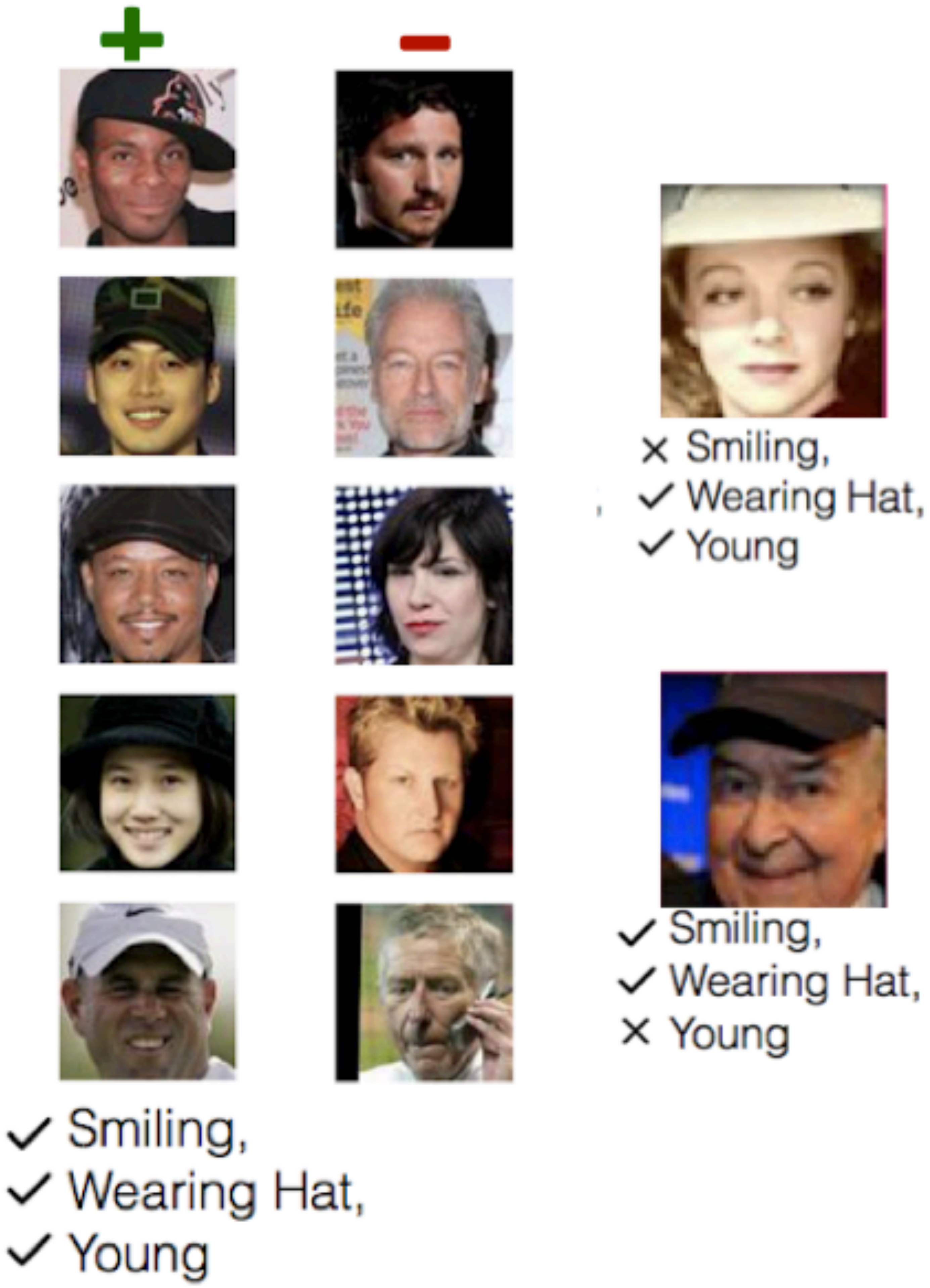
CS330 Review Session: Bayesian Meta-Learning

Stanford
Artificial
Intelligence

Why Bayesian Meta-Learning?



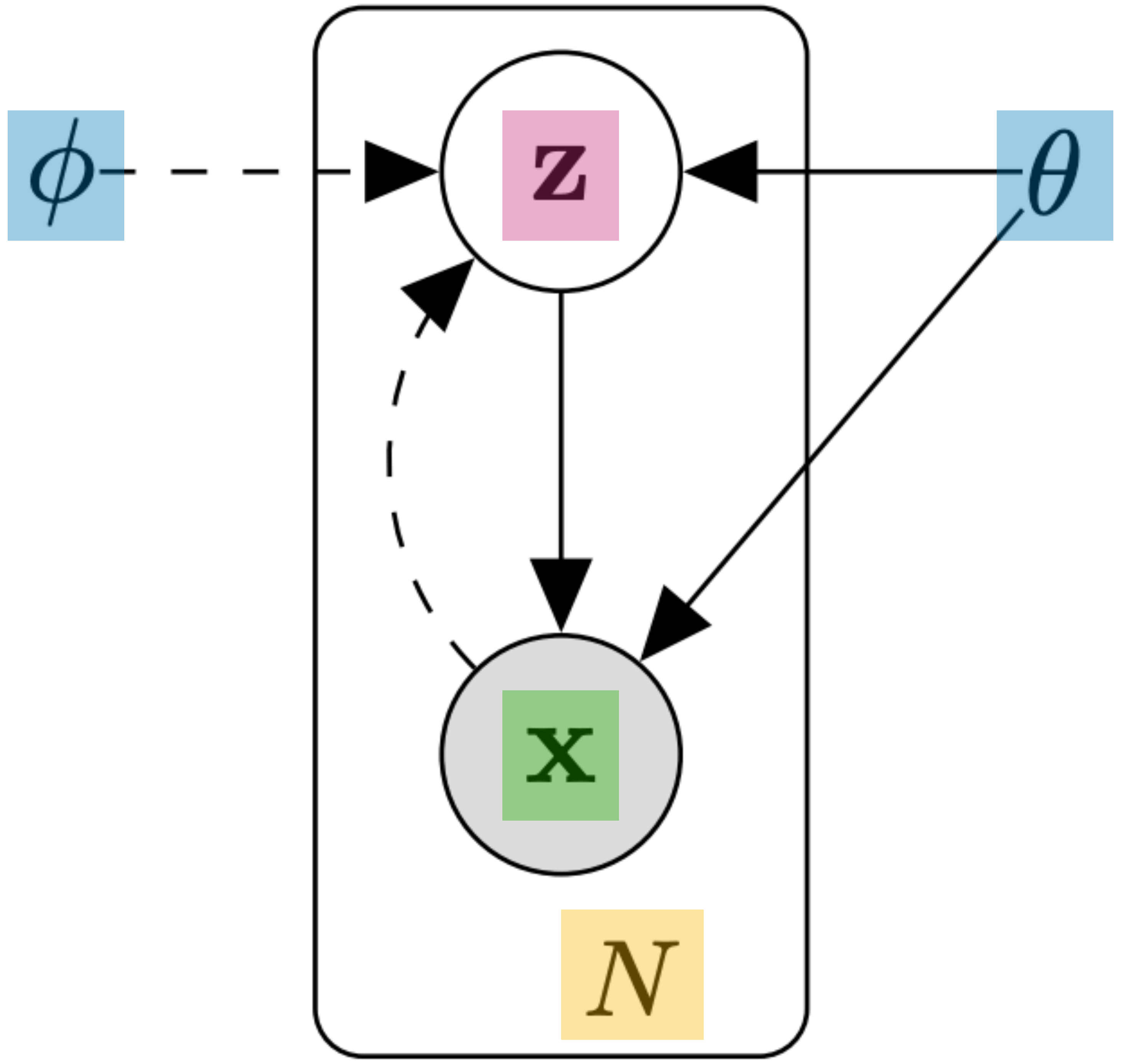
- Deterministic methods learn a point estimate (e.g. one classifier).
- Bayesian methods show multiple hypotheses. Useful for:
  - safety-critical settings
  - active learning
  - exploration

What We'll Cover Today

1. **Amortized Variational Inference**
2. ELBO Derivation for Black-Box Meta-Learning

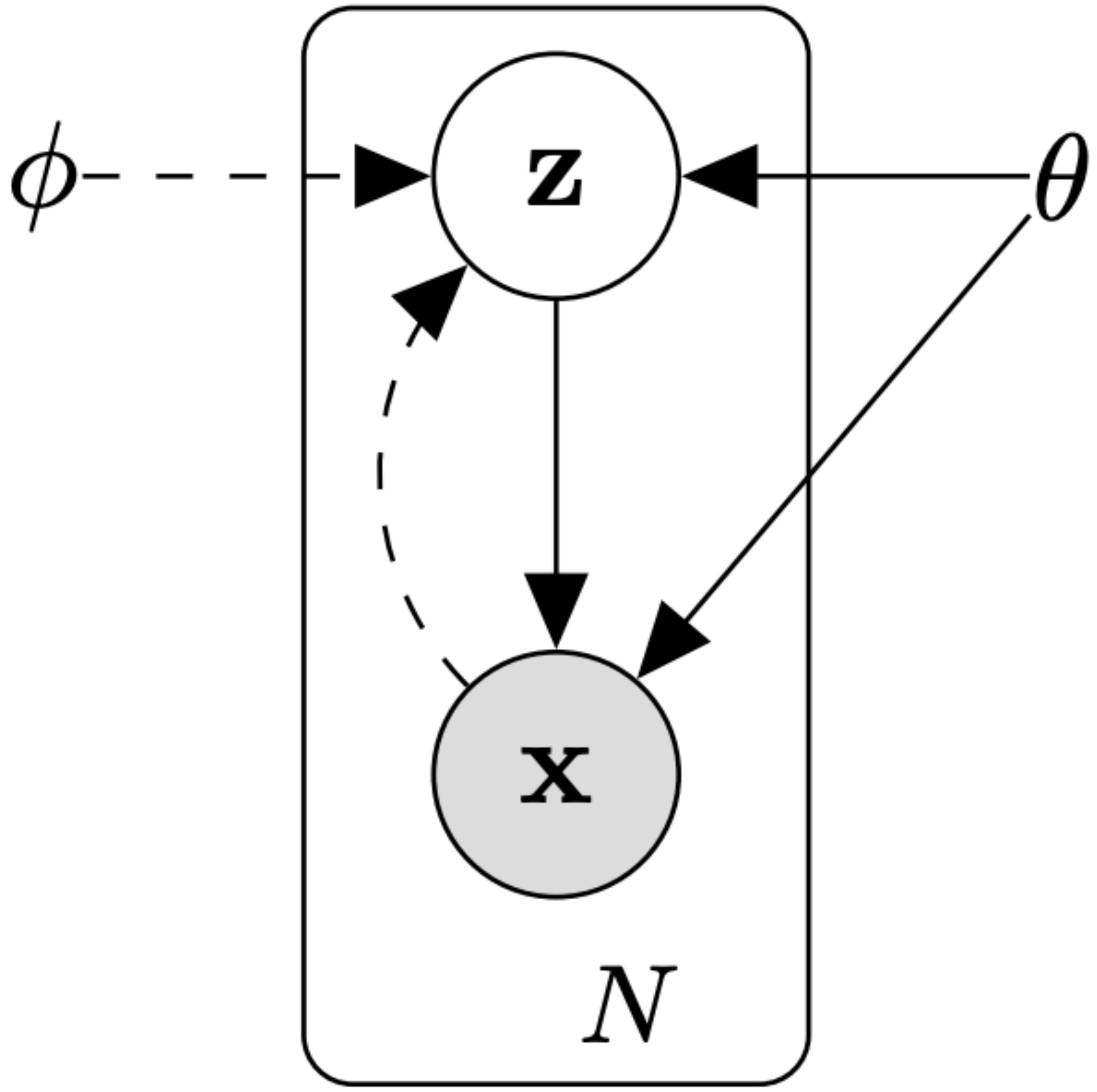
Out of scope: implemenation, MAML-based methods

How to Read Plate Notation



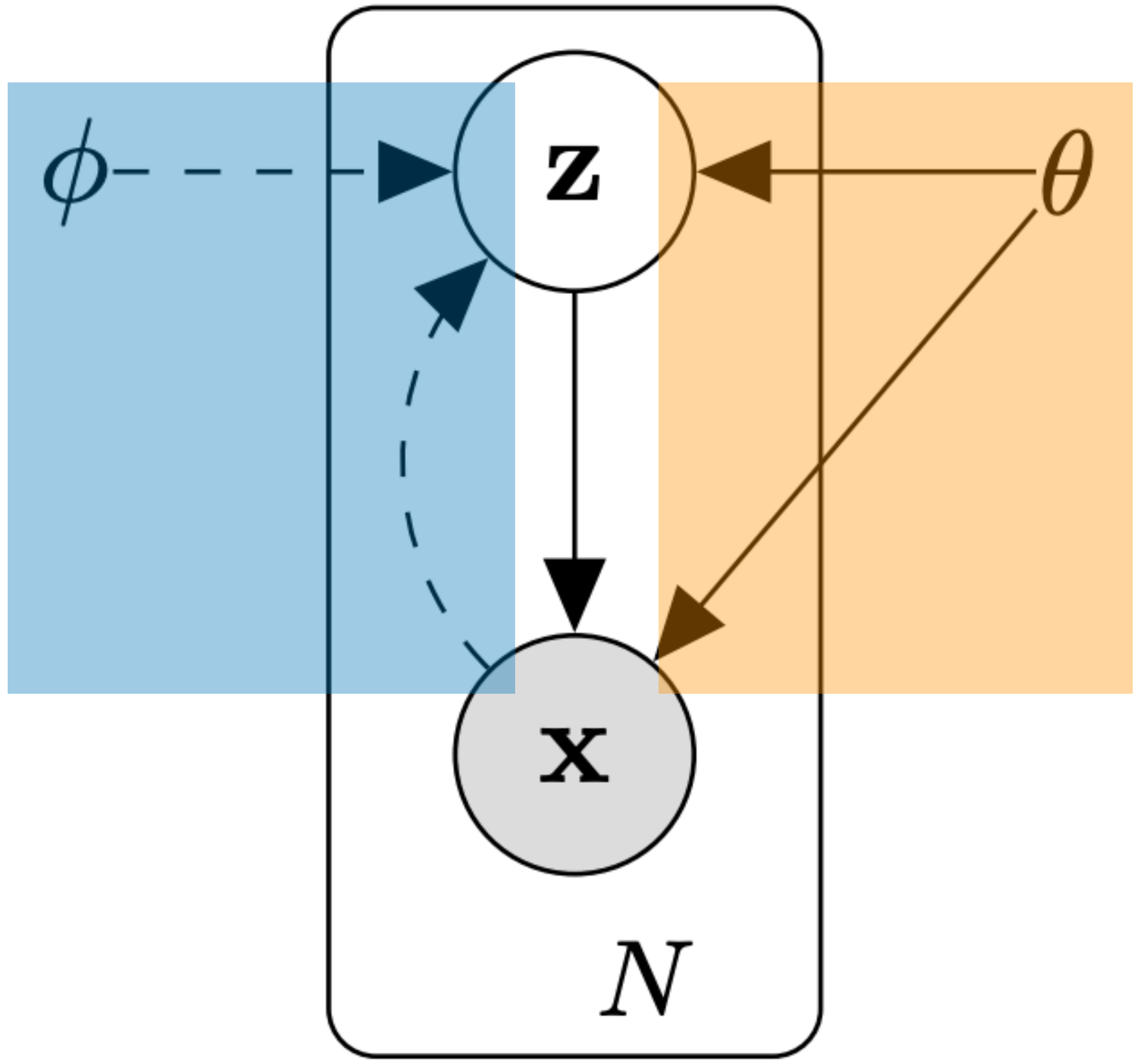
- We see N datapoints generated through the same process
- The parameters outside the plate are shared
- x is observed, z is unobserved

How to Read Plate Notation



- We see N datapoints generated through the same process
- The parameters outside the plate are shared
- x is observed, z is unobserved
- We sample x as $p_{\theta}(z)p_{\theta}(x \quad z)$ —often just $p(z)$
- Given x, we infer z as $q_{\phi}(z \quad x)$

Evidence Lower Bound (ELBO)



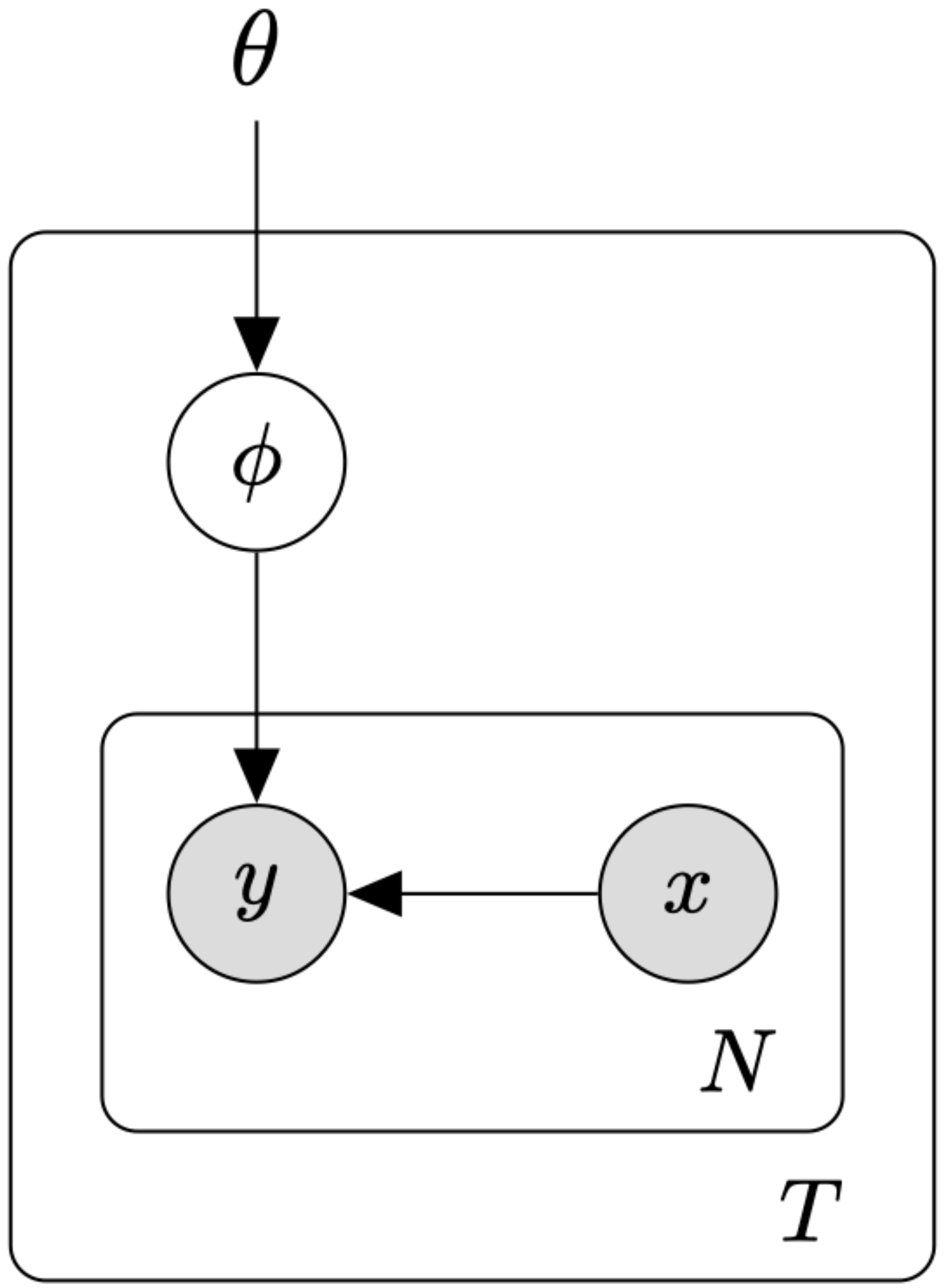
$$
\begin{aligned}
\log p(x) &\geq \mathbb{E}_{q(z|x)}\left[\log p(x,z)\right] + \mathscr{H}(q(z\,|\,x)) \\
&= \mathbb{E}_{q(z|x)}\left[\log p(x\,|\,z)\right] - D_{KL}\left(q(z\,|\,x)||p(z)\right)
\end{aligned}
$$

What We'll Cover Today

1. Amortized Variational Inference
2. **ELBO Derivation for Black-Box Meta-Learning**

Out of scope: implemenation, MAML-based methods

A Black-box Bayesian Model

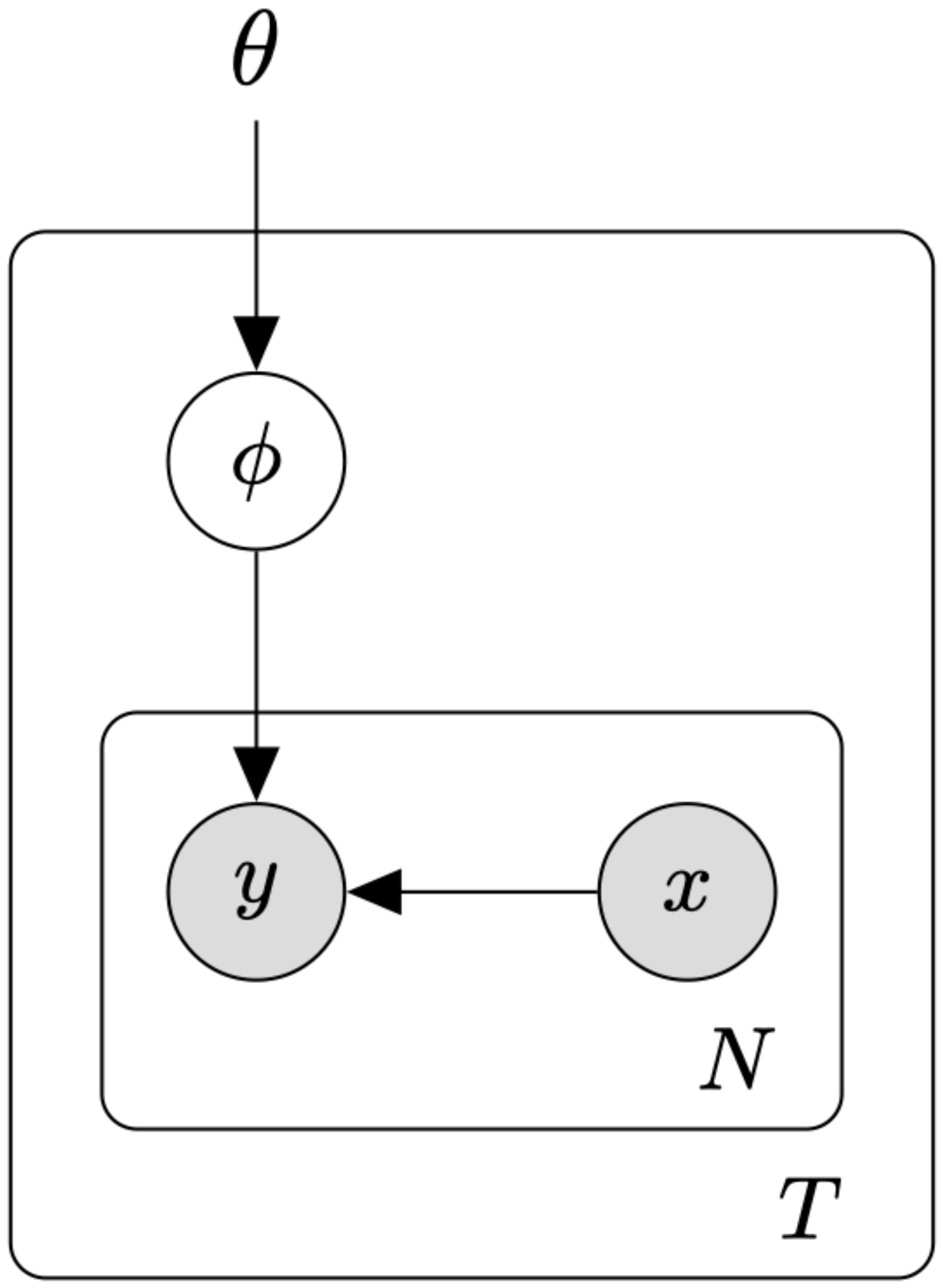


- Two plates: T tasks, N datapoints per task
- Global parameters θ model task parameters ϕ
    - ϕ = model parameters, model inputs…
- Shorthand: $X = (x_1, \ldots, x_N), Y = (y_1, \ldots, y_N)$
- For task i, labels predicted as $p(y \quad x, \phi_i)$.

To make predictions on a new task:

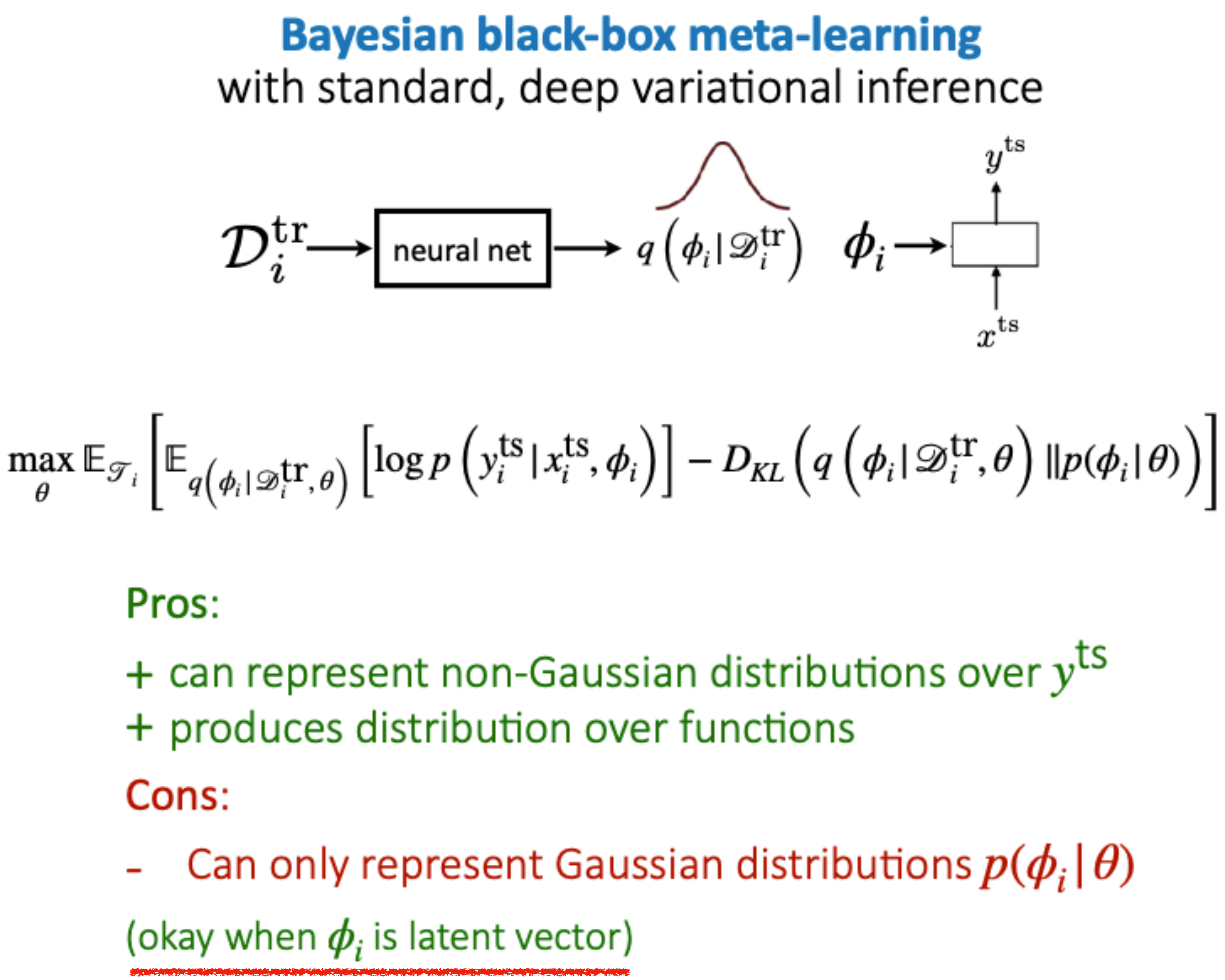
(1) infer $q(\phi \quad X, Y)$ (2) predict $p(y \quad x^{new}, \phi)$

Evidence Lower Bound (ELBO)



(whiteboard)

Parameterization


neural net
$\mathcal{D}_i^{\mathrm{tr}}$
$q\left(\phi_i|\mathcal{D}_i^{\mathrm{tr}}\right)$
ϕ_i
y^{ts}
x^{ts}

Bayesian black-box meta-learning

with standard, deep variational inference

$$\max_\theta \mathbb{E}_{\mathcal{T}_i}\left[\mathbb{E}_{q\left(\phi_i|\mathcal{D}_i^{\mathrm{tr}},\theta\right)}\left[\log p\left(y_i^{\mathrm{ts}}|x_i^{\mathrm{ts}},\phi_i\right)\right] - D_{KL}\left(q\left(\phi_i|\mathcal{D}_i^{\mathrm{tr}},\theta\right)\|p(\phi_i|\theta)\right)\right]$$

Pros:

+ can represent non-Gaussian distributions over y^{ts}
+ produces distribution over functions

Cons:

- Can only represent Gaussian distributions $p(\phi_i|\theta)$

(okay when ϕ_i is latent vector)

- ϕ = network weights
  - "Hypernetwork"
  - Learned prior $p(\phi \quad \theta)$ is important
- ϕ = inputs to a network
  - Meaning of ϕ is entirely learned
  - Simple prior $p(\phi)$ suffices

What We Covered Today

1. Amortized Variational Inference
2. ELBO Derivation for Black-Box Meta-Learning

Out of scope: implemenation, MAML-based methods