What is Your Metric Telling You?

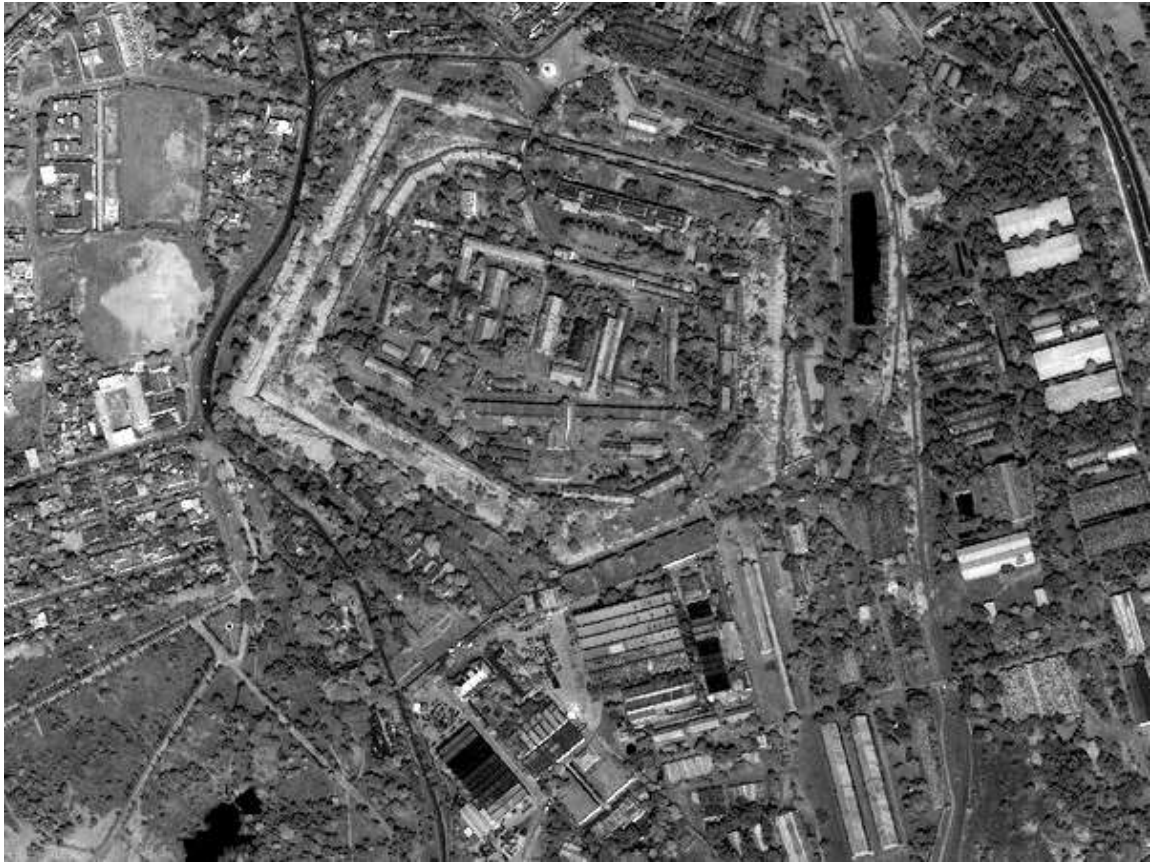
Evaluating Classifier Calibration under Context-Specific Definitions of Reliability

Eric Heim, Senior Machine Learning Research Scientist
AI Division, Software Engineering Institute, Carnegie Mellon University

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

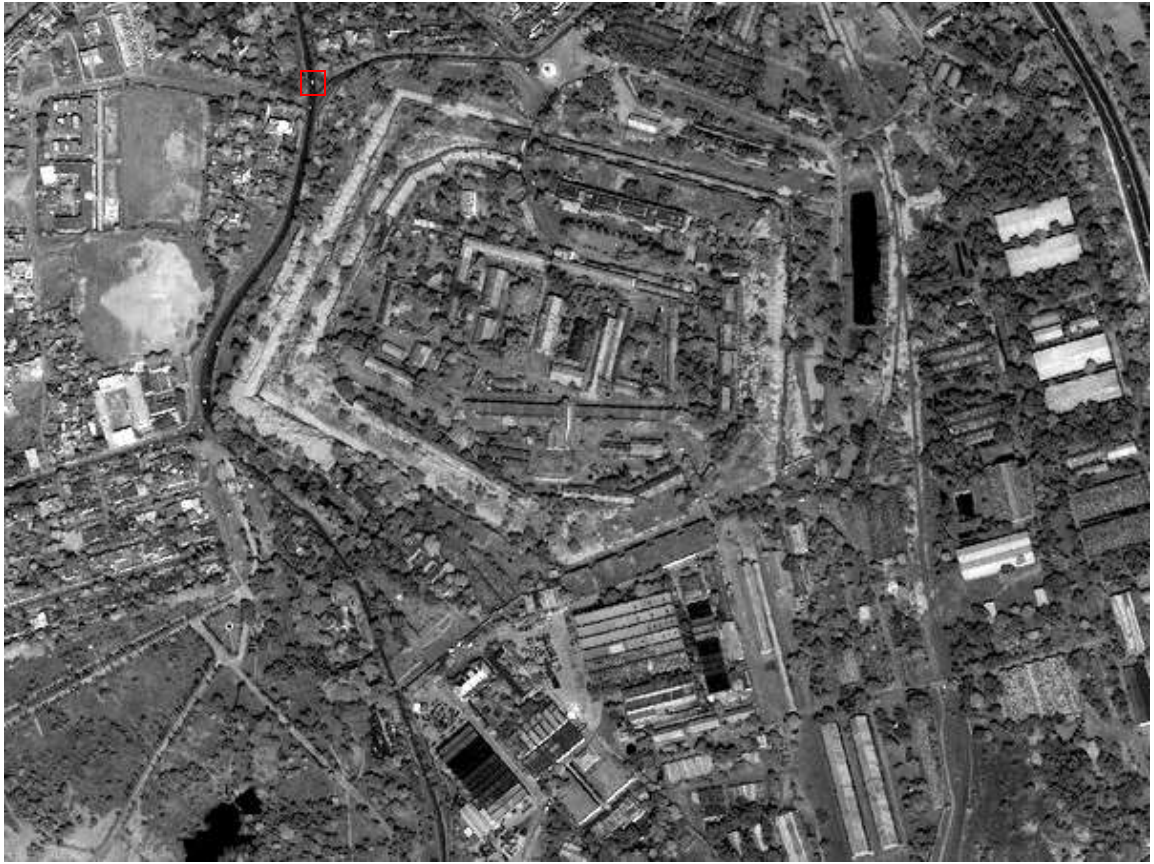
DM22-0394

An Argument for Uncertainty in Machine Learned Models



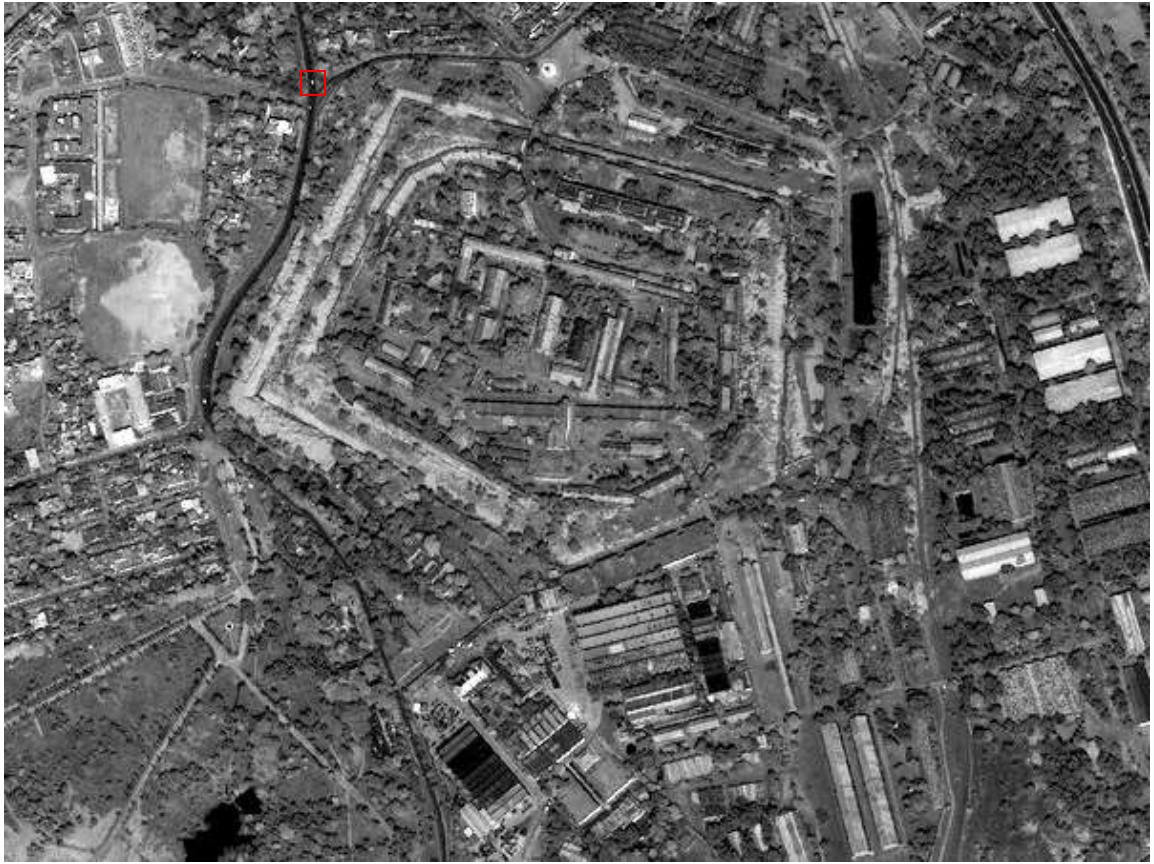
An Argument for Uncertainty in Machine Learned Models

Predicted Class
Civilian Vehicle



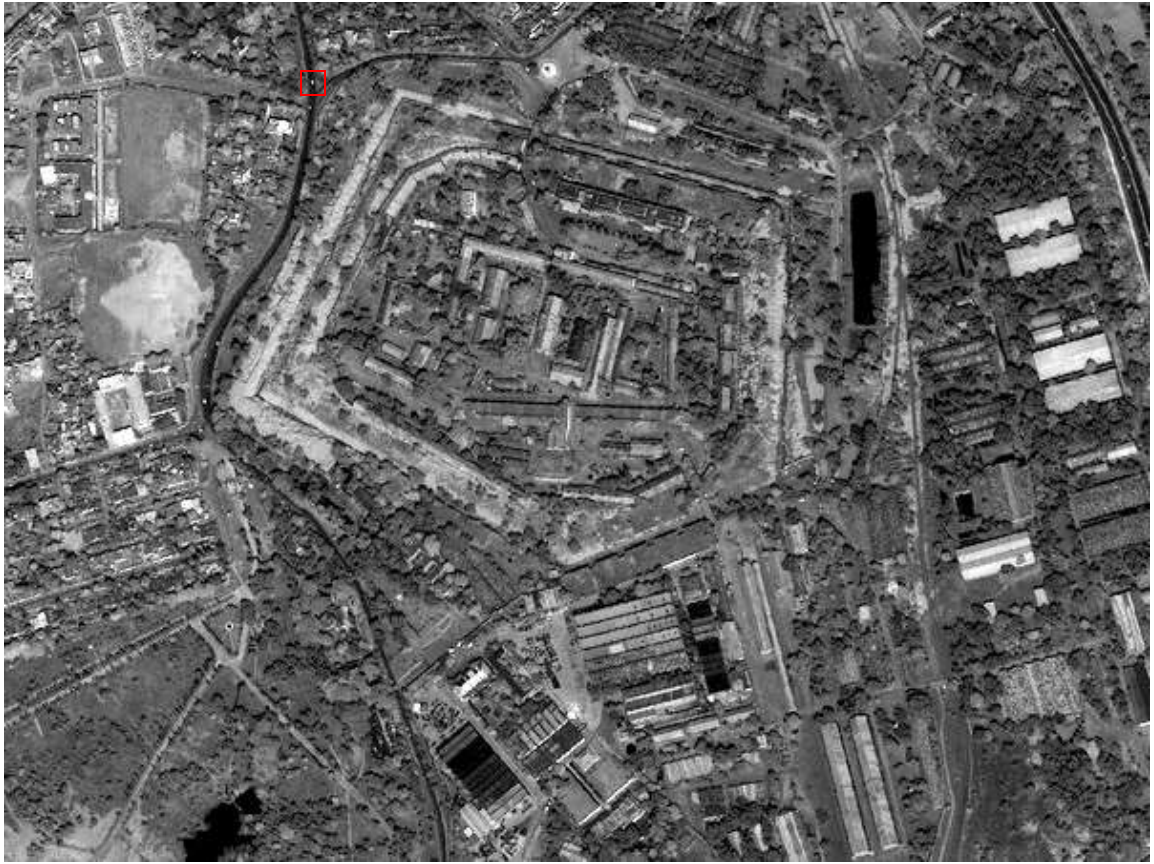
An Argument for Uncertainty in Machine Learned Models

Predicted Class	Confidence
Civilian Vehicle	0.95



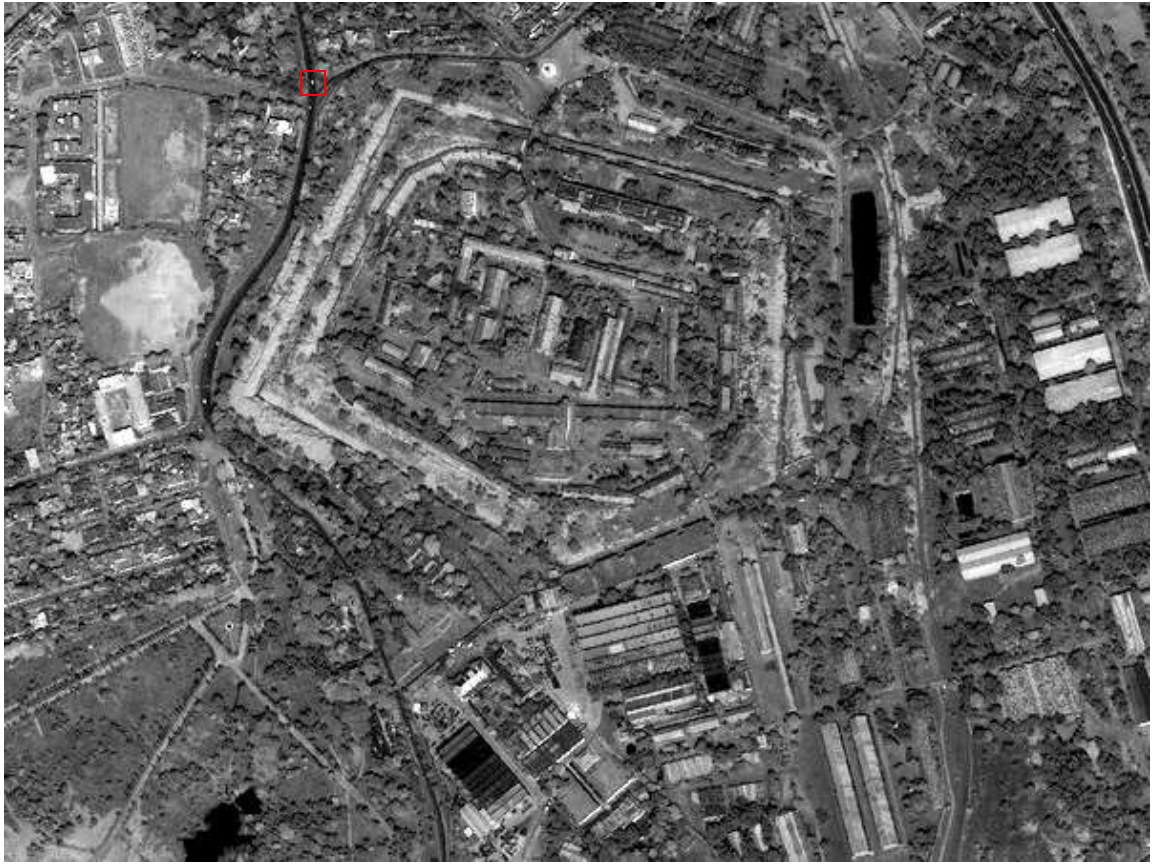
An Argument for Uncertainty in Machine Learned Models

Predicted Class	Confidence
Civilian Vehicle	0.6



An Argument for Uncertainty in Machine Learned Models

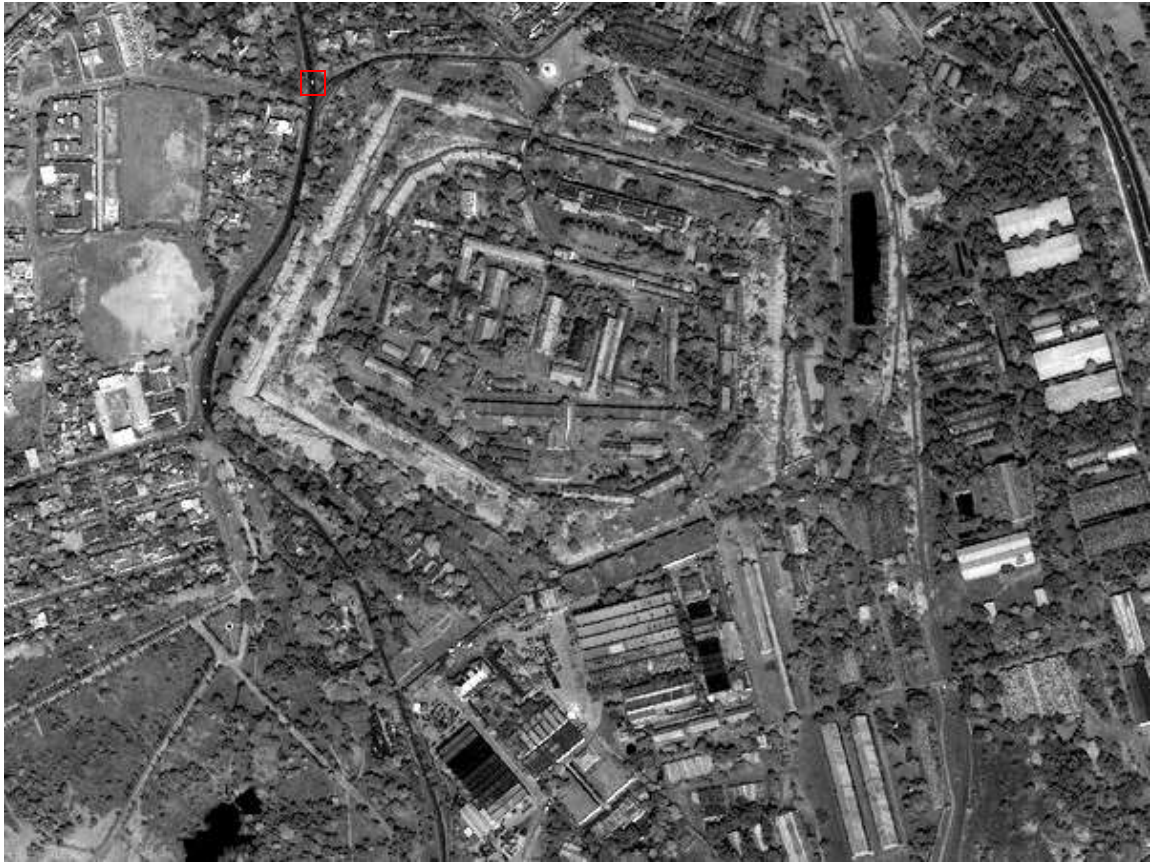
Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	0.35
…	…



An Argument for Uncertainty in Machine Learned Models

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	0.35
…	…

Confidence/Uncertainty can lead to more informed decision-making.

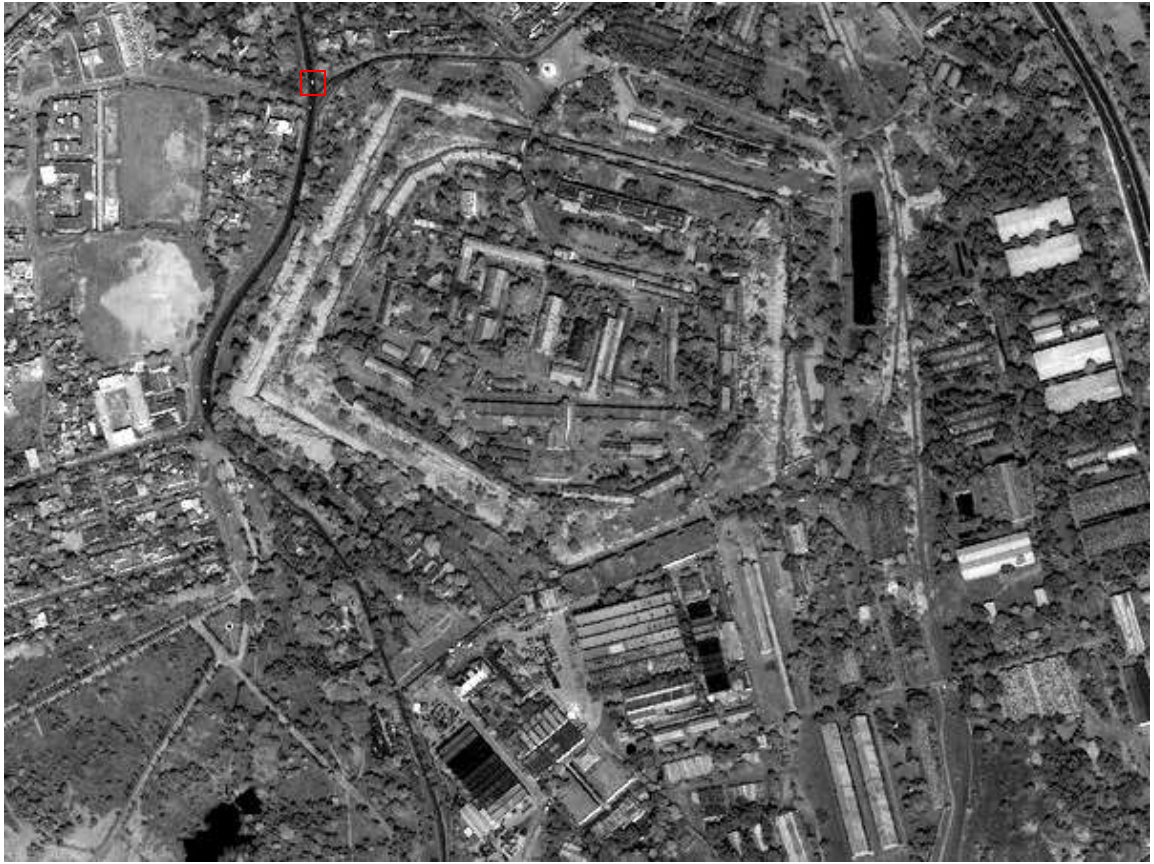


An Argument for Uncertainty in Machine Learned Models

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	0.35
…	…

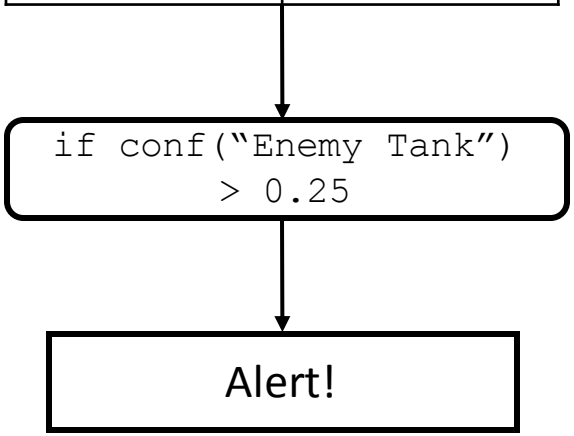
```
if conf("Enemy Tank")
        > 0.25
```

Alert!



An Argument for Uncertainty in Machine Learned Models

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	0.35
…	…



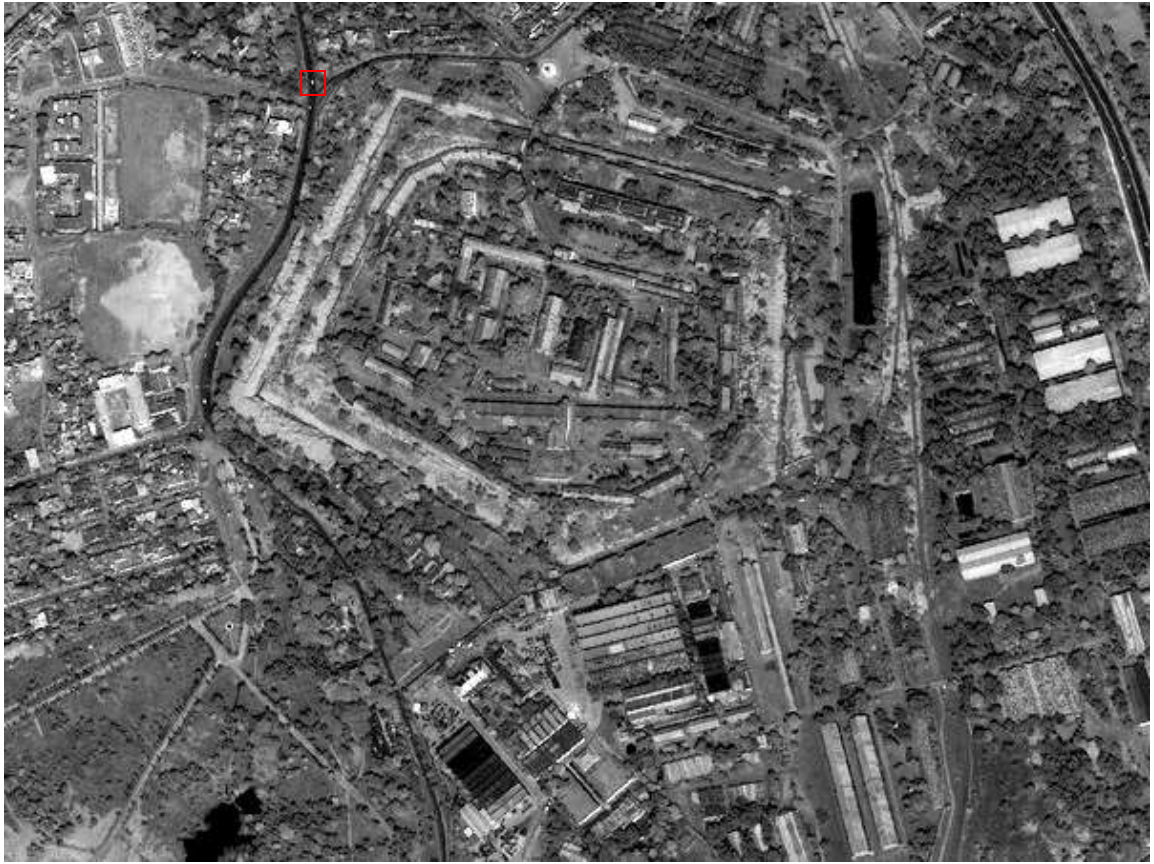
Confidence/Uncertainty can be used to make AI Systems more reliable.

An Argument for Uncertainty in Machine Learned Models

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	0.35
…	…

Why is the model uncertain about this instance?

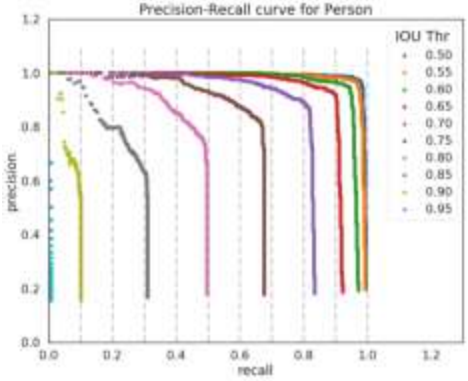
Uncertainty can be used to identify how to improve models.



A Quick Detour into T&E: Context-Focused Evaluation

Often ML models/methods are taken (almost) directly from academic research…
…However, the evaluations are often general and on benchmark data sets.

Because of this, the full evaluation methodology should not be directly adopted.



mAP: Average area under precision recall curve over different IoU thresholds.

In practice: Requirements on IoU, precision, and/or recall.

More specifically: Evaluations of ML models should reflect:
1. How models will be used in practice
2. Specific scenarios of importance to the application of the model

By focusing on these, evaluations can measure important characteristics of how the model will function in the context it will be deployed.

What we've done for T&E in UQ: *Context-focused Calibration Metrics*

Classifiers are often used to inform decisions…
…despite this how classifier calibration is evaluated assume a particular decision rule.

Most works assume the *Top-1* decision rule

"Of all the times my model outputs maximum confidence x, it should be right x percent of the time."

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	0.35
…	…

What we've done for T&E in UQ: *Context-focused Calibration Metrics*

Classifiers are often used to inform decisions…
…despite this how classifier calibration is evaluated assume a particular decision rule.

Most works assume the *Top-1* decision rule

"Of all the times my model outputs maximum confidence 0.6, it should be right 60% percent of the time."

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	???
…	???

Other interpretations of classifier outputs are more appropriate to facilitate decision making in different contexts or focused evaluation.

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	0.35
…	…

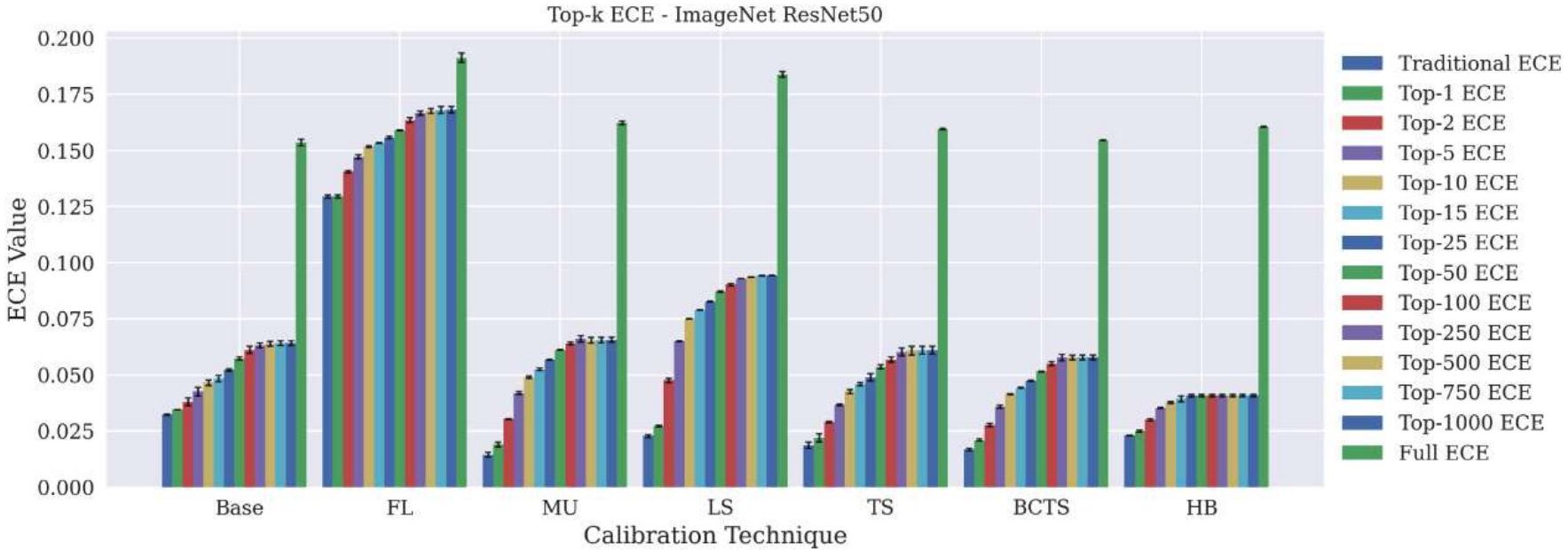
VS.

Confidence in class
Very Low
Low
Medium
High
Very High

Selected Result from [Kirchenbauer, Oaks, and Heim; 2022 (In Review)]

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

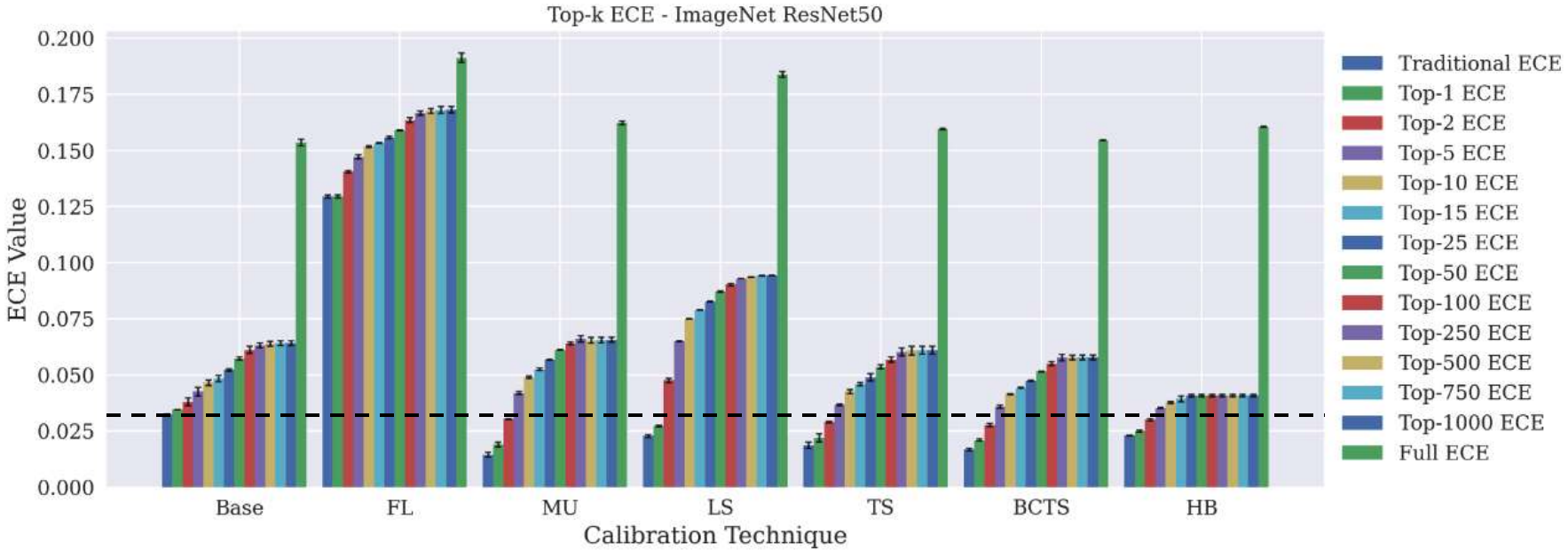
Experiment #1: Top-k ECE



Selected Result from [Kirchenbauer, Oaks, and Heim; 2022 (In Review)]

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

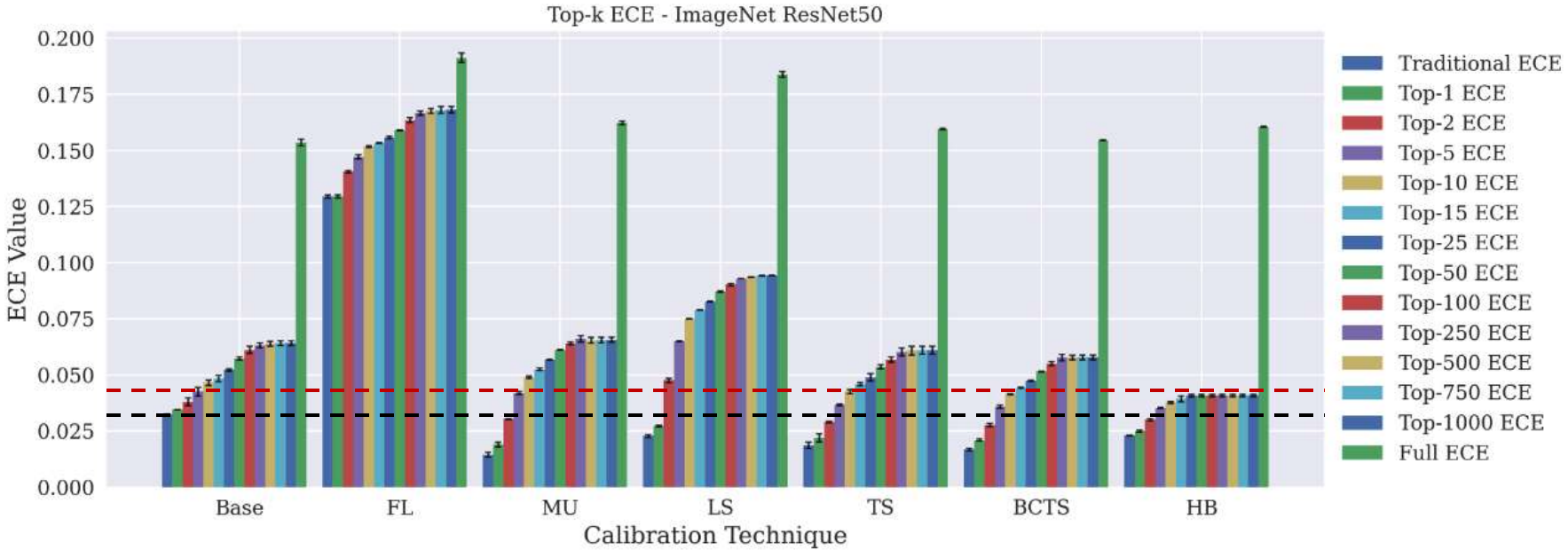
Experiment #1: Top-k ECE



Selected Result from [Kirchenbauer, Oaks, and Heim; 2022 (In Review)]

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

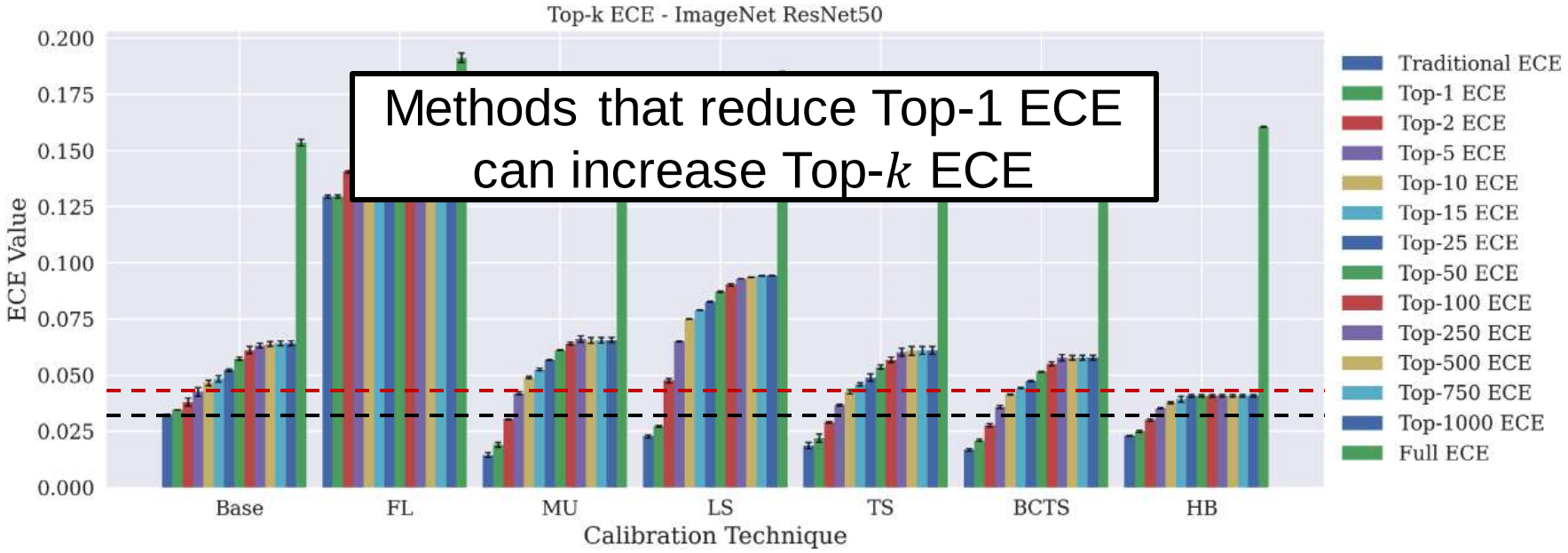
Experiment #1: Top-k ECE



Selected Result from [Kirchenbauer, Oaks, and Heim; 2022 (In Review)]

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

Experiment #1: Top-k ECE



Main Takeaways

Quantifying the Uncertainty of ML Models….

1. Facilitates better decision making
2. Can potentially reveal failure cases

Based on the need for **context-specific metrics for evaluation**, we developed a framework for measuring calibration error for a variety of cases.

We show that many popular calibration techniques fail at calibrating models in these cases.

Next for Uncertainty Quantification:

1. Use our calibration metrics in specific applications.
2. Learn models that can be calibrated according to context-specific definitions of calibration.
3. Use uncertainty to identify failure modes in models (out of distribution, noisy, anomalous instances)

More Broadly:

1. Further define processes for establishing ML model evaluation **in context**.

This work: Evaluating classifiers for context-specific calibration

How do you evaluate your classifier for it's ability to accurately express uncertainty?

First, you need to define how to interpret the confidence outputs of classifiers.

This work: Evaluating classifiers for context-specific calibration

How do you evaluate your classifier for it's ability to accurately express uncertainty?

First, you need to define how to interpret the confidence outputs of classifiers.

This work: Evaluating classifiers for context-specific calibration

How do you evaluate your classifier for it's ability to accurately express uncertainty?

First, you need to define how to interpret the confidence outputs of classifiers.

How do we understand these values?

This work: Evaluating classifiers for context-specific calibration

How do you evaluate your classifier for it's ability to accurately express uncertainty?

First, you need to define how to interpret the confidence outputs of classifiers.

Classifier Calibration: Classifier outputs match the frequency of class labels.

This work: Evaluating classifiers for context-specific calibration

How do you evaluate your classifier for it's ability to accurately express uncertainty?

First, you need to define how to interpret the confidence outputs of classifiers.

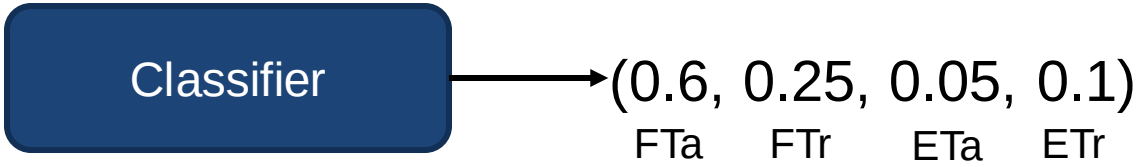
For all possible inputs that the classifier outputs (0.6,0.4)…
60% of the inputs should be a Civilian Vehicle,
40% of the inputs should be an Enemy Tank.

Classifier Calibration: Classifier outputs match the frequency of class labels.

This work: Evaluating classifiers for context-specific calibration

Modern machine learning literature has focused on evaluating classifier calibration according to their **Top-1 Expected Calibration Error (ECE)** (Guo et al; 2017)

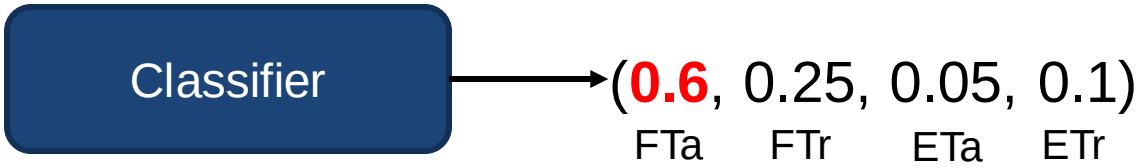
Classes = {Friendly Tank, Friendly Truck, Enemy Tank, Enemy Truck}



This work: Evaluating classifiers for context-specific calibration

Modern machine learning literature has focused on evaluating classifier calibration according to their **Top-1 Expected Calibration Error (ECE)** (Guo et al; 2017)

Classes = {Friendly Tank, Friendly Truck, Enemy Tank, Enemy Truck}



Top-1 Expected Calibration Error (ECE)

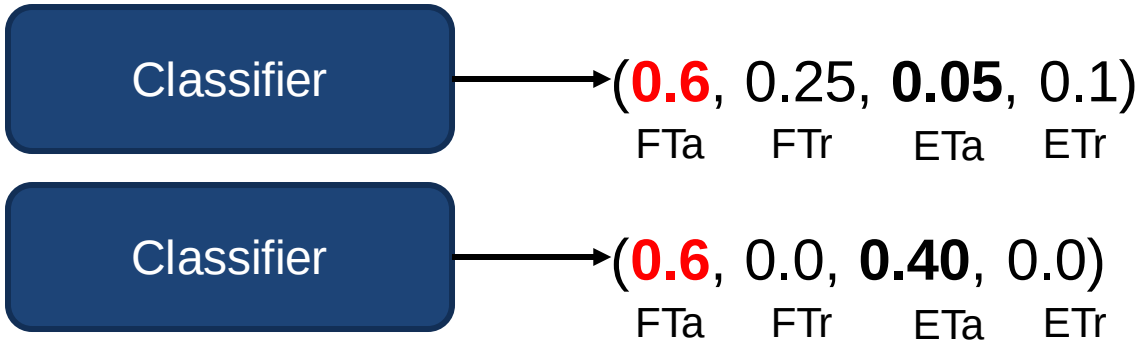
Considers only the most confident class in evaluating for calibration

For all possible inputs that the classifier outputs 0.6 as the most confident class…
60% of the those inputs should be that class.

This work: Evaluating classifiers for context-specific calibration

Modern machine learning literature has focused on evaluating classifier calibration according to their **Top-1 Expected Calibration Error (ECE)** (Guo et al; 2017)

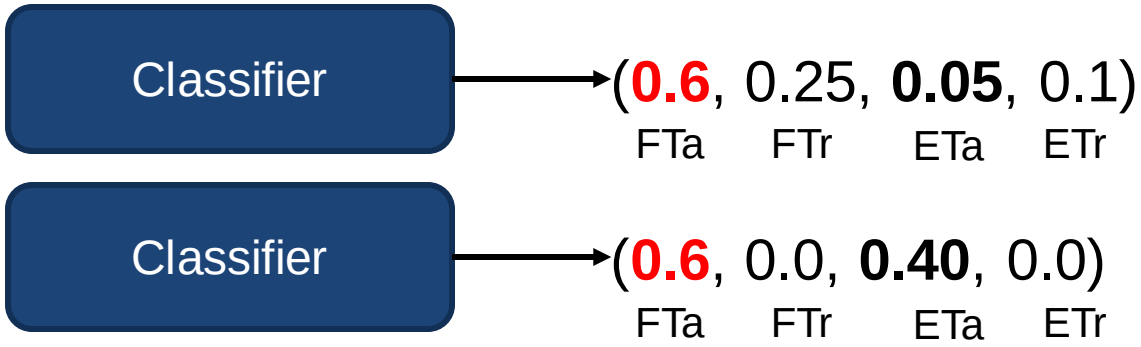
Classes = {Friendly Tank, Friendly Truck, Enemy Tank, Enemy Truck}



This work: Evaluating classifiers for context-specific calibration

Modern machine learning literature has focused on evaluating classifier calibration according to their **Top-1 Expected Calibration Error (ECE)** (Guo et al; 2017)

Classes = {Friendly Tank, Friendly Truck, Enemy Tank, Enemy Truck}



According to Top-1 ECE, these two classifiers *are considered the same*.
However, the two outputs can mean very different things with *mission context*.

This motivates the need for calibration metrics that are able to map to contexts and use cases that match realistic usage of deployment settings.

A statistical framework for context-specific ECE

We adopt the framing of ECE from (Vaicenavicius et al; 2019)

Canonical Calibration(Reliability) Condition

$$\mathbb{P}[Y \in \cdot \mid g(X)] = g(X)$$

A statistical framework for context-specific ECE

We adopt the framing of ECE from (Vaicenavicius et al; 2019)

Canonical Calibration(Reliability) Condition

$$\mathbb{P}[Y \in \cdot \mid g(X)] = g(X)$$

$$\mathbb{P}[Y \in \cdot \mid g(X) = (0.7, 0.1, 0.2))] = (0.7, 0.1, 0.2)$$

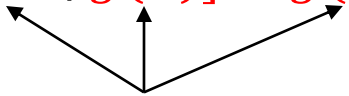
Of all the times g outputs $(0.7, 0.1, 0.2)$ ...
… 70% of the time it was class A
… 10% of the time it was class B
… 20% of the time it was class C

A statistical framework for context-specific ECE

We adopt the framing of ECE from (Vaicenavicius et al; 2019)

Canonical Calibration(Reliability) Condition

$$\mathbb{P}[Y \in \cdot \mid g(X)] = g(X)$$



"Lens"

A statistical framework for context-specific ECE

We adopt the framing of ECE from (Vaicenavicius et al; 2019)

Top-1 Reliability Condition

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$$

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X) = 0.7] = 0.7$$

Of all the times $g's$ highest confidence is 0.7

... 70% of the time it was the class with the highest confidence

A statistical framework for context-specific ECE

We adopt the framing of ECE from (Vaicenavicius et al; 2019)

Top-1 Reliability Condition

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$$

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X) = 0.7] = 0.7$$

Of all the times $g's$ highest confidence is 0.7
... 70% of the time it was the class with the highest confidence

Informally, a lens is a transformation of 1) The classifier's output 2) The probability simplex over classes that results in an induced classification problem.

A statistical framework for context-specific ECE

We adopt the framing of ECE from (Vaicenavicius et al; 2019)

Top-1 Reliability Condition

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$$

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X) = 0.7] = 0.7$$

Of all the times $g's$ highest confidence is 0.7
... 70% of the time it was the class with the highest confidence

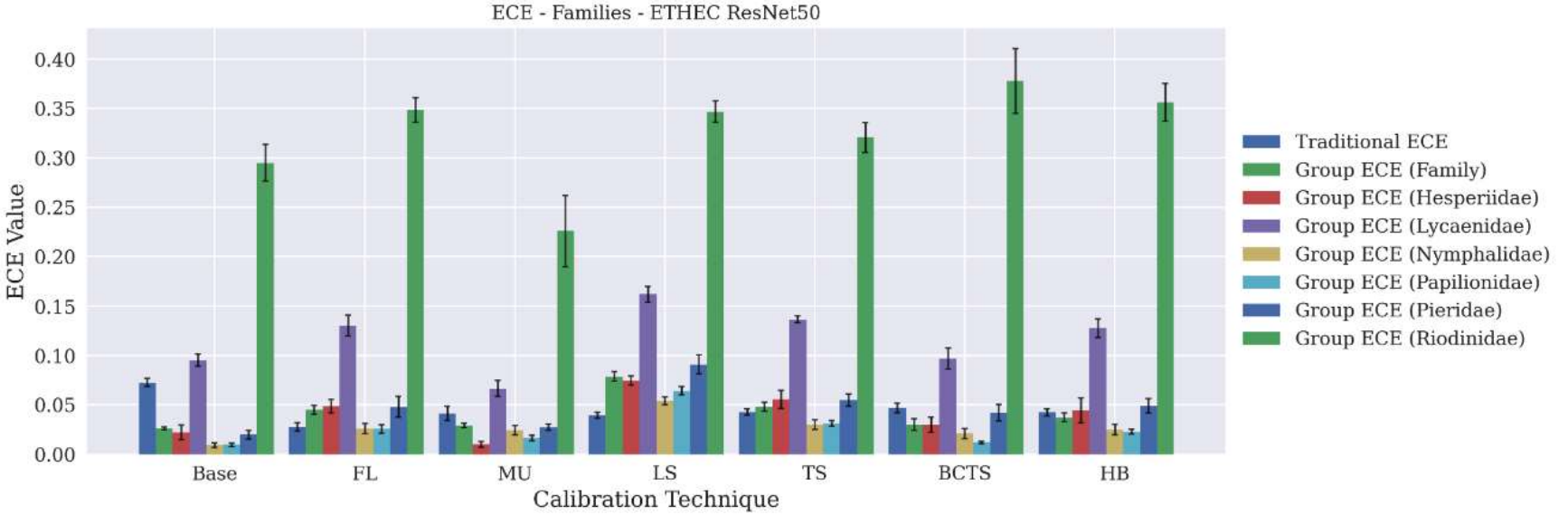
Informally, a lens is a transformation of 1) The classifier's output 2) The probability simplex over classes that results in an induced classification problem.

By designing lensing functions that result induced classification problems that map to specific mission contexts, you can evaluate classifiers in more application-specific ways.

Selected Results from [Kirchenbauer, Oaks, and Heim; 2022 (In Review)]

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

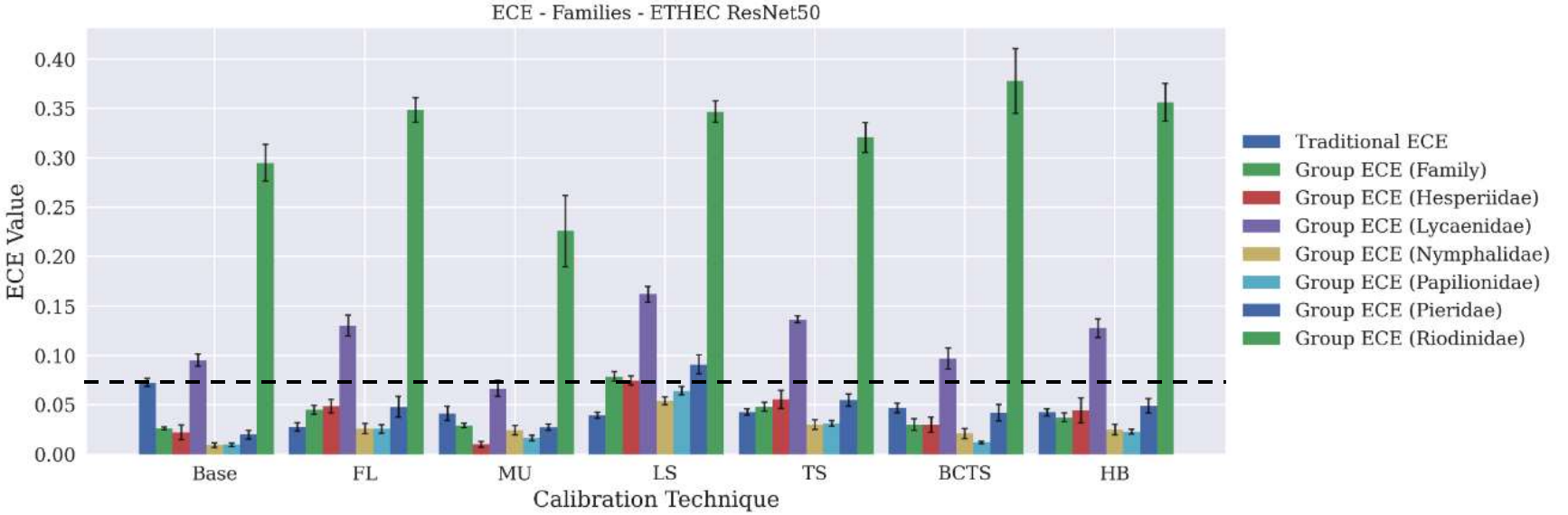
Experiment #2: Group-wise ECE



Selected Results from [Kirchenbauer, Oaks, and Heim; 2022 (In Review)]

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

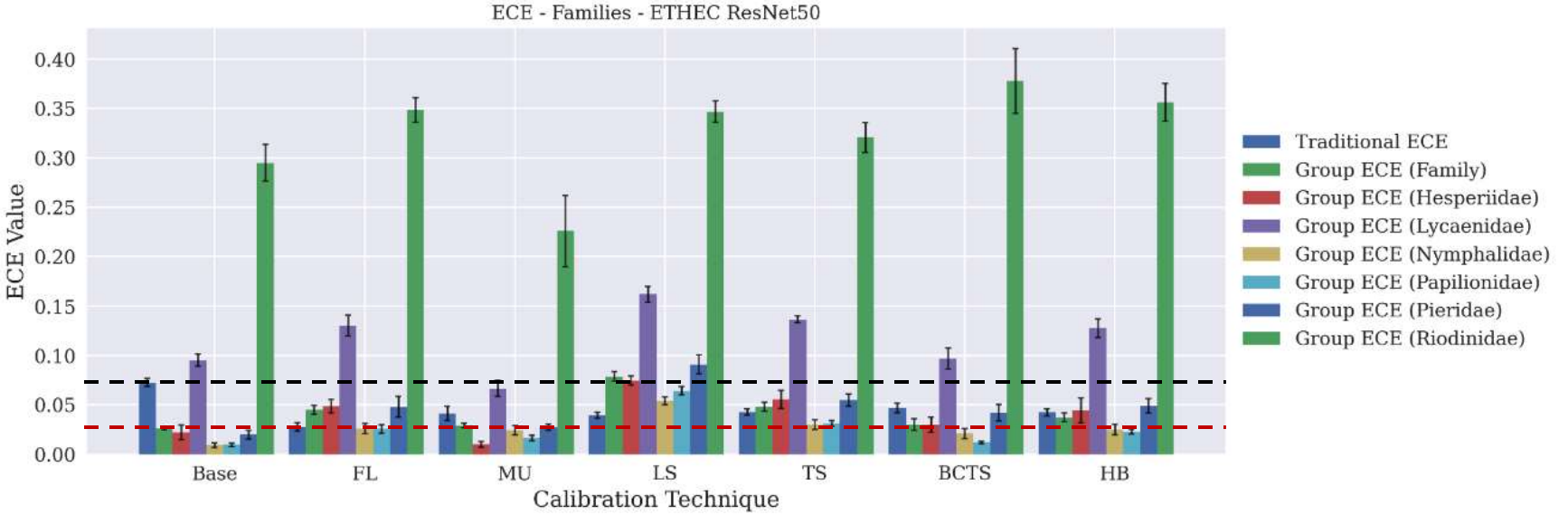
Experiment #2: Group-wise ECE



Selected Results from [Kirchenbauer, Oaks, and Heim; 2022 (In Review)]

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

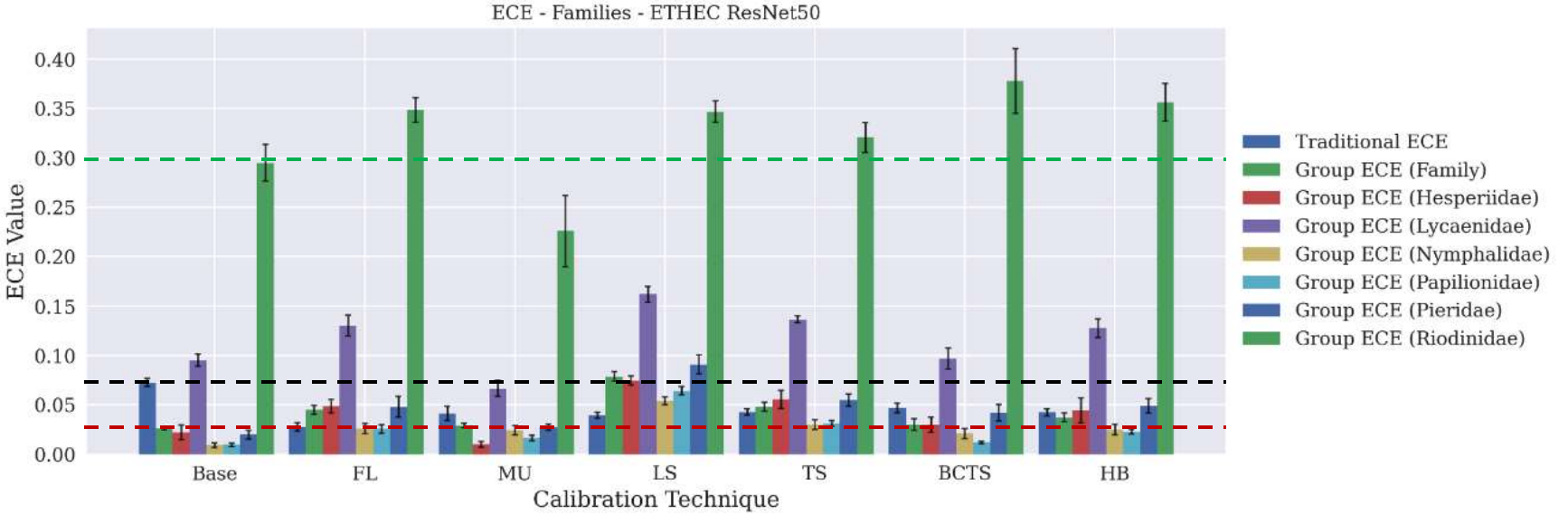
Experiment #2: Group-wise ECE



Selected Results from [Kirchenbauer, Oaks, and Heim; 2022 (In Review)]

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

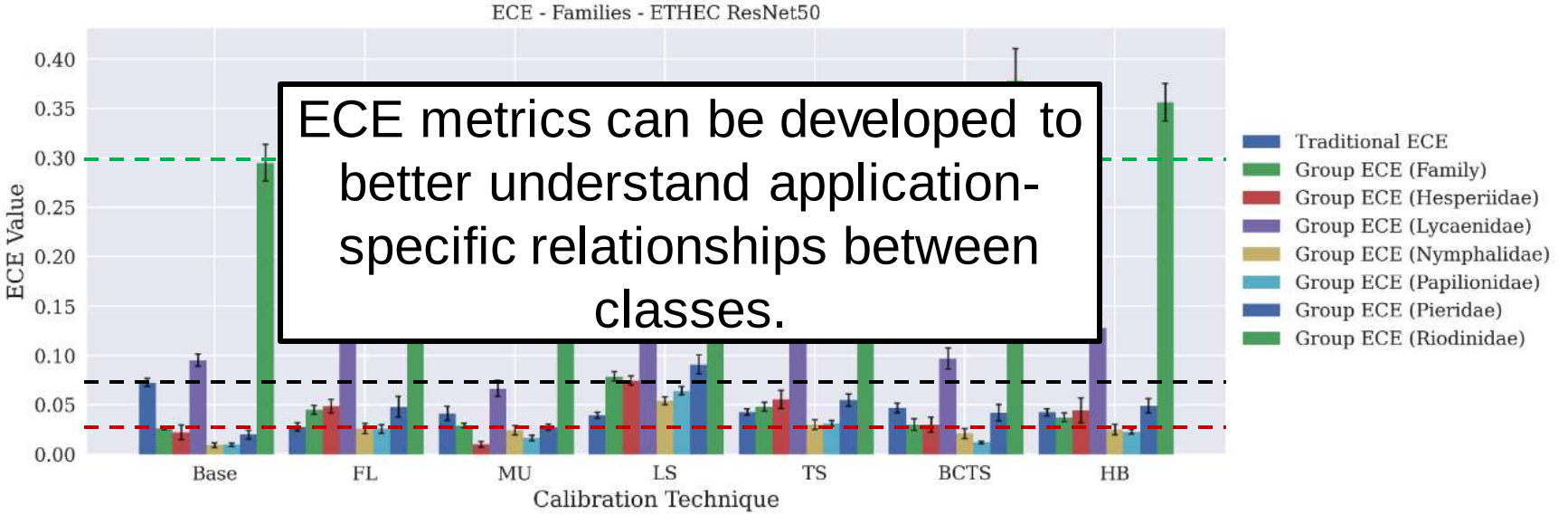
Experiment #2: Group-wise ECE



Selected Results from [Kirchenbauer, Oaks, and Heim; 2022 (In Review)]

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

Experiment #2: Group-wise ECE

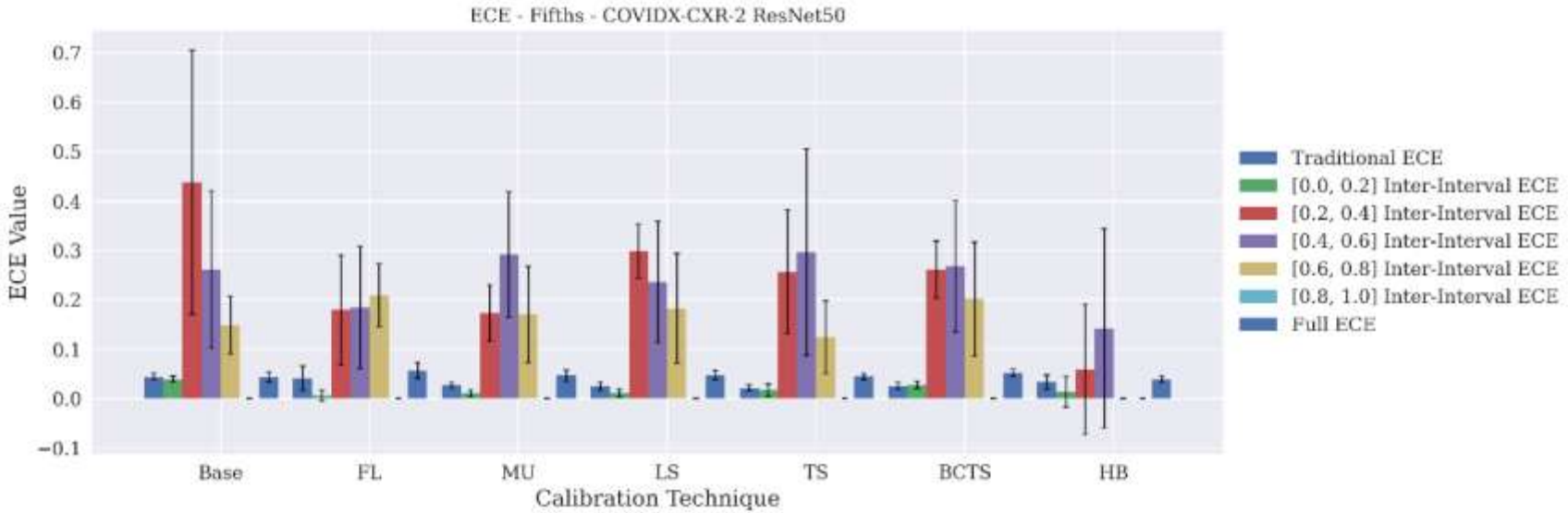


ECE metrics can be developed to better understand application-specific relationships between classes.

Selected Results from [Kirchenbauer, Oaks, and Heim; 2022 (In Review)]

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

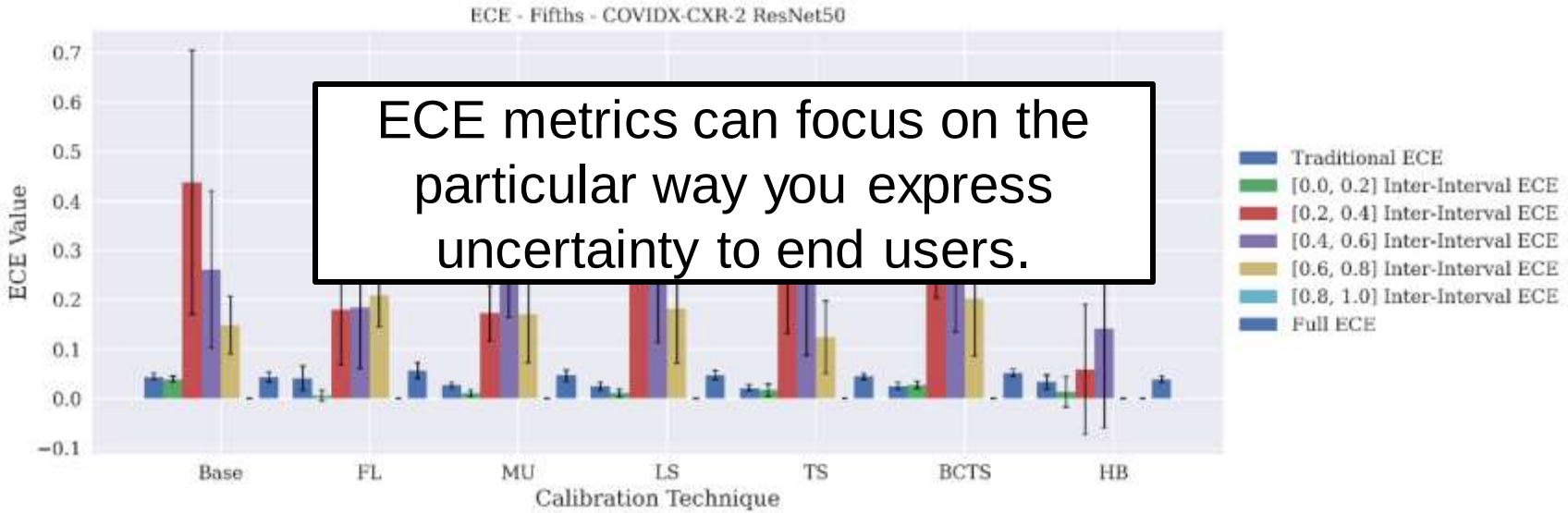
Experiment #3: Likert Category ECE



Selected Results from [Kirchenbauer, Oaks, and Heim; 2022 (In Review)]

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

Experiment #3: Likert Category ECE



ECE metrics can focus on the particular way you express uncertainty to end users.