IESTI01 - TinyML

Machine Learning for The physical world

Guest Lecturer, **Daniel Situnayake**
Edge Impulse

July 14th, 2021

Machine learning for the physical world

AI on embedded and edge devices

Daniel Situnayake
@dansitu

- Edge Impulse
- TinyML (O'Reilly)
- Google TensorFlow
- TinyML Foundation
- Tiny Farms

Agenda for the next 60 minutes

- Intro to embedded machine learning
- Real world case studies
- Why you should be working on embedded ML
- Resources for getting started

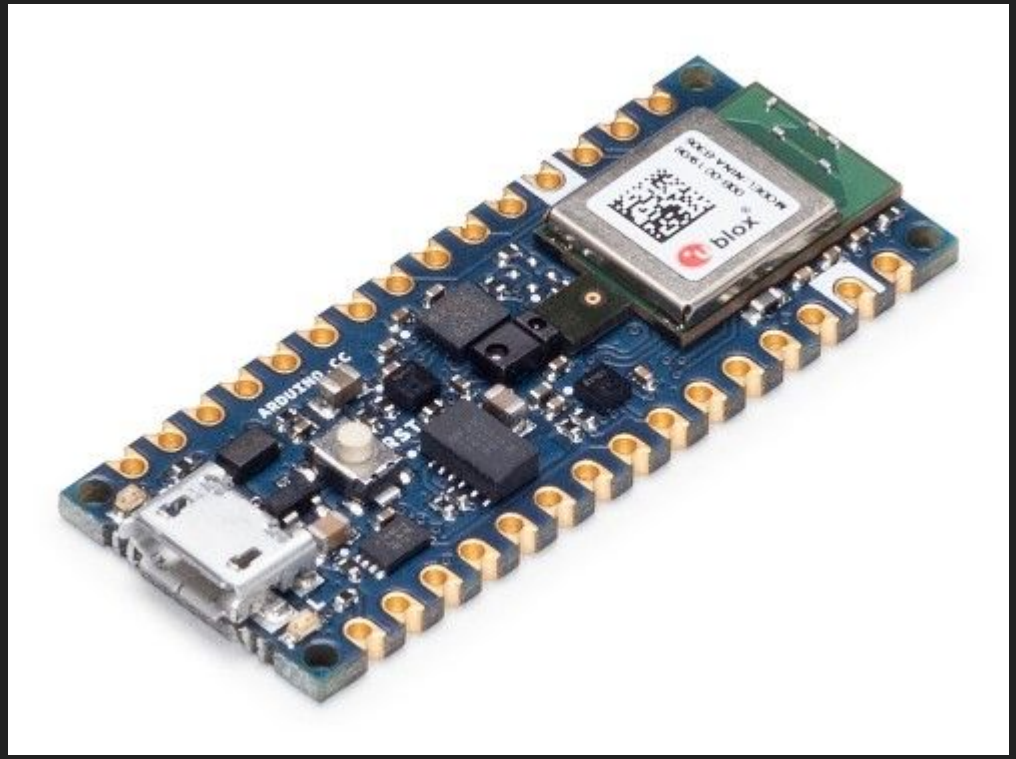
Please ask lots of questions!

What are edge and embedded devices?

Embedded devices

- Microcontrollers (MCUs) and Digital signal processors (DSPs)
- No/minimal operating system, run code on bare metal
- Program in C/C++
- May include hardware acceleration for fancy math
- May not even have FPU!

- Optimized for cost, energy efficiency, and size



Nordic nrf52840

- 256KB RAM
- 64MHz clock speed
- 10s of µA at 3V
- 1MB flash storage

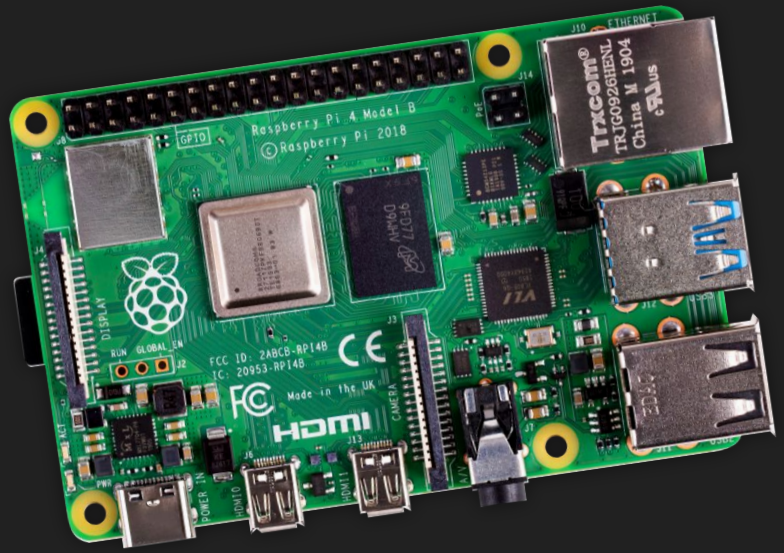
WE-I Plus ASIC +
Synopsys ARC EM9D DSP

- 2MB RAM
- 400MHz clock speed
- 10s of µA at 3V
- 2MB flash storage

System on a Chip devices (SoCs)

- Real operating system (often Linux)
- Program in whatever language you like
- May include hardware acceleration (even GPUs)
- May have additional MCUs or DSPs on board

- Optimized for general purpose performance first, then cost, energy efficiency, and size



Broadcom BCM2711

- Quad core 64-bit SoC
- 2-8GB RAM
- 1.5GHz clock speed
- 3A at 5V (whole system)

ML accelerators

- Silicon designed to run ML
- Specific APIs required for use
- Some paired with MCUs, some designed to pair with SoCs

- Optimized for ML performance, cost, energy efficiency, and size, depending on application

NVidia Jetson Nano

- 128-core GPU
- 2-4GB RAM
- 2A at 5V (whole system)

Why run ML on these?

B L E R P

Bandwidth

Latency

Economics

Reliability

Privacy

Jeff Bier

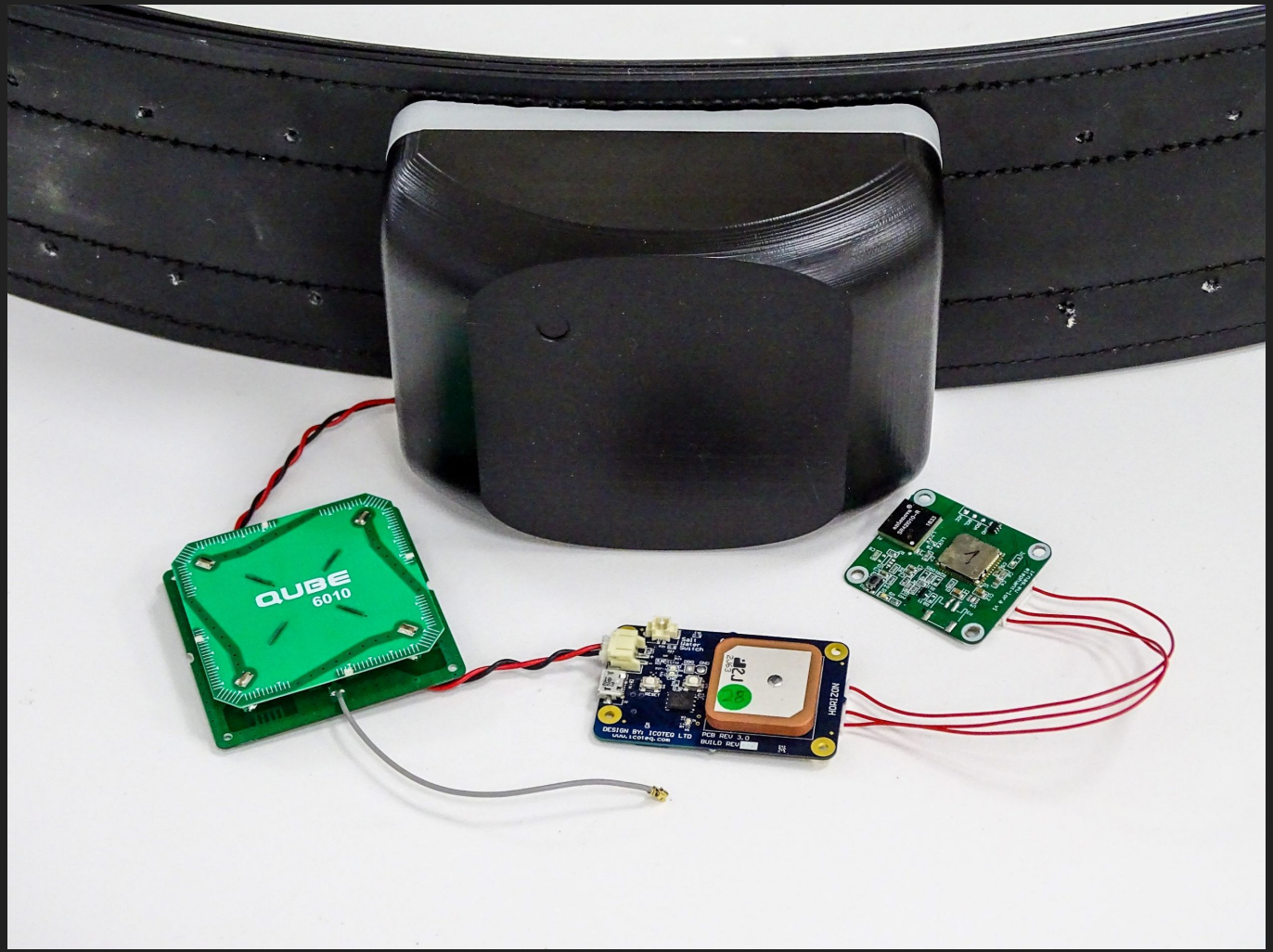
High-level use cases

- Smart sensors (bandwidth reduction)
- Privacy-focused
- Low latency (safety, high throughput)
- Intelligent objects

Common types of data

- Images (including infrared)
- Audio
- Time series sensor (vibration, temperature, etc)
- Positional (GPS)
- Combinations of the above
- + Many other things

In the real world?


QUBE
6010
DESIGN BY: ICOTEQ LTD
WWW.ICOTEQ.COM
PCB REV 3.0
BUILD REV
NOZTECH


RAM-1 MV A2
Code: 90 00 02
Date: 12.01.2021
ID: 352656105

Monitoring of installation point > 1 kV - Slovenia


Hi, how can I help?

What’s needed to make TinyML work?

Tiny models

Optimized kernels

Relevant metrics

Efficient runtime

Suitable datasets

Model compression

Machine learning workflow

1. Figure out what you want to do
2. Obtain a dataset
3. Train a model
4. Tweak the model and add more data until it works (or you give up)
5. Deploy the model
6. Monitor to make sure it is working

Keep it simple

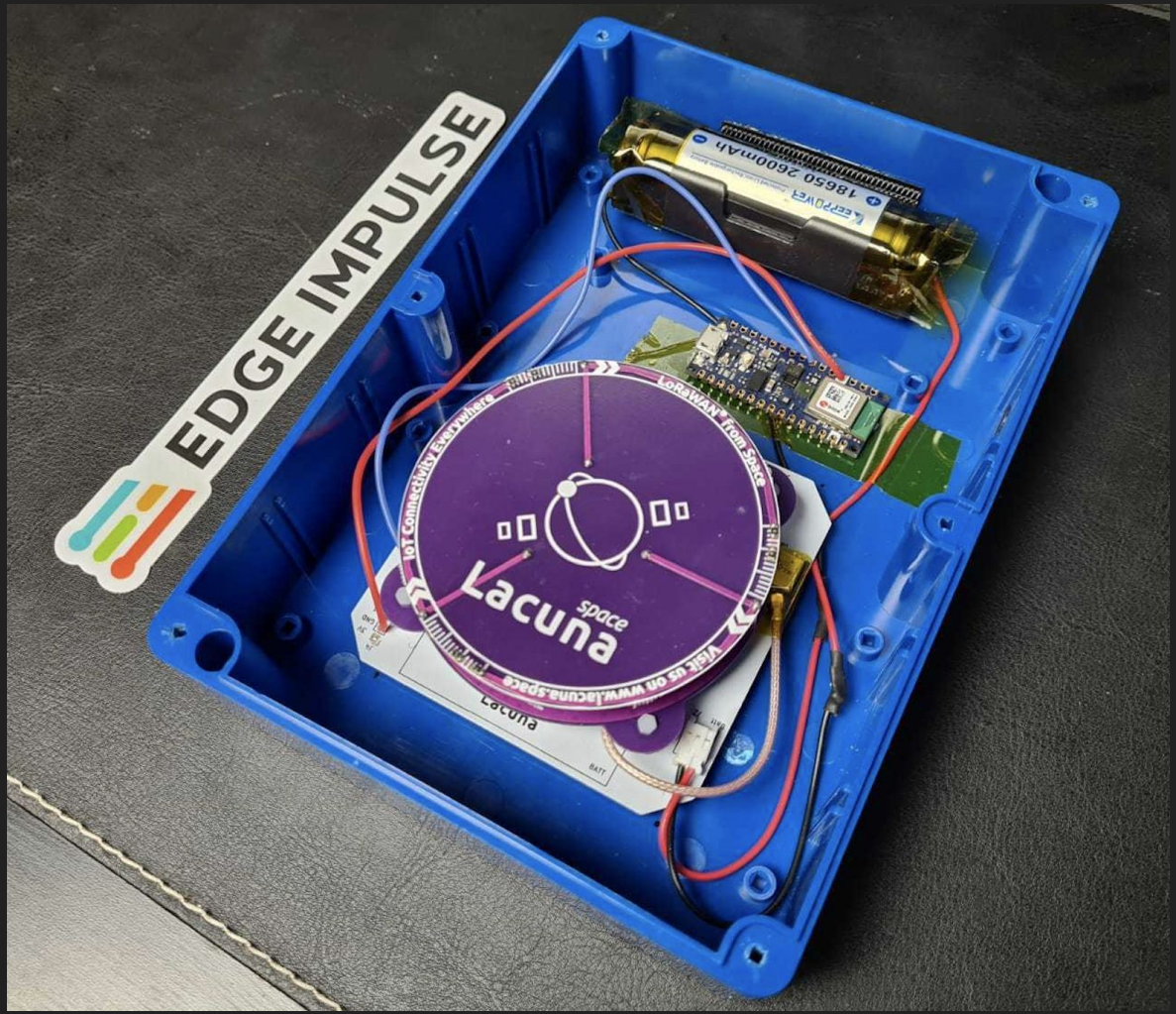
From production sensors and environment, if possible

Apply optimizations to reduce model size

Make sure it runs fast enough on your target

Hard with less connectivity!


© Rita Ivanauskas


EDGE IMPULSE
IoT Connectivity Everywhere
LoRaWAN from Space
space
Lacuna
Visit us on www.lacuna.space
Lacuna

The **practical** state of the art in 2021

Tiny models

- Model architectures that work with all sensor types (time series, image, etc.)
- DSP pipelines that pair well with models
- Basic model optimization (int8 quantization)

Accelerated hardware

- DSP and SIMD for signal processing and ML (e.g. Arm CMSIS)
- Ultra-low-power MCUs with DSP (e.g. Eta Compute)
- Embedded NN accelerators for specific purpose (e.g. Syntiant)

End-to-end tooling

- Ingestion and management for any type of data
- No-code model training and version tracking
- Simple deployment with target-specific optimizations

What do we still need?

Opportunities for R&D

- Software support for SOTA optimization (pruning, BNNs)
- Embedded-specific deep learning architectures
- Smarter division of embedded DSP/ML workload
- Public and commercial datasets for industrial use

Fast forward 5 years

Transformative technology: Embedded ML in 2025

- Demanded by industry and consumers
- Present in a majority of embedded devices
- Feature support in most MCU architectures
- Nearly all embedded engineers will have touched it
- Advanced ML Ops tools critical to development and deployment

Why **you** should be working in embedded ML

Massive scale impact

Huge research opportunities

Touches almost every sector

Where ML meets the real world

Building a better future

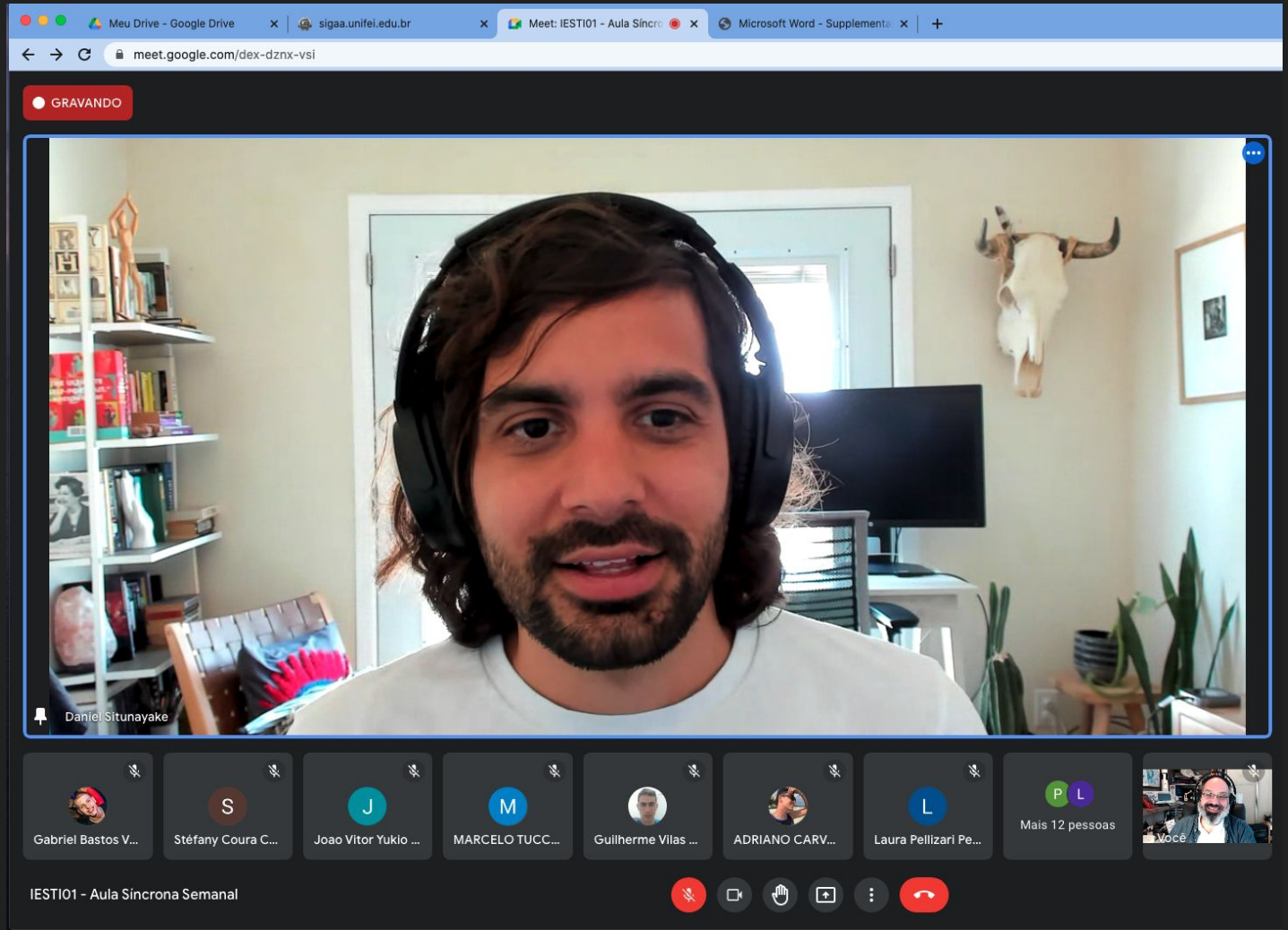
Vision on embedded

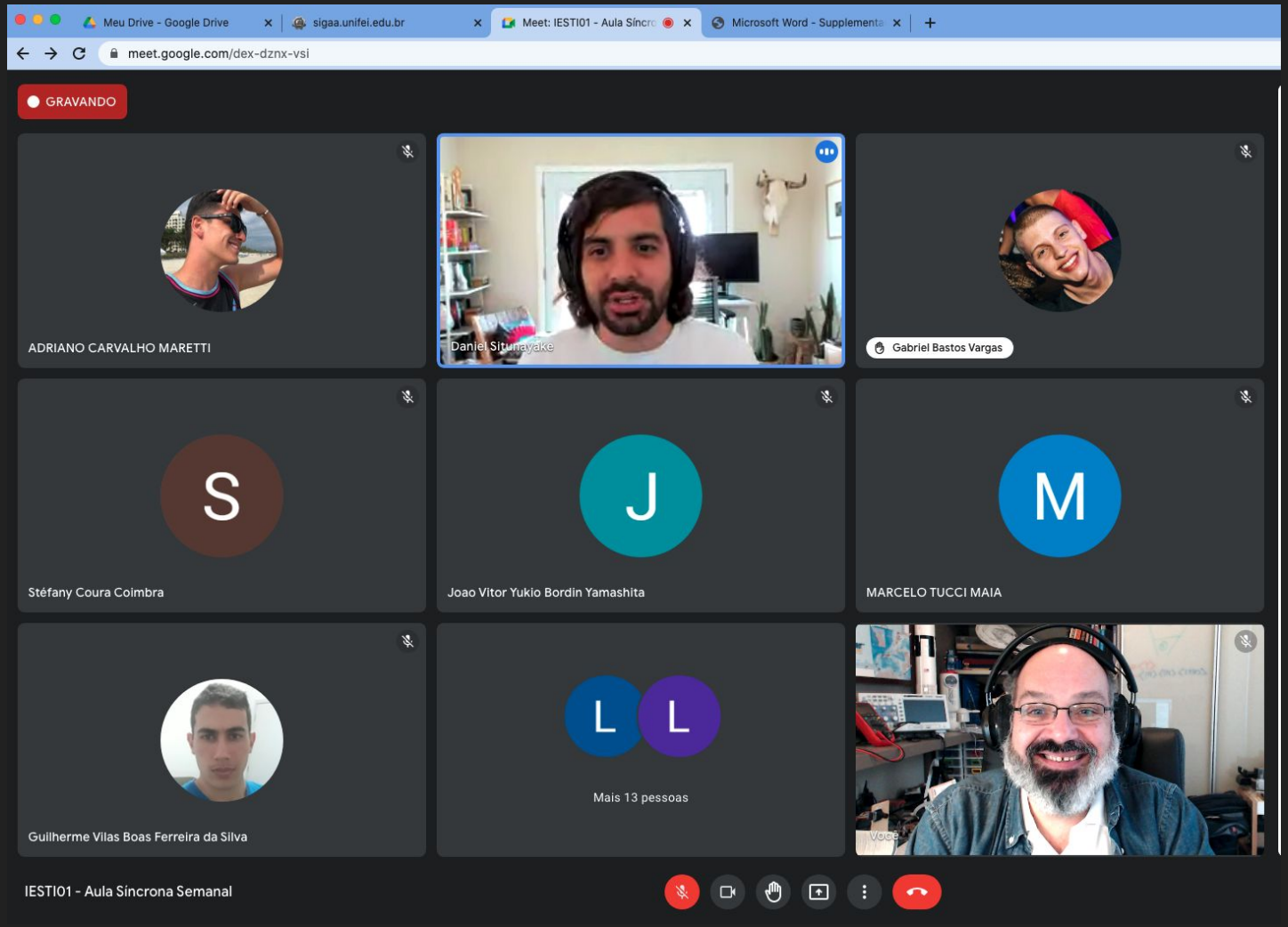
Questions?

Daniel Situnayake

dan@edgeimpulse.com

@dansitu


Meu Drive - Google Drive
sigaa.unifei.edu.br
Meet: IESTI01 - Aula Síncro
Microsoft Word - Supplementa
meet.google.com/dex-dznx-vsi
GRAVANDO
Daniel Situnayake
Gabriel Bastos V...
Stéfany Coura C...
Joao Vitor Yukio ...
MARCELO TUCC...
Guilherme Vilas ...
ADRIANO CARV...
Laura Pellizari Pe...
Mais 12 pessoas
Você
IESTI01 - Aula Síncrona Semanal


Meu Drive - Google Drive
sigaa.unifei.edu.br
Meet: IESTI01 - Aula Síncro
Microsoft Word - Suppleme
meet.google.com/dex-dznx-vsi
GRAVANDO
ADRIANO CARVALHO MARETTI
Daniel Situnayake
Gabriel Bastos Vargas
S
Stéfany Coura Coimbra
J
Joao Vitor Yukio Bordin Yamashita
M
MARCELO TUCCI MAIA
Guilherme Vilas Boas Ferreira da Silva
L
L
Mais 13 pessoas
Você
IESTI01 - Aula Síncrona Semanal

Thanks

And stay safe!