Noema

Hardware-Efficient Template Matching for Neural Population Pattern Detection

Ameer Abdelhadi Eugene Sha Ciaran Bannon

Hendrik Steenland Andreas Moshovos

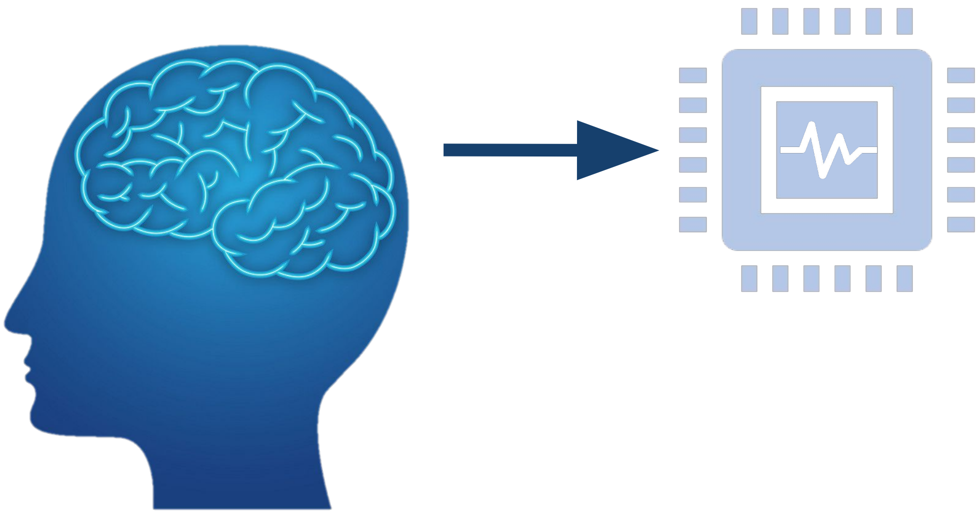
The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

NeuroTek
Innovative Technology

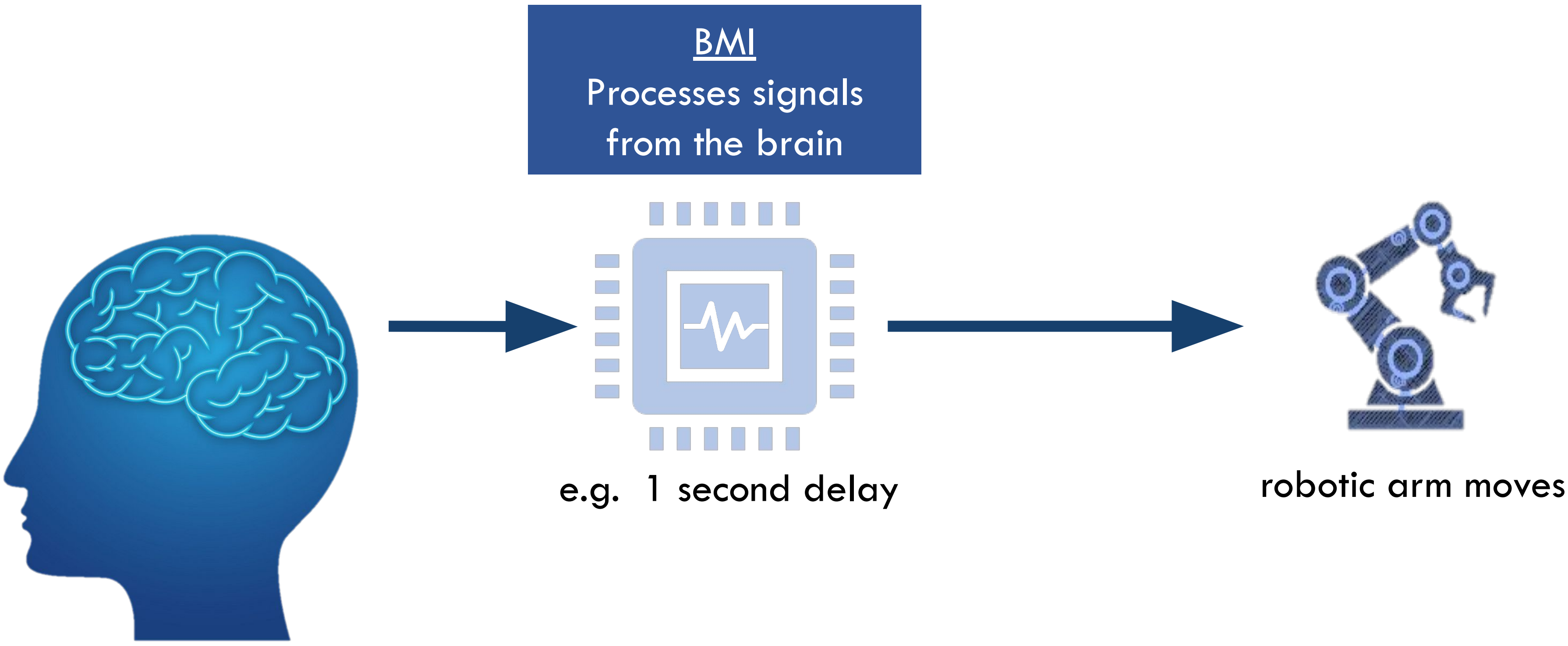
Brain Machine Interfaces (BMIs)

BMI

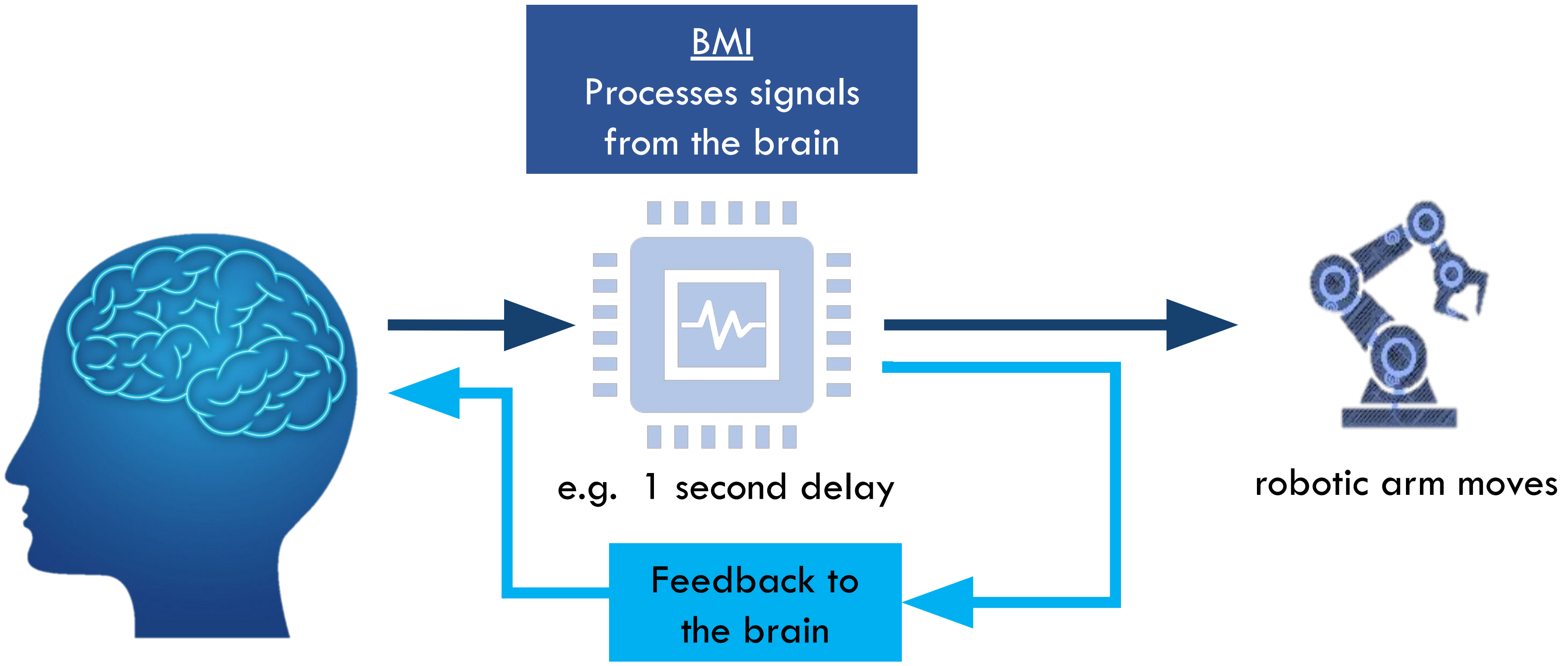
Processes signals from the brain



Brain Machine Interfaces (BMIs)



Brain Machine Interfaces (BMIs)

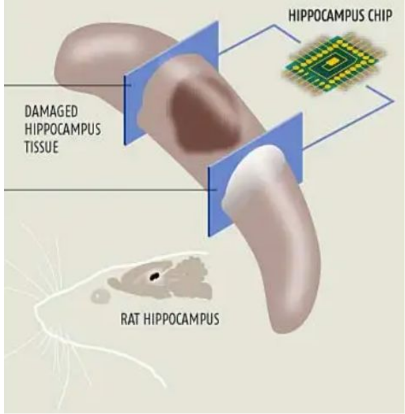


Applications of Brain-machine Interfaces

Repair Brain Function

Interface brain regions which no longer connect,

e.g., *Alzheimer's*



Replacement of damaged hippocampus with a chip [1]

Drive Effectors

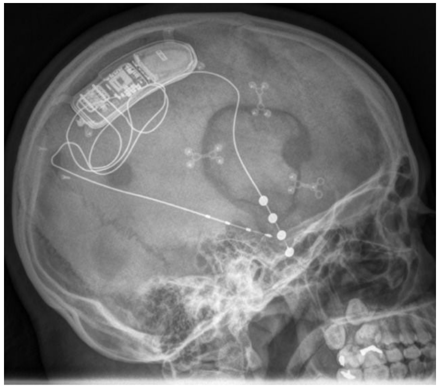
Greater accuracy and dexterity,

e.g., robotic limbs

Woman controls robotic arm with 100-channel Utah array [2]

Anticipate & prevent harmful neural activity

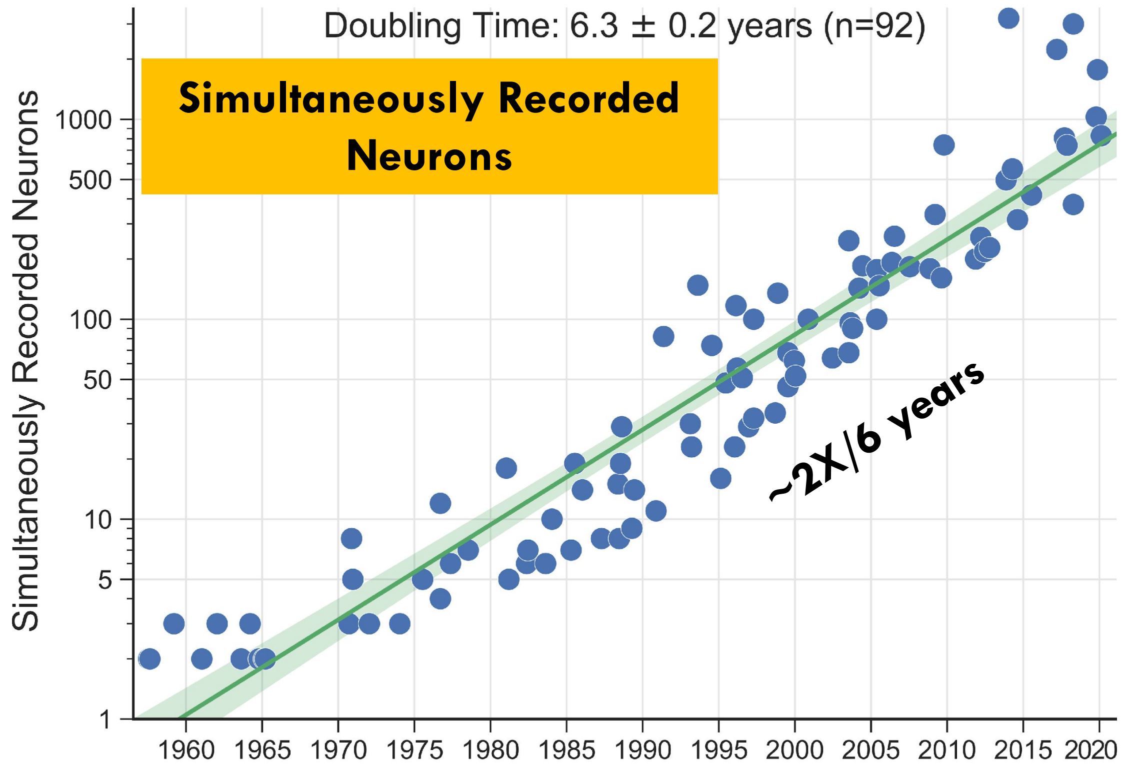
e.g., *epilepsy*



Responsive neurostimulator system for epilepsy [3]

[1] https://www.newscientist.com/article/dn3488-worlds-first-brain-prosthesis-revealed/ (Hippocampus repair)
[2] https://continuum.utah.edu/web-exclusives/the-bionics-man/ (Utah Array)
[3] Critical review of the responsive neurostimulator system for epilepsy (Thomas and Jobst, 2015)

The Challenge and Opportunity for Architecture: Capture Capability Growing Exponentially

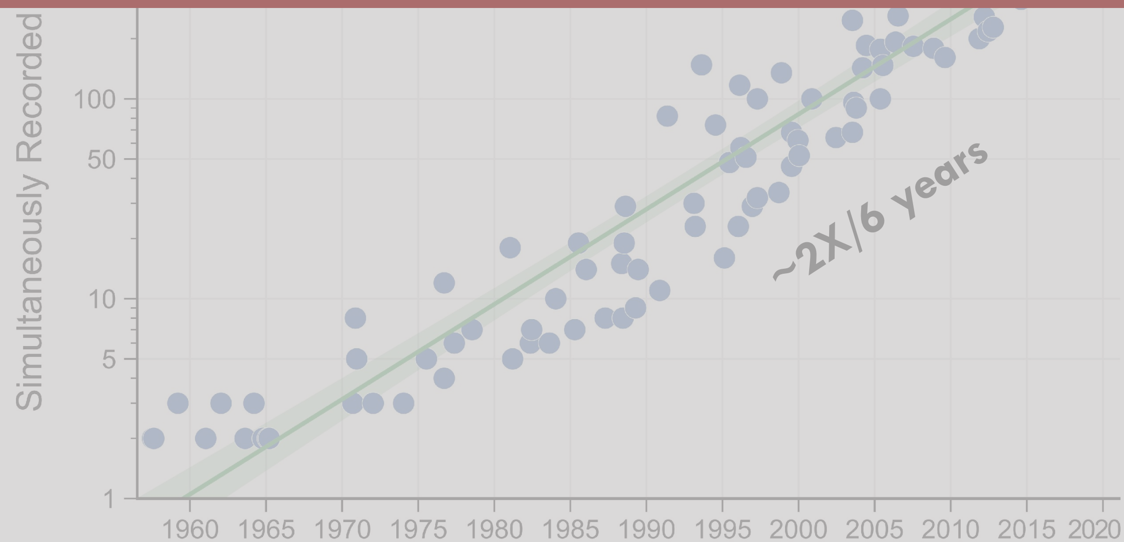


Constraints for a ***portable implanted device***

1. Fast (real-time, <5ms detection latency)
2. Low-power & low-area
3. Scalable

The Challenge and Opportunity for Architecture:

Existing solutions can't cope



Constraints for a ***portable implanted device***

1. Fast (real-time, <50ms overall latency)
2. Low-power & low-area
3. Scalable

The Challenge and Opportunity for Architecture:

Existing solutions can't cope

Limited number of neurons

Not real-time

High power

Physically large

Constraints for a ***portable implanted device***

1. Fast (real-time, <50ms overall latency)
2. Low-power & low-area
3. Scalable

The Challenge and Opportunity for Architecture:

Existing solutions can't cope

Limited number of neurons

Not real-time

High power

Physically large

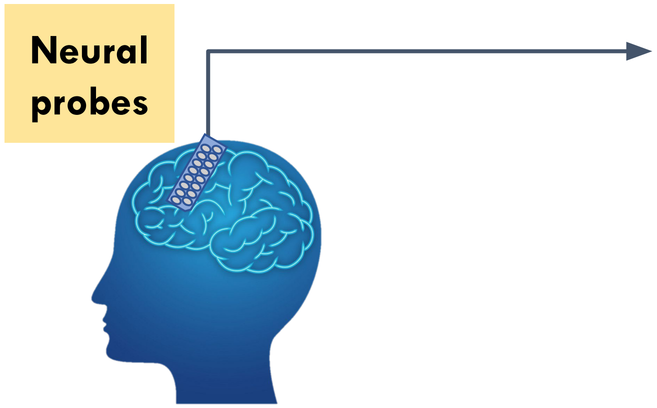
Brain activity decoding is
memory intensive &
computationally expensive

3. Scalable

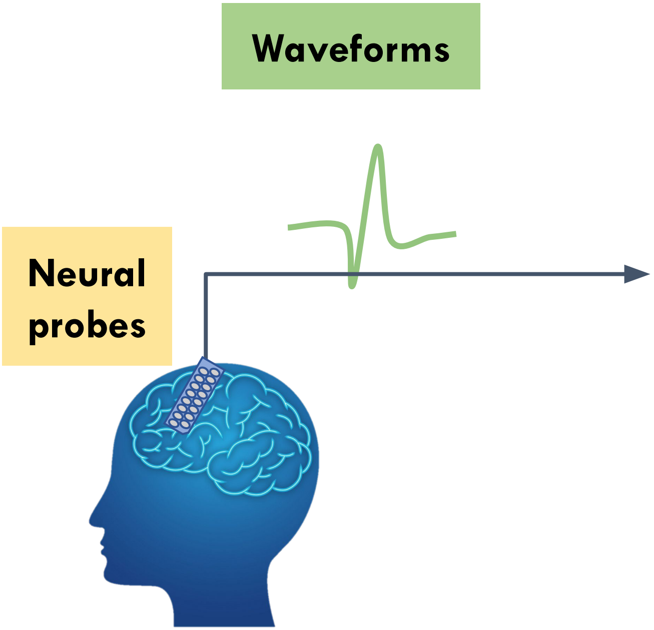
Roadmap

- Input to the system
- Template matching
- Baseline design & Noema
- Results

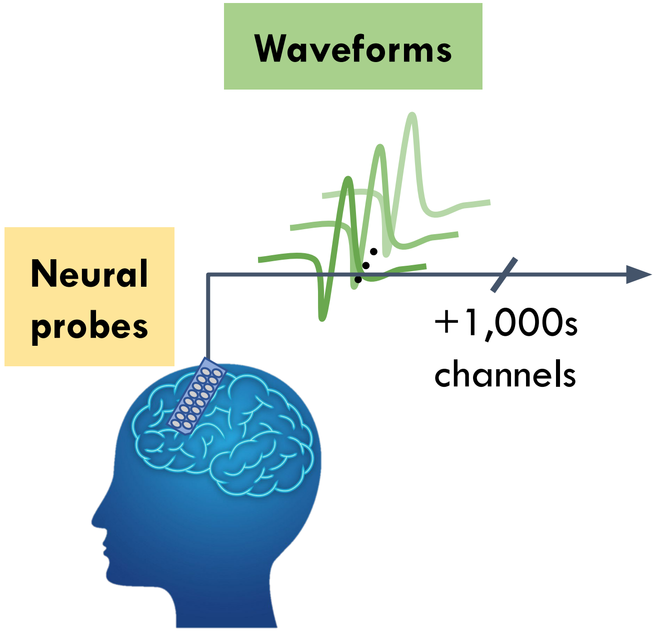
The Raw Input Data



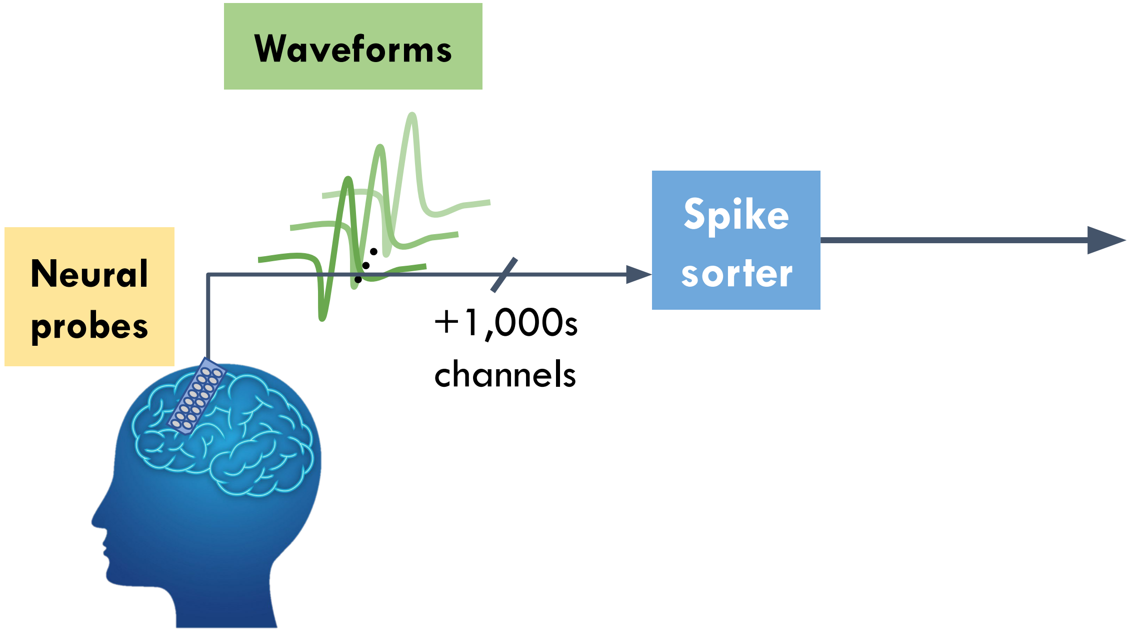
The Raw Input Data



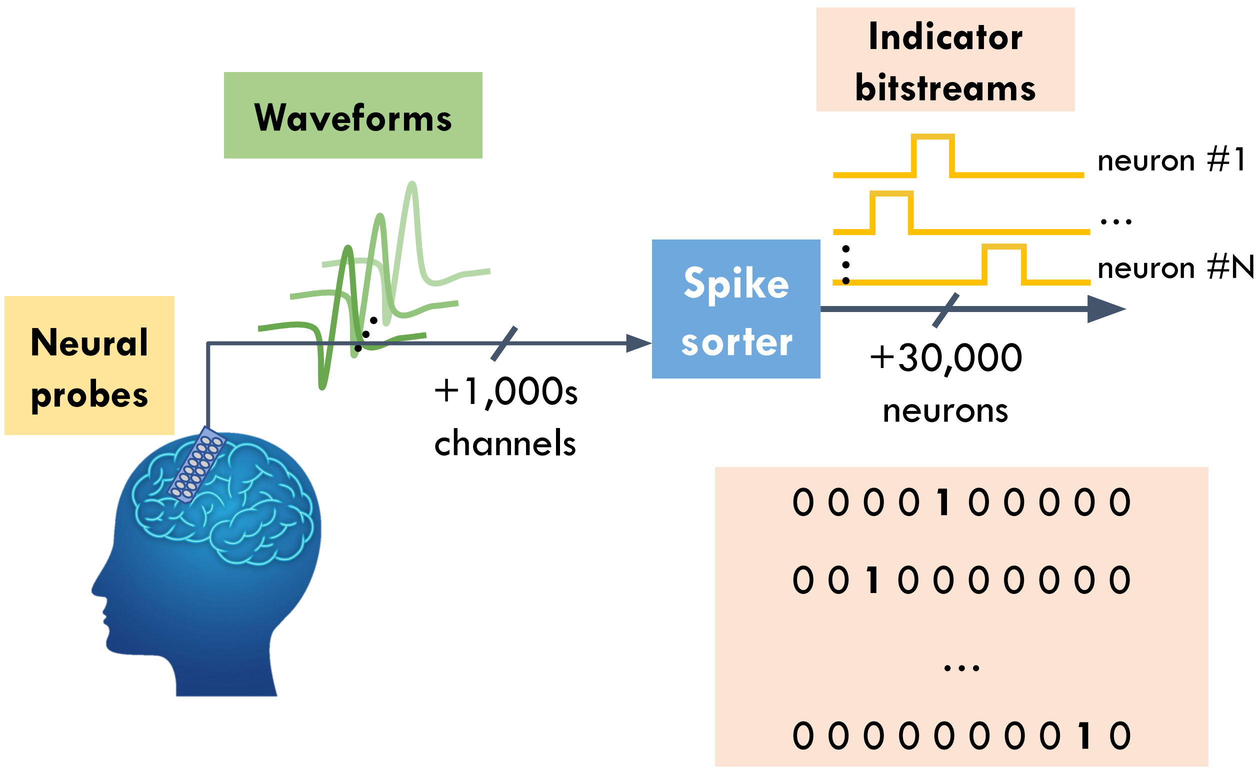
The Raw Input Data



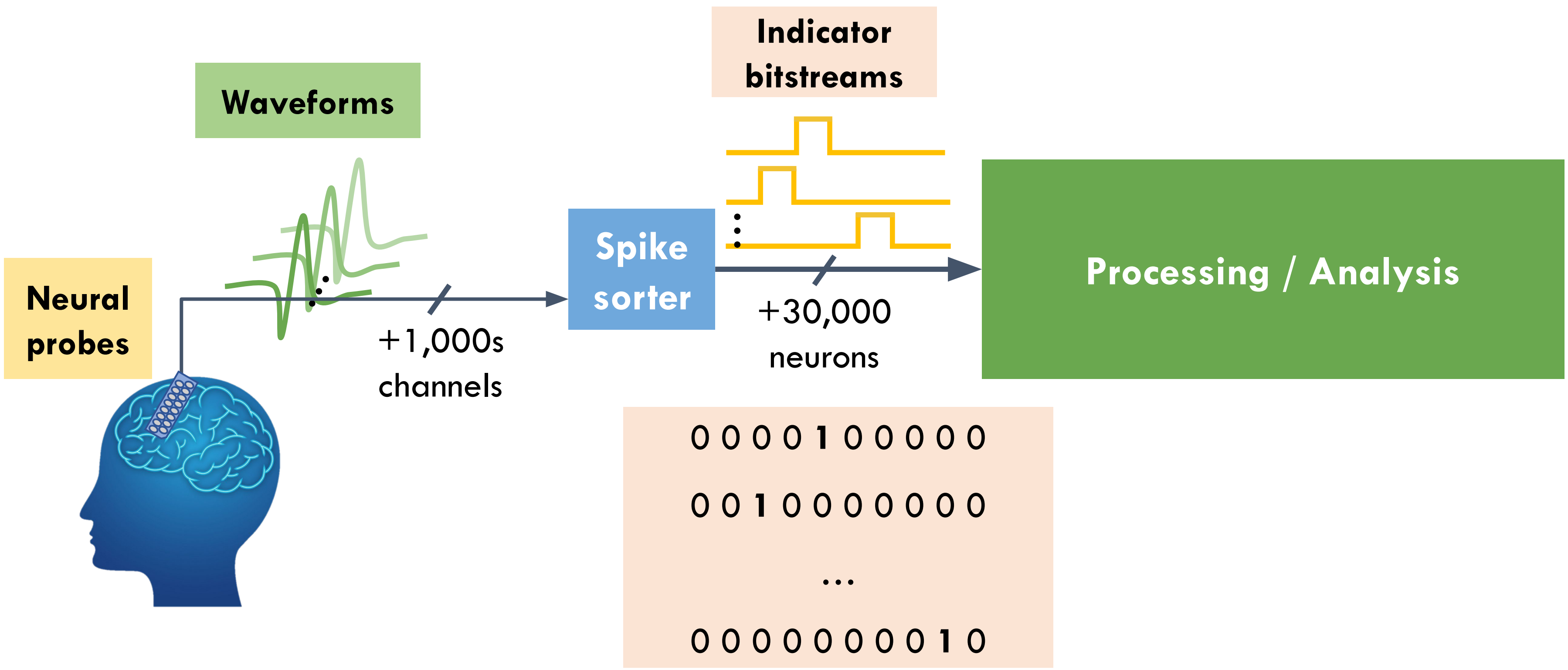
The Raw Input Data



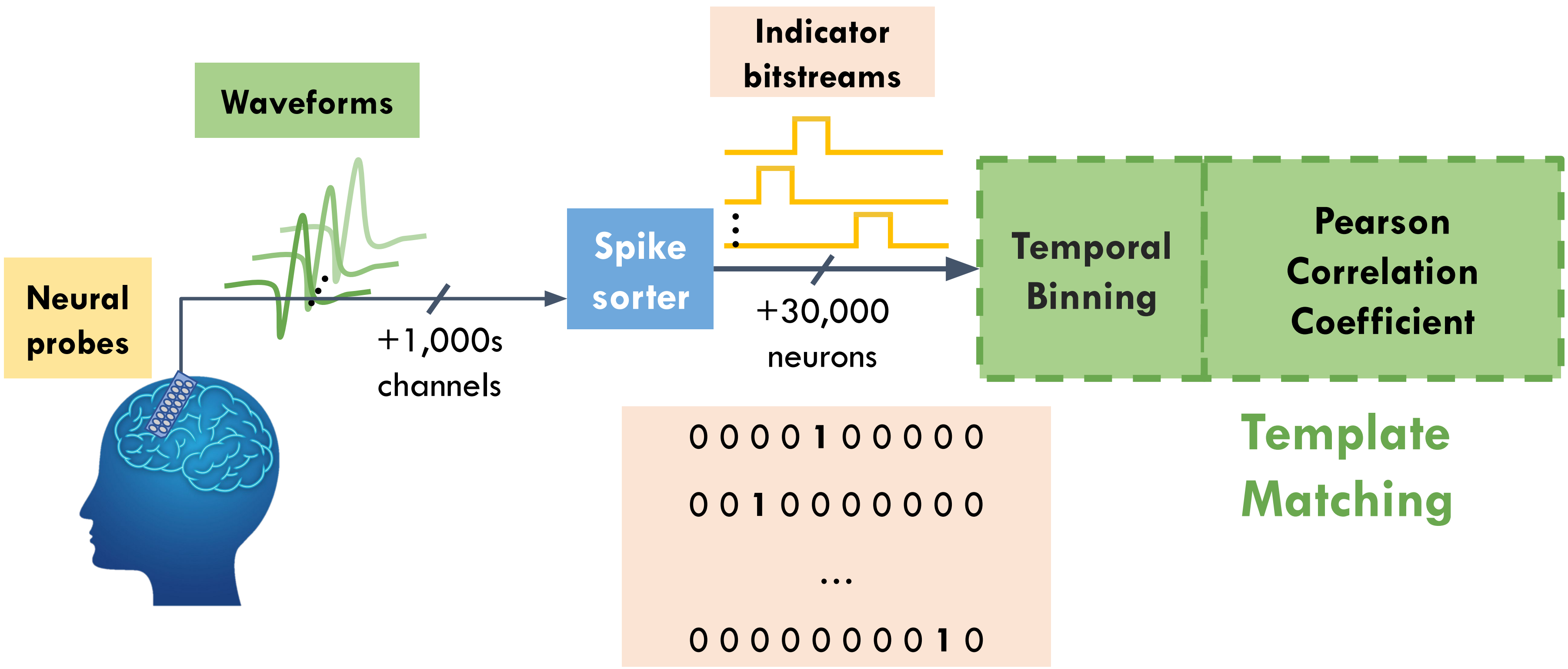
The Raw Input Data



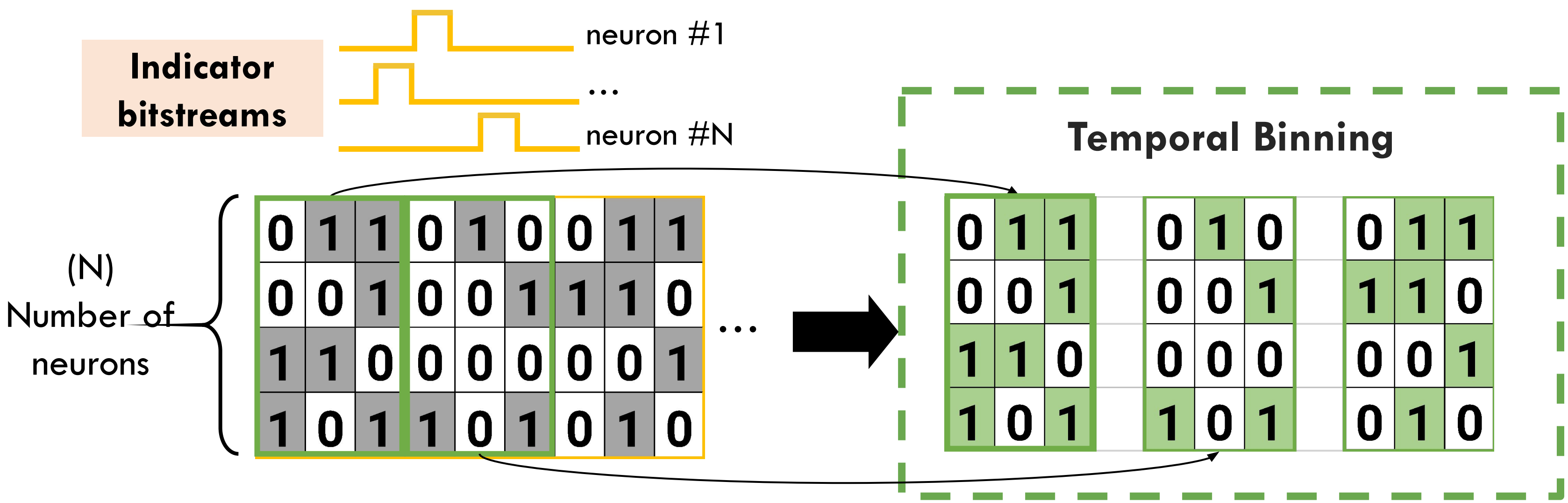
The Raw Input Data



The Raw Input Data

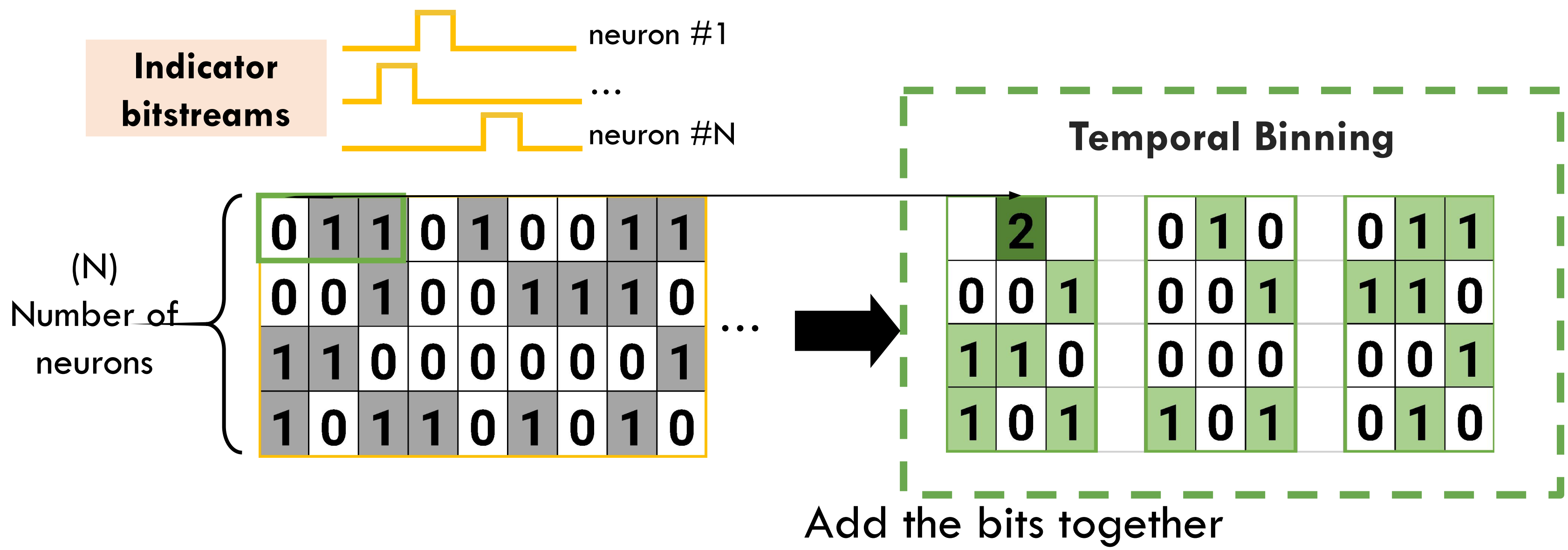


Temporal Binning

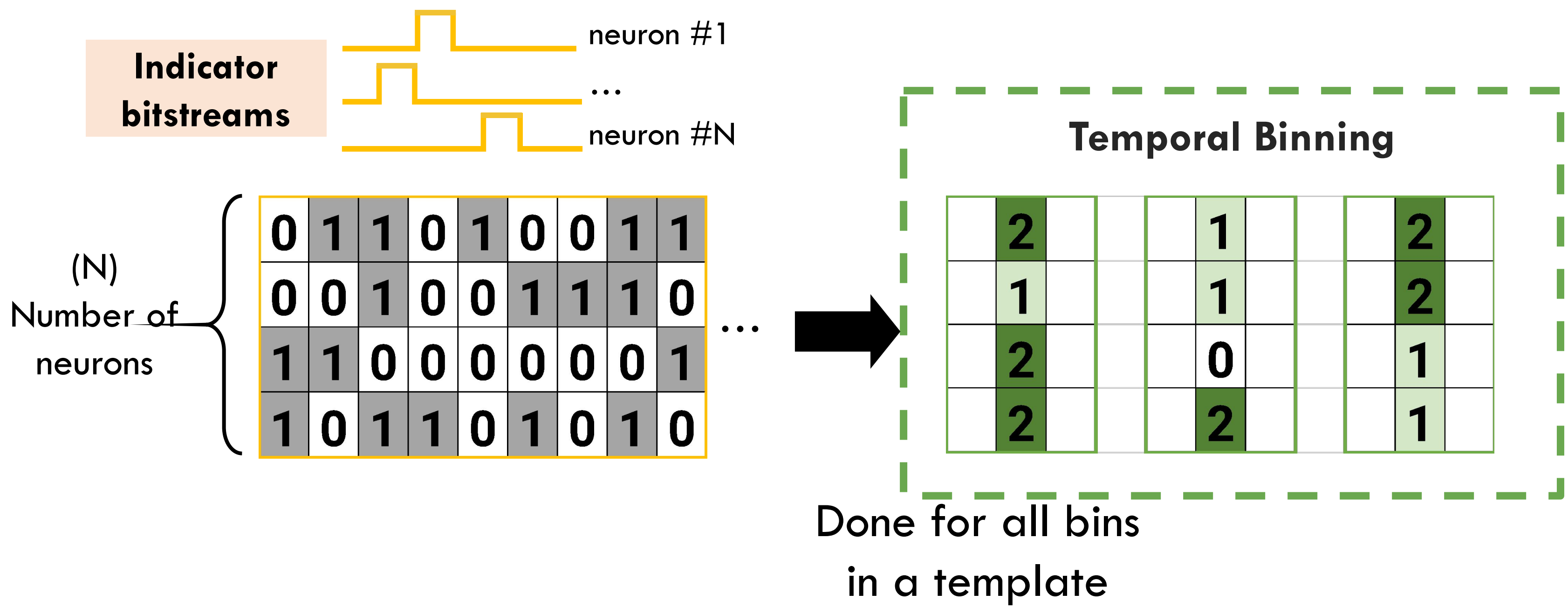


Divide incoming input into groups of 3 (example bin size)

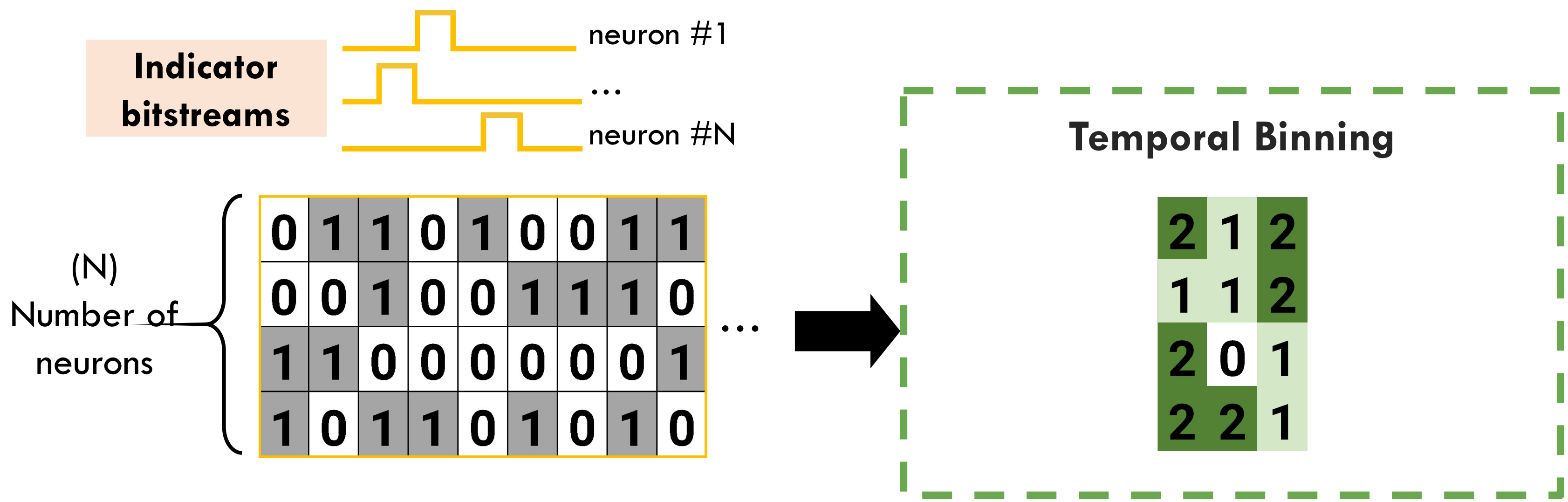
Temporal Binning



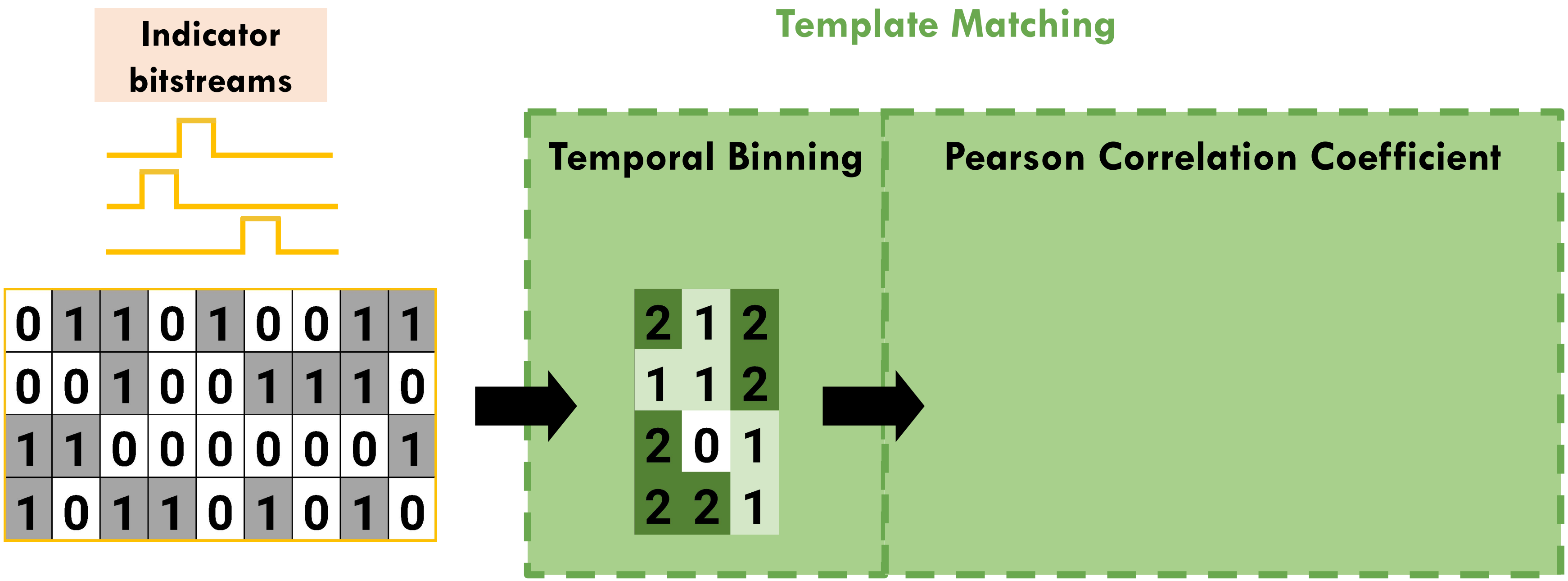
Temporal Binning



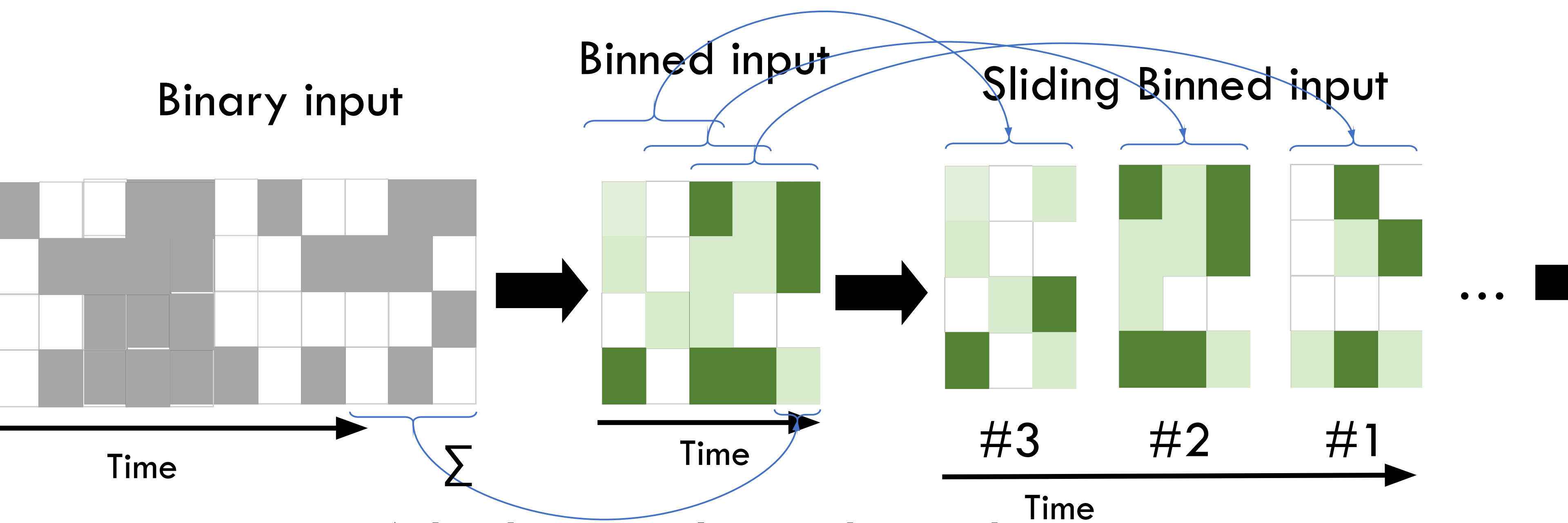
Temporal Binning



Temporal Binning - Overview

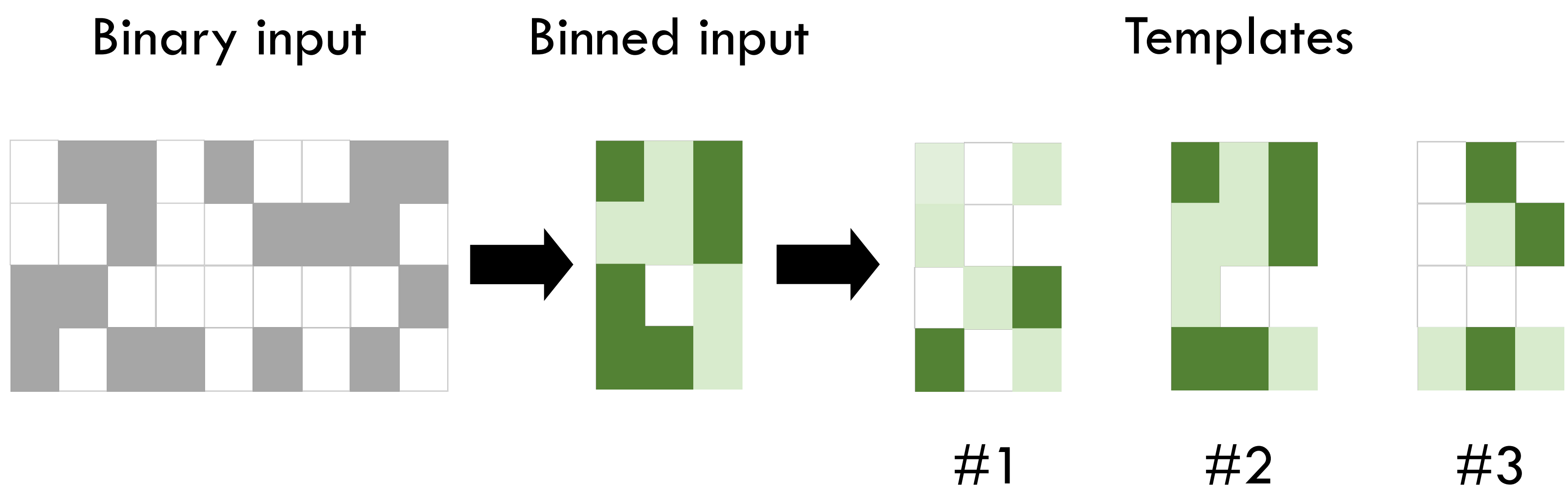


Template Matching



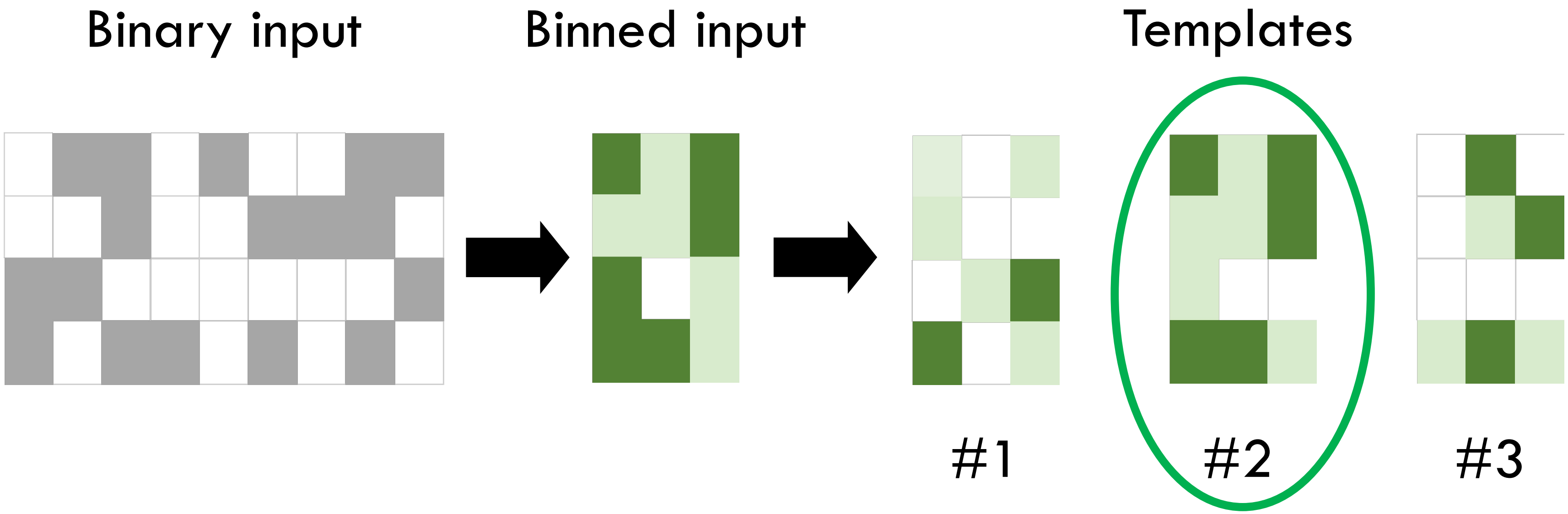
Which template does the input most closely resemble?

Template Matching



Which template does the input most closely resemble?

Template Matching



How do neuroscientists determine this?

Pearson Correlation Coefficient (PCC)

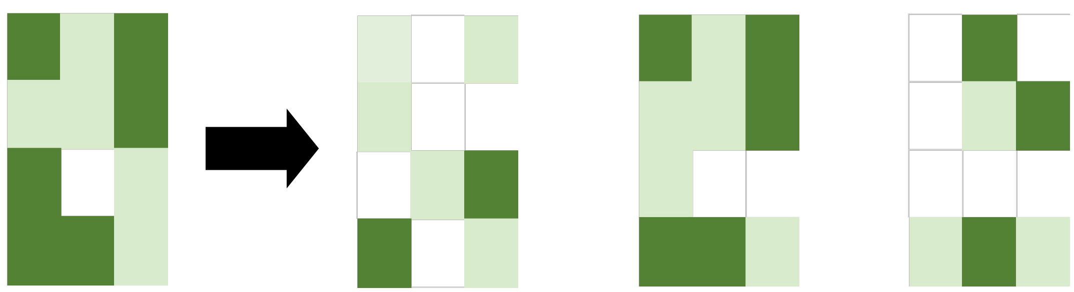
Widely used metric to measure
the "closeness" of two matrices

$$r(X,Y) = \frac{\sum_{i=1}^{L} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{L} (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{L} (y_i - \overline{y})^2}}$$

PCC Example

Binned input

Templates



#1
Move
right arm

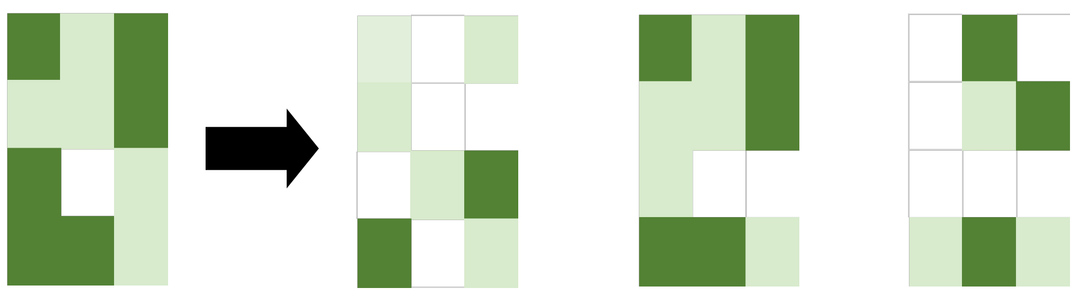
#2
Move left
arm

#3
Move left
leg

PCC Example

Binned input

Templates



#1
Move
right arm

#2
Move left
arm

#3
Move left
leg

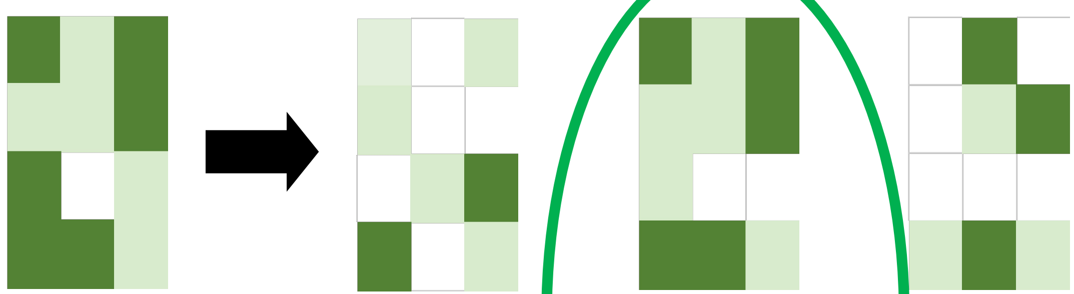
0.135 0.857 0.196

PCC scores (*r*)

PCC Example

Binned input

Templates



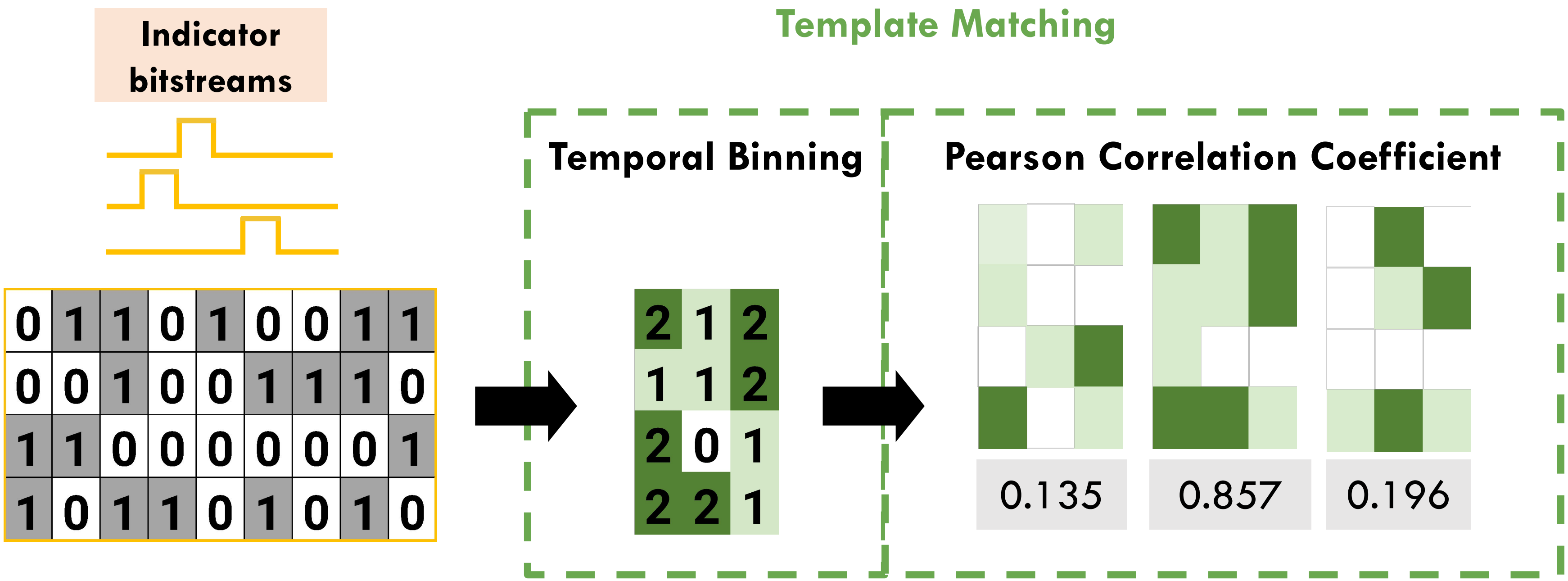
#1 Move right arm — 0.135

#2 Move left arm — 0.857

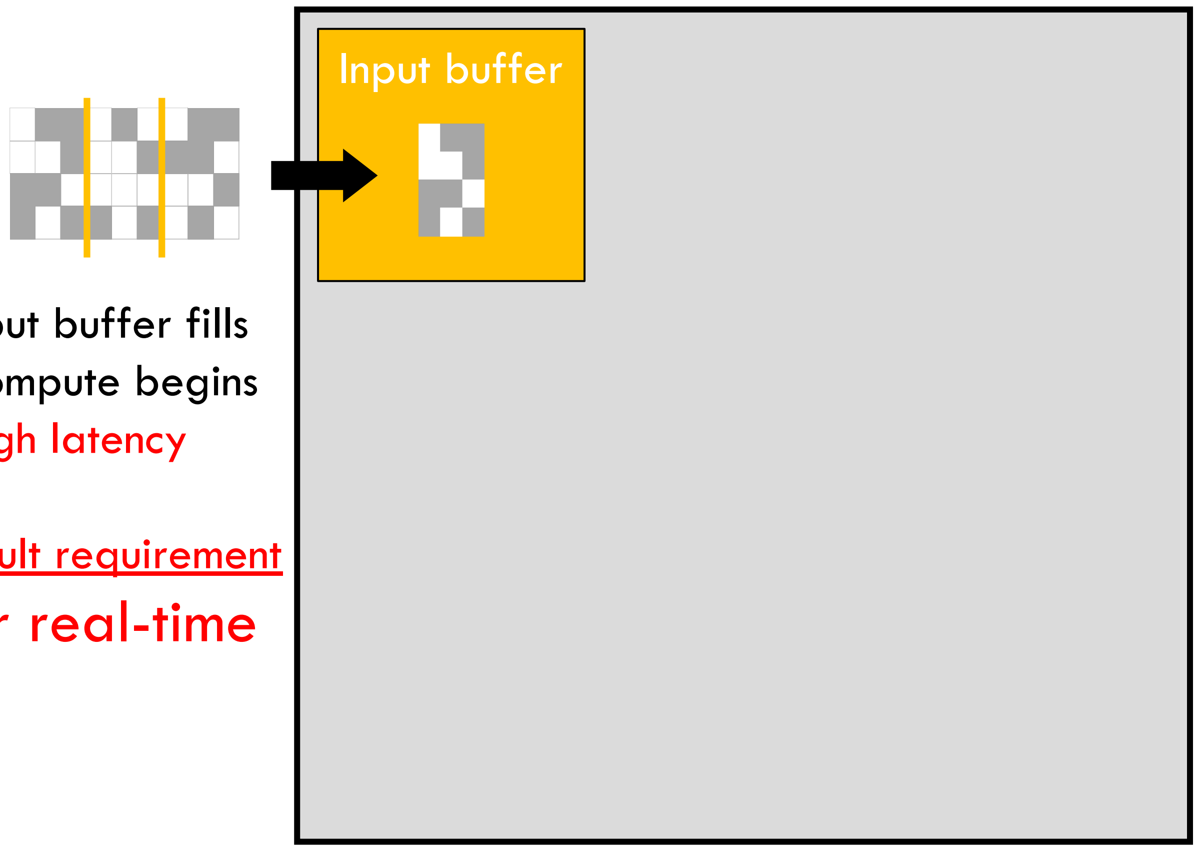
#3 Move left leg — 0.196

PCC scores (r)

Template Matching Overview



Costs of baseline template matching design



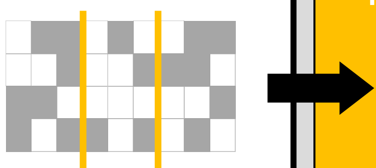
Entire input buffer fills before compute begins

 High latency

Most difficult requirement

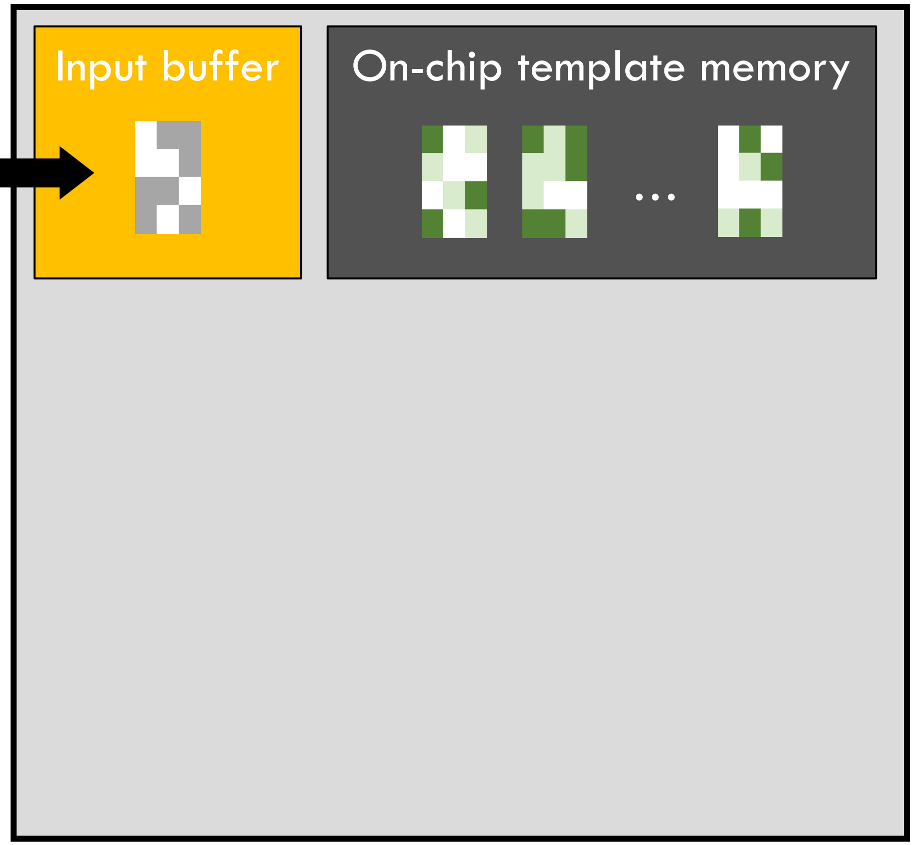
5ms for real-time

Costs of baseline template matching design



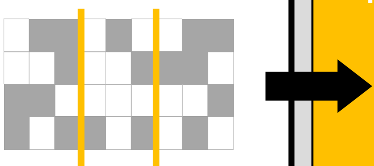
Entire input buffer fills before compute begins
 High latency

<u>Most difficult requirement</u>
5ms for real-time



Storage of input + templates
 Large memory cost
e.g. +1.24 Gb each

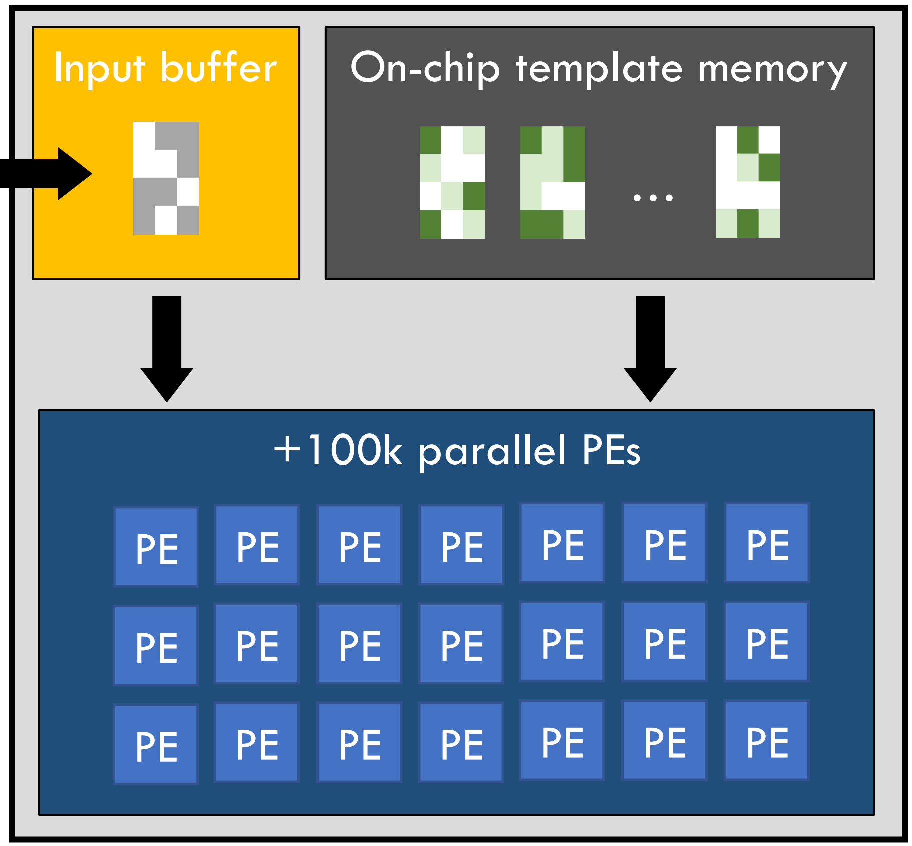
Costs of baseline template matching design



Entire input buffer fills before compute begins
□ High latency

<u>Most difficult requirement</u>

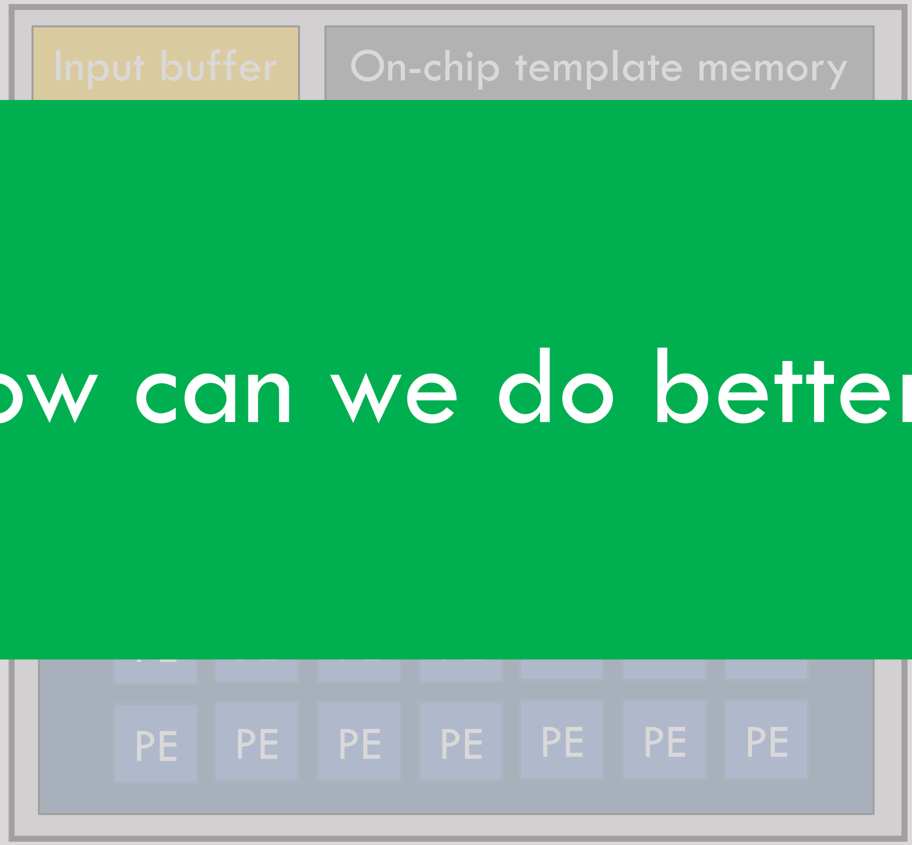
5ms for real-time



Storage of input + templates
□ Large memory cost
e.g. +1.24 Gb each

Many processing elements
□ Large area cost

Costs of baseline template matching design



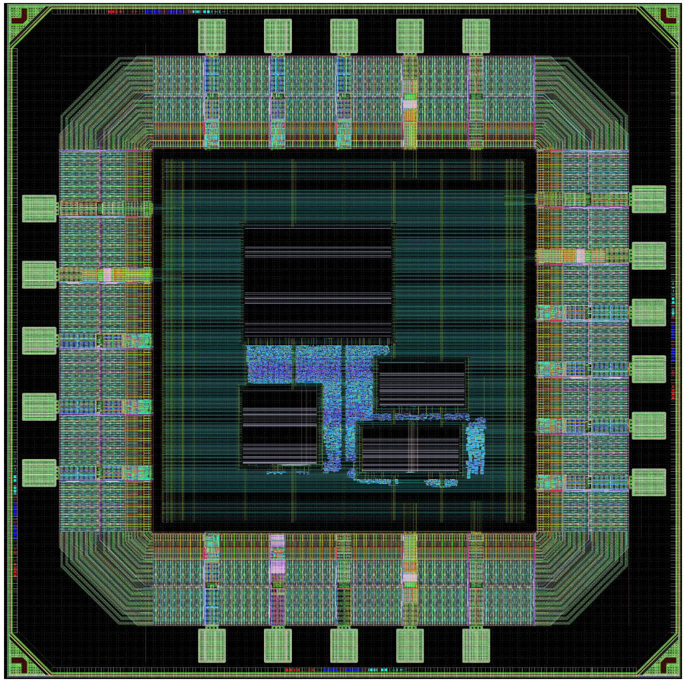
Storage of input +

How can we do better?

Large area cost

Noema: Custom Hardware Accelerator

- **50mm²** chip in **65nm** technology
- Only **24µsec** latency
- **30K** neurons, **9 sec** template
  - **1.2W** power consumption
  - **10×** more neurons than ever recorded
- **Linear** scalability



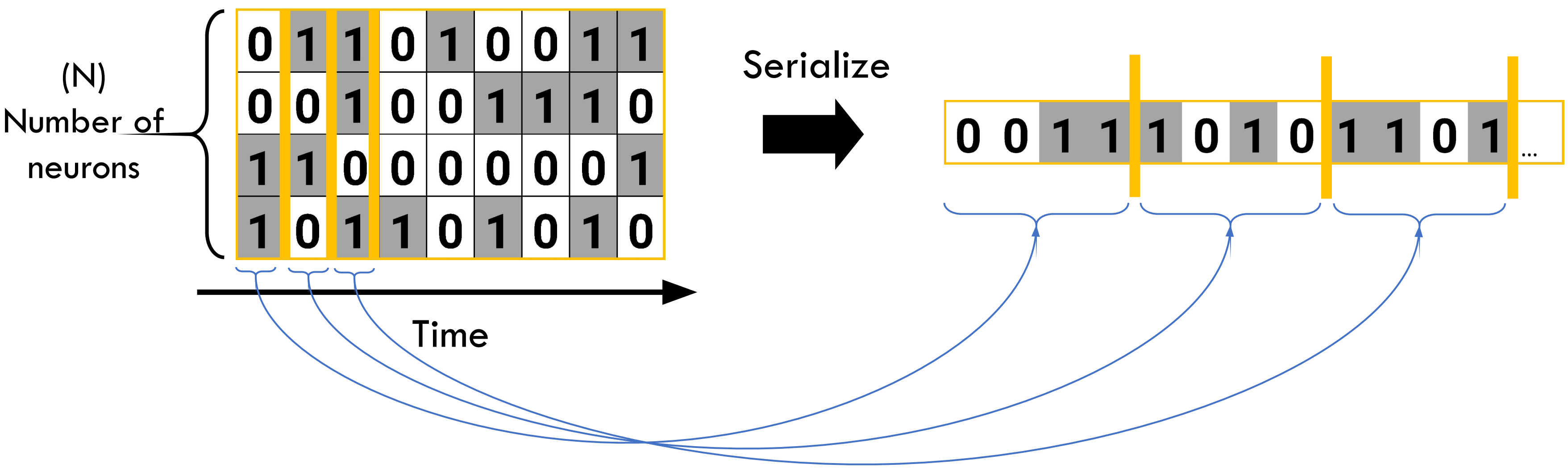
Input Serialization & PCC Reformulation

(N)
Number of
neurons

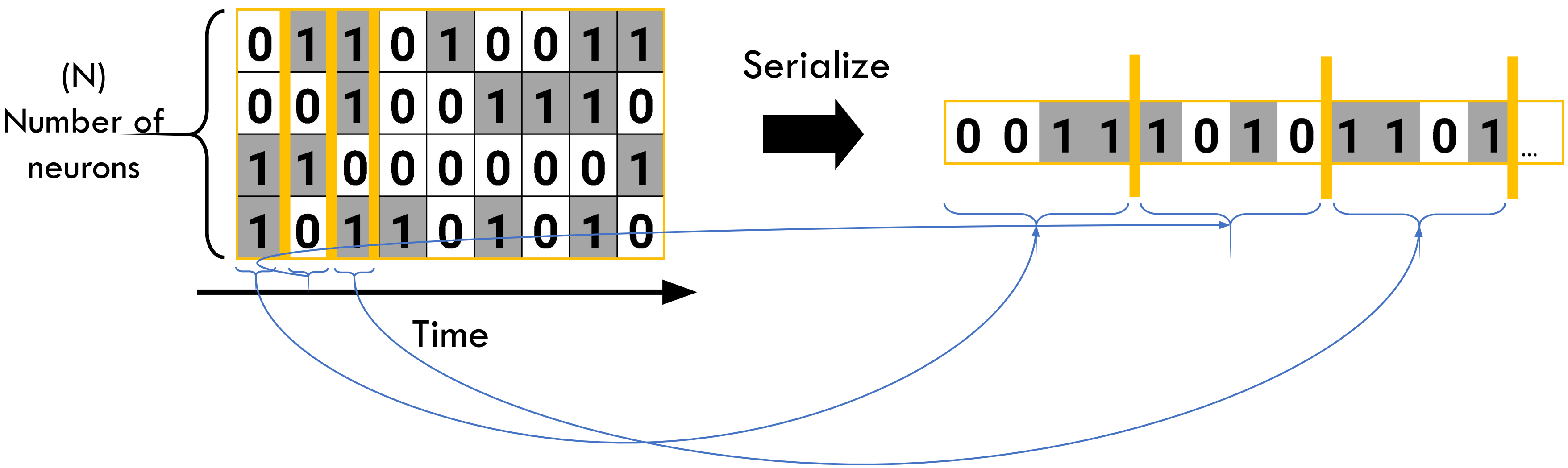
0	1	1	0	1	0	0	1	1
0	0	1	0	0	1	1	1	0
1	1	0	0	0	0	0	0	1
1	0	1	1	0	1	0	1	0

Time

Input Serialization & PCC Reformulation



Input Serialization & PCC Reformulation



$$r(X, Y) = \frac{\sum_{i=1}^{L} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{L} (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{L} (y_i - \overline{y})^2}}$$

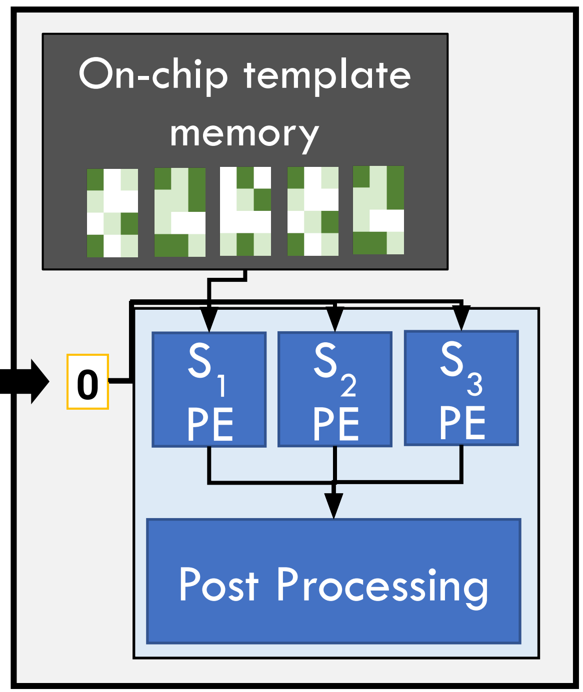
Reformulation

$$r[t]^2 = \frac{(C_1 S_1[t] - C_2 S_2[t])^2}{C_3 \left(C_1 S_3[t] - S_2[t]^2\right)}$$

Noema's innovations

Bit-serial input

- No buffering overhead
- Compute immediately when received



Near-memory bit-serial PEs

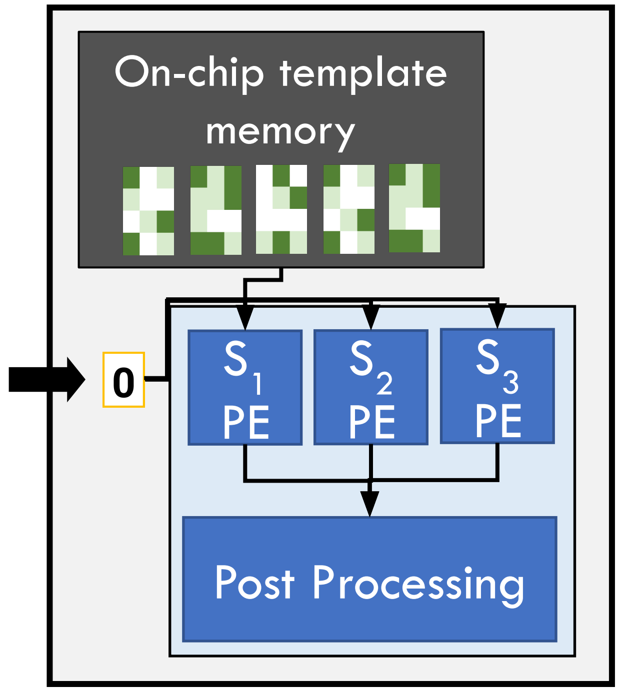
- Based on reformulated PCC
- Tiny, easy to scale

$$r[t]^2 = \frac{(C_1 S_1[t] - C_2 S_2[t])^2}{C_3\left(C_1 S_3[t] - S_2[t]^2\right)}$$

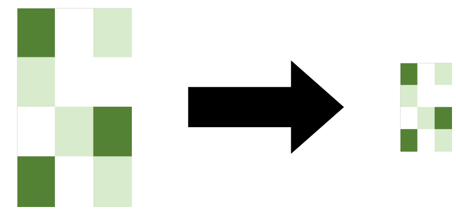
Noema’s innovations

Bit-serial input

- No buffering overhead
- Compute immediately when received



Simple memory compression (~2.8x)



Near-memory bit-serial PEs

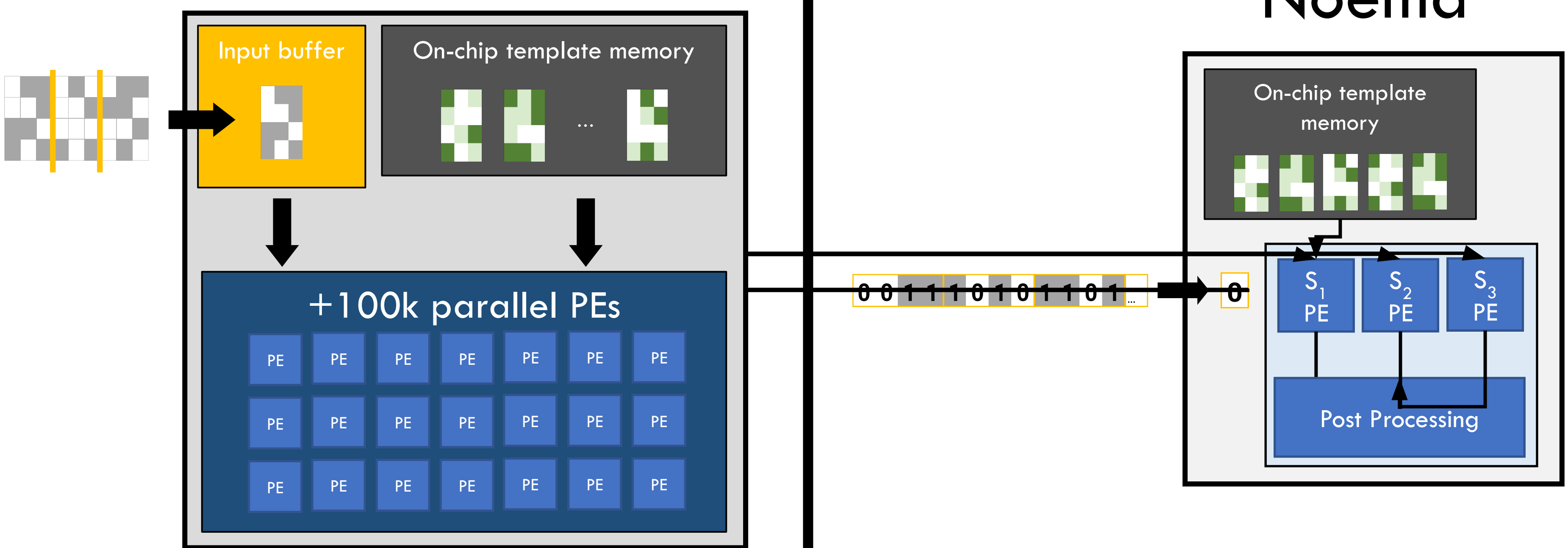
- Based on reformulated PCC
- Tiny, easy to scale

$$r[t]^2 = \frac{(C_1 S_1[t] - C_2 S_2[t])^2}{C_3\left(C_1 S_3[t] - S_2[t]^2\right)}$$

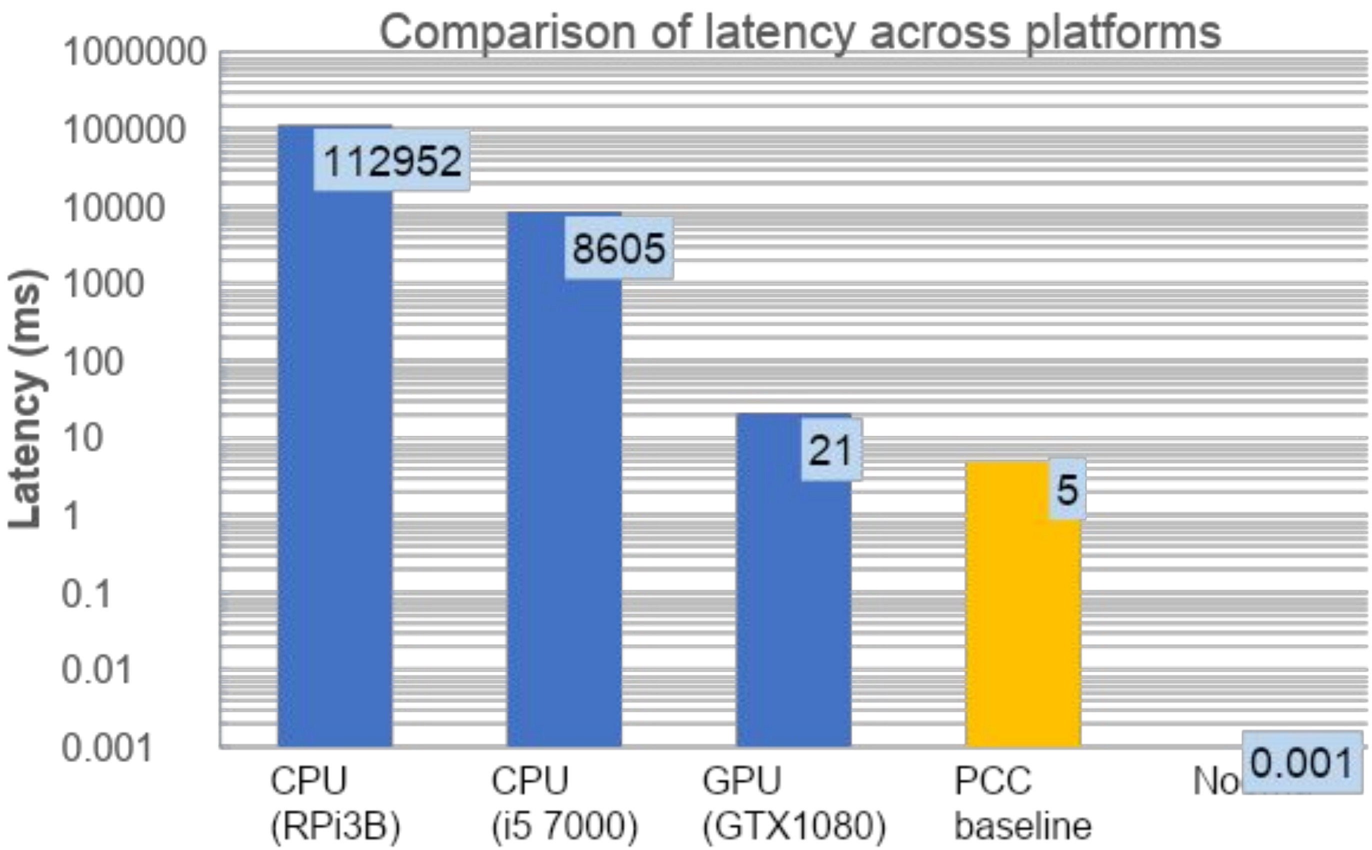
Baseline to Noema overview

Baseline

Noema

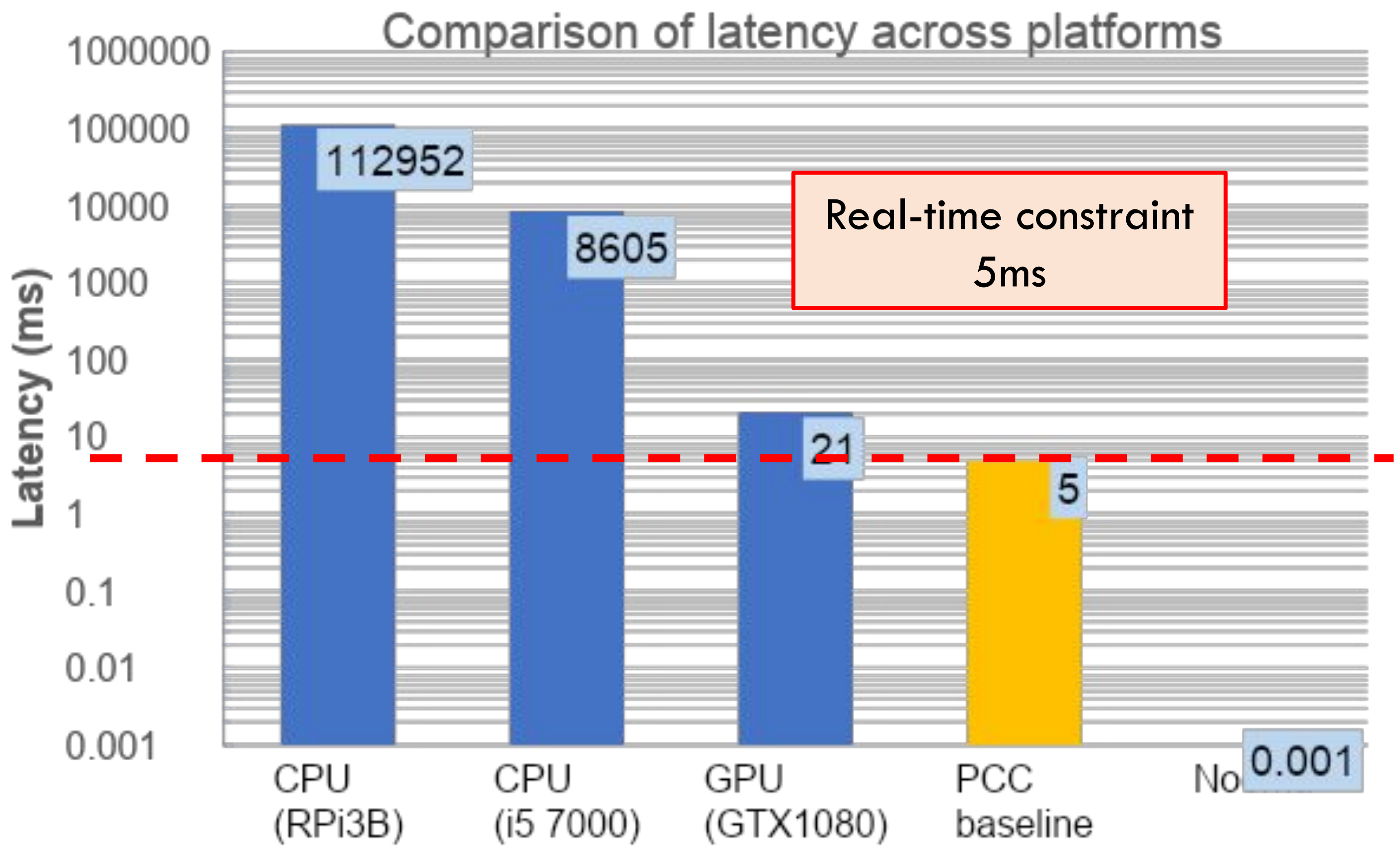


Performance Results



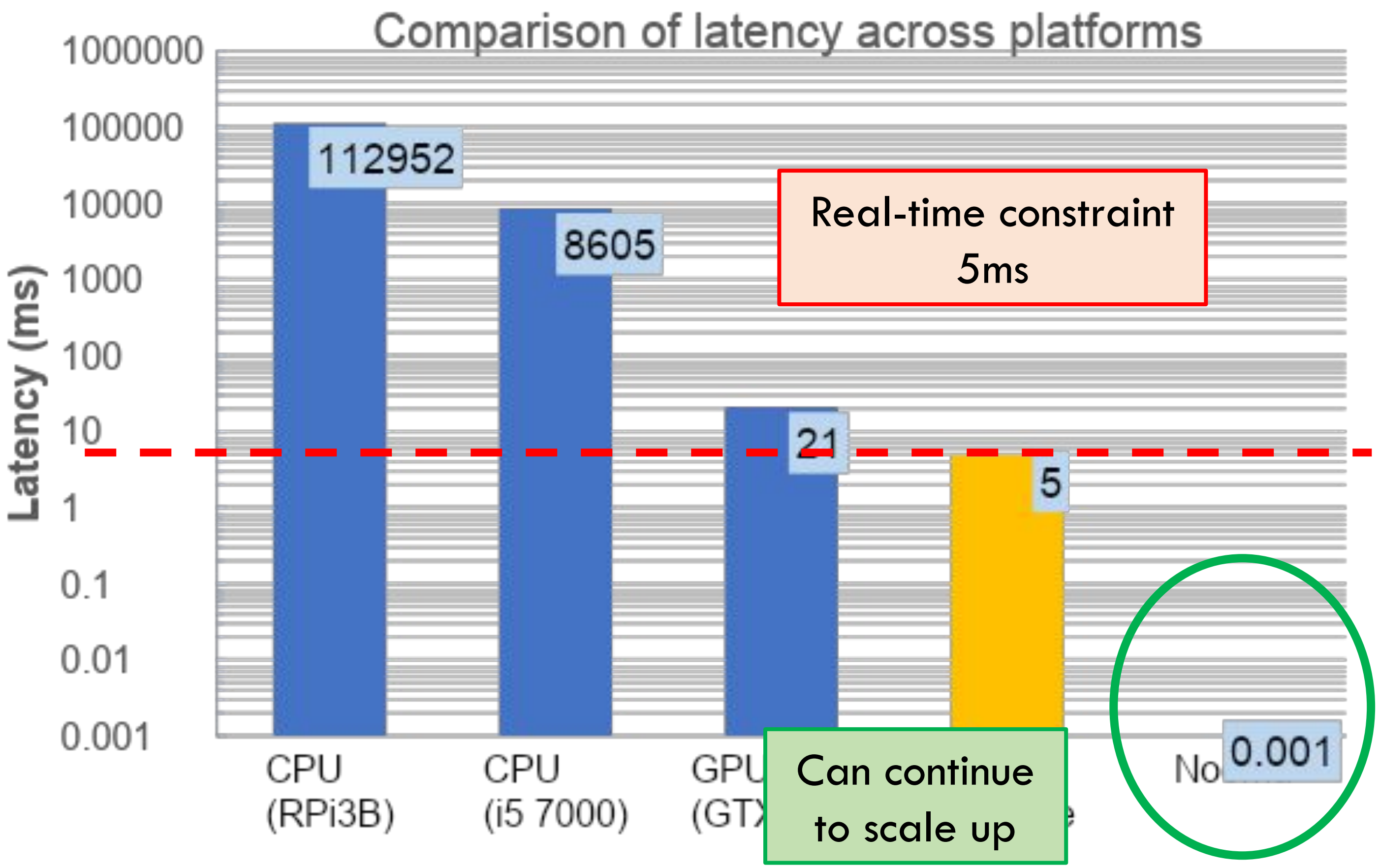
* For the most demanding configuration tested (see paper for details)

Performance Results



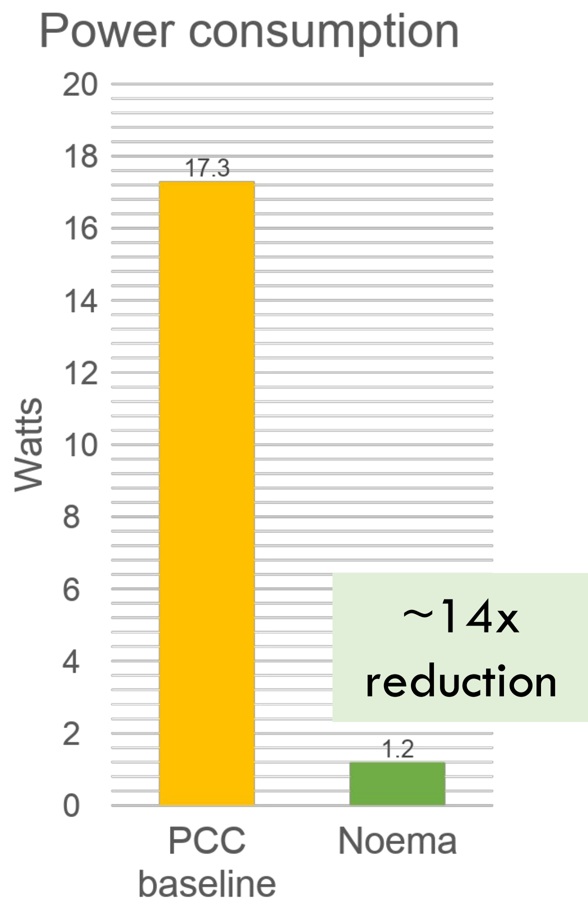
* For the most demanding configuration tested (see paper for details)

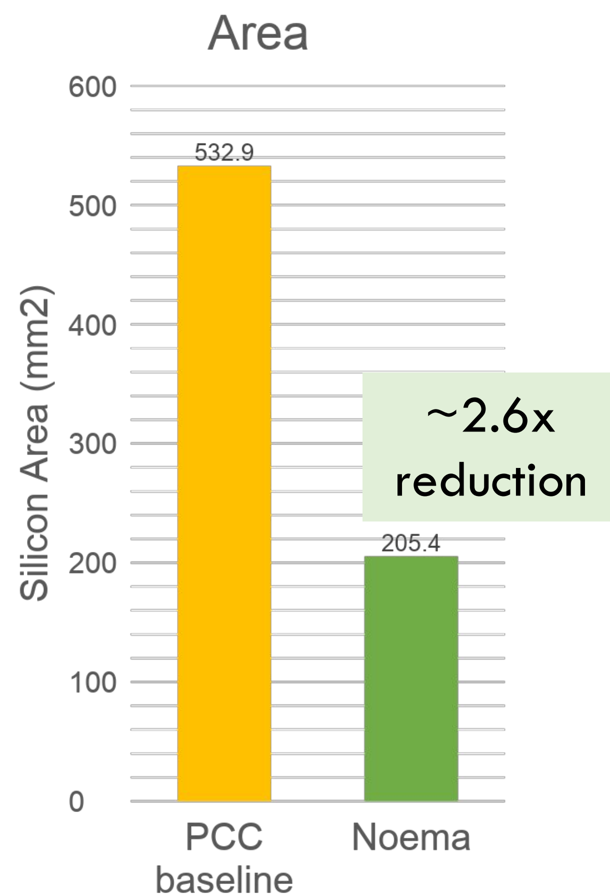
Performance Results



* For the most demanding configuration tested (see paper for details)

Power & Area Results

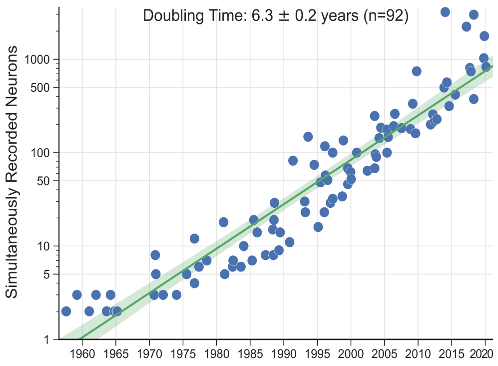




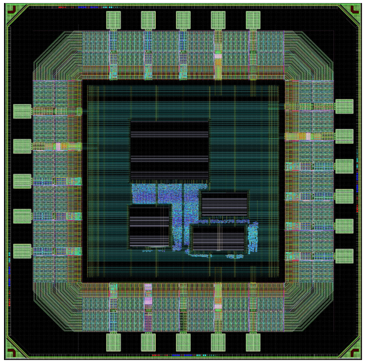
* For the most demanding configuration tested (see paper for details)

Noema: Key Takeaways

- Exponential growth in data



- Current solutions are not sufficient
- Our baseline solution can meet the demand
- Noema can scale to meet ***future*** demand
  - 14x less power, 2.6x smaller



Contact Information

Ameer Abdelhadi – ameer.abdelhadi.@utoronto.ca

Eugene Sha – eugene.sha@mail.utoronto.ca

Ciaran Bannon – ciaran.bannon@utoronto.ca

Andreas Moshovos – moshovos@eecg.toronto.edu

Hendrik Steenland – wsneurotek@gmail.com