# Sequential Density Ratio Estimation for Simultaneous Optimization of Speed and Accuracy

Akinori F. Ebihara[1]
Taiki Miyagawa[1,2],
Kazuyuki Sakurai[1],
Hitoshi Imaoka[1]

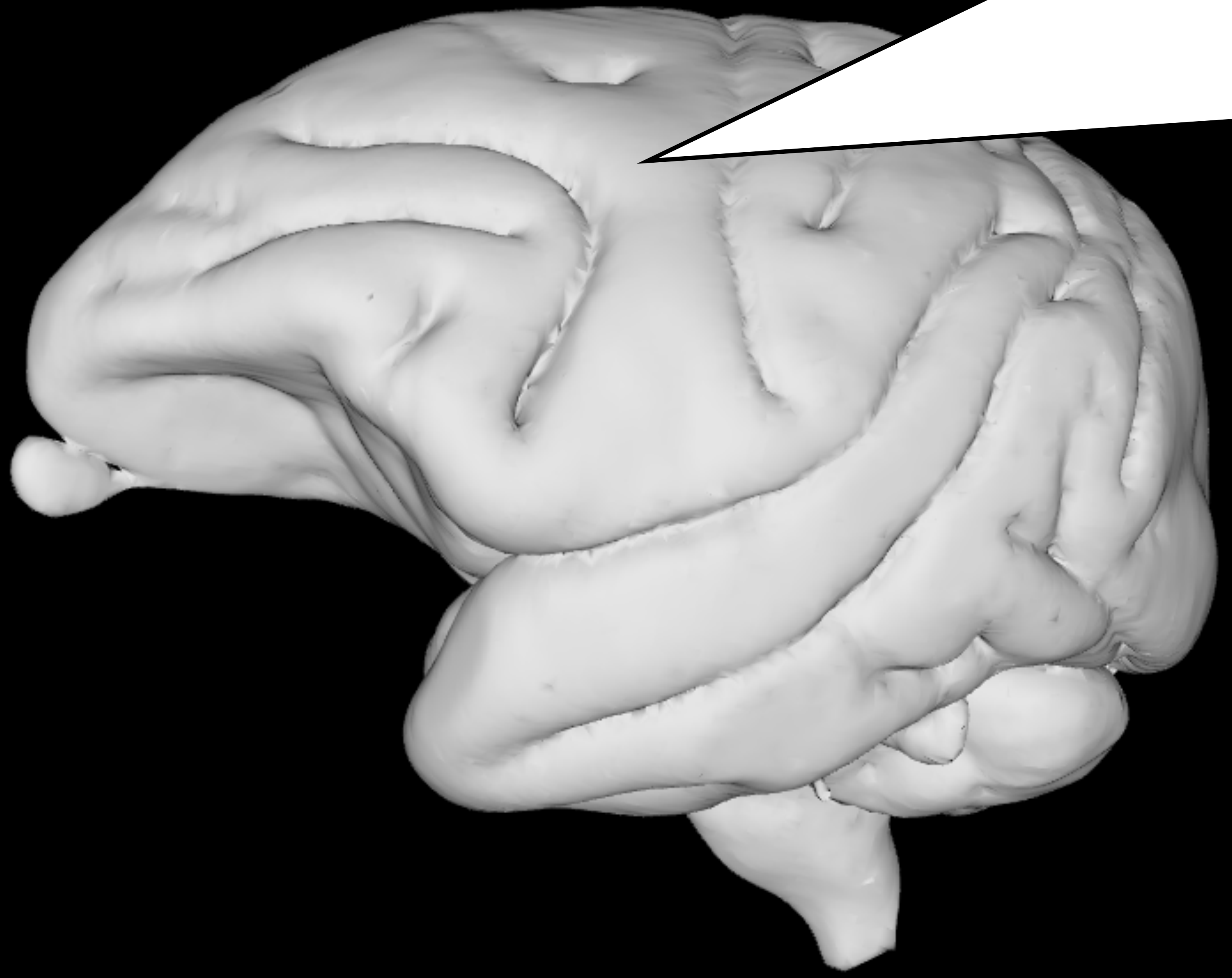
[1]NEC Corporation, Japan
[2]RIKEN Center for Advanced Intelligence Project (AIP)

Evidence:
Kicker's motion
You
Left
Right

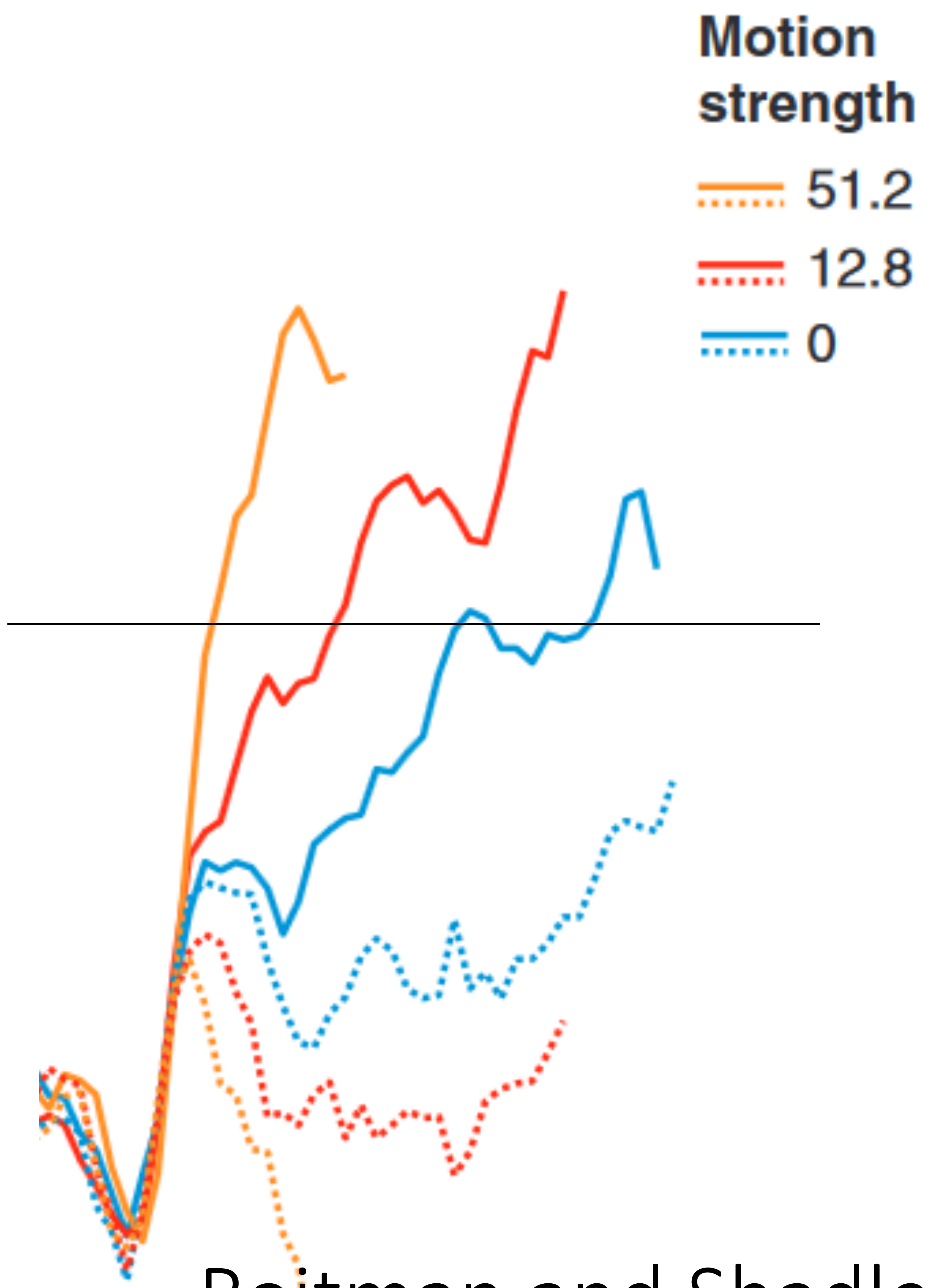
# Mind the sampling cost!


Left
Confidence level
Right
Time

# Parietal lobe neurons are thought to mediate evidence accumulation at the decision making process

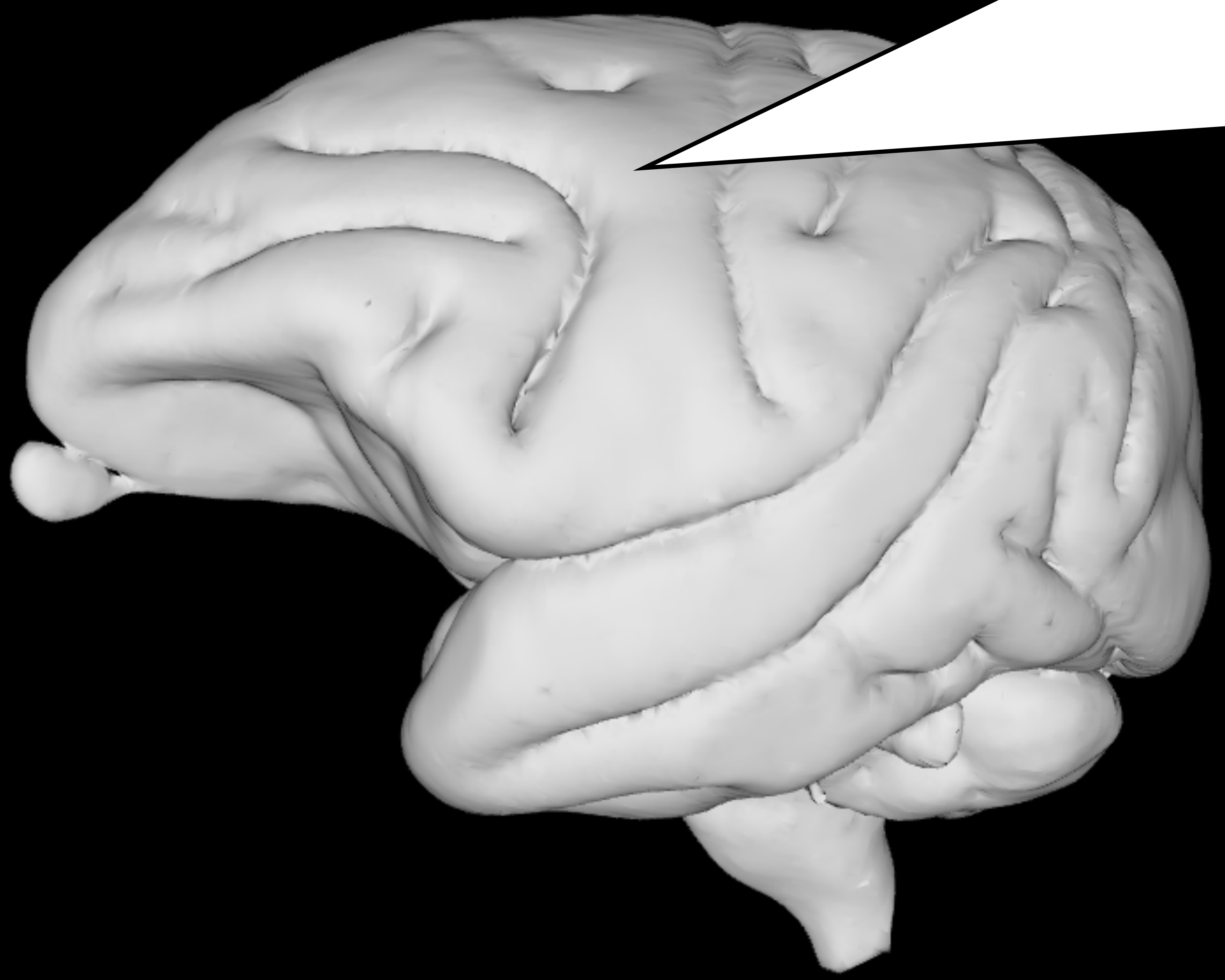


Lateral intraparietal neural activity

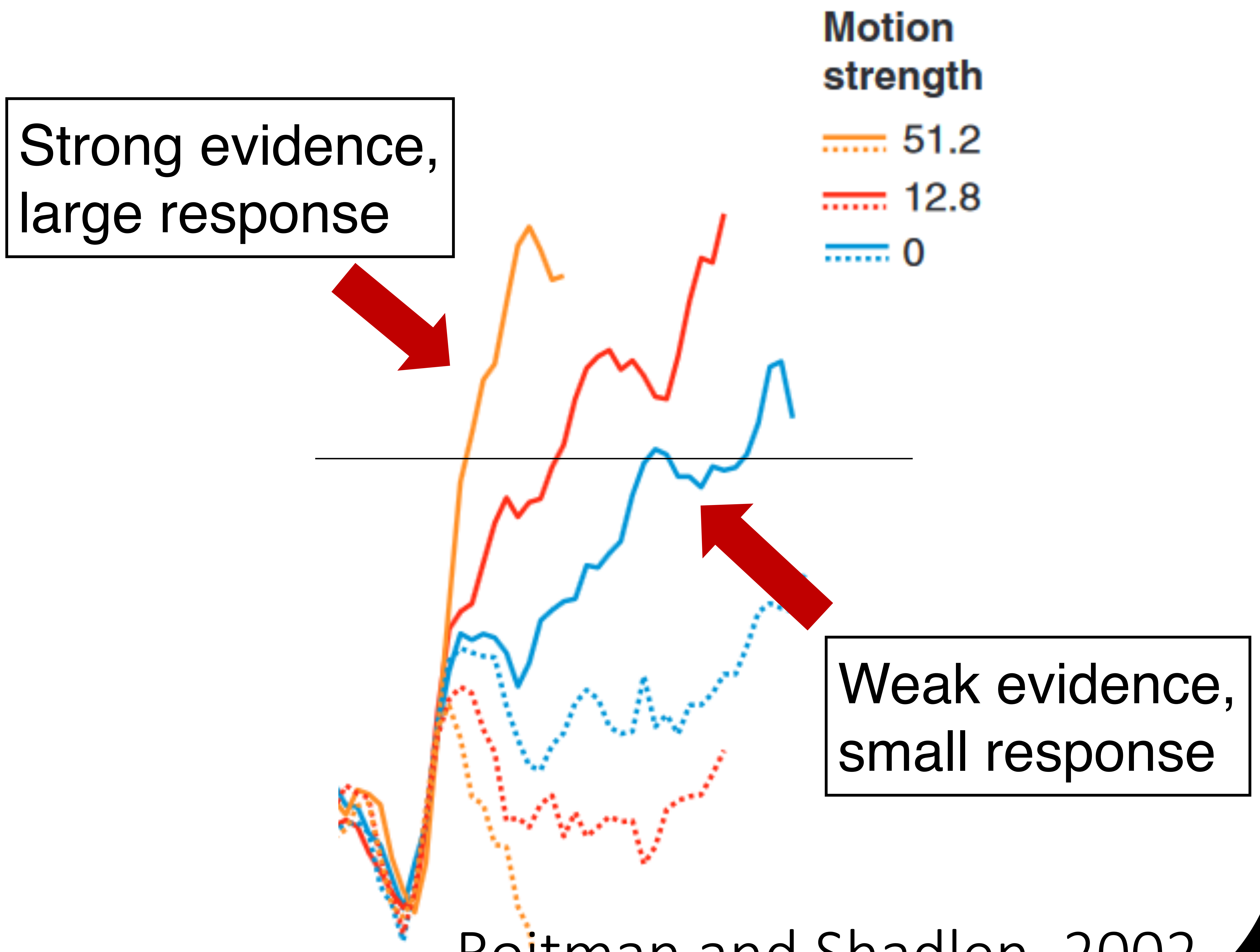



Roitman and Shadlen, 2002

# Parietal lobe neurons are thought to mediate evidence accumulation at the decision making process

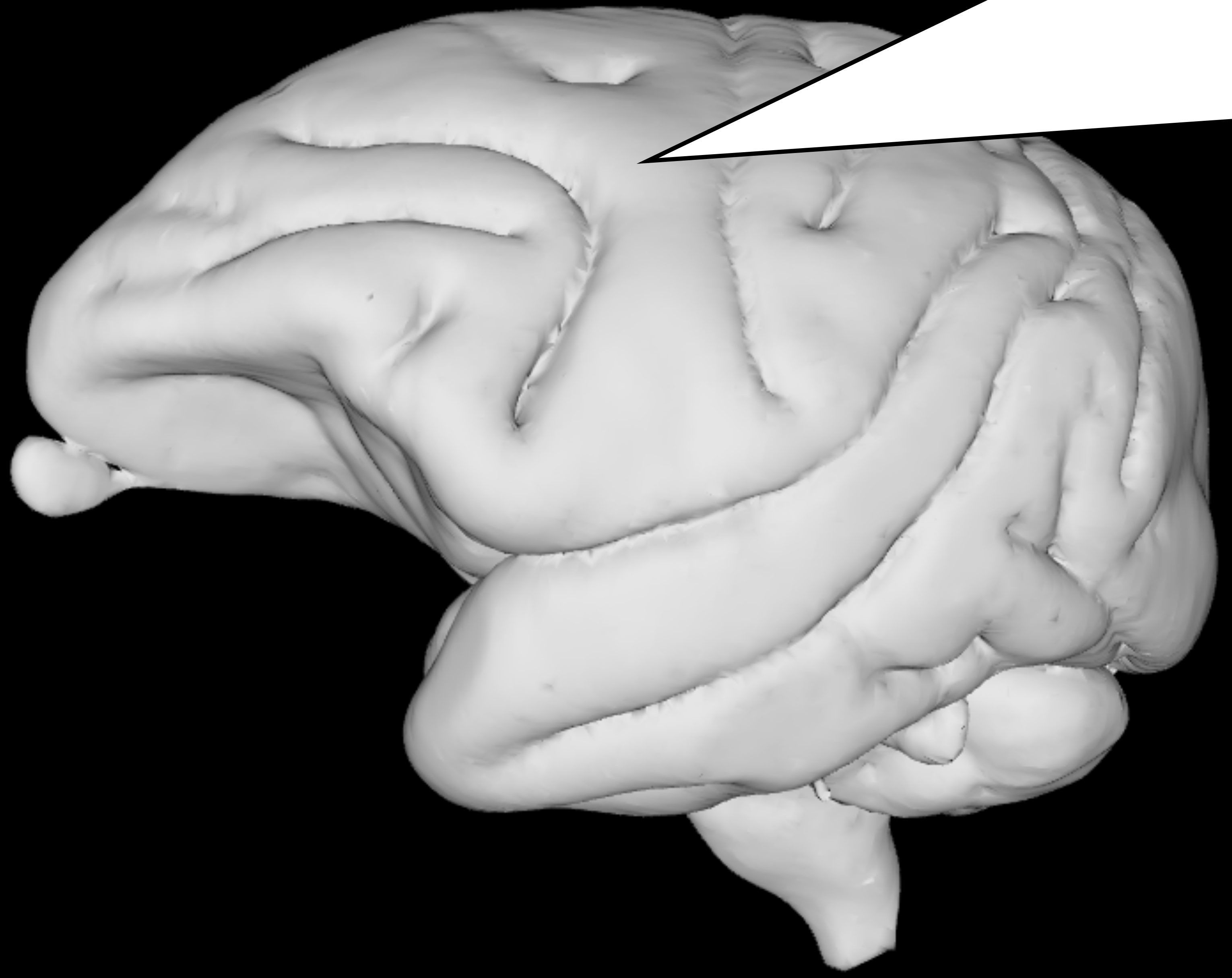


Lateral intraparietal neural activity



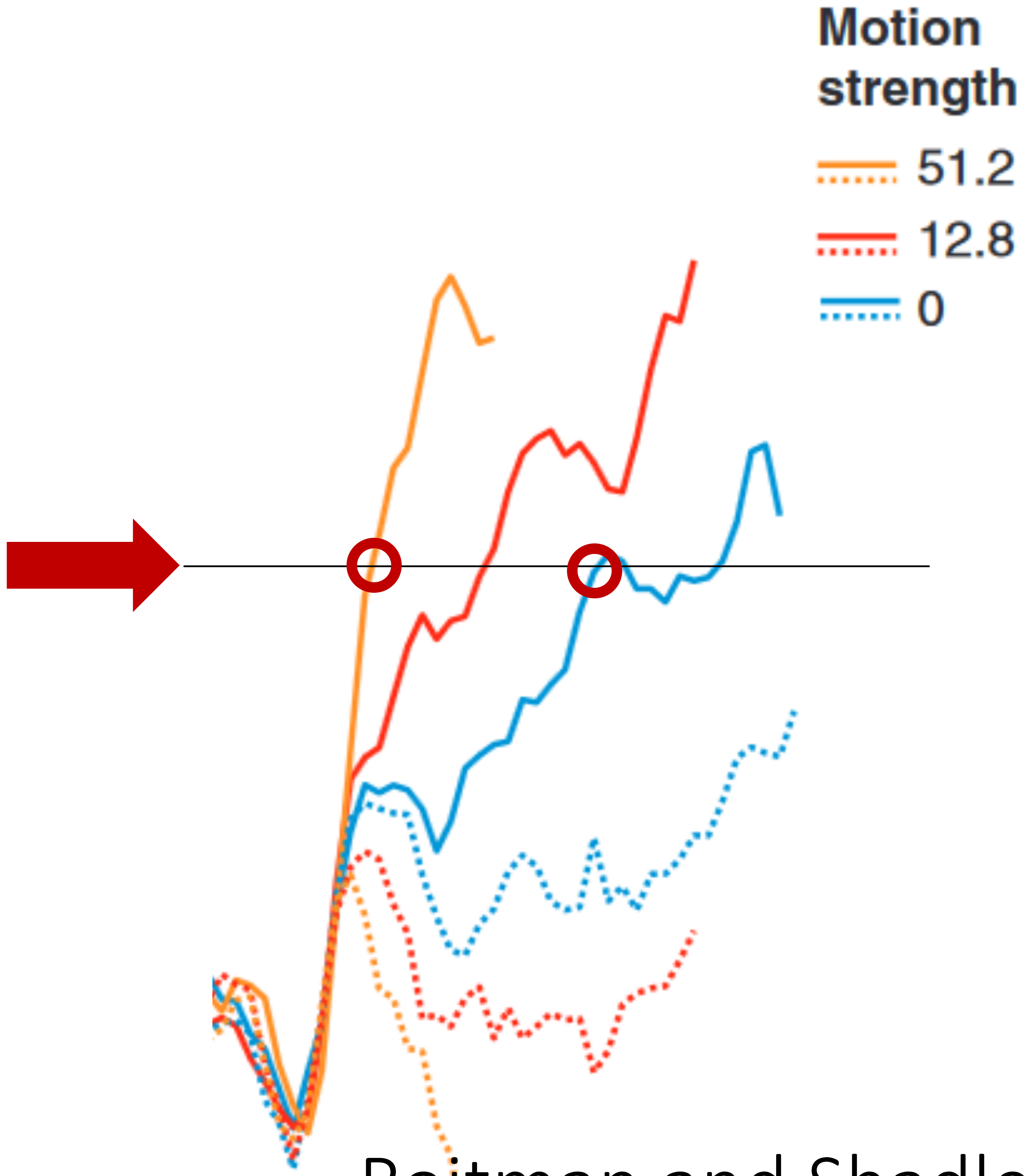

Roitman and Shadlen, 2002

# Parietal lobe neurons are thought to mediate evidence accumulation at the decision making process



Lateral intraparietal neural activity

Fixed decision boundary




Roitman and Shadlen, 2002

# Sequential Probability Ratio Test (SPRT) best explains the neural activity during the decision making process

**Glossary**

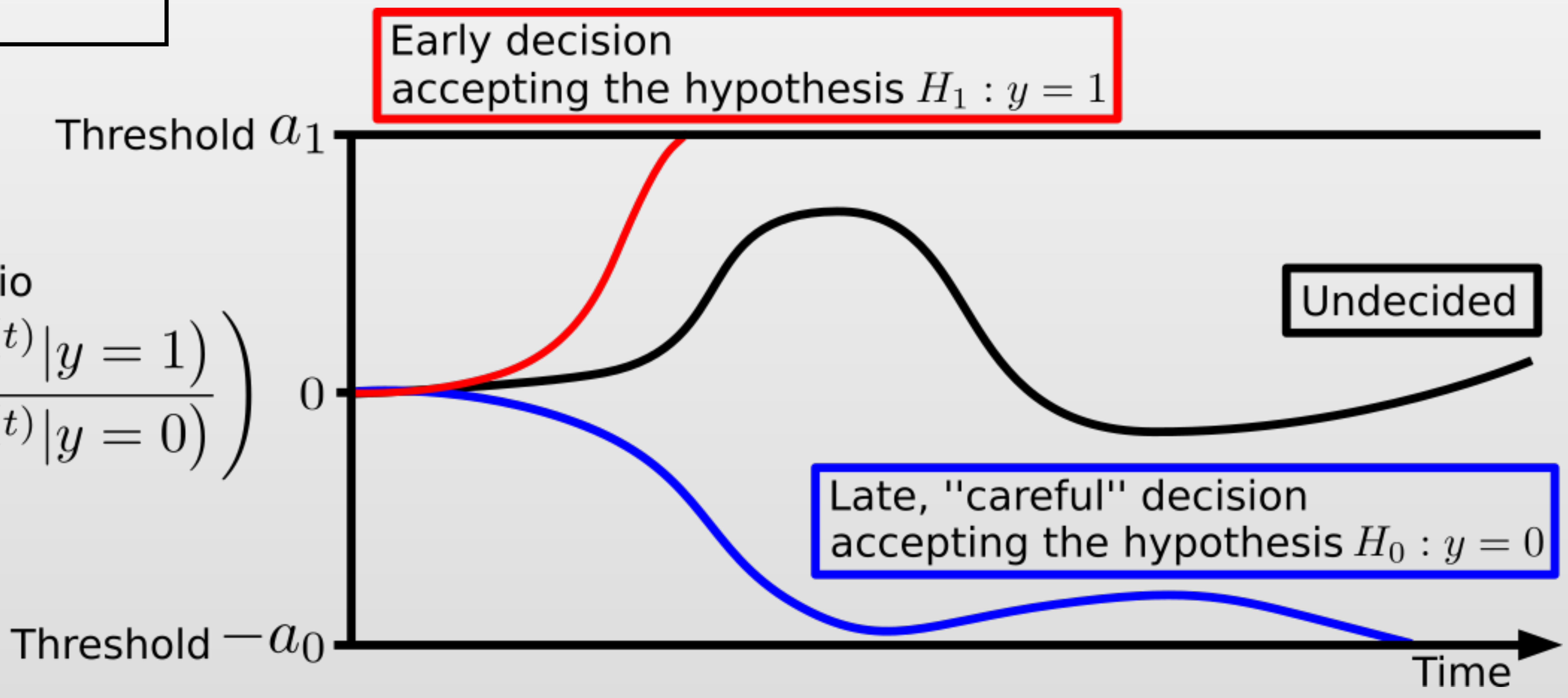
Maximum timestamp $T \in \mathbb{N}$

Sequential data $X^{(1,T)} := \{x^{(t)}\}_{t=1}^{T}$

Class label $y \in \{1,0\}$

Decision value: Log-likelihood ratio

$$\lambda_t = \log \left( \frac{p\left(x^{(1)}, x^{(2)}, ..., x^{(t)} | y = 1\right)}{p\left(x^{(1)}, x^{(2)}, ..., x^{(t)} | y = 0\right)} \right)$$




Wald, 1947, Kira et al. 2017

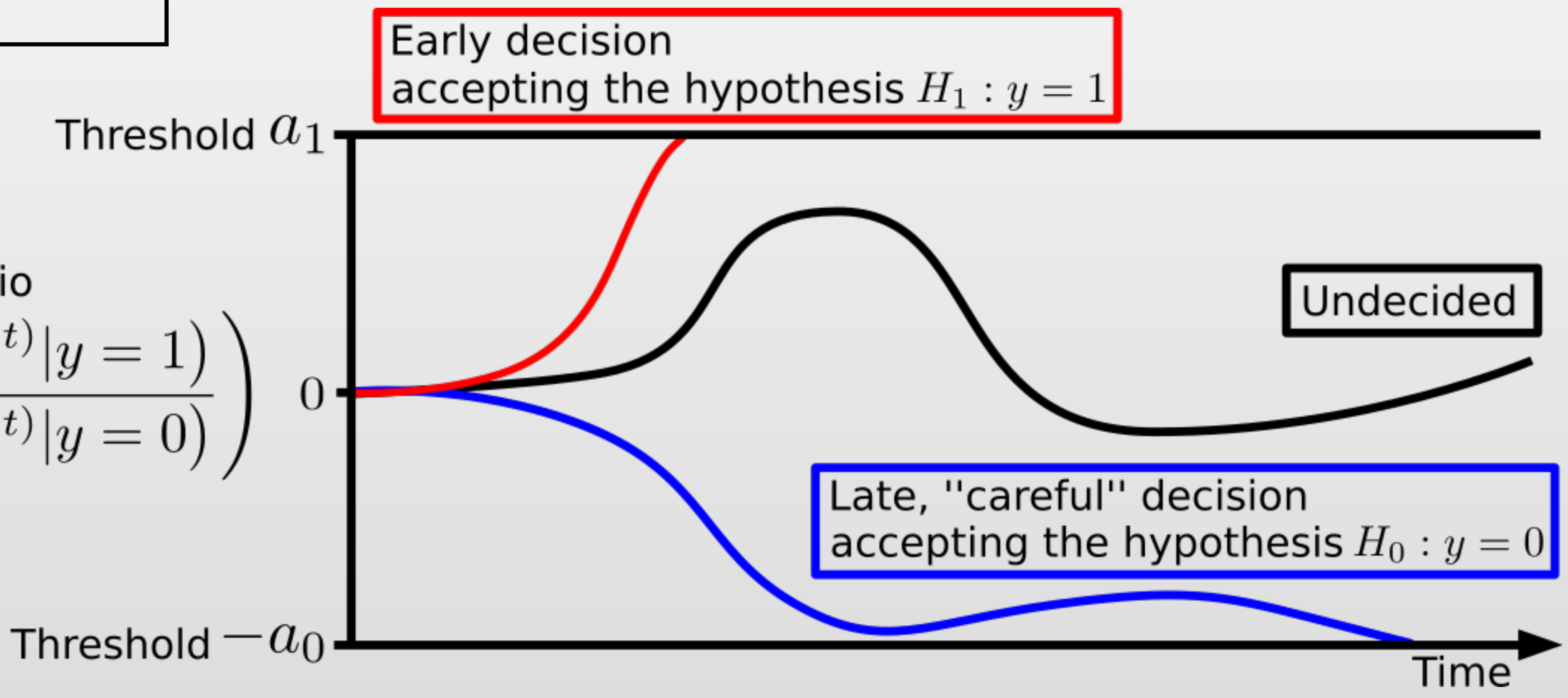
# Sequential Probability Ratio Test (SPRT) best explains the neural activity during the decision making process

Glossary

Maximum timestamp $T \in \mathbb{N}$
Sequential data $X^{(1,T)} := \{x^{(t)}\}_{t=1}^{T}$
Class label $y \in \{1,0\}$

Decision value: Log-likelihood ratio

$$\lambda_t = \log\left(\frac{p\left(x^{(1)}, x^{(2)}, ..., x^{(t)} | y = 1\right)}{p\left(x^{(1)}, x^{(2)}, ..., x^{(t)} | y = 0\right)}\right)$$




SPRT achieves accuracy **equivalent to the Neyman-Pearson test, known as the most powerful statistical test**
SPRT reaches the threshold **faster than any existing sequential algorithms**

Wald, 1947, Kira et al. 2017

# Two strict assumptions hamper SPRT from real-world applications

Assumtion 1:  samples are i.i.d.

SPRT-compatible toy model

Independent and identically distributed **(i.i.d.)**

Real-world scenarios

Highly **correlated**

# Two strict assumptions
# hamper SPRT from real-world applications

Assumtion 2:  Likelihood is known

SPRT-compatible toy model

Real-world scenarios

Likelihood is **calculable**

Likelihood is **unknown**

# The TANDEM formula to compute the log-likelihood ratio under Nth-order Markov process

Glossary

Maximum timestamp $T \in \mathbb{N}$

Sequential data $X^{(1,T)} := \{x^{(t)}\}_{t=1}^{T}$

Class label $y \in \{1,0\}$

Order of Markov process $N \in \{0,1,...,T-1\}$

$$\log\left(\frac{p(x^{(1)}, x^{(2)}, \ldots, x^{(t)} | y=1)}{p(x^{(1)}, x^{(2)}, \ldots, x^{(t)} | y=0)}\right)$$

$$= \sum_{s=N+1}^{t} \log\left(\frac{p(y=1 | x^{(s-N)}, \ldots, x^{(s)})}{p(y=0 | x^{(s-N)}, \ldots, x^{(s)})}\right) - \sum_{s=N+2}^{t} \log\left(\frac{p(y=1 | x^{(s-N)}, \ldots, x^{(s-1)})}{p(y=0 | x^{(s-N)}, \ldots, x^{(s-1)})}\right)$$

$$-\log\left(\frac{p(y=1)}{p(y=0)}\right)$$

# The TANDEM formula to compute the log-likelihood ratio under Nth-order Markov process

**Glossary**

Maximum timestamp $T \in \mathbb{N}$

Sequential data $X^{(1,T)} := \{x^{(t)}\}_{t=1}^{T}$

Class label $y \in \{1,0\}$

Order of Markov process $N \in \{0,1,...,T-1\}$

$$\log\left(\frac{p(x^{(1)},x^{(2)},\ldots,x^{(t)}\,|\,y=1)}{p(x^{(1)},x^{(2)},\ldots,x^{(t)}\,|\,y=0)}\right)$$

$$= \sum_{s=N+1}^{t} \log\left(\frac{p(y=1\,|\,x^{(s-N)},\ldots,x^{(s)})}{p(y=0\,|\,x^{(s-N)},\ldots,x^{(s)})}\right) - \sum_{s=N+2}^{t} \log\left(\frac{p(y=1\,|\,x^{(s-N)},\ldots,x^{(s-1)})}{p(y=0\,|\,x^{(s-N)},\ldots,x^{(s-1)})}\right)$$

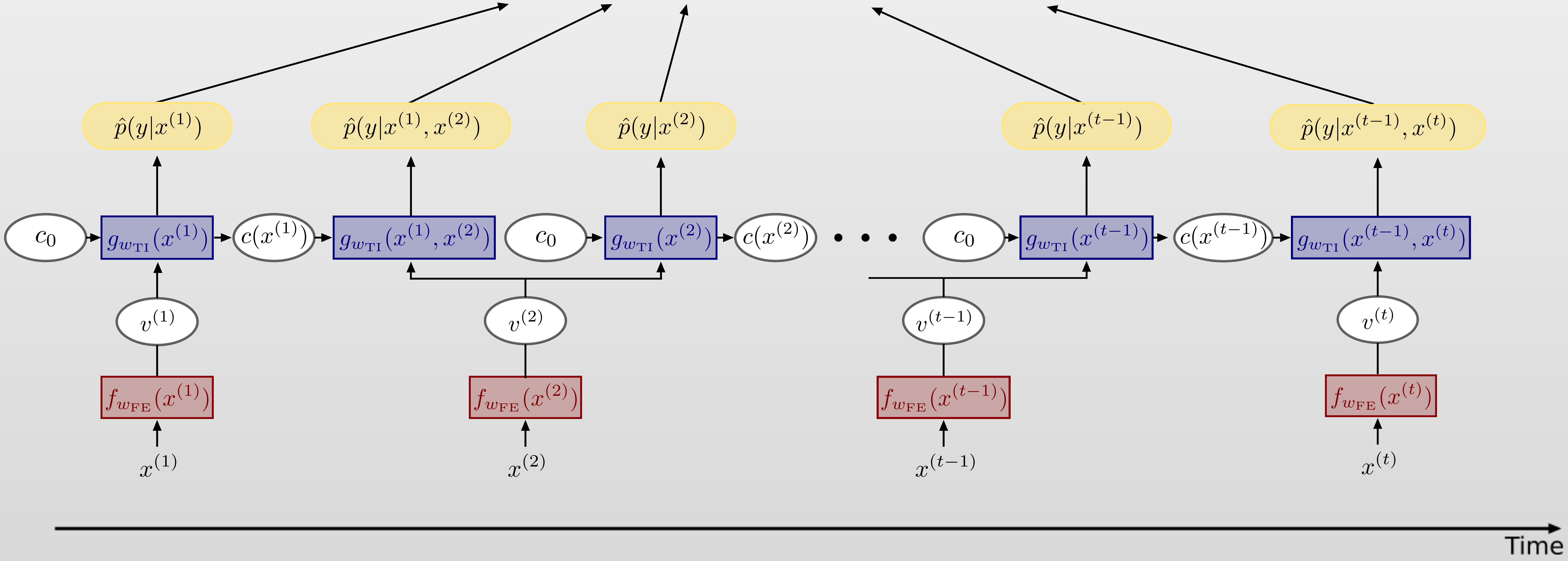
$$-\log\left(\frac{p(y=1)}{p(y=0)}\right)$$

**Prior term**

**Terms work in "TANDEM"**

# The SPRT-TANDEM network to explicitly calculate the TANDEM formula

$$\hat{\lambda}_t = \log\left(\frac{\hat{p}(x^{(1)}, x^{(2)}, ..., x^{(t)}|y=1)}{\hat{p}(x^{(1)}, x^{(2)}, ..., x^{(t)}|y=0)}\right)$$
$$= \sum_{s=2}^{t} \log\left(\frac{\hat{p}(y=1|x^{(s-1)}, x^{(s)})}{\hat{p}(y=0|x^{(s-1)}, x^{(s)})}\right) - \sum_{s=3}^{t} \log\left(\frac{\hat{p}(y=1|x^{(s-1)})}{\hat{p}(y=0|x^{(s-1)})}\right) - \log\left(\frac{\hat{p}(y=1)}{\hat{p}(y=0)}\right)$$



# Loss for log-likelihood ratio estimation (LLLR) to correctly estimate the log-likelihood ratio

Glossary

| | |
|---|---|
| Maximum timestamp | $T \in \mathbb{N}$ |
| Sequential data | $X^{(1,T)} := \{x^{(t)}\}_{t=1}^{T}$ |
| Class label | $y \in \{1,0\}$ |
| Dataset size | $M \in \mathbb{N}$ |
| Order of Markov process | $N \in \{0,1,...,T-1\}$ |
| Sigmoid function | $\sigma$ |

$$L_{\mathrm{LLR}} = \frac{1}{MT} \sum_{i=1}^{M} \sum_{t=1}^{T} \left| y_i - \sigma\left( \log\left( \frac{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=1)}{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=0)} \right) \right) \right|$$

# Loss for log-likelihood ratio estimation (LLLR) to correctly estimate the log-likelihood ratio

$$L_{\mathrm{LLR}} = \frac{1}{MT}\sum_{i=1}^{M}\sum_{t=1}^{T}\left| y_i - \sigma\left(\log\left(\frac{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=1)}{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=0)}\right)\right)\right|$$

$$= \frac{1}{M_0 T}\sum_{i=1}^{M_0}\sum_{t=1}^{T}\sigma\left(\log\left(\frac{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=1)}{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=0)}\right)\right)$$

$$+ \frac{1}{M_1 T}\sum_{i=1}^{M_1}\sum_{t=1}^{T}\left| 1 - \sigma\left(\log\left(\frac{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=1)}{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=0)}\right)\right)\right|$$

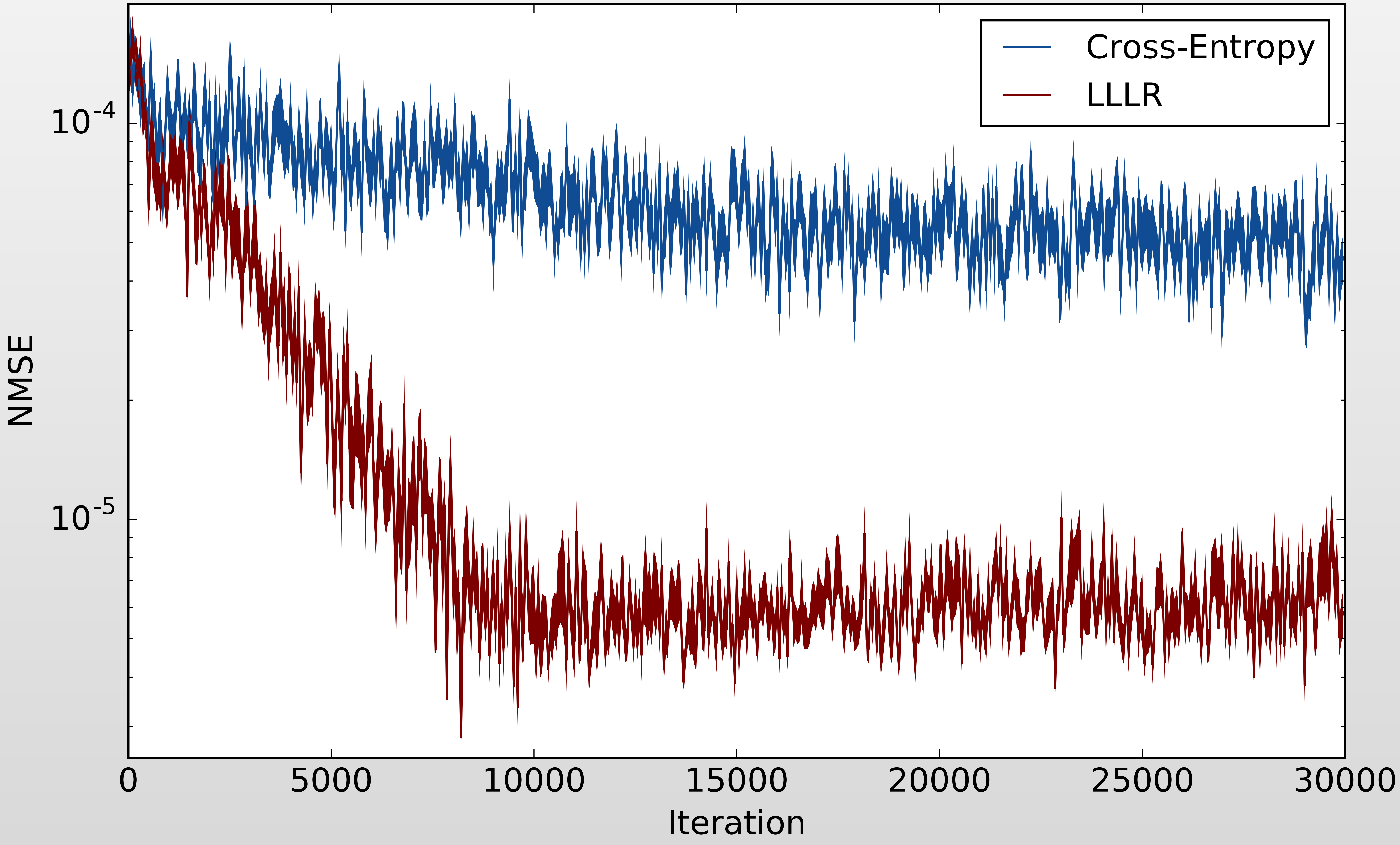
$$L_{\mathrm{KLIEP}} = \frac{1}{M_0 T}\sum_{i=1}^{M_0}\sum_{t=1}^{T}\log\left(\frac{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=1)}{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=0)}\right)$$

$$+ \frac{1}{M_1 T}\sum_{i=1}^{M_1}\sum_{t=1}^{T} - \log\left(\frac{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=1)}{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=0)}\right)$$

**Glossary**

- Maximum timestamp $T \in \mathbb{N}$
- Sequential data $X^{(1,T)} := \{x^{(t)}\}_{t=1}^{T}$
- Class label $y \in \{1,0\}$
- Dataset size $M \in \mathbb{N}$
- **Dataset size of class 0** $M_0 \in \mathbb{N}$
- **Dataset size of class 1** $M_1 \in \mathbb{N}$
- Order of Markov process $N \in \{0,1,\ldots,T-1\}$
- Sigmoid function $\sigma$

# LLLR effectively estimates the true probability density ratio compared with cross-entropy loss



# LLLR is combined with the multiplet cross-entropy loss

**Glossary**

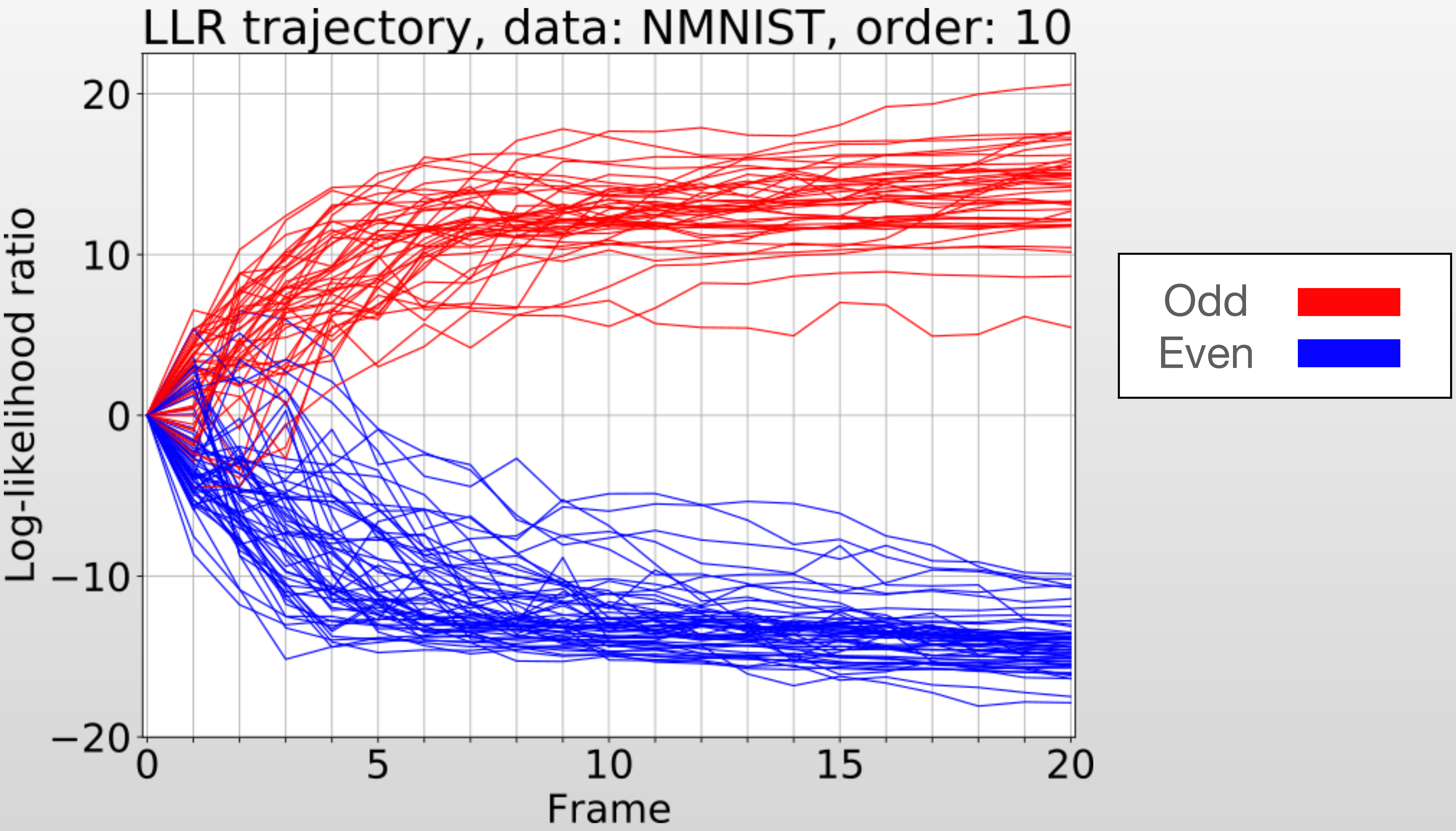
| | |
|---|---|
| Maximum timestamp | $T \in \mathbb{N}$ |
| Sequential data | $X^{(1,T)} := \{x^{(t)}\}_{t=1}^{T}$ |
| Class label | $y \in \{1,0\}$ |
| Dataset size | $M \in \mathbb{N}$ |
| Order of Markov process | $N \in \{0,1,...,T-1\}$ |
| Sigmoid function | $\sigma$ |

$$L_{\mathrm{LLR}} = \frac{1}{MT} \sum_{i=1}^{M} \sum_{t=1}^{T} \left| y_i - \sigma \left( \log \left( \frac{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=1)}{\hat{p}(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)} \,|\, y=0)} \right) \right) \right|$$

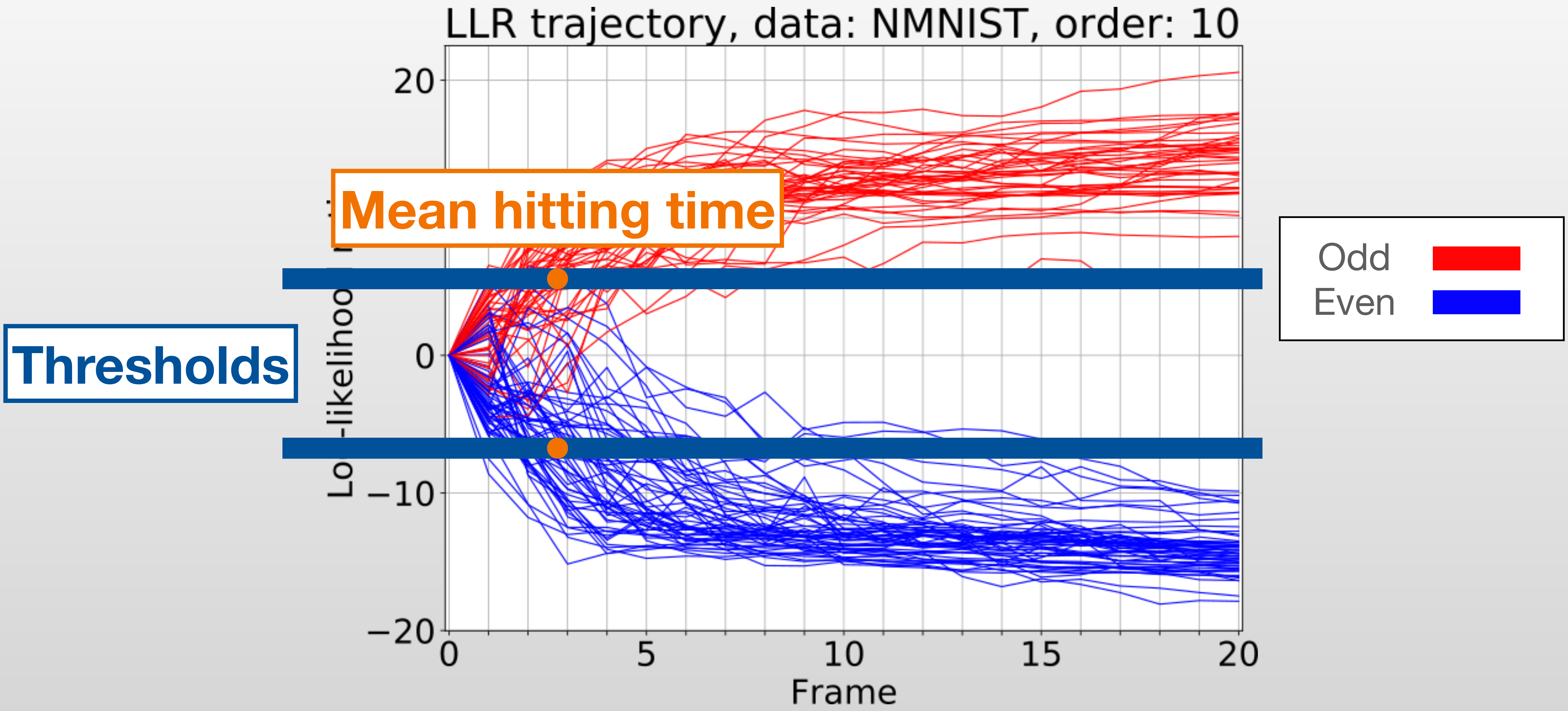
$$L_{\mathrm{multiplet}} = \sum_{k=1}^{N+1} \frac{1}{M(T-N)} \sum_{i=1}^{M} \sum_{t=k}^{T-(N+1-k)} \left( -\log \hat{p}(y_i \,|\, x_i^{(t-k+1)}, \ldots, x_i^{(t)}) \right)$$

# SPRT-TANDEM outperforms other baselines on the Nosaic-NMIST database

# Log-likelihood ratio trajectory shows the two hypotheses are separated as the evidence is accumulated

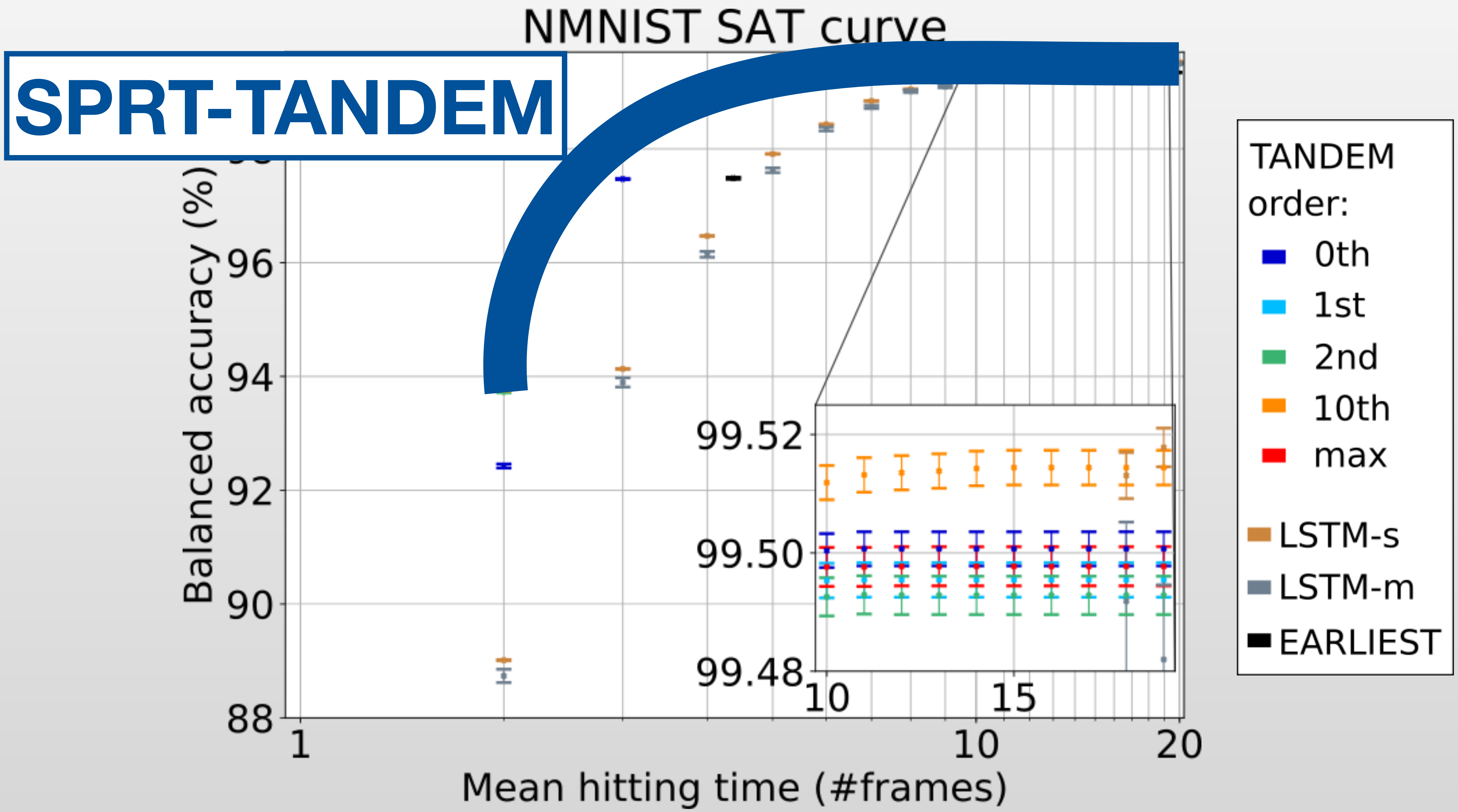


# Log-likelihood ratio trajectory shows
# the two hypotheses are separated as the evidence is accumulated



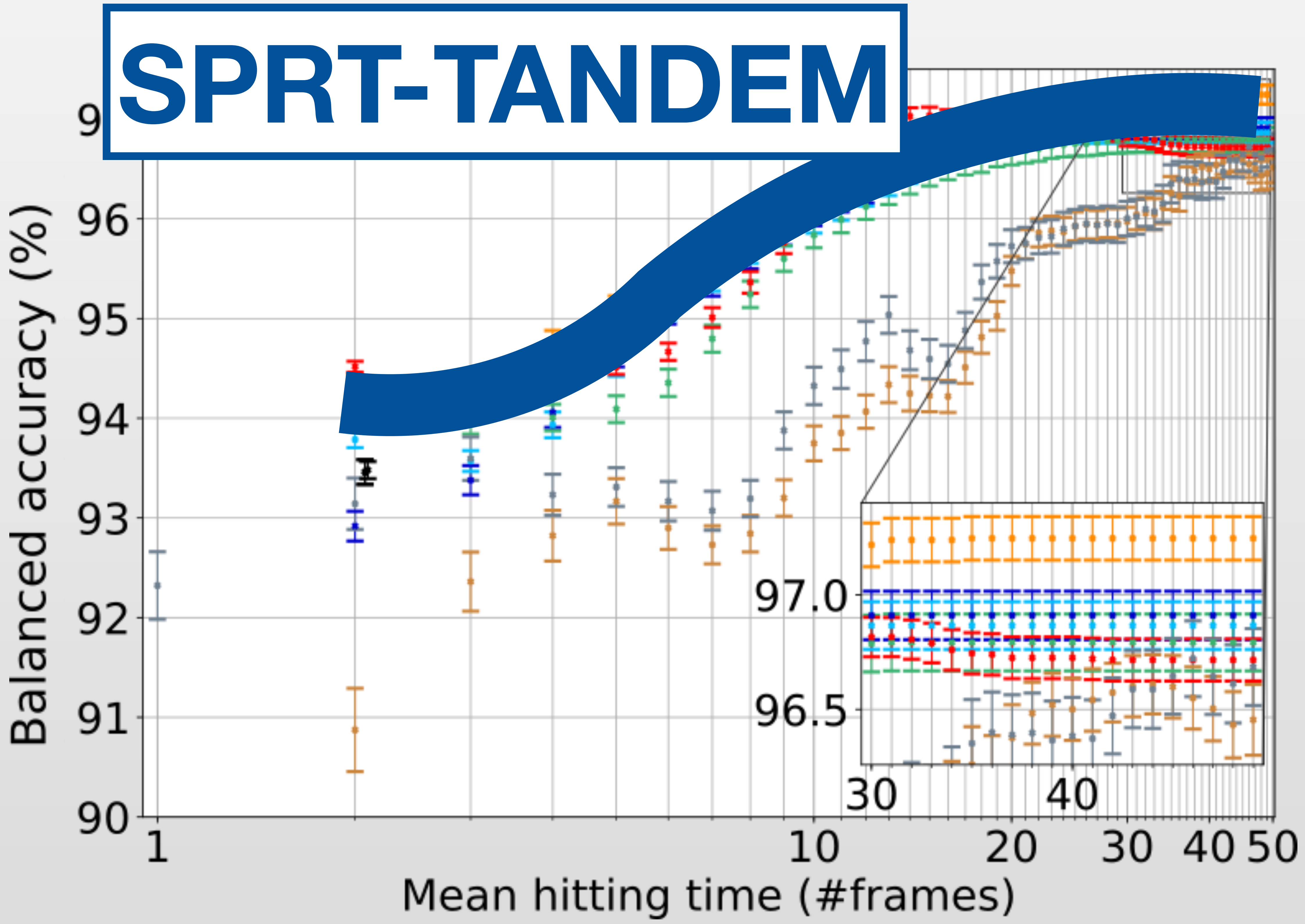
# SPRT-TANDEM outperforms other baselines on the Nosaic-NMIST database

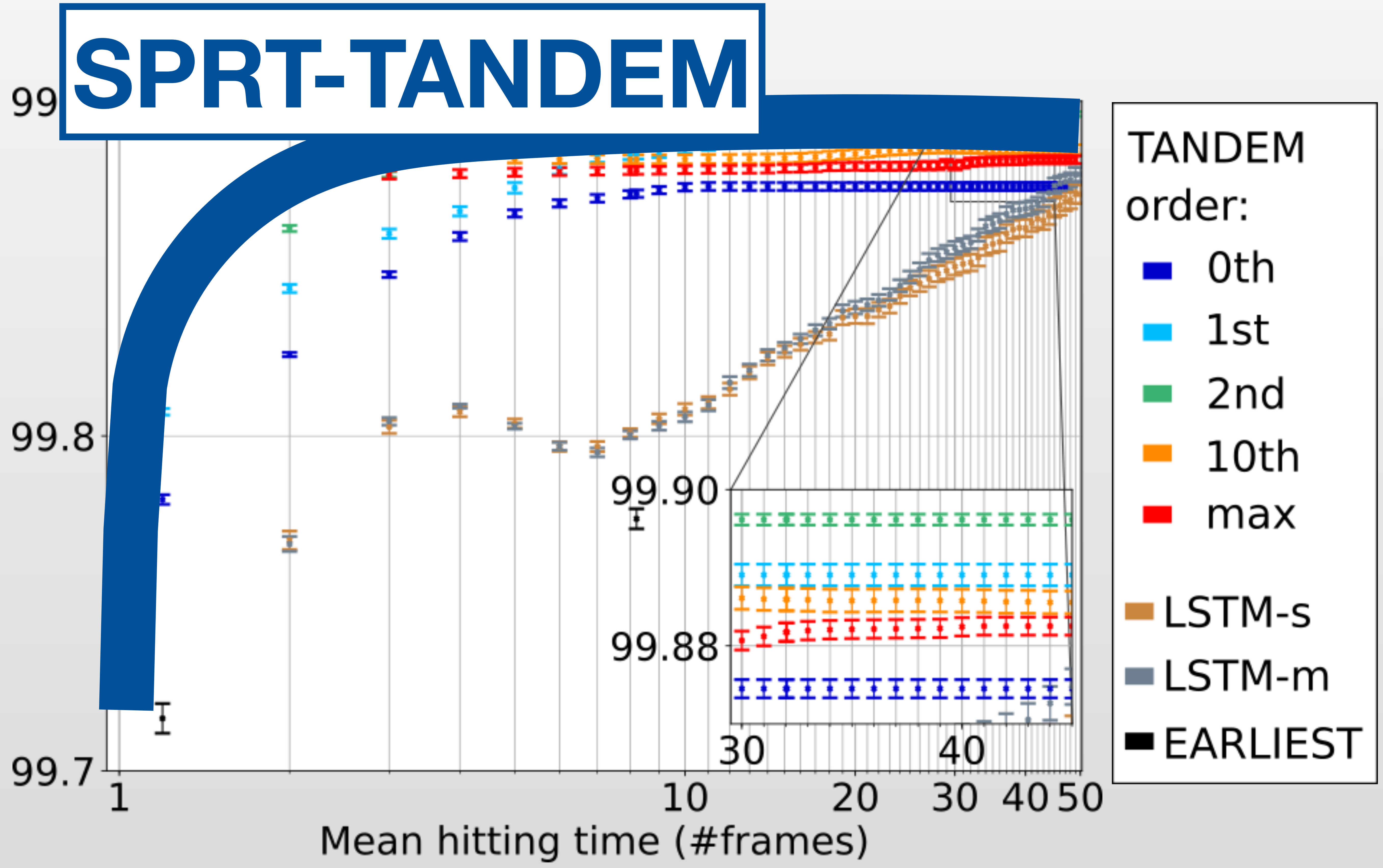
Achieves statistically significantly better accuracy at given mean hitting time



# Performance on UCF and SiW databases confirmed applicability of SPRT-TANDEM under real-world scenarios

Achieves statistically significantly better accuracy at given mean hitting time



# Conclusions

- Invented the **SPRT-TANDEM** framework that optimizes speed and accuracy simultaneously by using the **TANDEM formula**, **SPRT-TANDEM network**, and the **LLLR**. Also introduced the **Nosaic-MNIST database**.

# Contacts


- **Akinori F. Ebihara**

aebihara@nec.com
https://github.com/Akinori-F-Ebihara
https://twitter.com/non_iid