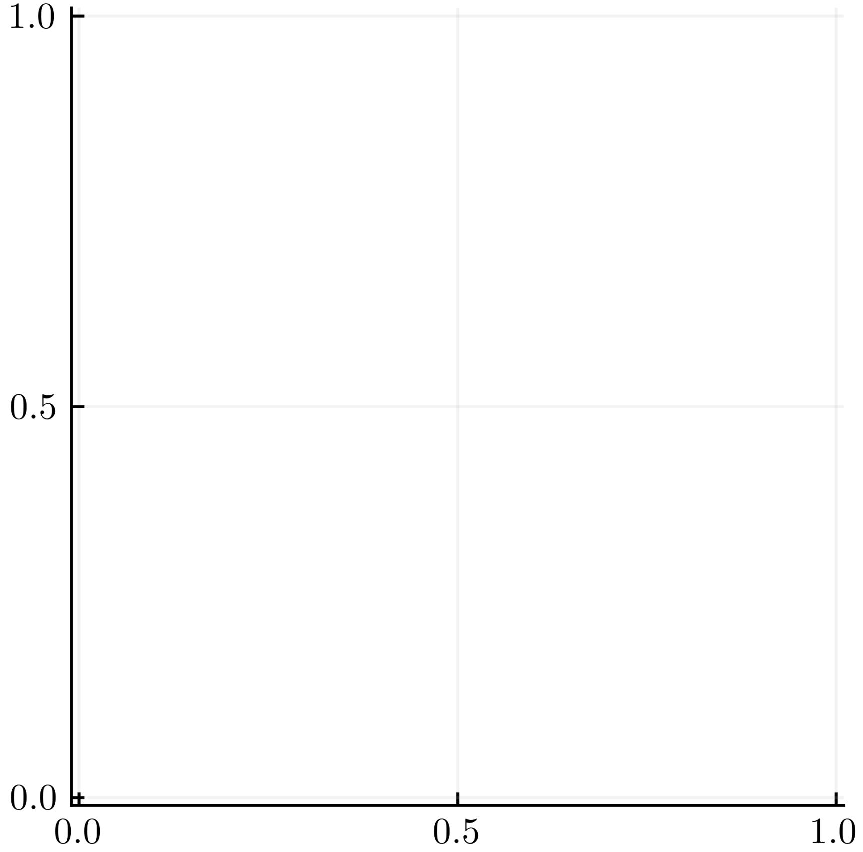
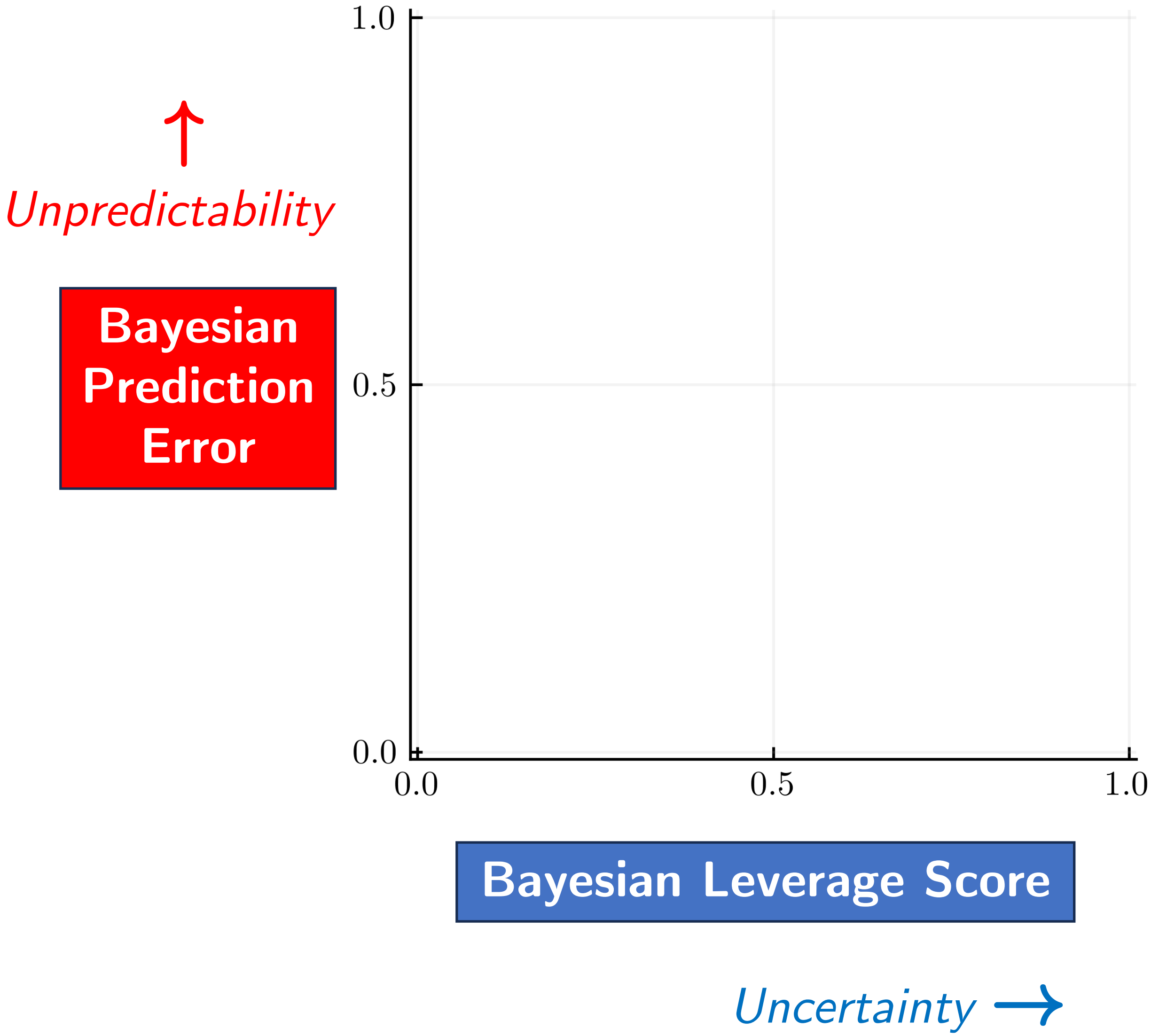
Memory Maps to Understand Models

Dharmesh Tailor

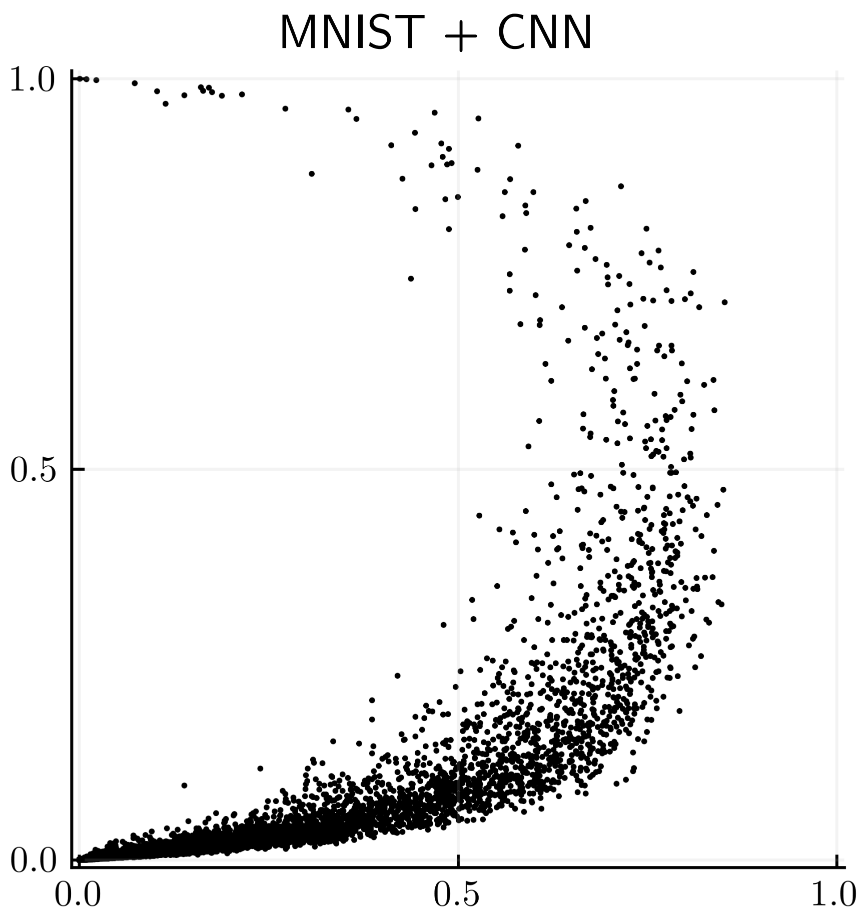
A visual representation for generic models



A visual representation for generic models



A visual representation for generic models



↑

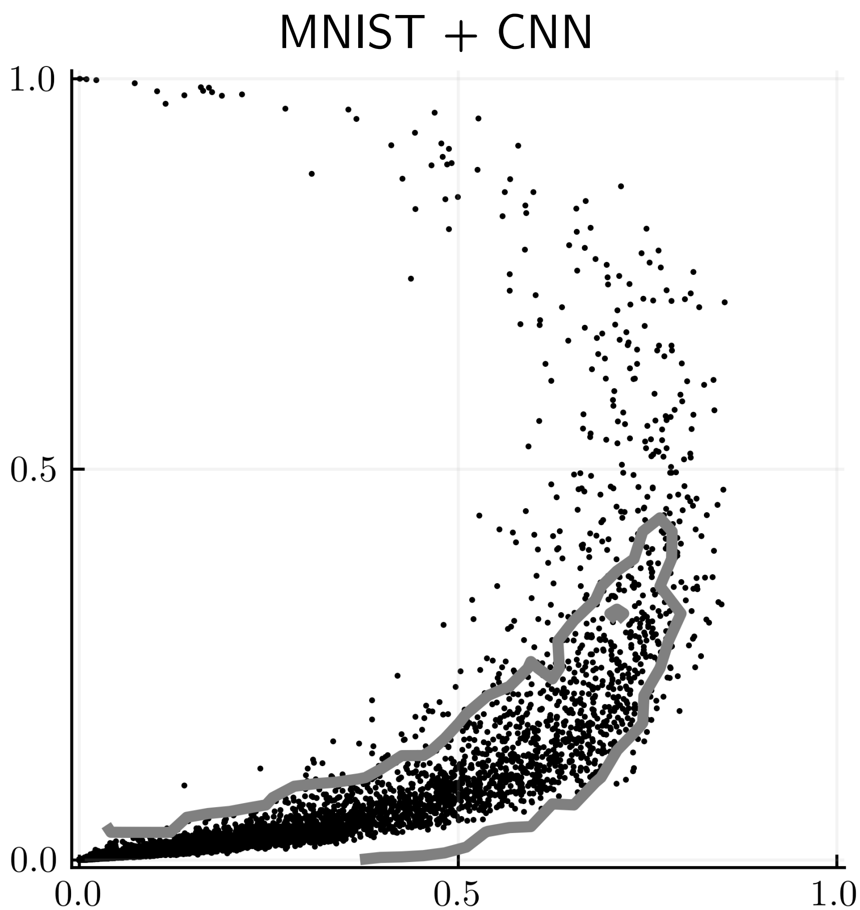
Unpredictability

Bayesian Prediction Error

Bayesian Leverage Score

Uncertainty →

A visual representation for generic models



Unpredictability

Bayesian Prediction Error

Bayesian Leverage Score

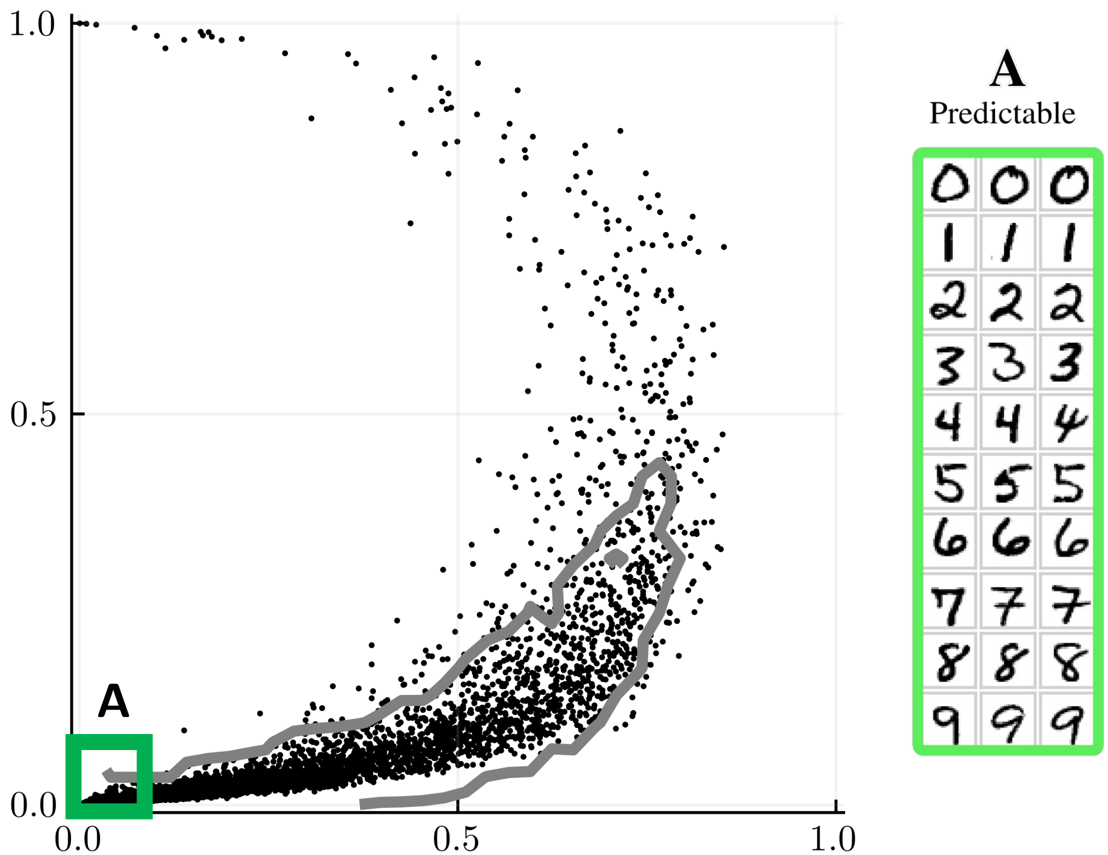
Uncertainty →

A visual representation for generic models

MNIST + CNN

Unpredictability ↑

Bayesian Prediction Error



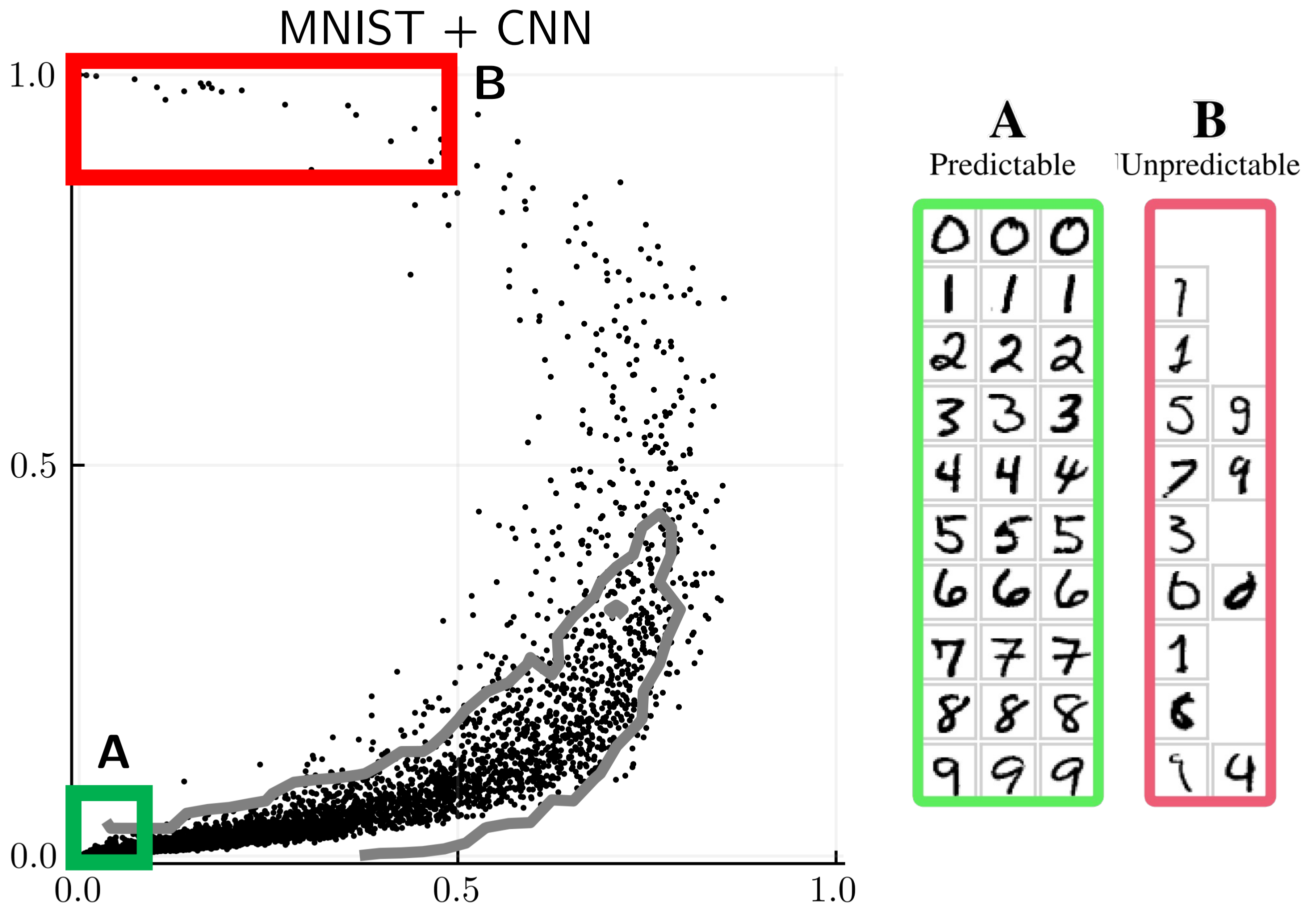
Bayesian Leverage Score

Uncertainty →

A visual representation for generic models

Unpredictability ↑

Bayesian Prediction Error



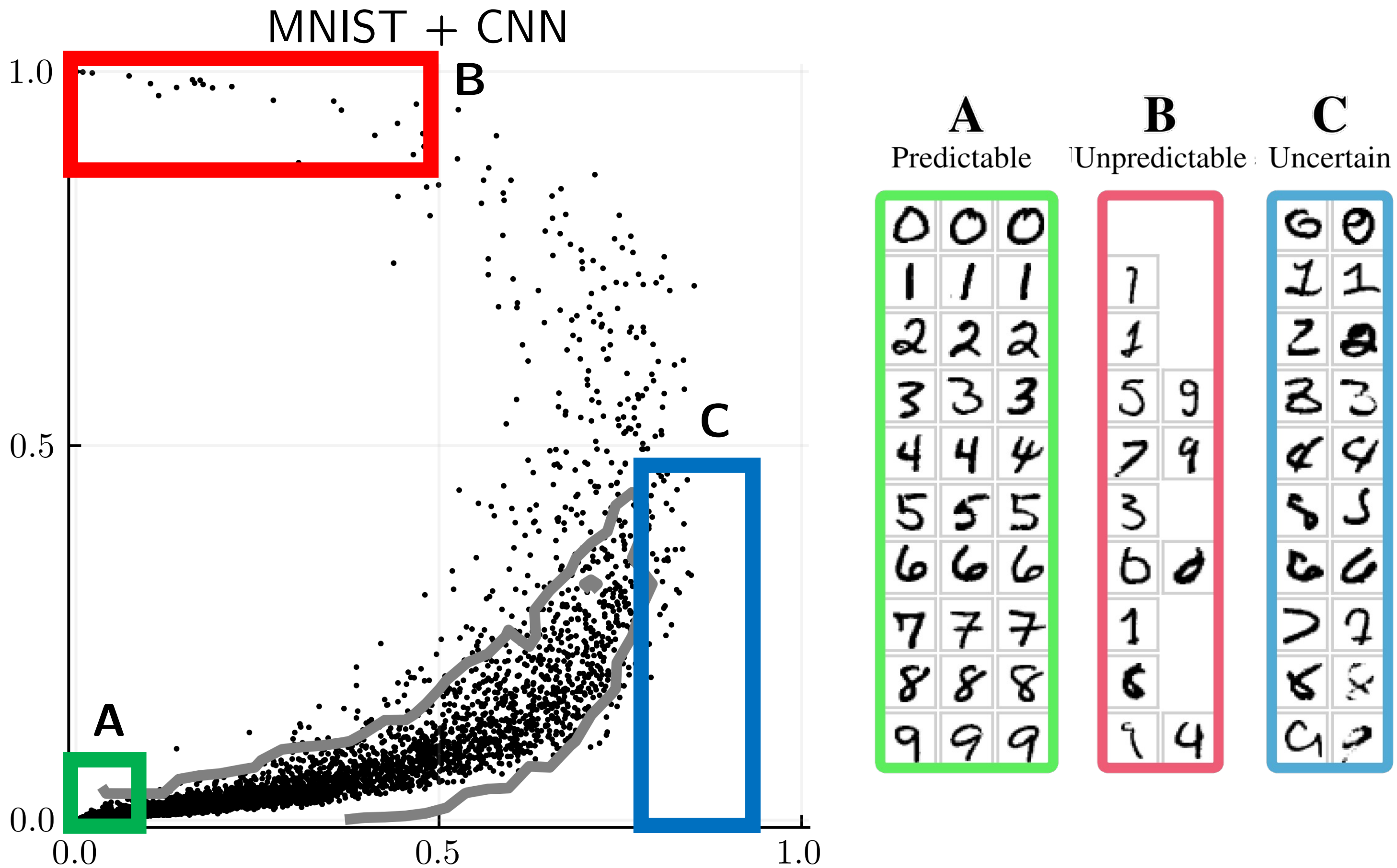
Bayesian Leverage Score

Uncertainty →

A visual representation for generic models

Unpredictability ↑

Bayesian Prediction Error

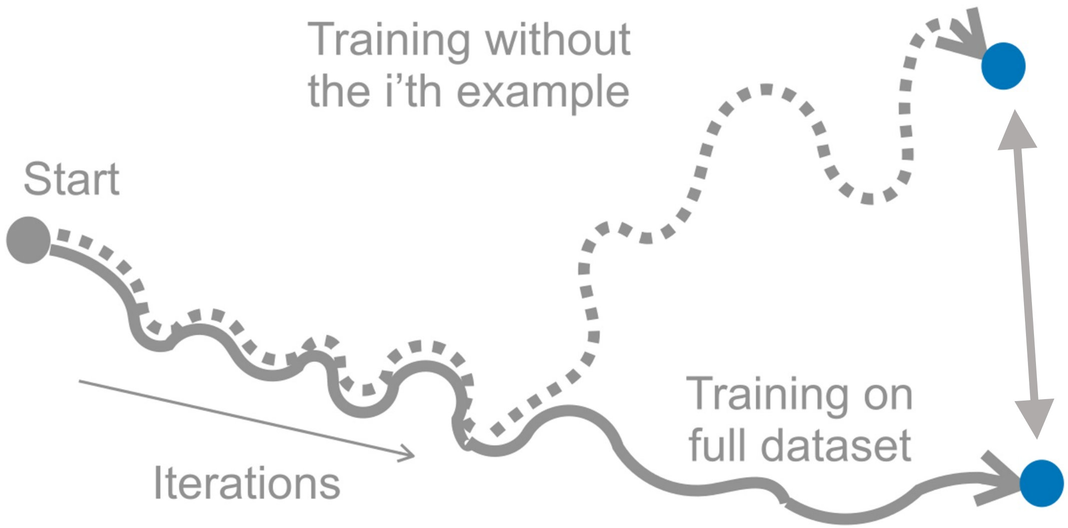


Bayesian Leverage Score

Uncertainty ⟶

Why ***memory*** maps?

Our visual representation is derived from the **sensitivity** of the model to its training data



prediction error

*

variance

Examples with high sensitivity characterize the model's *memory* since the model changes a lot if these examples are removed or perturbed heavily

Nickl, P., Xu, L., Tailor, D., Möllenhoff, T., & Khan, M. E. (2023). The Memory Perturbation Equation: Understanding Model's Sensitivity to Data. NeurIPS, 2023.

Influence in linear regression

R Dennis Cook (& others) in 1970s

$$\boldsymbol{\theta}_*^{\setminus i} - \boldsymbol{\theta}_* = \mathbf{H}_*^{-1}\mathbf{x}_i \frac{e_i}{1 - v_i}$$

$$f_i(\boldsymbol{\theta}_*^{\setminus i}) - f_i(\boldsymbol{\theta}_*) = e_i \cdot \frac{v_i}{1 - v_i}$$

Prediction error (residual)

$$e_i = \mathbf{x}_i^\top \boldsymbol{\theta}_* - y_i$$

Prediction variance (leverage)

$$v_i = \mathbf{x}_i^\top \mathbf{H}_*^{-1} \mathbf{x}_i$$

Diagnostic tool for models: 2D scatter plot of residual-leverage

R Dennis Cook. Detection of Influential Observation in Linear Regression. Technometrics, 1977.

Extension to generic models

Bayesian Learning Rule unifies many popular learning algorithms (*e.g.* SGD, Newton's method, Adam) as specific instances of a **natural-gradient descent** to solve a **generalized Bayesian objective**

$$\boldsymbol{\lambda}_t \leftarrow (1-\rho)\boldsymbol{\lambda}_{t-1} - \rho \sum_{j=0}^{N} \tilde{\mathbf{g}}_j(\boldsymbol{\lambda}_{t-1})$$

BLR as inference in conjugate Bayesian model

$$q_t \propto \underbrace{(q_{t-1})^{1-\rho}\,(p_0)^{\rho}}_{\text{Prior}} \prod_{j=1}^{N} \underbrace{e^{\langle -\rho \tilde{\mathbf{g}}_j(\boldsymbol{\lambda}_{t-1}), \mathbf{T}(\boldsymbol{\theta})\rangle}}_{\text{Likelihood}}$$

Mohammad Emtiyaz Khan and Håvard Rue. The Bayesian Learning Rule. JMLR, 2023.

Specialization of existing sensitivity measures

ResNet-FRN-20 + CIFAR-10

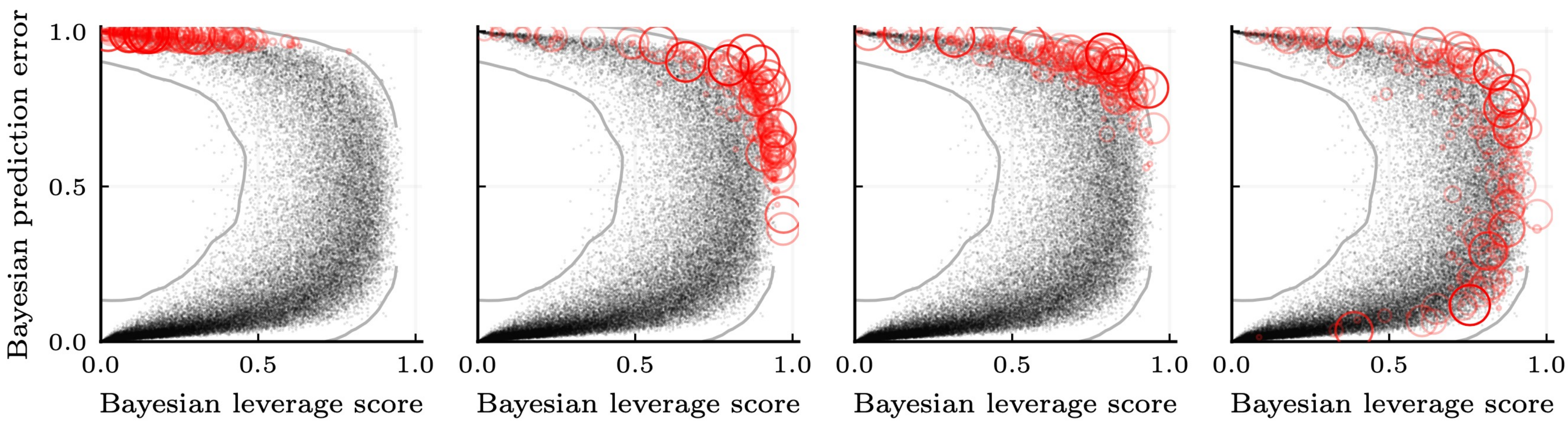
ranking per measure of top-1% examples

Influence function
[Cook & Weisberg, 1980]

Cook's distance
[Cook, 1977]

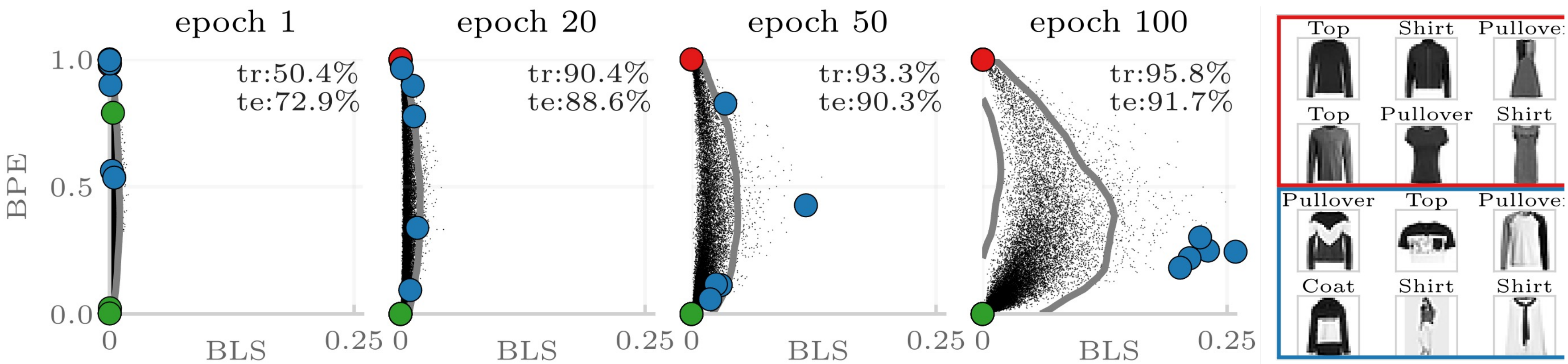
Gradient norm
[Katharopoulos & Fleuret, 2018]

Feldman's label memorization
[Feldman & Zhang, 2020]

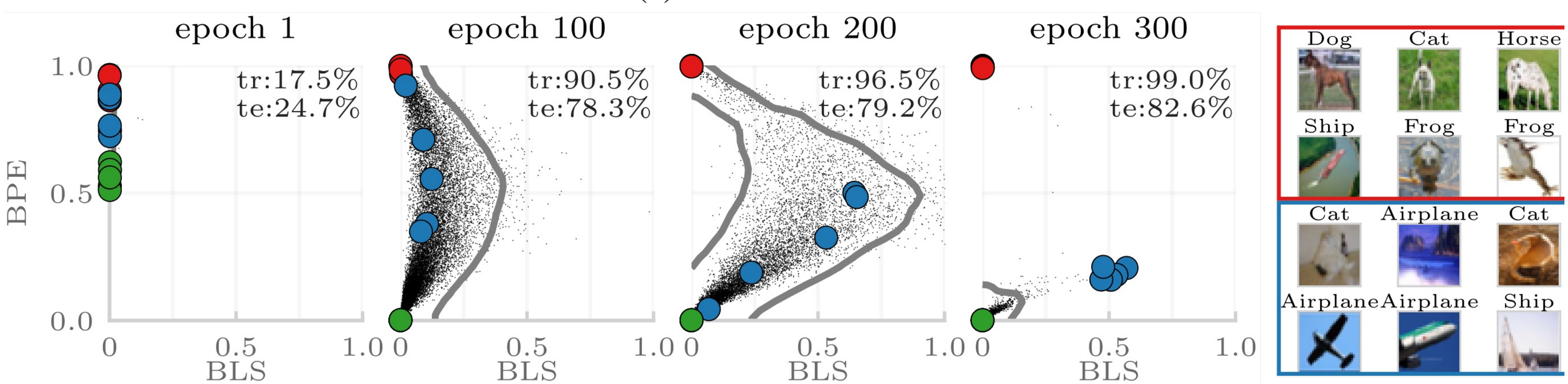


- Cook, R. D. and Weisberg, S. Characterizations of an empirical influence function for detecting influential cases in regression. Technometrics. 1980.
- Cook, R. D. Detection of influential observation in linear regression. Technometrics. 1977.
- Katharopoulos, A. and Fleuret, F. Not all samples are created equal: Deep learning with importance sampling. ICML. 2018.
- Feldman, V. and Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. NeurIPS. 2020.

Analyzing training trajectories

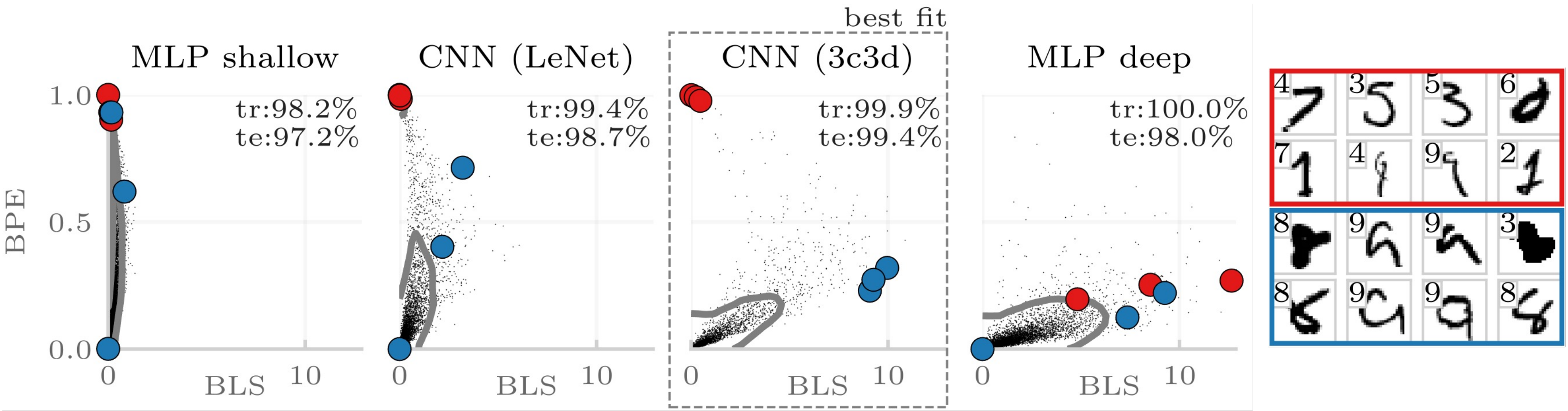


(a) FMNIST + LeNet

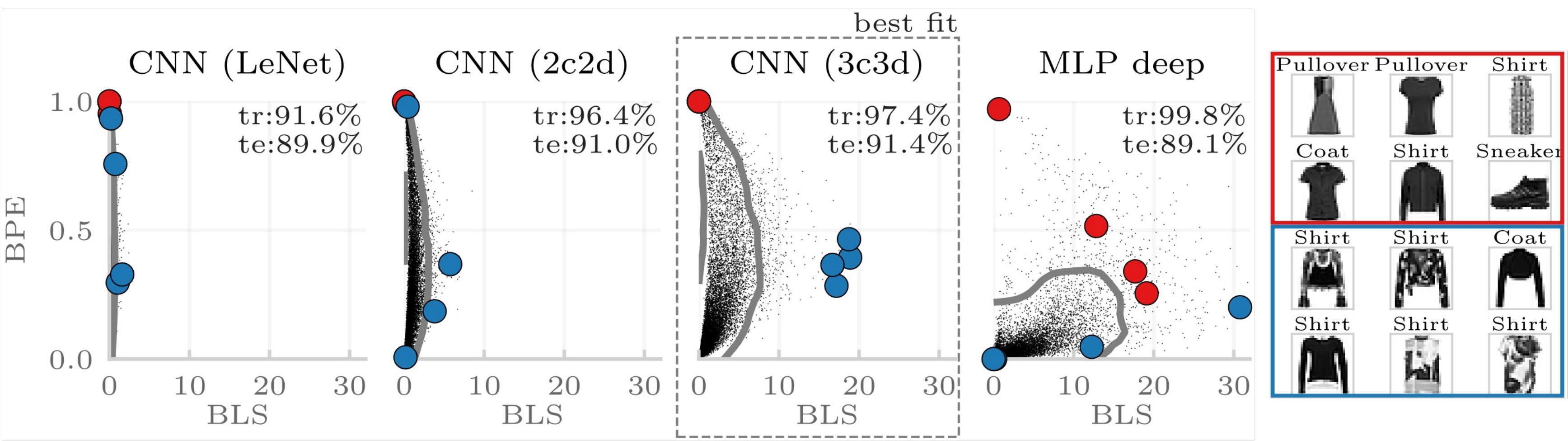


(b) CIFAR10 + ResNet-20

Understanding model complexity and diagnosing overfitting



(a) MNIST



(b) FMNIST

Acknowledgements

Paul Chang
(Aalto University)

Siddharth Swaroop
(Harvard University)

Eric Nalisnick
(University of Amsterdam)

Arno Solin
(Aalto University)

Emtiyaz Khan
(RIKEN AIP)