UNIVERSITÉ DE SHERBROOKE

RAMS

# Neural Signal Compression System for a Seizure-Predicting Brain Implant in CMOS 28nm

FastML Workshop, October 3, 2022

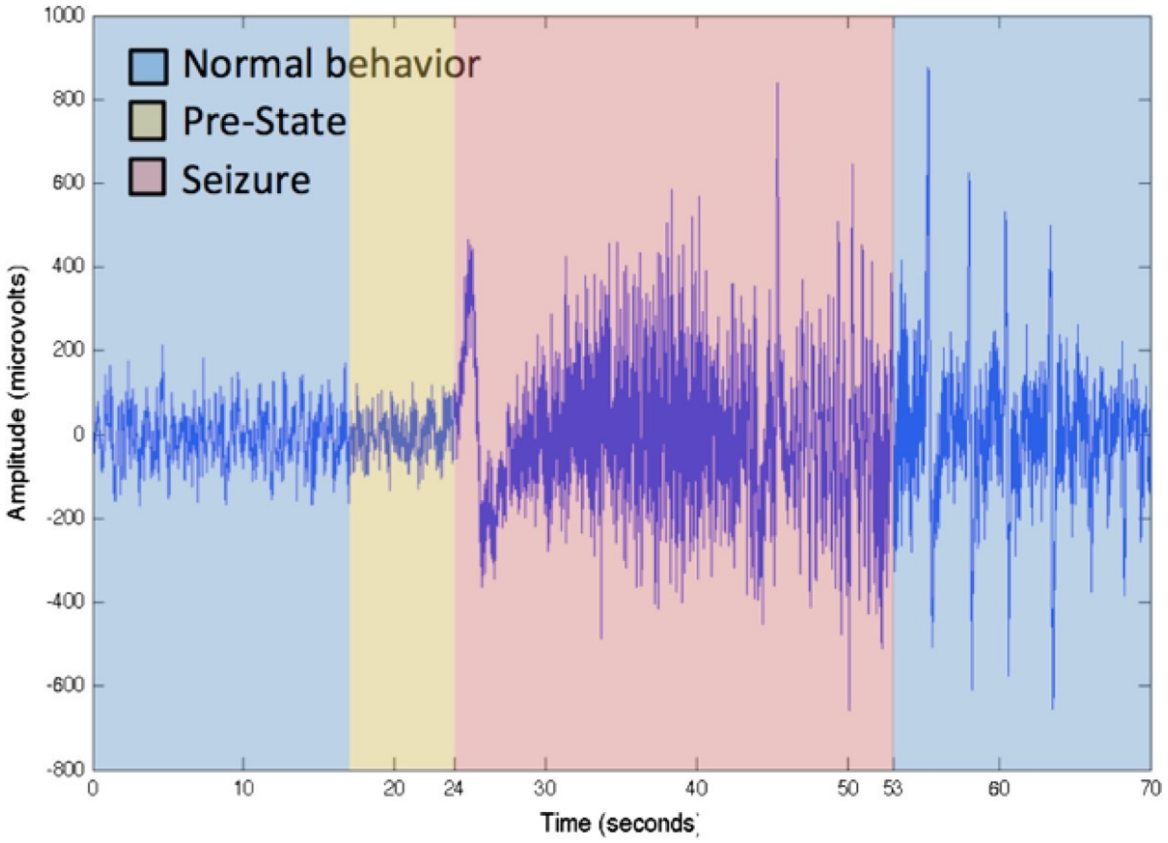
*William Lemaire ; Berthié Gouin-Ferland ; Vincent Gauthier ; Esmaeil Ranjbar Koleibi ; Montassar Dridi ; Maher Benhouria ; Konin Koua ; Takwa Omrani ; Sébastien Roy ; Réjean Fontaine*

CMC MICROSYSTEMS

CRSNG NSERC

# Epilepsy

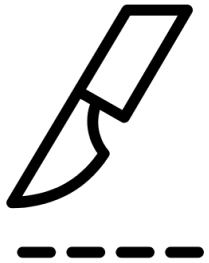
**Prevalence:** ~1% worldwide [2]



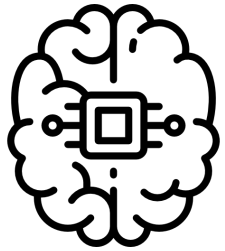

[1] Turkey N Alotaiby, Eurasip J. Adv. Signal Process, 2014

# Treatment of Refractory Epilepsy

**Antiepilleptic drugs**
**70%**
[1] Mogul et al., Annu Rev Biomed Eng, 2014



**Resective brain surgery**
**1%**
[2] Engel et al., Curr Opin Neurol 2018



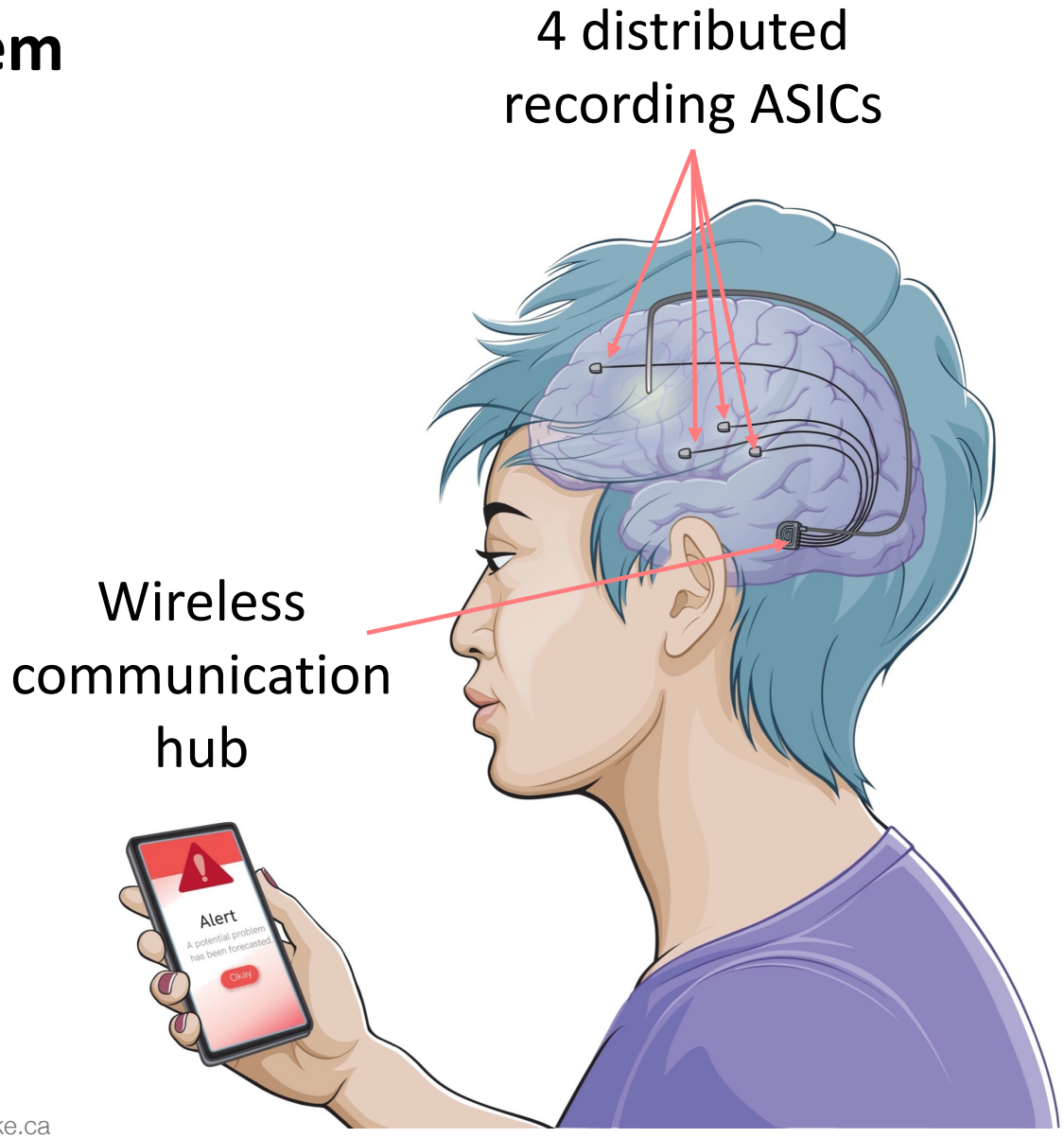
**Implant-based seizure prediction**

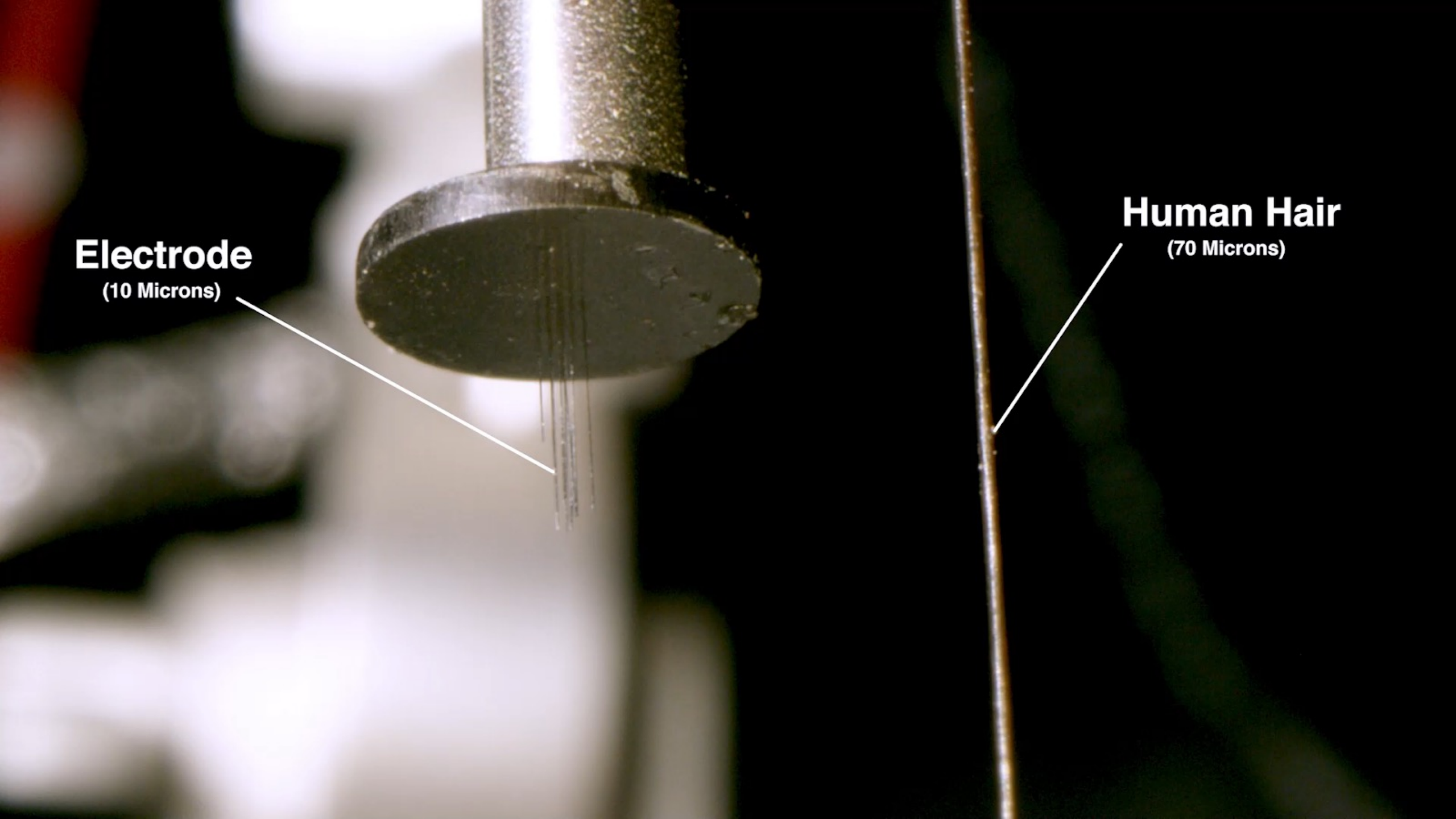
# Proposed system

## Goal

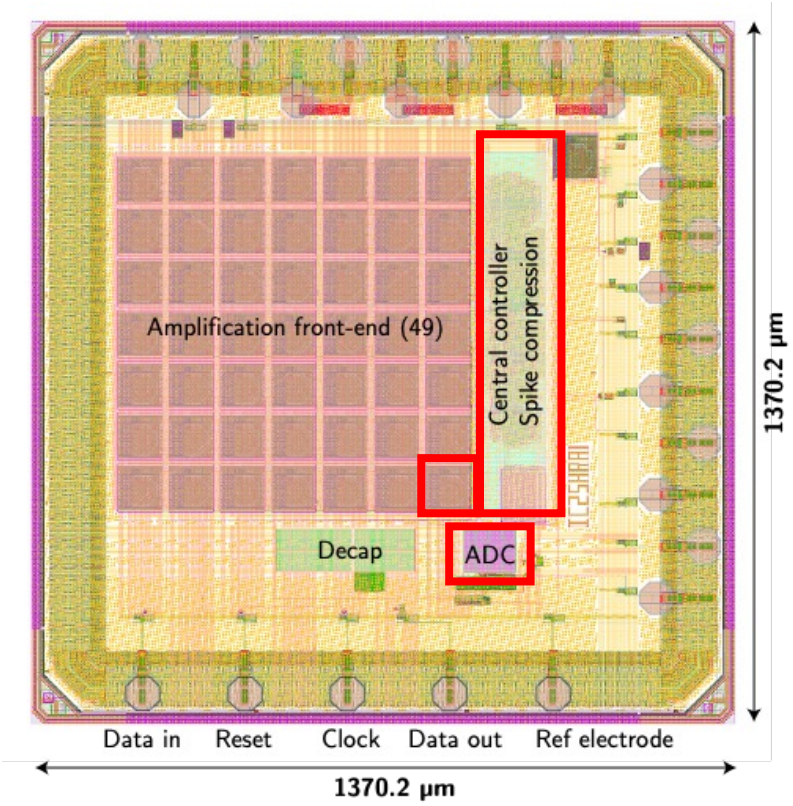
Forecast **1 day** in advance [1]

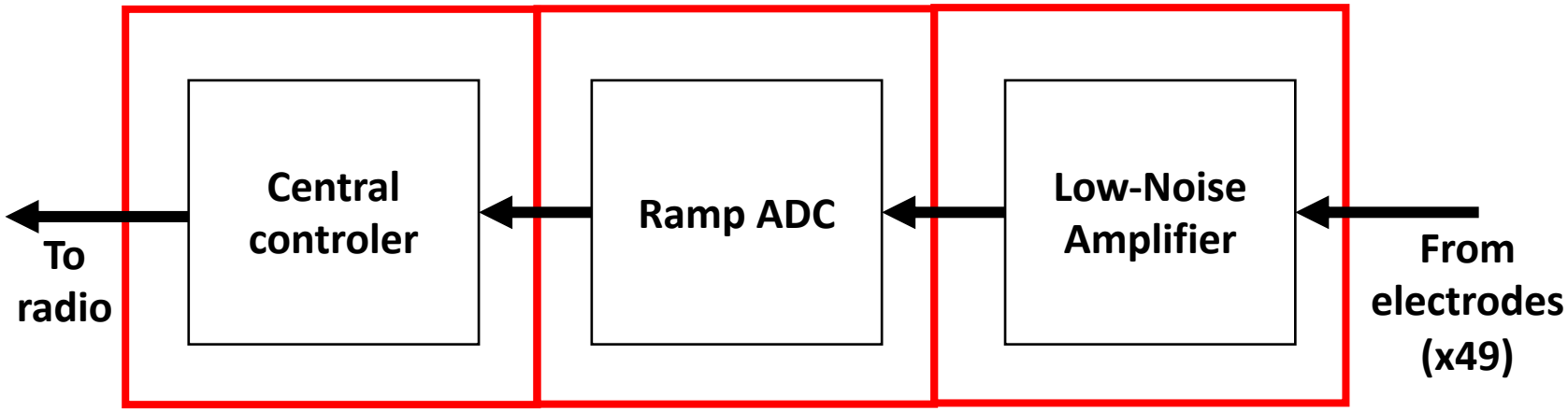
[1] Maturana et al., Nature, 2020




Electrode
(10 Microns)
Human Hair
(70 Microns)

# 28-nm CMOS ASIC




28 nm CMOS
49 electrodes, 90 µm pitch

# Design Challenge

## Power

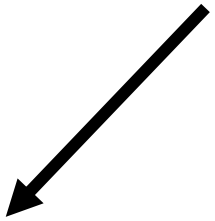
< 1 mW
(49 electrodes)

## Area

< 2 x 2 $mm^2$

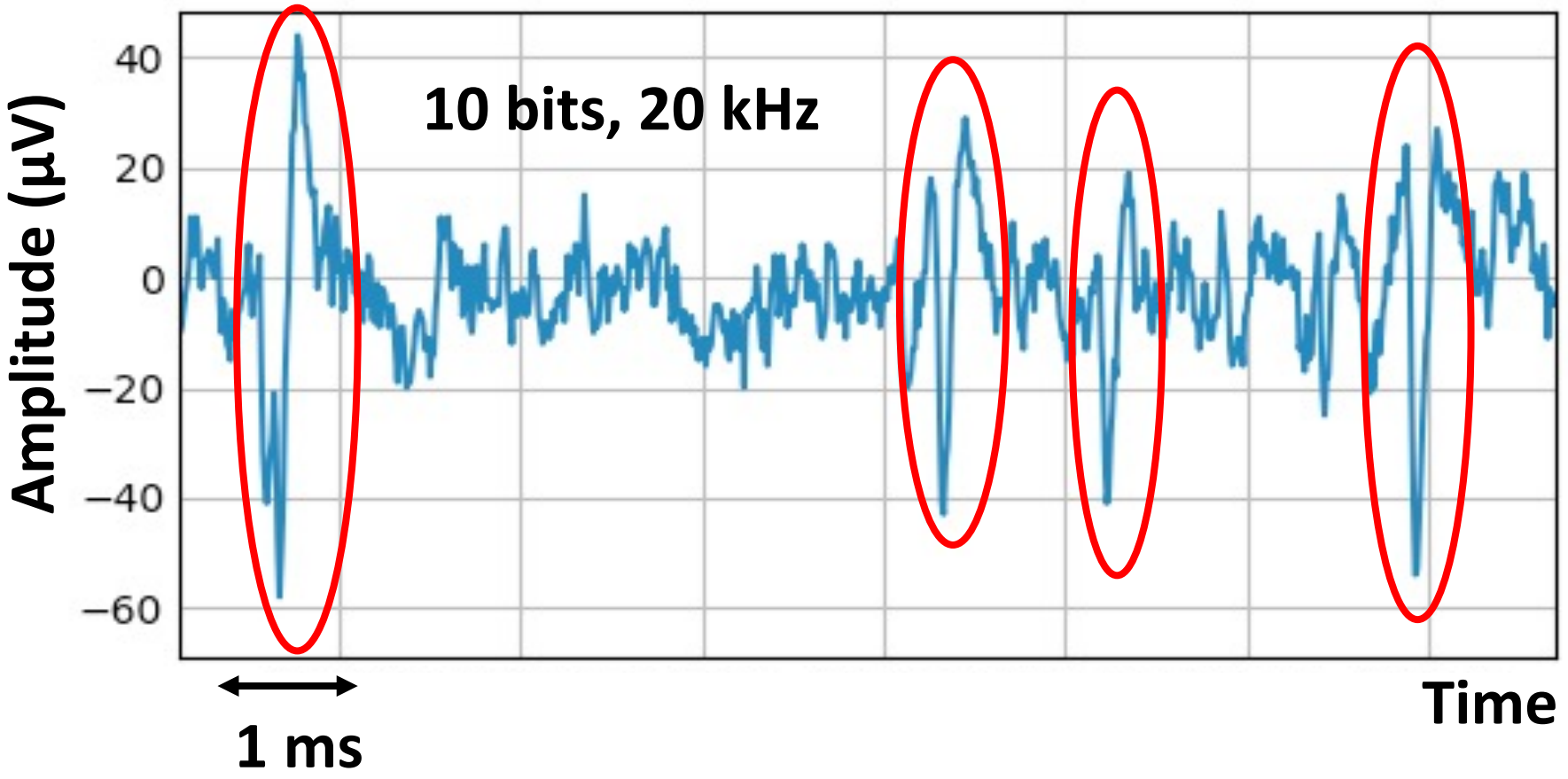
## Data Rate

Bluetooth low energy:
< 2 Mbit/s (~20 mW)

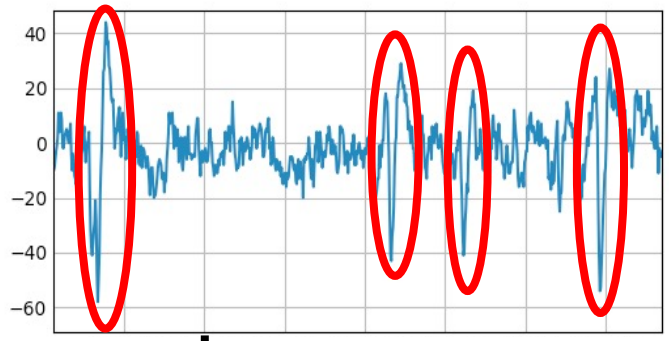


**The problem**
4 ASICs x 49 electrodes x 20 kHz x 10 bits =
40 Mbit/s

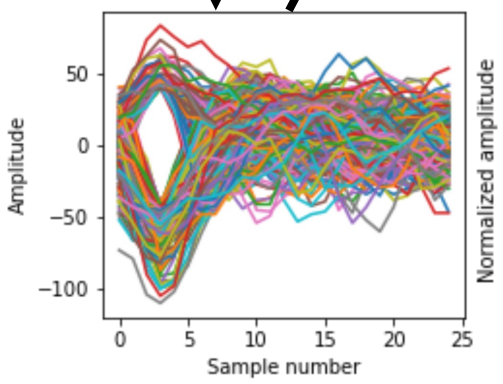
# Solution: Signal Compression

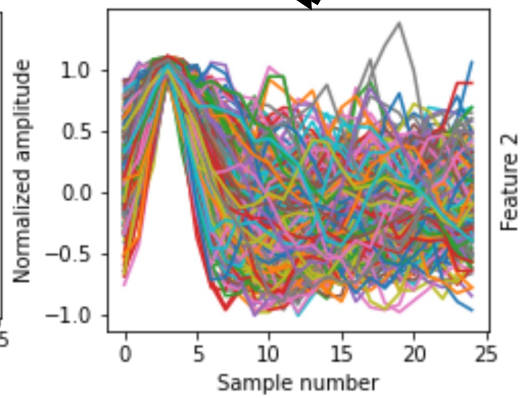


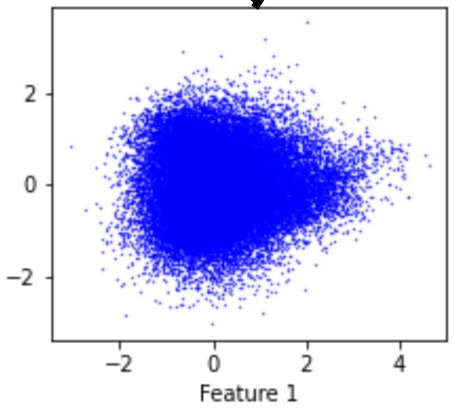
# Proposed algorithm

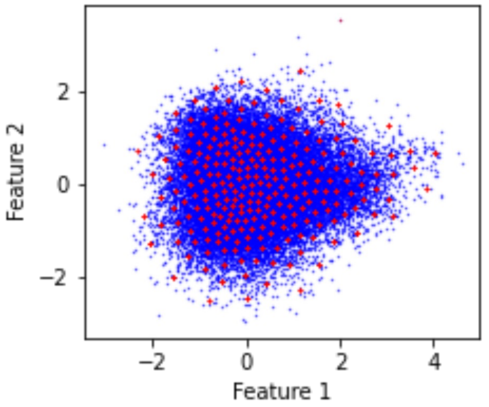


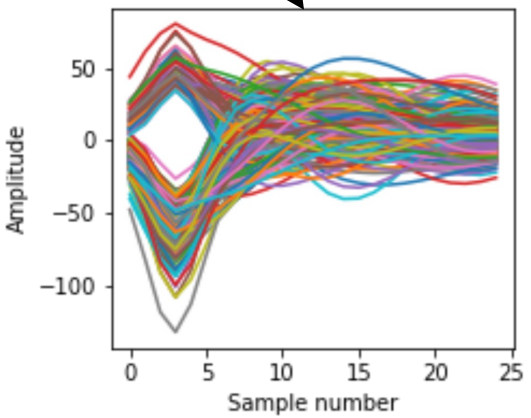

Detection → Normalization → Feature extraction → Quantization → Reconstruction



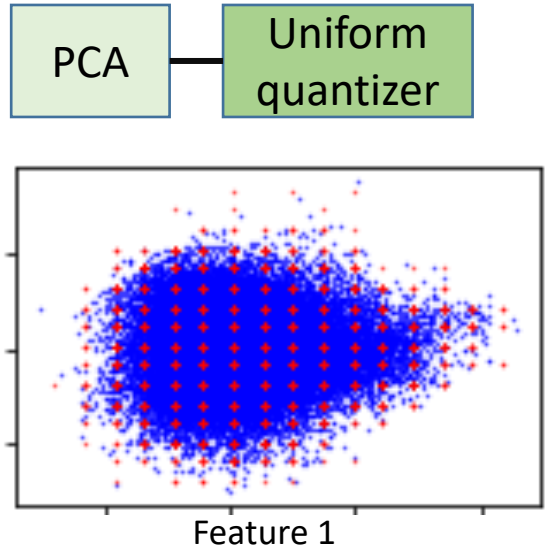



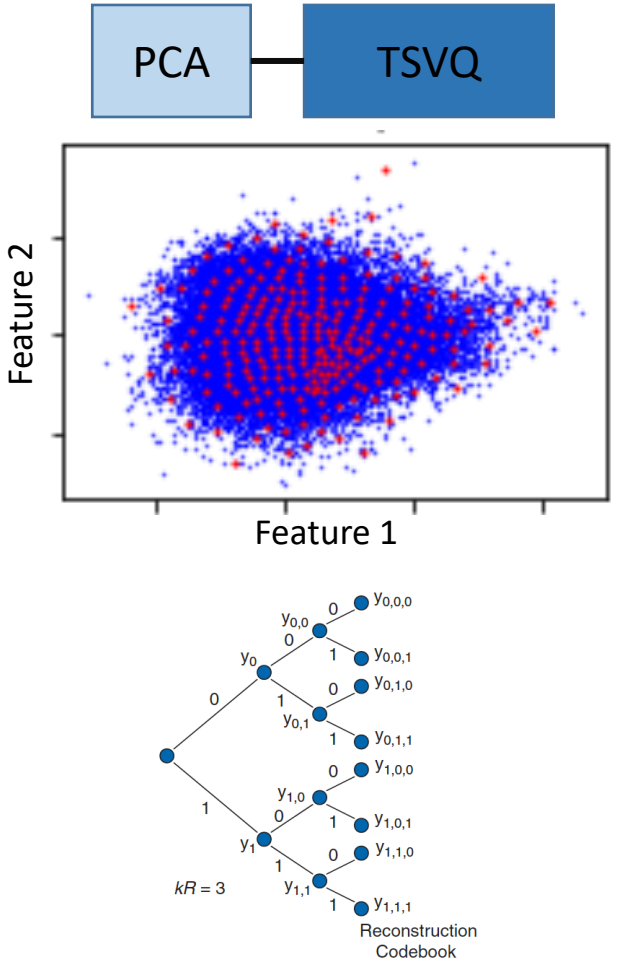



# Compare algorithms

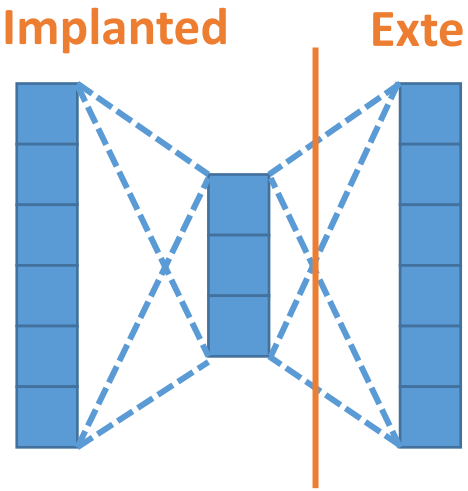
## Principal Component Analysis + Uniform Quantizer

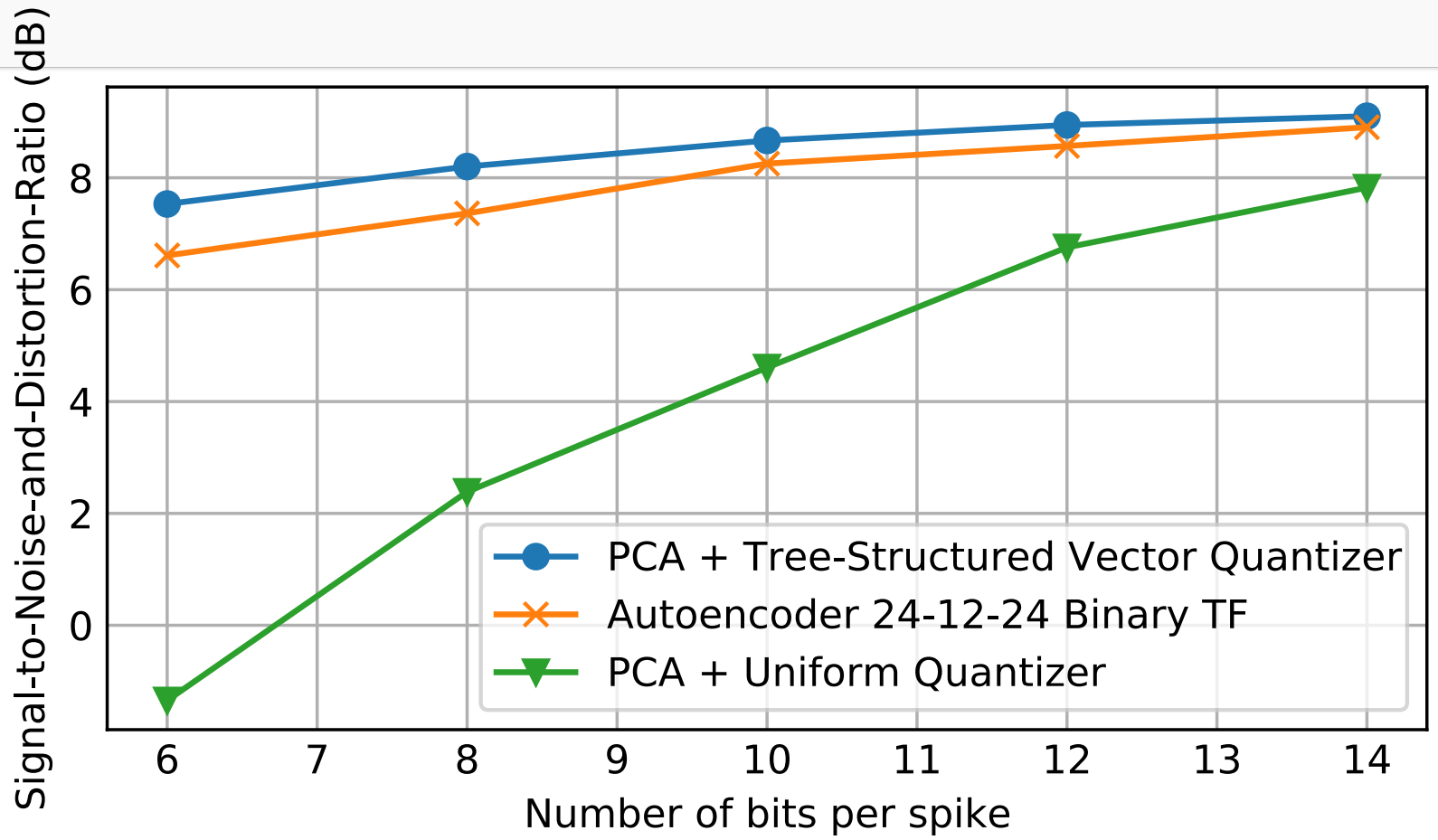



## Principal component Analysis + Tree-structured Vector Quantizer (TSVQ)



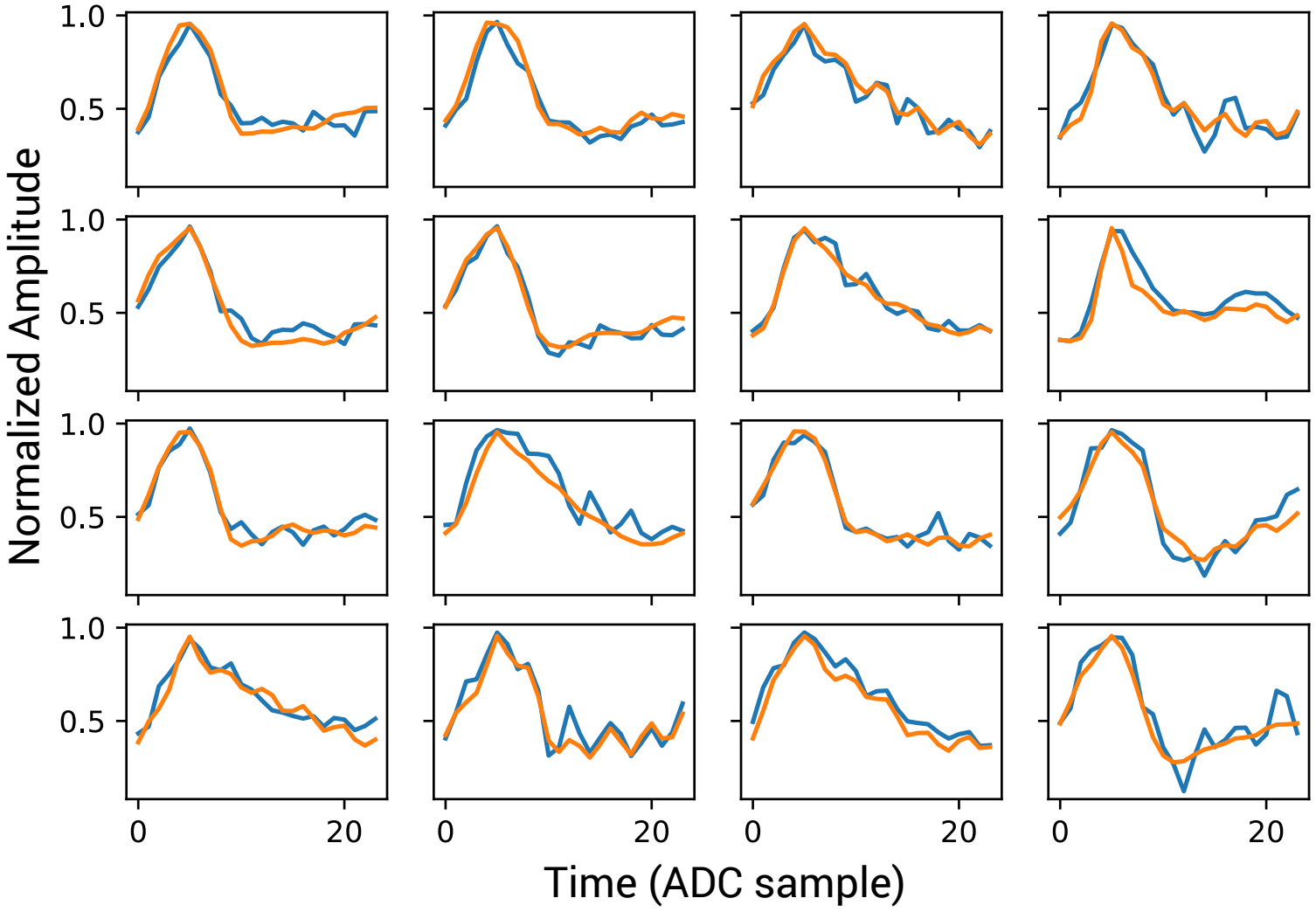

## Quantized Autoencoder






| | Memory usage | Mult per Spike | SNDR @ 10-bits |
|---|---|---|---|
| PCA4 + Uniform quantizer | 0.6 kb | 96 | 2.2 |
| Autoencoder 24-12-24 | 1.1 kb | 288 | 7.2 |
| PCA4 + Tree-Structured Quantizer | 393.0 kb | 192 | 8.1 |

# Autoencoder Compressed Spikes Examples



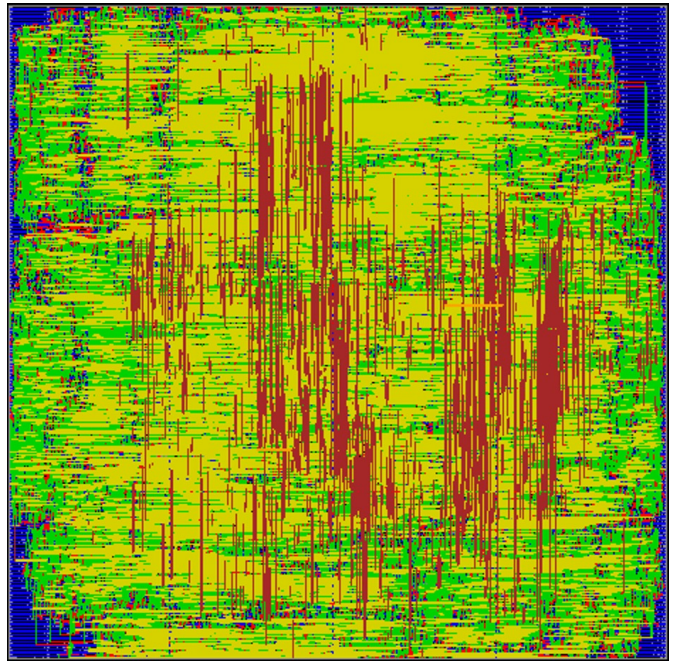

Original spike
Compressed spike

# Preliminary HLS4ML implementation + Catapult

## Area

< 2 x 2 mm$^2$

**Autoencoder layout in 28 nm**




## Power

< 1 mW
(49 electrodes)

**Autoencoder power in 28 nm**

| | Power (µW) |
|---|---|
| **Dynamic** | 31 |
| **Leakage** | 107 |
| **Total** | **138** |

## Data Rate

< 2 Mbit/s (~20 mW)

| | Data Rate | Compression Ratio (CR) |
|---|---|---|
| **Raw Signal** | 40 Mbit/s | - |
| **Raw Spikes** | 2.4 Mbit/s | 16 |
| **Compressed Spikes** | 0.3 Mbit/s | 125 |

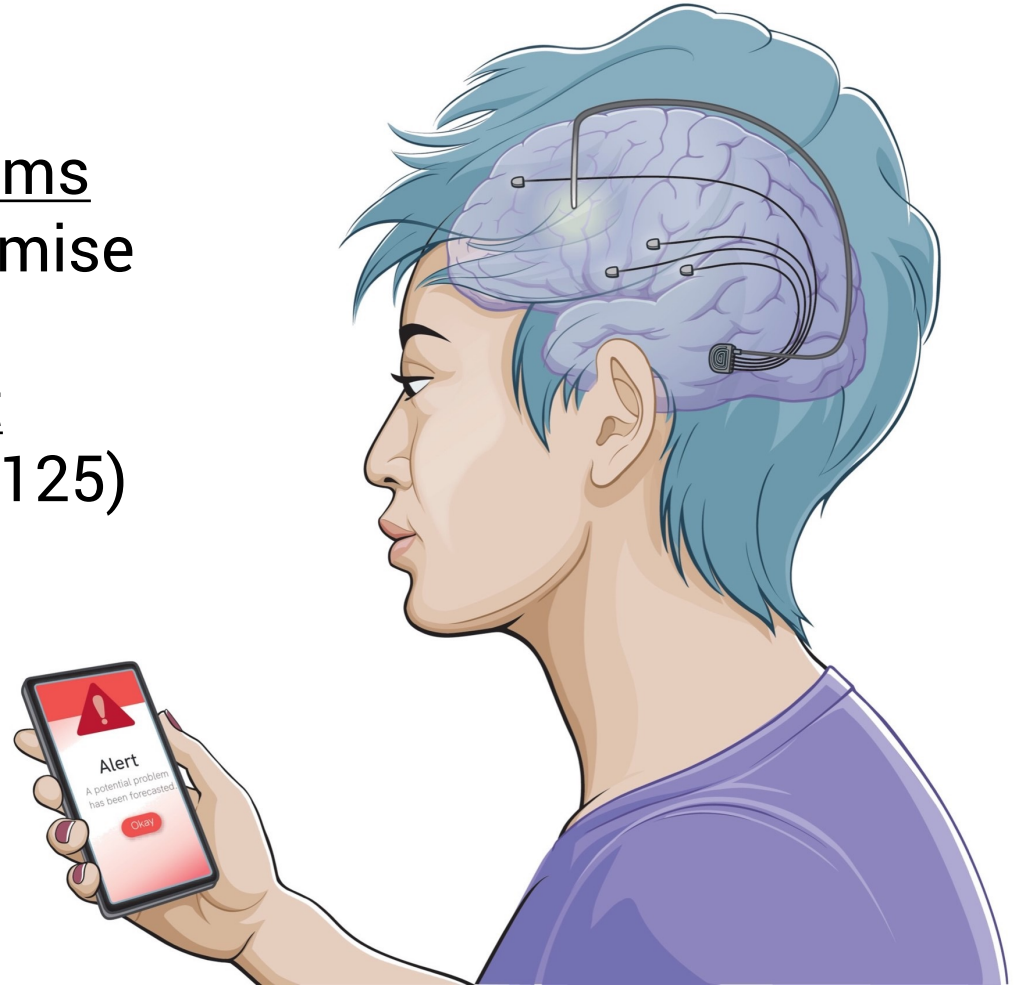
# Conclusion

1. Developing a seizure-predicting implant
   - Bandwidth problem

2. Compared 3 neural spike compression algorithms
   - Quantized autoencoders offer best compromise

3. Implemented in 28nm with HLS4ML + Catapult
   - Achieve specs (< 10 mW, < 2 x 2 $mm^2$, CR ~125)
   - 2 weeks from idea to layout

4. Next steps: demonstrate on silicon



# Thank you!


# Dataset

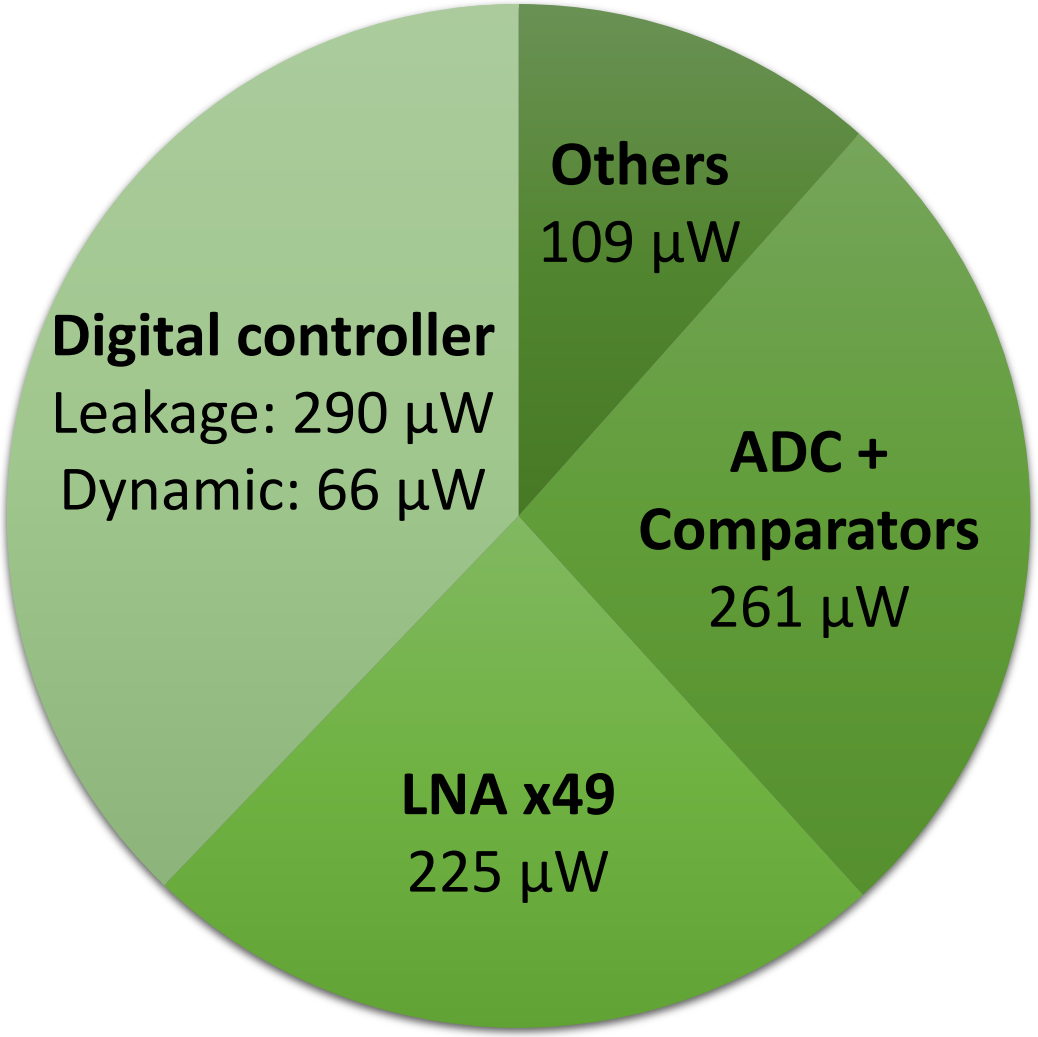
- 4 electrodes probe in CA1 region of rat hippocampus
- 12 x 600s recordings
- 263 300 spikes
- 75/25 train/validation

# Bandwidth calculation

- Raw: 49 electrodes x 4 ASICS x 20 kHz x 10 bits/sample = 39 200 kbit/s
- Spikes: 49 electrodes x 4 ASICS x 50 Spikes/s/electrode x 24 samples/spike x 10 bit/sample = 2 352 kbit/s


william.lemaire@usherbrooke.ca

# Power Consumption




**ASIC Total**
828 µW