tinyML® Summit

Enabling Ultra-low Power Machine Learning at the Edge

Products and applications enabled by tinyML

March 28 – 29, 2023

www.tinyML.org

MAYO CLINIC

LOW-ENERGY PHYSIOLOGIC BIOMARKERS

WEARABLE ML INFERENCE WITH A GAP9 RISCV PROCESSOR

Christopher L. Felton

tinyML Summit 2023
27-Mar-2023, San Francisco

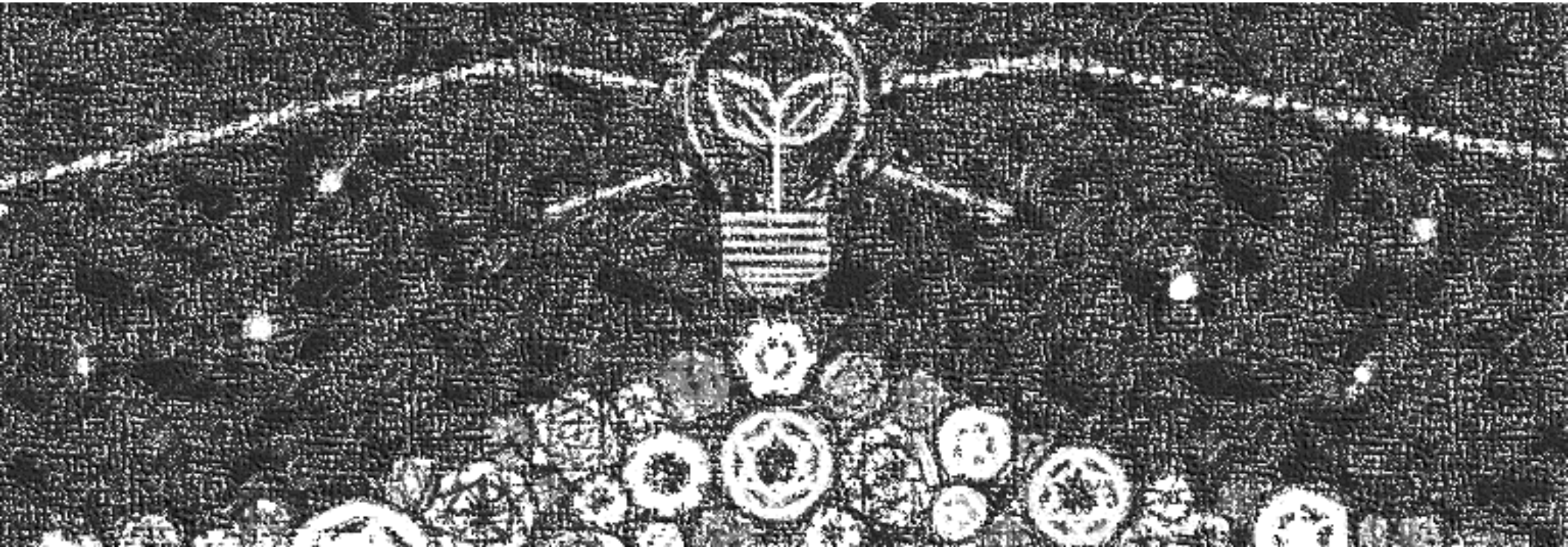
MAYO TEAM PATH TO TINYML

- Disclosures
- Background (Wearable Platform, Data, Models)
- The Path to Low-Energy Machine Inference
- Towards Generative Physiologic Signals

DISCLOSURE

NO CONFLICTS

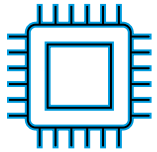
Past work supported by Mayo IR&D and Government funding – no explicit funding for the presented and updated content



MAYO TEAM ACKNOWLEDGEMENTS

MULTIDISCIPLINARY TEAM REPRESENTED IN PRESENTATION

Study Design; Human Subject Testing; Data Analysis; Model Training; Model Evaluation; Real-Time Lightweight Implementations; Hardware Platforms.



SPPDG

Clifton Haider, Ph.D

Lincoln Kirchoff

Spencer Wright

BACE

David Holmes, Ph.D

HIP

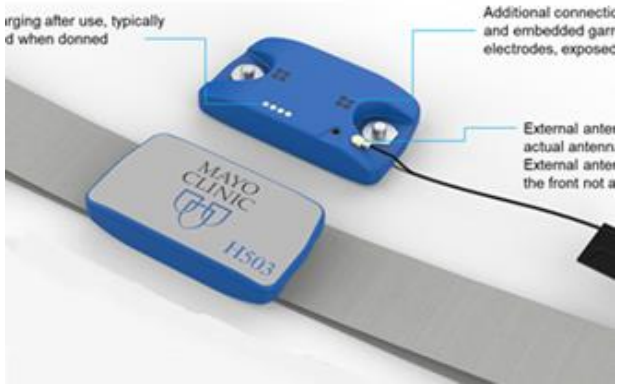
Michael Joyner, M.D

Timothy Curry, M.D, Ph.D

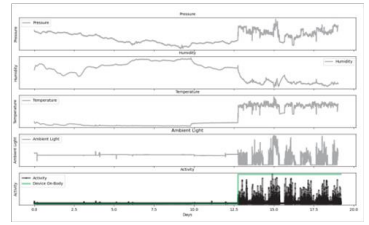
Christopher Johnson (CJ)

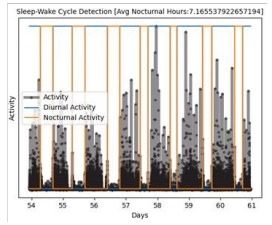
BACKGROUND

Scalable and Modular Wearable Platform



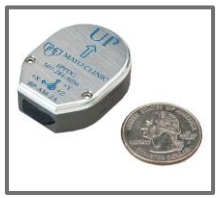
Data Collection

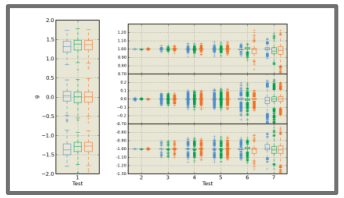




WEARABLE DEVELOPMENT TIMELINE

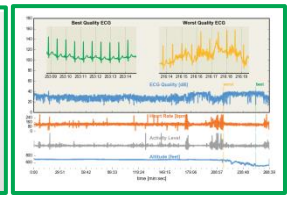
Past Research and Development
Micro-electronic and integrated circuit package miniaturization





Precision and accuracy
Long term calibration and accuracy of MEMs sensor transducers, in addition to 30-60 day run-times.

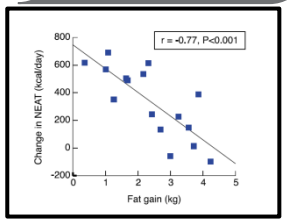
Optical characterization and optical sensing development



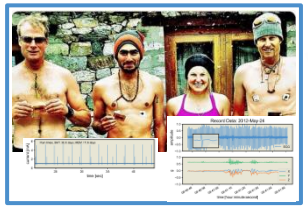
Multi-sensor Platform
Modular and scalable architecture to support a plethora of physiologic and environmental sensors, a device capable of full wave recording and real-time processing.

Present and Future Development
Full system integration, and advanced embedded algorithms

2007 | **2010** | **2012** | **2017** | **2018** | **2019** | **202***



Non-exercise activity thermogenesis
Wearable device to measure a subject's activity over extended time durations.



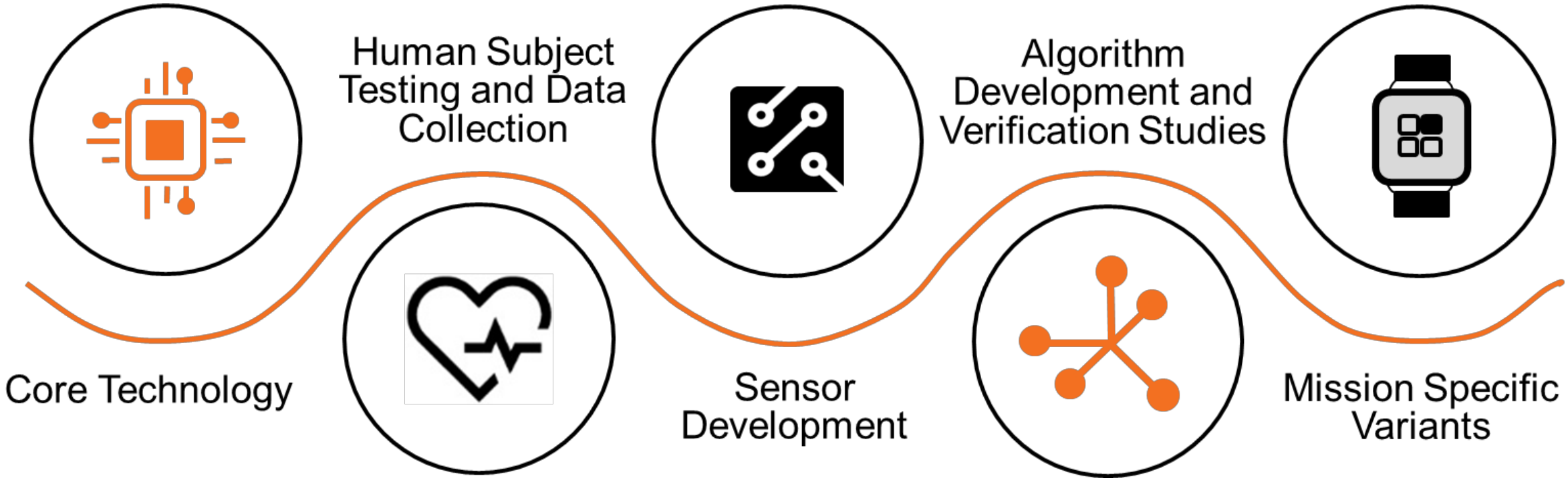
Longest Remote Recording Device
15-day run-time, ECG and motion data recording, from a 2012 Everest summit expedition on elite climbers and base camp researchers.

Core Technology and Variants
Sponsored to develop the core technology; mission specific variants; as well as real-time analysis and predictive algorithms.

Special Purpose
Device variants with a targeted use.

DEVELOPMENT PILLARS

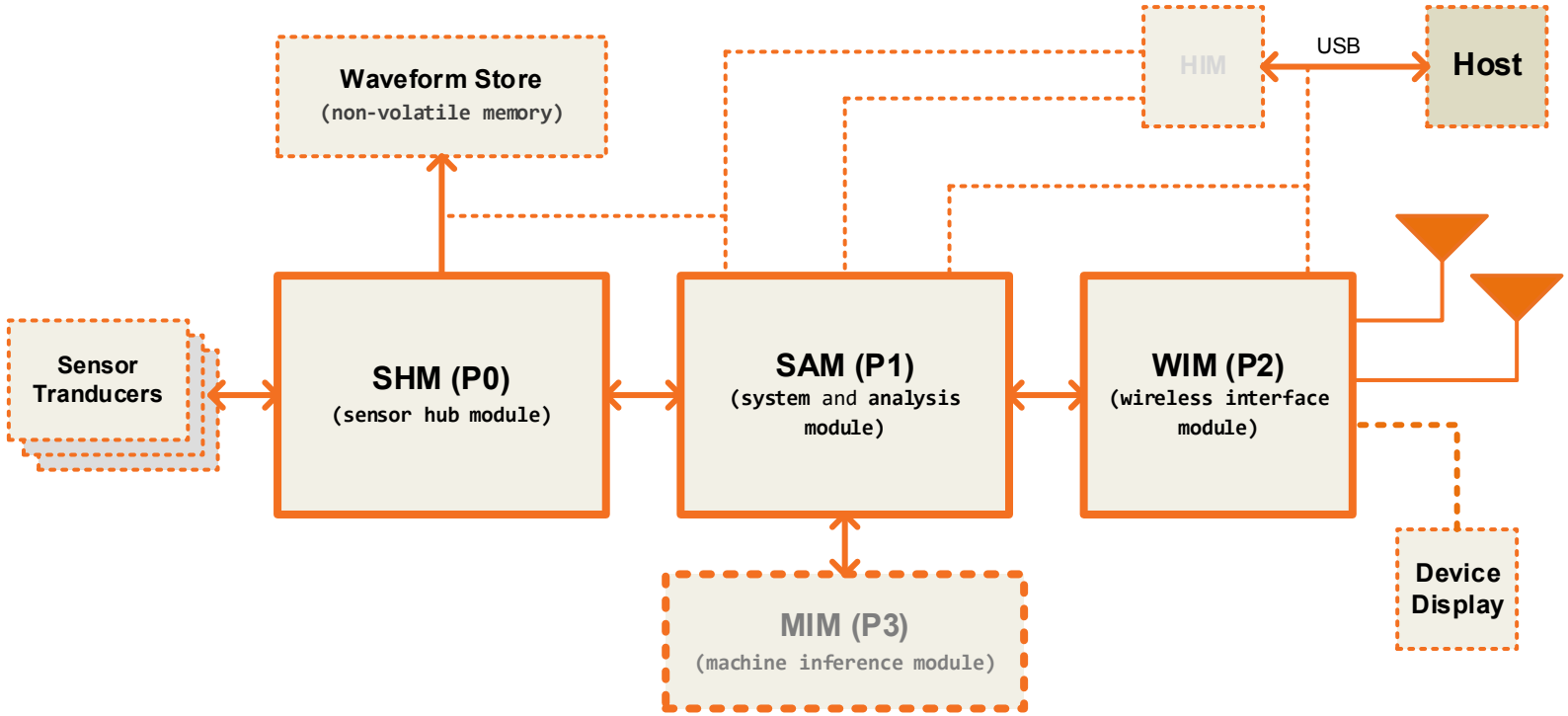
Use wearable prototypes to collect physiologic signals, evaluate signals to estimate biomarkers, iterate design.



WEARABLE PLATFORM AND DEVICE VARIANTS

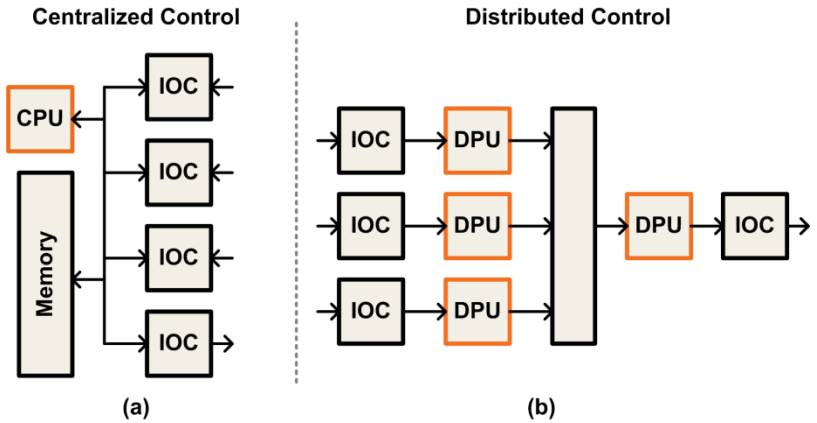
- Rapid prototype device variants from core-technology
- Modular to include various sensor transducers
- Scalable long-runtimes (limited features) to many sensor streams
- Lightweight algorithm development target
- Test and characterize sensor transducers

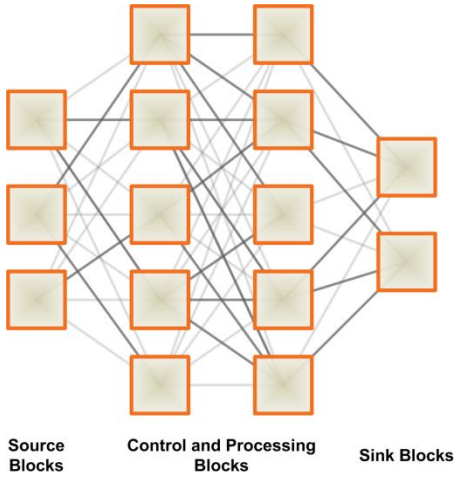
WEARABLE PLATFORM ARCHITECTURE



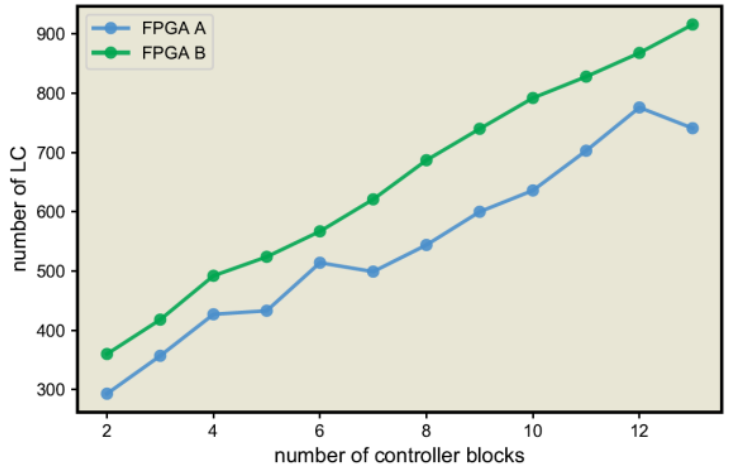
SHM: HDL METHODOLOGY

- Hardware Description Methodology
- Leveraged a Python based hardware description language (HDL), MyHDL, remove redundant layers of programmability. The methodology fully exploits an executable elaboration phase to provide system level parameretization. Further, energy conservation is achieved by reduced toggle rates on the programmable logic, i.e. a field programmable gate array (FGPA) , when compared to a central processing unit (CPU).





	FPGA A		FPGA B		FPGA C	
Core	LC	FF	LC	FF	LC	FF
I2C [19]	359	193	339	181	298	193
I2C crux	205	101	240	89	183	101
Savings	43%	48%	29%	51%	39%	48%
UART [4]	281	171	253	171	286	171
UART crux	145	91	129	84	174	155
Savings	48%	47%	49%	51%	39%	9%



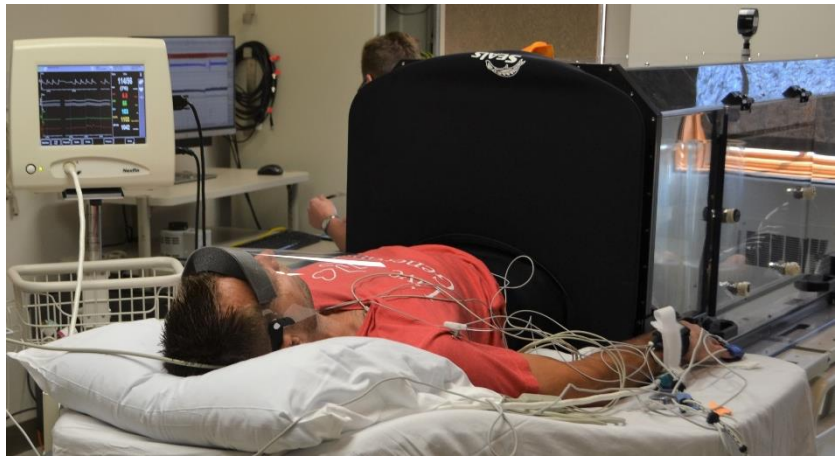
PHYSIOLOGIC DATASETS AND TARGETS

Human Subject Studies

Clinical Data Collection

HYPOVOLEMIA HUMAN SUBJECT TESTS

- Hypovolemia (blood loss) is the leading cause of death in trauma victims.
- Traditional vitals signs yield little insight into severity, as the body compensates for blood loss and maintains pulse and blood pressure.
- To gain a better understanding of hypovolemia, Lower Body Negative Pressure (LBNP) human studies are performed.
- The Compensatory Reserve Metric (CRM) was developed to characterize hypovolemia progression from these physiologic signals.



- ECG
- Same Side PPG
- Finger Clamp PPG

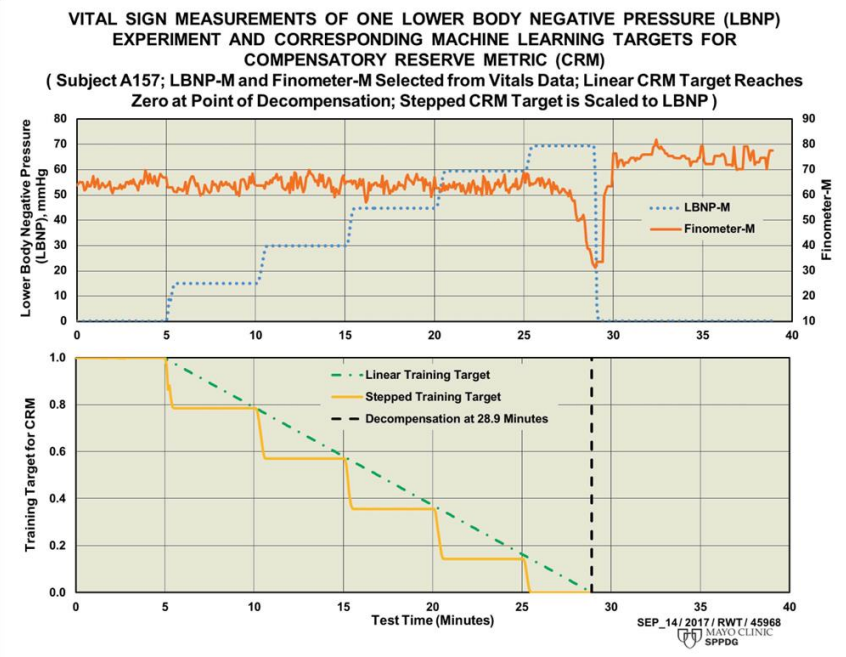
Mayo Clinic completed 100 LBNP studies to support analysis, model validation and development

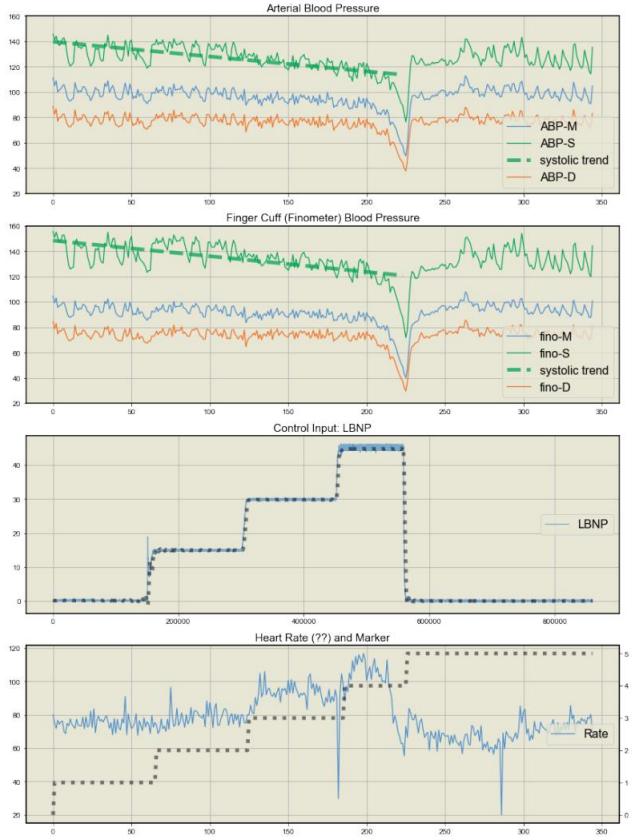
LOWER BODY NEGATIVE PRESSURE (LBNP) MODEL FOR HYPOVOLEMIA (CRM/CRA)

Vitals signs (MAP) remain stable throughout blood loss, despite increasing LBNP.

CRM is defined as 100% during first five minutes resting baseline, and 0% at decompensation. CRM target could be linear or stepped.

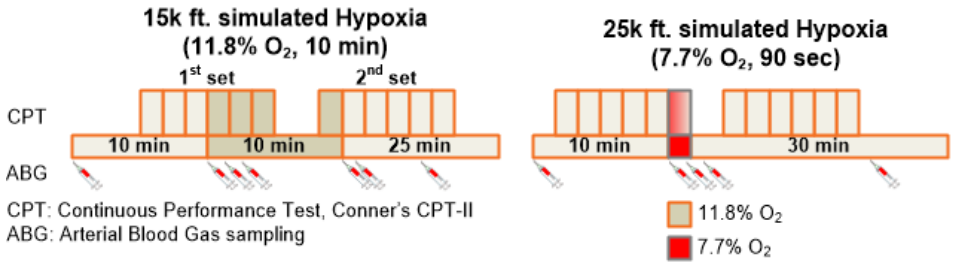
Protocol must include end point to accurately define slope of CRM





HYPOXIC-HYPOXIA HUMAN SUBJECT TESTS

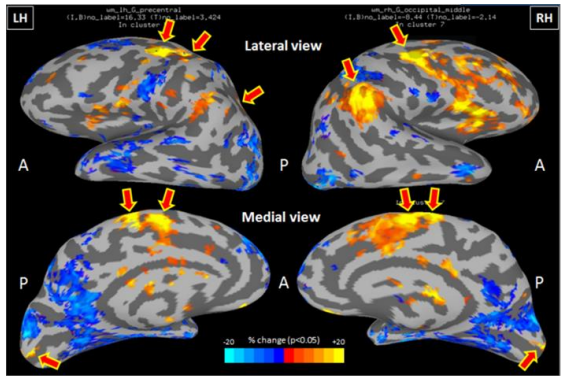
- Goals:
  - Concurrent collection of ECG waveforms using laboratory reference and wearable omni-device.
  - Quantification of physiologic response to hypoxic conditions.
    - 15k feet experiment and 25k feet experiment.
  - Cognitive assessments at baseline and under hypoxic hypoxia conditions.



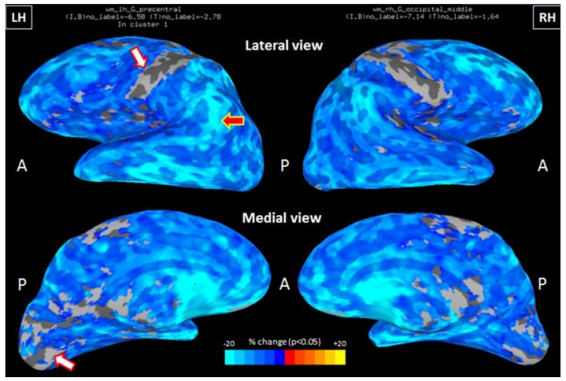
- There is a lack of objective measures to predict, prevent, monitor, and mitigate cerebral hypoxia and hypoxia-related physiological and cognitive dysfunction.
- Goal: characterize regional versus global cerebral oxygenation and oxygen consumption during mental tasks.

HUMAN SUBJECT PHYSIOLOGIC WAVEFORMS

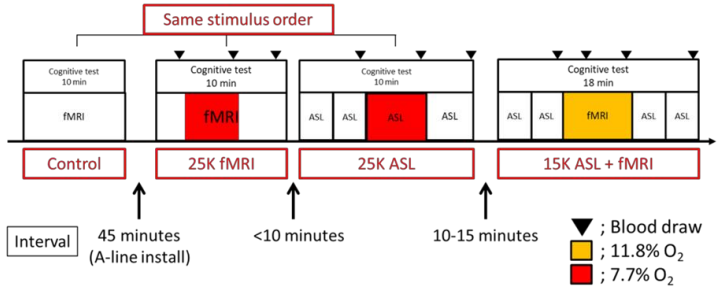
Cognitive Task CBF Baseline

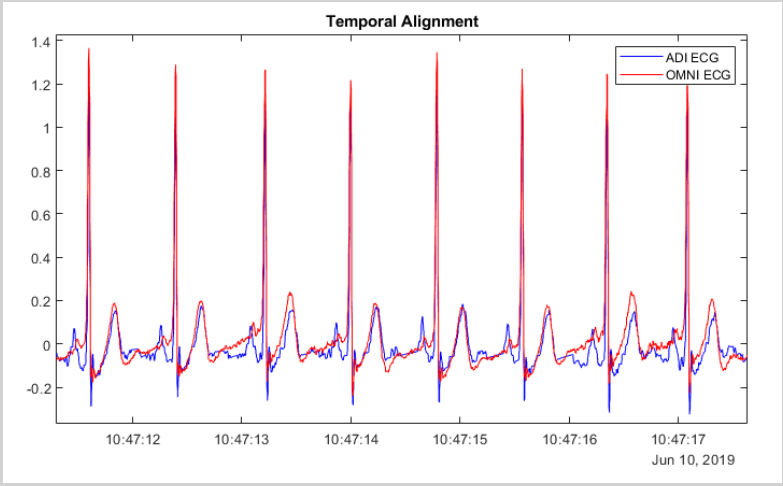


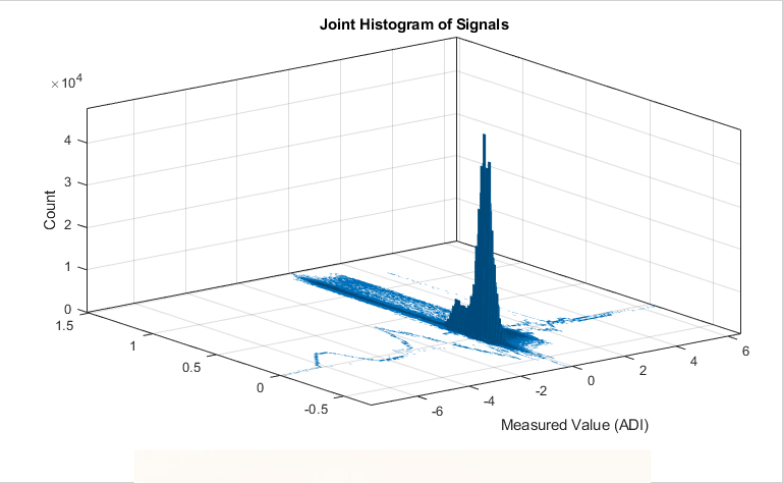
Hypoxia effect on CBF

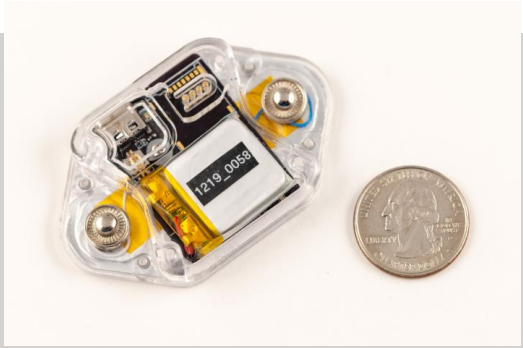


Phase 3 Protocol

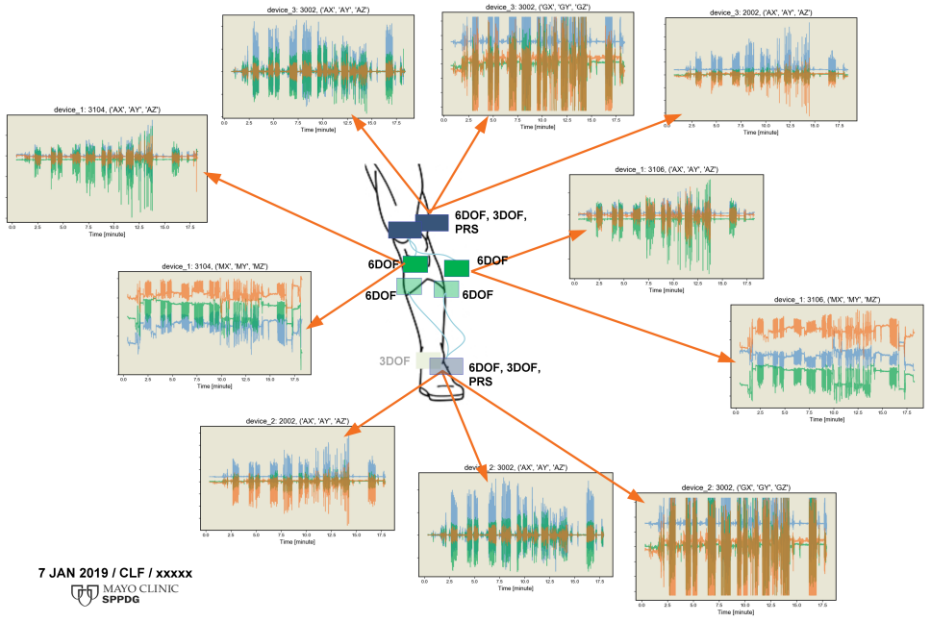


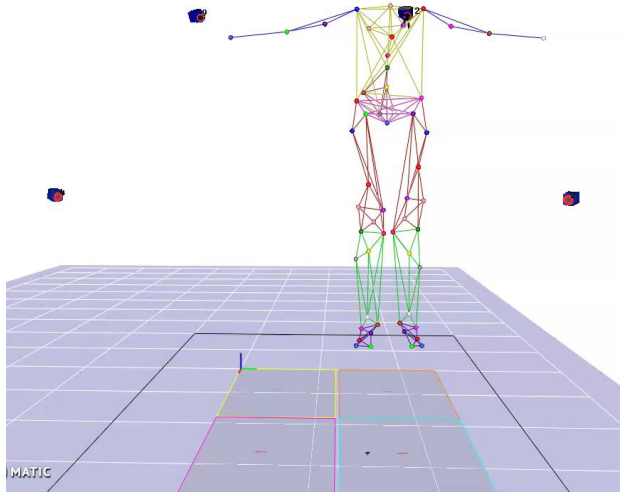






LOWER EXTREMITY INJURY PREDICTION DATASETS





THE PATH TO LOW-ENERGY MACHINE INFERENCE

First Models … Small Models … Reduced Models

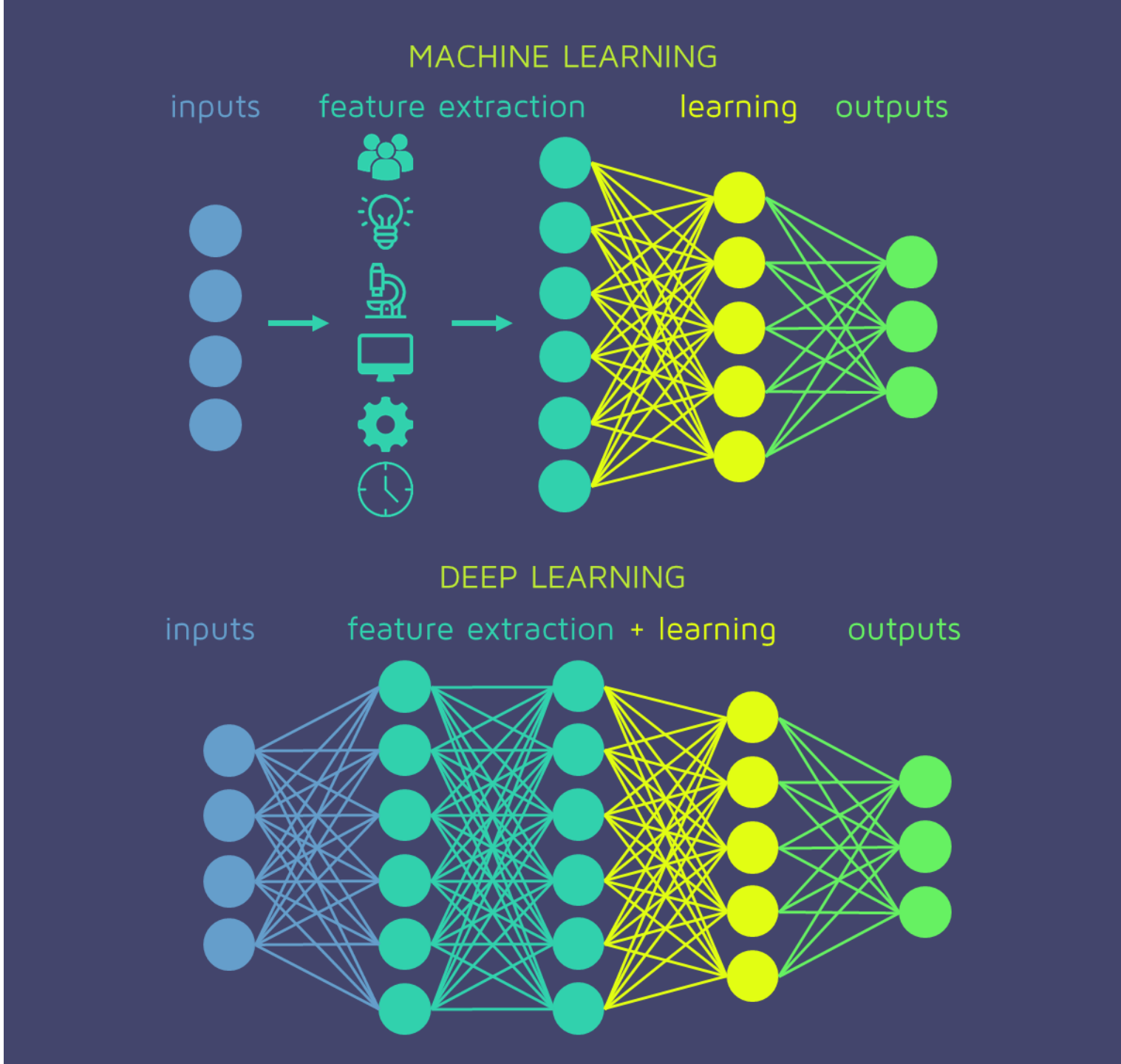
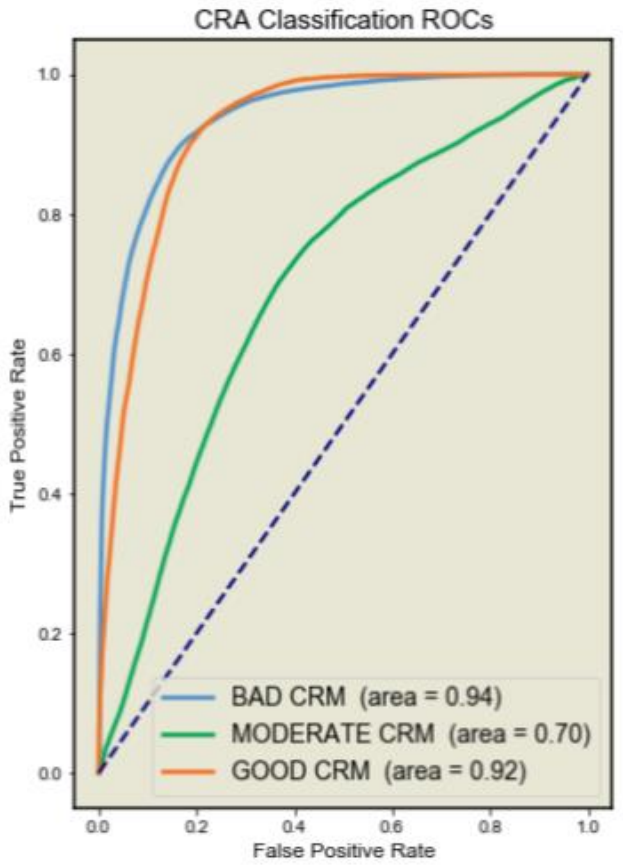
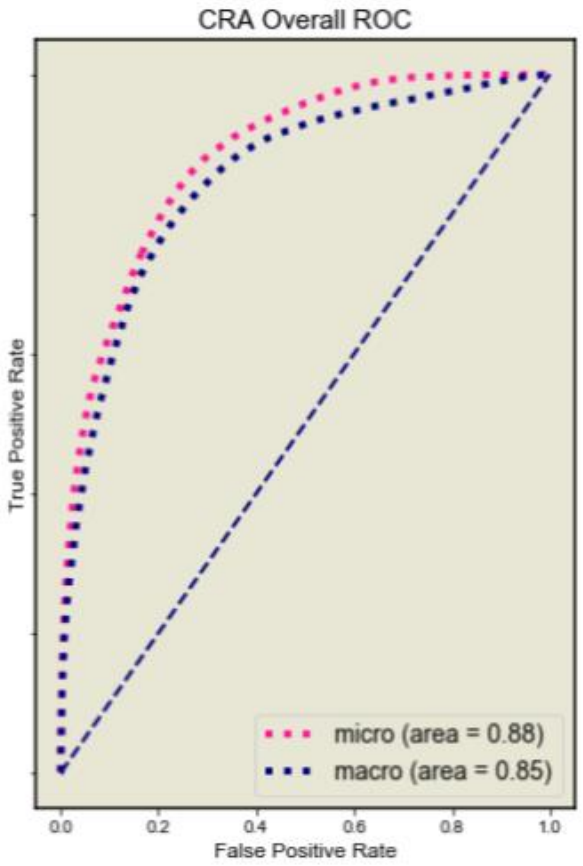


Figure from: What is the difference …
https://quantdare.com/what-is-the-difference-between-deep-learning-and-machine-learning/#:~:text=We%20refer%20to%20shallow%20learning,learning%20that%20are%20not%20deep.

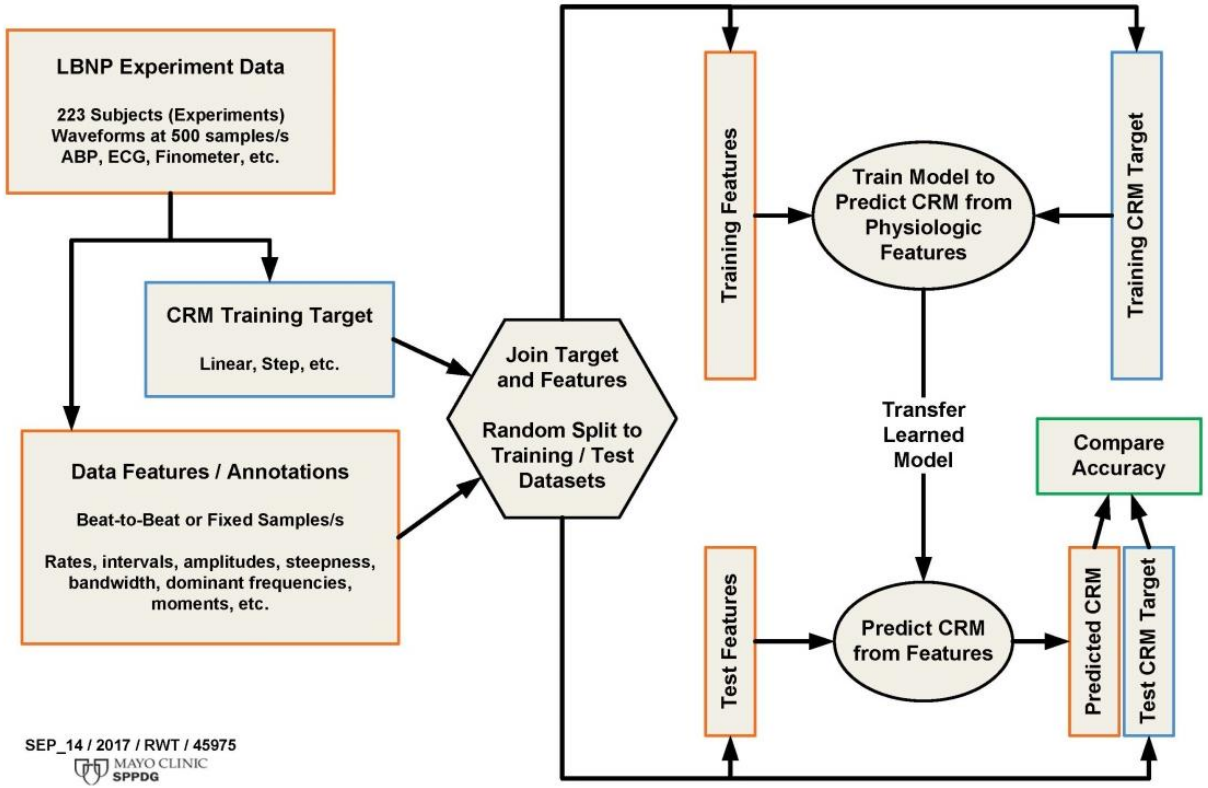
TRADITIONAL MACHINE-LEARNING

- First endeavors into lightweight machine-learning
- Featured engineered
- Use of traditional machine-learning < 1K Bytes coefficients
  - LSVM and Ridge classifiers
- Additional processing, pre and post





CRM MODEL FRAMEWORK AND ARCHITECTURE

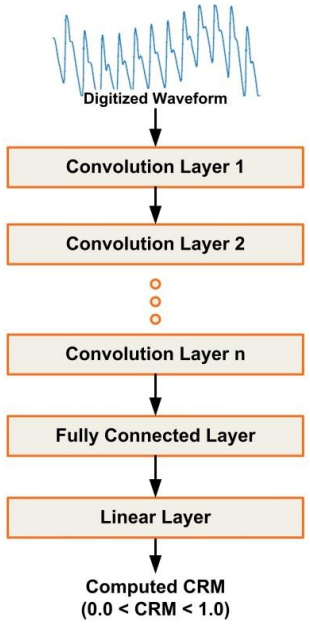


SEP_14 / 2017 / RWT / 45975

MAYO CLINIC SPPDG

SIMPLIFIED CONVOLUTIONAL NEURAL NETWORK (CNN) ARCHITECTURE FOR COMPUTING COMPENSATORY RESERVE METRIC (CRM)

(Hyper-Parameter Optimization Determines Number of Convolution Layers, Number of Filters, Activation Functions, Learning Rate, Pooling Type, and 15 Other Parameters)



Convolutional Layer Parameters Include
(Layers are Independent)

- Number of Filters
- Kernel Size
- Activation Function

Model Parameters Include

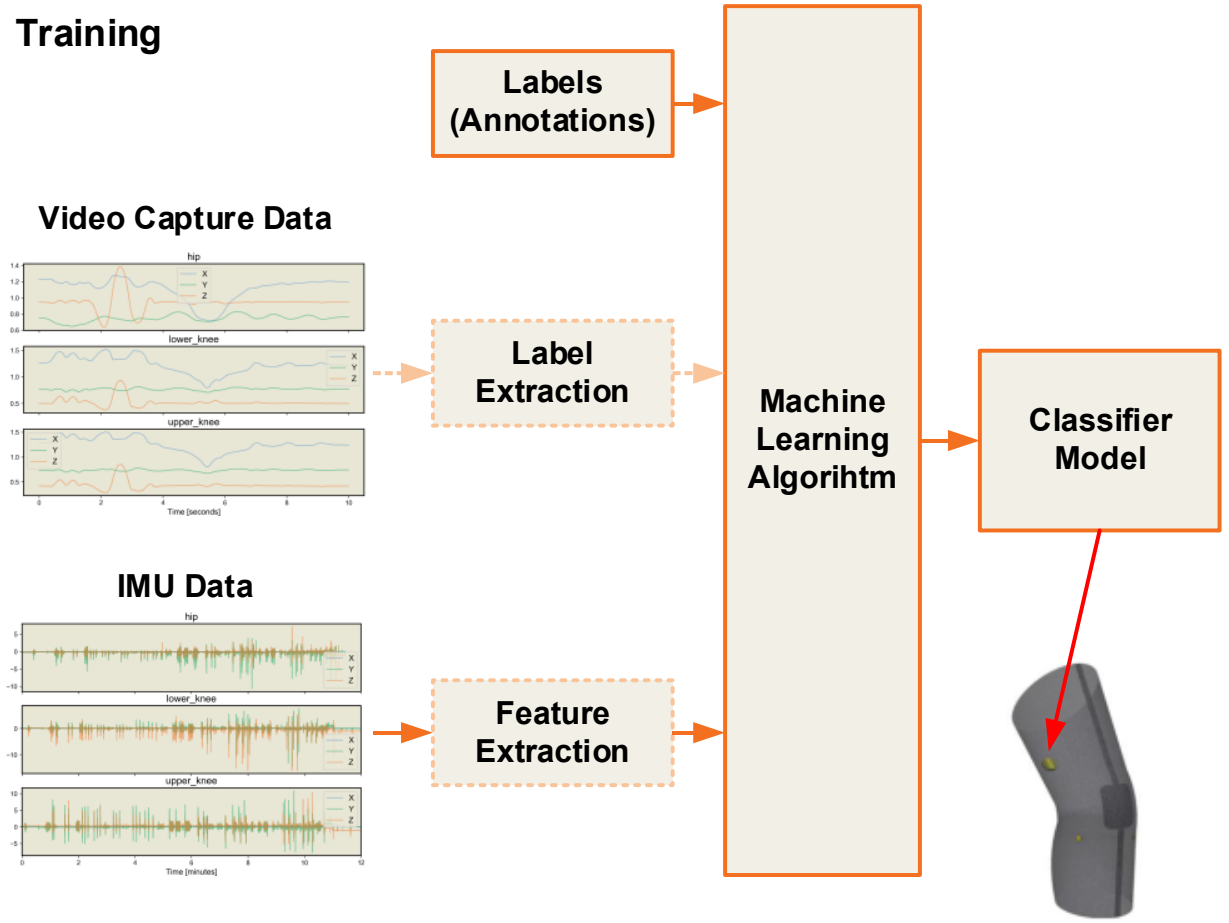
- Input Waveform Size
- Number of Convolutional Layers
- Dimension of Fully Connected Layer
- Learning Rate
- Optimizer
- Loss Function
- Training Duration (epochs)

Illustration omits pooling and dropout layers, which manage model complexity and over-fitting, respectively

DEC_05 / 2018 / RWT / 46203

MAYO CLINIC SPPDG

MODEL RAMEWORK INJURY PREVENTION

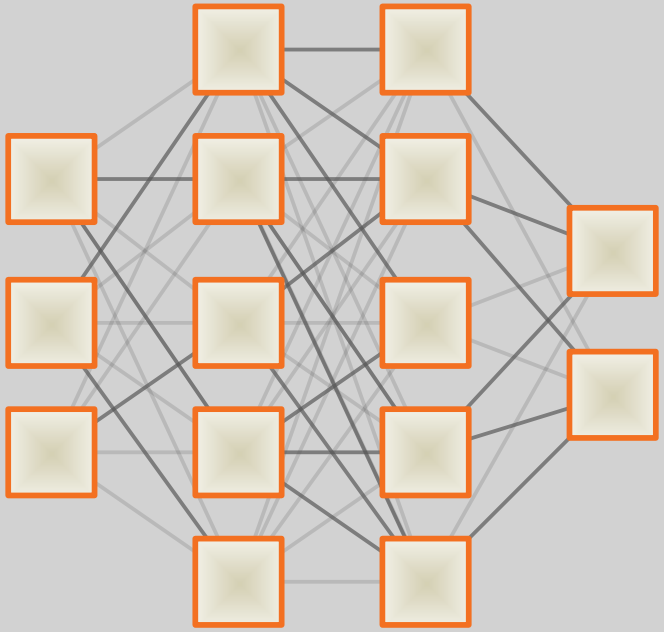


LOW-ENERGY MACHINE INFERENCE

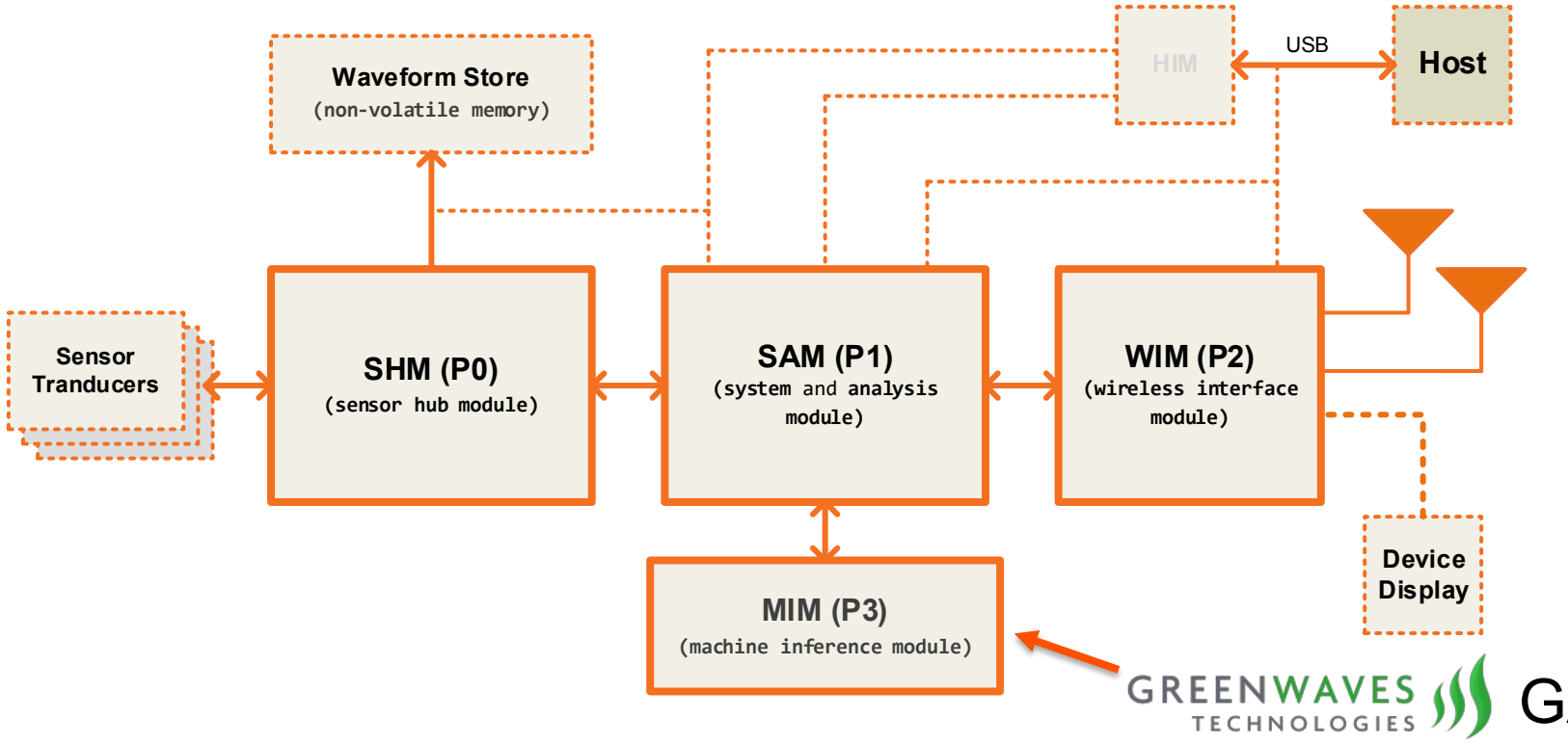
Adding a Machine Inference Processing Node: Towards Lightweight Low-Energy Machine Inference on a Wearable

Low-Energy Machine-Inference

Pipelines to Reduce Models



WEARABLE PLATFORM ARCHITECTURE



GAP9 SELECTION

GAP9	Alternates
• Small size, WLCSP	• Some

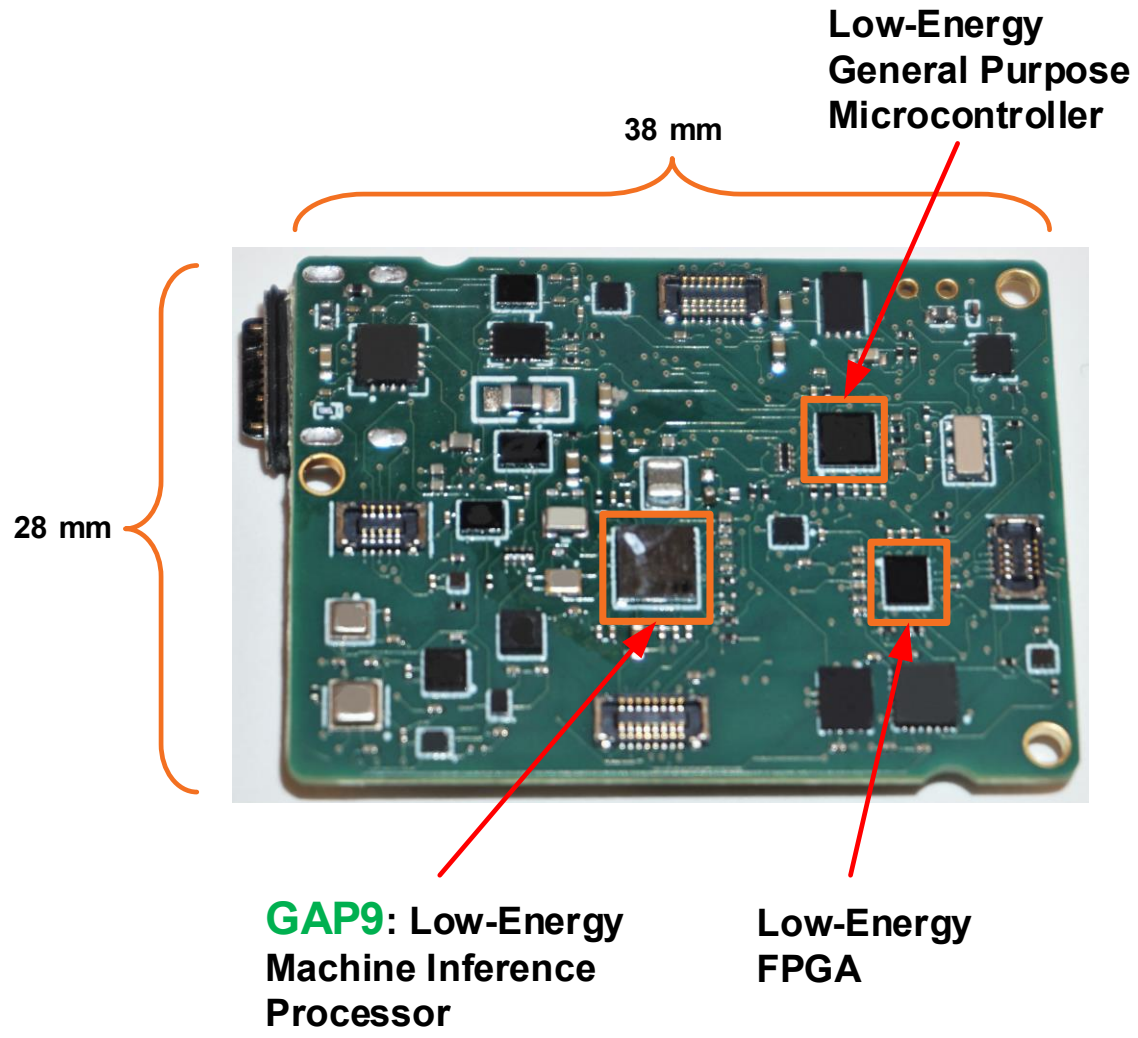
• RISCV
  • Multi-CPU
  • NE Accelerator
	• Few
	  • None
	  • Some

• Low-Energy
  • Joule per inference
  • ML Accelerators
	• Some
	  • Not at high parallelization

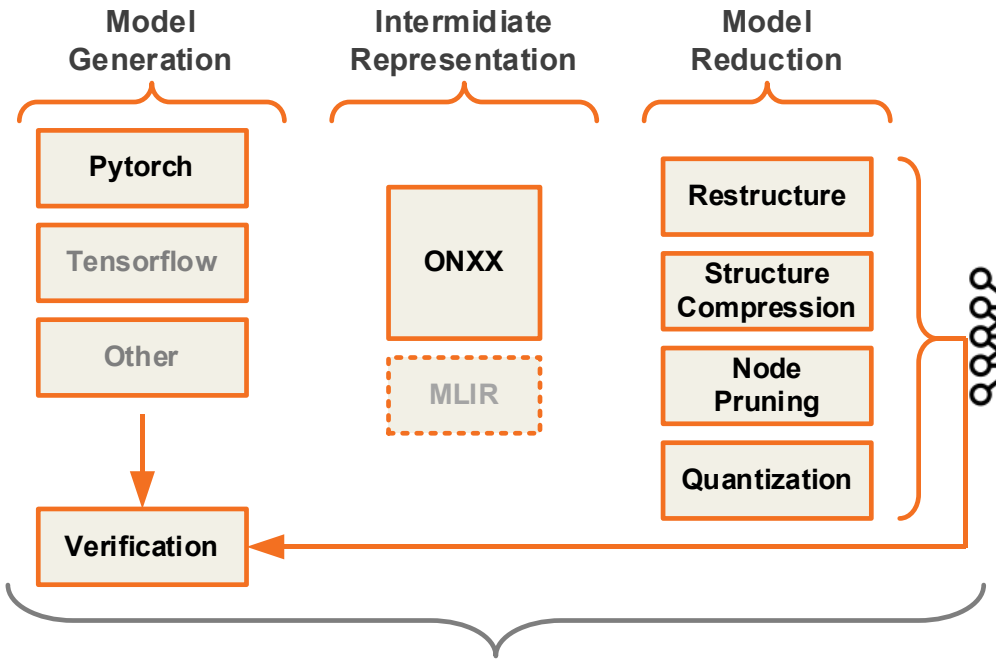
• Toolflow	• Some

MACHINE INFERENCE (GPA9) ADDITION

- Small footprint, did not increase our device footprint
  - Opportunities for smaller specific variants
- GPA9 design (streaming) fits into the overall architecture

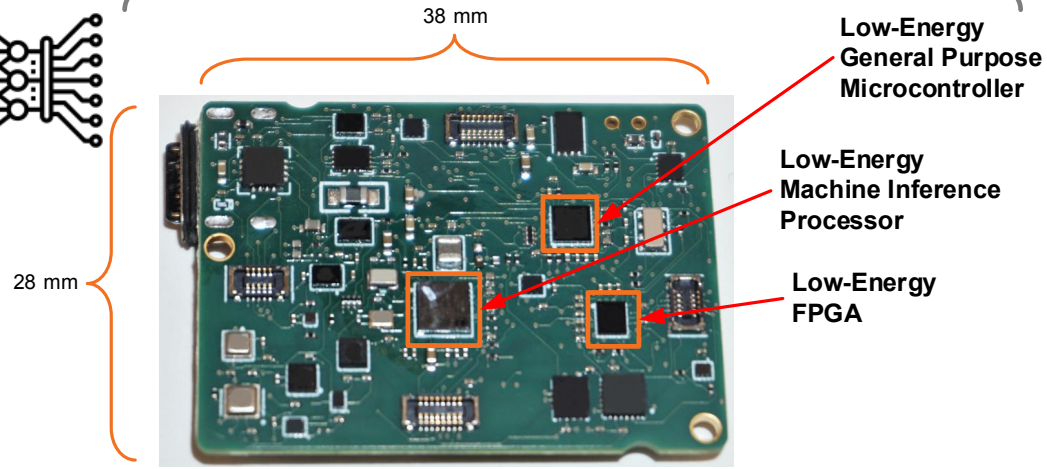


CONCEPTUAL PROCESSING PIPELINE

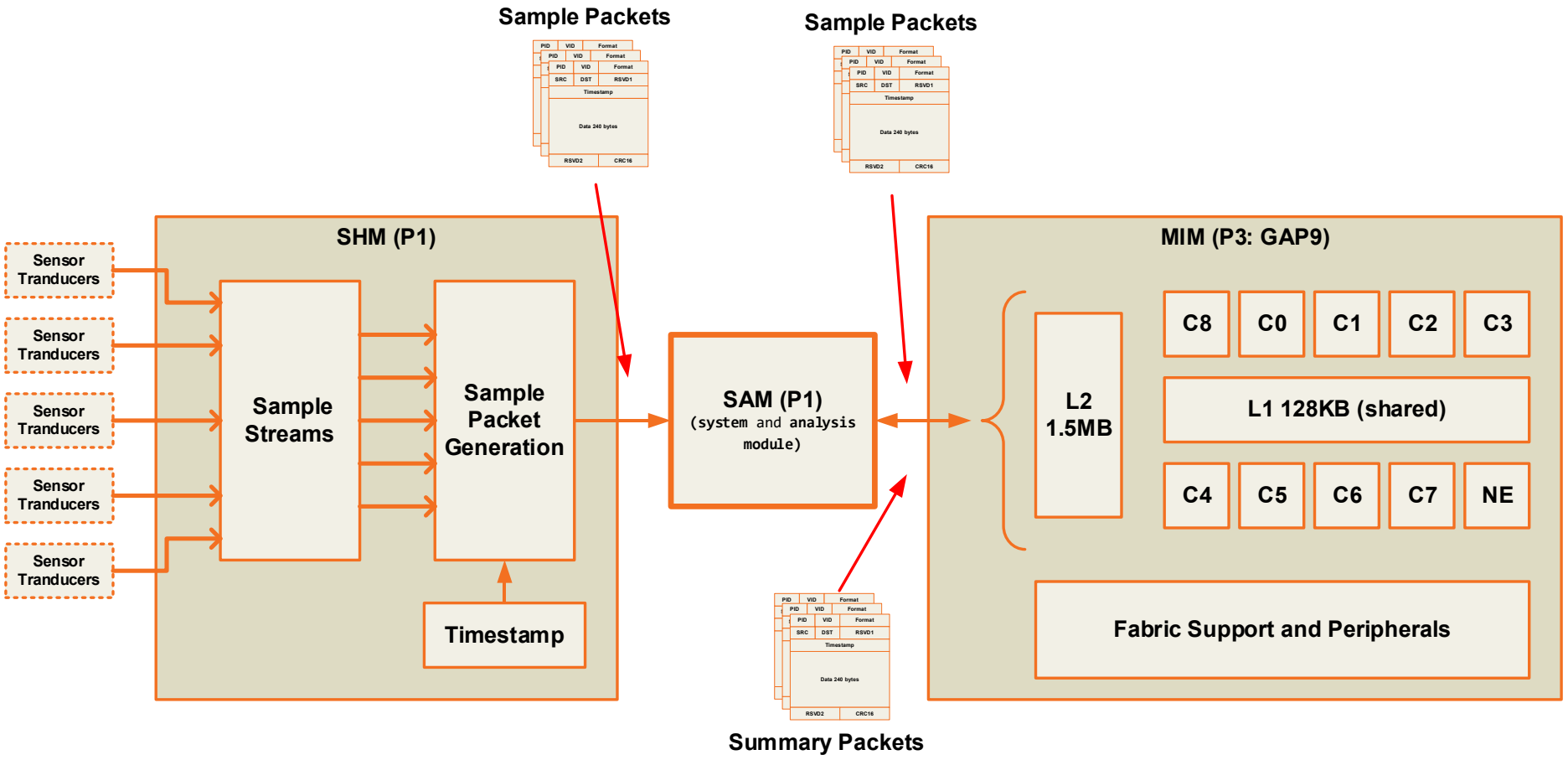


Model Reduction/Minimization Pipeline for Low-Energy Wearable Processing

Wearable Development Platform (omni-platform)
Multivariate/Multimodal Sensor Streams,
Low-Energy, Scalable and Modular

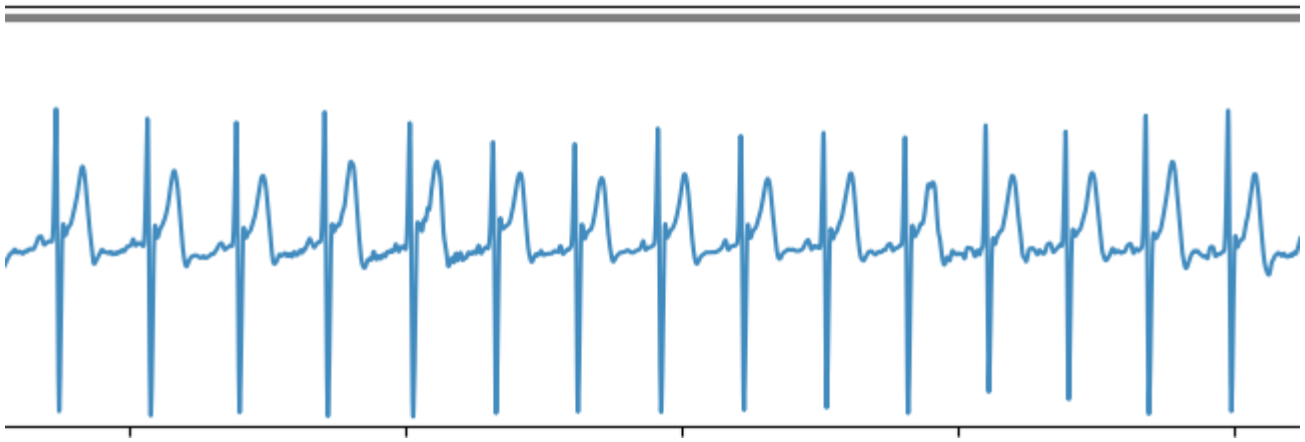


DATFLOW TO THE GAP9



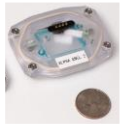
TOWARDS GENERATIVE (DENOISING) PHYSIOLOGIC SIGNALS

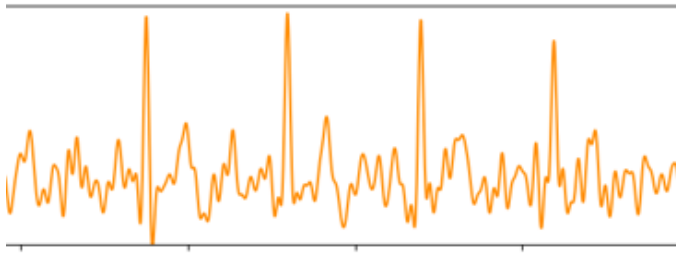
Real-time generative denoising of signals capture in free-living environments

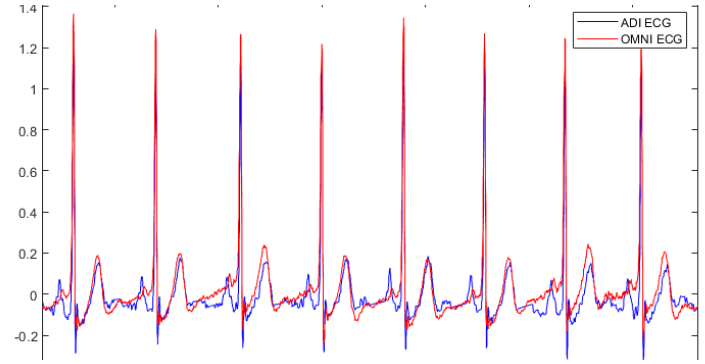


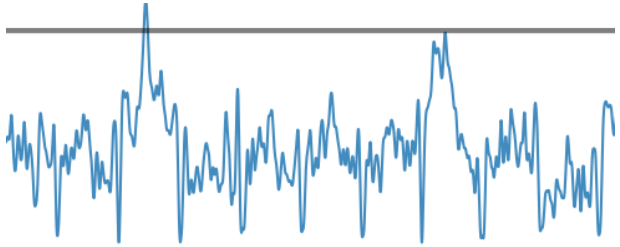
FREE-LIVING, FREE-ROAMING, HIGH-ACTIVITY

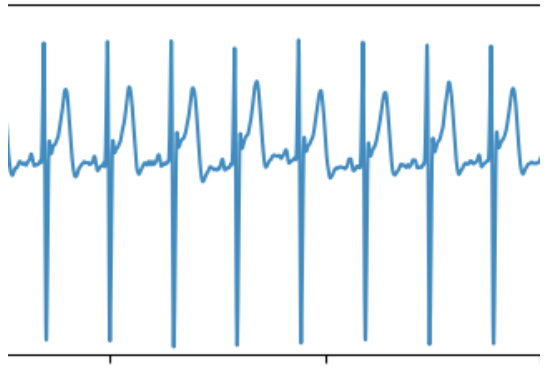
"Clipart courtesy FCIT"







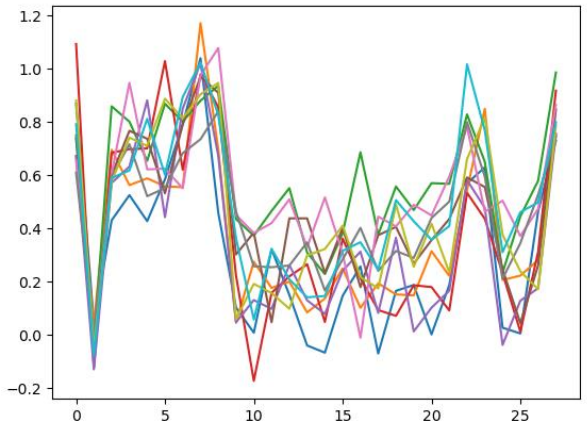




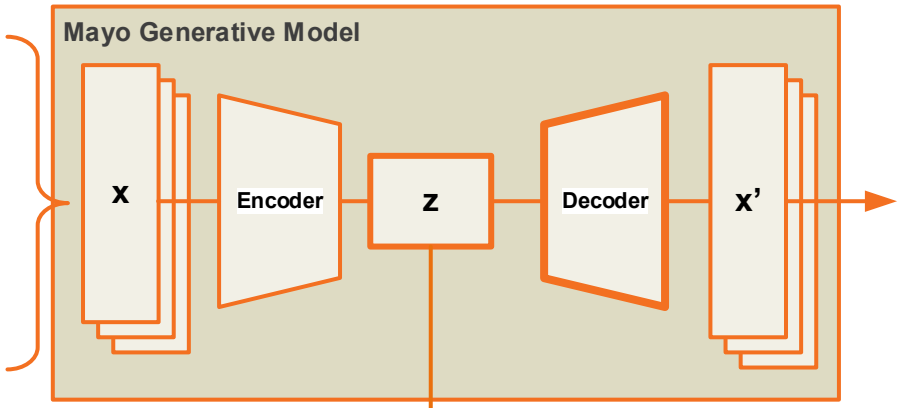
GENERATIVE (DENOISING) MODELS

Model creation to generatively create the full-waveforms from low-fidelity high artifact physiologic signals

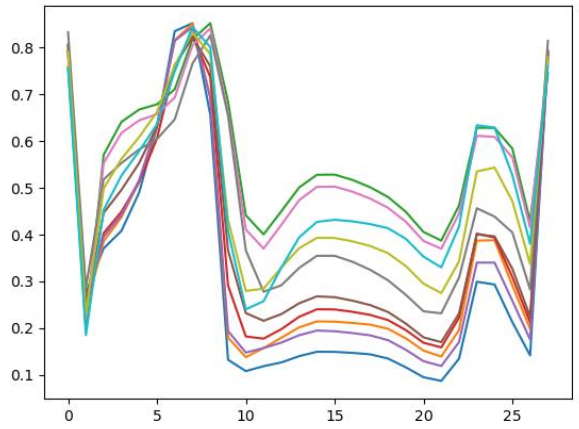
Multimodal / Multivariate
Inputs omni-wearable sensor stream inputs



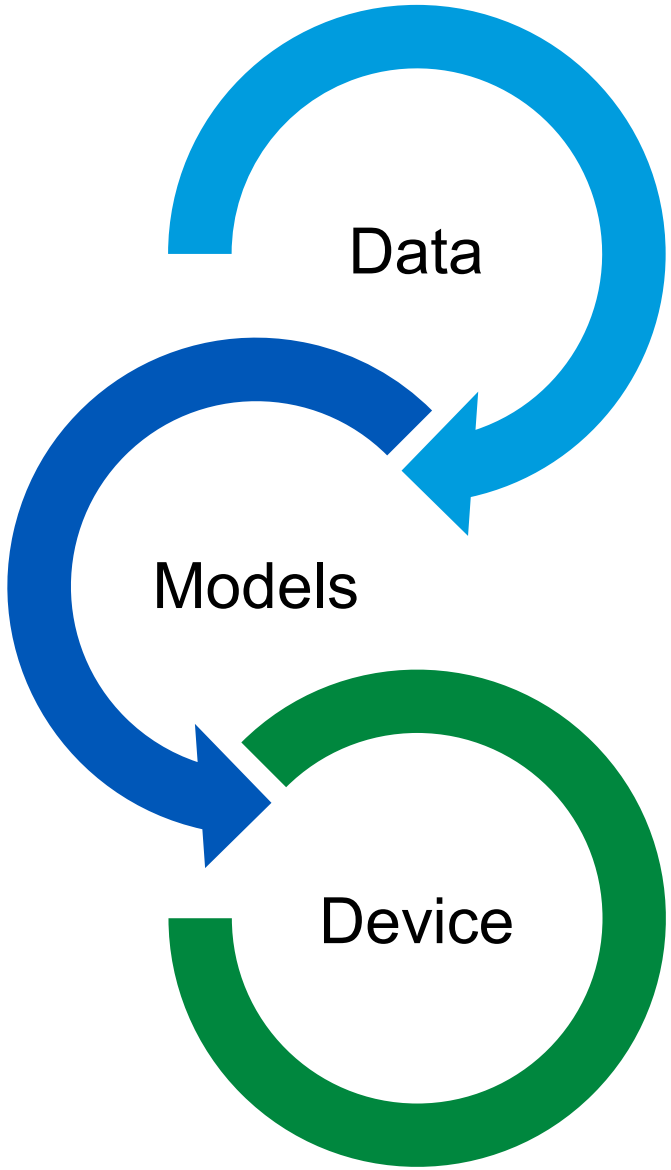
The **generative model** will be the decoder of the autoencoder structure



Regenerative / Generative Signals



Description representation parameters



CONCLUSION

DATA, MODELS, WEARABLE DEVICE

Building Physiologic Datasets

- Hypothesize Biomarkers
  - Design study to induce condition
  - Record physiologic signals for emulated conditions

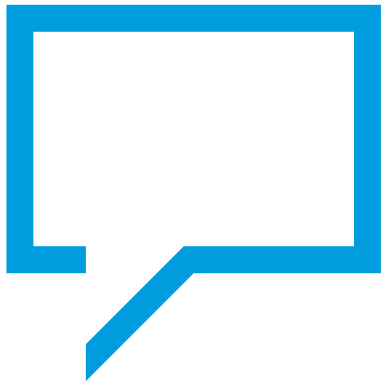
Build Models from Signals

- Clinical reference for conditions
- Unbounded resource models

Wearable Device

- Lightweight and low-energy (tinyML)

QUESTIONS & ANSWERS



Copyright Notice

www.tinyml.org