# Gaussian Process Assisted Meta-learning

DATAWorks 2026

**Anna R. Flowers**[1] Christopher T. Franck[1] Robert B. Gramacy[1] Justin A. Krometis[23]

[1] Department of Statistics, Virginia Tech [2] Department of Mathematics, Virginia Tech; [3] Virginia Tech National Security Institute

April 21, 2026

# Motivation for meta-learning

- We consider data $D_N$ used to train a machine learning (ML) model.

# Motivation for meta-learning

- We consider data $D_N$ used to train a machine learning (ML) model.
- We want to choose $n$ new data points and then re-train the model, and want it to be as good as possible.

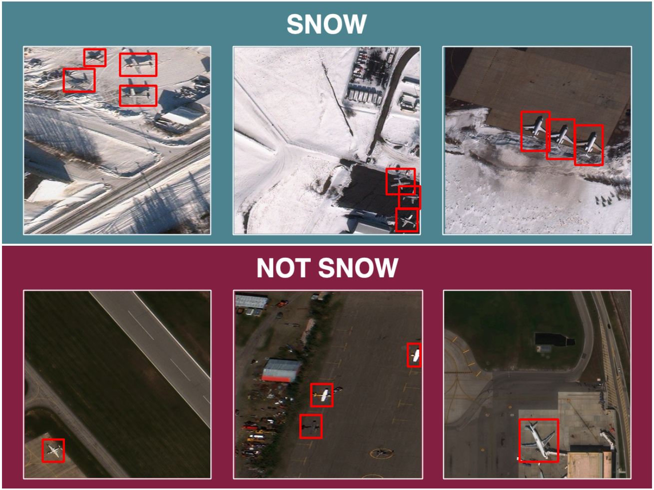
# Motivation for meta-learning

- We consider data $D_N$ used to train a machine learning (ML) model.
- We want to choose $n$ new data points and then re-train the model, and want it to be as good as possible.
- Some images may be more valuable than others based on their **metadata**, e.g., the weather in the picture: snow or not snow.

# Motivation for meta-learning

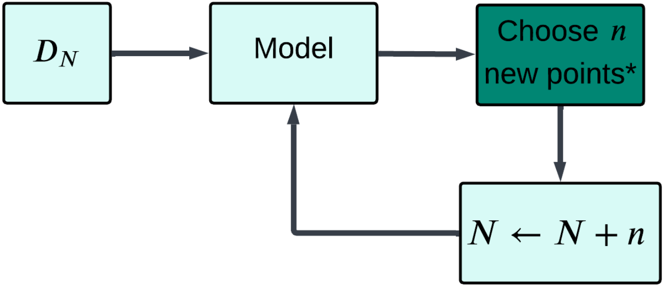
- We consider data $D_N$ used to train a machine learning (ML) model.
- We want to choose $n$ new data points and then re-train the model, and want it to be as good as possible.
- Some images may be more valuable than others based on their **metadata**, e.g., the weather in the picture: snow or not snow.
- We use **active learning** to learn the optimal metadata balance.

# Motivation for meta-learning: RarePlanes[1]




[1] Shermeyer et al. (2021)

# Active learning framework




- *How do we choose these new points?

# Gaussian Process Regression

- Let $X$ be an $r \times d$ matrix of inputs and $Y$ be a corresponding vector of outputs. A **Gaussian process** (Rasmussen and Williams 2006; Gramacy 2020) prior specifies $Y \sim \mathcal{N}_r(0, \Sigma(X))$, where

$$\Sigma(x_i, x_j) = \tau^2 \left( \exp\left( -\frac{\left\| x_i - x_j \right\|^2}{\theta} \right) + g\mathbb{I}_{\{i=j\}} \right). \tag{1}$$

# Gaussian Process Regression

- Let $X$ be an $r \times d$ matrix of inputs and $Y$ be a corresponding vector of outputs. A **Gaussian process** (Rasmussen and Williams 2006; Gramacy 2020) prior specifies $Y \sim \mathcal{N}_r(0, \Sigma(X))$, where

$$\Sigma(x_i, x_j) = \tau^2 \left( \exp\left( -\frac{\|x_i - x_j\|^2}{\theta} \right) + g\mathbb{I}_{\{i=j\}} \right). \qquad (1)$$

- For new data $\mathcal{X}$, the distribution of $\mathcal{Y}(\mathcal{X})$ conditional on $(X, Y)$ is

$$\begin{aligned} \mathcal{Y}(\mathcal{X})|X, Y \sim \mathcal{N}_r(\mu_{\mathcal{X}}, \Sigma_{\mathcal{X}}), & \\ \mu_{\mathcal{X}} = \Sigma(\mathcal{X}, X)\Sigma(X)^{-1}Y & \\ \Sigma_{\mathcal{X}} = \Sigma(\mathcal{X}) - \Sigma(\mathcal{X}, X)\Sigma(X)^{-1}\Sigma(\mathcal{X}, X^\top) & \qquad (2) \end{aligned}$$

# Idea: GP-Assisted Meta-learning (GPAML)[2]

- Our idea is to use a GP to learn about ML model performance.

# Idea: GP-Assisted Meta-learning (GPAML)[2]

- Our idea is to use a GP to learn about ML model performance.
- We view metadata categories as **buckets** from which we can sample.

# Idea: GP-Assisted Meta-learning (GPAML)[2]

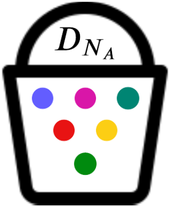
- Our idea is to use a GP to learn about ML model performance.
- We view metadata categories as **buckets** from which we can sample.
- Then, we can use the GP to model how ML model performance changes as metadata balance changes.

# Idea: GP-Assisted Meta-learning (GPAML)[2]

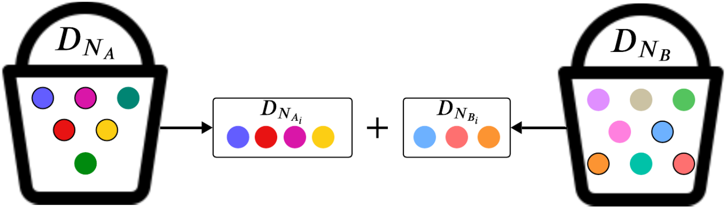
- Our idea is to use a GP to learn about ML model performance.
- We view metadata categories as **buckets** from which we can sample.
- Then, we can use the GP to model how ML model performance changes as metadata balance changes.
- We perform a data subsampling experiment to vary the metadata balance.

[2] Flowers et al. (2025)

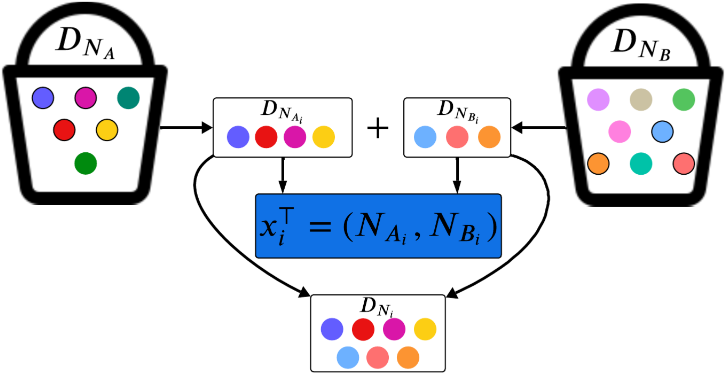
# Step 1: Learning about model performance



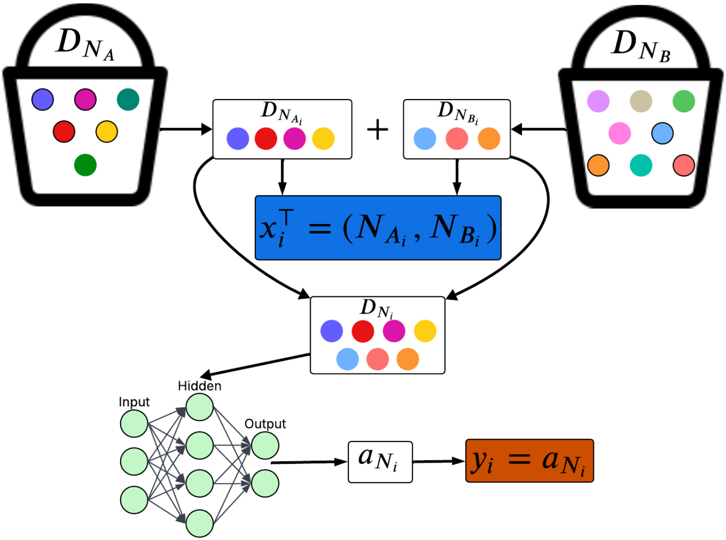

# Step 1: Learning about model performance



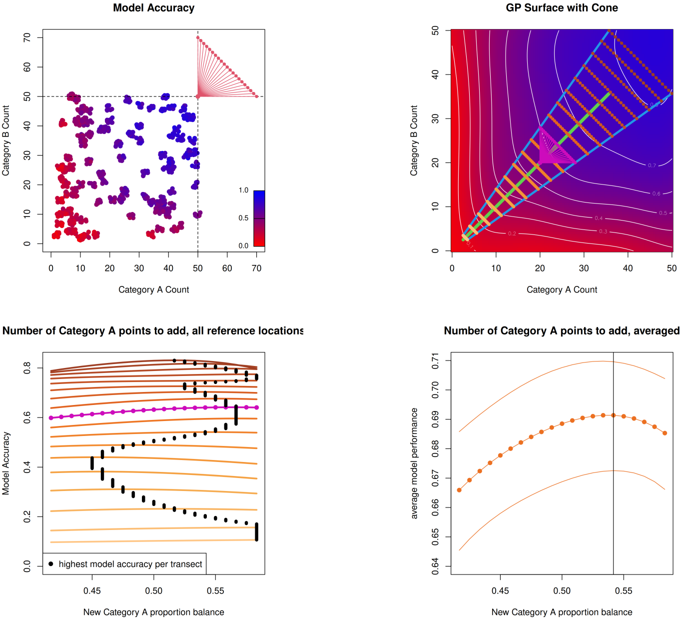
# Step 1: Learning about model performance



# Step 1: Learning about model performance

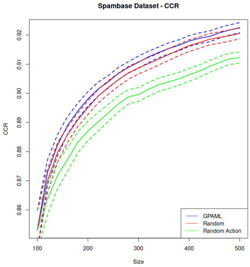


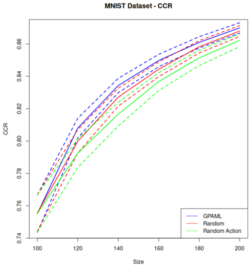
# Step 2: Using the GP to make decisions

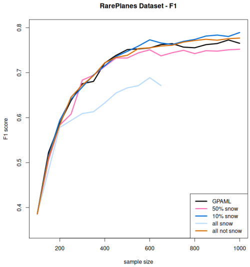


# Results

GPAML performs comparably to or better than alternative active learning methods on Spambase (Hopkins and Suermondt 1999), MNIST (LeCun 1998), and RarePlanes (Shermeyer et al. 2021) datasets.



# Thanks for listening!

- For more information, come talk to me at the poster session or check out our paper on arXiv (Flowers et al. 2025).

# References I

Flowers, Anna R, Christopher T Franck, Robert B Gramacy, and Justin A Krometis. 2025. "Gaussian Process Assisted Meta-Learning for Image Classification and Object Detection Models." *arXiv Preprint arXiv:2512.20021*.

Gramacy, Robert B. 2020. *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. CRC press.

Hopkins, Reeber, Mark, and Jaap Suermondt. 1999. "Spambase." UCI Machine Learning Repository.

LeCun, Yann. 1998. "The MNIST Database of Handwritten Digits." *Http://Yann. Lecun. Com/Exdb/Mnist/*.

Rasmussen, Carl Edward, and Christopher KI Williams. 2006. *Gaussian Processes for Machine Learning*. Vol. 2. 3. MIT press Cambridge, MA.

# References II

Shermeyer, Jacob, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. 2021. "RarePlanes: Synthetic Data Takes Flight." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 207–17.