Architectural Techniques to Build Energy-Efficient Brain Implants

ARM Research Summit: Biotechnology Track

Abhishek Bhattacharjee

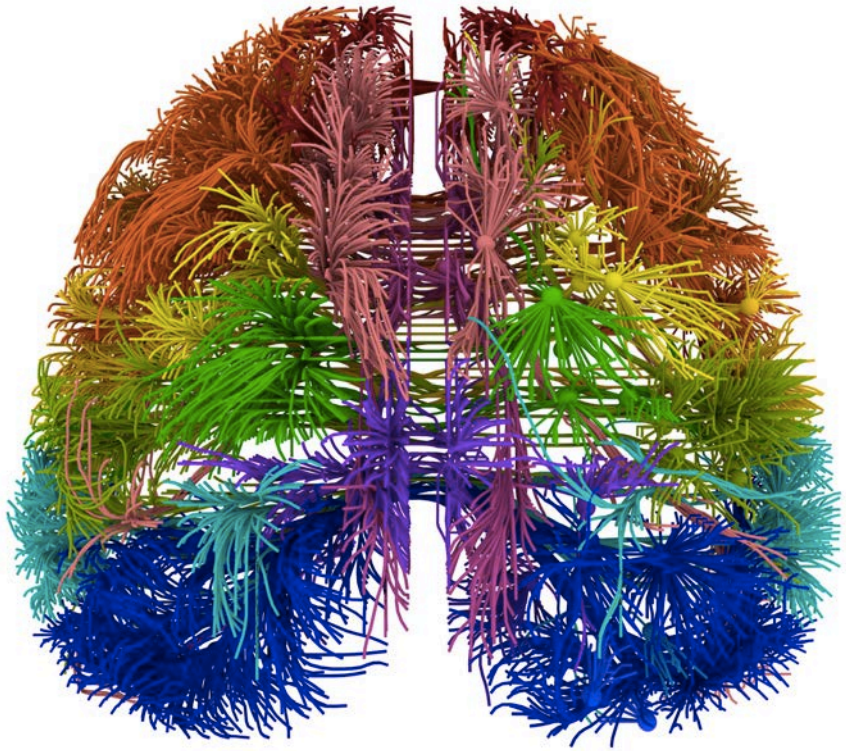
Associate Professor

Department of Computer Science

Rutgers University

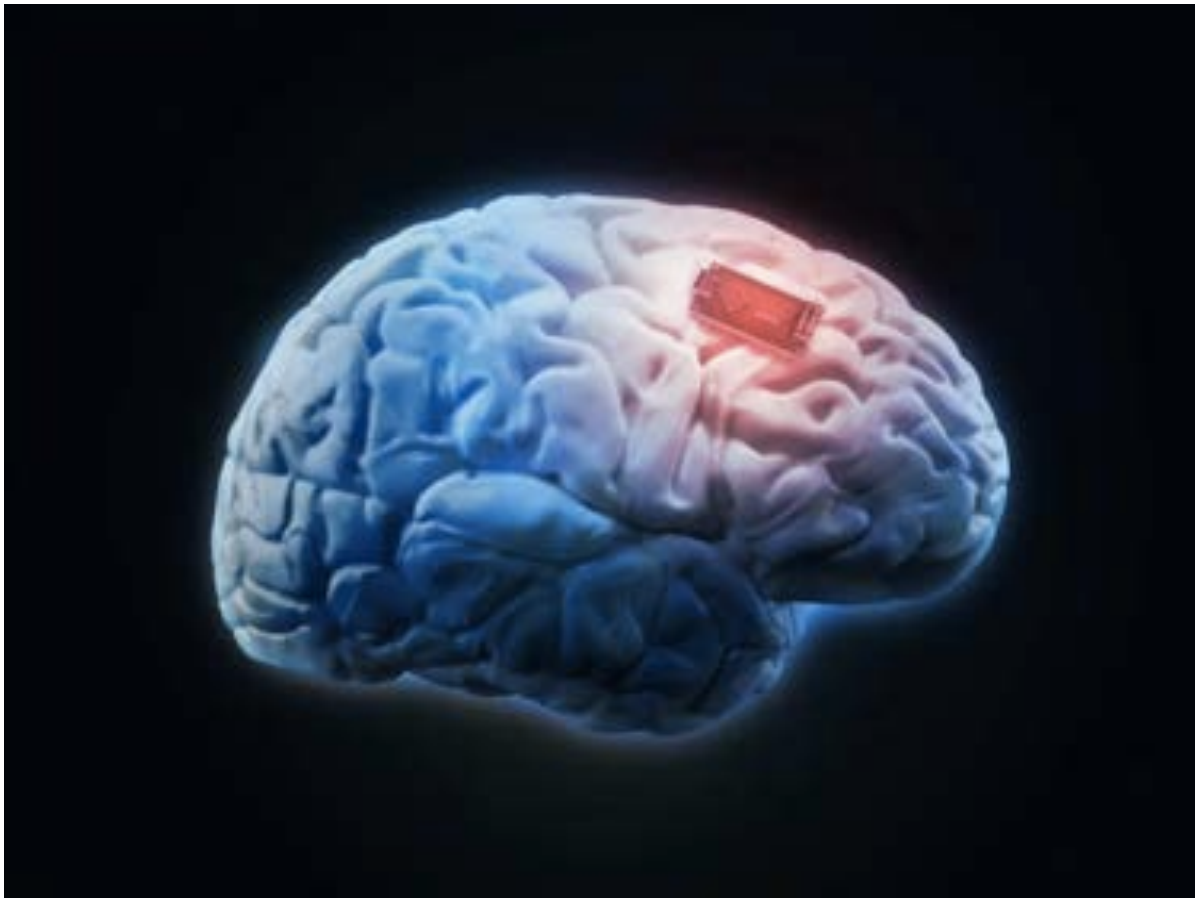
https://media.wired.com/photos/5932492aedfced5820d0f637/master/w_660,c_limit/Rama-mouse.jpg

How does neuronal activity affect behavior and how do we treat disorders?



http://blogs.discovermagazine.com/d-brief/files/2014/04/brain-map-top-.jpg

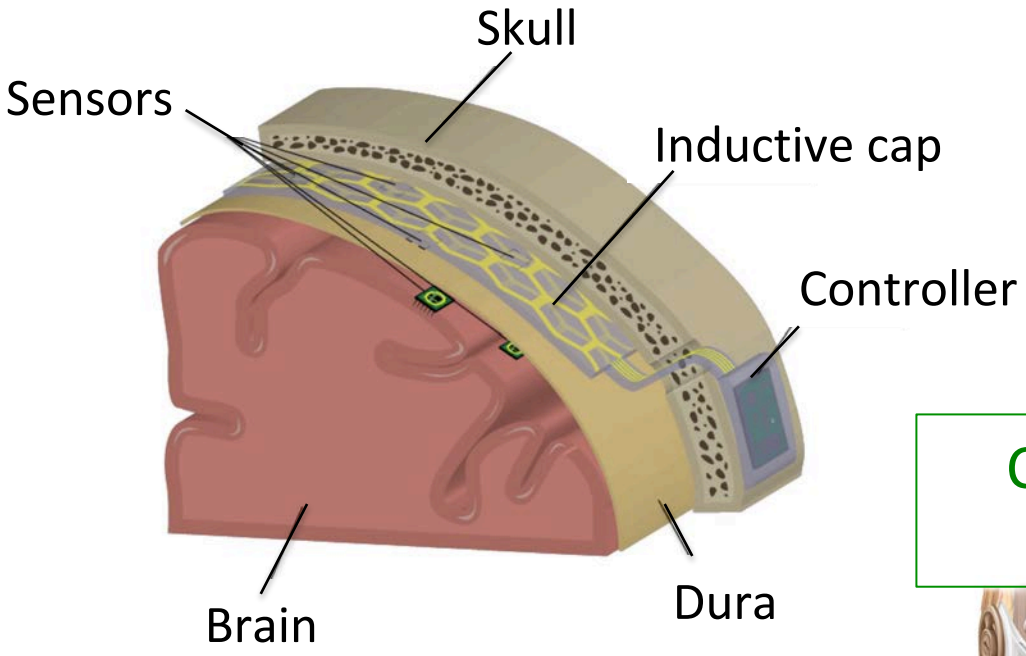
Implants to understand the brain and treat disorders

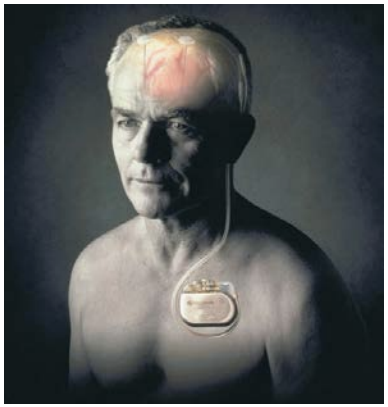


https://spectrum.ieee.org/image/Mjk4MTQ1NQ.jpeg

Brain implants are already being used to treat neurological conditions

Deep-brain stimulation
Over 40K users





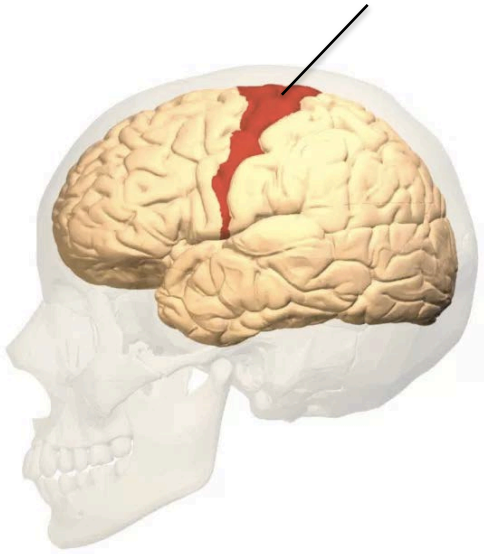
Cochlear and retinal implants
Over 50K users

http://borton.engin.brown.edu/; https://www.scmp.com/sites/default/files/styles/486w/public/2013/07/05/06d190bf66c18164db7f0abd0f8b7147.jpg?itok=FBB_j8Wq

Brain implants are already being used to treat neurological conditions

Motor cortex implants

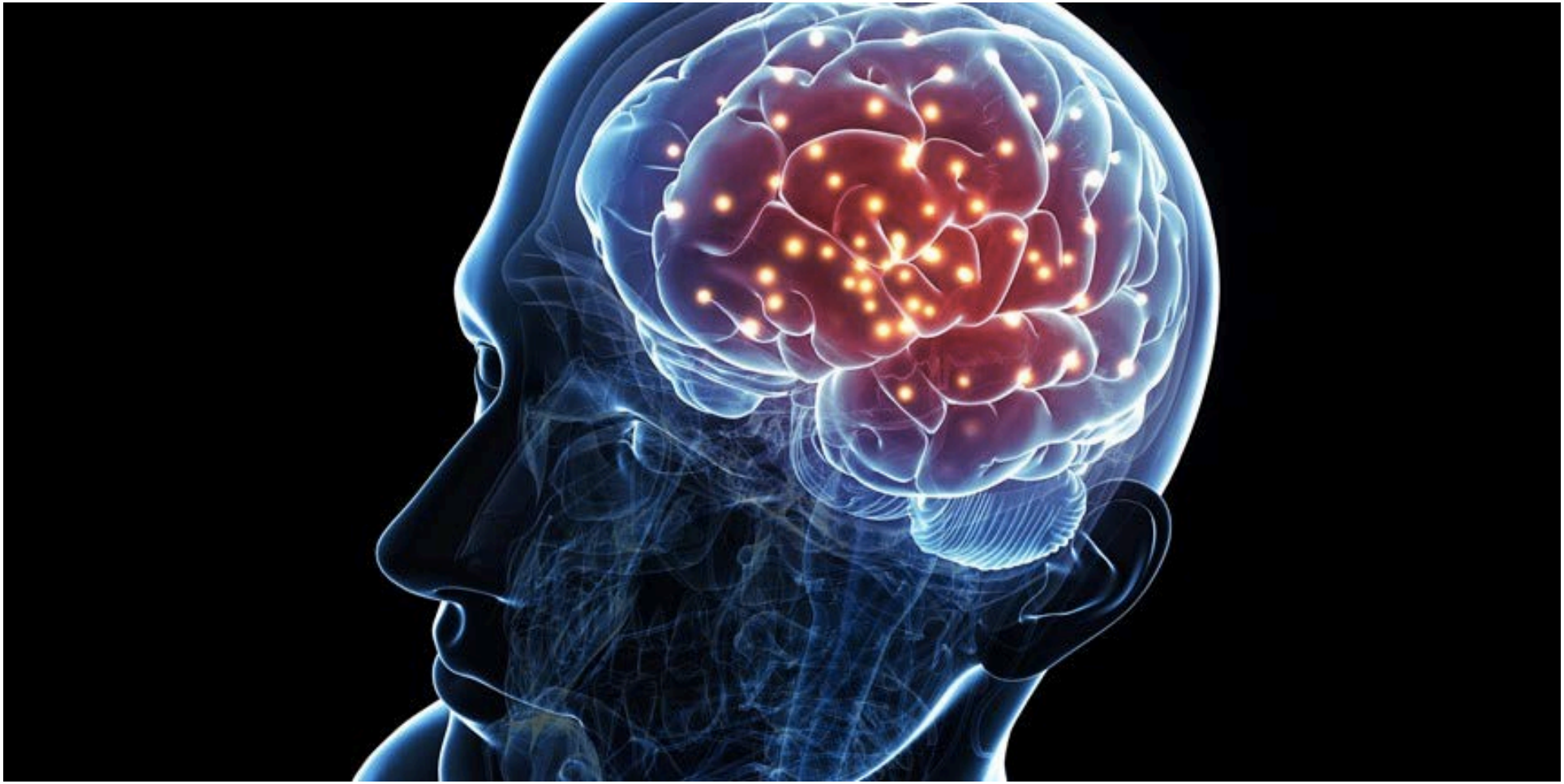
Implant reads from motor cortex



https://i.ytimg.com/vi/QRt8QCx3BCo/maxresdefault.jpg

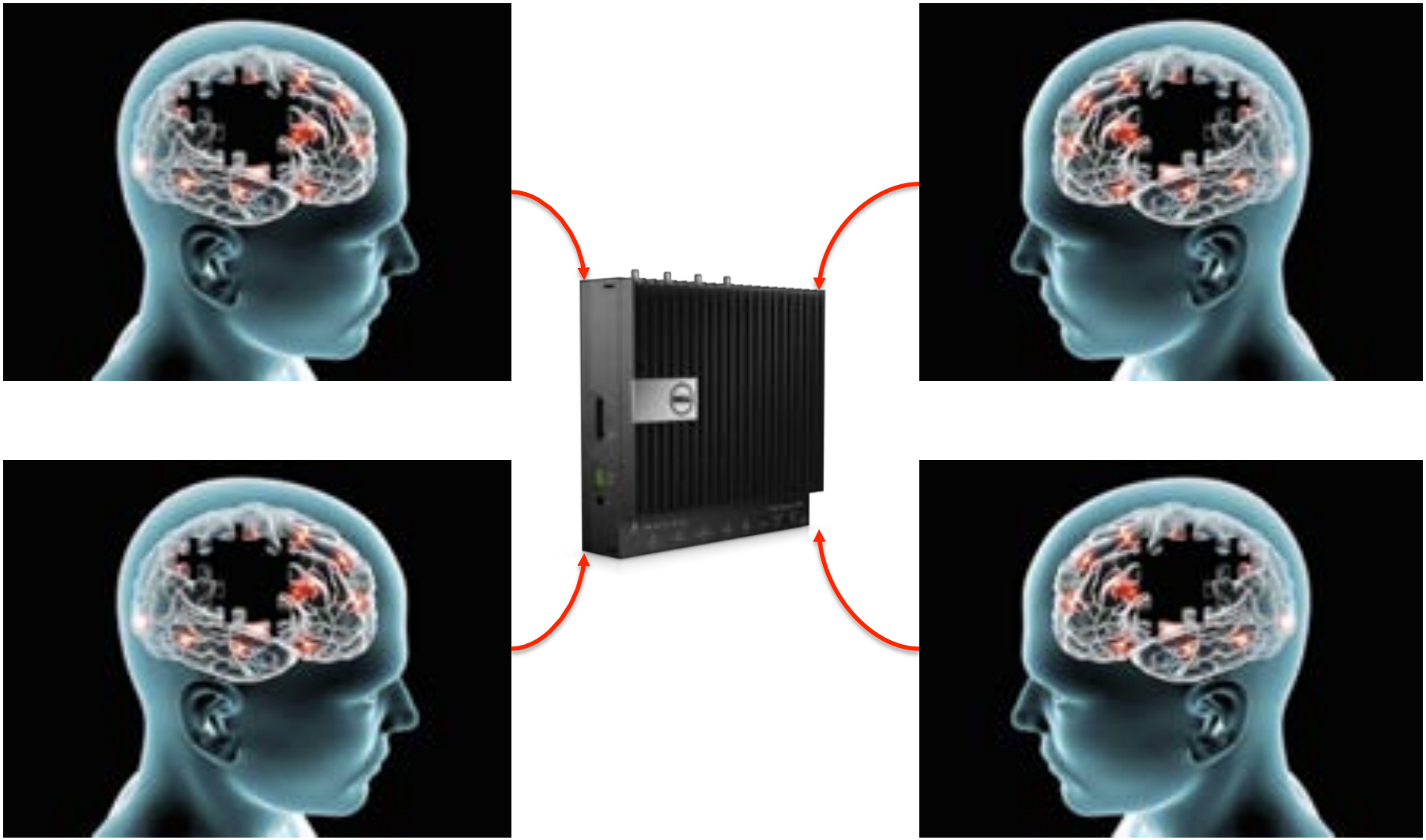
http://www.sciencemag.org/sites/default/files/styles/article_main_large/public/images/sn-ratsense.jpg?itok=s8CyJ3yK

We want to monitor and stimulate multiple brain sites



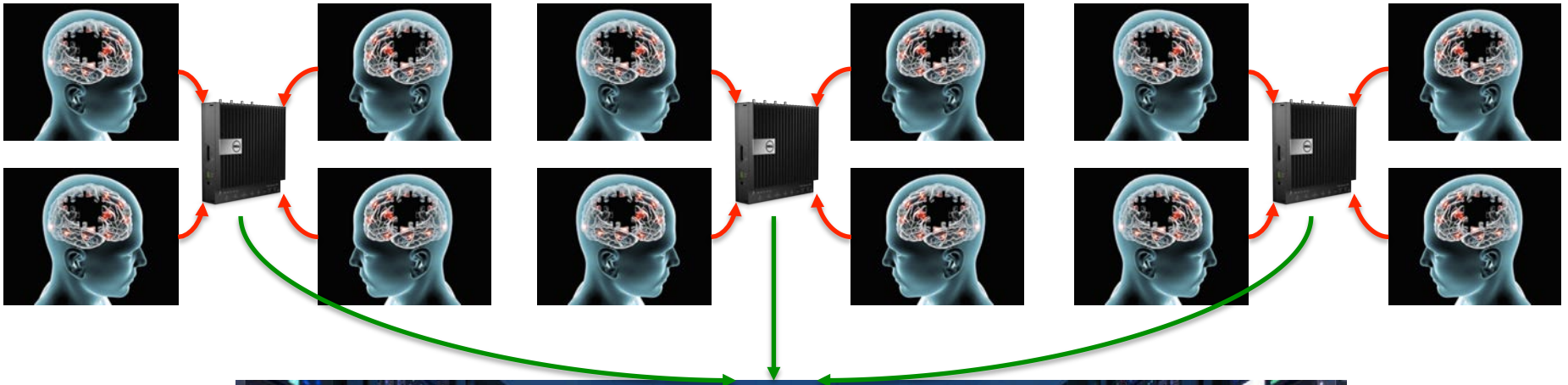
https://img.purch.com/rc/300x200/

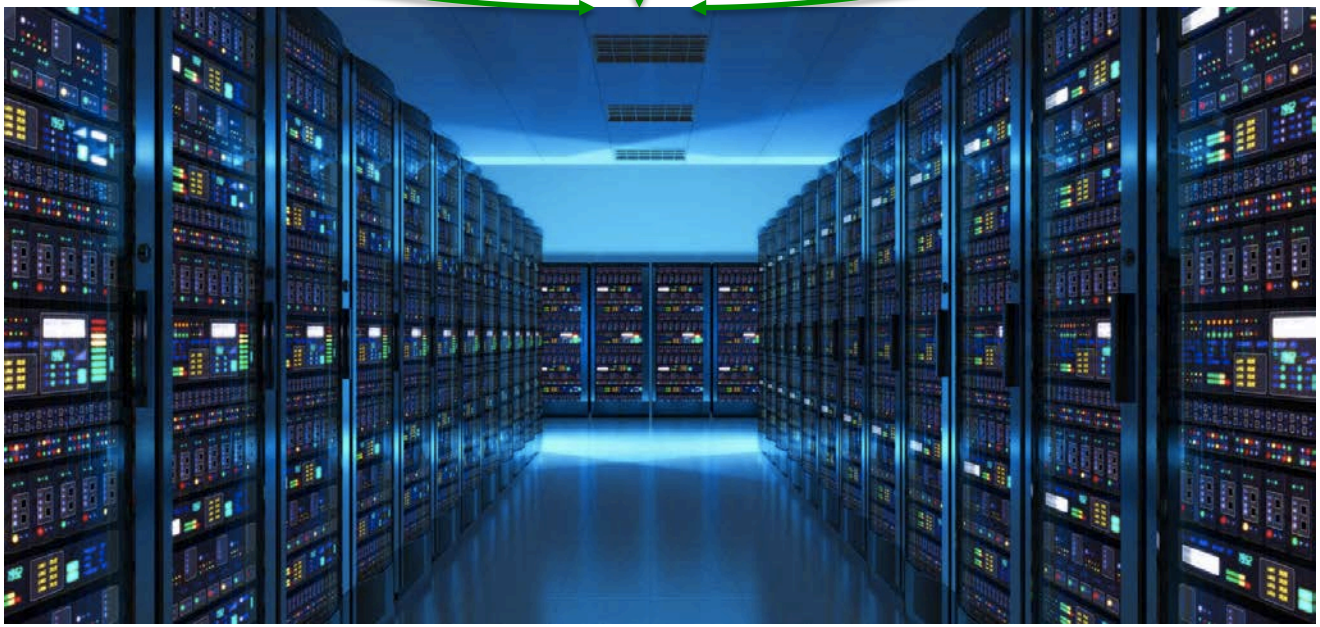
Send neuronal data from implants to base stations in our surroundings



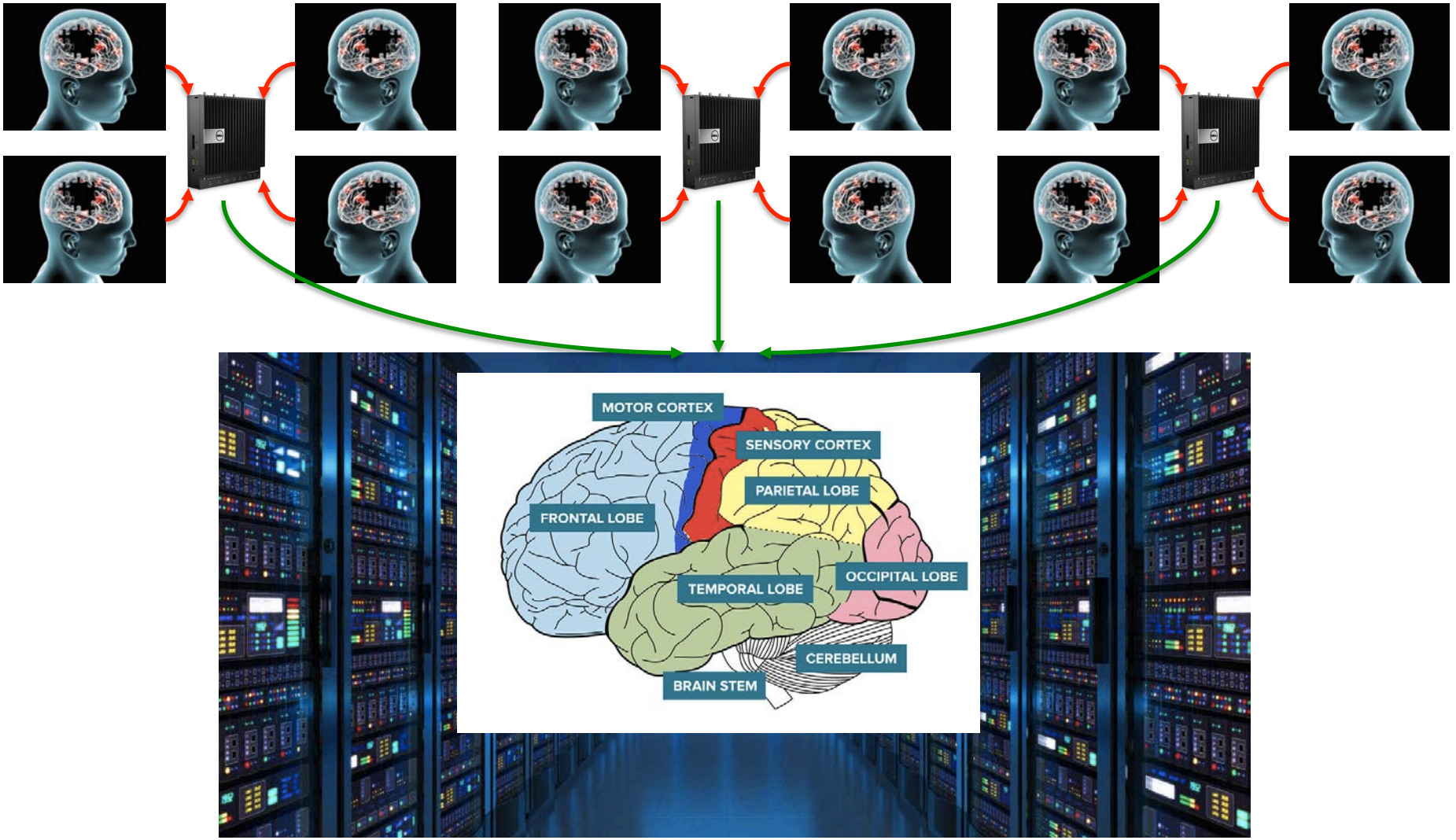
https://spectrum.ieee.org/image/Mjk4MTQ1NQ.jpeg

The neural data will be relayed to clusters and datacenters





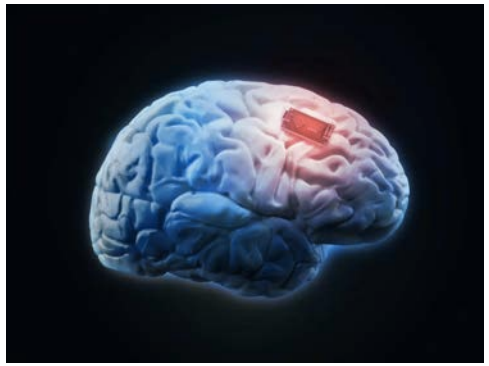
Server-scale systems process neuronal data and model the brain

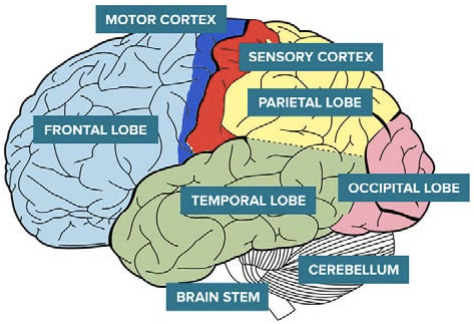


My work reduces implant energy and improves server performance

Implants

Servers





Implemented on real systems

Monkeys, pigs, sheep

BRAINGATE
TURNING THOUGHT INTO ACTION

AMD

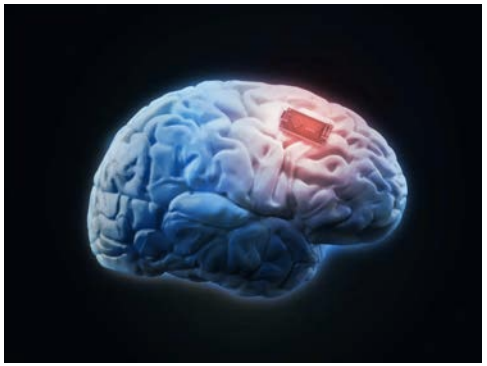
Ryzen chips

Linux

Kernel 4.14 series

My work reduces implant energy and improves server performance

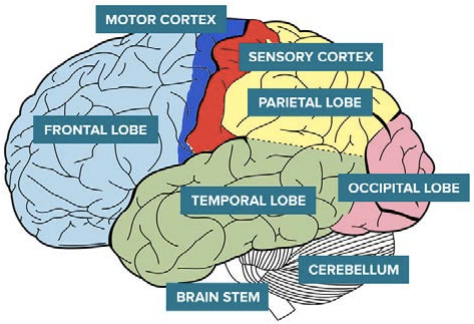
Implants



Monkeys, pigs, sheep

BRAINGATE
TURNING THOUGHT INTO ACTION

Servers



Adoption on systems

AMD

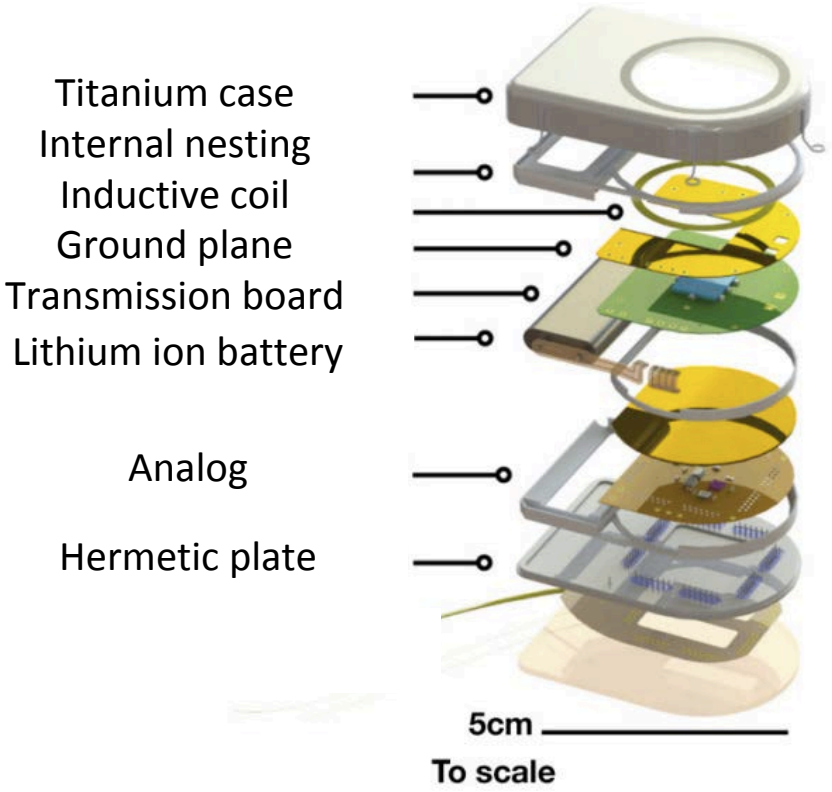
Ryzen chips

Linux

Kernel 4.14 series

Implants are energy-constrained

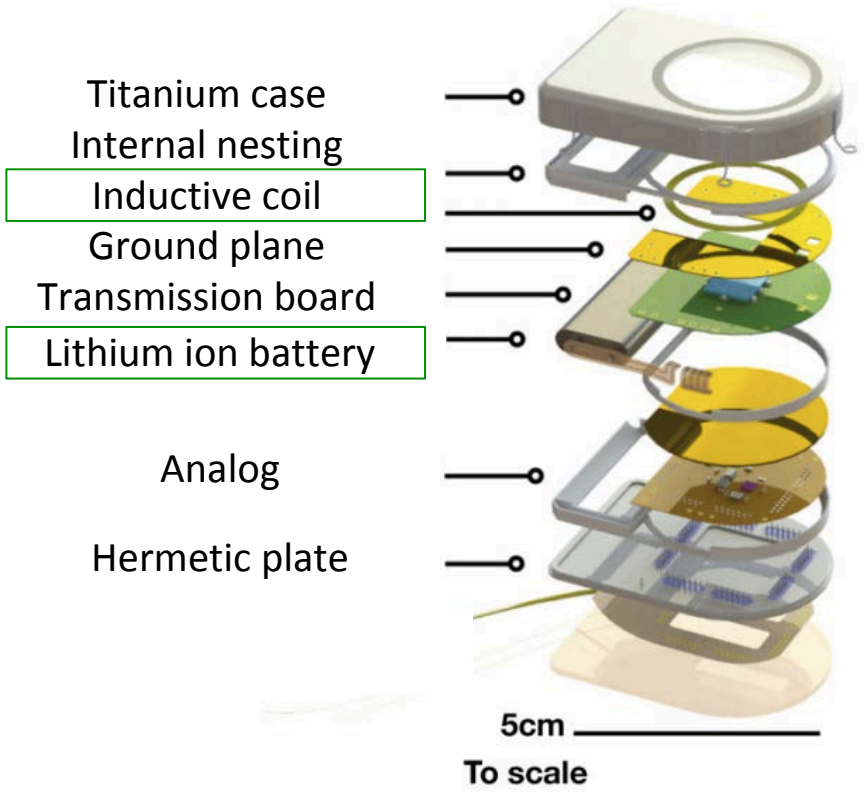
Cerebellar implant



http://borton.engin.brown.edu/

Implants are energy-constrained

Cerebellar implant



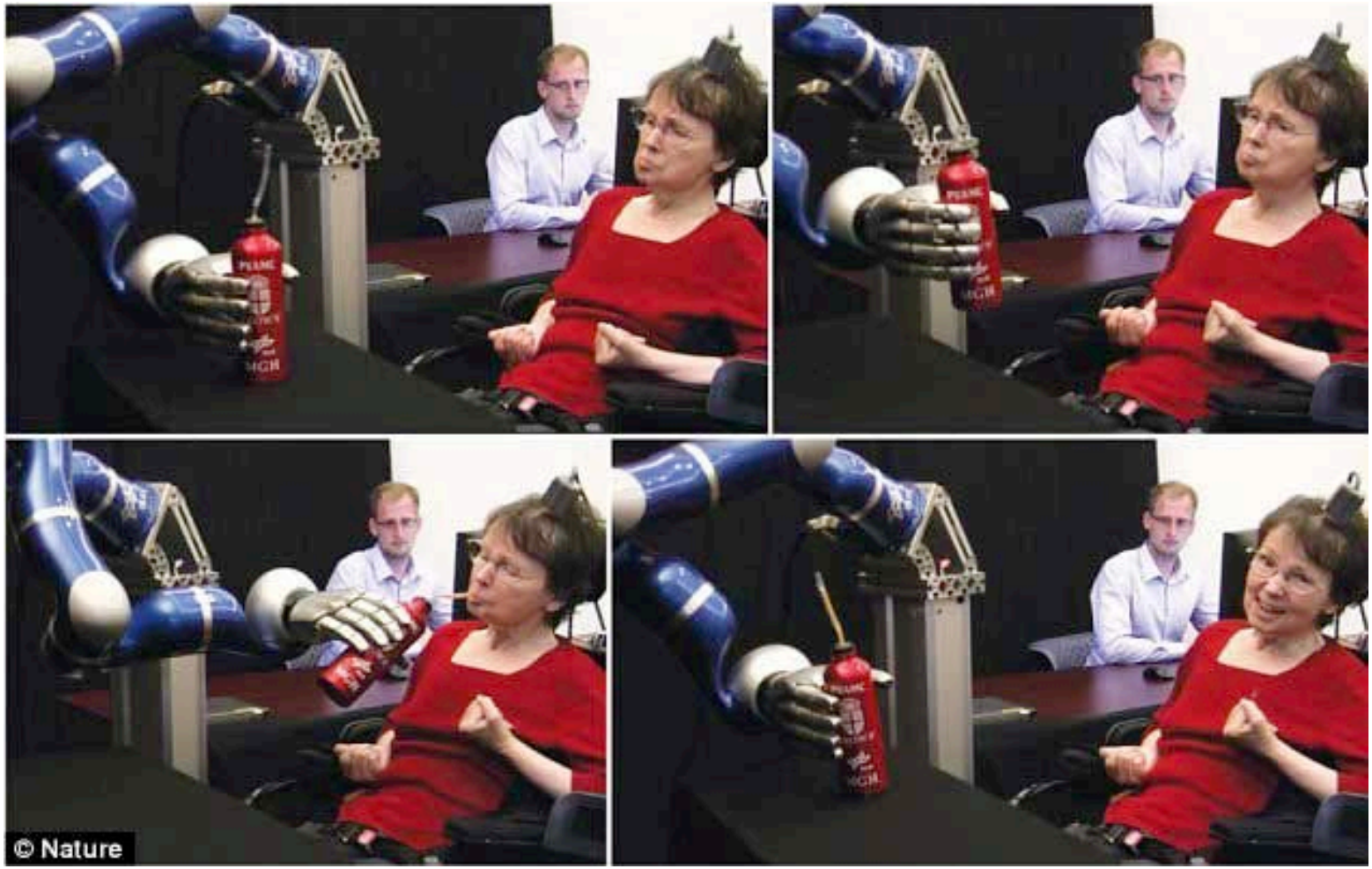
http://borton.engin.brown.edu/

Implants are thermally constrained

1°C increases damage brain tissue

[Mrosek, Anesthesiology Report, ‘12]
[Rutherford, Lancet Neuro, ‘10]
[Liu, Scientific Reports, ’16]

But sufficient performance for real-time responses is needed



Co-opt hardware pieces on the server to save energy on implants

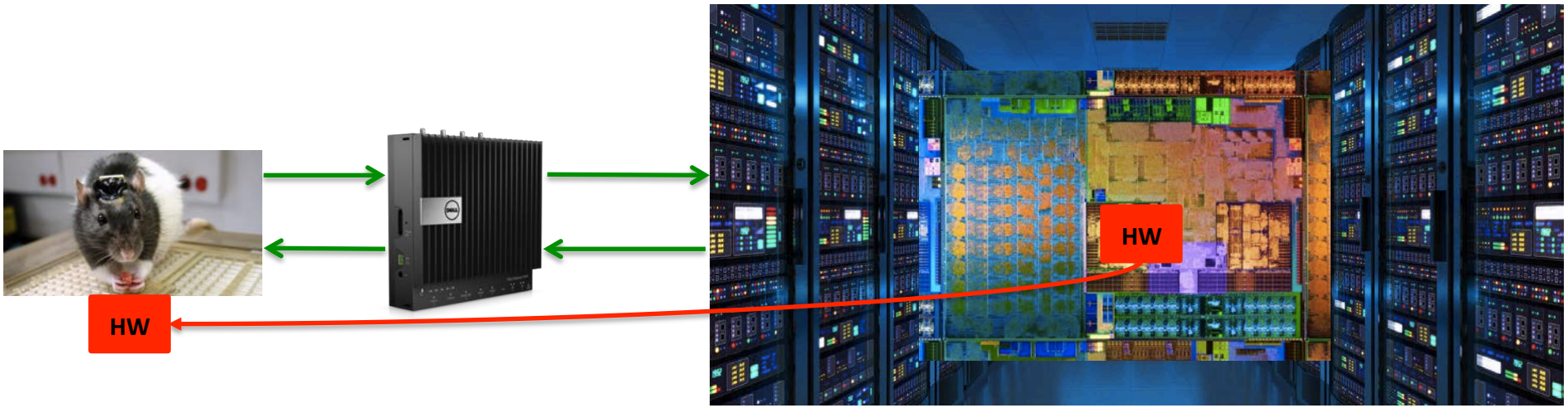
Implants | Base station | Server-scale

Co-opt hardware pieces on the server to save energy on implants

Implants | Base station | Server-scale

Co-opt hardware pieces on the server to save energy on implants

Implants | Base station | Server-scale



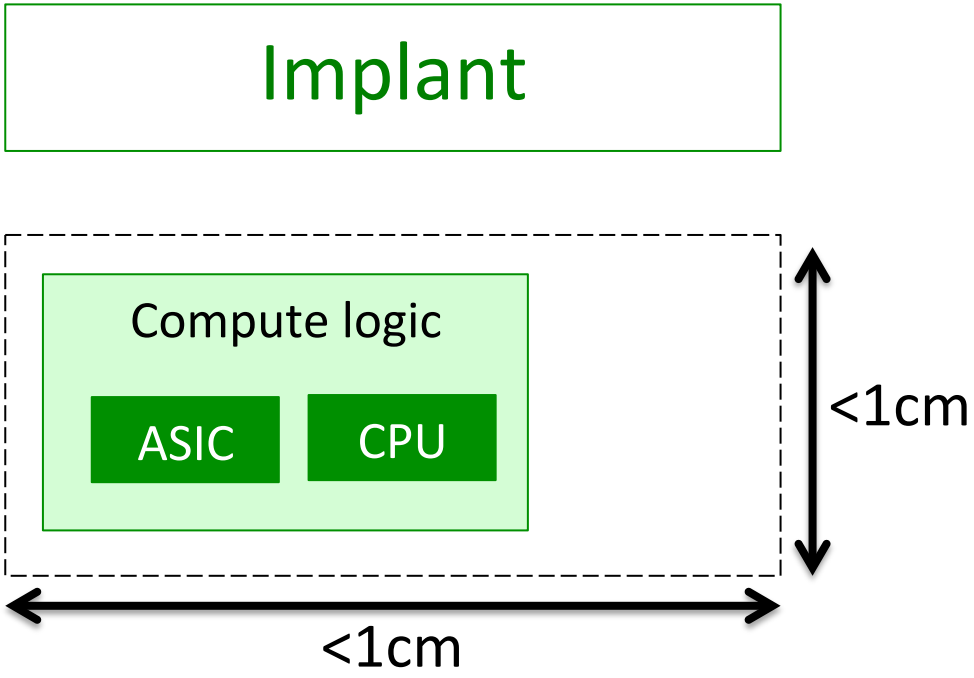
Today: Use **Hardware Perceptrons** to Save 20-35% energy on implants

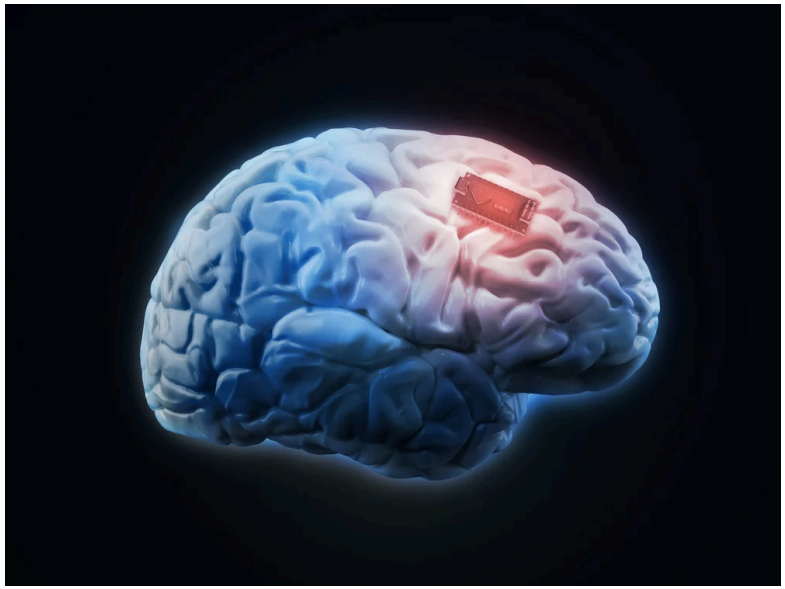
Implants | Base station | Server-scale

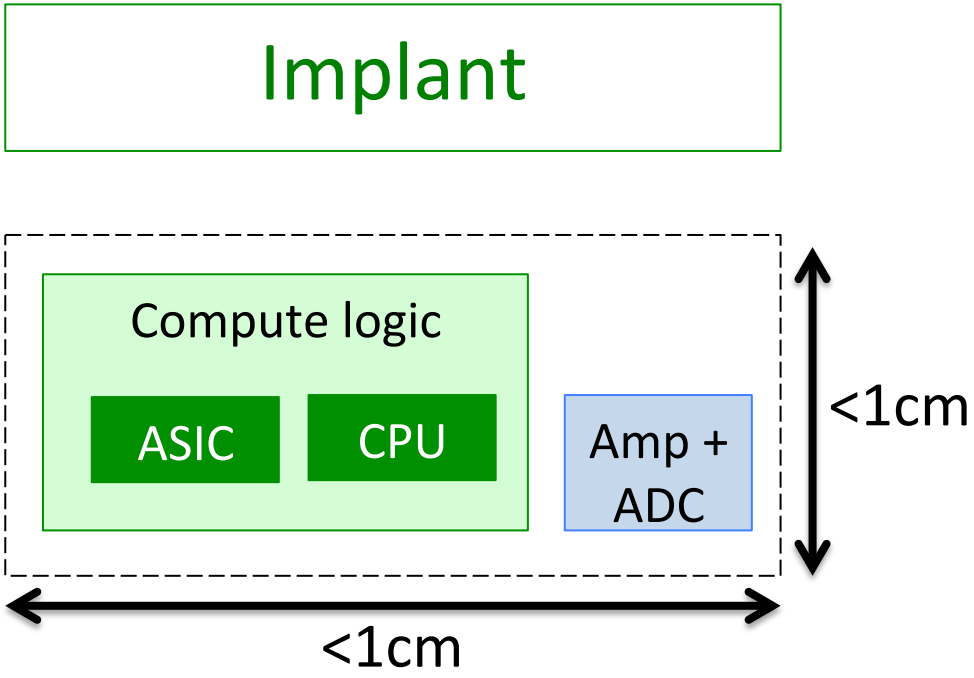
Pushing into real systems

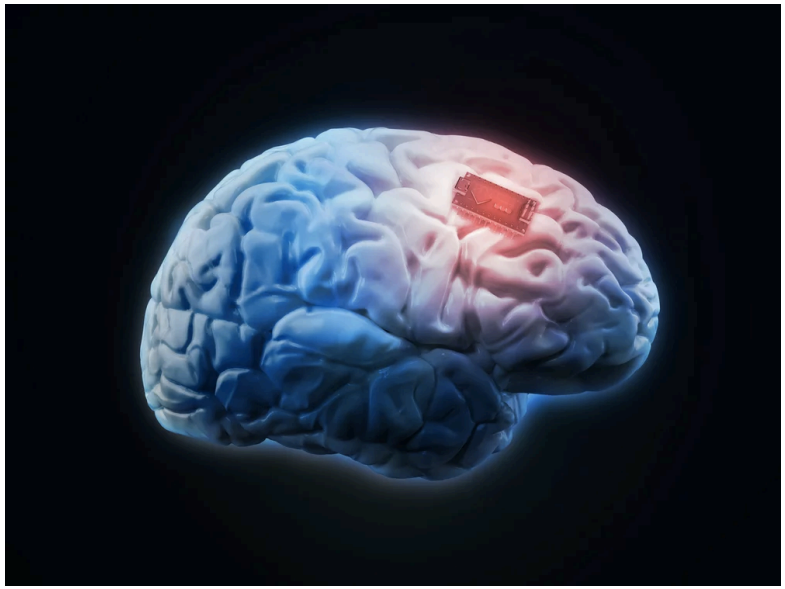
Monkeys, pigs, sheep

BRAINGATE
TURNING THOUGHT INTO ACTION


Implant
Compute logic
ASIC
CPU
<1cm
<1cm




Implant
Compute logic
ASIC
CPU
Amp +
ADC
<1cm
<1cm



Implant

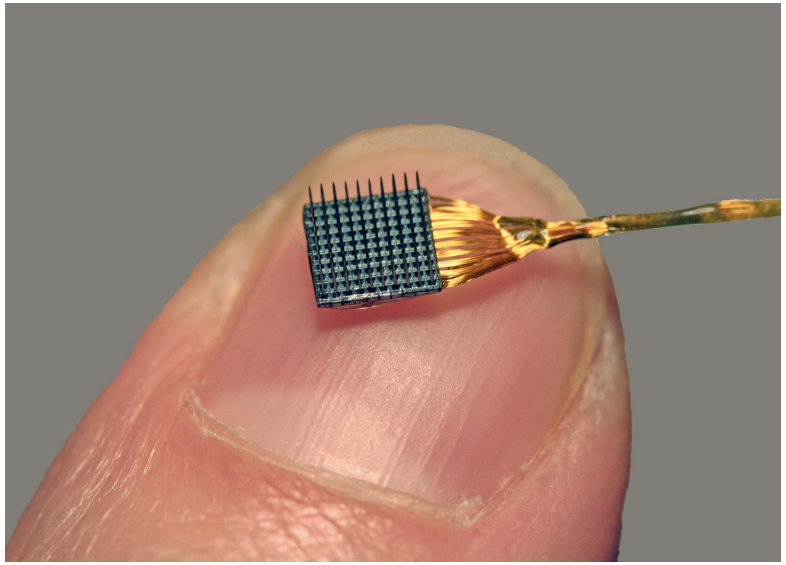
Compute logic

ASIC

CPU

Amp + ADC

Utah array



Implant

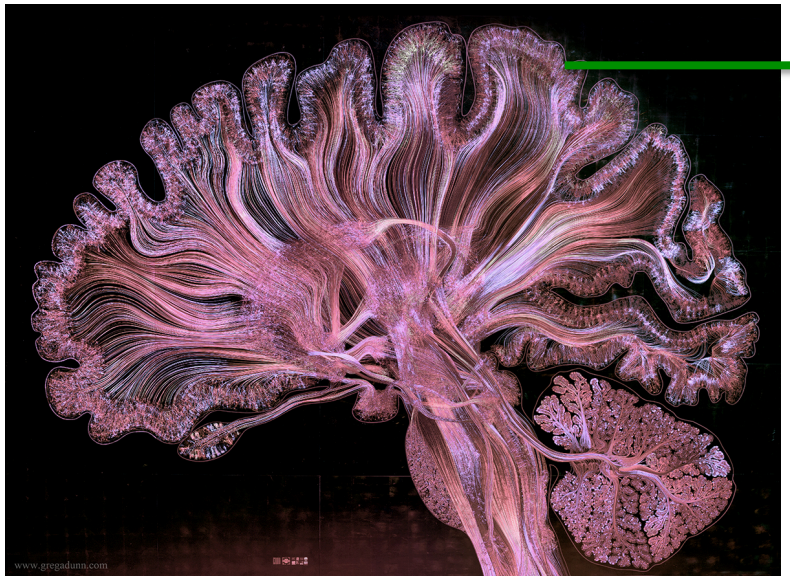
Compute logic

ASIC

CPU

Amp + ADC

Utah array



Utah array probes 1-2mm

Implant

Compute logic

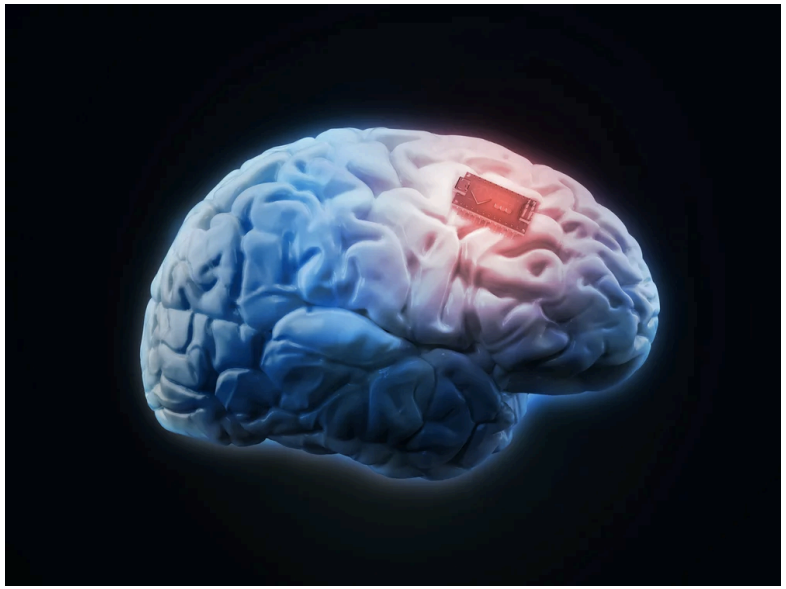
ASIC

CPU

Radio

Amp + ADC

Utah array



Implant

External system

1 device ~ 100Mbps

Compute logic

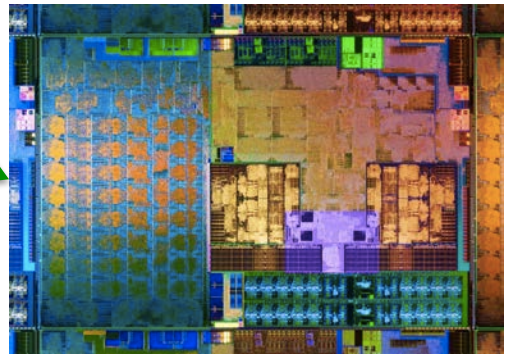
ASIC

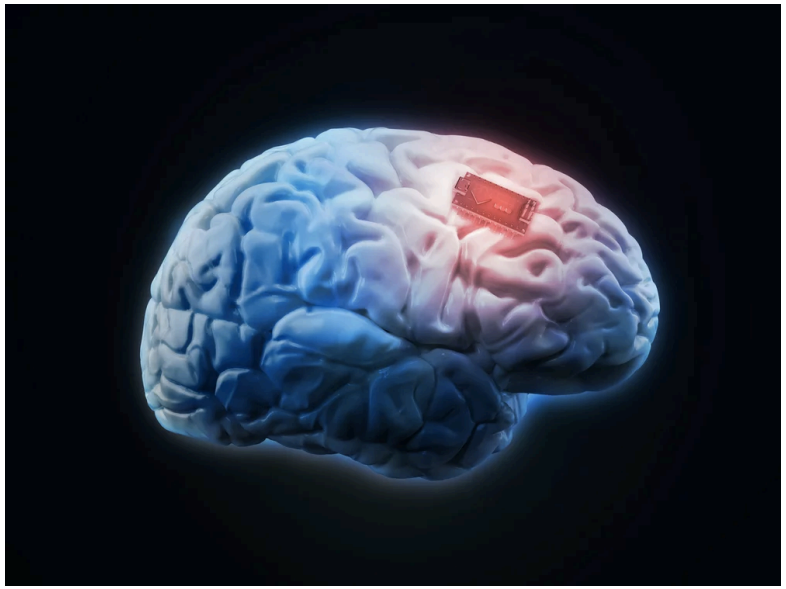
CPU

Radio

Amp + ADC

Utah array





Implant

10mW

Compute logic
4mW

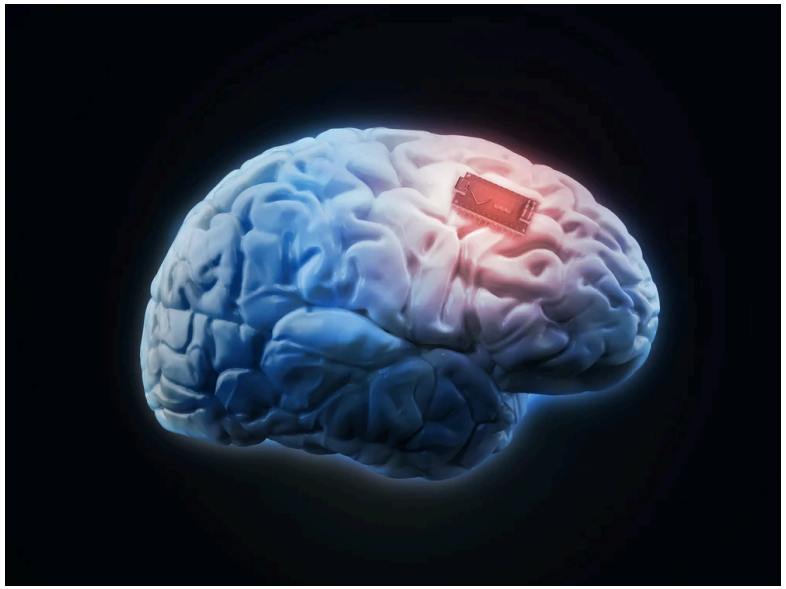
ASIC

CPU

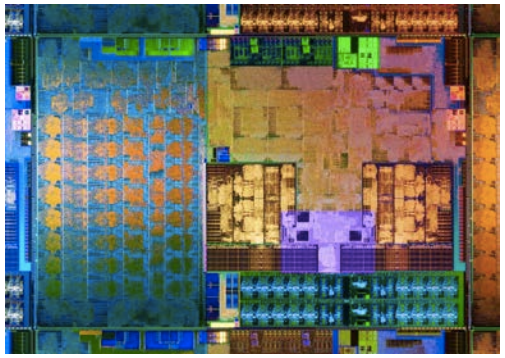
4mW
Radio

2mW
Amp + ADC

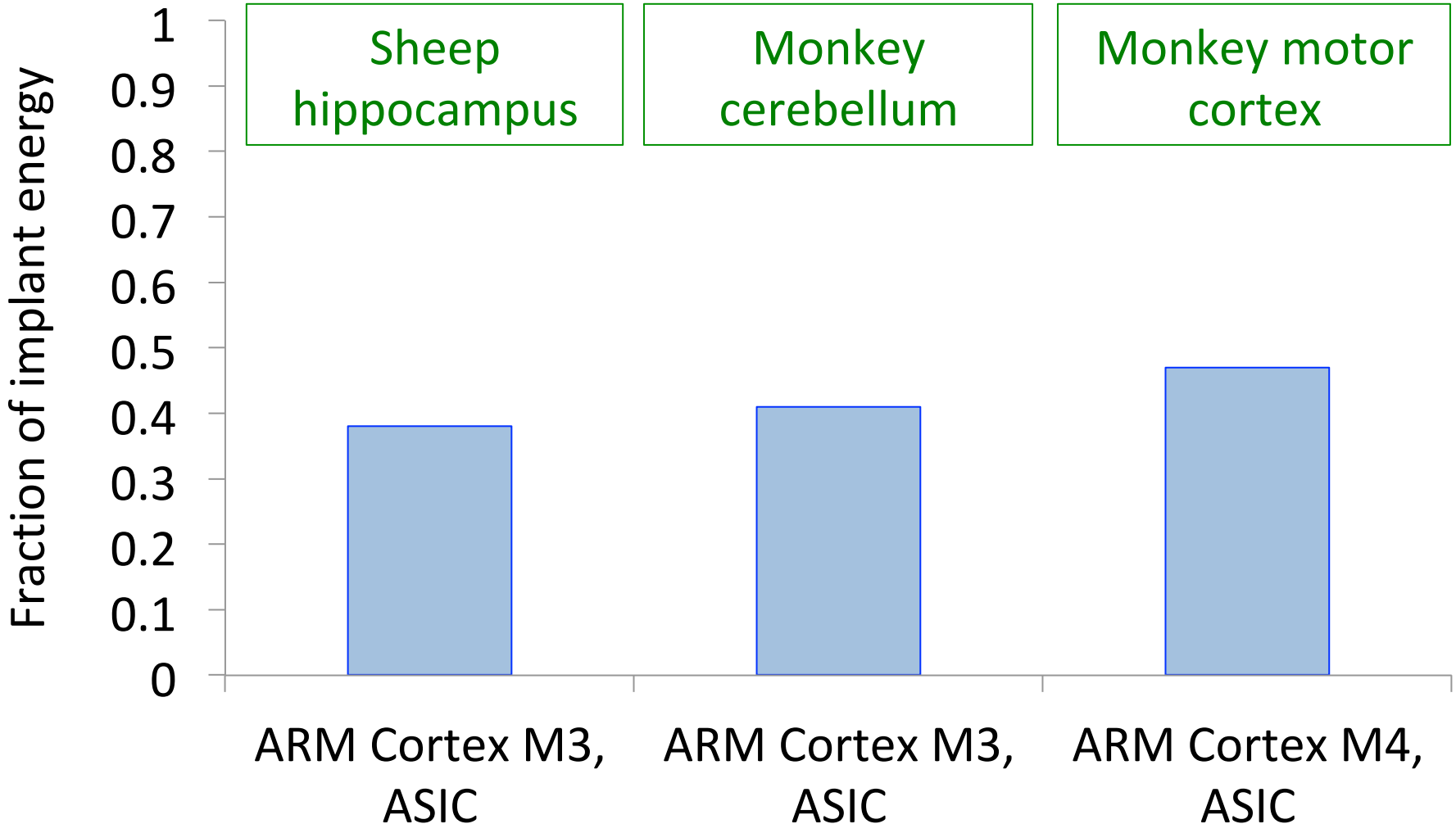
Utah array



External system

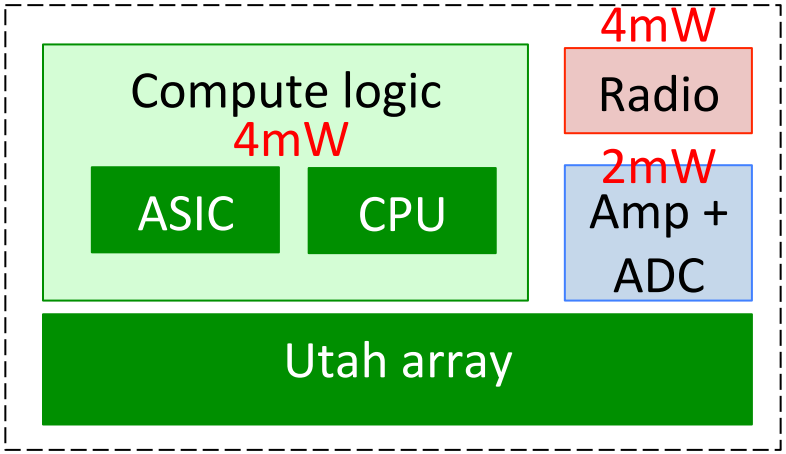


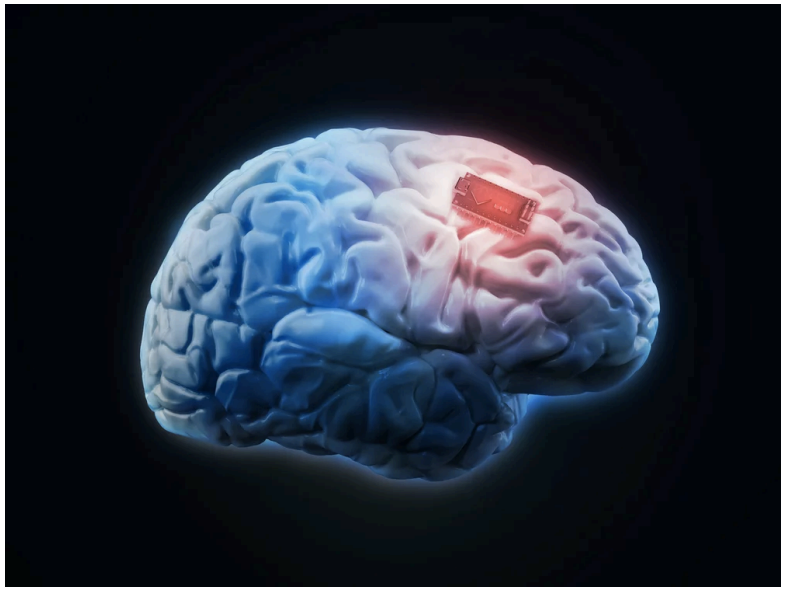
Fraction of implant energy spent on CPU + ASIC



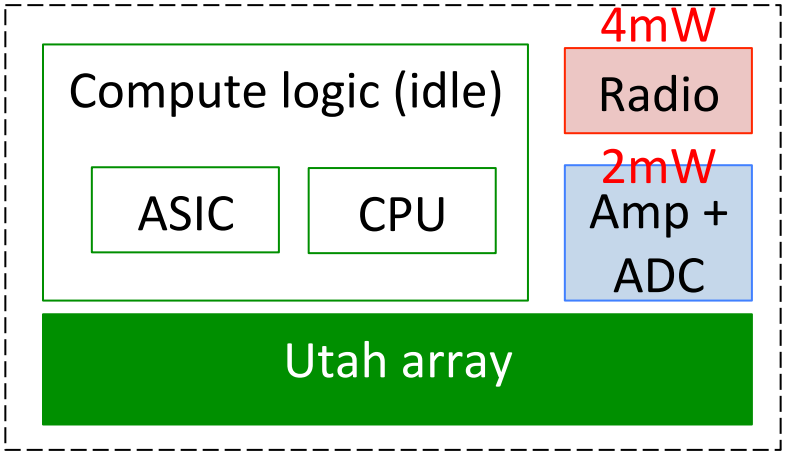
This data was profiled at the Princeton Neuroscience Institute

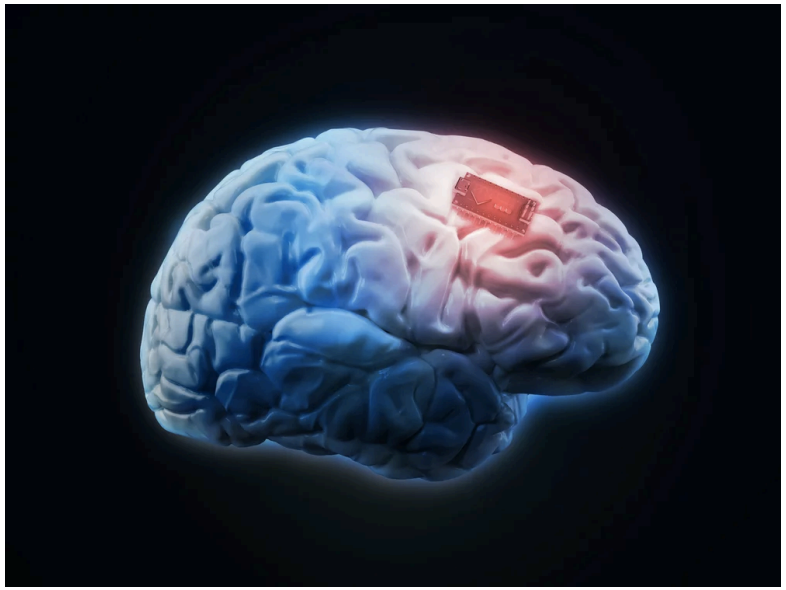
Can we use low power modes?





Save processing energy in the absence of “interesting” neuronal activity



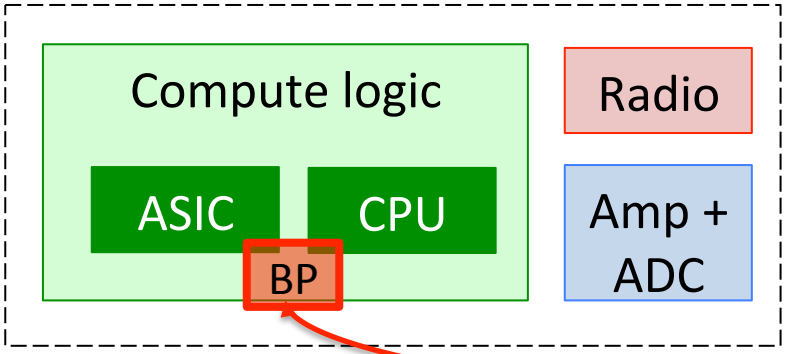


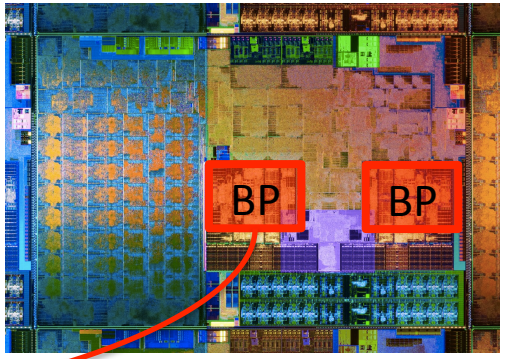
Two caveats

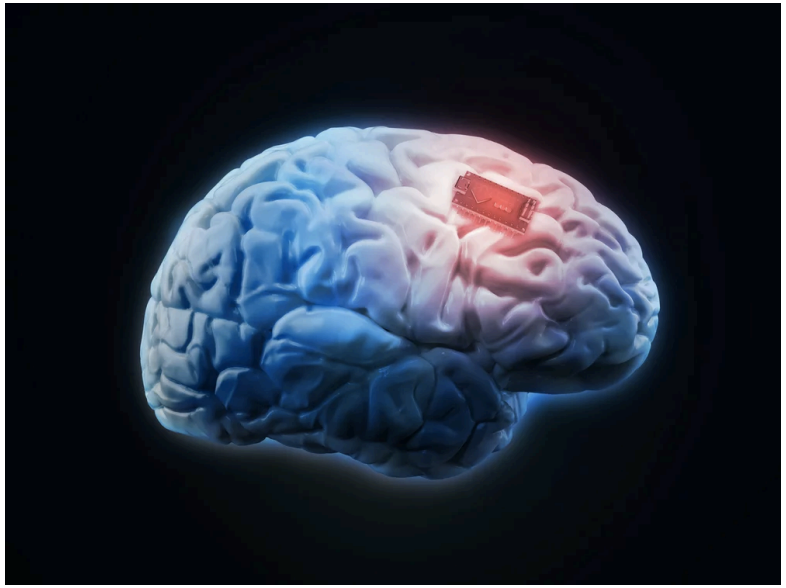
Lose neuronal samples to prolong battery life

Fast wakeup of compute logic for
recording and stimulation

Predict neuronal activity using single-layer perceptron branch predictors (BPs)







What neuronal activity is “interesting”?

Why do we need to predict?

Why is prediction hard?

How are branches similar to neurons?

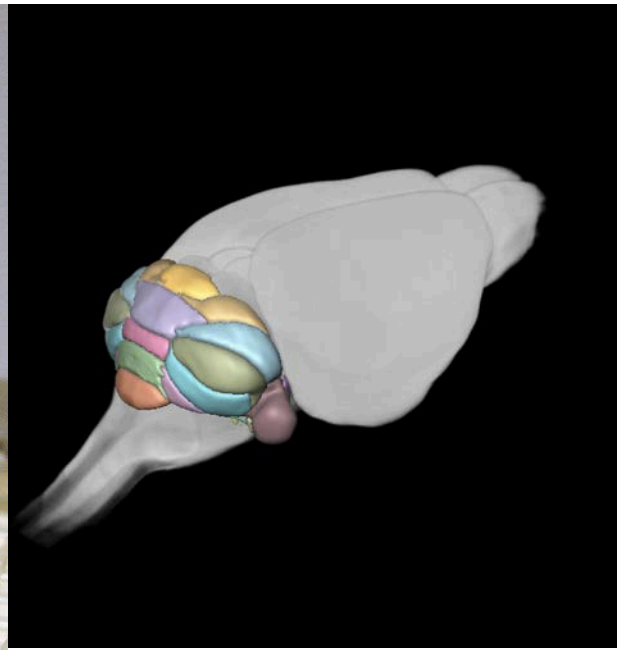
What neuronal activity is “interesting”?

Why do we need to predict?

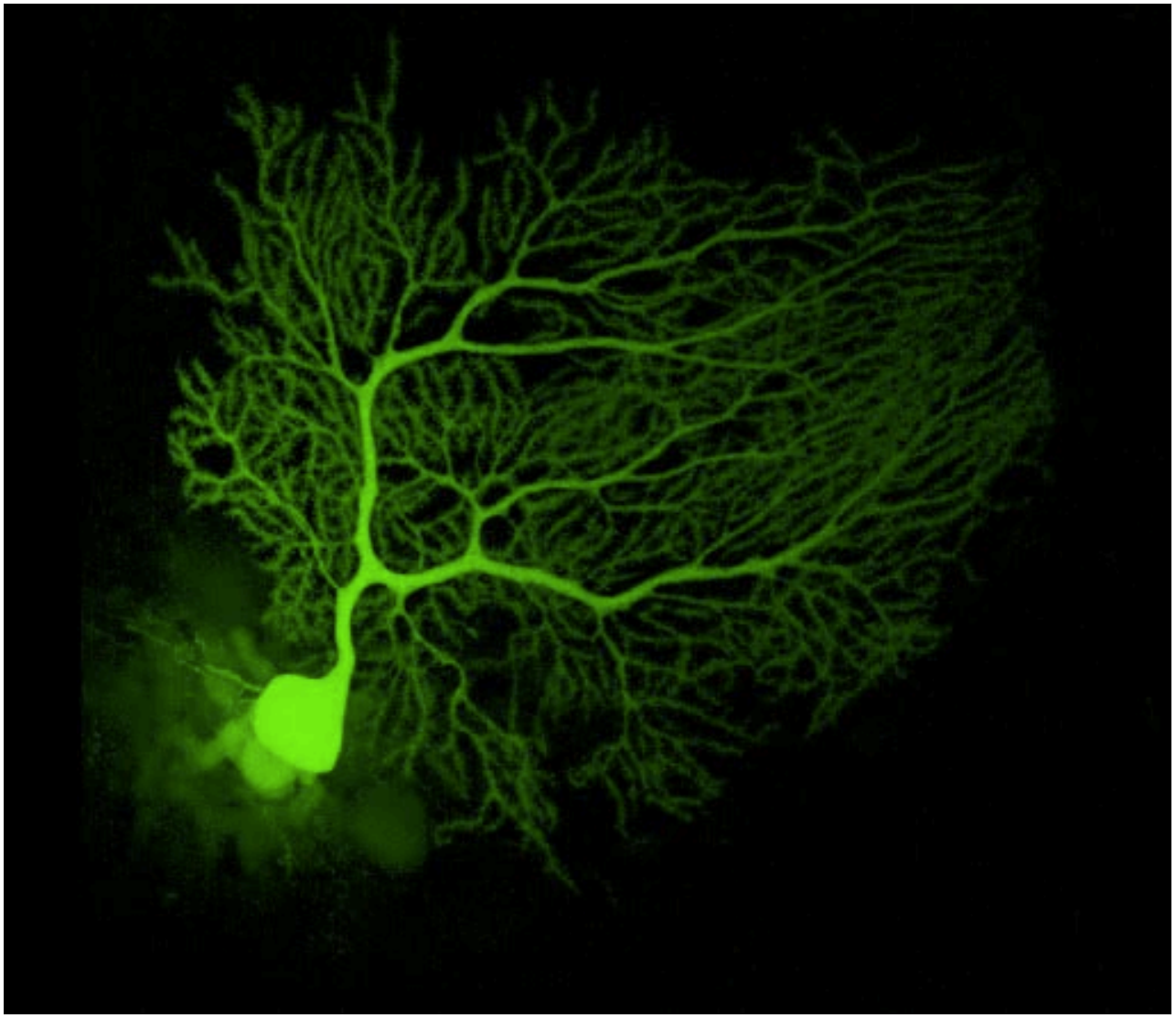
Why is prediction hard?

How are branches similar to neurons?

We are designing implants to monitor and stimulate the cerebellum in mice

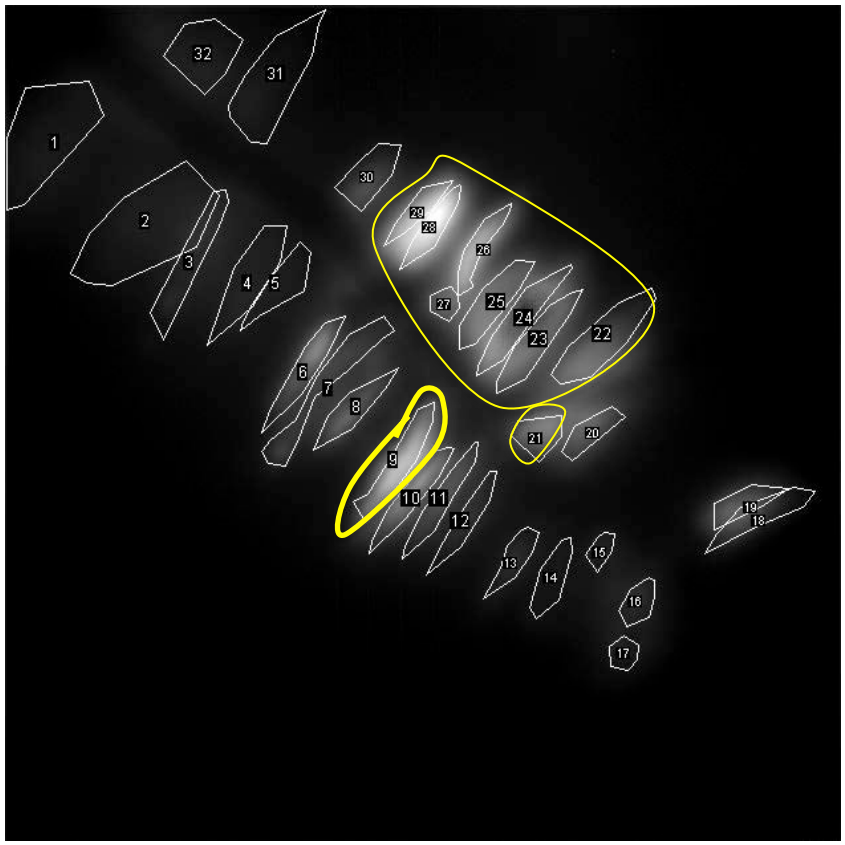


We care about synchronized activity among Purkinje neurons



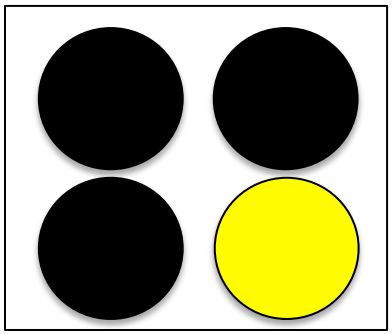
https://ntls.co/images/NTLS/blog/purkinjecells.jpg

Calcium imaging of Purkinje activity on lobule 6 of mouse cerebellum

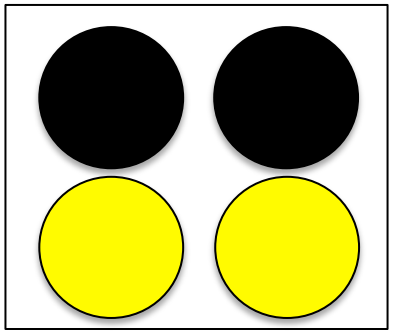


This data was collected at the Princeton Neuroscience Institute

Power management strategy



- No synchronization → low power



- Synchronization → nominal operation

Spiking (S) channel **Quiet (Q)** channel

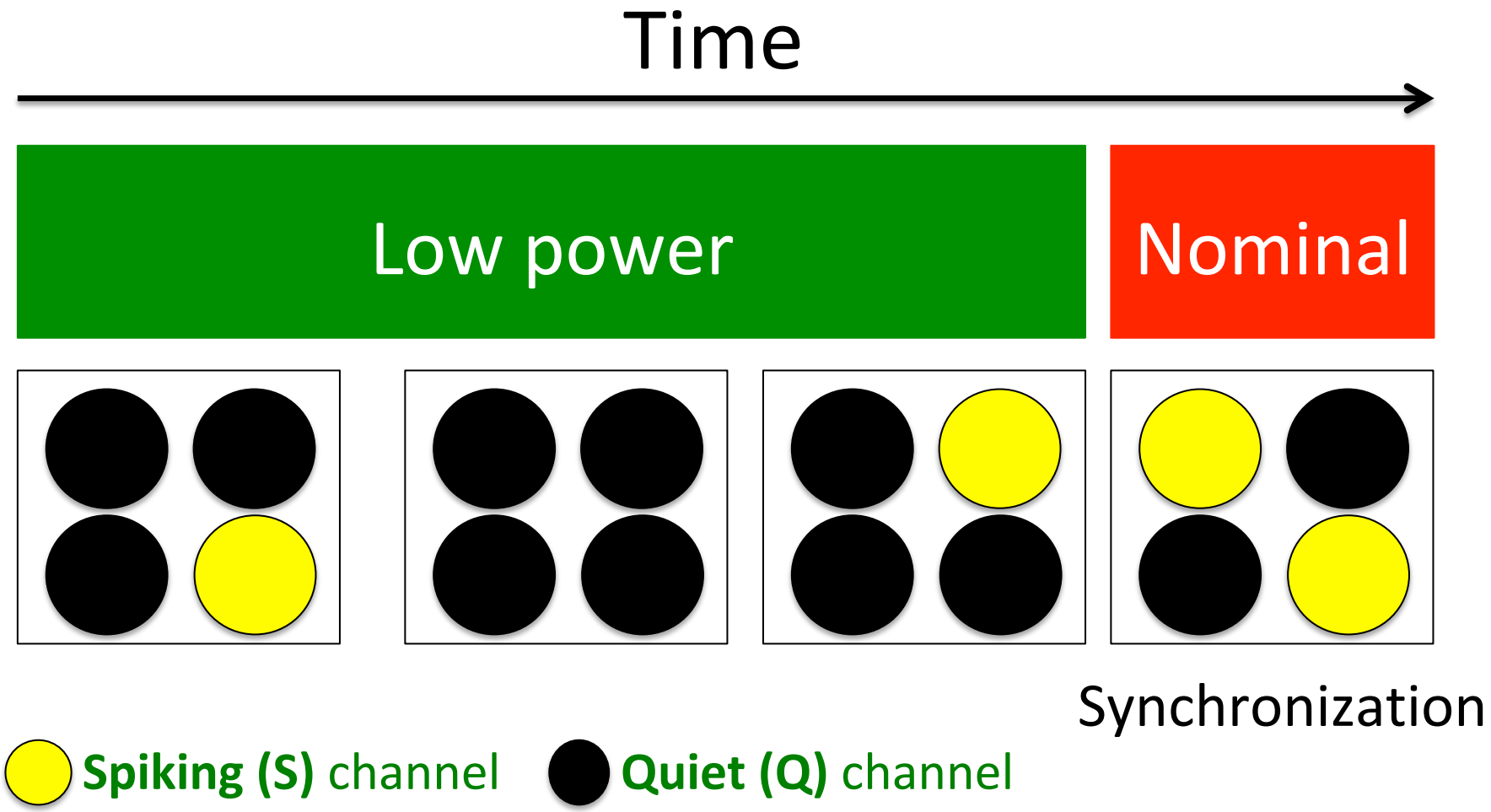
What neuronal activity is “interesting”? ✓

Why do we need to predict?

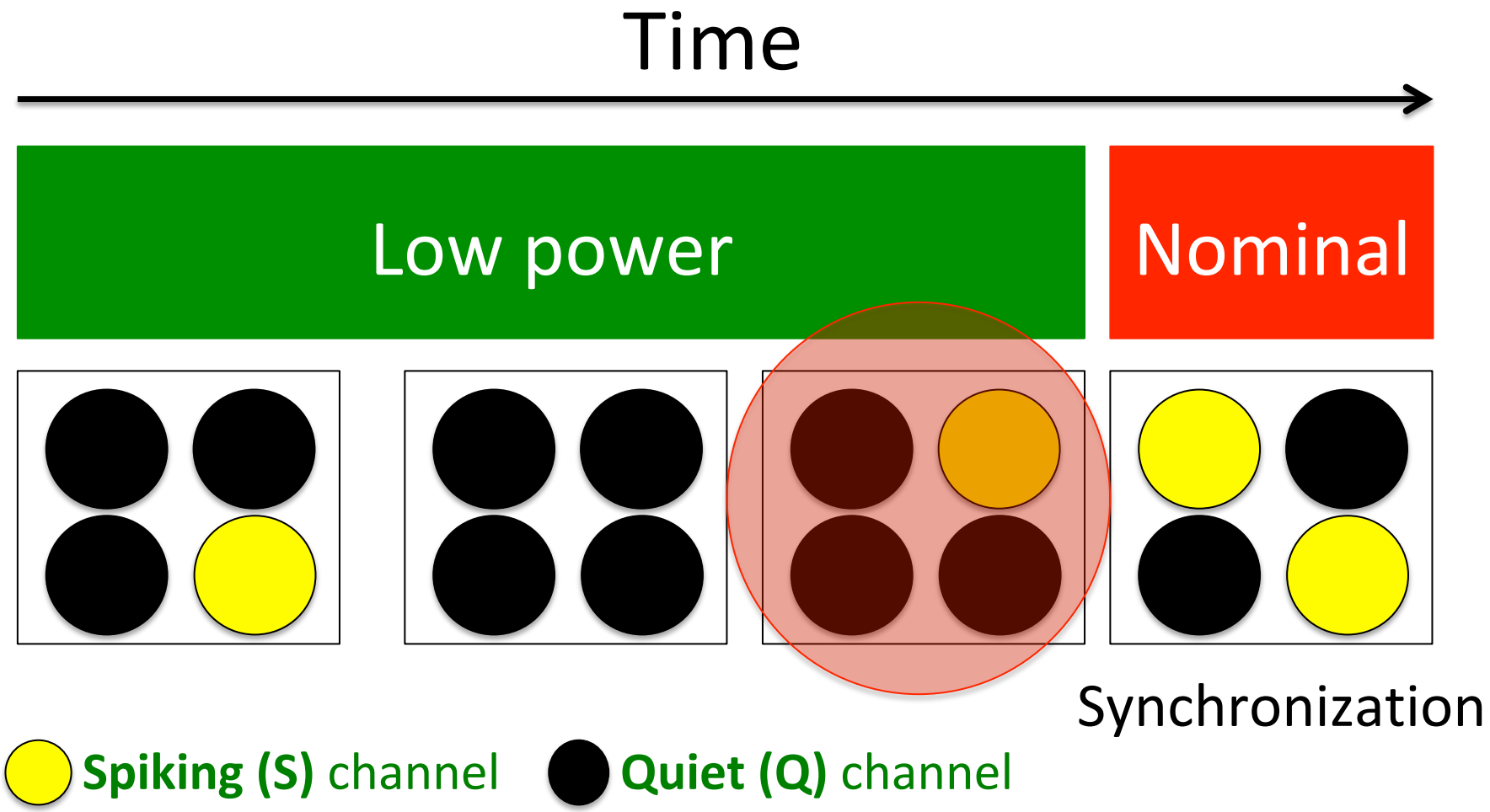
Why is prediction hard?

How are branches similar to neurons?

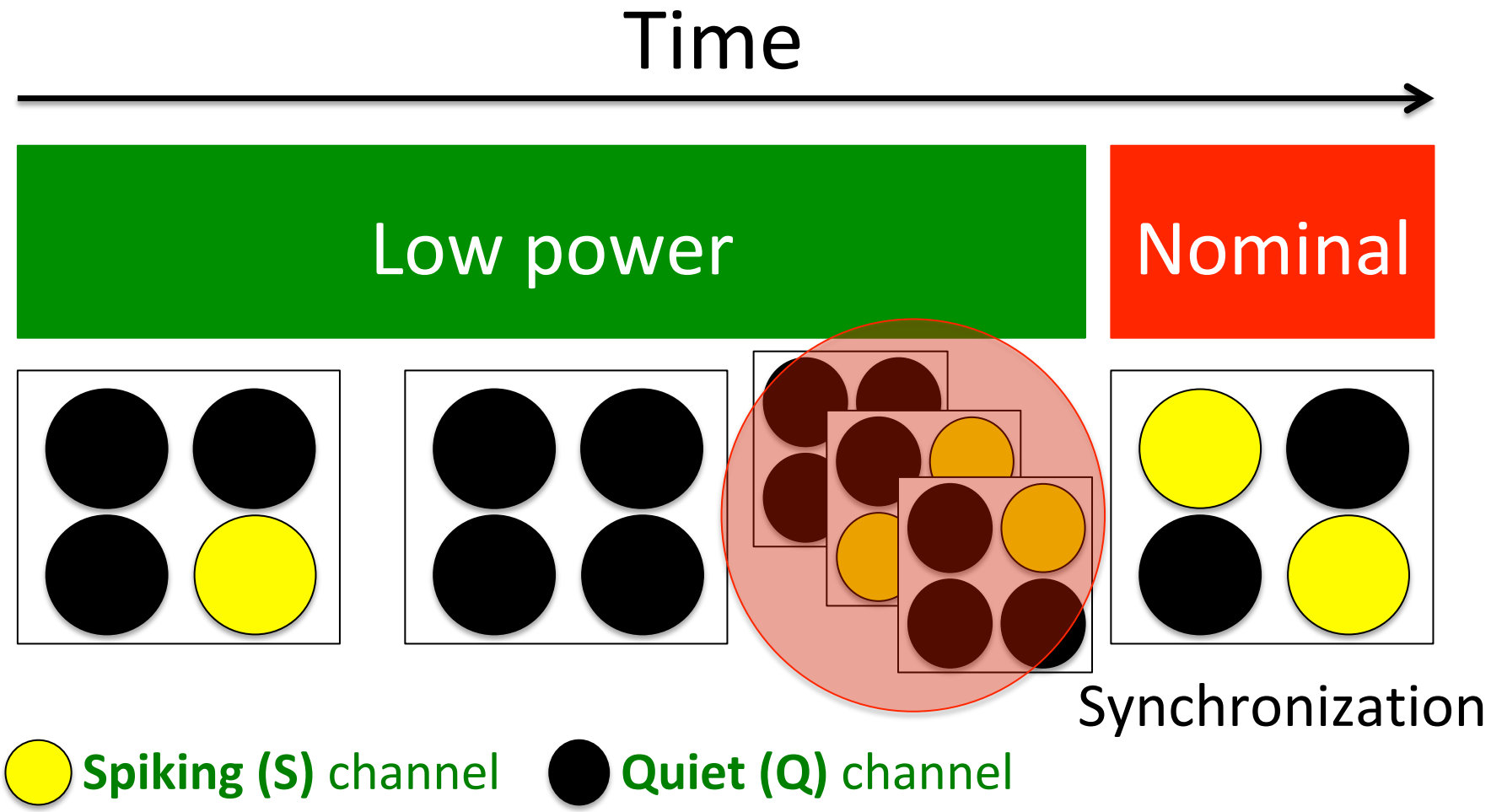
If we were interested in just synchronization, we could be reactive



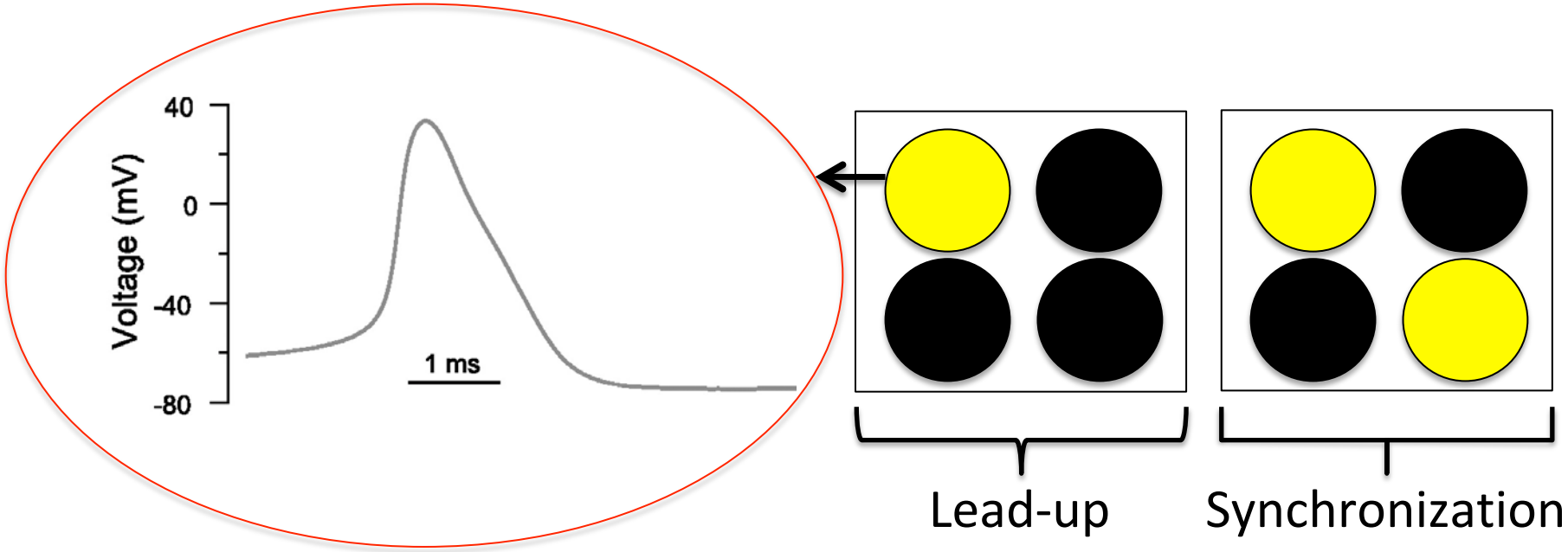
But we care about lead-up activity too



Neuroscientists don’t know what lead-up activity looks like

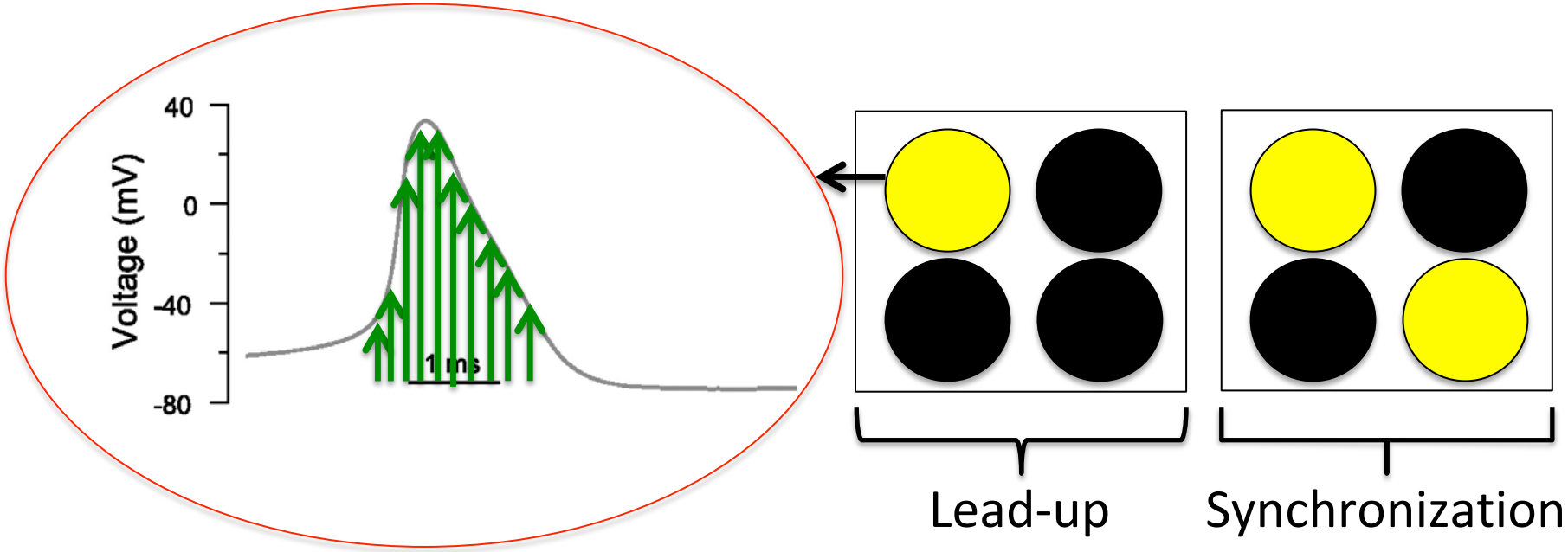


We want to process and react to lead-up in millisecond timescales



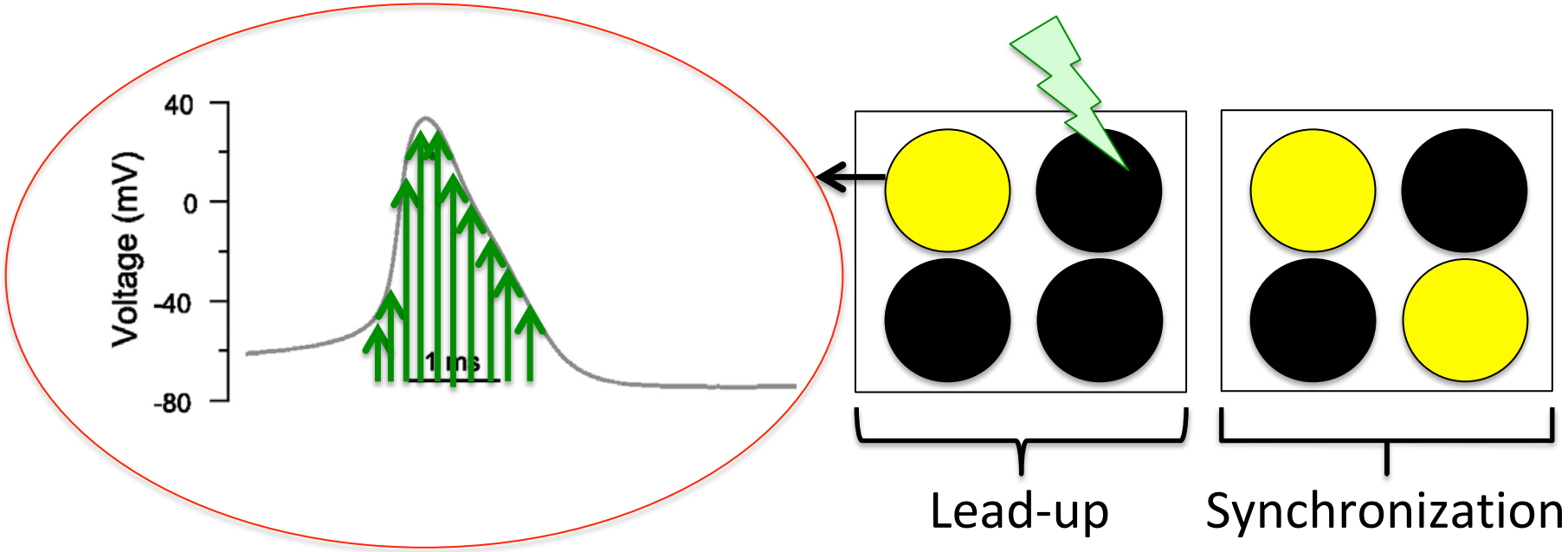
Spiking (S) channel **Quiet (Q)** channel

We want to process and react to lead-up in millisecond timescales



Spiking (S) channel **Quiet (Q)** channel

We want to process and react to lead-up in millisecond timescales



Spiking (S) channel **Quiet (Q)** channel

What neuronal activity is “interesting”? ✓

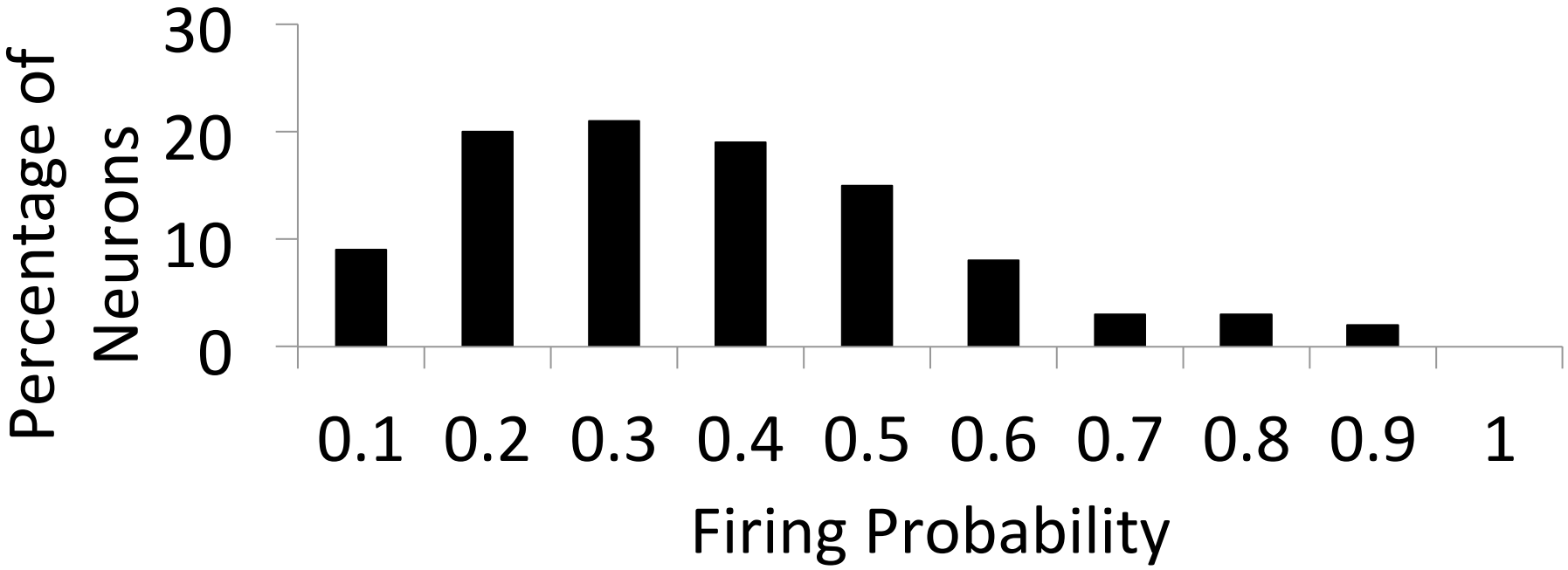
Why do we need to predict? ✓

Why is prediction hard?

How are branches similar to neurons?

Behavior varies among neurons

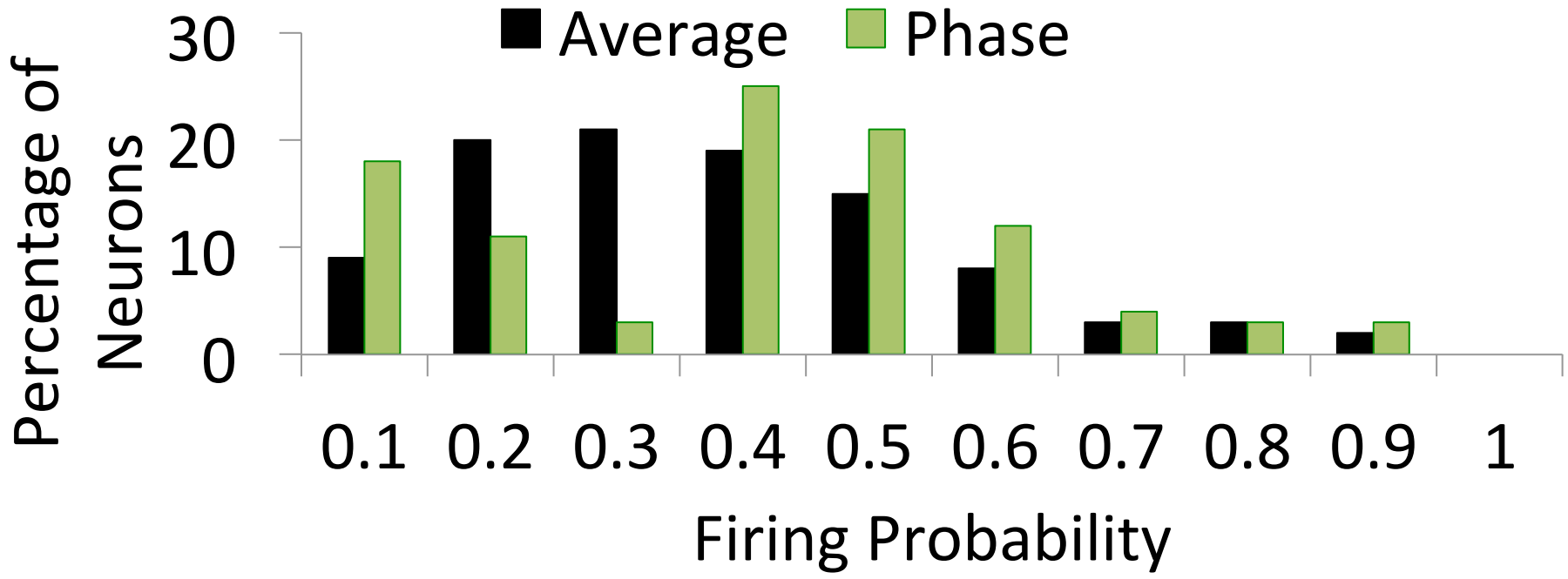
Craniotomy on lobule 6 of cerebellum
Average (26 minutes)



This data was profiled at the Princeton Neuroscience Institute

Behavior varies for the same neuron over time

Craniotomy on cerebellum lobule 6, 20-40 psi air puffs on whiskers
Average (26 minutes), Phase (5 seconds)



This data was profiled at the Princeton Neuroscience Institute

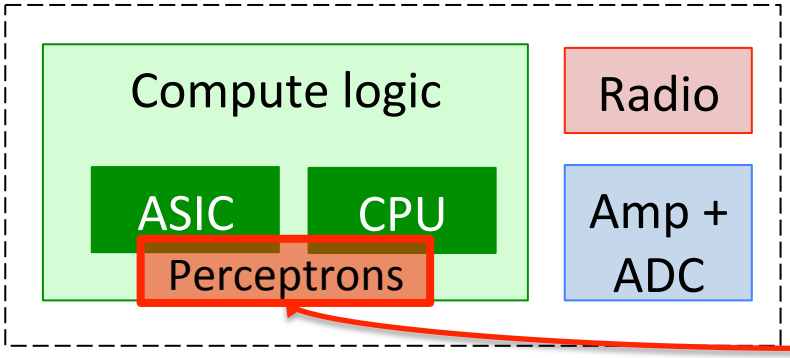
What neuronal activity is “interesting”? ✓

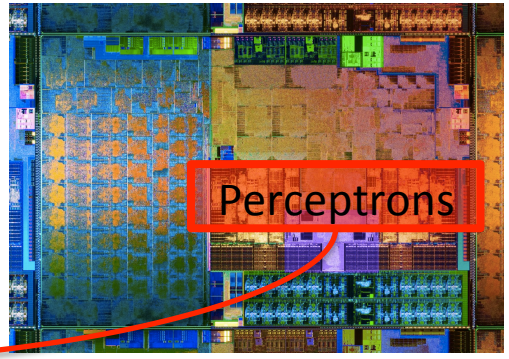
Why do we need to predict? ✓

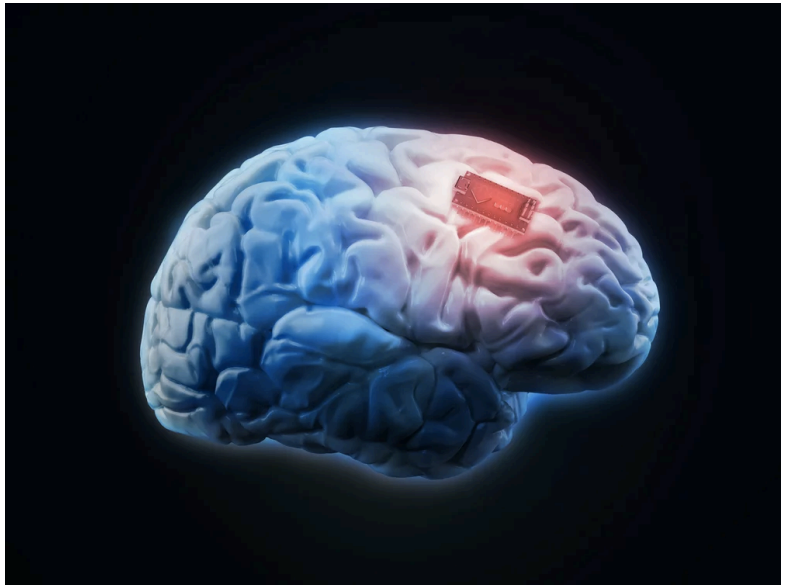
Why is prediction hard? ✓

How are branches similar to neurons?

Co-opt single-layer perceptron BPs to predict lead-up + synchronized activity







How are branches similar to neurons?

Binary behavior

Correlations

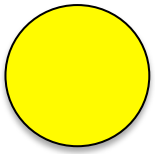
Program branches are either **taken** or **not taken**

```
for(i=0; i < 10; i++)
        {
          /* stuff */
        }
/* other stuff */
```

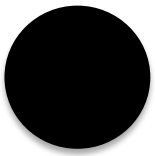
Taken (T)

Not taken (NT)

Biological neurons are **spiking** or **quiet**



Spiking (S) channel



Quiet (Q) channel

Program branches are **correlated**

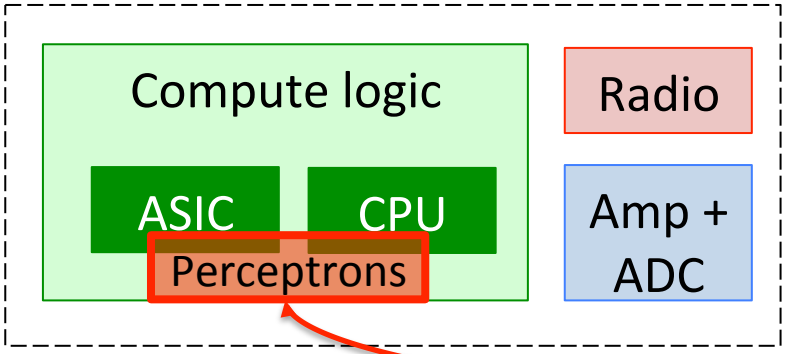
```
p = findNode(foo);
if ( p is parent )
  do something;
do other stuff;
if ( p is a child)
  do something else;
```

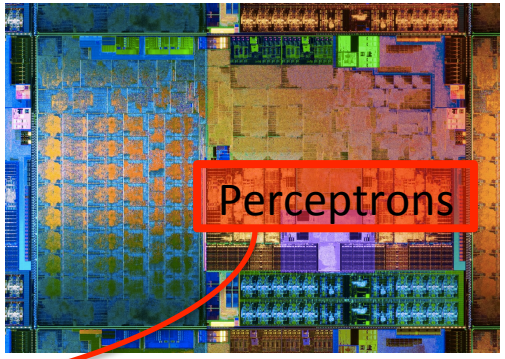

Branch 2 is opposite of branch 1

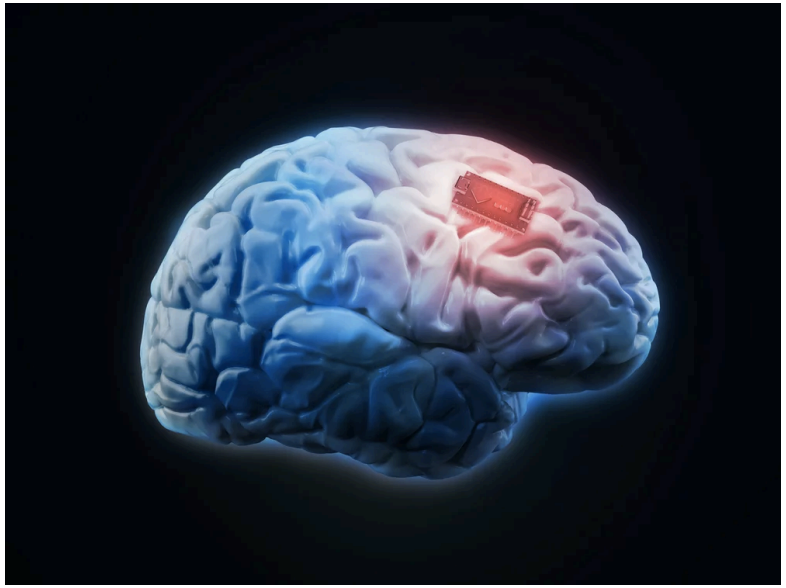
Microgrids of tens of Purkinje neurons are **co-activated**

Ilker Ozden et al. [Journal of Neuroscience, 2009]
Sullivan et al. [Journal of Neurophysiology, 2005]
Tank et al. [Science, 1998]

Co-opt single-layer perceptron BPs to predict lead-up + synchronized activity







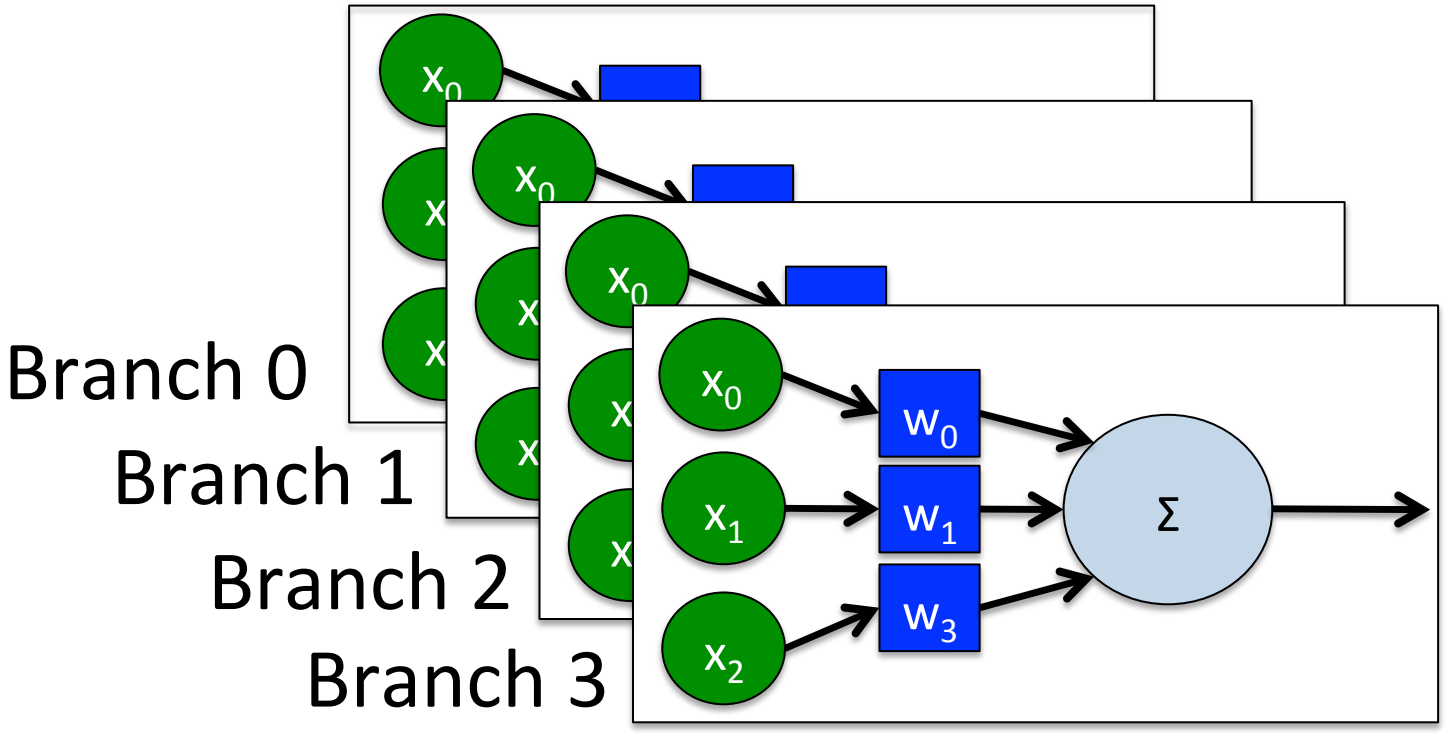
What neuronal activity is “interesting”? ✓

Why do we need to predict? ✓

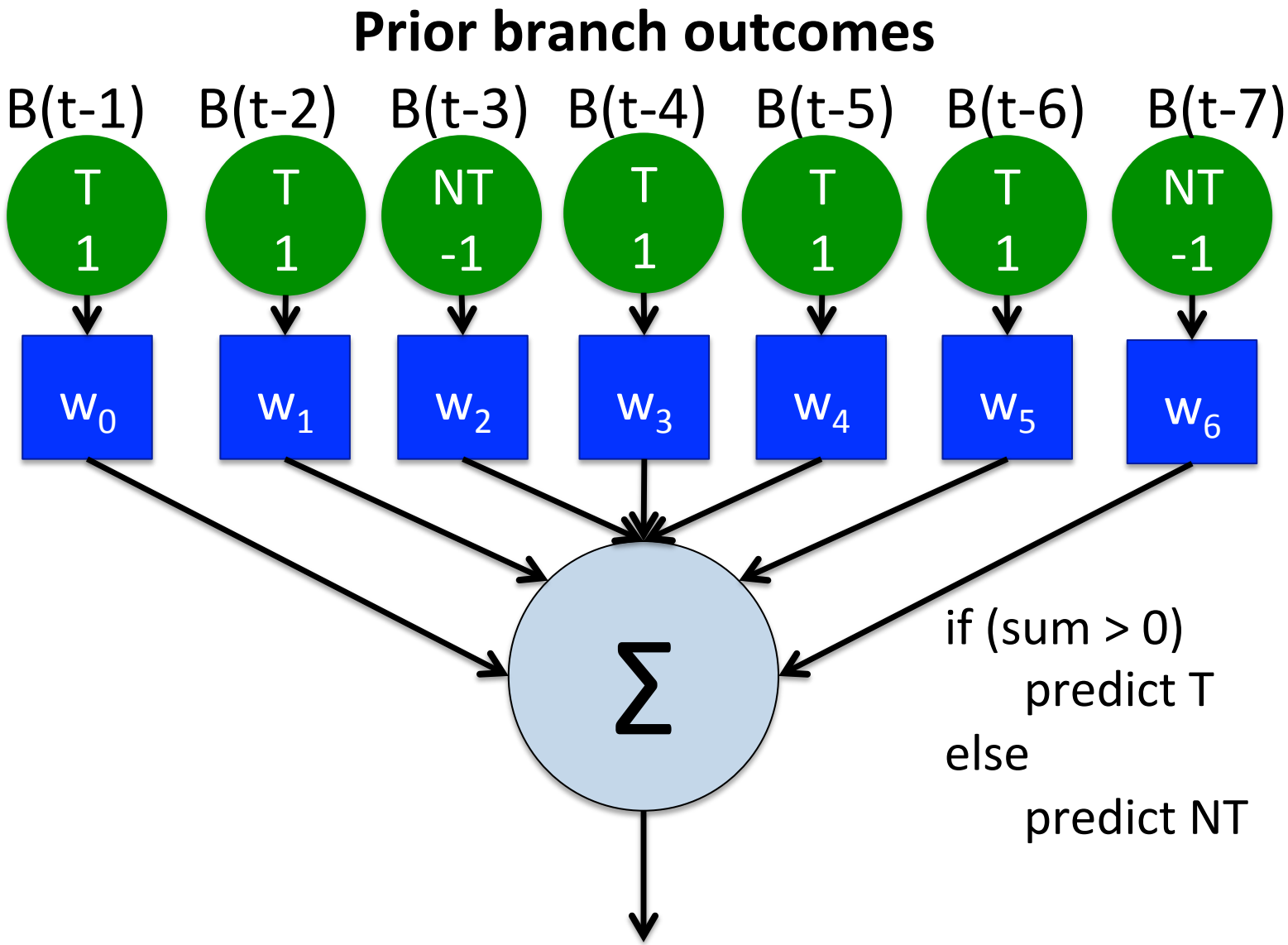
Why is prediction hard? ✓

How are branches similar to neurons? ✓

Predict program branches using perceptrons

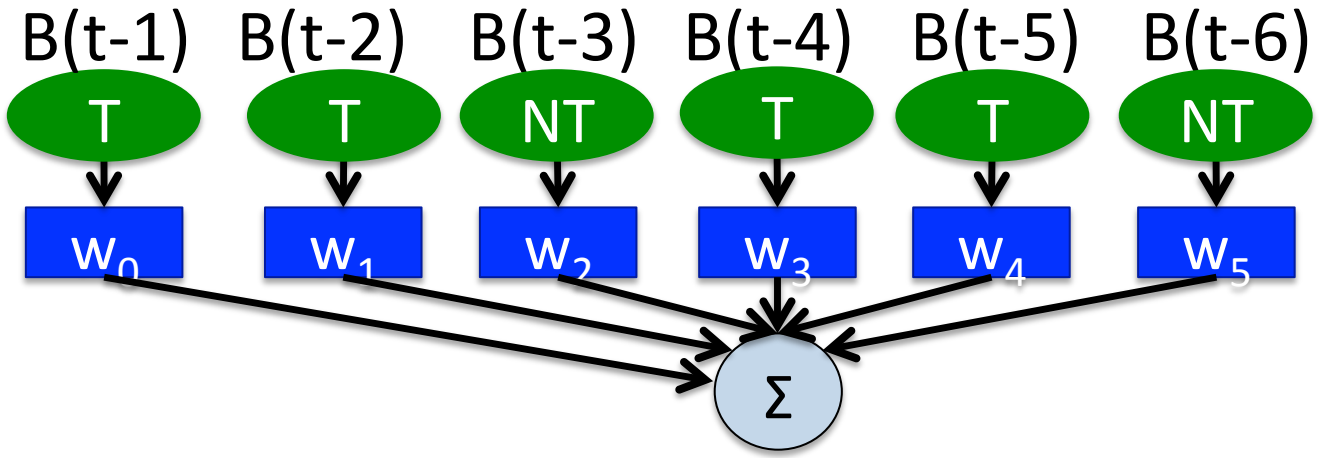


Perceptron branch predictor for one branch at time t



Predictor for one branch at time t

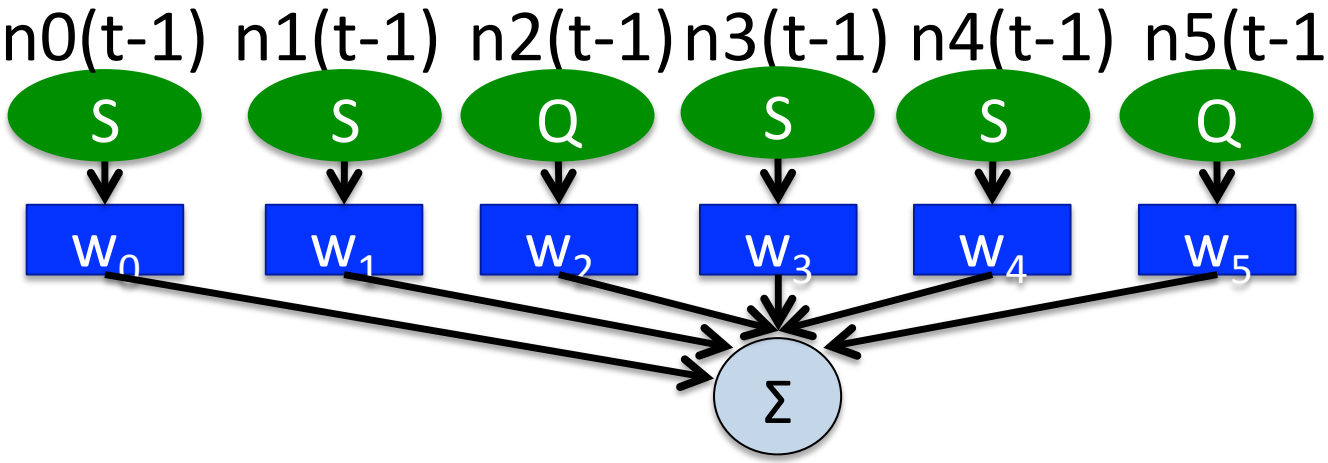
Prior branch outcomes




```
if (sum > 0)
    predict T
else
    predict NT
```

Predictor for one neuron at time t

Prior neuron channel activity



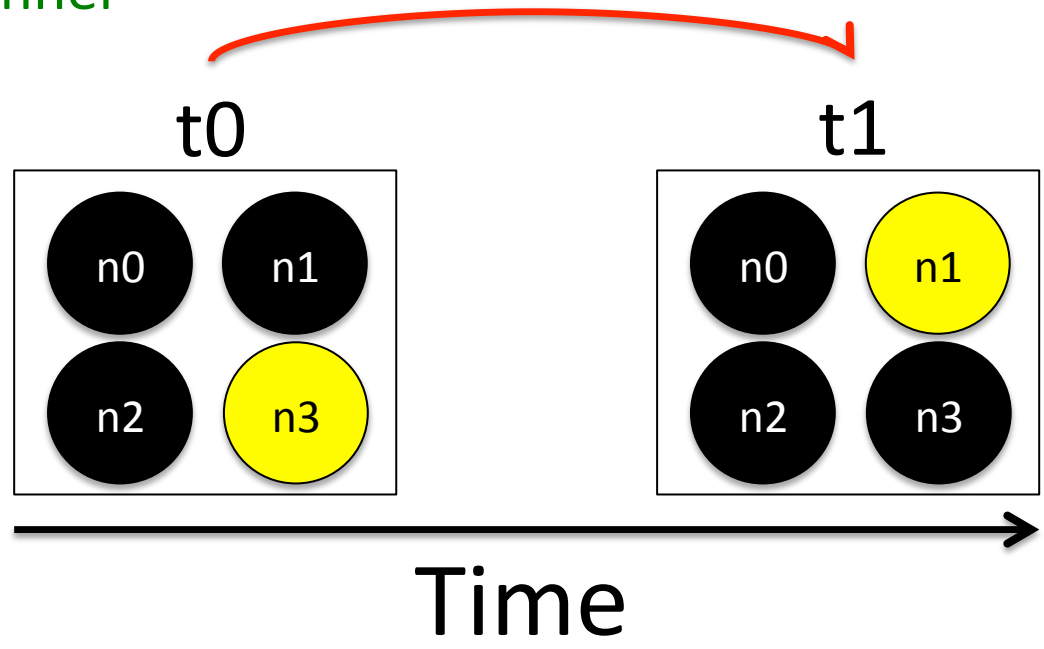

```
if (sum > 0)
    predict S
else
    predict Q
```

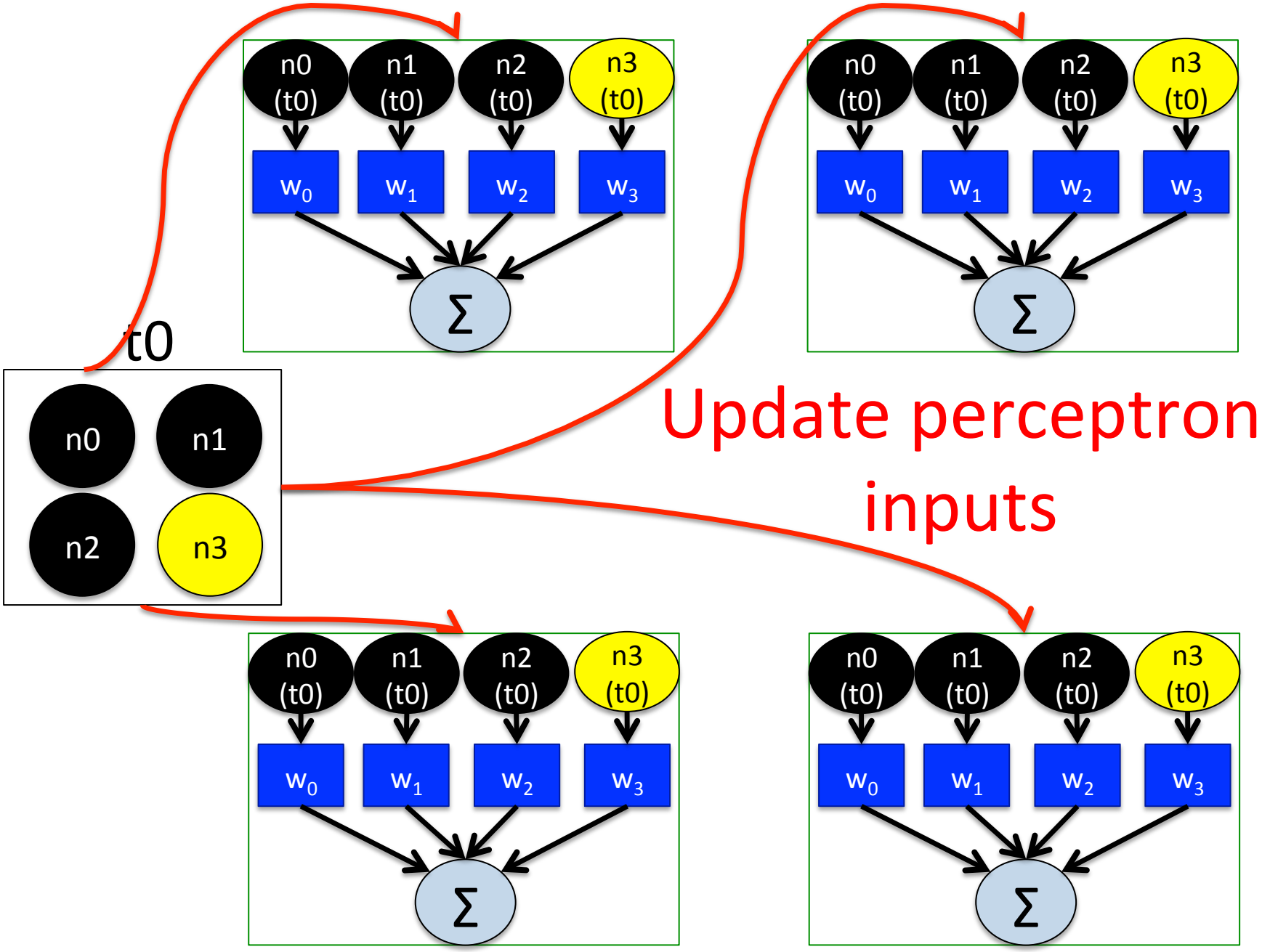
Using perceptrons to predict neuron behavior

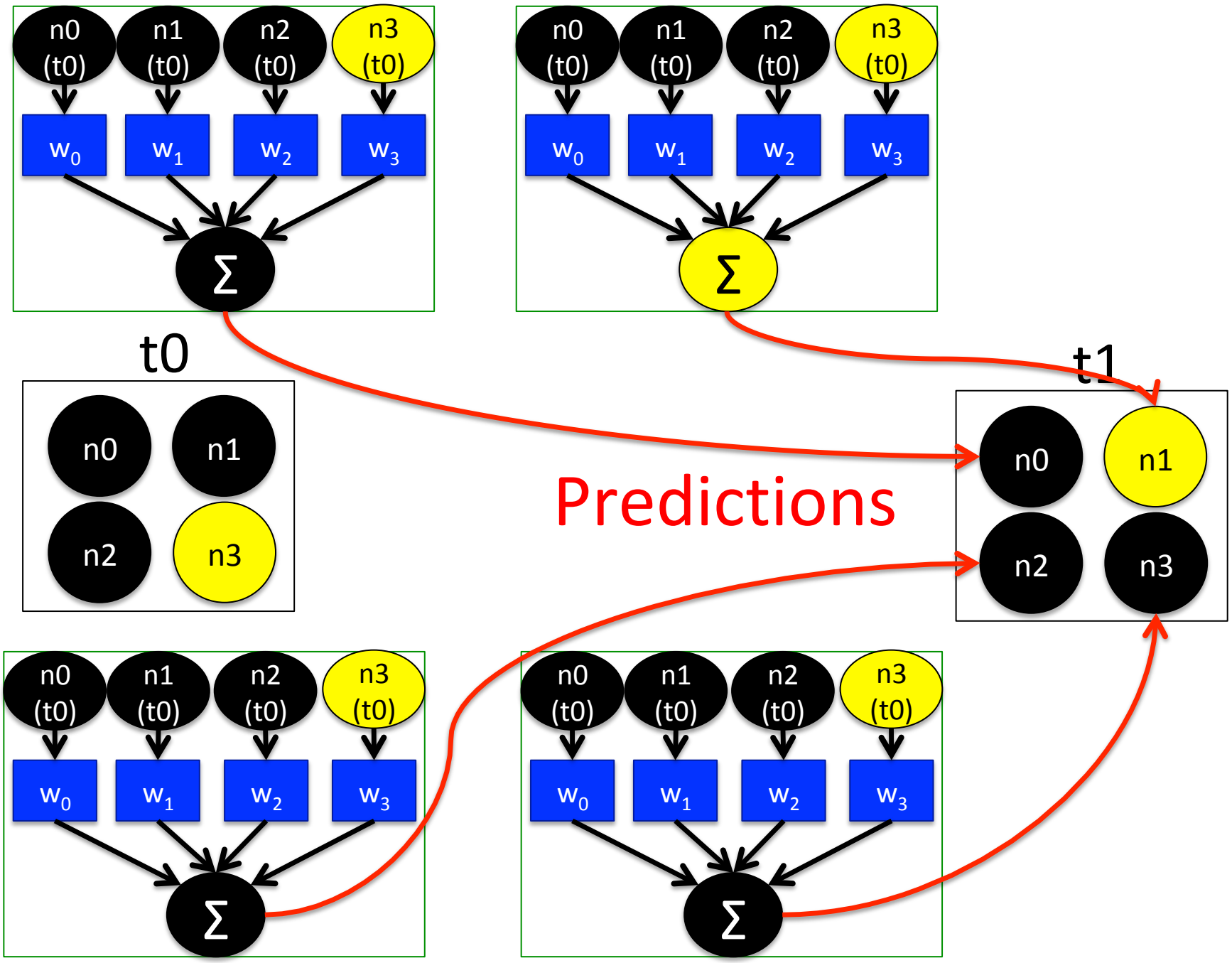
Predictions for t1 in t0

Spiking (S) channel

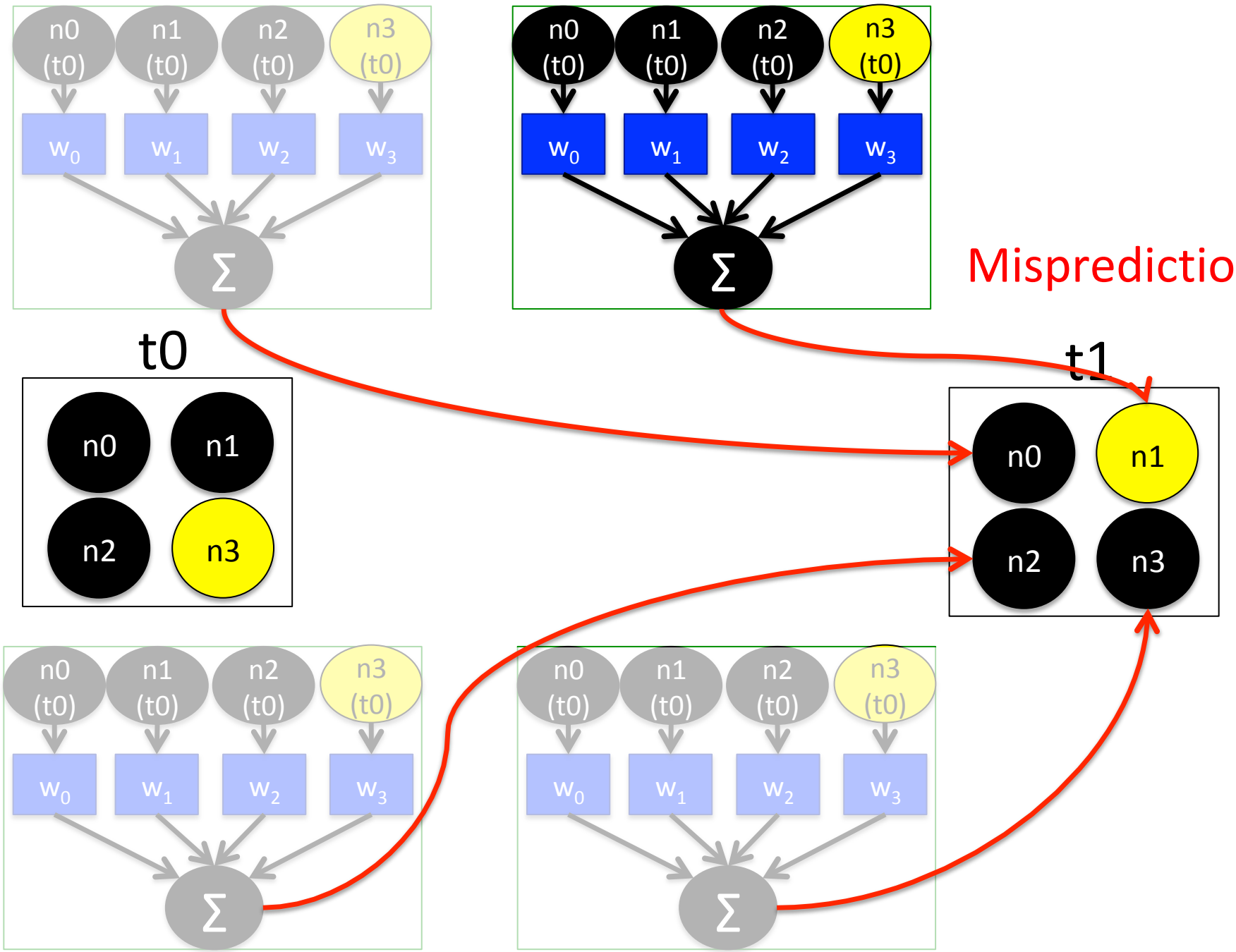
Quiet (Q) channel




t0
n0
n1
n2
n3
n0 (t0)
n1 (t0)
n2 (t0)
n3 (t0)
w0
w1
w2
w3
Σ
Update perceptron inputs

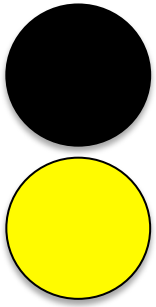

n0 (t0)
n1 (t0)
n2 (t0)
n3 (t0)
w0
w1
w2
w3
Σ
t0
n0
n1
n2
n3
Predictions
t1

We train the perceptrons online;
e.g., on mispredictions


n0 (t0)
n1 (t0)
n2 (t0)
n3 (t0)
w0
w1
w2
w3
Σ
Misprediction!
t0
n0
n1
n2
n3
t1
n0
n1
n2
n3

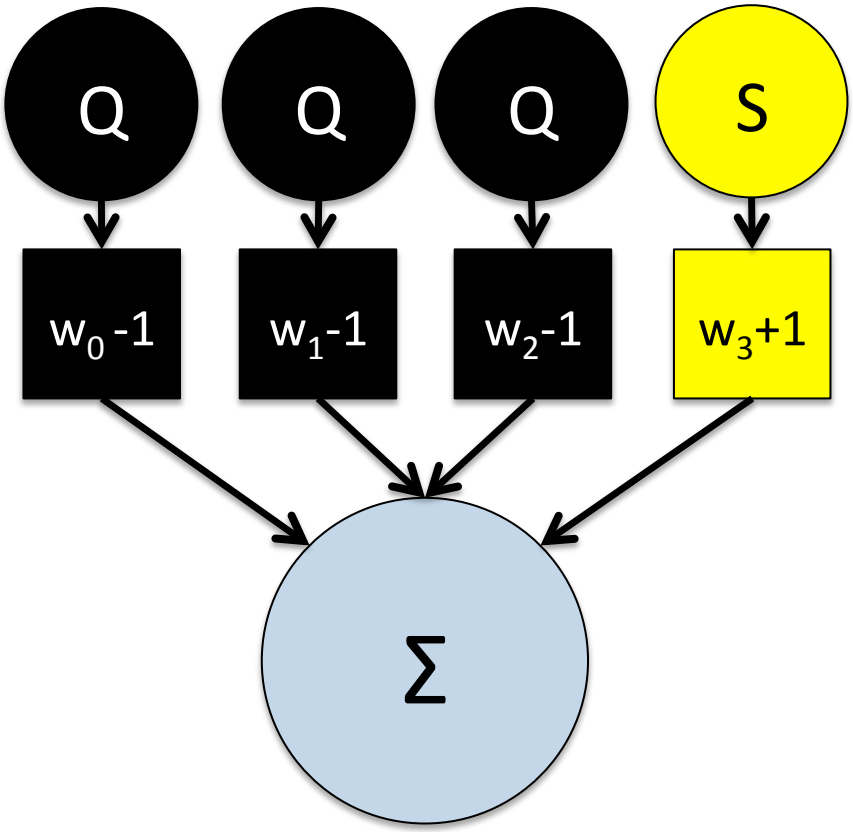
t0 → predict **quiet**

t1 → actual outcome is **spike**



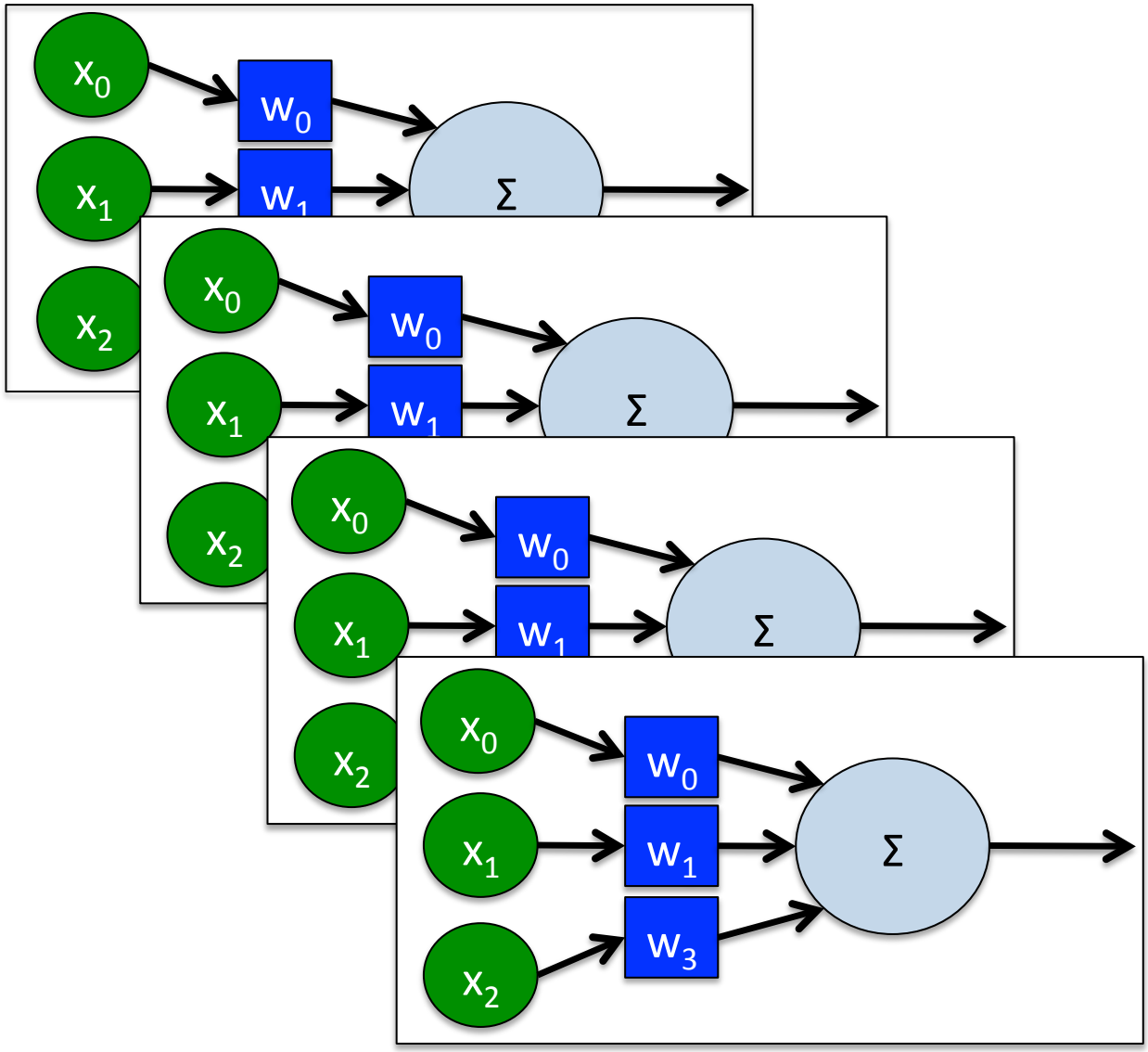
Weights-1 for neuron channels that were **quiet** in t0

Weights+1 for neuron channels that were **spiking** in t0

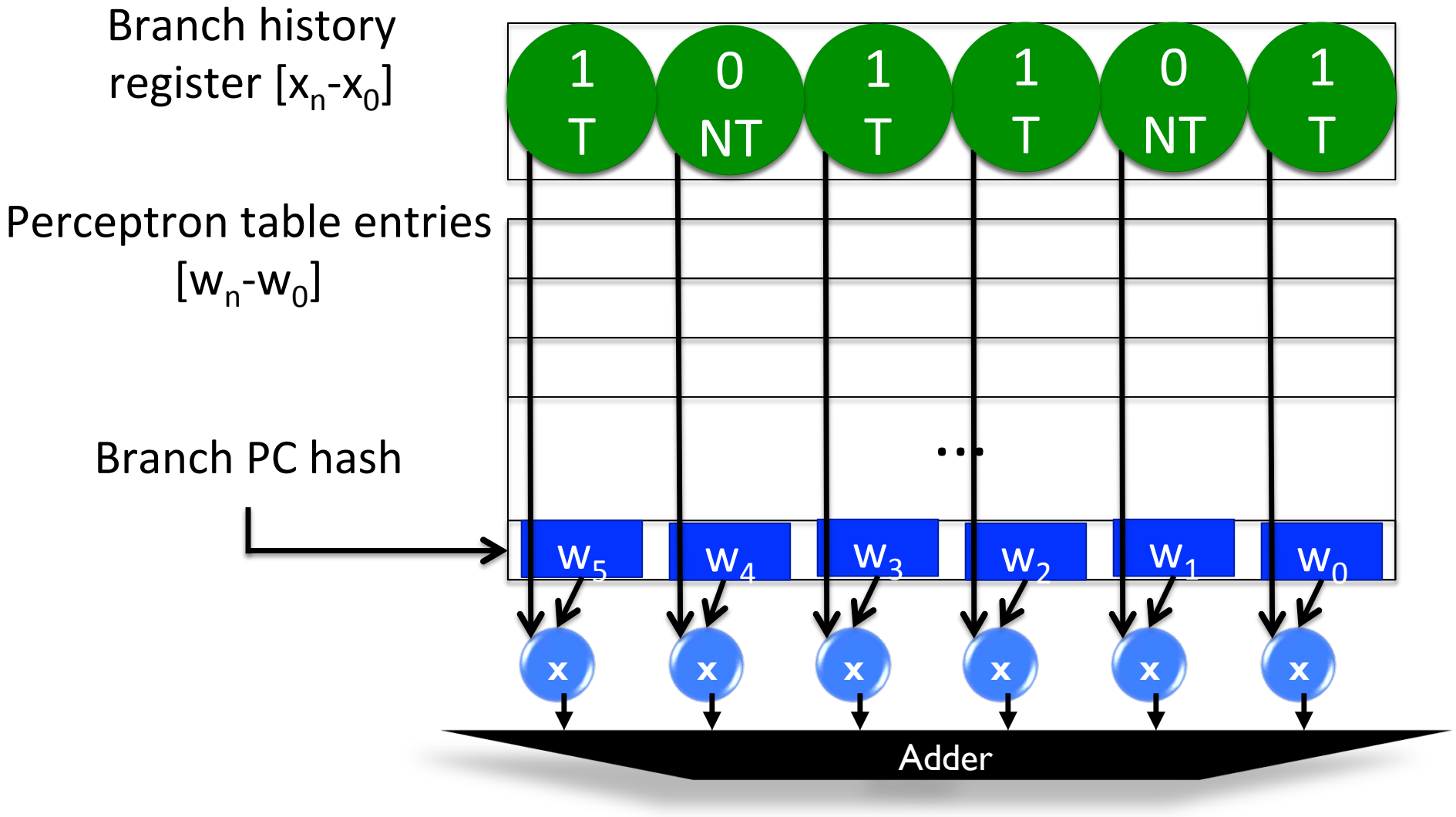


So what does the hardware look like?

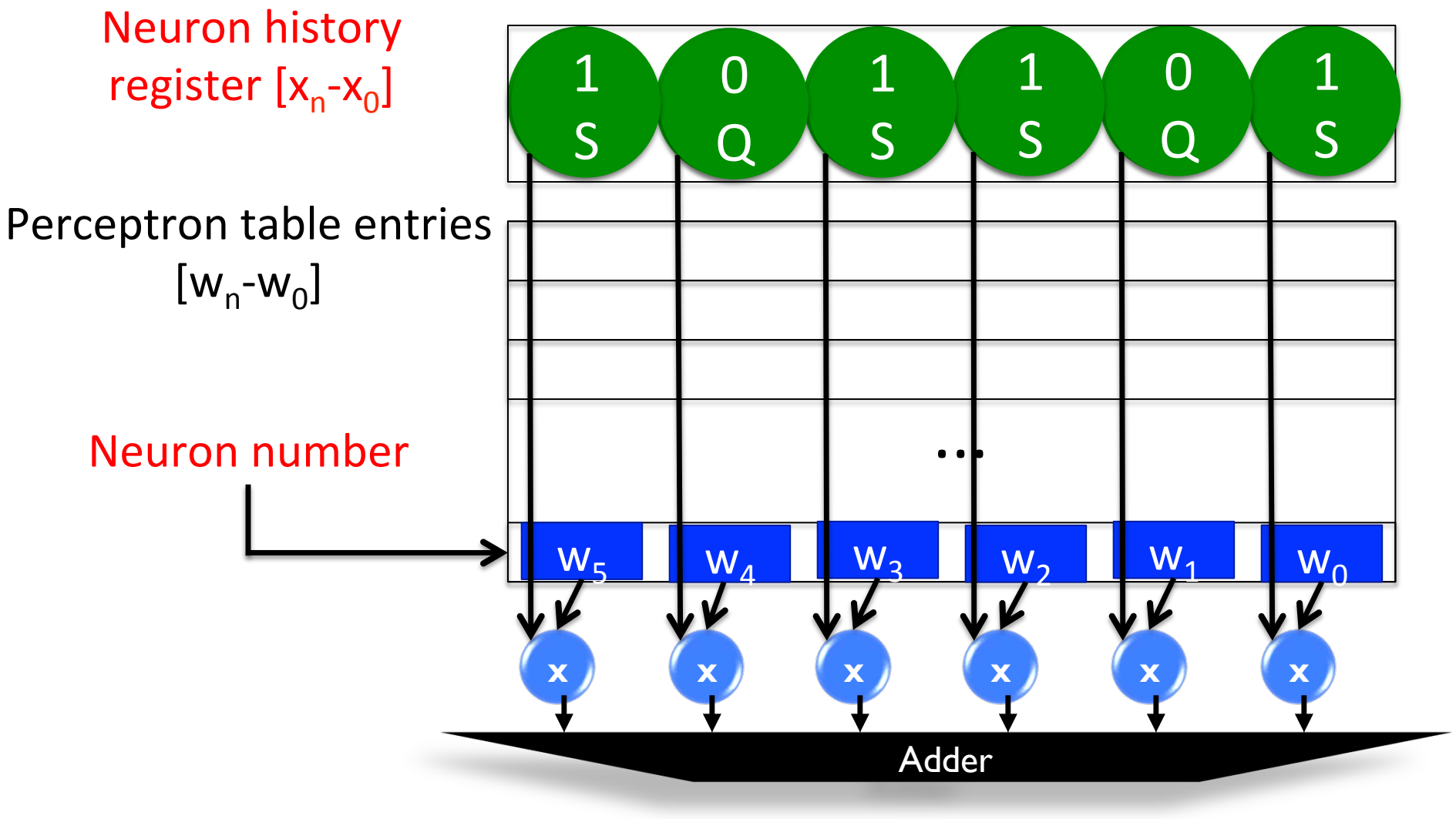
Multiple perceptrons implemented in hardware



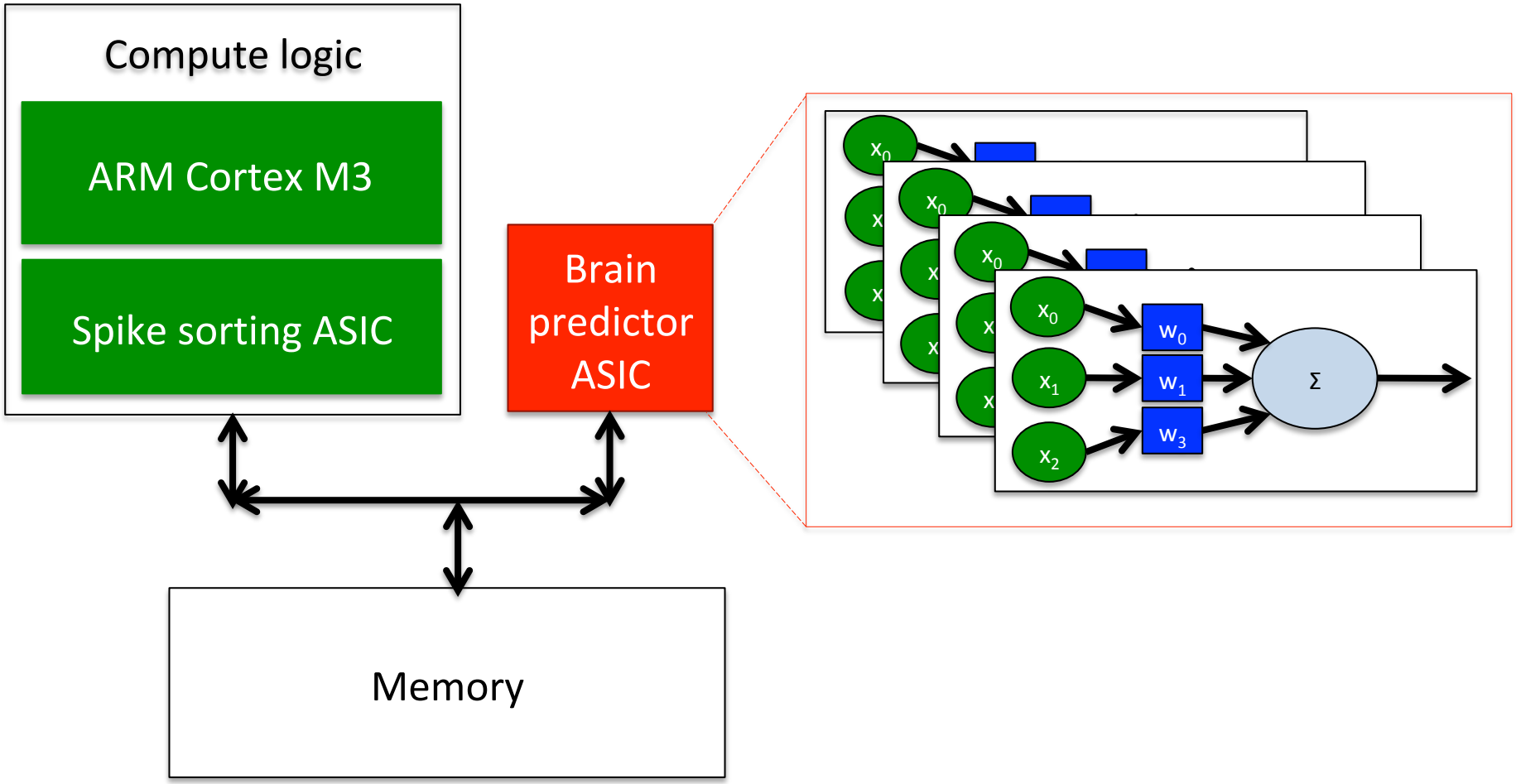
Perceptron branch predictor implementation



Neuronal prediction requires minor hardware changes



We add a perceptron-based ASIC to wake up the compute logic



More details on misprediction

Predict synch, Outcome no synch →
wasted energy

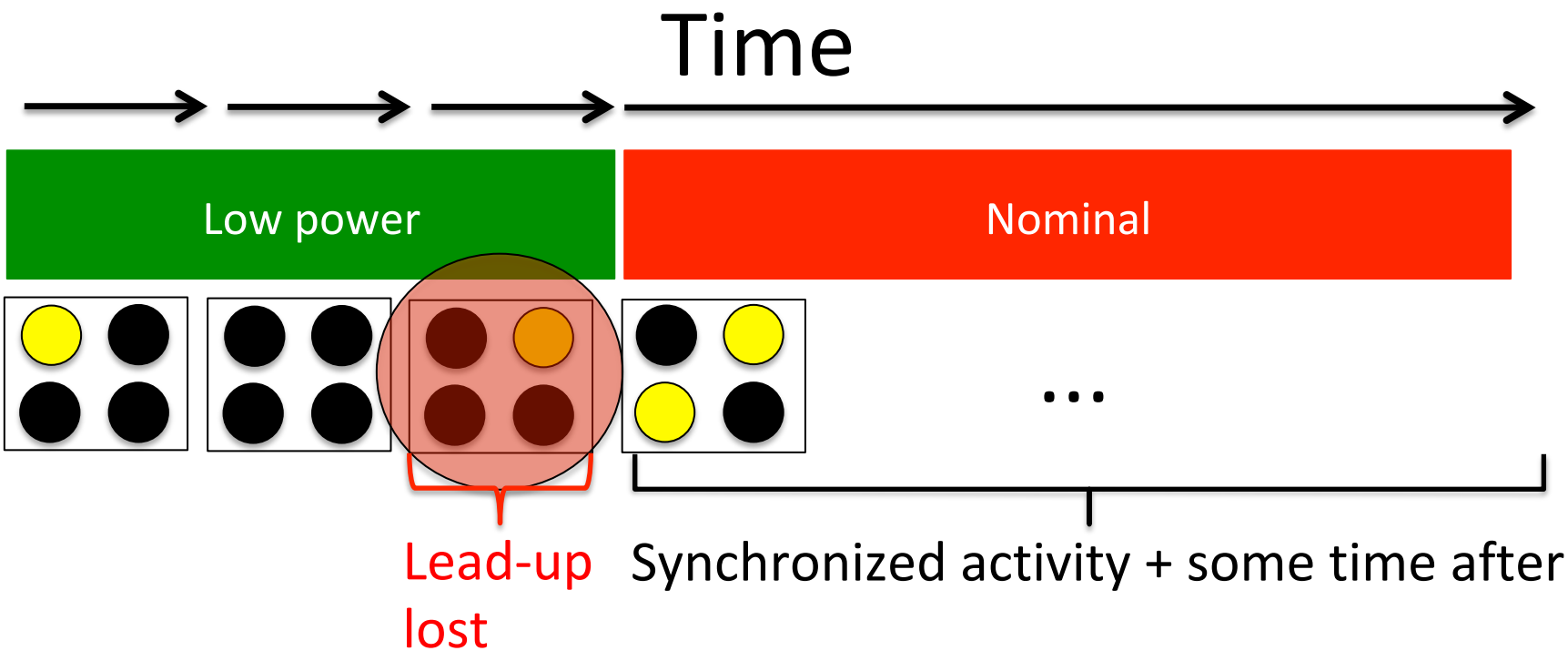
Predict no synch, Outcome synch →
Lose lead-up and synchronization activity

More details on misprediction

Predict synch, Outcome no synch →
wasted energy

**Predict no synch, Outcome synch →
Lose lead-up and synchronization activity**

Use reactive approach to capture synchronization



Spiking (S) channel

Quiet (Q) channel

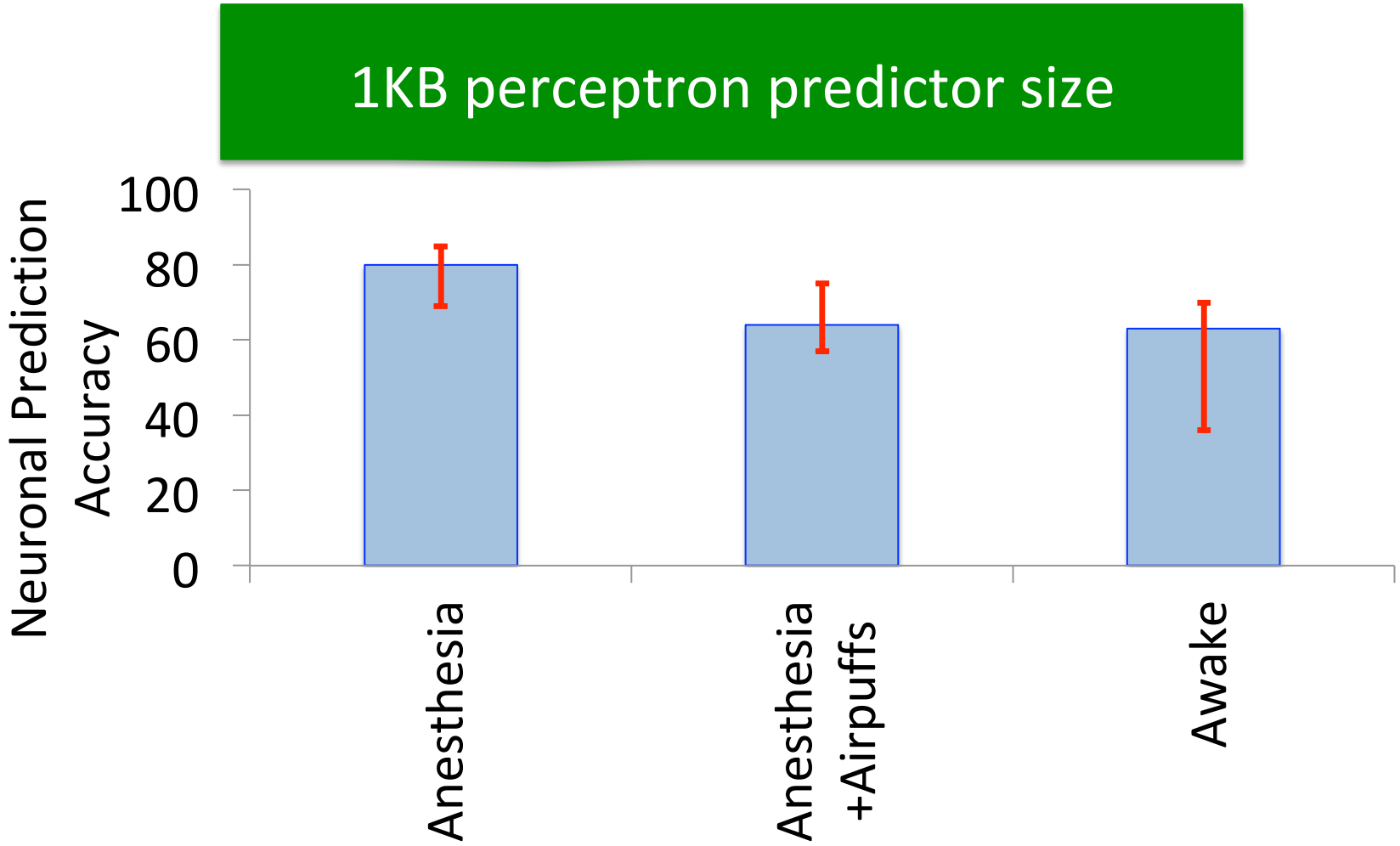
Details of the implants we built

- ARM Cortex M3
  - 2-issue, 4-stage, in-order pipeline with forwarding
  - 32KB instruction & data caches
  - FP support
  - 6/4 R/W-port register file
- ASIC to process 100 channels of data
- 100 channels, 20 bit samples at 50 kilo-samples

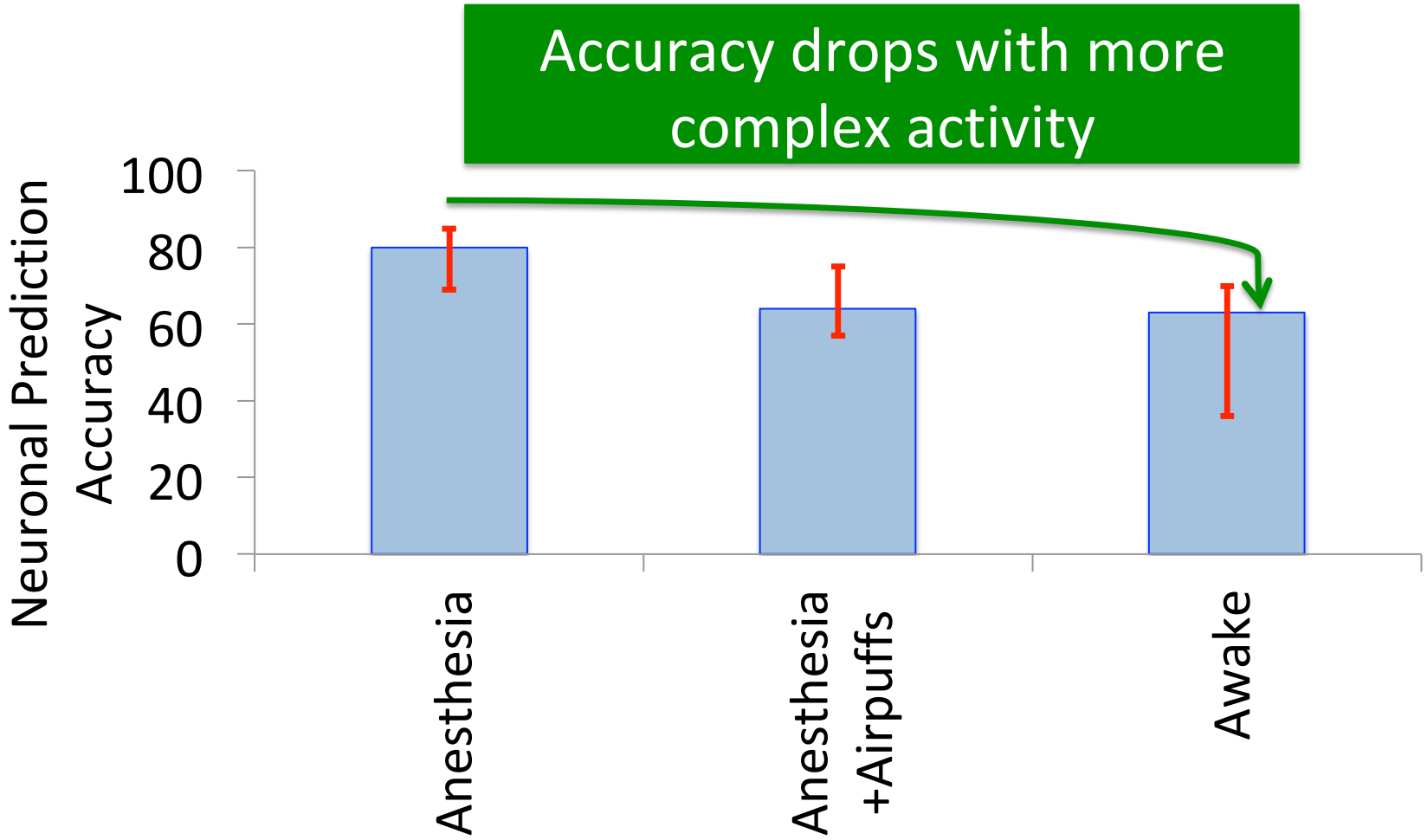
Details of our wet lab experiments

- Mouse experiments
  - 2mm craniotomies on lobule 6 of cerebellum
  - Mice on post-natal days 21-42
  - Anesthesia with xylamine/ketamine

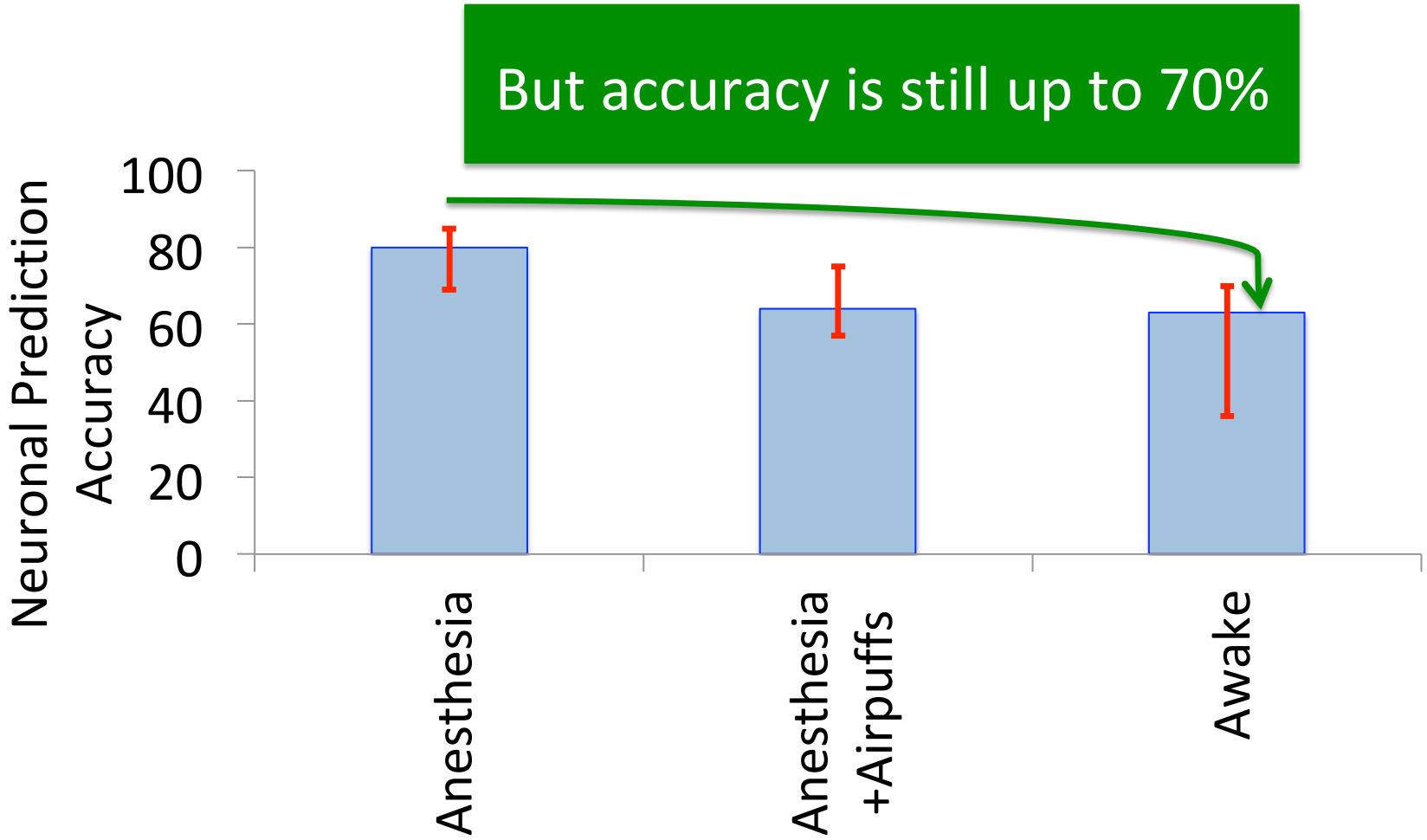
Perceptron prediction accuracy



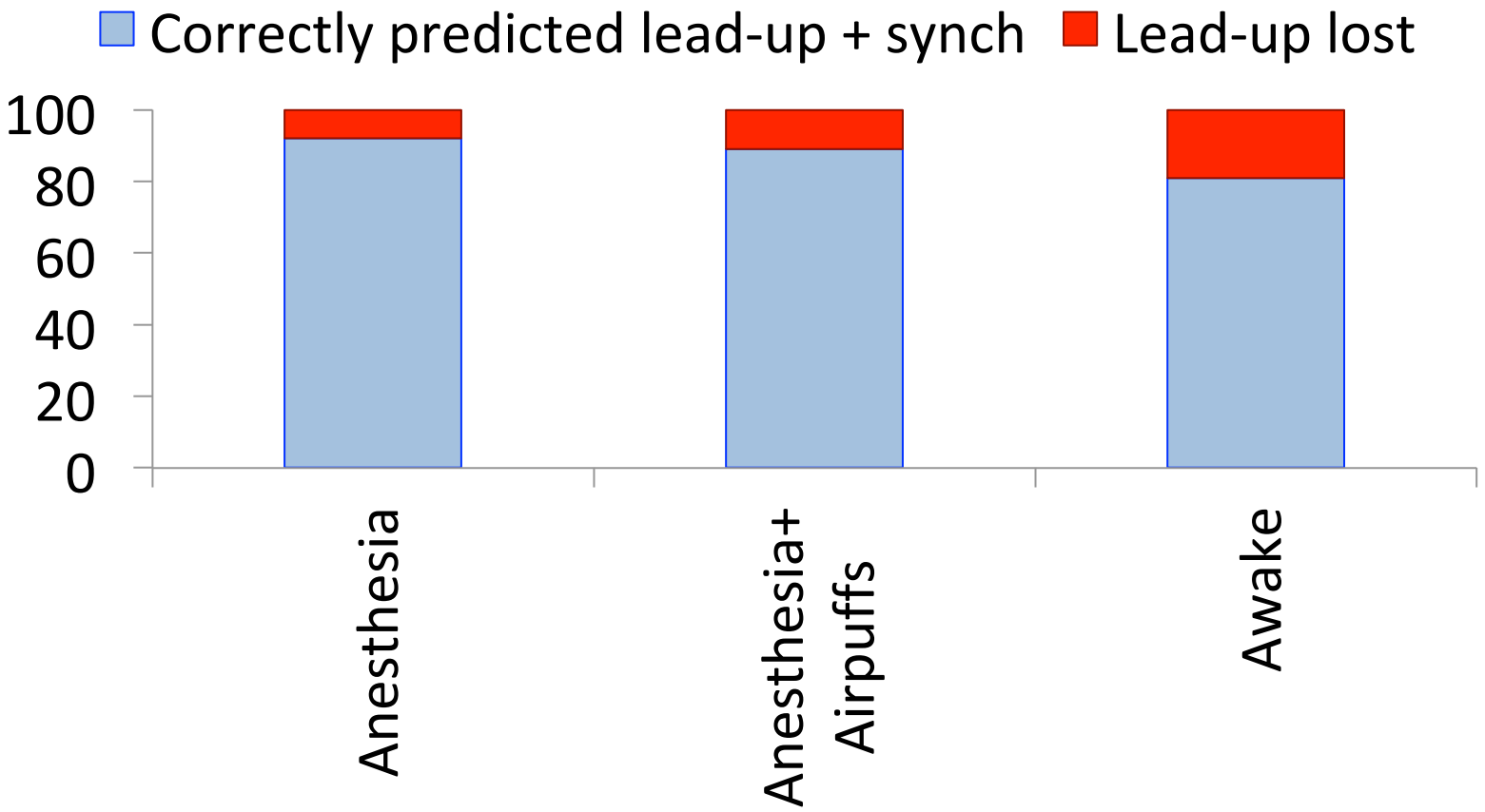
Perceptron prediction accuracy



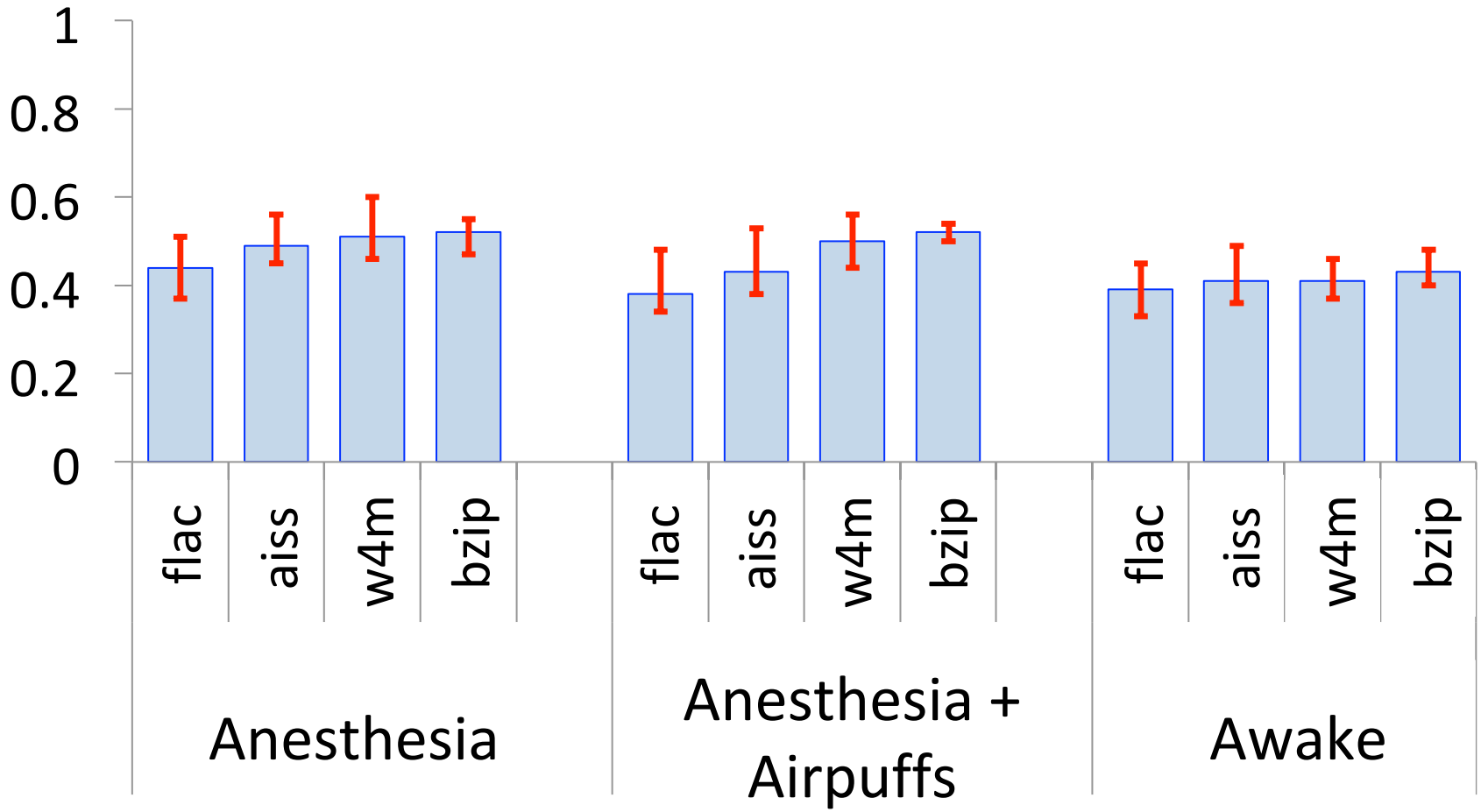
Perceptron prediction accuracy



Percentage of lead-up + synchronization predicted correctly and incorrectly



Fraction of compute logic energy saved



What does saving implant energy buy us?

Longer implant battery lifetimes

Mouse implants: 25-35% longer battery life

Place sensors on more parts of the brain

Monkey implants: 15% increase in sensors

Architectural Techniques to Build Energy-Efficient Brain Implants

ARM Research Summit: Biotechnology

Abhishek Bhattacharjee

Associate Professor

Department of Computer Science

Rutgers University

Prediction versus sampling