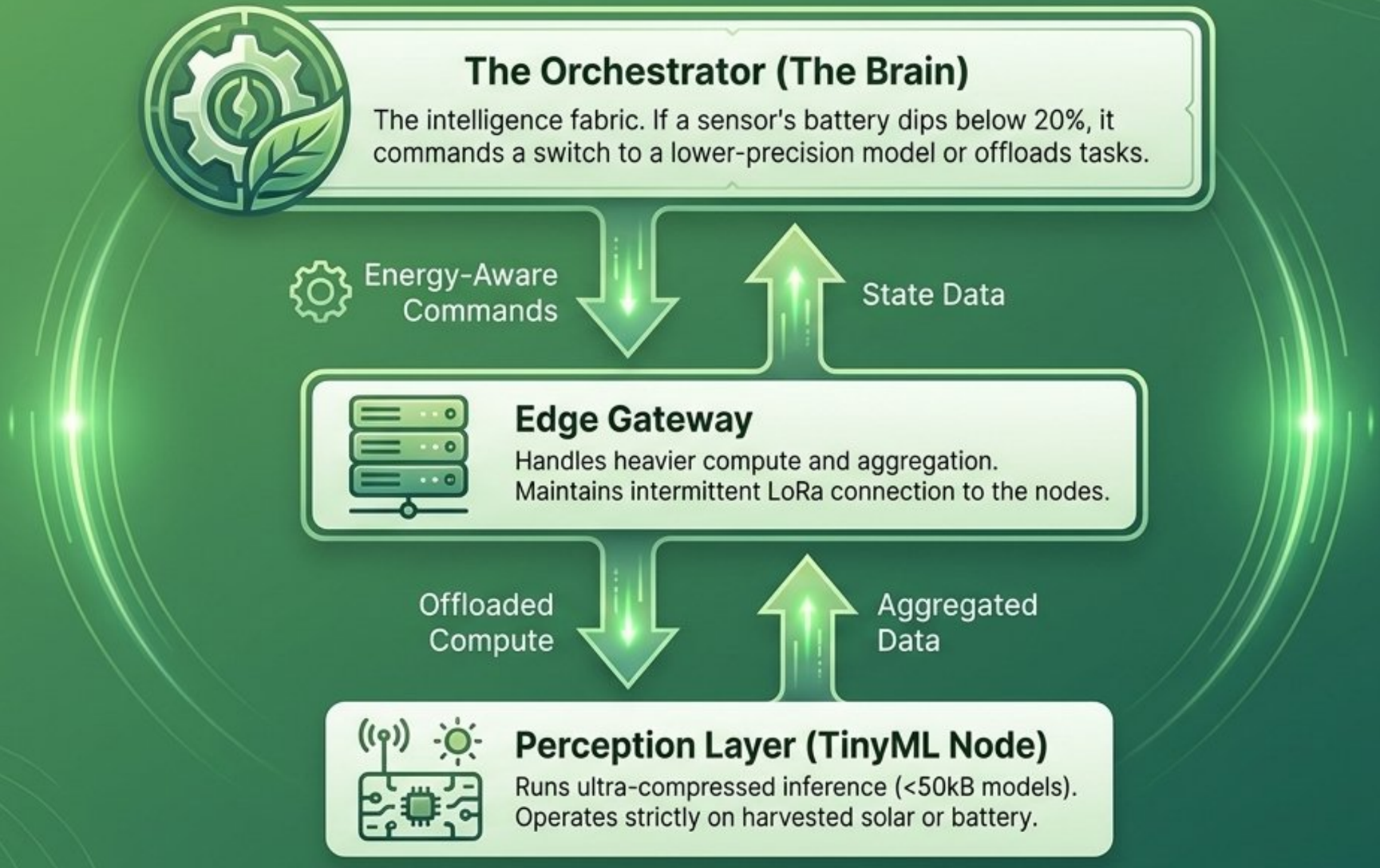
Enter **Green AI Orchestration** (e.g., the RAIDO framework).

AI cannot be static. Green orchestration introduces a dynamic layer that monitors energy budgets, sensor activity, and contextual needs in real-time.

It dictates when an edge node should run a complex model, when it should switch to an ultra-compressed fallback, or when it should go to sleep entirely based on solar availability.

Gemini Notebook

The Green AI Orchestration Architecture



Gemini Notebook

The Philosophy of Frugal AI

Moving beyond technical efficiency to prioritize environmental and social equity.



Efficiency

Achieving **maximum utility** with minimal hardware, memory, and energy.



Affordability

Utilizing **low-cost microcontrollers** ($1-$5) rather than high-end GPUs, eliminating financial barriers to entry.



Inclusivity

Designing AI that works **offline**, ensuring communities without high-bandwidth infrastructure can still benefit.

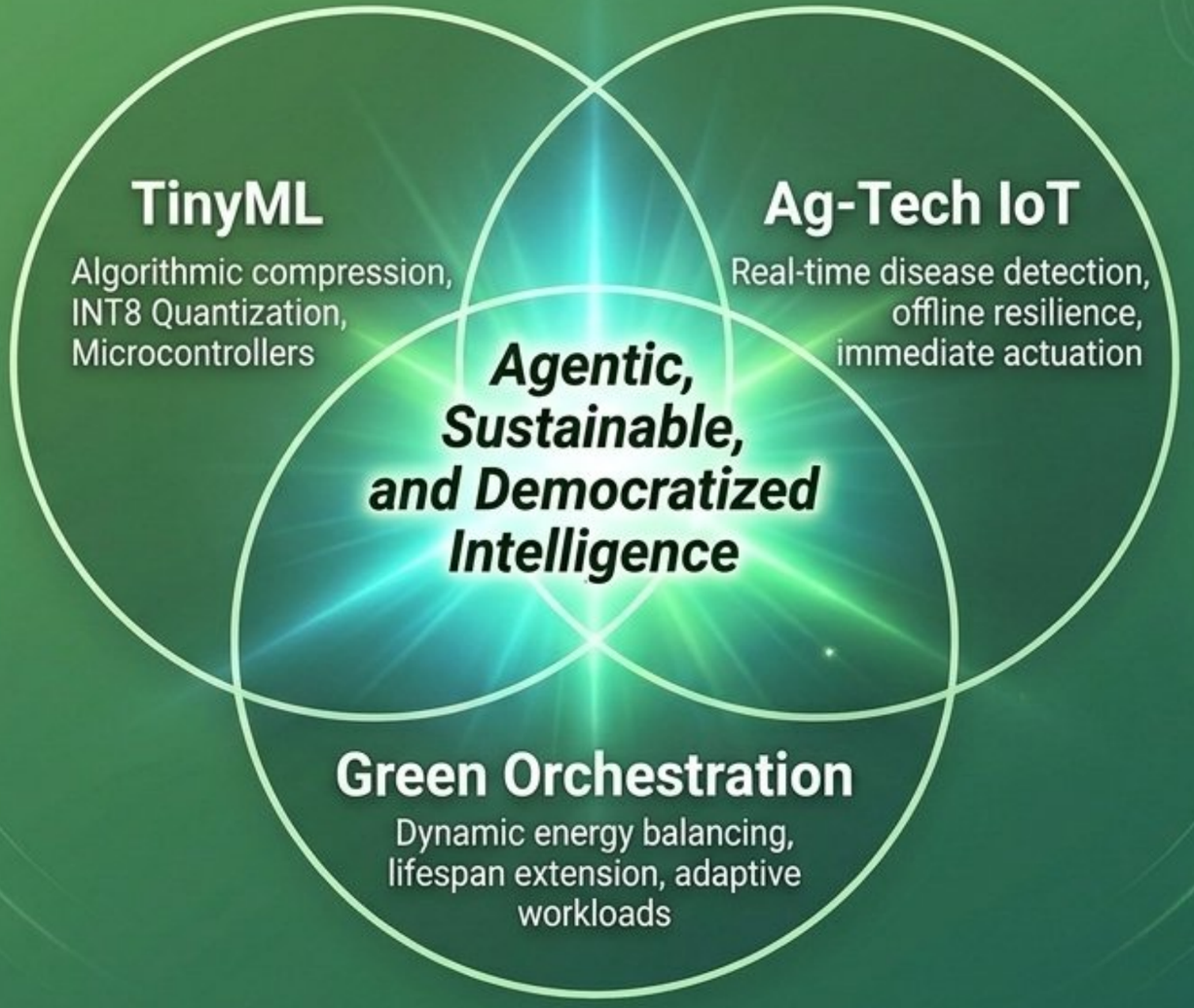


Environmental Responsibility

Extending **device lifecycles**, minimizing e-waste, and actively reducing the **carbon footprint** of AI inference.

Gemini Notebook

The Frugal Edge AI Paradigm



A structural redefinition of intelligent systems tailored for the realities of the physical world.

Gemini Notebook

ARTIFICIAL INTELLIGENCE | SUSTAINABLE FUTURE

The Future of AI is Tiny

> True technological scale isn't found in building larger data centers. It is found in embedding ***frugal, sustainable intelligence*** into the very fabric of our environment.



Frugal Innovation

Smarter solutions with less resources, more impact.

Sustainable Intelligence

AI that respects our planet and preserves its resources.



Global Impact

Small, intelligent steps for a healthier and more resilient world.



SMALL TECHNOLOGIES • BIGGER CHANGE

مركز الملكة رانيا للدراسات الأردنية
وخدمة المجتمع – جامعة اليرموك

Queen Rania Center for Jordanian Studies
and Community Services – Yarmouk University

Thank You

Together for a stronger community
and a brighter future

Education & Training

Consulting & Expertise

Community Service

Sustainable Development



Knowledge · Innovation · Community Impact