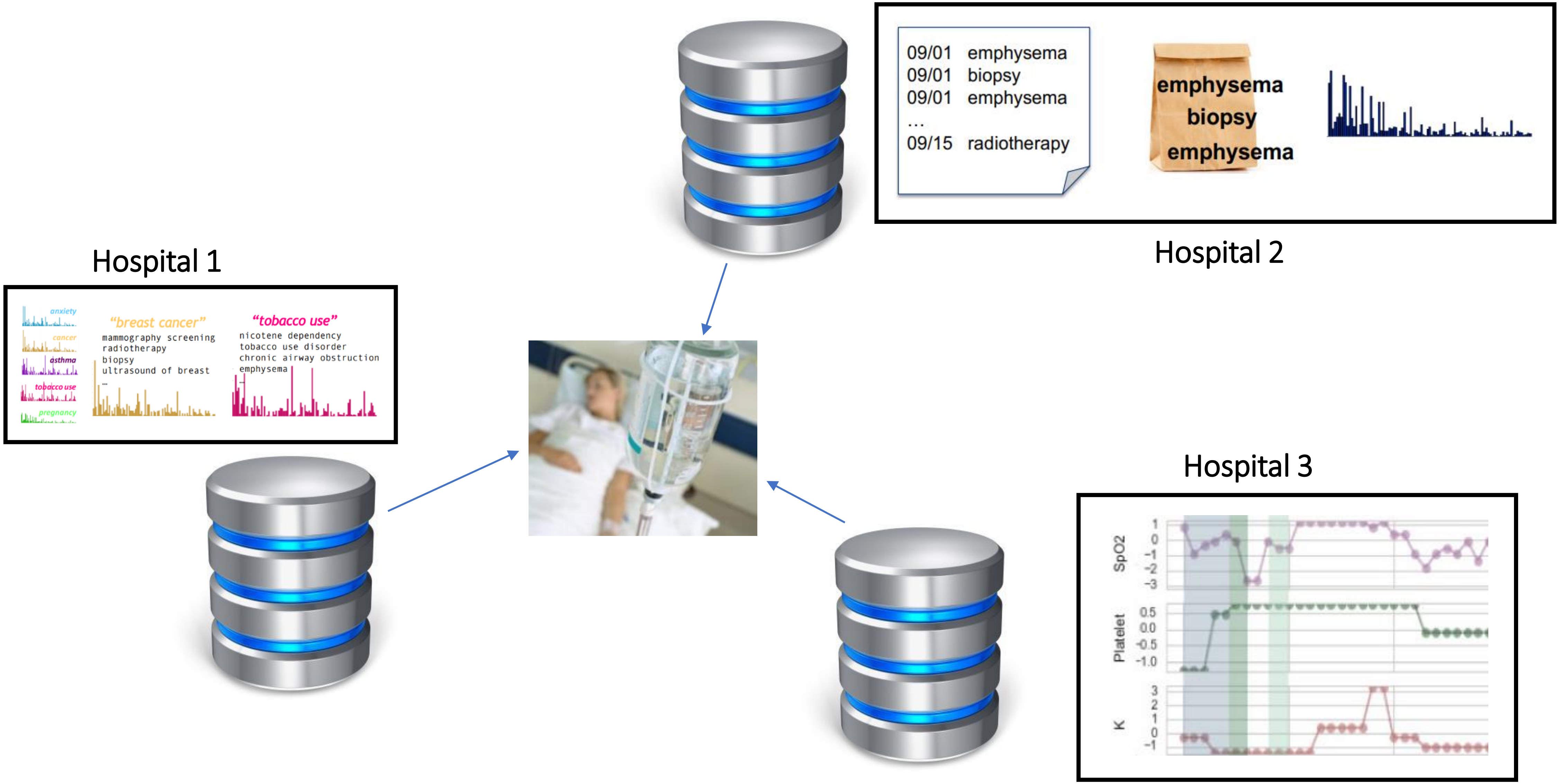
Learning Task-Agnostic Embedding of Multiple Black-Box Experts for Multi-Task Model Fusion

Nghia Hoang (MIT-IBM Watson AI Lab), **Thanh Lam** (National University of Singapore),
Bryan Kian Hsiang Low (National University of Singapore), Patrick Jaillet (MIT)

Roadmap

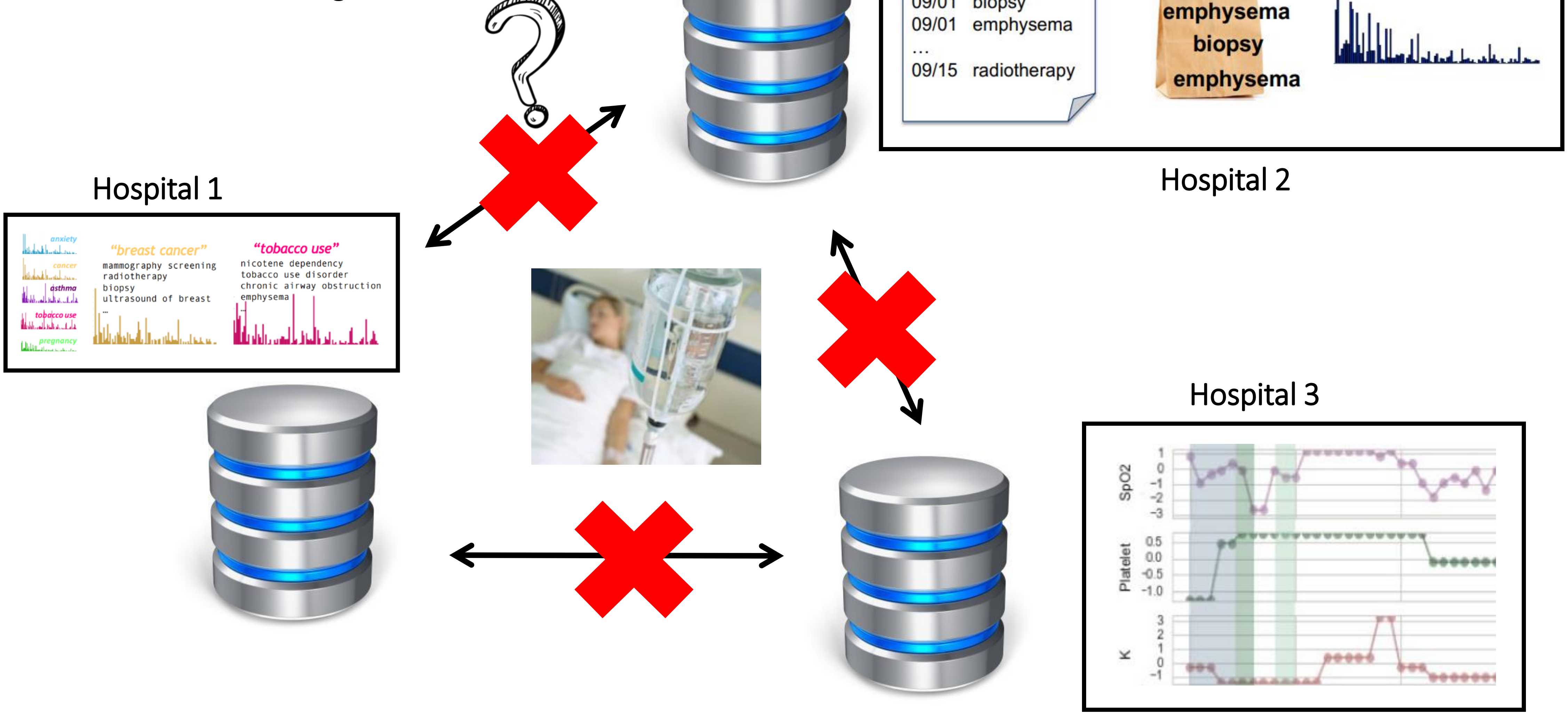
- ❑ Multi-Task Collective Learning
- ❑ Related Literature
- ❑ Model Decomposition via Task-Agnostic Embedding
- ❑ Model Fusion via PAC-Bayes Adaptation
- ❑ Empirical Results

Collective Learning: Sharing Information improves Performance

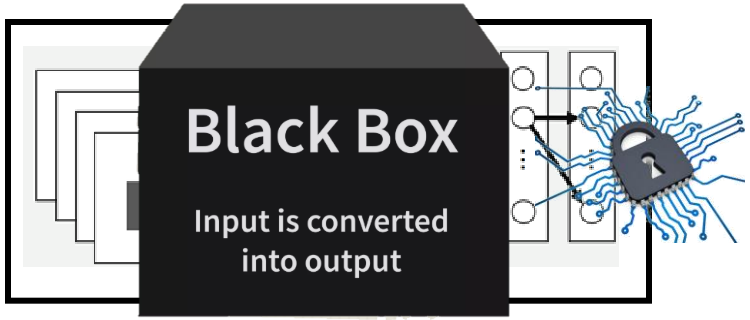


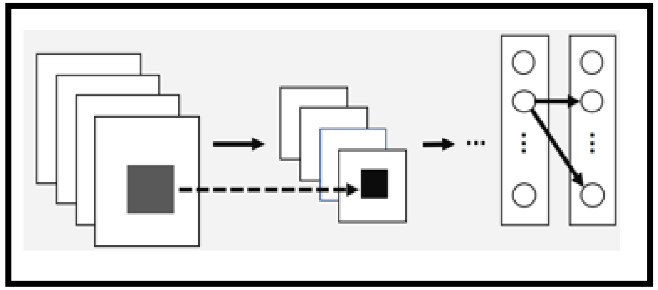
Issue: Raw information (data) is private & cannot be shared

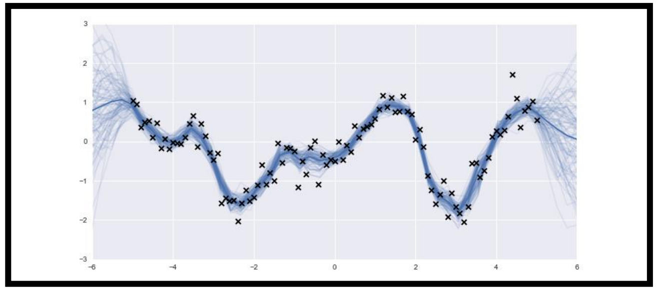
Federated Learning (McMahan, 2016) addresses this when models are homogeneous

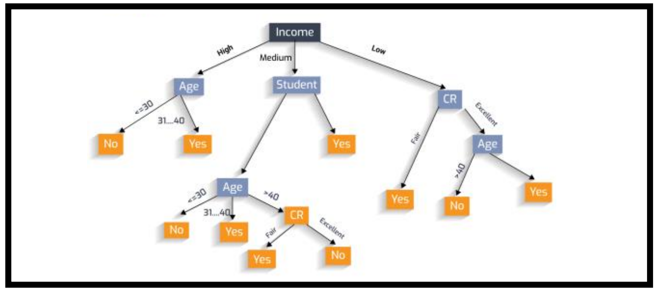


Issue: What if models are parameterized differently ?









Black-box setting happens when:

(a) Models have different parameterization / solve different tasks

(b) Models parameterization cannot be released

Our Focus ↓

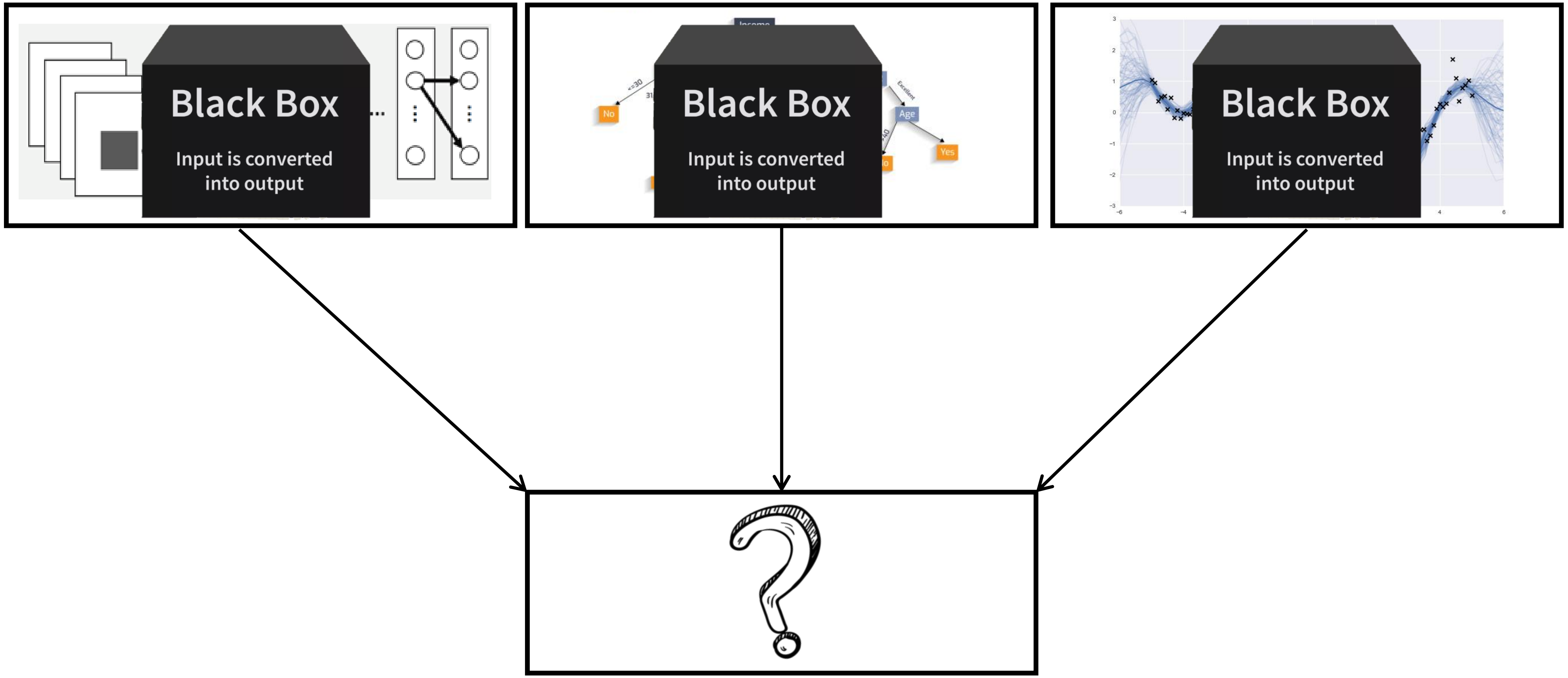
Why? (a) – to fit different on-board computation capabilities / different (related) tasks

(b) – to avoid adversarial attack (Ian Goodfellow, 2014)

Heterogeneous Models:

1. Deep Neural Network (DNN)
2. Gaussian Process (GP)
3. Decision Tree (DT)
4. Human Cognitive Reasoning etc

Idea: Model Fusion using Task-Agnostic Model Embedding



Model Fusion: Synthesizing New Model from Observing How Related Models Make Predictions (Without Accessing Local Data) – existing literature will be discussed next!

Roadmap

- ❑ Multi-Task Collective Learning
- ❑ Related Literature
- ❑ Model Decomposition via Task-Agnostic Embedding
- ❑ Model Fusion via PAC-Bayes Adaptation
- ❑ Empirical Results

MIT-IBM Watson AI Lab

Model Agnostic Meta Learning (Finn et al., 2017)

Training

Model 1: "cat-bird"

Model 2: "flower-bike"

Testing

Idea: sample tasks & learn a base model which Can be adapted to solve any task with little data

Caveat: Existing meta learning algorithm assumes **data can be centralized** for learning

Model Fusion (Hoang et al., 2019)

Model Fusion (recap.): Synthesizing New Model from Observing How Related Models Make Predictions
(Without Accessing Local Data) – existing literature will be discussed next!

A new study that emerged from Federated Learning that allows a certain degree of model agnosticity:

Collective Online Learning of Gaussian Processes for Massive Multi-Agent Systems (AAAI-19)
(Hoang, Hoang, Low & How) – combine different sparse approximations of Gaussian processes

Collective Model Fusion for Multiple Black-Box Experts (ICML-19)
(Hoang, Hoang, Low & Kingsford) – assemble different black-box models into a product of expert (PoE) model

Bayesian Non-parametric Federated Learning of Neural Networks (ICML-19)
(Yurochkin, Agrawal, Ghosh, Greenewald, Hoang & Khazaeni) – combine neural networks with different no. of hidden units

Statistical Model Aggregation via Parameter Matching (NeurIPS-19)
(Yurochkin, Agrawal, Ghosh, Greenewald & Hoang) – generalize the above to a wider class of model (including GP & DNN)

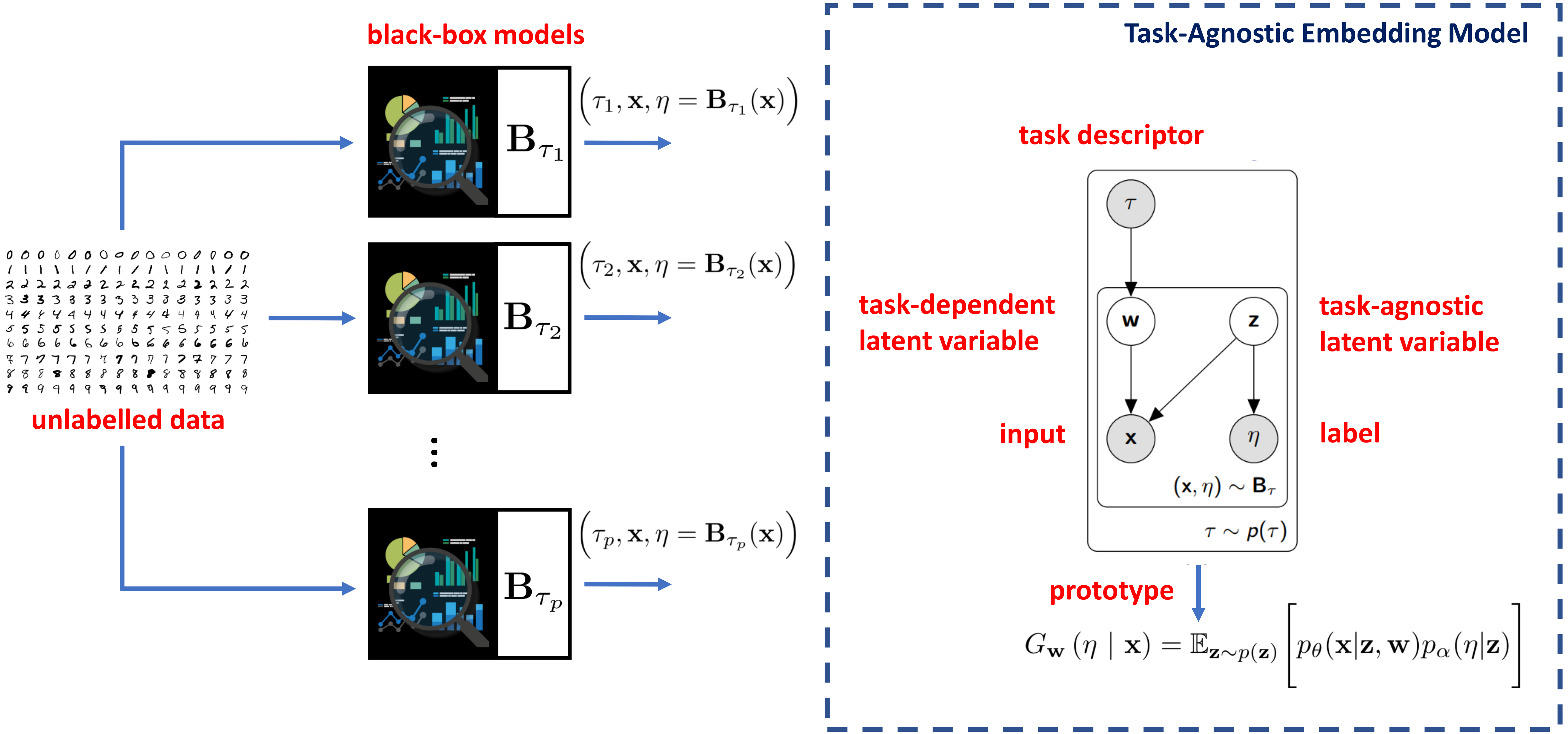
Learning Task-Agnostic Embedding of Multiple Black-Box Experts for Multi-Task Model Fusion (ICML-20)
(Hoang, Lam, Low & Jaillet) ← **TODAY's FOCUS: A new perspective of model fusion for multi-task setting**

Roadmap

- ❑ Multi-Task Collective Learning
- ❑ Related Literature
- ❑ Model Decomposition via Task-Agnostic Embedding
- ❑ Model Fusion via PAC-Bayes Adaptation
- ❑ Empirical Results

MIT-IBM Watson AI Lab

Task-Agnostic Embedding Model



Learning Task-Agnostic Embedding (without labeled data)

Generative Network Parameterization:

$$p(\mathbf{w}, \mathbf{z}, \mathbf{x}, \eta, \tau; \theta, \gamma, \alpha) \triangleq p_\theta(\mathbf{x}|\mathbf{w}, \mathbf{z}) p_\gamma(\mathbf{w}|\tau) p_\alpha(\eta|\mathbf{z}) \boxed{p(\tau) p(\mathbf{z})} .$$

learnable parameters

Latent prior: encode domain knowledge ☺

- **Example: MNIST**
  - [1, 1, 0, 0, 0, 0, 0, 0, 0, 1] – 0/1/9 classifier
  - **w** – strokes weights, orientations, …
  - **z** – numeric value

Task-Agnostic Embedding Model

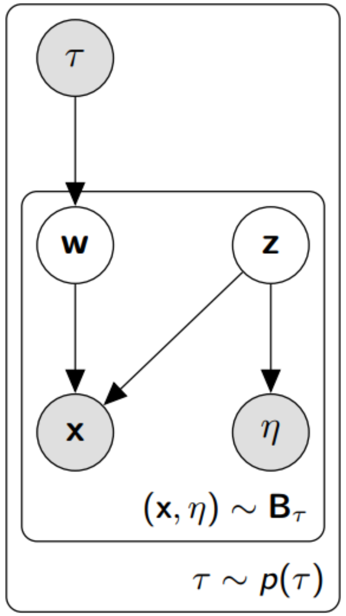
task descriptor

task-dependent latent variable

task-agnostic latent variable

input

label



prototype

$$G_{\mathbf{w}}(\eta \mid \mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \left[p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{w}) p_\alpha(\eta|\mathbf{z}) \right]$$

Learning Task-Agnostic Embedding (without labeled data)

Generative Network Parameterization:

$$p(\mathbf{w}, \mathbf{z}, \mathbf{x}, \eta, \tau; \theta, \gamma, \alpha) \triangleq p_\theta(\mathbf{x}|\mathbf{w}, \mathbf{z}) p_\gamma(\mathbf{w}|\tau) p_\alpha(\eta|\mathbf{z}) \boxed{p(\tau) p(\mathbf{z})} .$$

learnable parameters

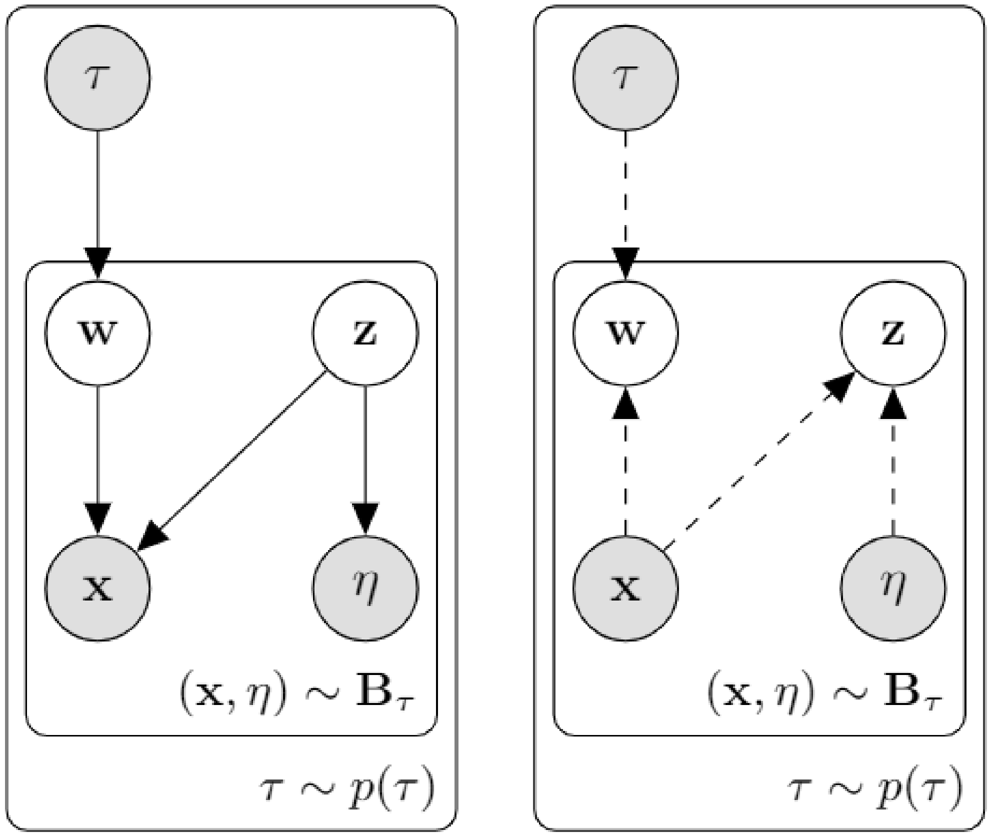
Latent prior: encode domain knowledge ☺

Inference Network Parameterization:

$$q_\phi(\mathbf{w}, \mathbf{z}|\mathbf{x}, \tau, \eta) \triangleq q_\phi(\mathbf{z}|\mathbf{x}, \eta) \, q_\phi(\mathbf{w}|\mathbf{x}, \tau)$$

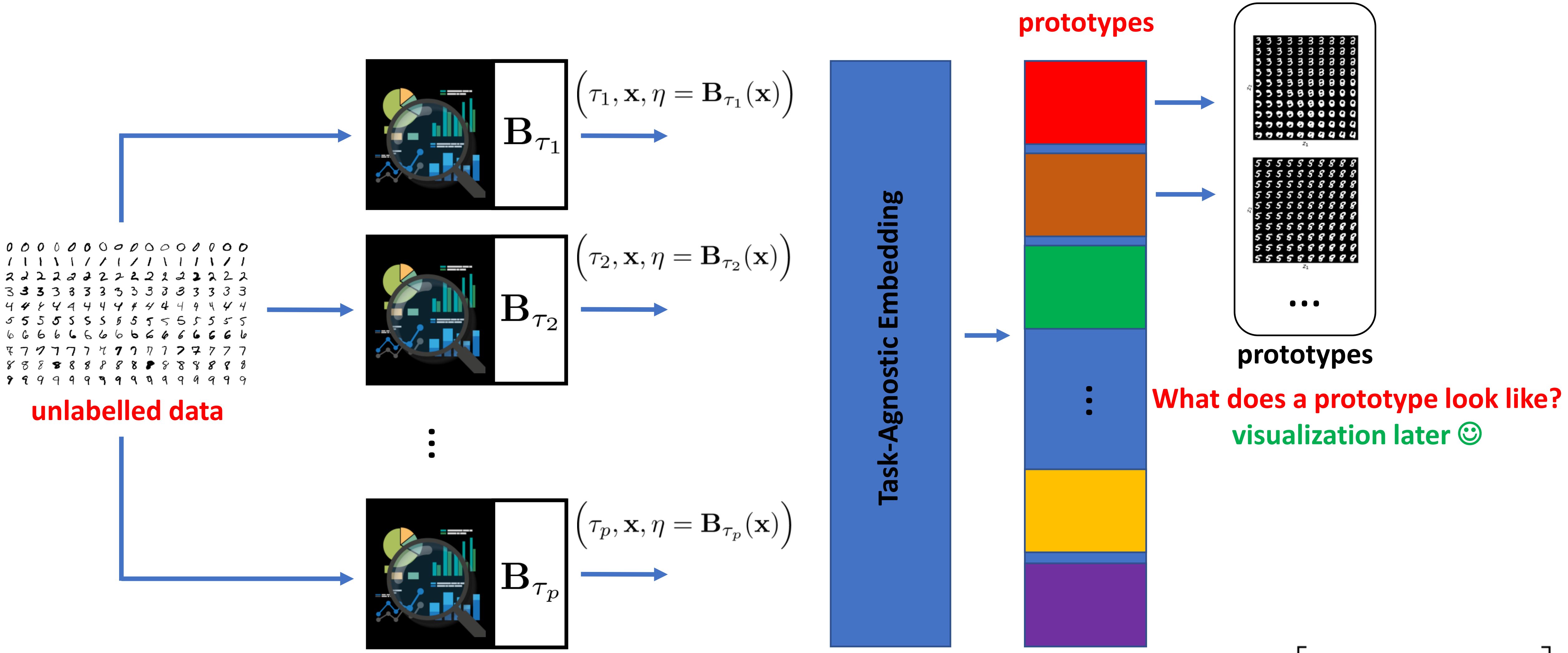
learnable parameters

Parameters can be learned end-to-end via optimizing the model evidence's lower-bound (Kingma et al., 2014) ☺



Generative Network   Inference Network

Task-Agnostic Embedding Model: From Model to Prototype ☺

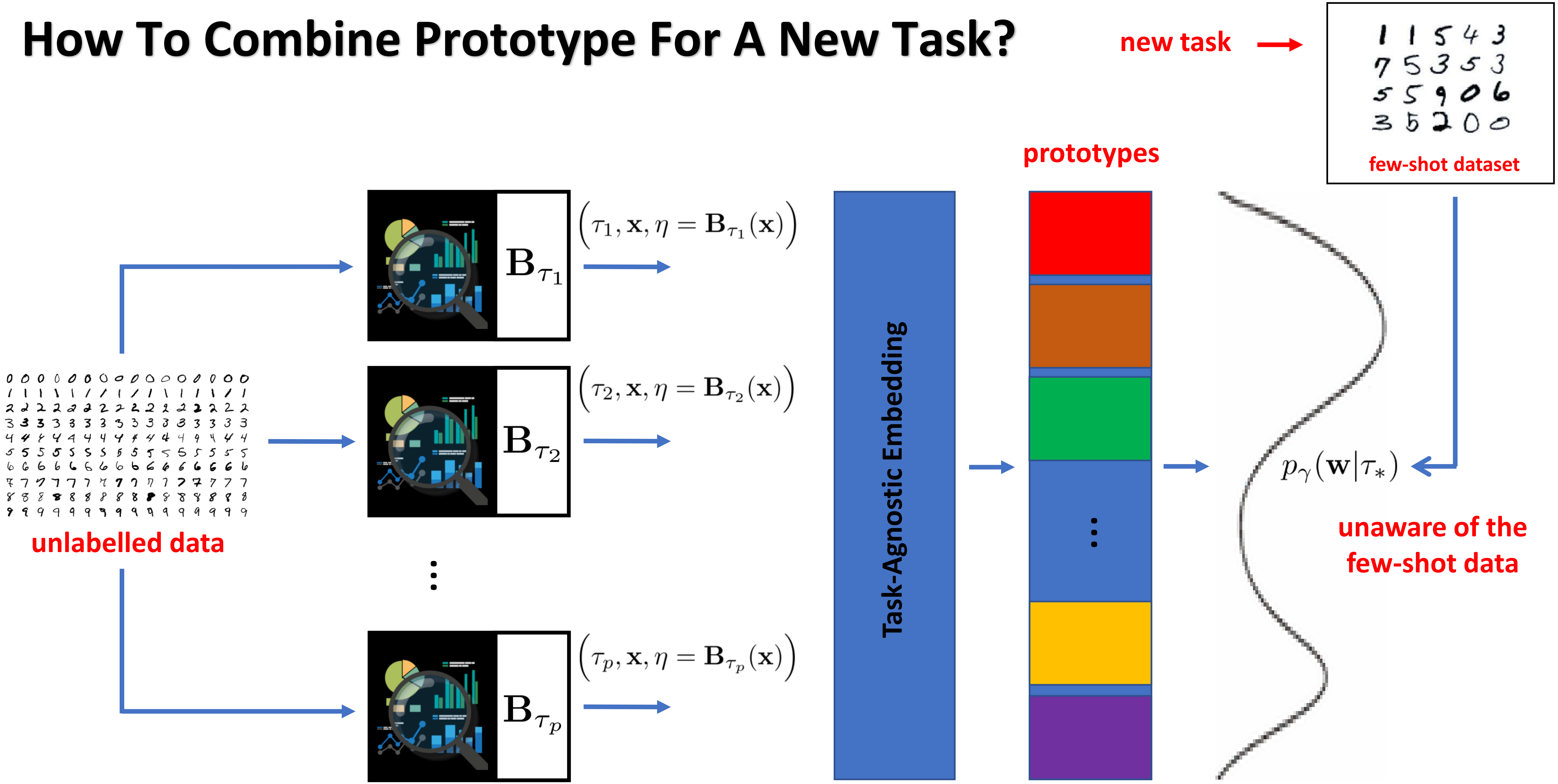


$$G_{\mathbf{w}}\left(\eta \mid \mathbf{x}\right) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}\left[p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{w}) p_\alpha(\eta|\mathbf{z})\right]$$

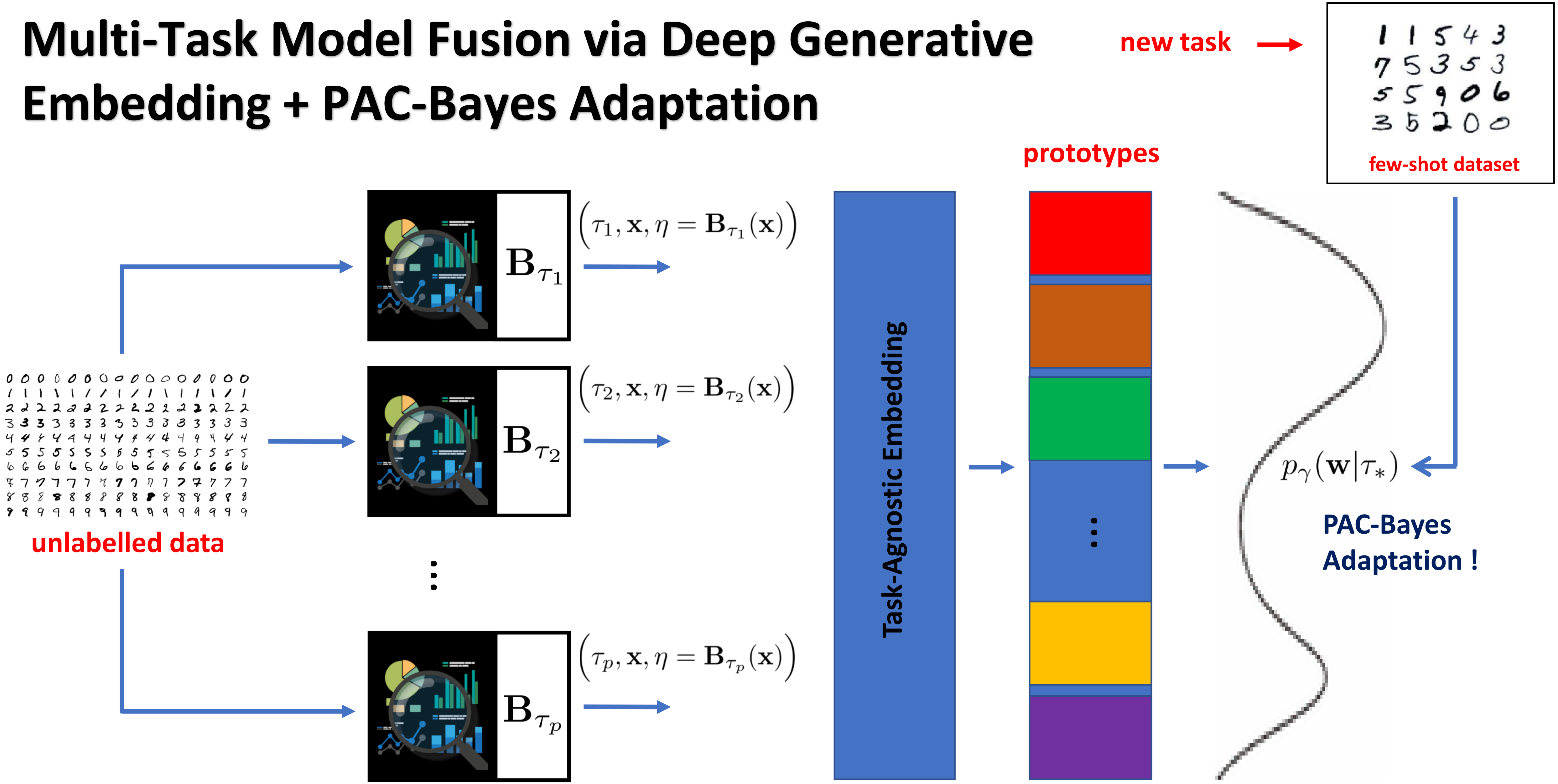
Roadmap

- ❑ **Multi-Task Collective Learning**
- ❑ **Related Literature**
- ❑ **Model Decomposition via Task-Agnostic Embedding**
- ❑ **Model Fusion via PAC-Bayes Adaptation**
- ❑ **Empirical Results**

How To Combine Prototype For A New Task?



Multi-Task Model Fusion via Deep Generative Embedding + PAC-Bayes Adaptation



Model Fusion via PAC-Bayes Adaptation

- **Goal: Optimize the prototype distribution for the new task**
- **Leverage on few-shot data**
  - minimize empirical loss on few-shot data – **may overfit ☹**
- **Add regularization term ☺**
- **Minimize PAC-Bayes Bound for Adaptation:**

posterior after adaptation

prior learnt from embedding

$$\underbrace{\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\mathbf{w}\sim q_\lambda(\mathbf{w})} g\left(G_{\mathbf{w}},(\mathbf{x}_k,y_k)\right)}_{\text{Empirical risk on the few-shot data}} + \underbrace{\sqrt{\frac{1}{2(K-1)}\left(\mathbb{KL}(q_\lambda(\mathbf{w})\|p_\gamma(\mathbf{w}|\tau_*)) + \log\frac{K}{\delta}\right)}}_{\text{Complexity term}}$$

Roadmap

- ❑ Multi-Task Collective Learning
- ❑ Related Literature
- ❑ Model Decomposition via Task-Agnostic Embedding
- ❑ Model Fusion via PAC-Bayes Adaptation
- ❑ Empirical Results

MIT-IBM Watson AI Lab

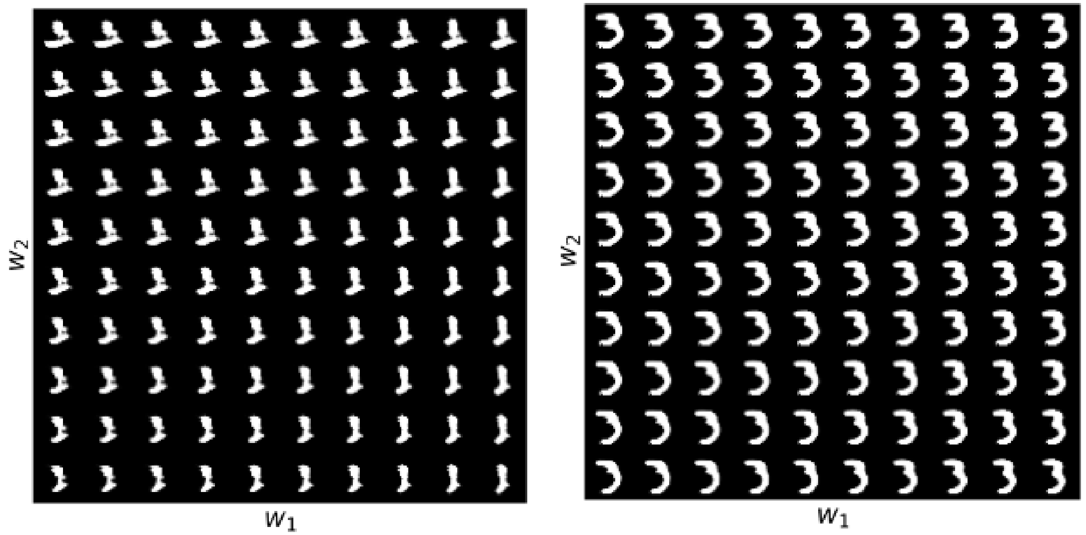
Empirical Results

observations?

❑ Task-Agnostic Decomposition

- ❑ Separate Task-Dependent and Task-Agnostic Information?

❑ Results:



Fix z:

- Same digit
- Different styles

Fix an arbitrary value of $\mathbf{z}$

Plot the $\mathbf{x}$ generated from $\mathbf{p}_\theta(\mathbf{x}|\mathbf{w},\mathbf{z})$ over the $\mathbf{w}$-space

Empirical Results

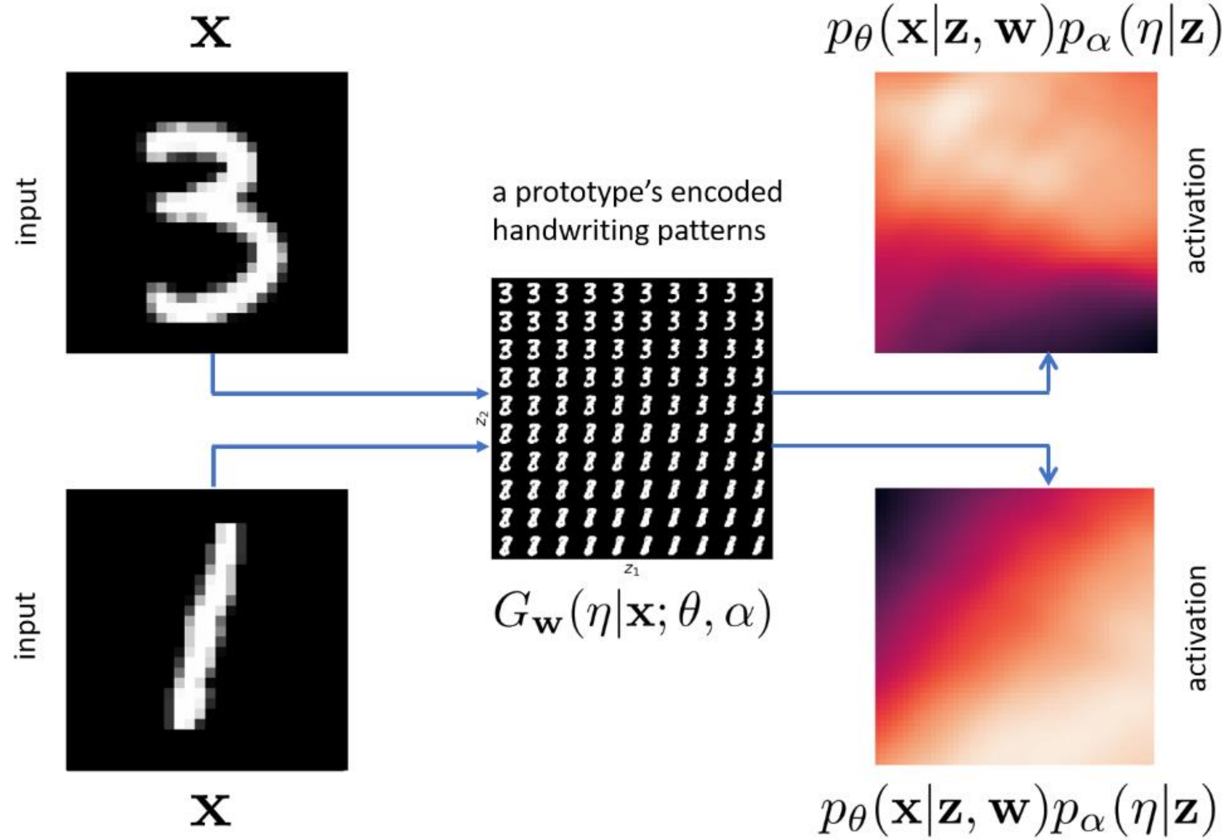
observations?

$$G_{\mathbf{w}}(\eta \mid \mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\left[p_\theta(\mathbf{x}|\mathbf{z},\mathbf{w})p_\alpha(\eta|\mathbf{z})\right]$$

❑ Prototype Visualization

- ❑ Prototypes are **task-agnostic** and will be **activated differently** depending on each input

❑ Results:



Fix an arbitrary value of $\mathbf{w}$

Plot the $\mathbf{x}$ generated from $\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{w},\mathbf{z})$ over the $\mathbf{z}$-space

Empirical Results

observations?

- ❑ **Multi-Task Model Fusion**
  - ❑ Qualitative results on standard meta-learning benchmarks
- ❑ **Comparison baseline: Modified-MAML:**
  - ❑ Data for different tasks are private
  - ❑ Original MAML requires data centralization
  - ❑ Modified-MAML only samples classes within the same task!
- ❑ **Other baselines: Ad-hoc Aggregation Methods (via + & max) & FS**
- ❑ **Dataset: MNIST, nMNIST & miniImageNet**

Empirical Results – MNIST & nMNIST (2-way) & Mini-Imagenet (5-way)

Multi-Task Model Fusion

- Qualitative results on standard meta-learning benchmarks (1-shot)

Results

number of black-boxes

N-MNIST-3-S

dataset name

S: test classes were seen
U: test classes not seen by any black-boxes

FUSION	B_+	B_{MAX}	B_{τ_*} (OURS)	B_{MAML}	FS
MNIST-2-S	$\mathbf{96.25 \pm 1.06}$	96.25 ± 1.06	$\mathbf{94.25 \pm 4.60}$	92.13 ± 1.60	80.75 ± 13.7
N-MNIST-3-S	99.02 ± 0.71	$\mathbf{99.12 \pm 0.71}$	$\mathbf{96.25 \pm 0.35}$	80.79 ± 2.06	77.11 ± 7.07
MINIIMAGENET-3-S	87.20 ± 3.75	$\mathbf{87.21 \pm 3.01}$	$\mathbf{87.10 \pm 1.02}$	41.38 ± 2.13	26.45 ± 0.55
MNIST-2-U	50.25 ± 0.35	50.75 ± 1.06	$\mathbf{78.56 \pm 2.70}$	73.92 ± 7.32	76.75 ± 5.30
N-MNIST-3-U	48.11 ± 4.95	48.25 ± 6.01	$\mathbf{94.02 \pm 1.41}$	77.25 ± 7.71	92.50 ± 0.70
MINIIMAGENET-3-U	21.80 ± 3.78	22.41 ± 2.02	$\mathbf{42.80 \pm 1.11}$	40.78 ± 2.01	26.17 ± 0.78

Take-Home Messages ☺

Thank You

- ❑ **A Model Fusion Perspective for Meta Learning in Private Data Setting**
  (a.k.a. where model fusion meets meta learning ☺)

Multi-Task Model Fusion via Deep Generative Embedding + PAC-Bayes Adaptation