

The Role of Bayesian Multilevel Models in AI Performance Measurement

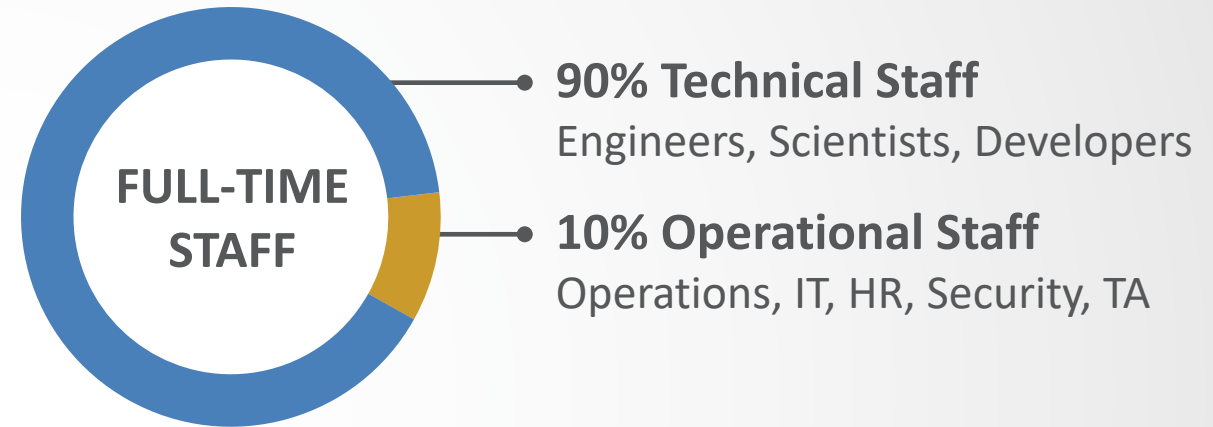
Austin Amaya | aamaya@morsecorp.com

Sean Dougherty | sdougherty@morsecorp.com

Company Overview



- Headquartered in Cambridge Tech Square/MIT
- Founded in Oct 2014
 - Management team together over 15 years
- MORSE Staff
 - 160 Employees
 - 20 interns and co-ops
- 100% employee owned



Introduction



- MORSE T&E needs to rank and select most performant object detection computer vision model
 - Performance estimate needs to align with production environment operating on whole images
 - Need to reframe simple counting metrics such as precision, recall, and F1 in terms of images not objects
- We use a Bayesian multilevel model to estimate performance
- This model estimates performance simultaneously within and across images
 - Estimate is robust to uneven distribution of information across images
 - Learns parameters common to and unique to all images
 - Uncertainty in performance estimate is native to the Bayesian framework
 - Estimates performance from small amounts of data

MORSE uses a multilevel Bayesian model that uses information efficiently and generates full uncertainty distribution to measure production-aligned CV model performance.

T&E Workflow

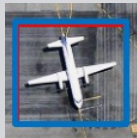
Object Detection AI Model

Model expected to detect and classify few-to-many counts of target classes in each image

Test and Evaluation

Observe AI model performance on a labeled test set and use counting metrics to characterize its ability to correctly detect and identify objects

True Positive



TP = 1

False Negative



FN = 1


False Positive




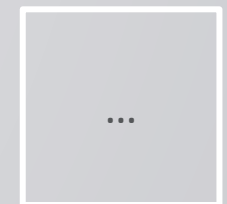
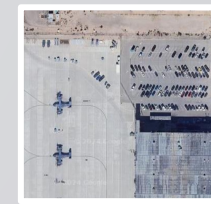
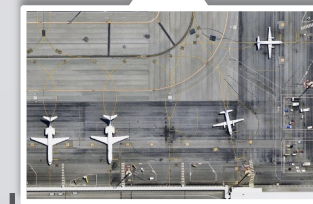
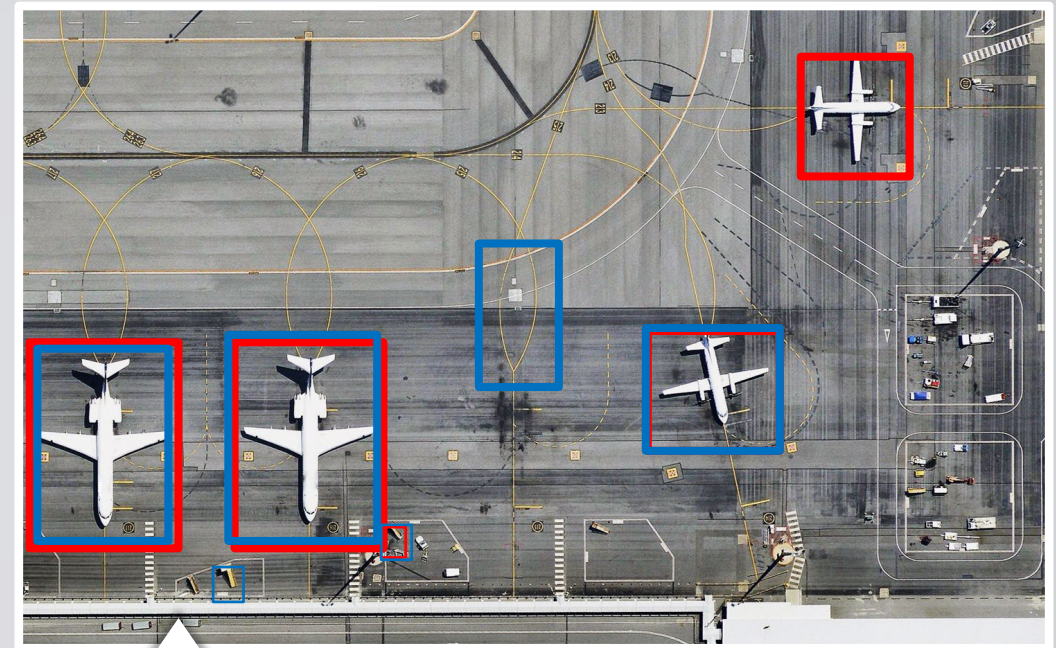
FP = 1

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FN}} \quad \text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FP}}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

 Ground Truth

 Model Detection

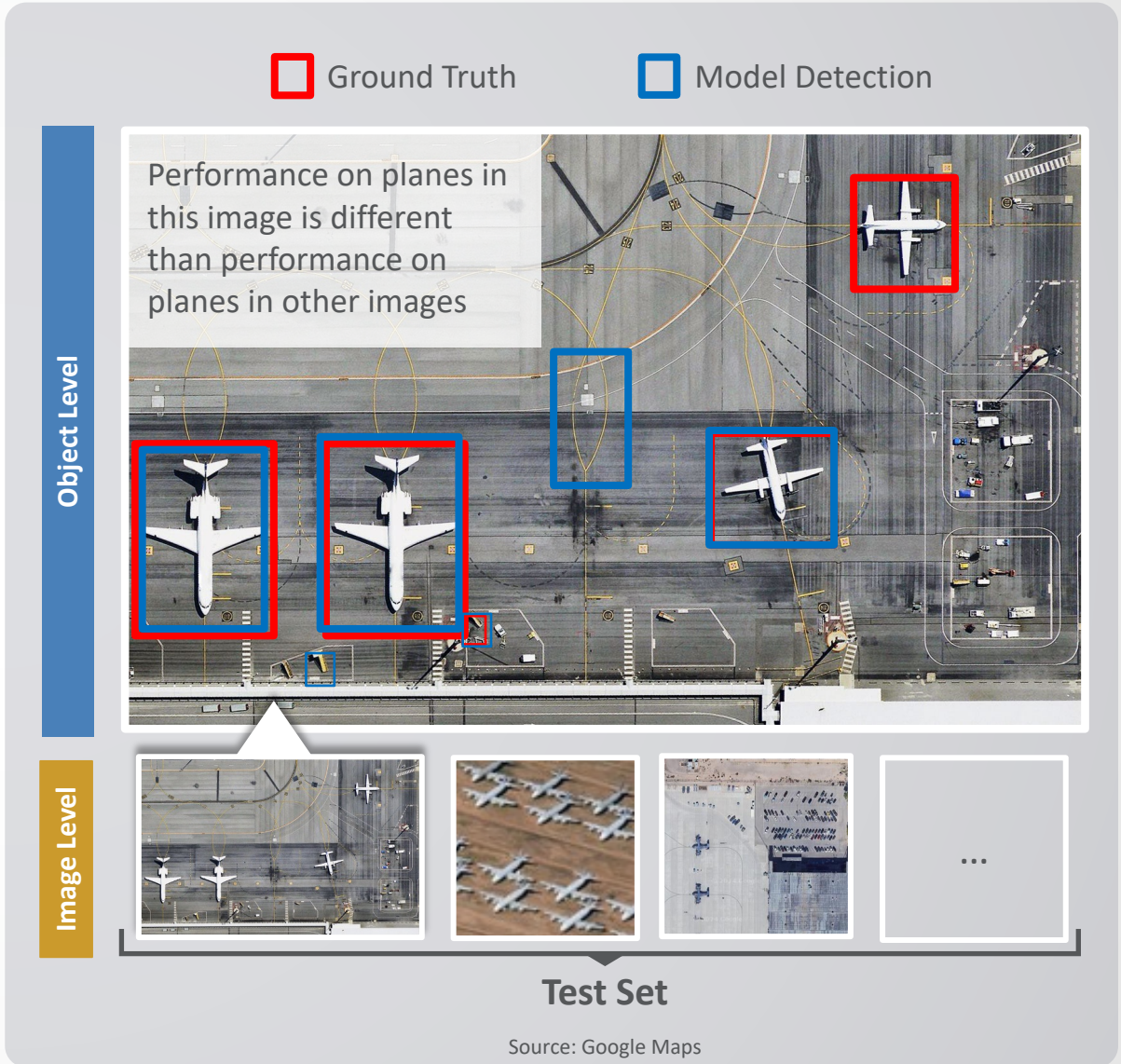


Test Set

Source: Google Maps

Hierarchical Data Structures

- Data in performance measurement applications often comes in some hierarchical structure
- Shared observing conditions induce data clusters with high intra-cluster correlation
- Define two levels in object detection data:
 - Image Level:** The collection of images on which AI performance is evaluated
 - Object Level:** The collection of ground truth labels and AI detections clustered by image



Object Detection in Correlated Data Clusters Experiment



To demonstrate how counting metrics interact with highly correlated data structures, we simulate object detection inference results and measure **recall**:

Simulated AI Inference Results

100 images
1862 ground truth labels

Image recall is driven by observing conditions varying between *Easy* and *Hard*

Observing conditions are constant *within* an image, inducing **correlated performance at the object level**



Image ground truth label count shown against fraction of those detected. Image label density displayed in top plot.

Object Detection in Correlated Data Clusters Experiment



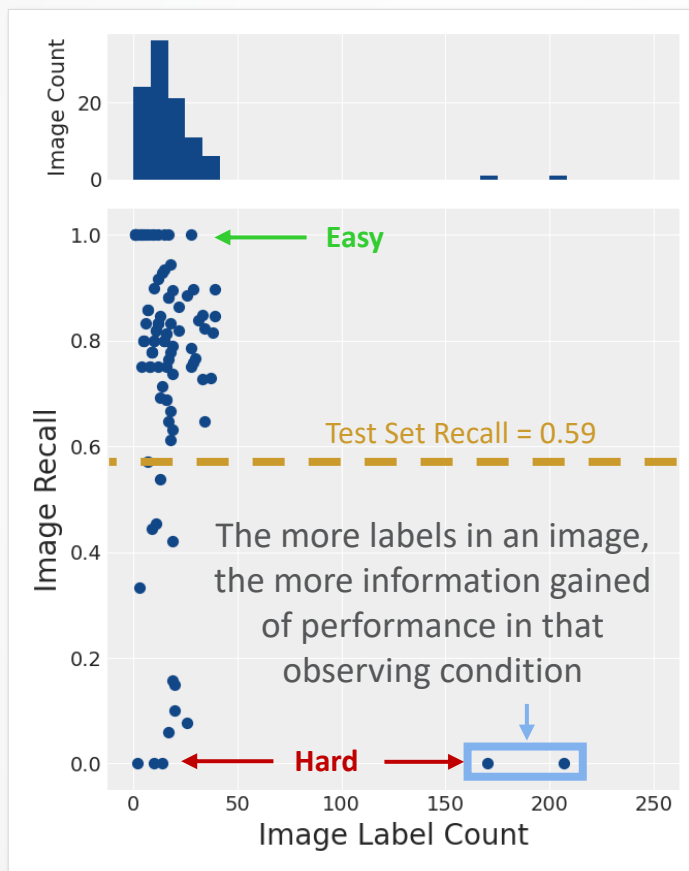
To demonstrate how counting metrics interact with highly correlated data structures, we simulate object detection inference results and measure **recall**:

Simulated AI Inference Results

100 images
1862 ground truth labels

Image recall is driven by observing conditions varying between *Easy* and *Hard*

Observing conditions are constant *within* an image, inducing **correlated performance at the object level**



Synthetic Inference Results

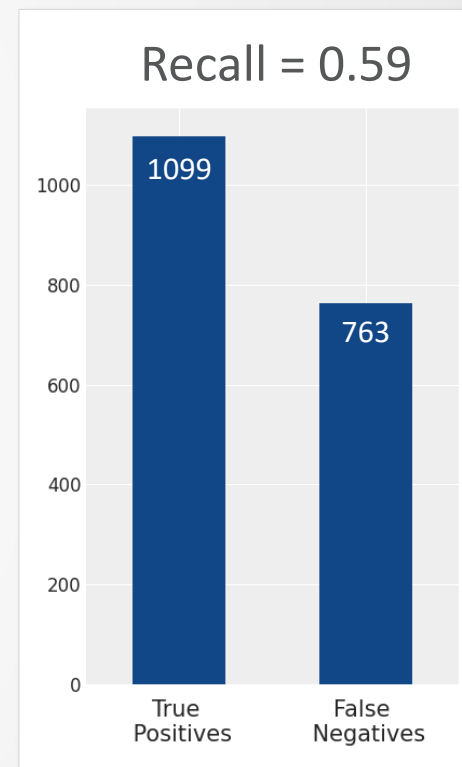
Image ground truth label count shown against fraction of those detected. Image label density displayed in top plot.

Test and Evaluation

Compute a naïve recall:

Given ground truth label j , of n labels, indicators $TP_j = 1$ and $FN_j = 0$ denote a detection and missed detection, respectively:

$$\text{Recall} = \frac{\sum_{j=1}^n TP_j}{\sum_{j=1}^n TP_j + \sum_{j=1}^n FN_j}$$



Count of how many ground truth objects were detected (True Positives) and not detected (False Negatives).

Recall Perspective at the Image Level

Given image i , of m images, with ground truth label counts n_i and true positives y_i

$$\text{Recall} = \frac{\sum_{j=1}^n \text{TP}_j}{\sum_{j=1}^n \text{TP}_j + \sum_{j=1}^n \text{FN}_j} = \frac{1}{n} \sum_{i=1}^m \left[n_i \times \frac{y_i}{n_i} \right]$$

We can express recall as the average of individual image recalls (y_i/n_i) weighted by the number of labels within that image (i.e., how much information contained in that image)

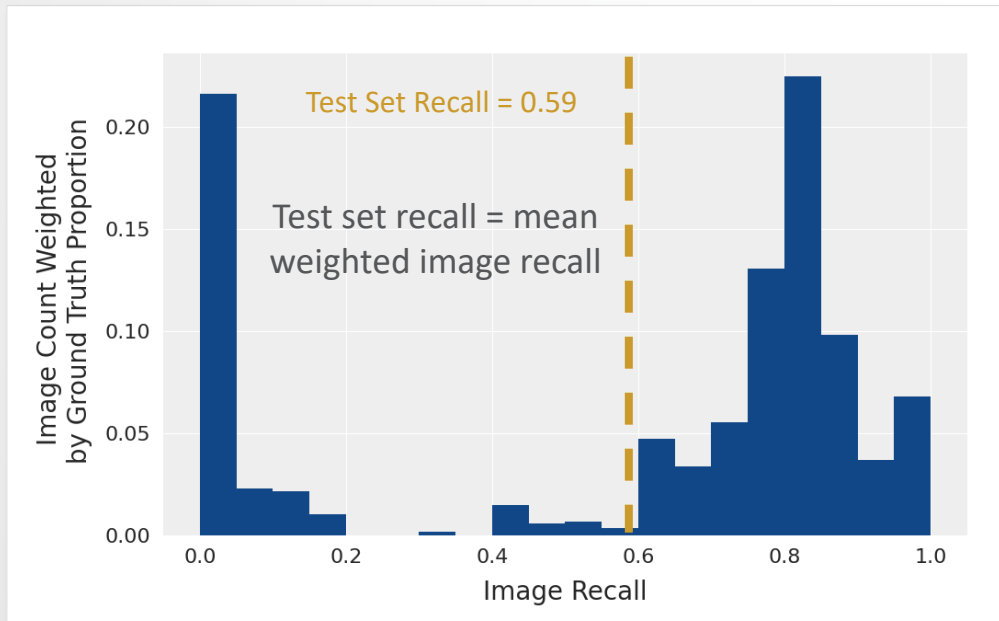


Image recall distribution weighted by ground truth label count.

Recall Perspective at the Image Level

Given image i , of m images, with ground truth label counts n_i and true positives y_i

$$\text{Recall} = \frac{\sum_{j=1}^n \text{TP}_j}{\sum_{j=1}^n \text{TP}_j + \sum_{j=1}^n \text{FN}_j} = \frac{1}{n} \sum_{i=1}^m \left[n_i \times \frac{y_i}{n_i} \right] \rightarrow \frac{1}{m} \sum_{i=1}^m \frac{y_i}{n_i}$$

All performance per image should be independent of the number of targets in an image
Object Detection Assumption

We can express recall as the average of individual image recalls (y_i/n_i) weighted by the number of labels within that image (i.e., how much information contained in that image)

The appropriate performance metric is image-level recall because models in production operate on single images

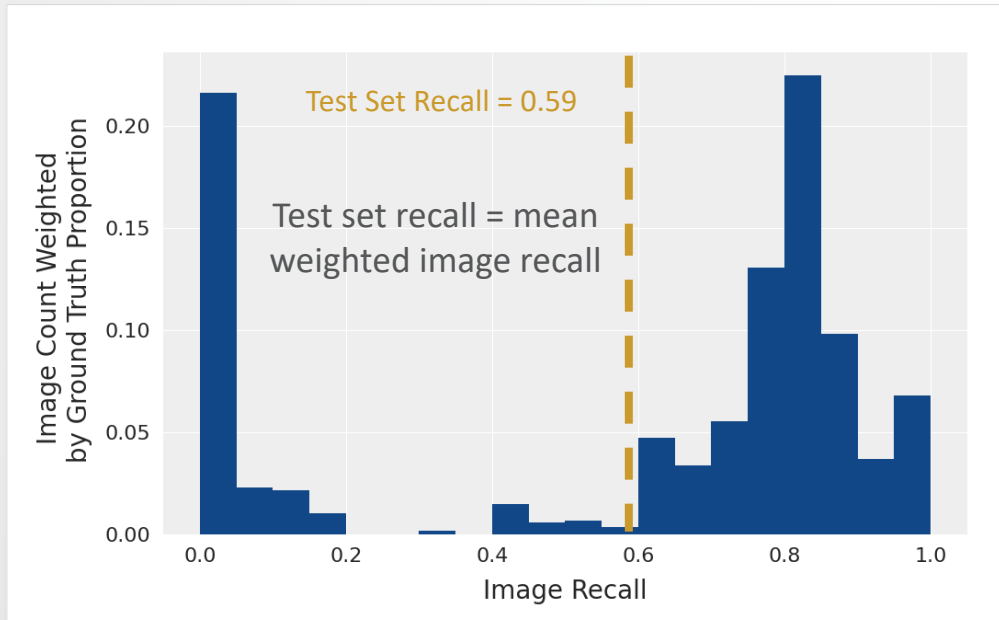
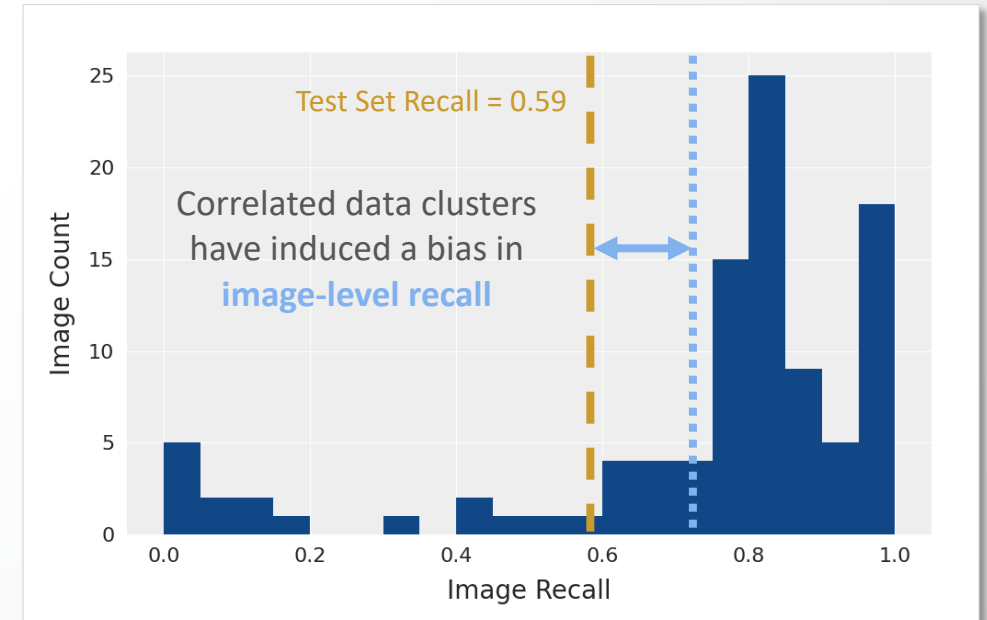


Image recall distribution weighted by ground truth label count.



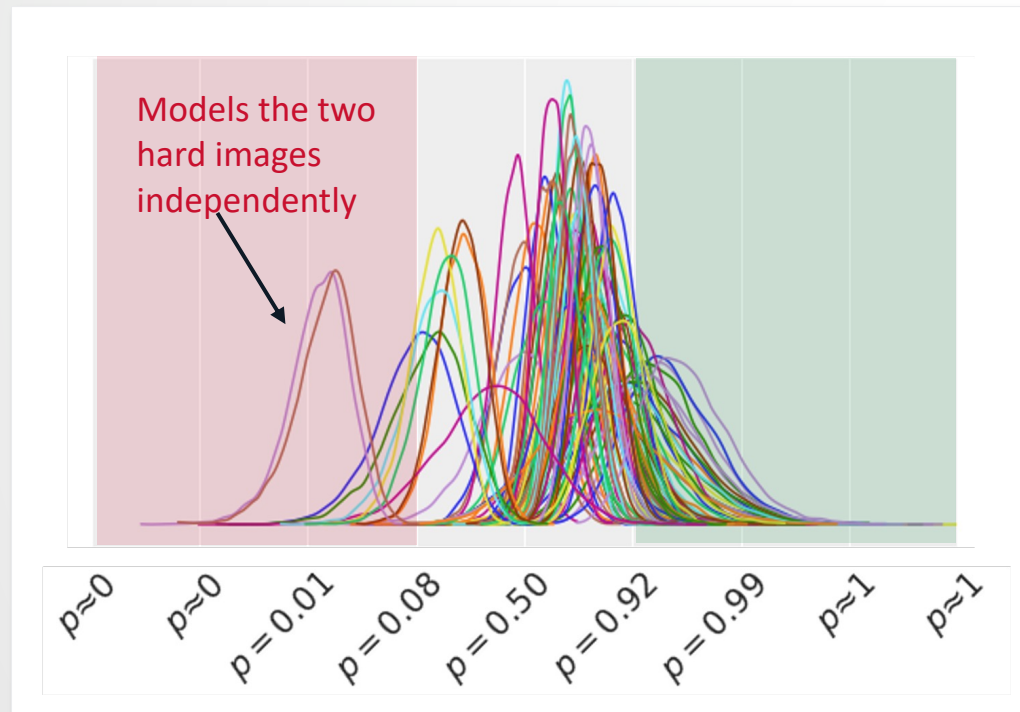
Unweighted image recall distribution.

Multilevel Bayesian Modeling

A multilevel Bayesian model estimates both image- and object-level AI model performance simultaneously by constraining image-specific parameters with pooled image-shared ones

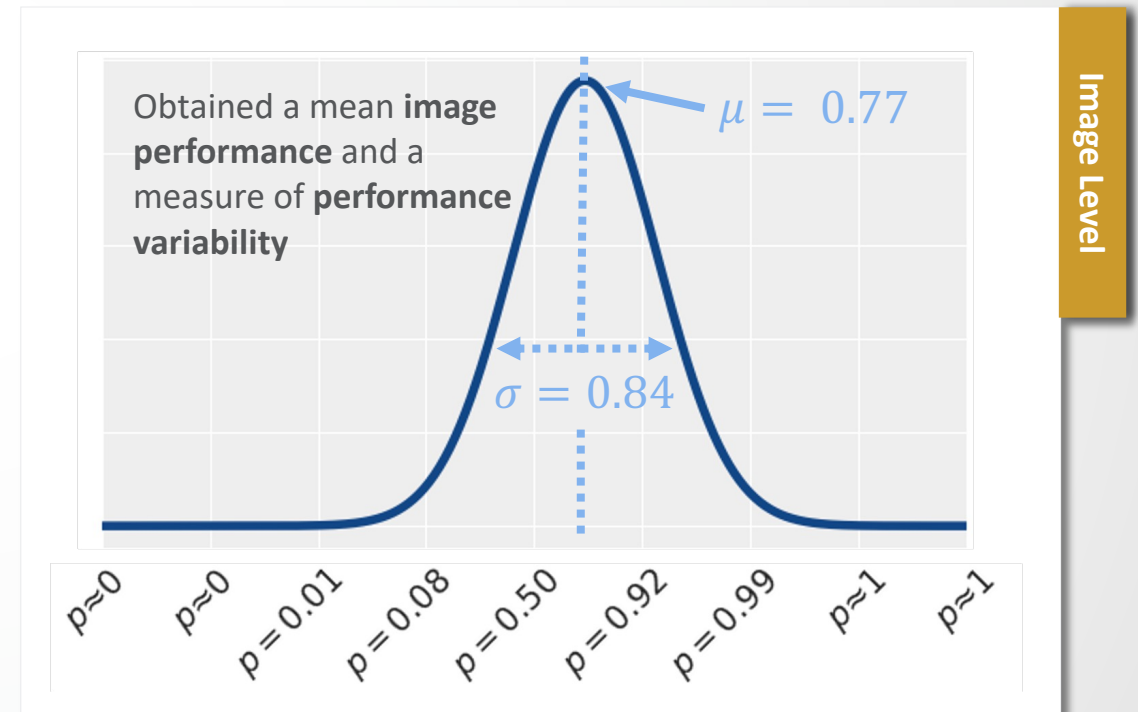
- Uses data efficiently to estimate performance
- Provides uncertainty estimate to performance metrics

Posteriors describe the AI model's performance on the **distribution of objects in each image**



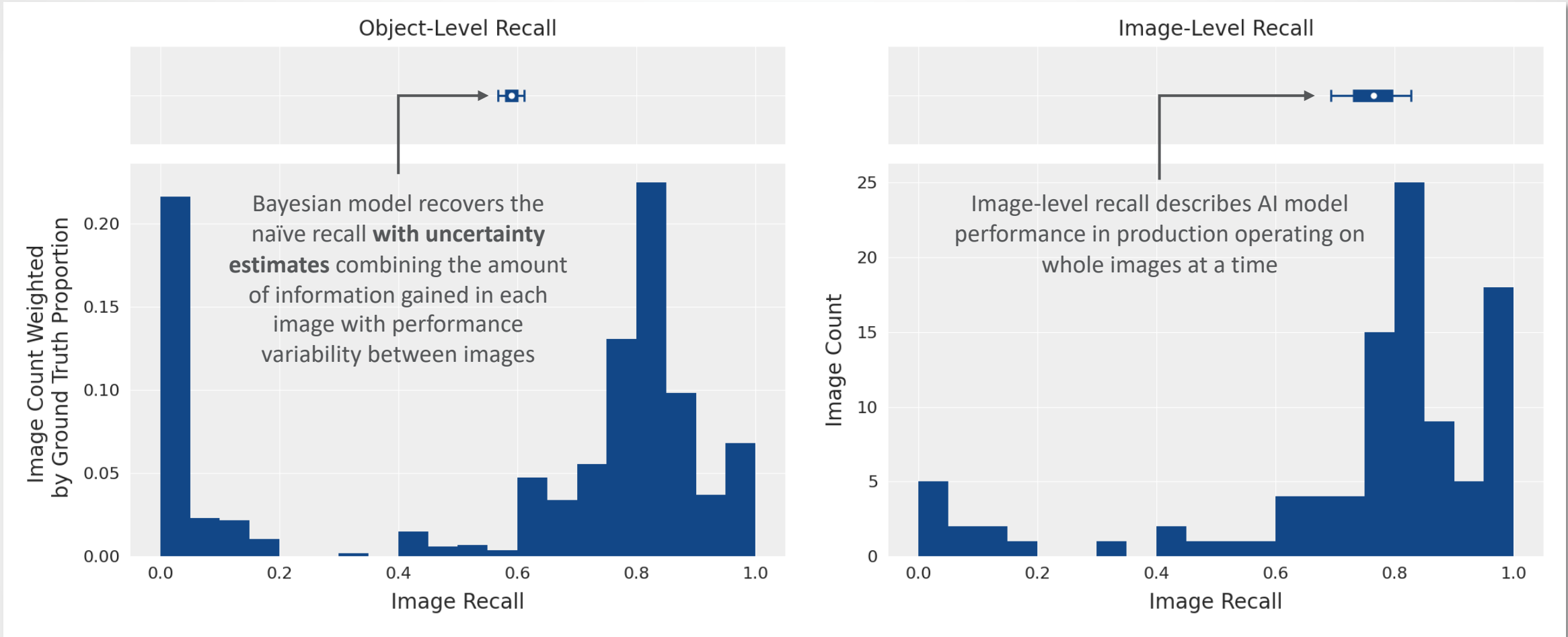
Posterior distributions for the recall of each image in the test set. Posteriors plotted in log-odds space, but the x-axis labels have been converted back to probability space for interpretability.

Posterior describes the AI model's performance on the **distribution of images**



Posterior distribution for the recall of each image in the test set. Posteriors plotted in log-odds space, but the x-axis labels have been converted back to probability space for interpretability.

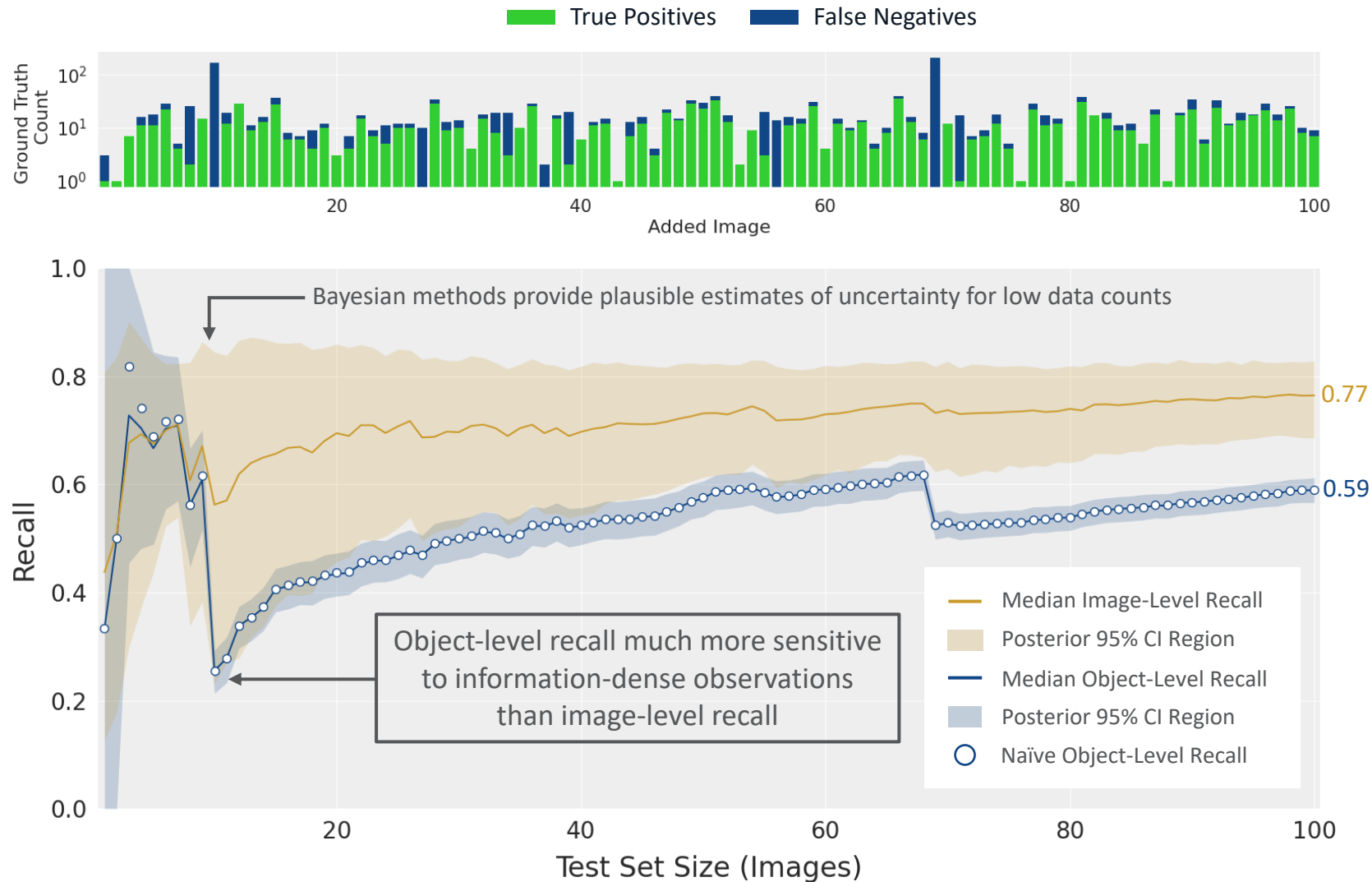
Image and Object Level Recall



Top: Test set naïve recall (point) with bound 68 and 95% credible intervals (bar and line, respectively) output from the posterior predictive distribution of the multilevel Bayesian model. **Bottom:** Image recall distribution weighted by ground truth label count.

Top: Test set mean image-level recall with bound 68 and 95% credible intervals (bar and line, respectively) from the multilevel Bayesian model. **Bottom:** Unweighted image recall distribution.

Robustness to Uneven Information Distributions



The image-level recall determined via a multilevel Bayesian model provides robust performance constraints in the face of an unbalanced distribution of information across different observing conditions

Top: Stacked histogram (in log-scale) of the true positives and false negatives of each image in the synthetic test set inference results.

Bottom: Object- and image-level recall posterior median and credible interval (CI) for cumulative test sets.

Summary



- MORSE T&E needs to rank and select most performant object detection computer vision model
 - Performance estimate needs to align with production environment operating on whole images
 - Need to reframe simple counting metrics such as precision, recall, and F1 in terms of images not objects
- We use a Bayesian multilevel model to estimate performance
- This model estimates performance simultaneously within and across images
 - Estimate is robust to uneven distribution of information across images
 - Learns parameters common to and unique to all images
 - Uncertainty in performance estimate is native to the Bayesian framework
 - Estimates performance from small amounts of data

MORSE uses a multilevel Bayesian model that uses information efficiently and generates full uncertainty distribution to measure production-aligned CV model performance.