# Learn-by-Calibrating: Using Calibration as a Training Objective

**Jay Thiagarajan**
*Lawrence Livermore National Labs*

**Bindya Venkatesh**
*Arizona State University*

**Deepta Rajan**
*IBM Research AI Almaden*

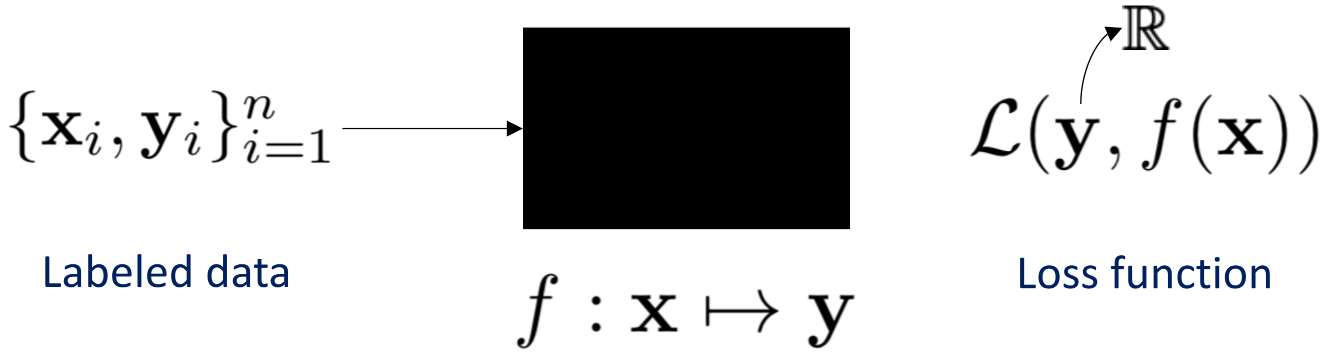
Lawrence Livermore National Laboratory

ASU
Arizona State University

IBM

ICASSP2020
Barcelona

# Predictive Models Can Emulate Complex Processes and Produce Powerful Surrogates



$$\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n} \longrightarrow$$

Labeled data

$$f : \mathbf{x} \mapsto \mathbf{y}$$

$$\mathcal{L}(\mathbf{y}, f(\mathbf{x})) \to \mathbb{R}$$

Loss function

- Supervised learning resorts to empirical risk minimization

$$\arg\max_{\phi} \log p(\phi|\mathcal{D}) \qquad = \arg\max_{\phi} \log p(\mathcal{D}|\phi) + \log p(\phi)$$

model parameters

training data

data likelihood

regularizer (e.g., weight decay)

# Choice of Loss Function is Driven by Assumptions on the Residual Structure in Observed Data

- Residuals from the model can be computed as $\mathbf{r} = (\mathbf{y} - f(\mathbf{x}))$
- Example: L2 metric
    - Assumes that the distribution is symmetric.
    - Optimal estimate is the conditional mean $E(\mathbf{y}|\mathbf{x})$
    - Susceptible to outlying data
- Robust alternatives: Huber, Vapnik's ε-sensitive loss etc.
- When data is heterogeneous, symmetric losses can be inappropriate.
    - Parameterized asymmetric loss functions such as quantile, quantile Huber can be used

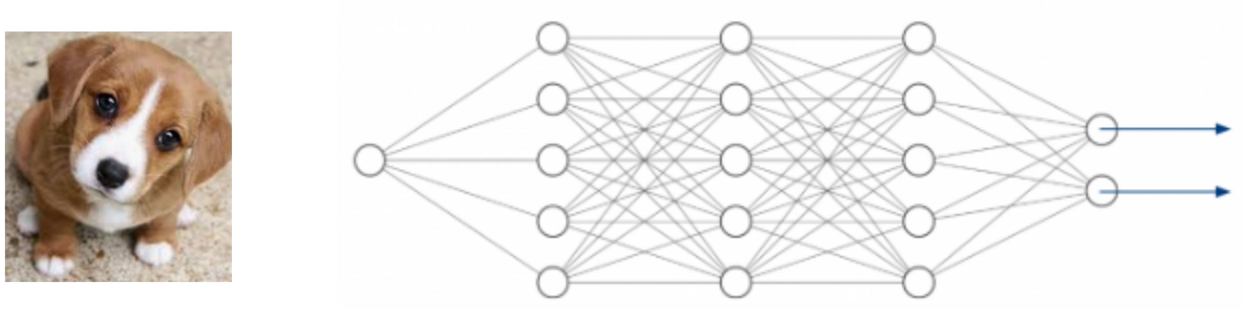
# In this Work, We Explore the Use of Calibration as a Learning Objective in Predictive Models

- Determining optimal parameters for parameterized asymmetric loss functions is challenging
    - In practice, carried out using cross-validation.
    - Model uncertainties (*epistemic*) can make this inferencing difficult.

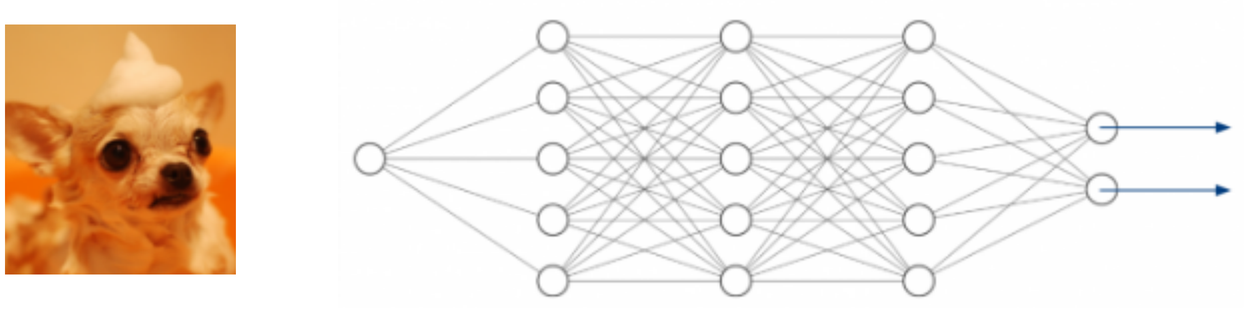
- Calibration is a popular idea in uncertainty quantification

*"Uncertainty Quantification refers to the scientific process of predicting outcomes based on finite amounts data to provide measures of confidence that are used to inform decisions"*

# In Classification, Calibration is Measured as Discrepancy Between Accuracy and Expected Confidence!

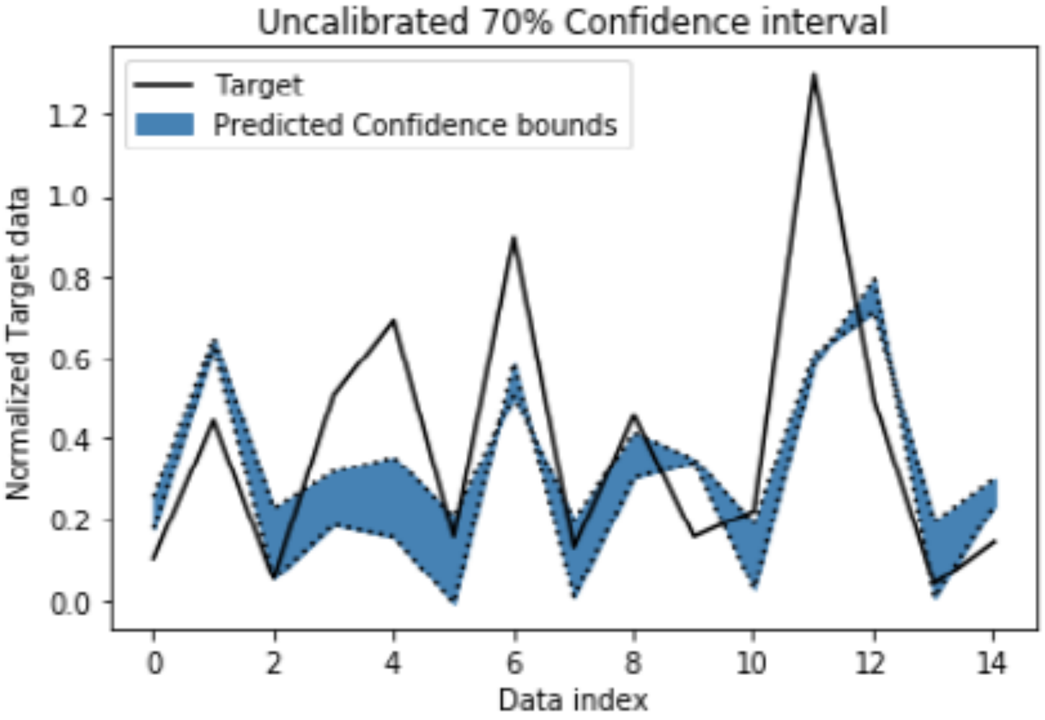


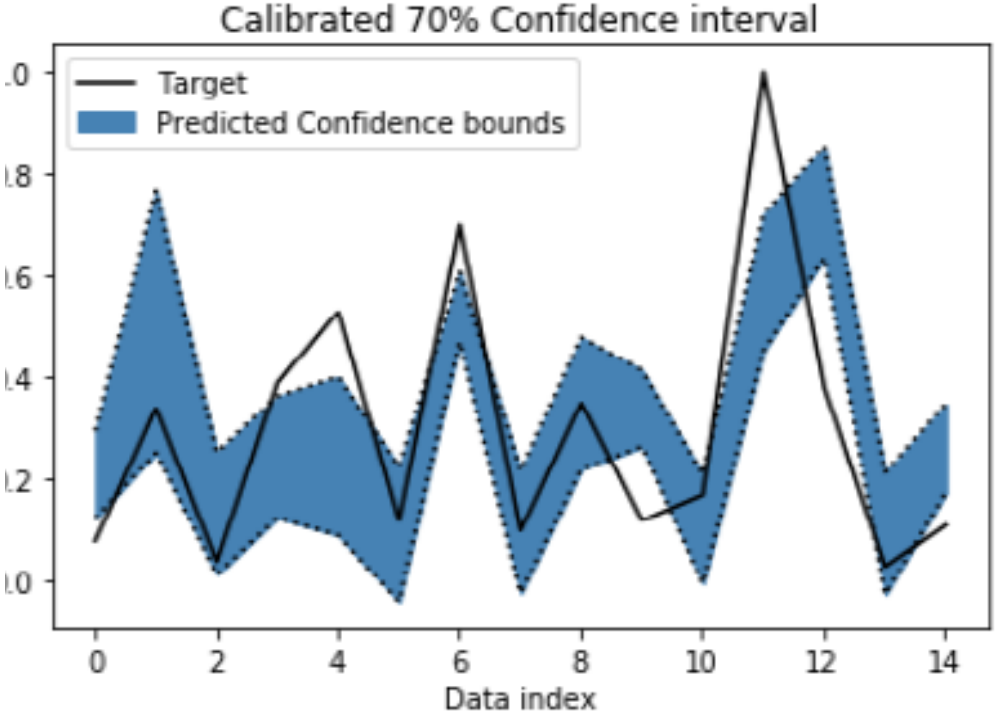
95% Dog
5% Cat



15% Dog
85% Cat

# In Regression Problems, We Often Consider the Notion of Interval Calibration to Evaluate Predictions





*The likelihood of the true target falling in the interval is consistent with the confidence level of the interval*

# In this Work, We Explore the Use of Calibration as a Learning Objective in Predictive Models

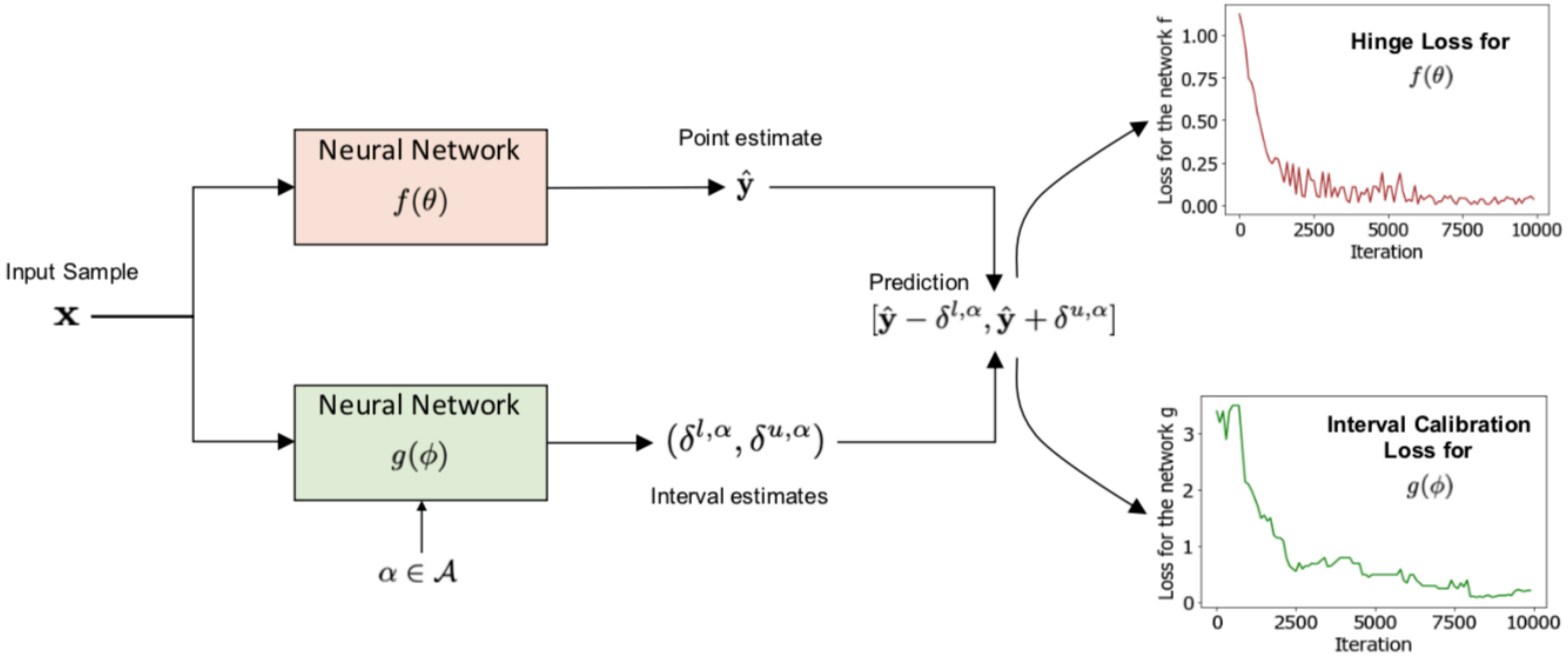
- While calibration is conventionally used for evaluating uncertainty estimators, we utilize it to construct loss functions that reflect the true data characteristics.

$$p\Big((\hat{\mathbf{y}} - \delta^{l}) \leq \mathbf{y} \leq (\hat{\mathbf{y}} + \delta^{u})\Big) = \alpha$$

$$\Downarrow$$

$$\mathrm{L}_{emce} = \left| \alpha - \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[(\hat{\mathbf{y}}_i - \delta_i^{l}) \leq \mathbf{y}_i \leq (\hat{\mathbf{y}}_i + \delta_i^{u})\right] \right|$$

# Our Approach is Comprised of Two Models to Obtain Mean and Interval Estimates



- No explicit distribution assumption on the residuals

# Our Approach is Comprised of Two Models to Obtain Mean and Interval Estimates

- A bi-level optimization problem

Interval Estimator

$$\begin{aligned} &\min_{\theta} \mathcal{L}_f\left(\theta; \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n}, g(\phi^*)\right) \\ &\text{s.t. } \ \phi^* = \arg\min_{\phi} \mathcal{L}_g\left(\phi; \{\mathbf{x}_i\}_{i=1}^{n}, f(\theta)\right) \end{aligned}$$

Mean Estimator

# Our Approach is Comprised of Two Models to Obtain Mean and Interval Estimates

Calibration Loss for Interval Estimator

Calibration

$$\sum_{\alpha\in A}\left(\left|\alpha-\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[\hat{\mathbf{y}}_i-\delta_i^{\alpha}\leq \mathbf{y}_i\leq \hat{\mathbf{y}}_i+\delta_i^{\alpha}\right]\right| + \lambda_1\left|(\hat{\mathbf{y}}_i+\delta_i^{\alpha})-\mathbf{y}_i\right| + \lambda_2\left|\mathbf{y}_i-(\hat{\mathbf{y}}_i-\delta_i^{\alpha})\right|\right)$$

Sharpness

Hinge Loss for Mean Estimator

$$\sum_{i=1}^{N} w_i\Big[\max(0,(\hat{\mathbf{y}}_i-\delta_i^{\alpha})-\mathbf{y}_i+\tau) + \max(0,\mathbf{y}_i-(\hat{\mathbf{y}}_i+\delta_i^{\alpha})+\tau)\Big]$$

# Conceptually, this synergistic optimization attempts to achieve calibration at all confidences simultaneously

- Both models are implemented as deep neural networks.
- Our formulation attempts to simultaneously achieve calibration at all confidence levels $\alpha \in \mathcal{A},\ \mathcal{A} = [0.1, 0.3, 0.5, 0.7, 0.9, 0.99]$
    - In practice, this is very challenging and hence we consider a single randomly chosen alpha in each epoch during training.

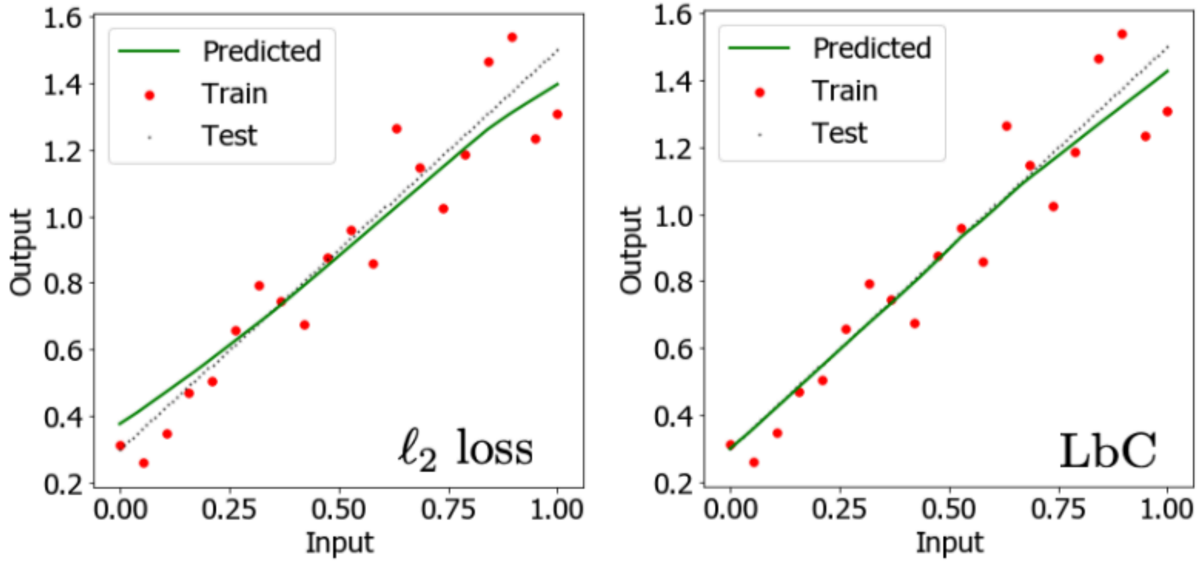
*Improved estimates from the mean estimator can increase calibration error by achieving higher likelihood for a given alpha*

↕

*Update intervals to become sharper in order to reduce the calibration error*

# A Synthetic Example

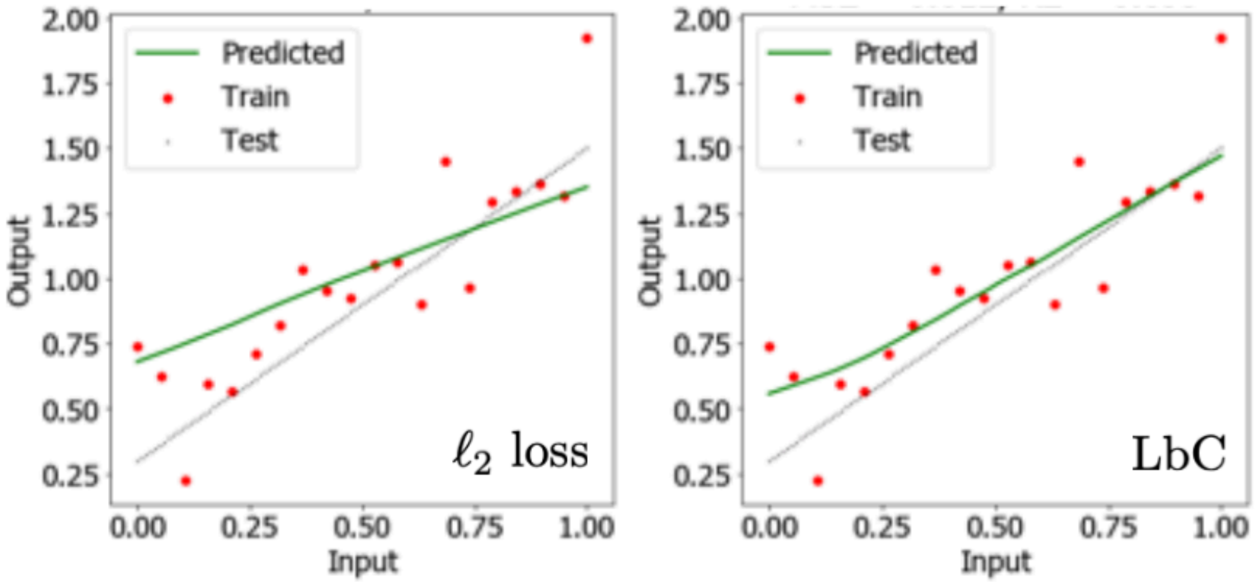
Symmetric Noise



| | $\ell_2$ | Huber | LbC |
|---|---|---|---|
| MSE | 2e-3 | 1e-3 | **8e-4** |
| R2 | 0.984 | 0.986 | **0.994** |

# A Synthetic Example

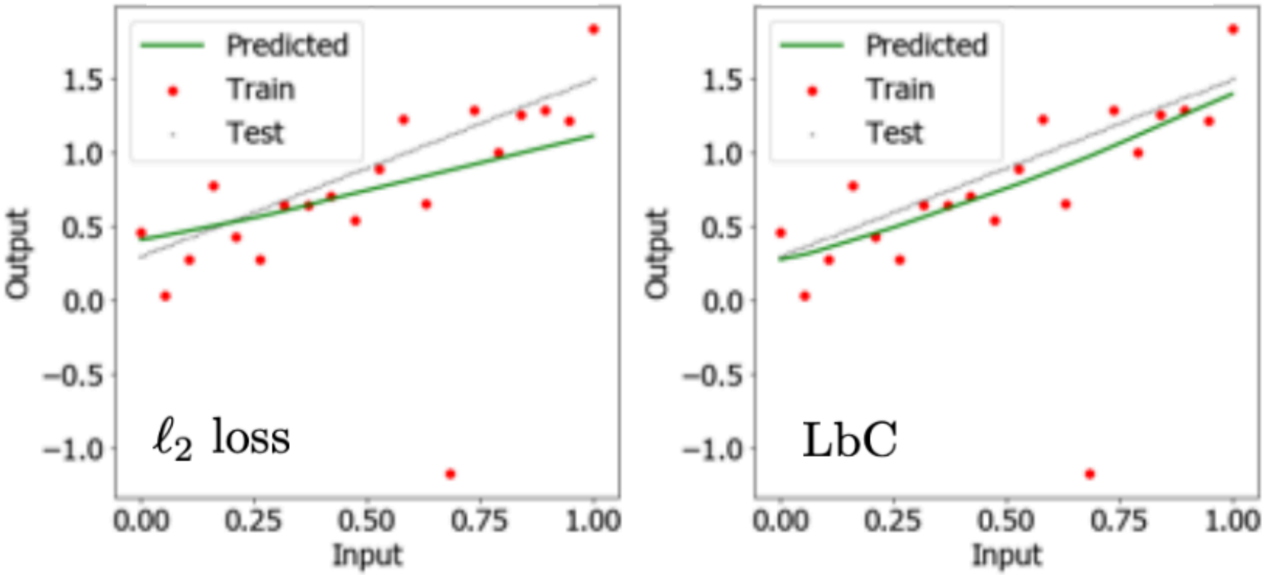
Asymmetric Noise



| | $\ell_2$ | Huber | LbC |
|---|---|---|---|
| MSE | 0.038 | 0.041 | **0.012** |
| R2 | 0.689 | 0.697 | **0.899** |

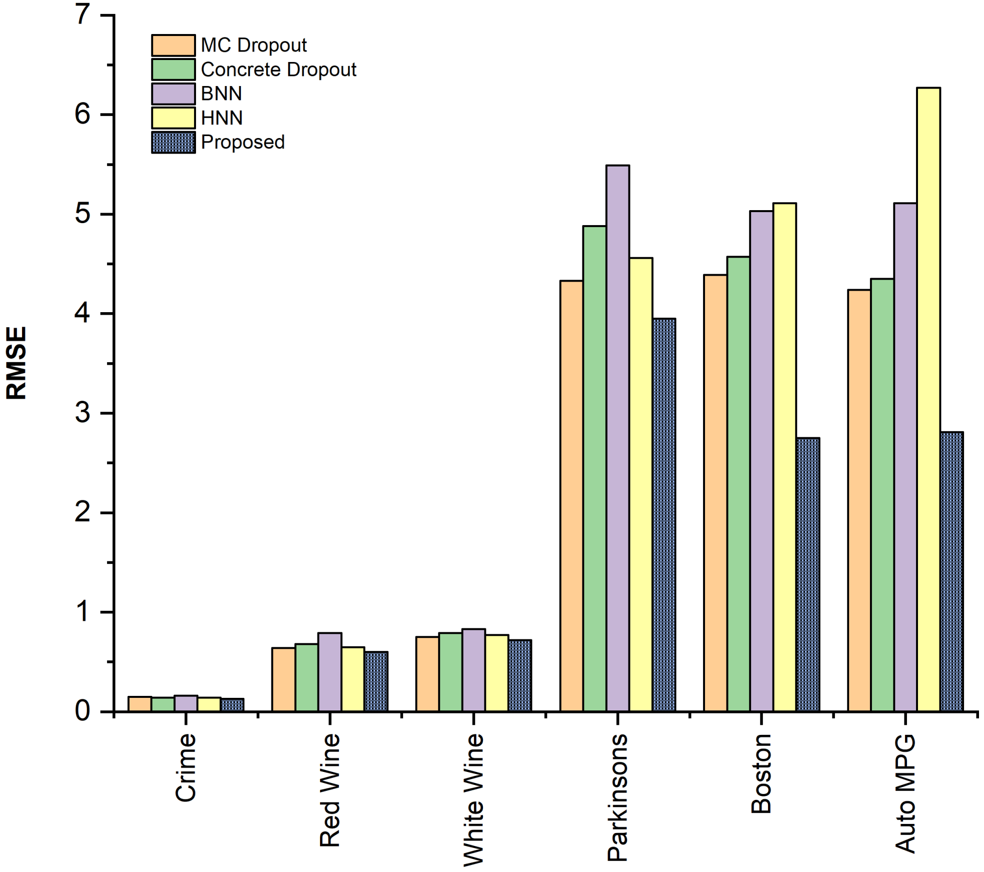
# A Synthetic Example

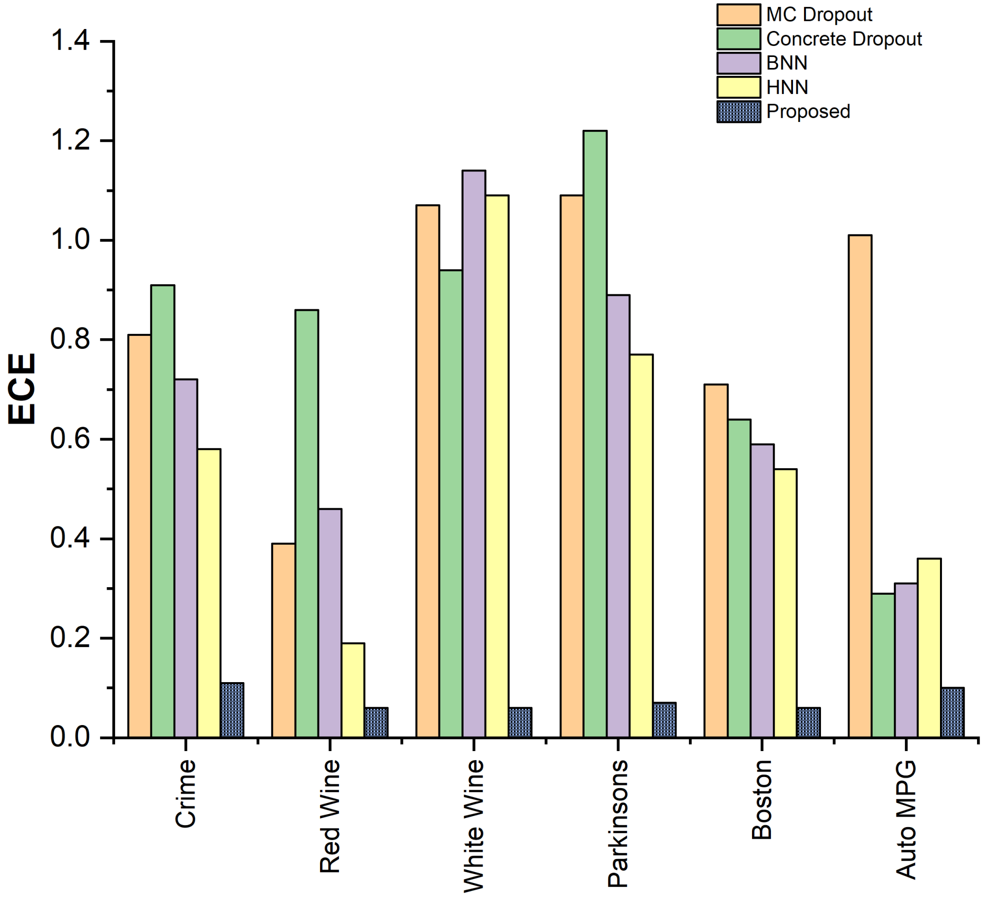
Asymmetric Noise + Outlier



| | $\ell_2$ | Huber | LbC |
|---|---|---|---|
| MSE | 0.043 | 0.039 | **0.014** |
| R2 | 0.652 | 0.688 | **0.885** |

# With Multiple Benchmark Regression Tasks, LbC Consistently Produces Improved Models





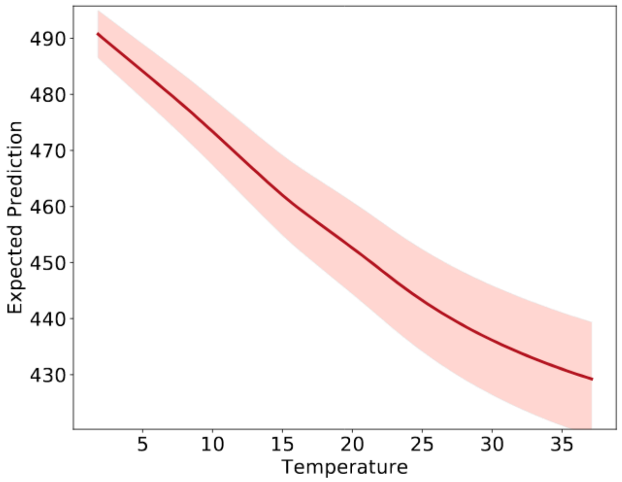
# Using Estimated Intervals to Gain Insights into Model Behavior – Enhanced Partial Dependence Plots

- PDP studies the marginal effect of each (or two) feature on the predicted outcome of a model – reveals global relationships.
- Formally, the PDP for a feature $\mathbf{x}^s$ can be estimated as

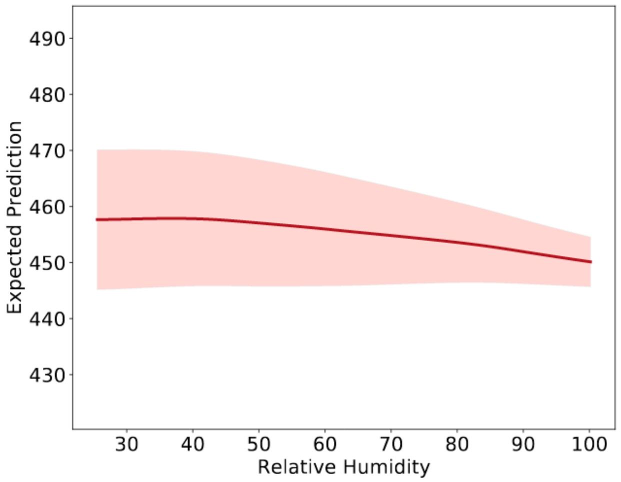
$$\mathrm{P}(\mathrm{x}^s) = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathcal{F}(\mathrm{x}^s, \mathrm{x}_i^c)$$

- We augment PDP with the interval estimates from LbC to obtain a better understanding of the dependencies.

# Enhanced Partial Dependence Plots Reveal More Complex Dependencies that are not Immediately Apparent

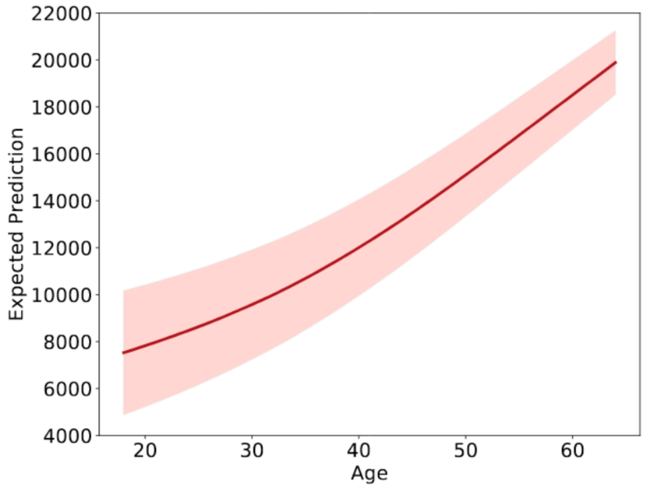


- Inverse relationship with temperature
- More sensitive at lower values



- Mean shows no apparent relationship
- Intervals reveal a complex relationship at lower values

Power Plant Dataset



- Intervals at Age 20 are large enough to overlap with cost at 35.
- Predictions are sensitive as Age variable grows

Insurance Cost Dataset

# Summary

- The notion of calibration from UQ can be effectively repurposed to train predictive models.
- A prior-free loss function that reflects the true data characteristics.
- Consistently produces superior generalization.
- LbC is found to be highly effective in small data regimes compared to standard neural networks.
- Well-calibrated intervals can shed light into the model’s behavior – Enhaced PDPs.

# Questions?

Contact: jjayaram@llnl.gov