Noise-Contrastive Estimation
for Multivariate Point Processes

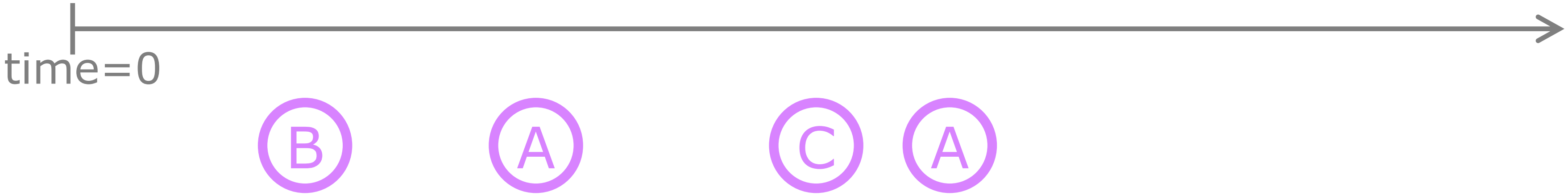
Hongyuan Mei, Tom Wan, Jason Eisner

Johns Hopkins University

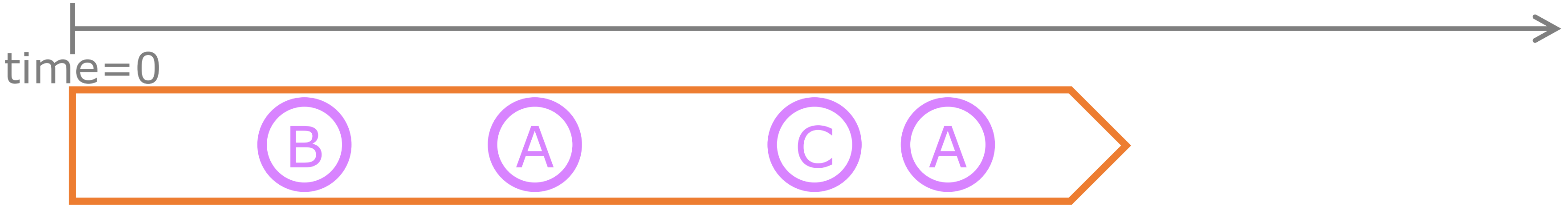
MLE: Max log prob of *data*

MLE: Max log prob of *data*

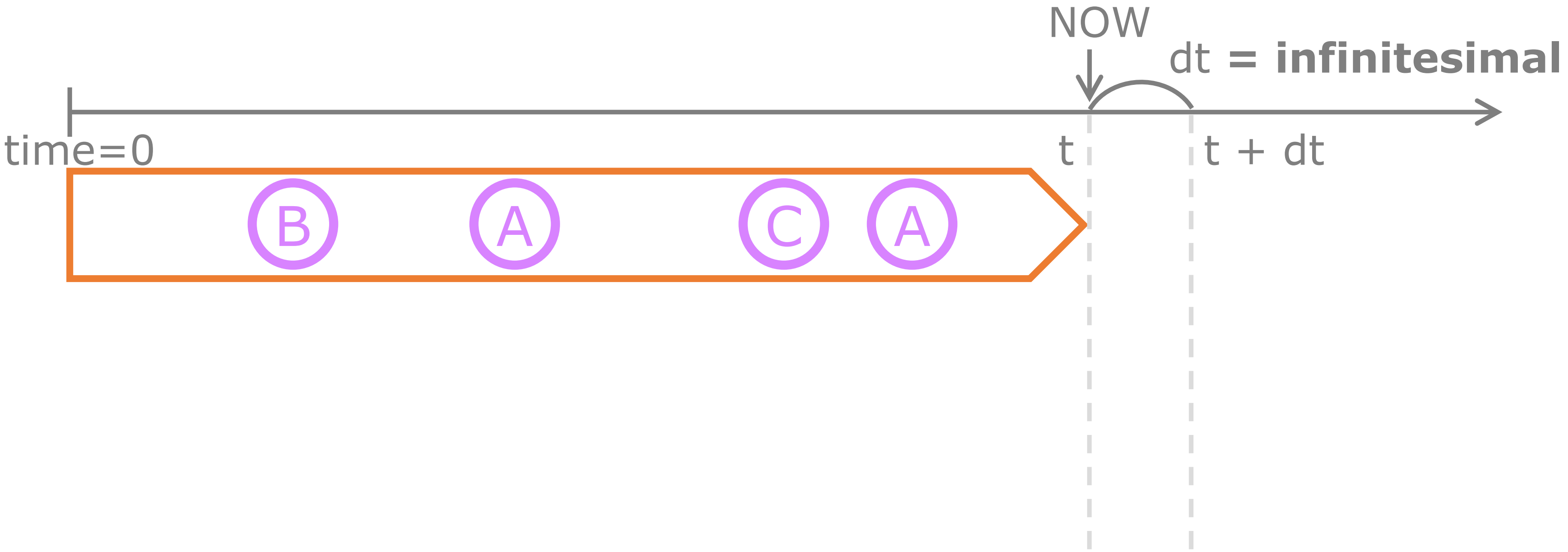
MLE: Max log prob of *data*



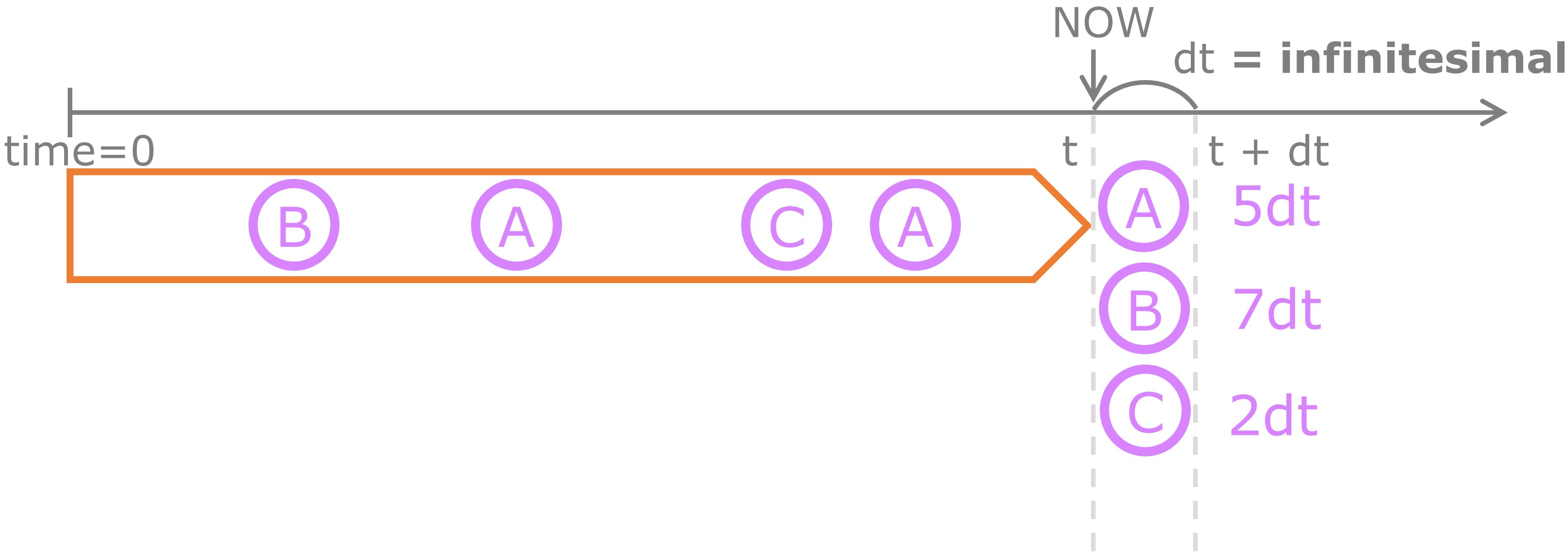
MLE: Max log prob of *data*



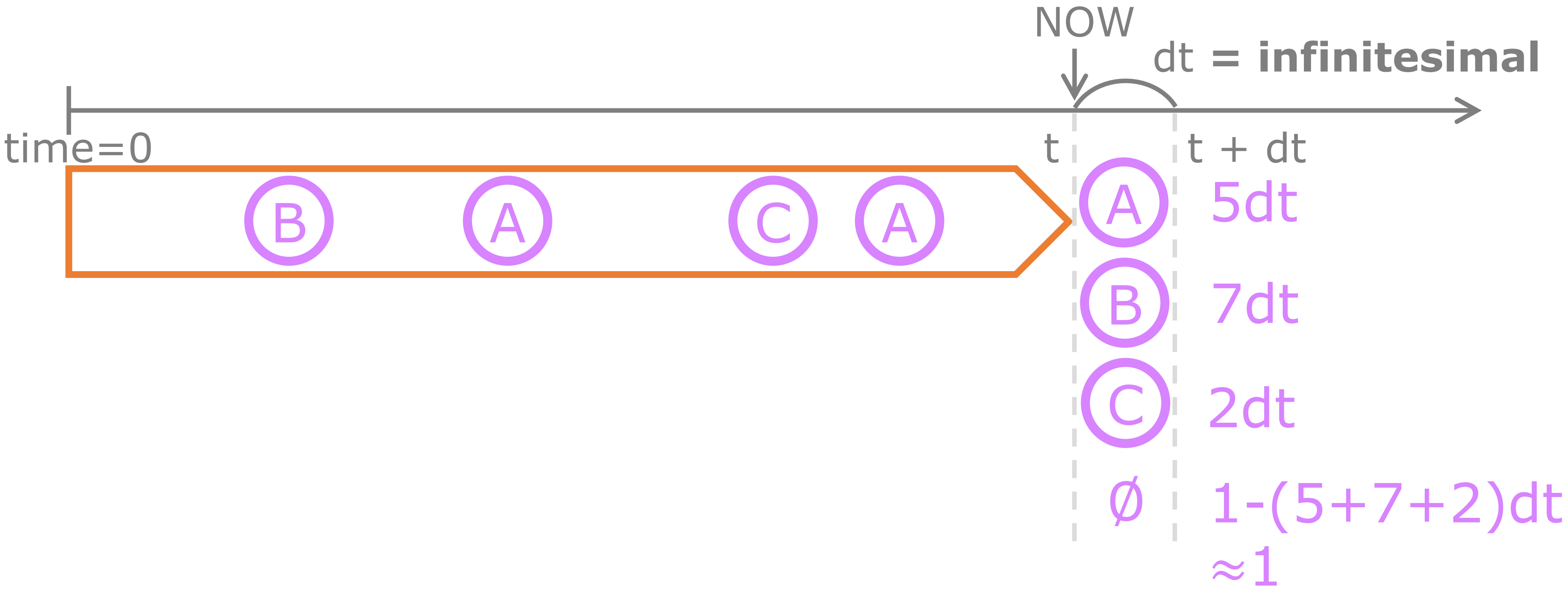
MLE: Max log prob of *data*



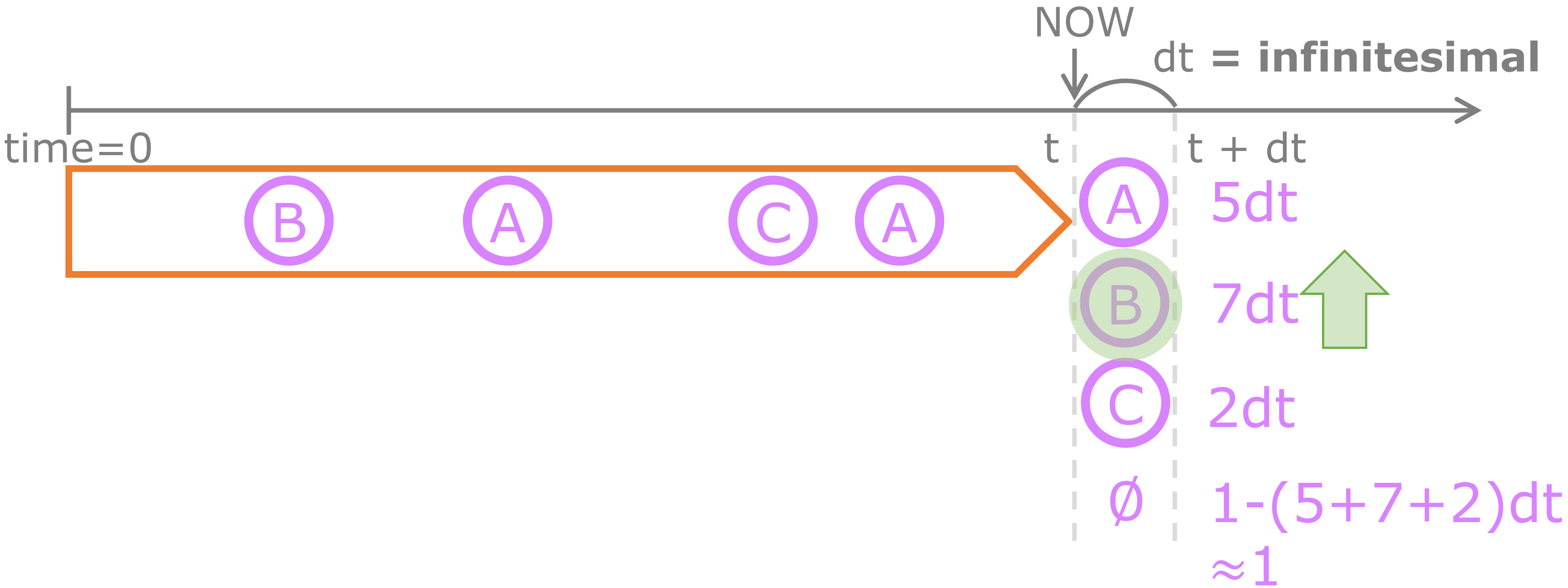
MLE: Max log prob of *data*



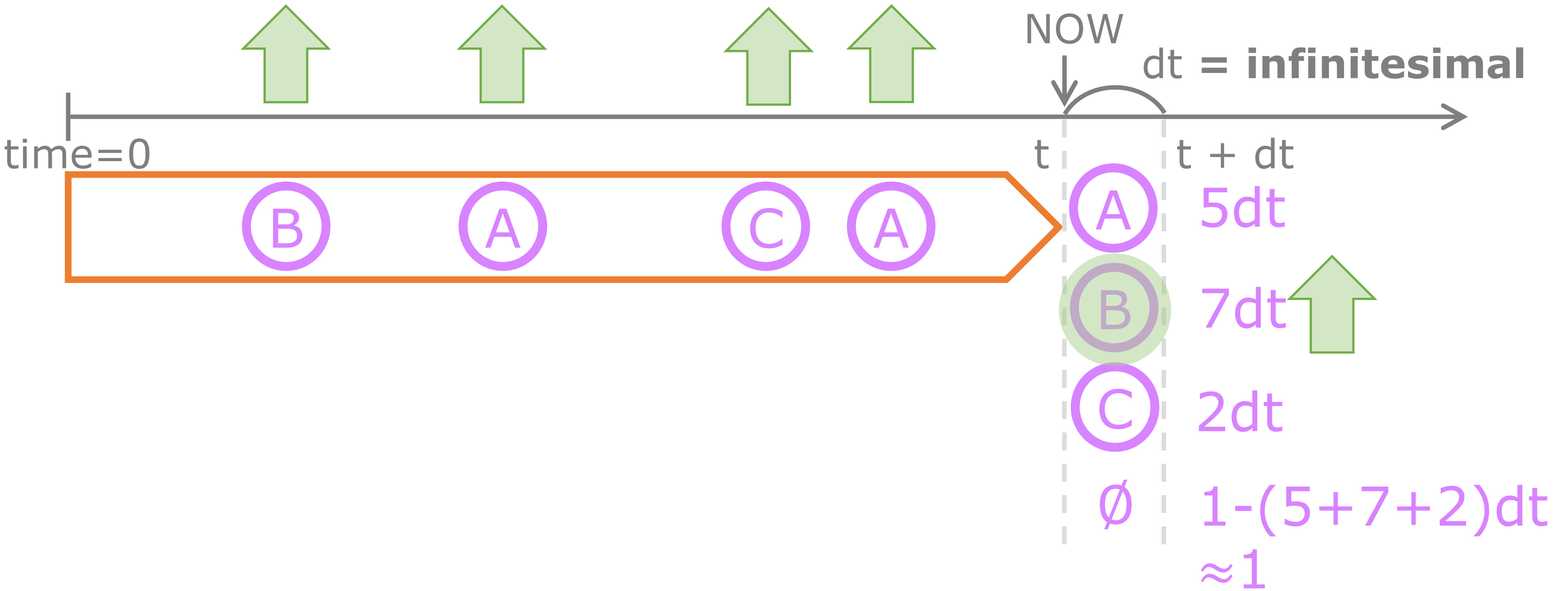
MLE: Max log prob of ***data***



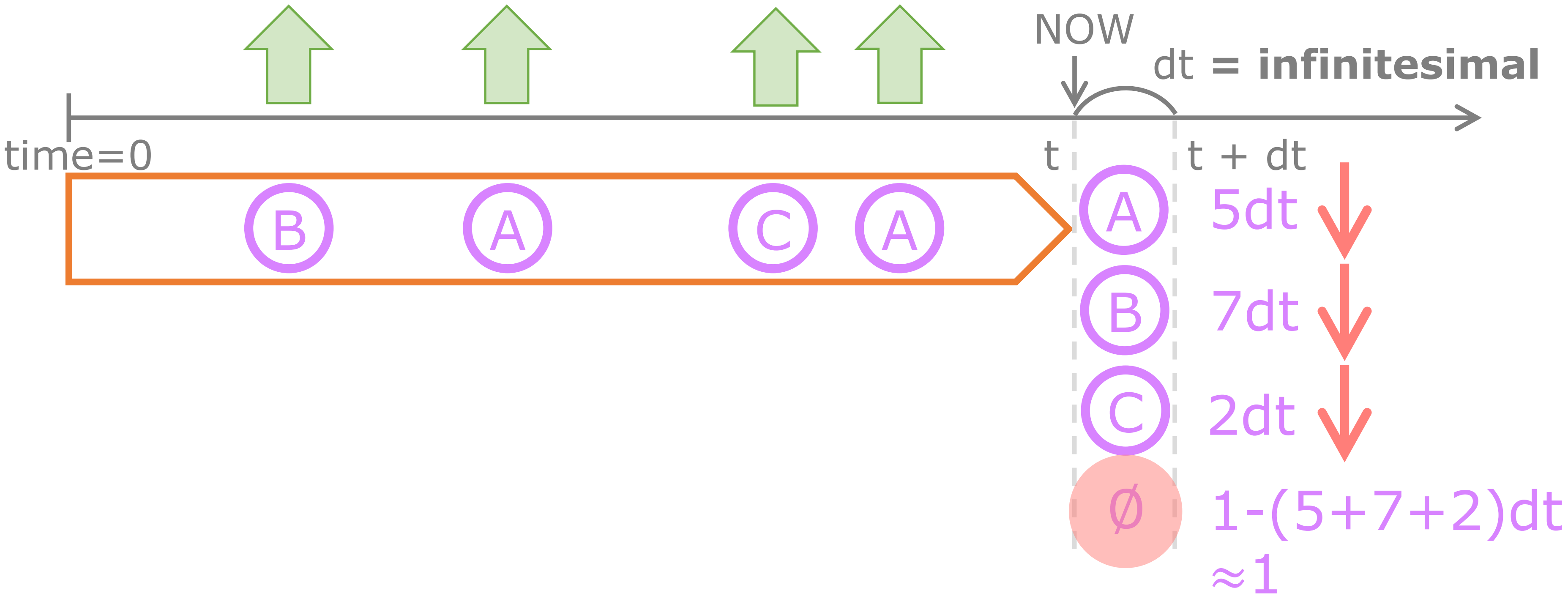
MLE: Max log prob of *data*



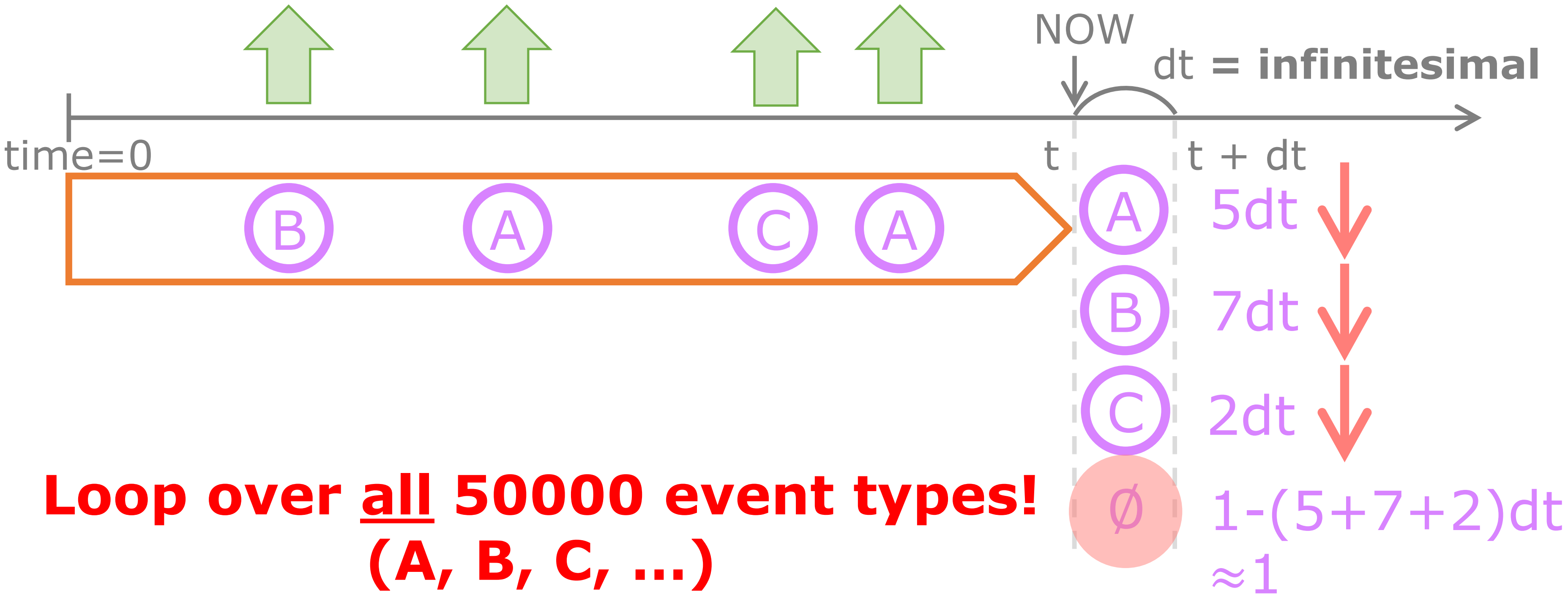
MLE: Max log prob of *data*



MLE: Max log prob of *data*

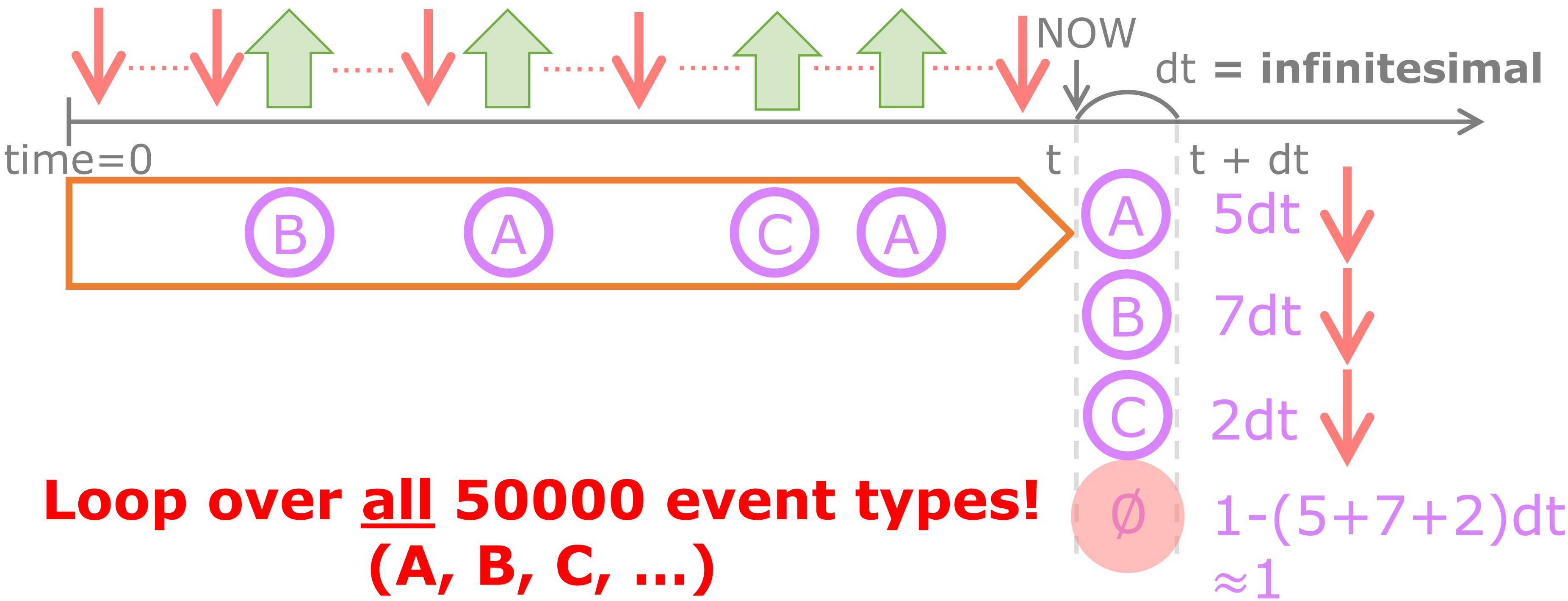


MLE: Max log prob of *data*



Loop over <u>all</u> 50000 event types!
(A, B, C, ...)

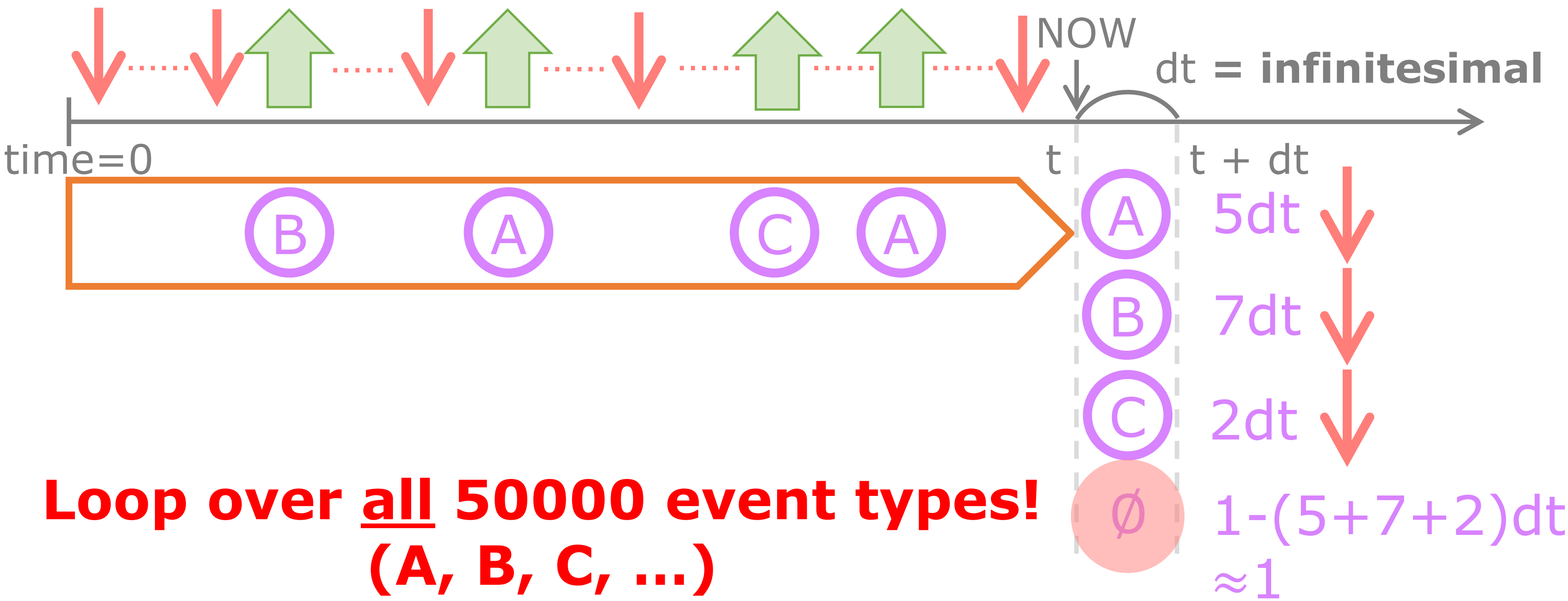
MLE: Max log prob of *data*



Loop over <u>all</u> 50000 event types!
(A, B, C, ...)

Integrate over <u>infinitely</u> many *non*-events!
(often approx by sampling)

MLE: Max log prob of *data*

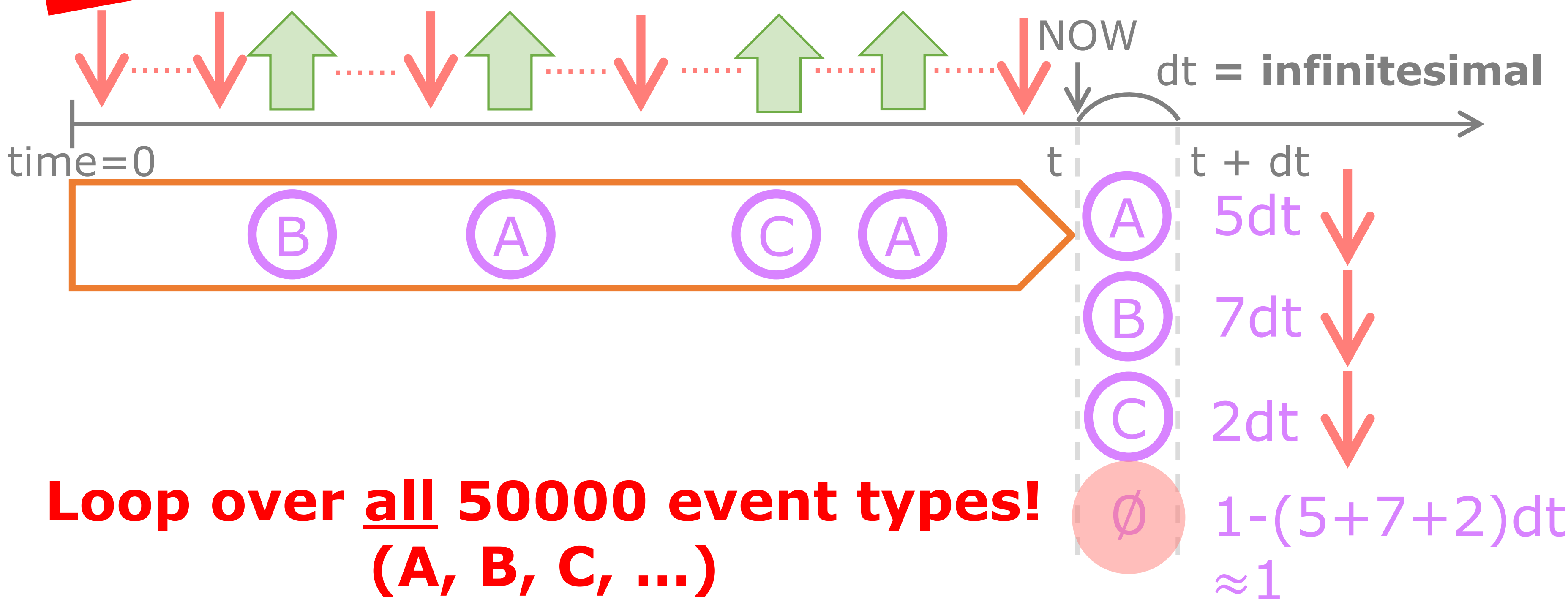


Loop over all 50000 event types!
(A, B, C, ...)

Integrate over infinitely many *non*-events!
(often approx by sampling)

SLOW

~~MLE: Max log prob of *data*~~



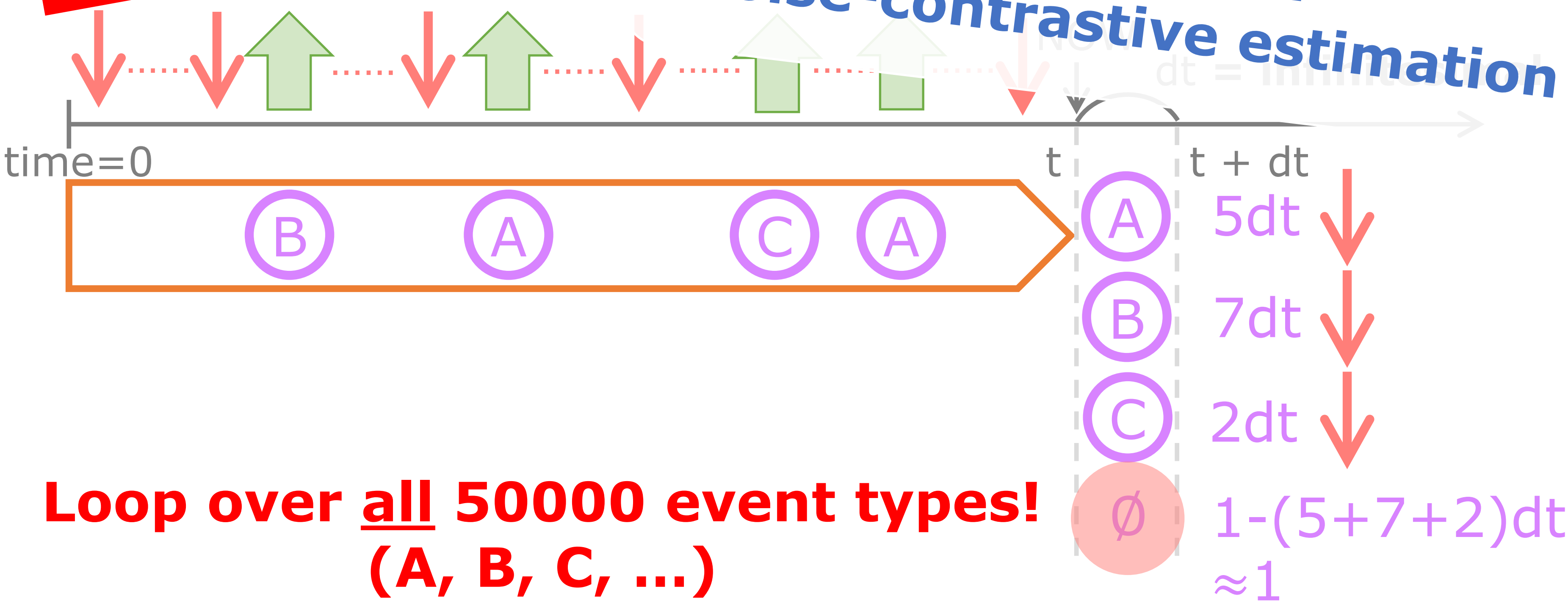
Loop over <u>all</u> 50000 event types!
(A, B, C, ...)

Integrate over <u>infinitely</u> many *non*-events!
(often approx by sampling)

SLOW

~~MLE: Max log prob of data~~

alternative: noise-contrastive estimation

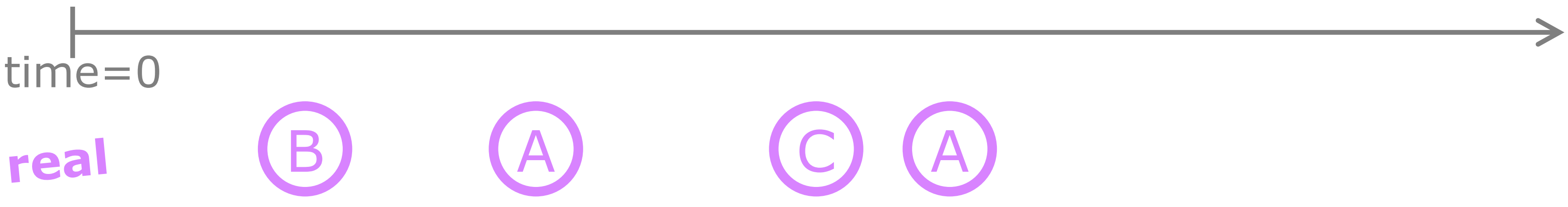


Loop over <u>all</u> 50000 event types!
(A, B, C, ...)

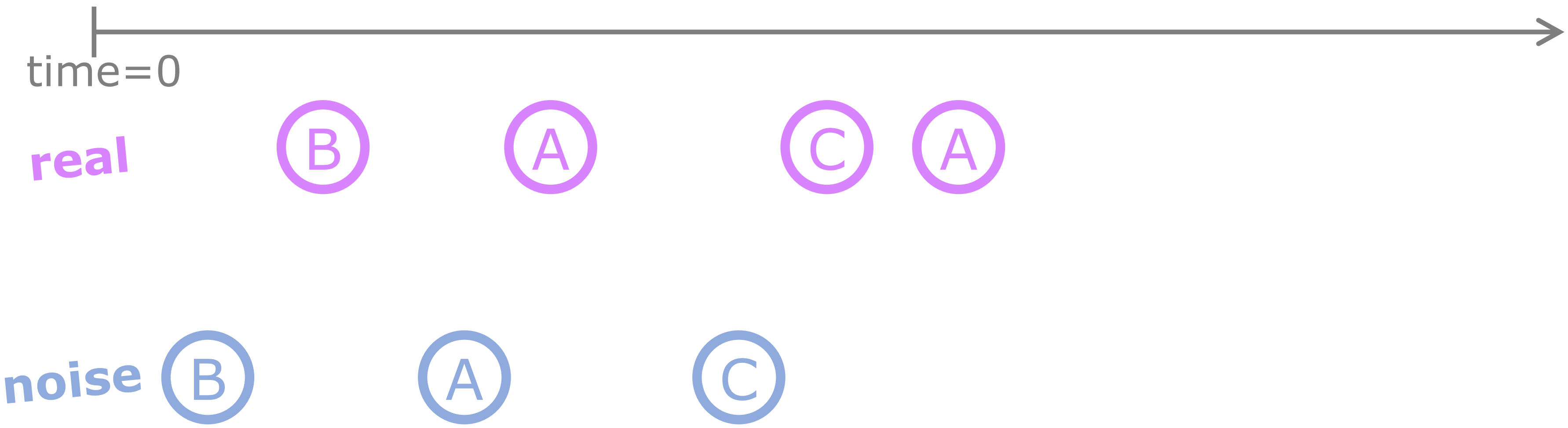
Integrate over <u>infinitely</u> many *non*-events!
(often approx by sampling)

SLOW

NCE: Max log prob of *correct discrimination*

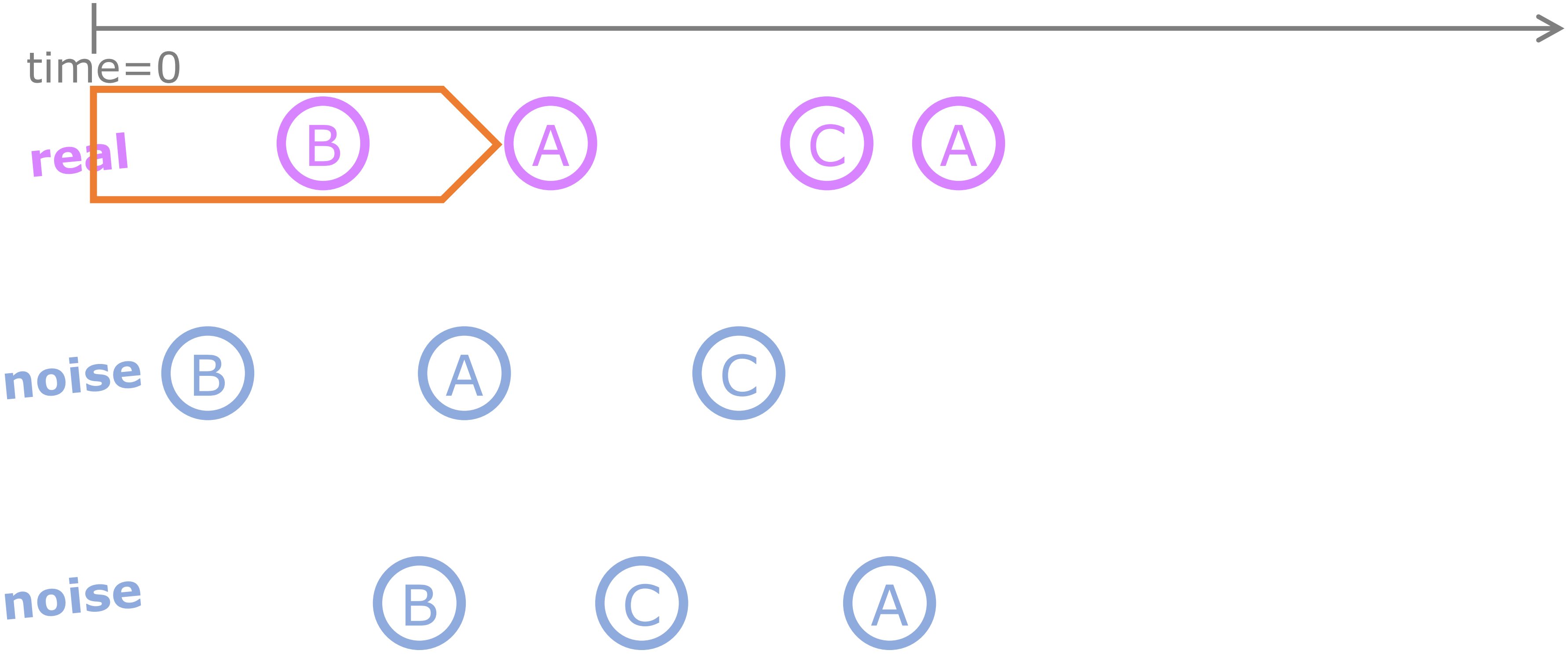


NCE: Max log prob of *correct discrimination*

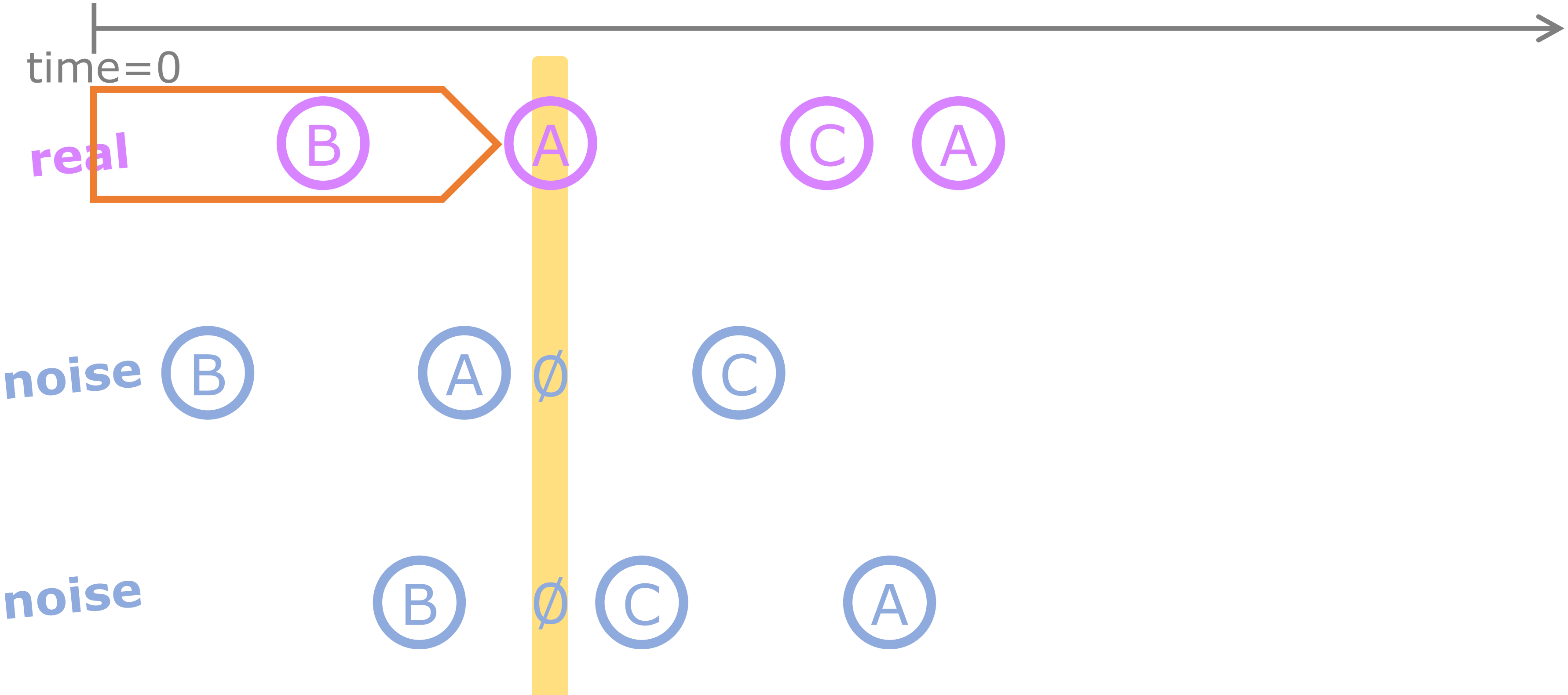


NCE: Max log prob of *correct discrimination*

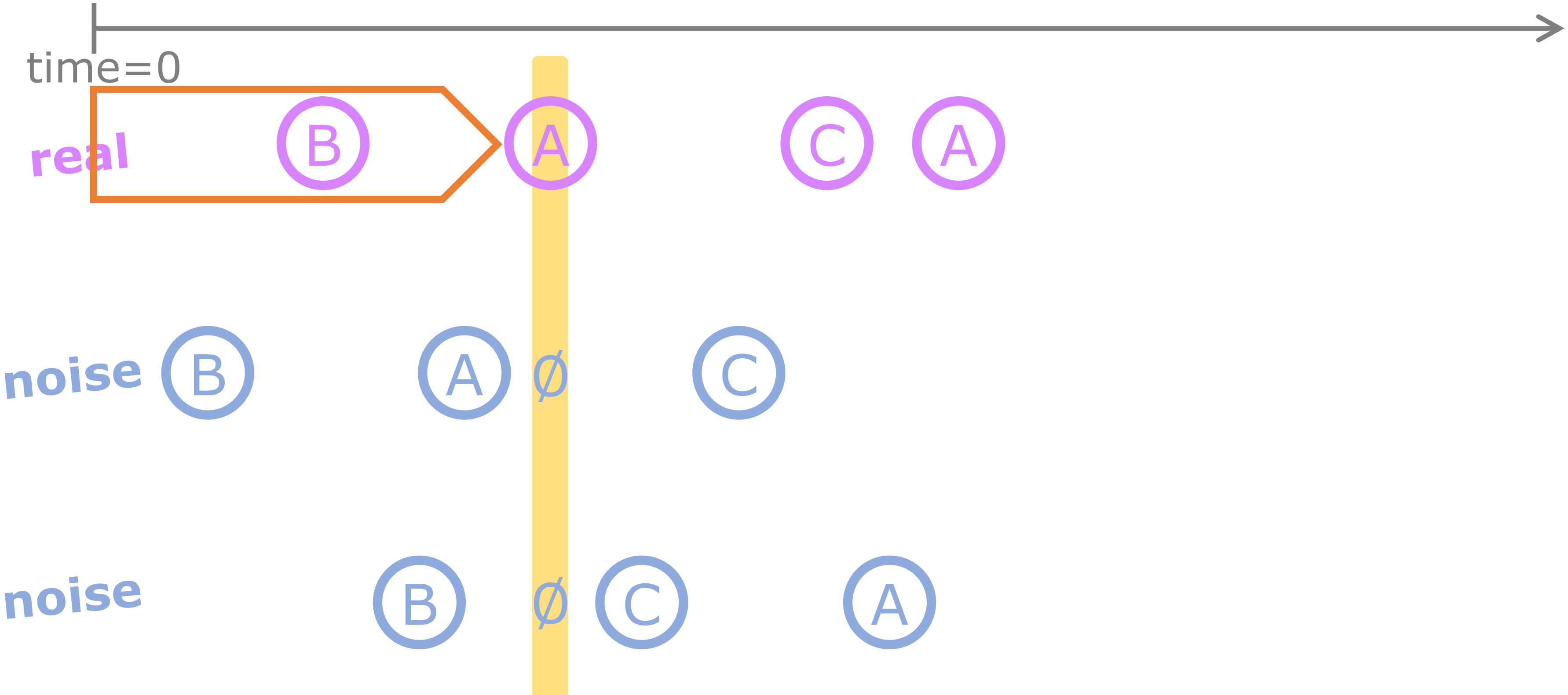
NCE: Max log prob of *correct discrimination*



NCE: Max log prob of *correct discrimination*

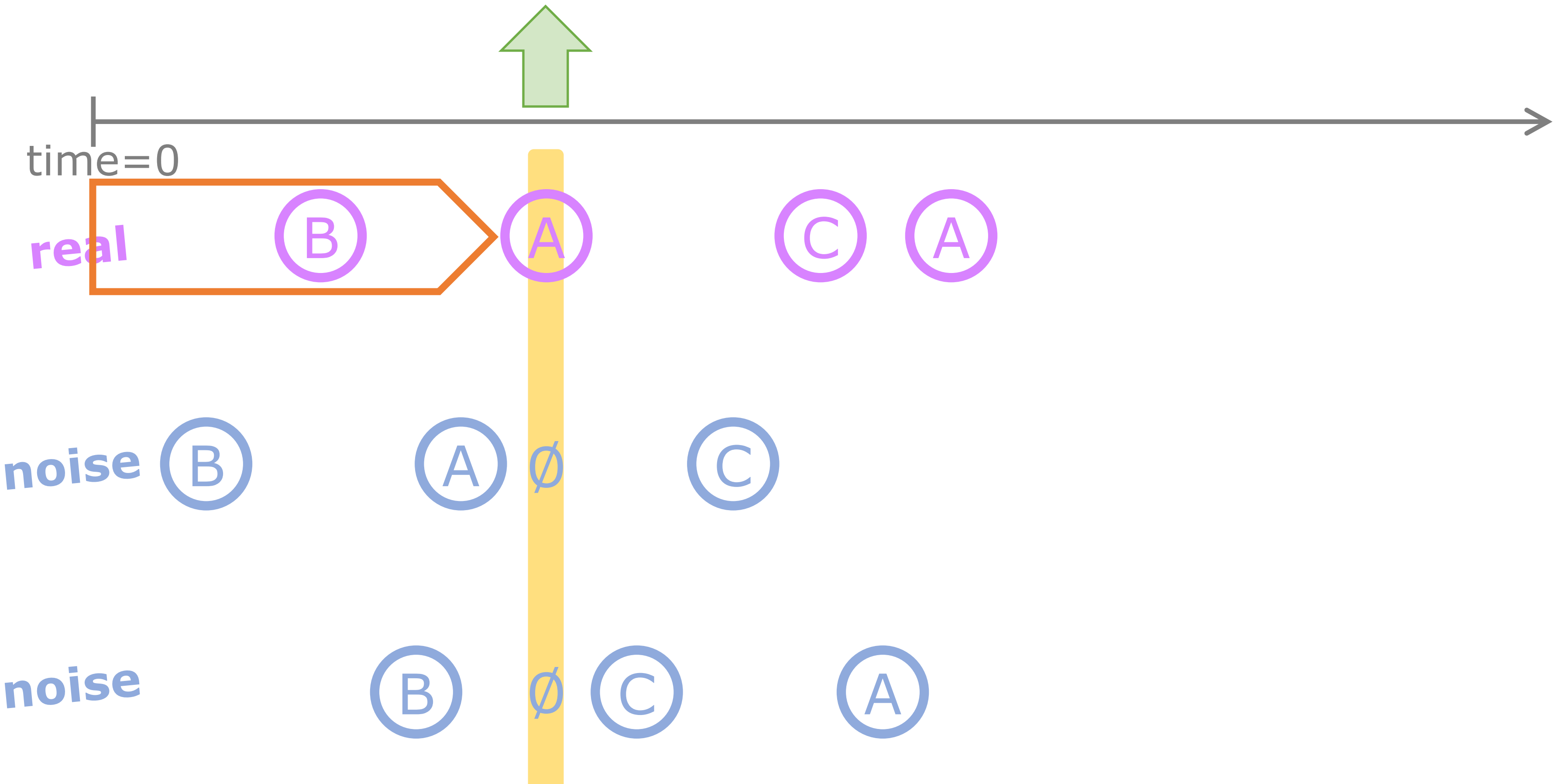


NCE: Max log prob of *correct discrimination*



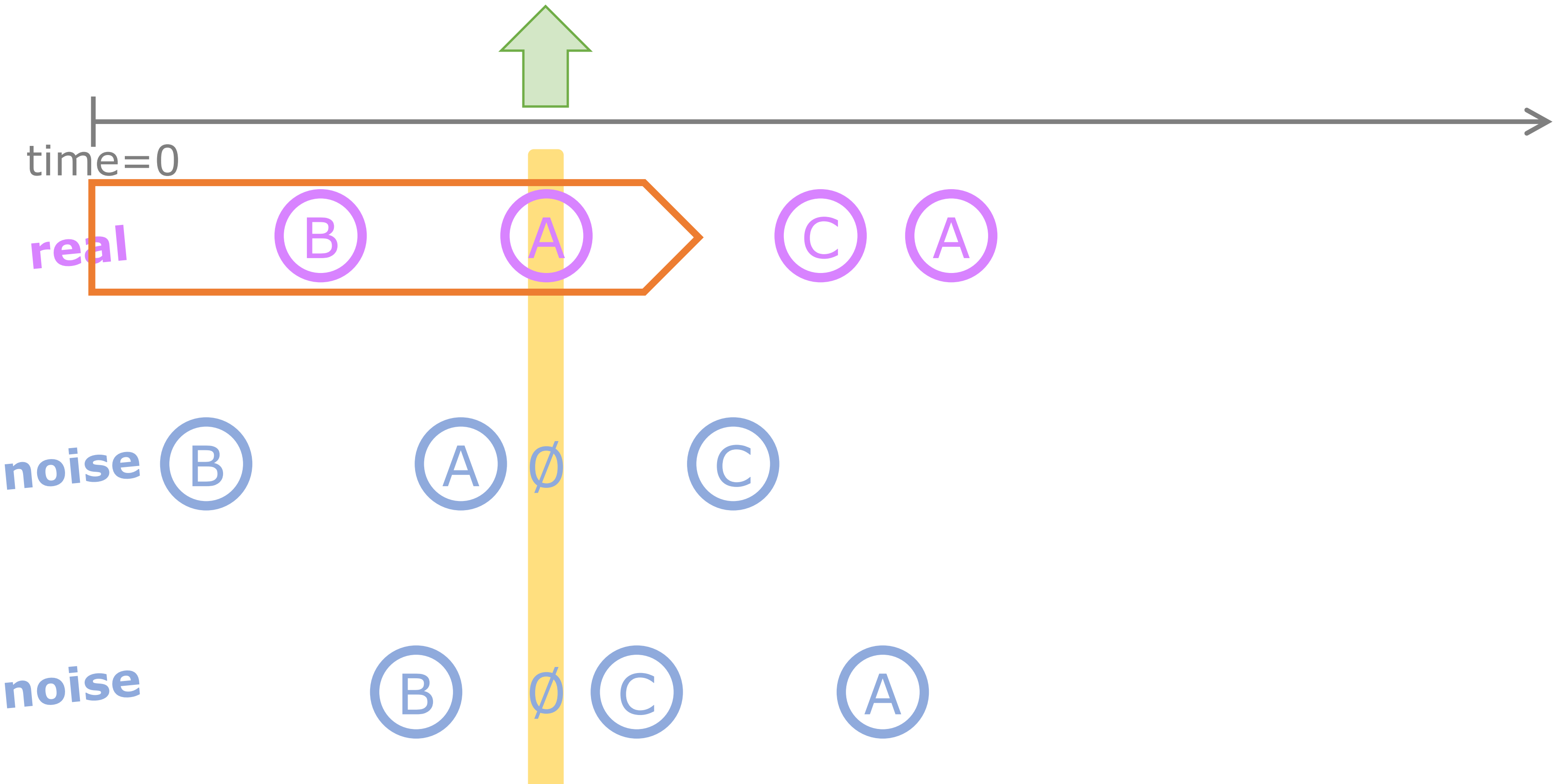
Which Is Real?

NCE: Max log prob of *correct discrimination*



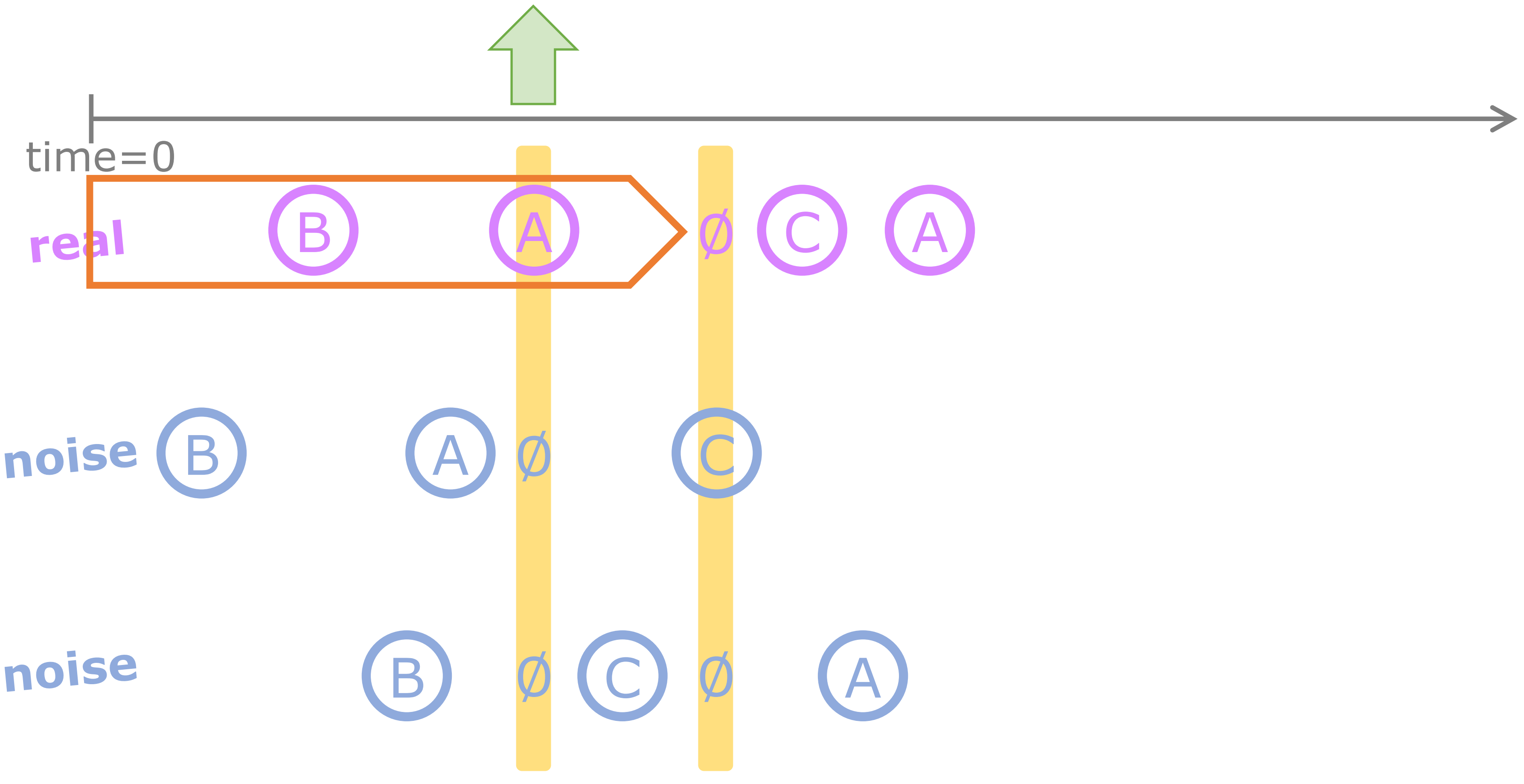
Which Is Real?

NCE: Max log prob of *correct discrimination*



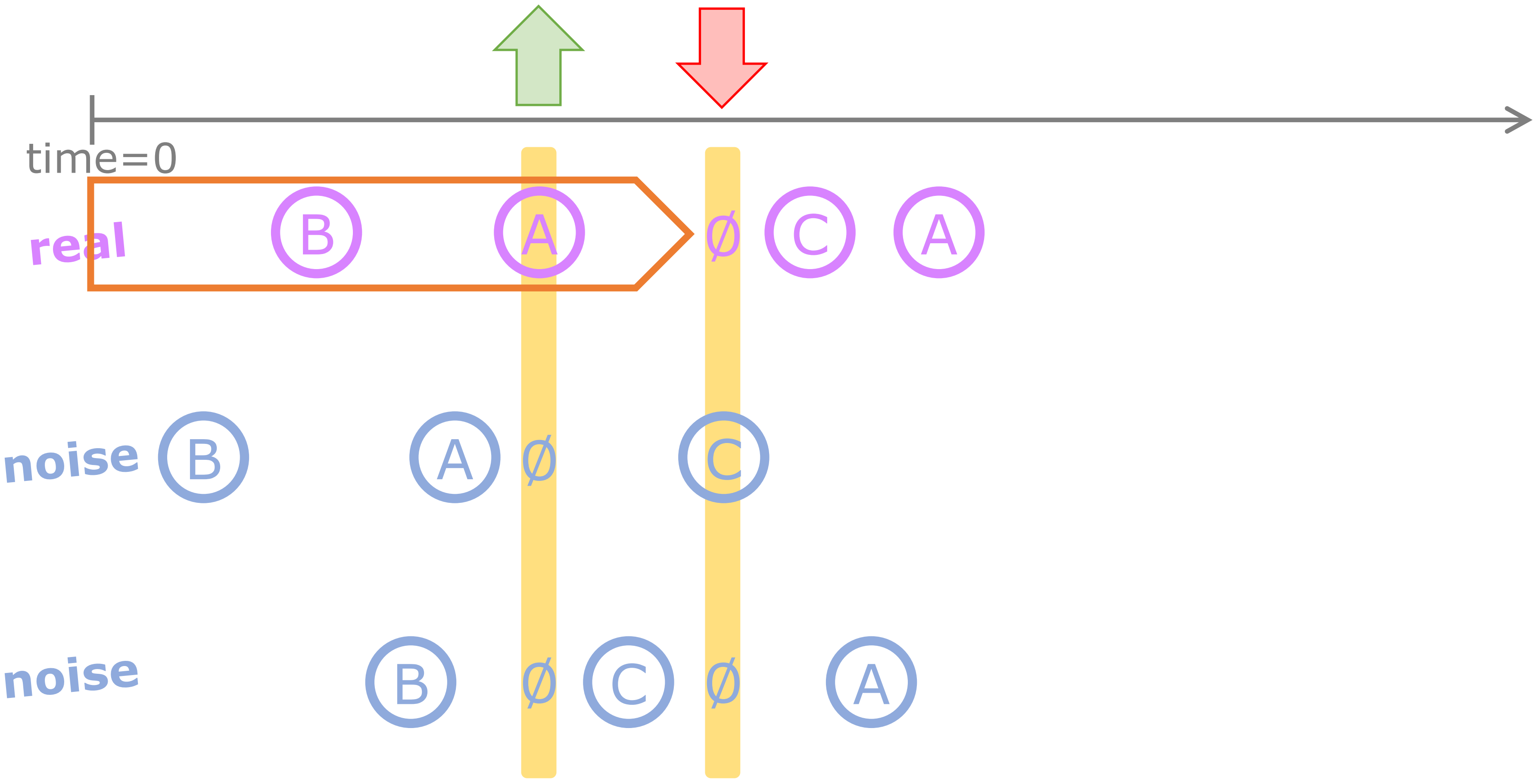
Which Is Real?

NCE: Max log prob of *correct discrimination*



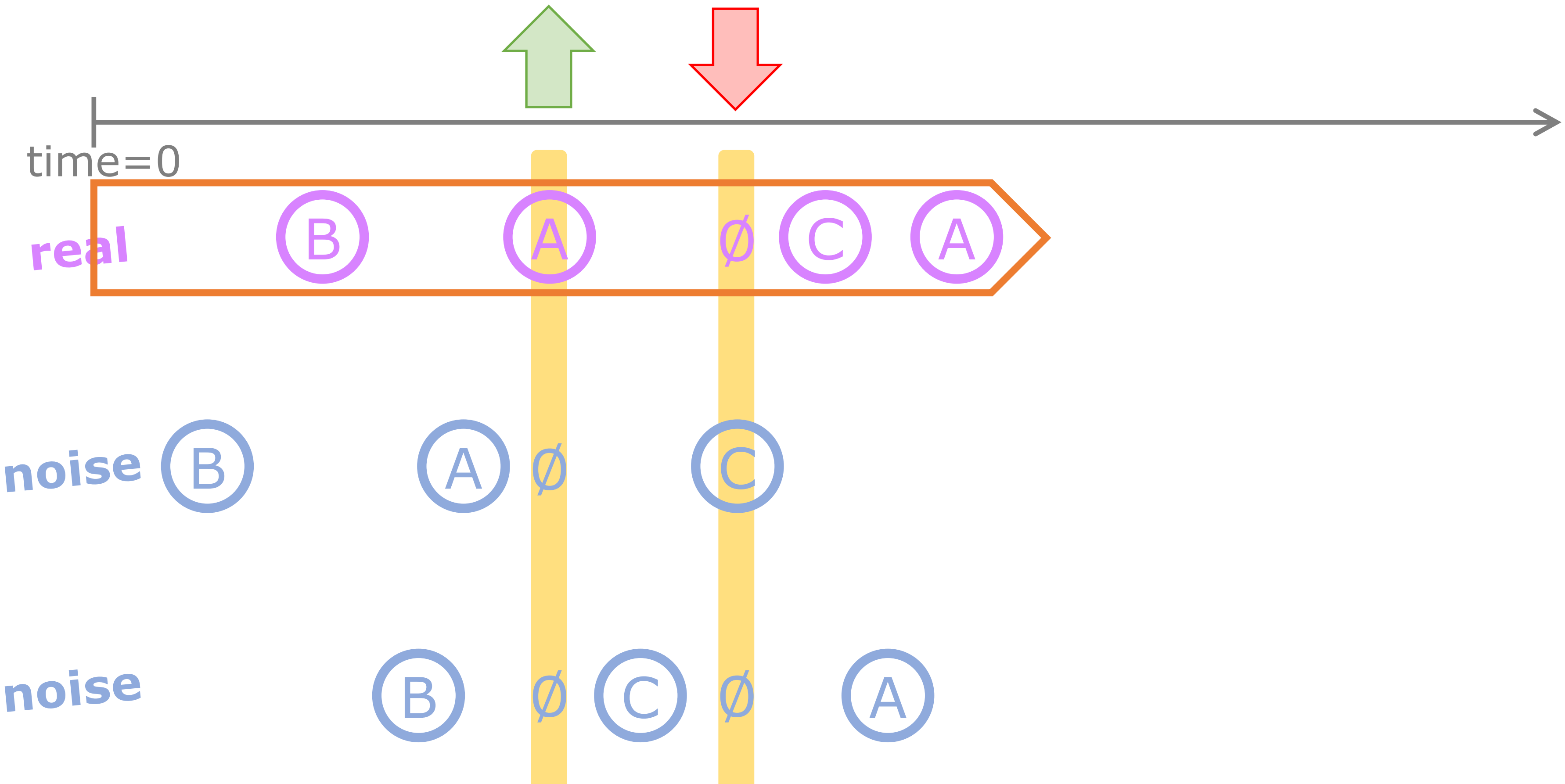
Which Is Real?

NCE: Max log prob of *correct discrimination*



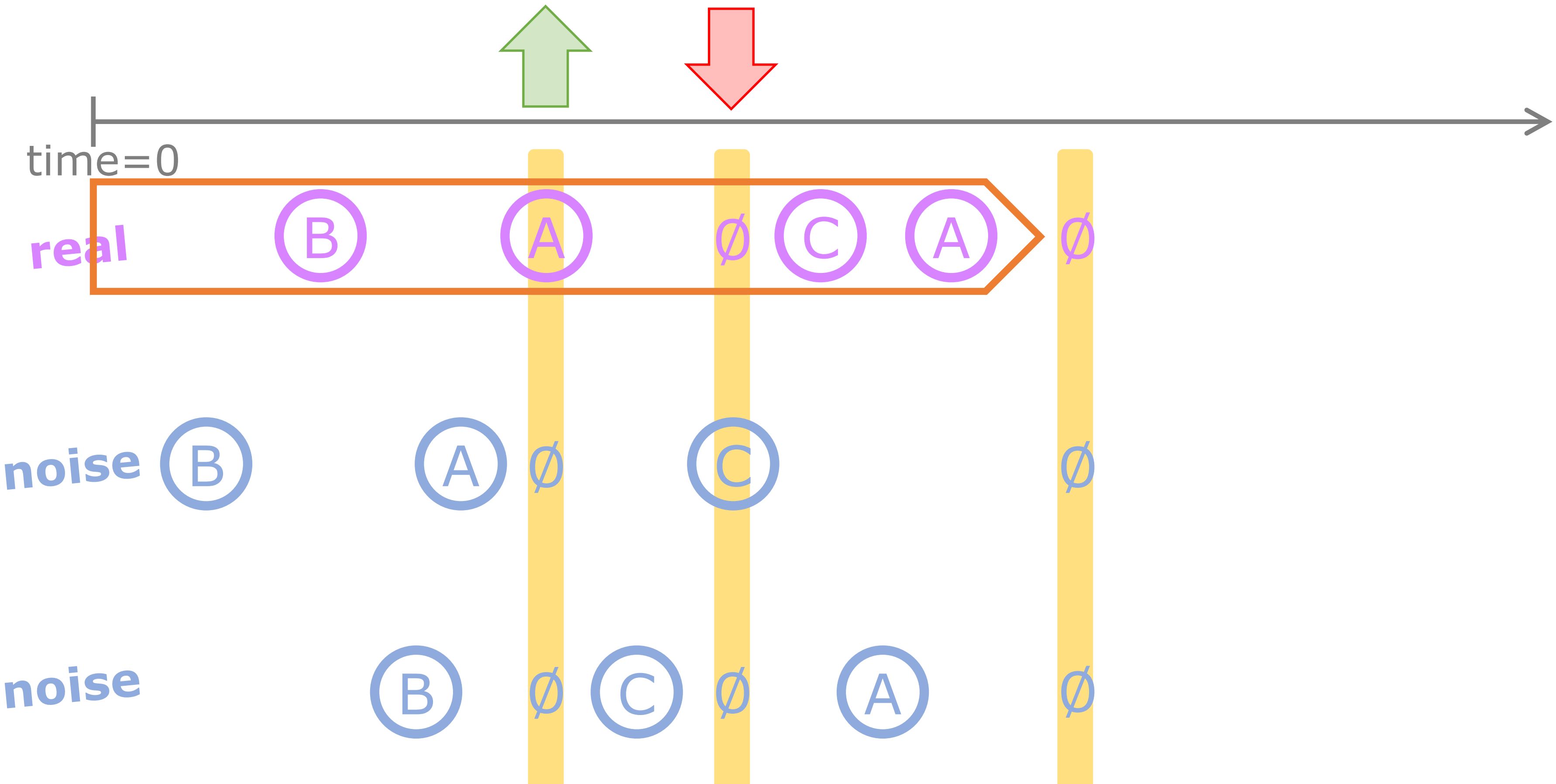
Which Is Real?

NCE: Max log prob of *correct discrimination*



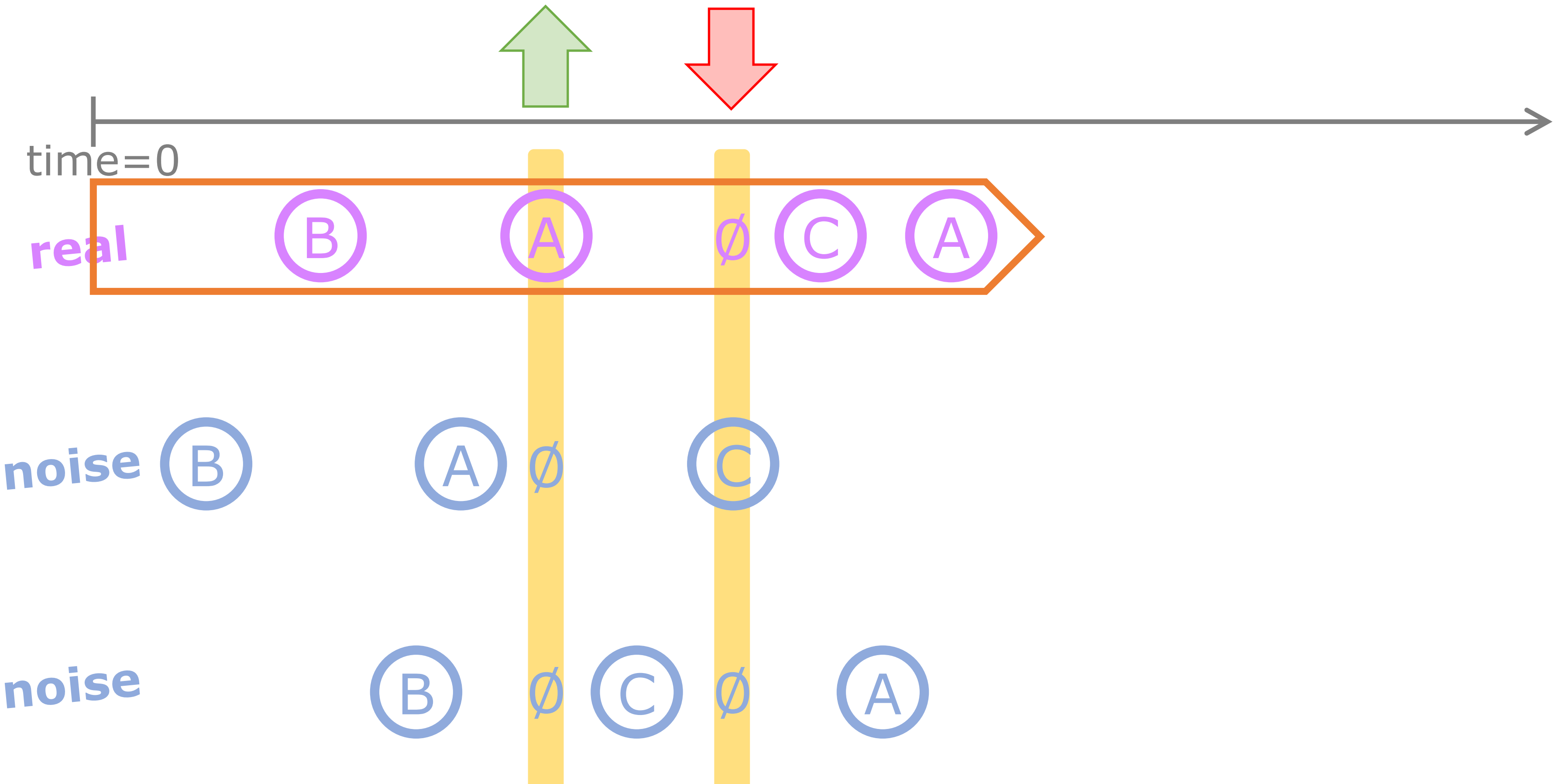
Which Is Real?

NCE: Max log prob of *correct discrimination*



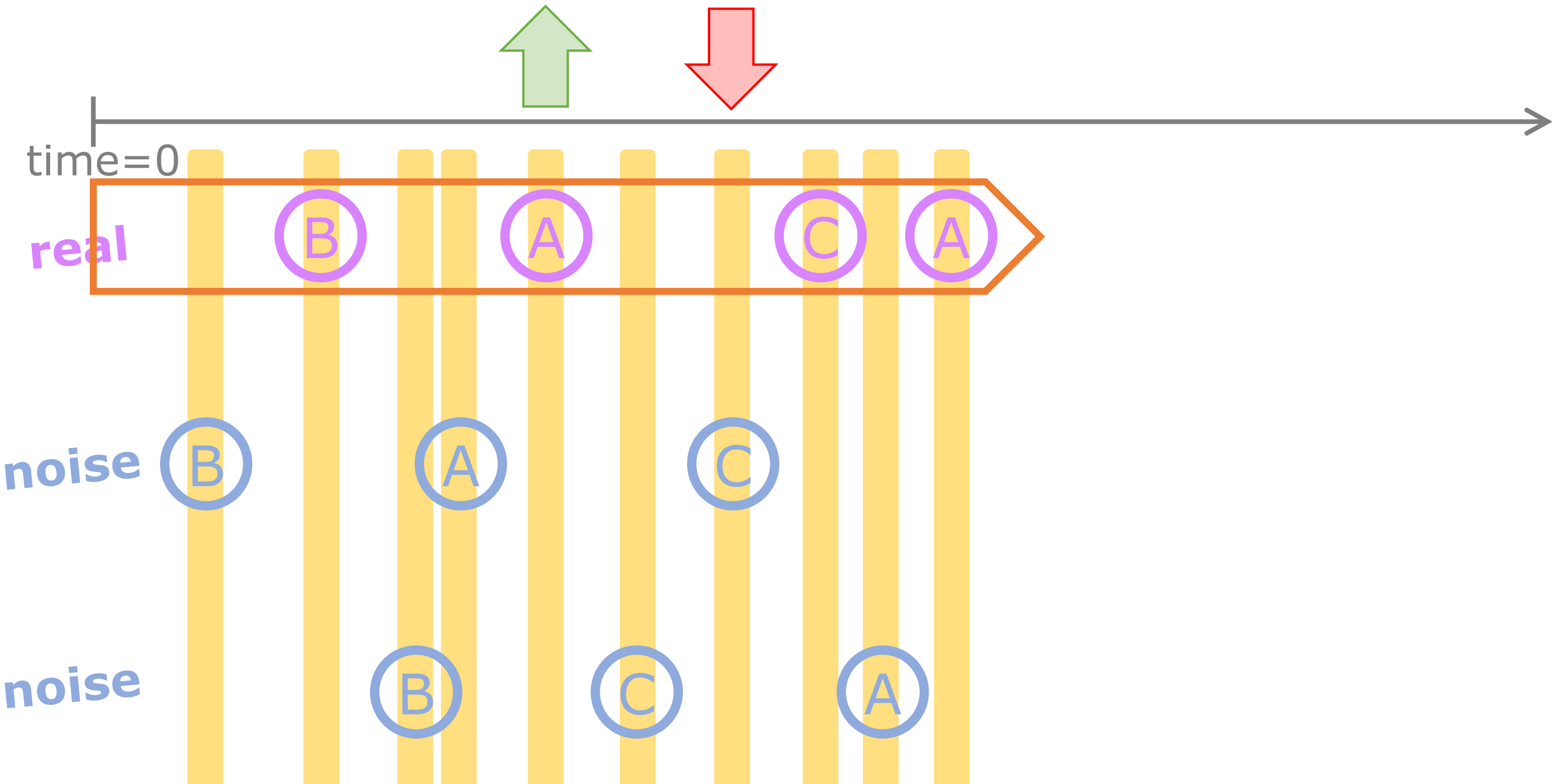
Which Is Real?

NCE: Max log prob of *correct discrimination*



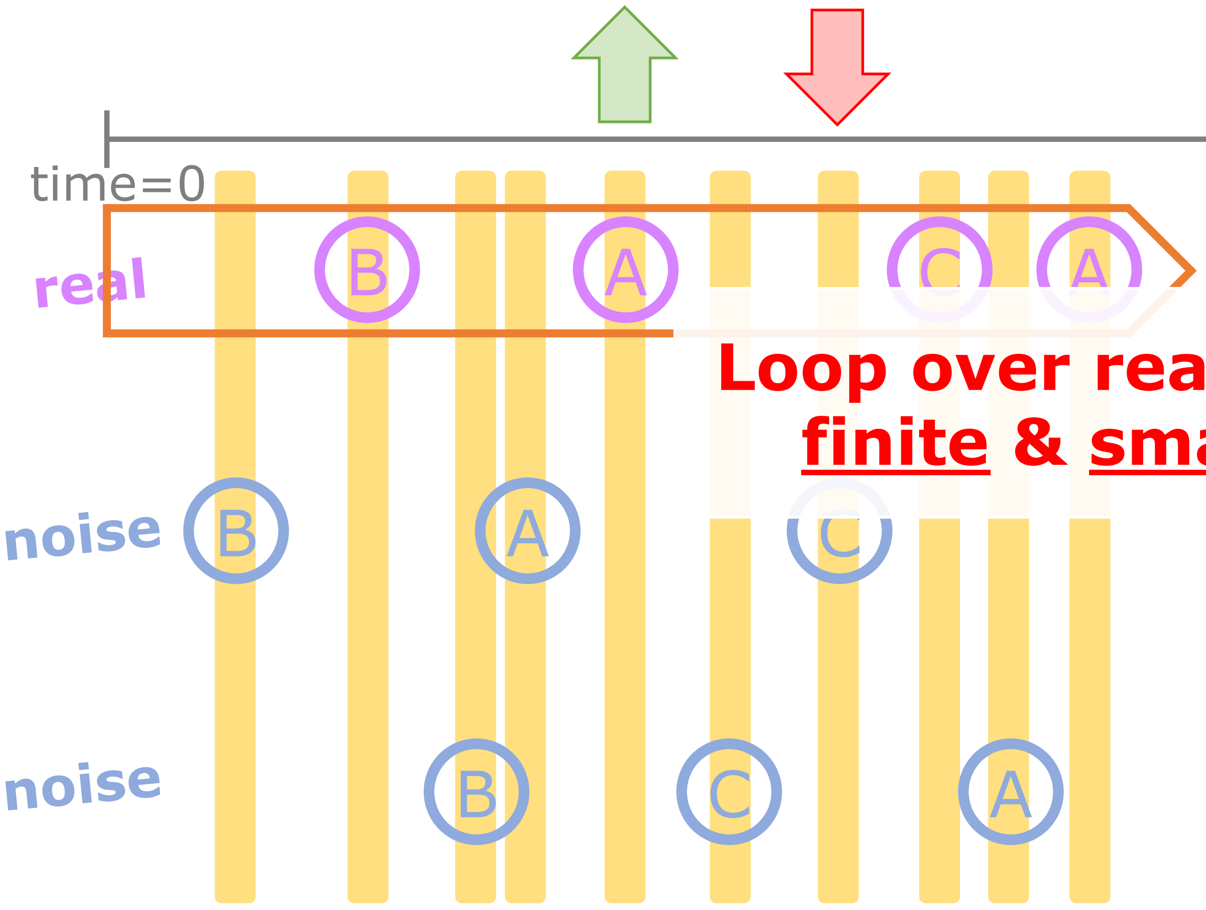
Which Is Real?

NCE: Max log prob of *correct discrimination*



Which Is Real?

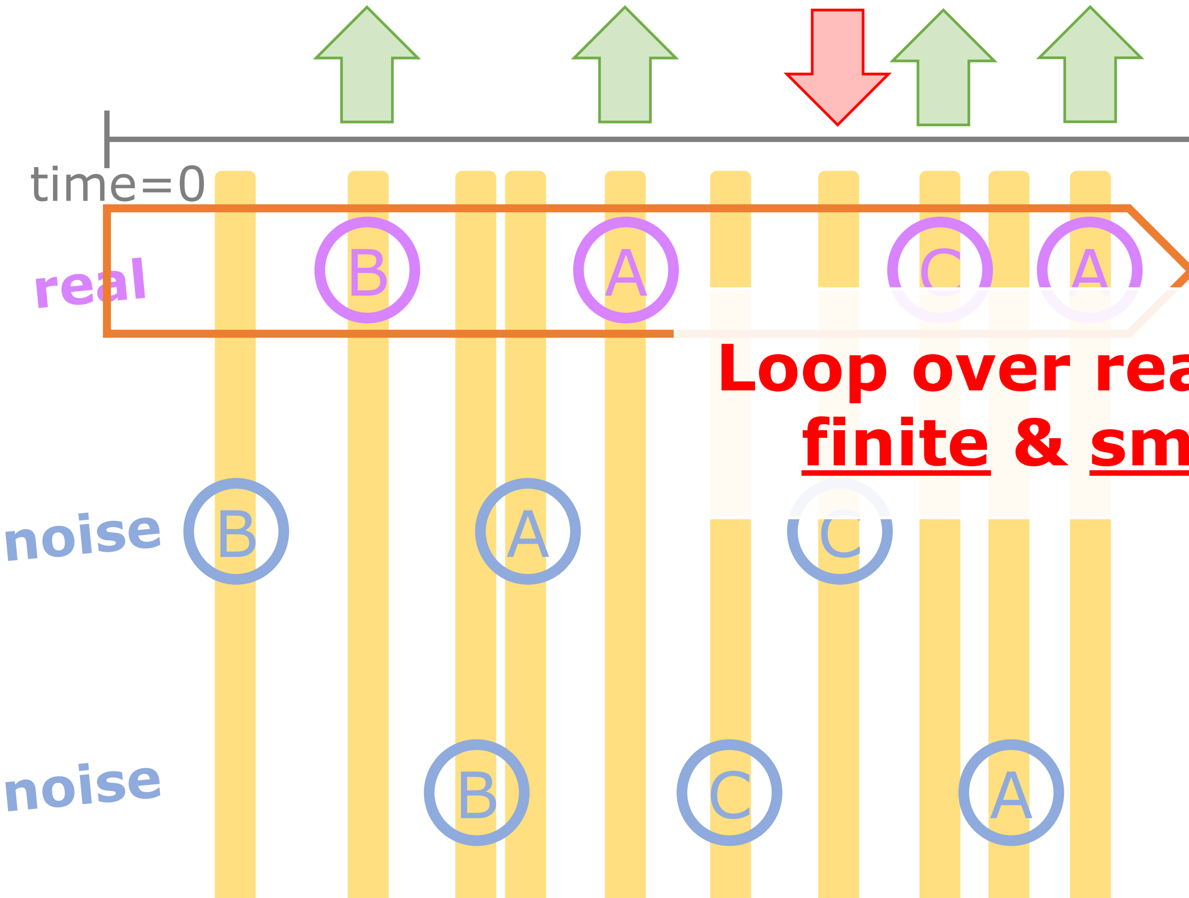
NCE: Max log prob of *correct discrimination*



Loop over real and noise events
finite & small → faster SGD

Which Is Real?

NCE: **Max log prob of** ***correct discrimination***

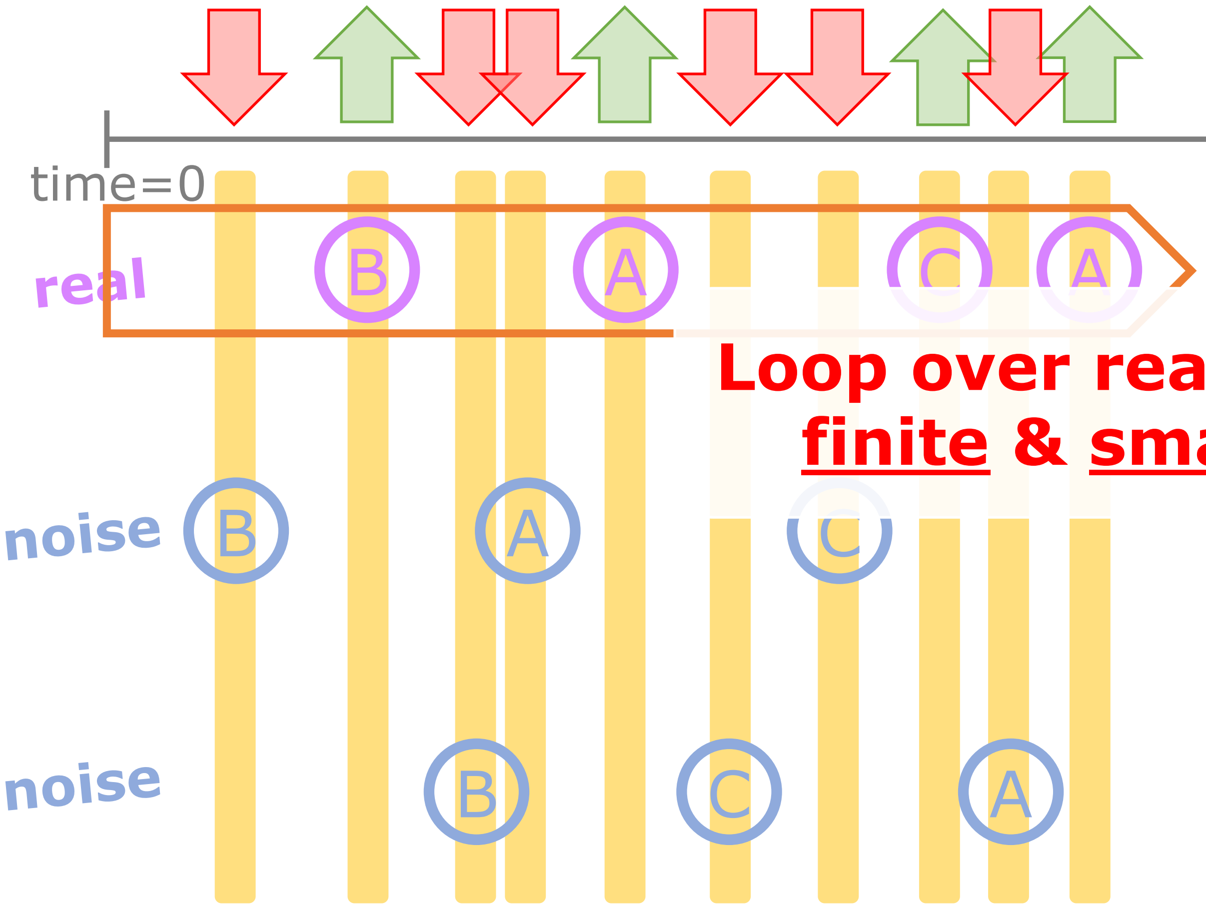


Loop over real and noise events

<u>finite</u> & <u>small</u> ➔ faster SGD

Which Is Real?

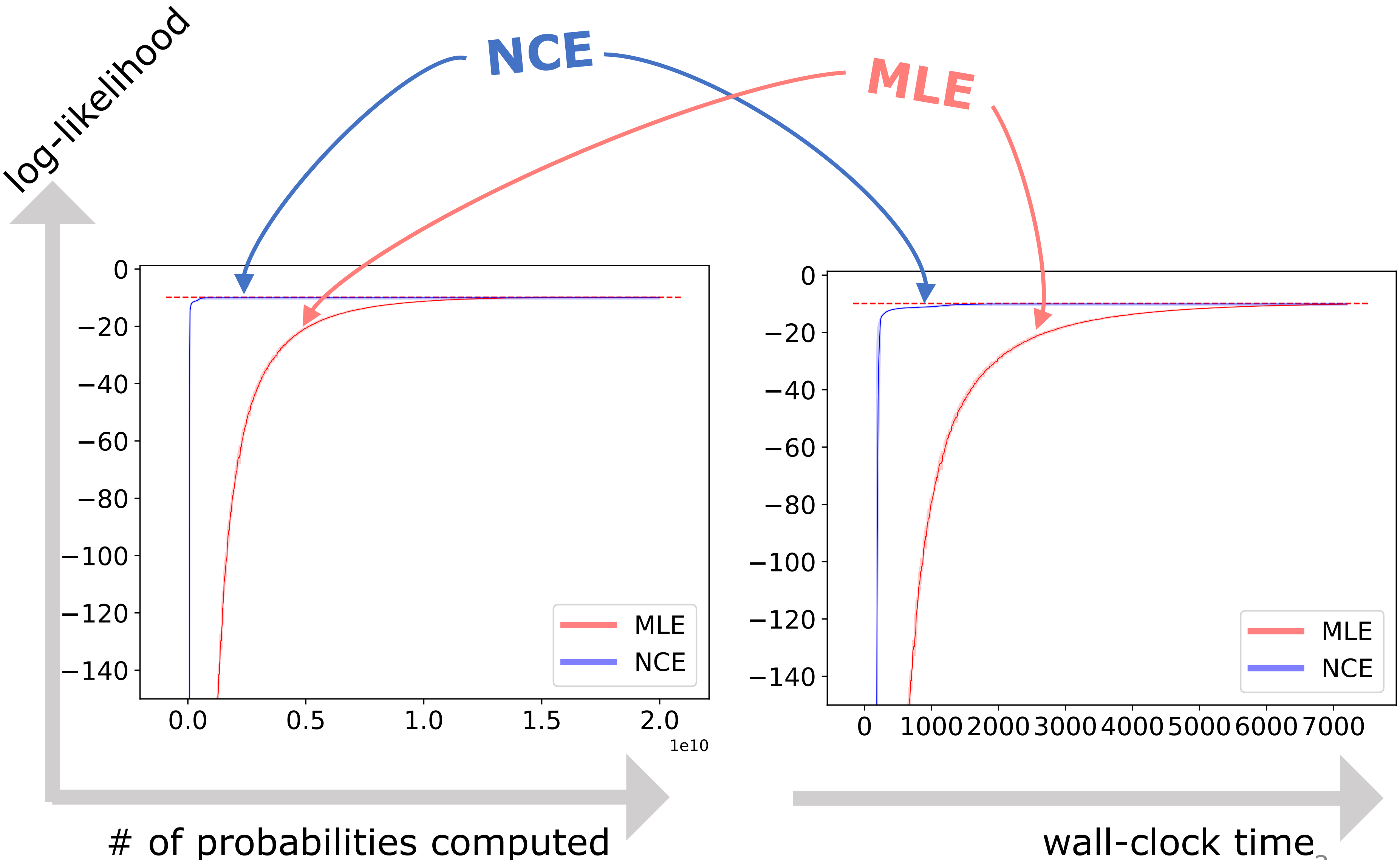
NCE: Max log prob of *correct discrimination*



Loop over real and noise events
<u>finite</u> & <u>small</u> → faster SGD

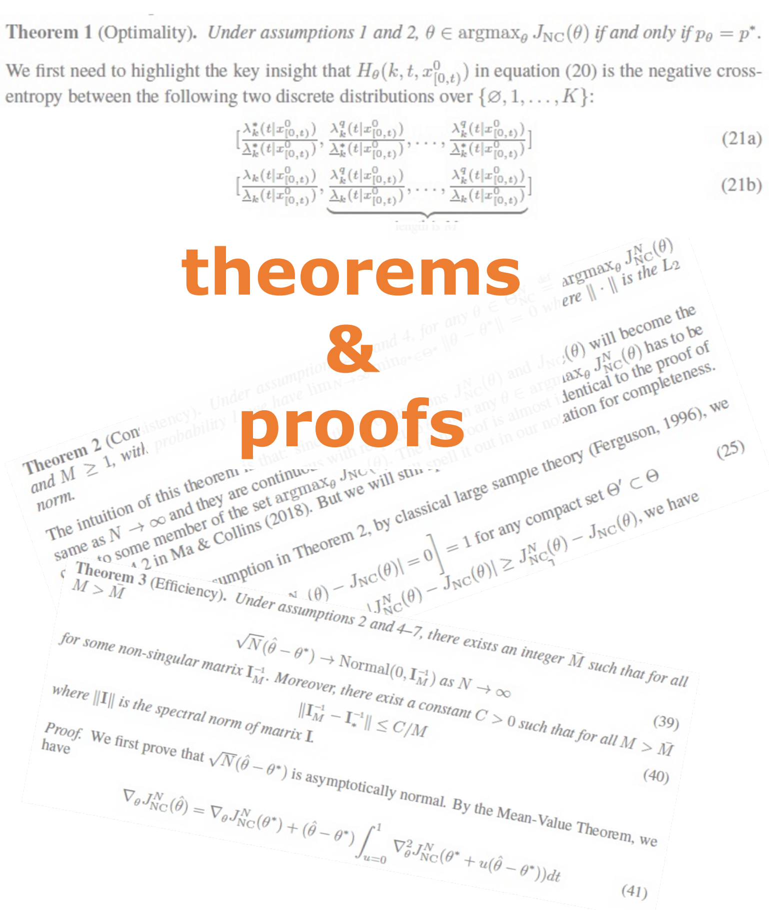
Which Is Real?

NCE vs MLE: what it typically looks like



NCE: More in paper

NCE: More in paper


Theorem 1 (Optimality). *Under assumptions 1 and 2, $\theta \in \operatorname{argmax}_\theta J_{\mathrm{NC}}(\theta)$ if and only if $p_\theta = p^*$.*

We first need to highlight the key insight that $H_\theta(k, t, x^0_{[0,t)})$ in equation (20) is the negative cross-entropy between the following two discrete distributions over $\{\varnothing, 1, \ldots, K\}$:

$$\left[\frac{\lambda^*_k(t \mid x^0_{[0,t)})}{\Delta^*_k(t \mid x^0_{[0,t)})}, \frac{\lambda^q_k(t \mid x^0_{[0,t)})}{\Delta^*_k(t \mid x^0_{[0,t)})}, \ldots, \frac{\lambda^q_k(t \mid x^0_{[0,t)})}{\Delta^*_k(t \mid x^0_{[0,t)})}\right] \quad (21a)$$

$$\left[\frac{\lambda_k(t \mid x^0_{[0,t)})}{\Delta_k(t \mid x^0_{[0,t)})}, \underbrace{\frac{\lambda^q_k(t \mid x^0_{[0,t)})}{\Delta_k(t \mid x^0_{[0,t)})}, \ldots, \frac{\lambda^q_k(t \mid x^0_{[0,t)})}{\Delta_k(t \mid x^0_{[0,t)})}}_{\text{length is } M}\right] \quad (21b)$$

Theorem 2 (Con... *and $M \geq 1$, with probability 1 ... $\operatorname{argmax}_\theta J^N_{\mathrm{NC}}(\theta)$ ... where $\|\cdot\|$ is the L_2 norm.*

The intuition of this theorem is ... $N \to \infty$ and they are continuous ... (θ) will become the same as ... to some member of the set $\operatorname{argmax}_\theta J_{\mathrm{NC}}(\theta)$ ... $\operatorname{argmax}_\theta J^N_{\mathrm{NC}}(\theta)$ has to be ... 2 in Ma & Collins (2018). But we will still spell it out in our notation for completeness.

... assumption in Theorem 2, by classical large sample theory (Ferguson, 1996), we ...

$$\ldots = 1 \text{ for any compact set } \Theta' \subset \Theta \quad (25)$$

... $J^N_{\mathrm{NC}}(\theta) - J_{\mathrm{NC}}(\theta)$, we have

Theorem 3 (Efficiency). *Under assumptions 2 and 4–7, there exists an integer $\bar{M}$ such that for all $M > \bar{M}$*

$$\sqrt{N}(\hat{\theta} - \theta^*) \to \mathrm{Normal}(0, \mathbf{I}_M^{-1}) \text{ as } N \to \infty$$

for some non-singular matrix $\mathbf{I}_M^{-1}$. Moreover, there exist a constant $C > 0$ such that for all $M > \bar{M}$

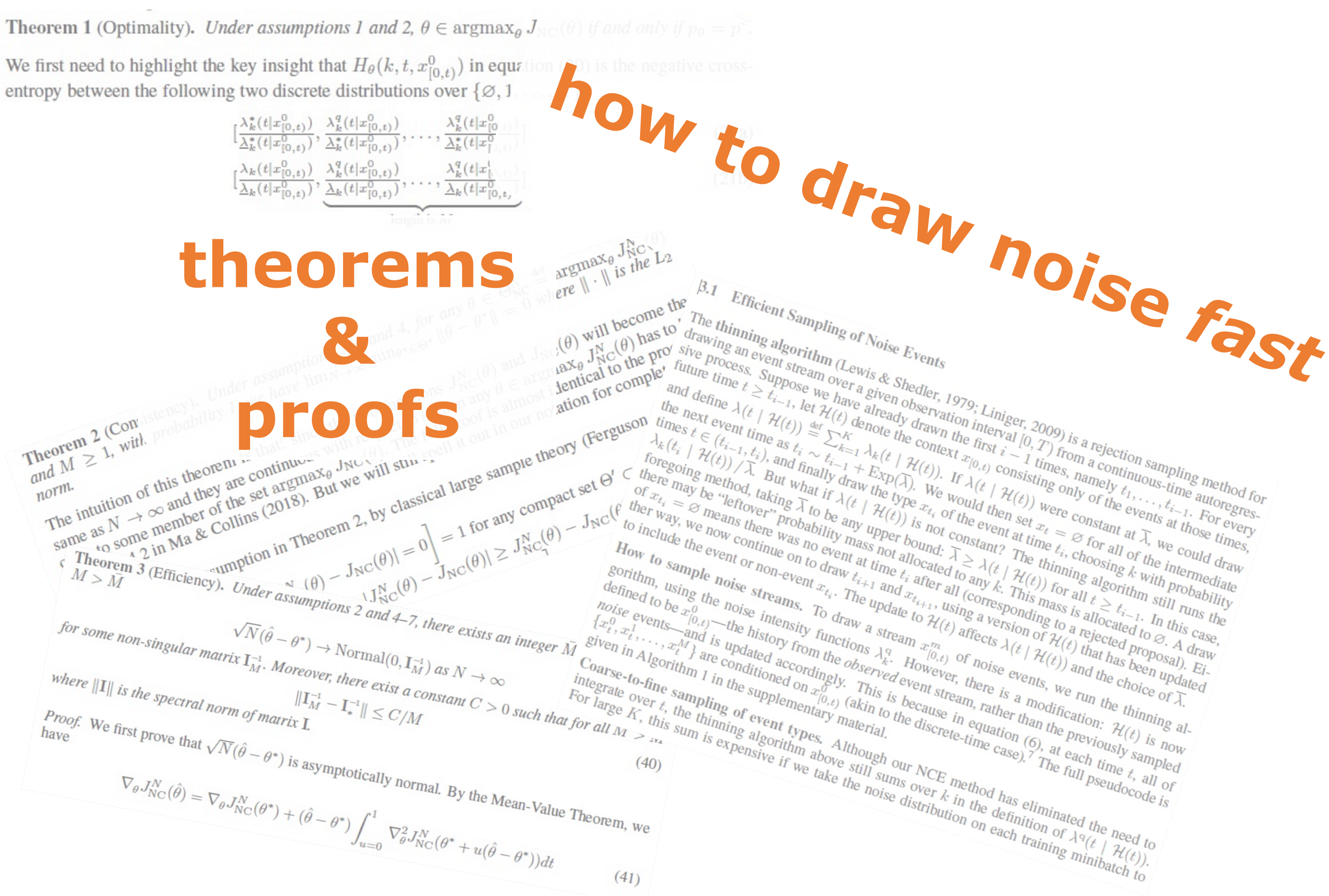
$$\|\mathbf{I}_M^{-1} - \mathbf{I}_*^{-1}\| \leq C/M \quad (39)$$

where $\|\mathbf{I}\|$ is the spectral norm of matrix $\mathbf{I}$. (40)

Proof. We first prove that $\sqrt{N}(\hat{\theta} - \theta^*)$ is asymptotically normal. By the Mean-Value Theorem, we have

$$\nabla_\theta J^N_{\mathrm{NC}}(\hat{\theta}) = \nabla_\theta J^N_{\mathrm{NC}}(\theta^*) + (\hat{\theta} - \theta^*) \int_{u=0}^{1} \nabla^2_\theta J^N_{\mathrm{NC}}(\theta^* + u(\hat{\theta} - \theta^*))dt \quad (41)$$

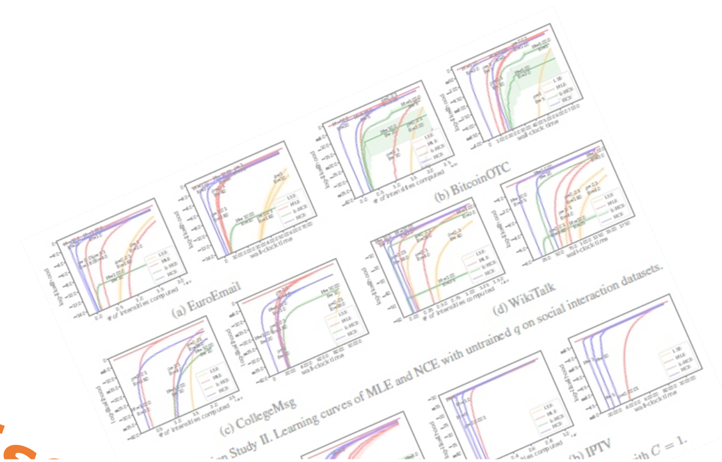
NCE: More in paper



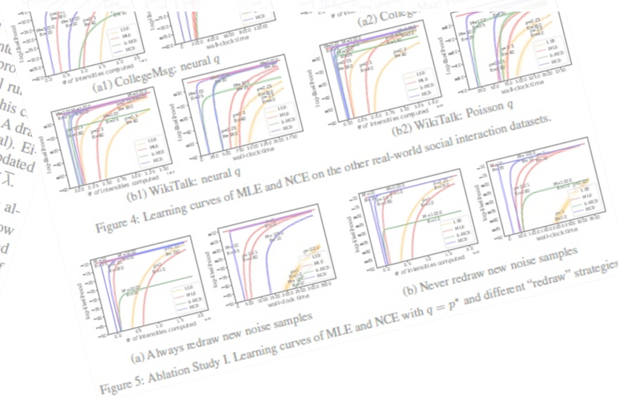
NCE: More in paper

how to draw noise

theorems
&
proofs



more results & analysis



THANK YOU

Hongyuan Mei, Tom Wan, Jason Eisner

Johns Hopkins University