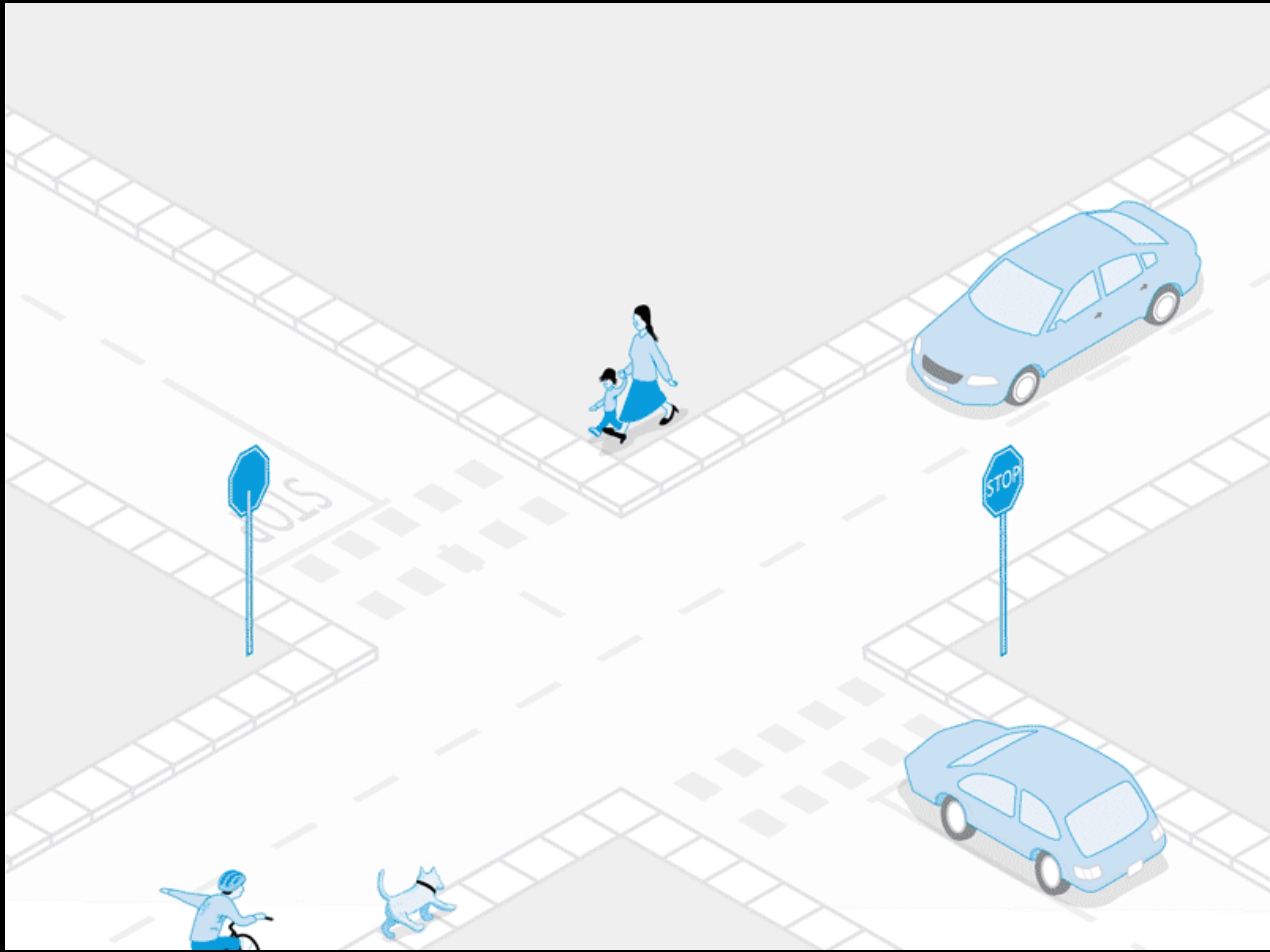
Ensembles

Accuracy and Calibration

Divyam Madaan, April 9, 2025

Make a decision

Neural network predicts “stop sign” with 95% confidence. Will you stop?



Ensembles

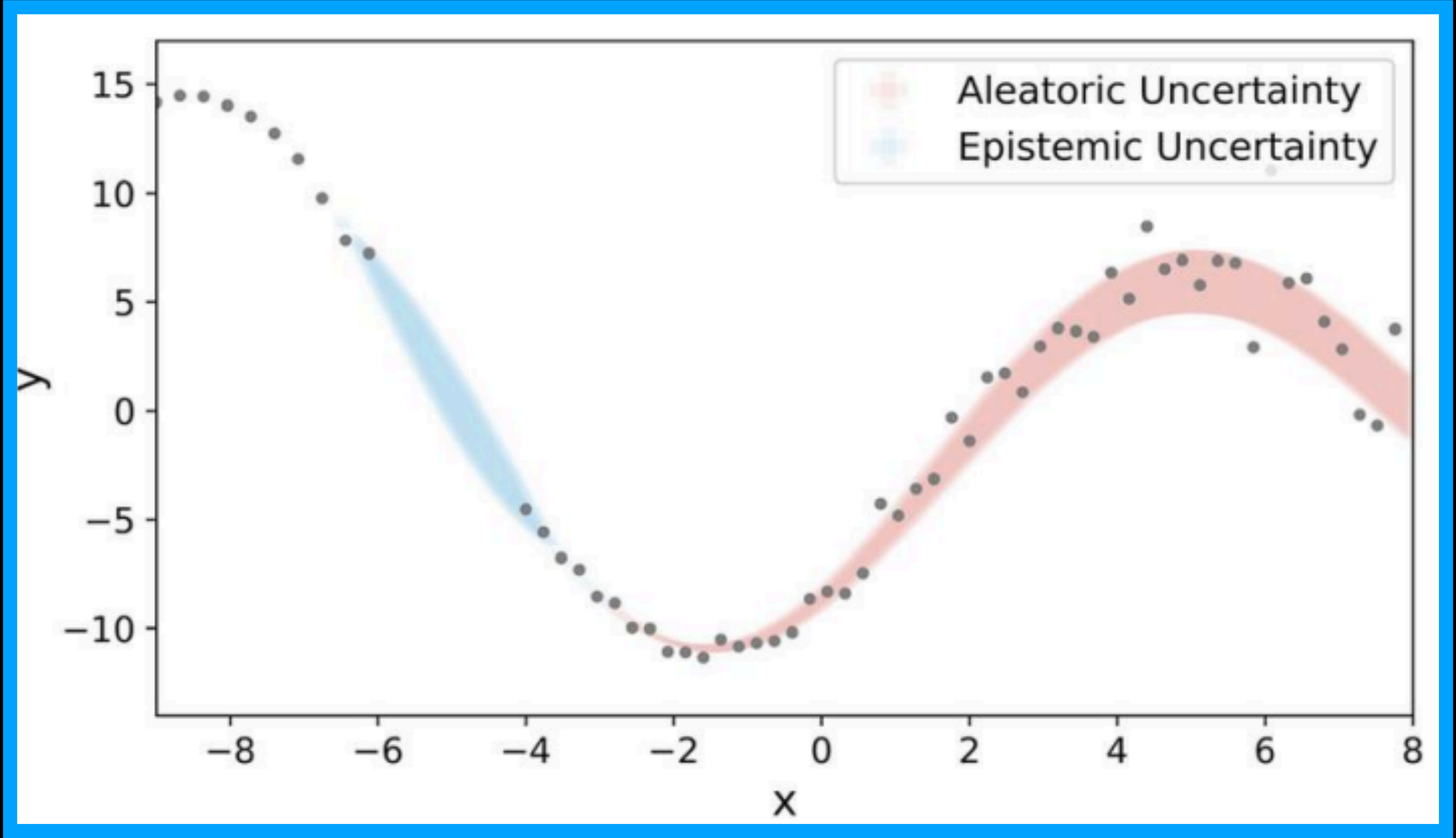
Given what we know from training, what's the best guess for a new x'?

$$p(y' | x', D, \beta) = \int \underbrace{p(y' \mid x', \theta, \beta)}_{\text{aleatoric}} \; \underbrace{q(\theta \mid D, \beta)}_{\text{epistemic}} \; d\theta \approx \frac{1}{K} \sum_{k=1}^{S} p(y' | x', \theta_k, \beta)$$

Aleatoric uncertainty — irreducible, stems from our data

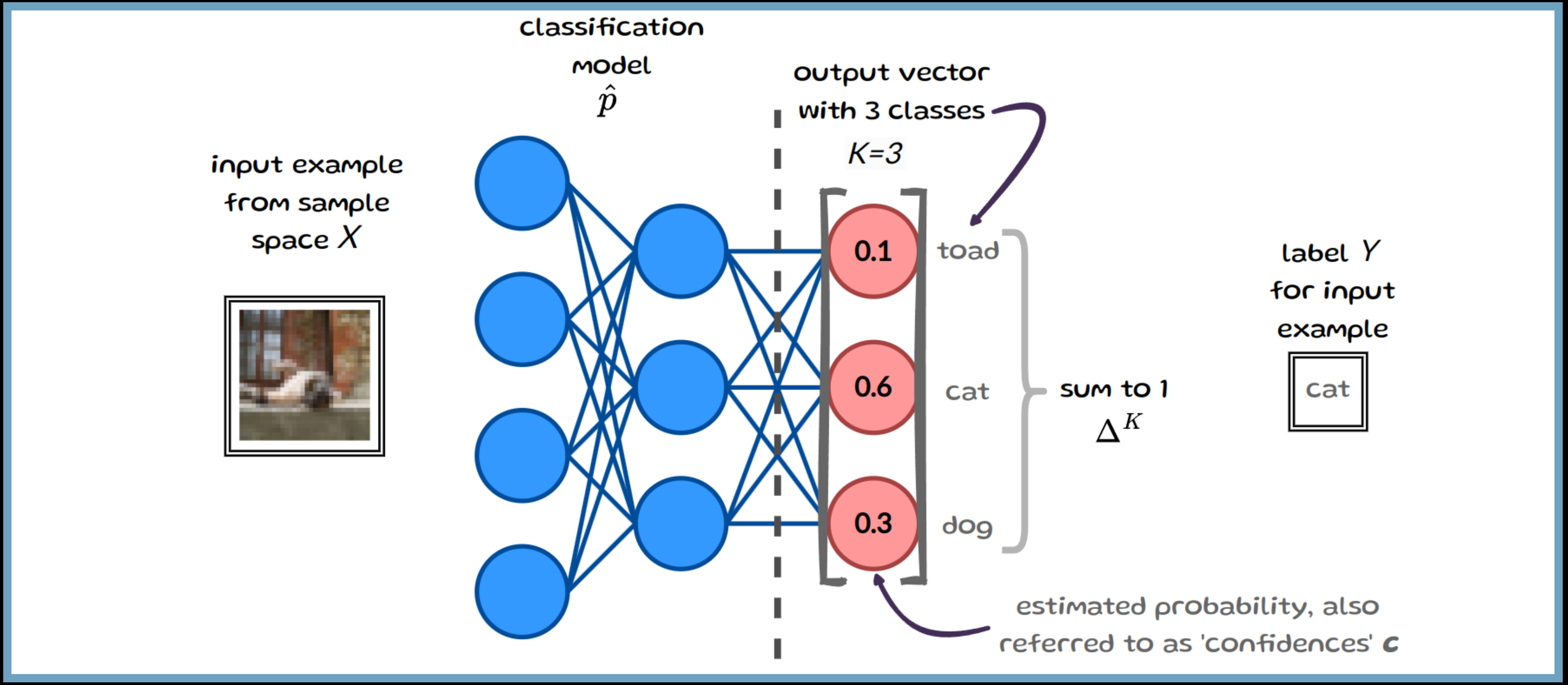
Epistemic — reducible, stems from our model

How to measure uncertainty?



How to measure uncertainty?

When the model is **c confident**, actual probability it is **correct** should also be **c**.



If the model predicts 100 samples with **confidence 0.8**, and is right on **80** of them, it is well-calibrated at **c = 0.8**

$$\mathbb{P}(Y = \arg\max(\hat{p}(X)) \mid \max(\hat{p}(X)) = c) = c \quad \forall c \in [0,1]$$

Expected Calibration Error

Weighted average over the absolute difference between *acc* and *confidence*.

Sample (i)	Estimated probabilities ($\hat{p}_i$)			Predicted Label ($\hat{y}_i$)	True Label (y_i)
	Class=C	Class=D	Class=T		
1	0.78	0.12	0.1	C	C
2	0.1	0.64	0.26	D	D
3	0.04	0.04	0.92	T	D
4	0.58	0.3	0.12	C	C
5	0.05	0.51	0.44	D	C
6	0.85	0.15	0	C	C
7	0.22	0.7	0.08	D	D
8	0.63	0.34	0.03	C	T
9	0.02	0.15	0.83	T	T

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \text{acc}(B_m) - \text{conf}(B_m) \right|$$

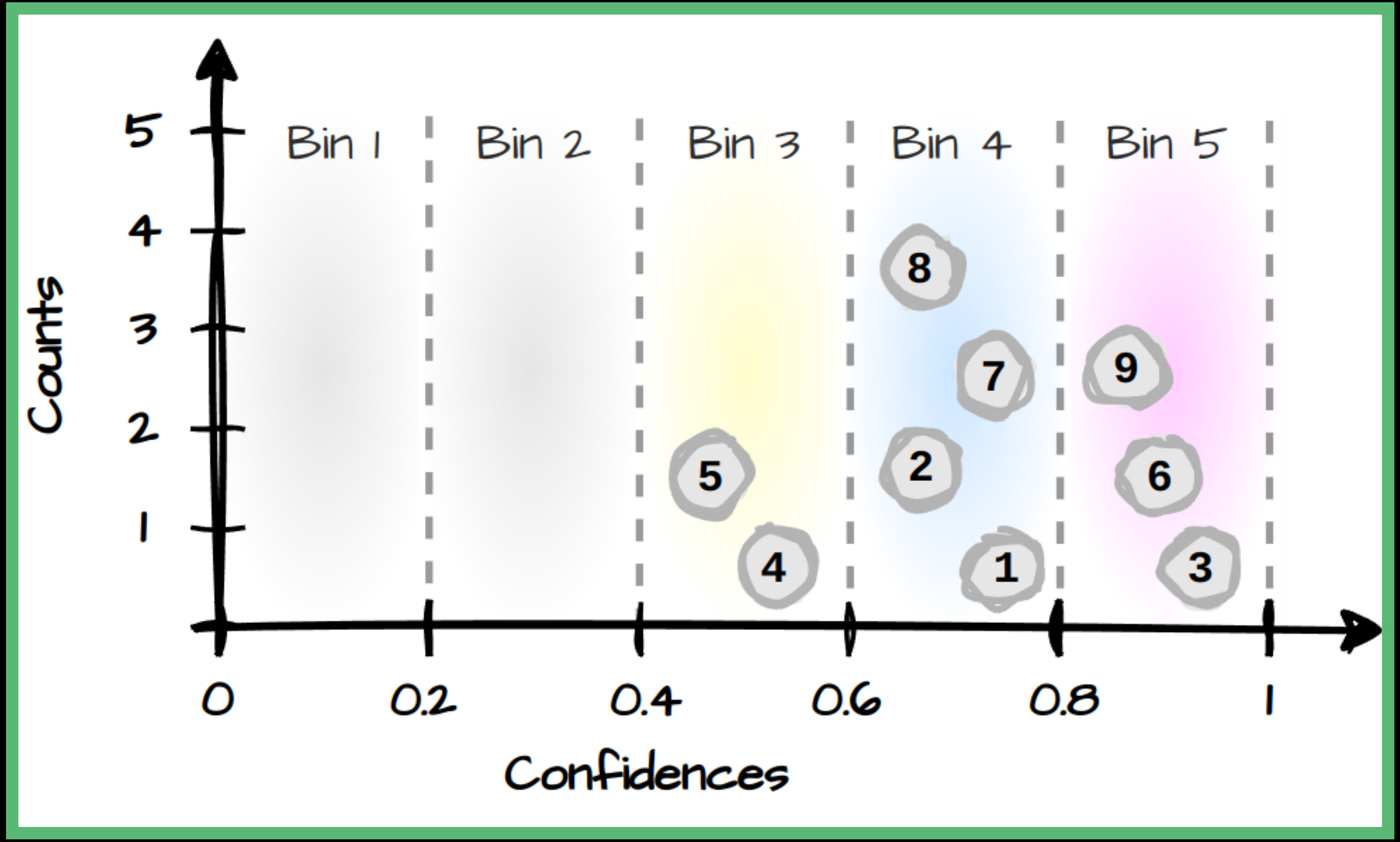
$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i)$$

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}(x_i)$$

Expected Calibration Error

Step 1: Bin samples based on the maximum probability across classes

Sample (i)	Max estimated probabilities ($\hat{p}_i$)	Predicted Label ($\hat{y}_i$)	True Label (y_i)
1	0.78	C	C
2	0.64	D	D
3	0.92	T	D
4	0.58	C	C
5	0.51	D	C
6	0.85	C	C
7	0.7	D	D
8	0.63	C	T
9	0.83	T	T



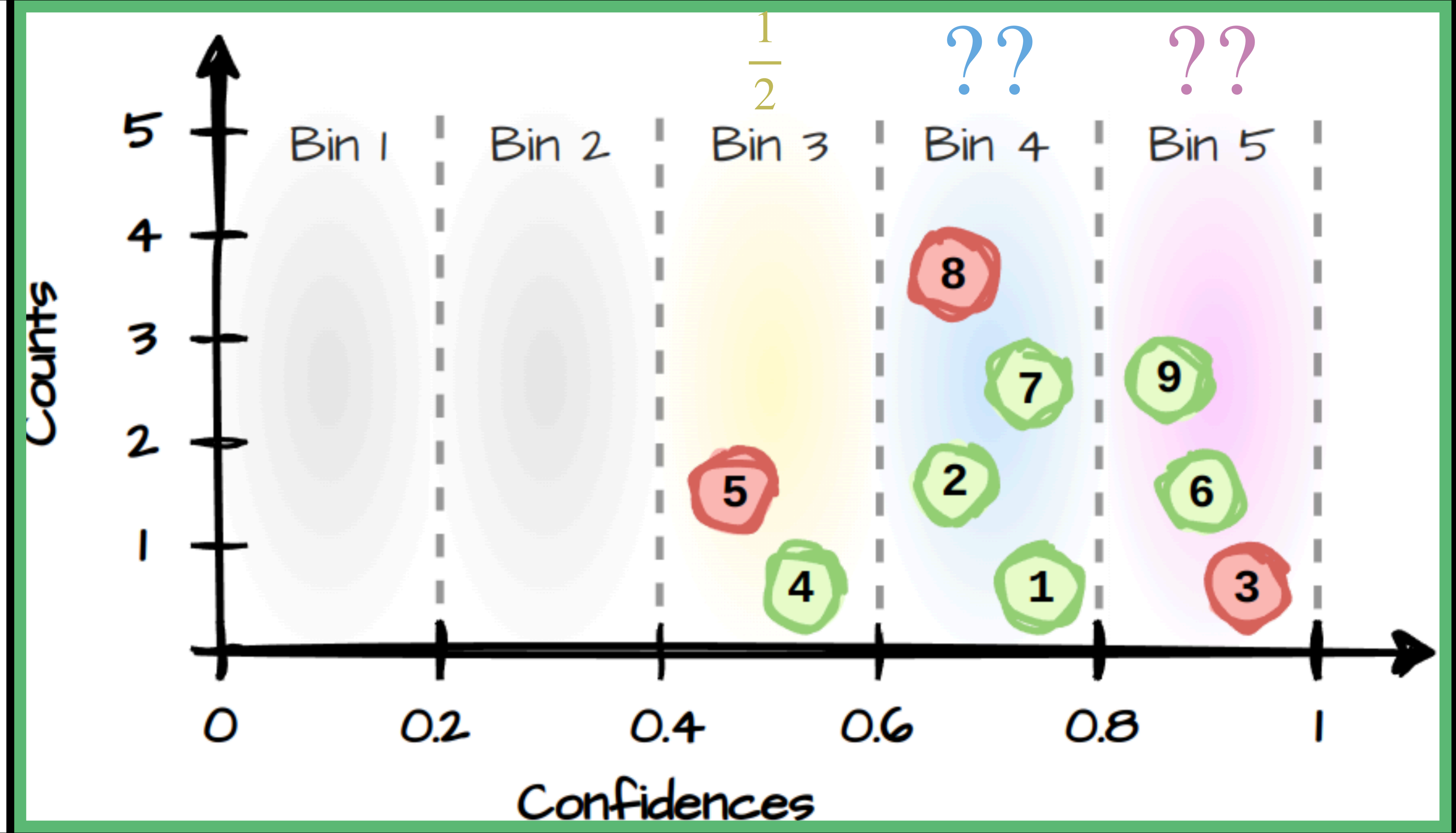
Example from Blogpost "Understanding Model Calibration: A Gentle Introduction & Visual Exploration"

Expected Calibration Error

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i)$$

Step 2: Accuracy is simply fraction of correctly predicted samples per bin

Sample (i)	Max estimated probabilities ($\hat{p}_i$)	Predicted Label ($\hat{y}_i$)	True Label (y_i)
1	0.78	C	C
2	0.64	D	D
3	0.92	T	D
4	0.58	C	C
5	0.51	D	C
6	0.85	C	C
7	0.7	D	D
8	0.63	C	T
9	0.83	T	T

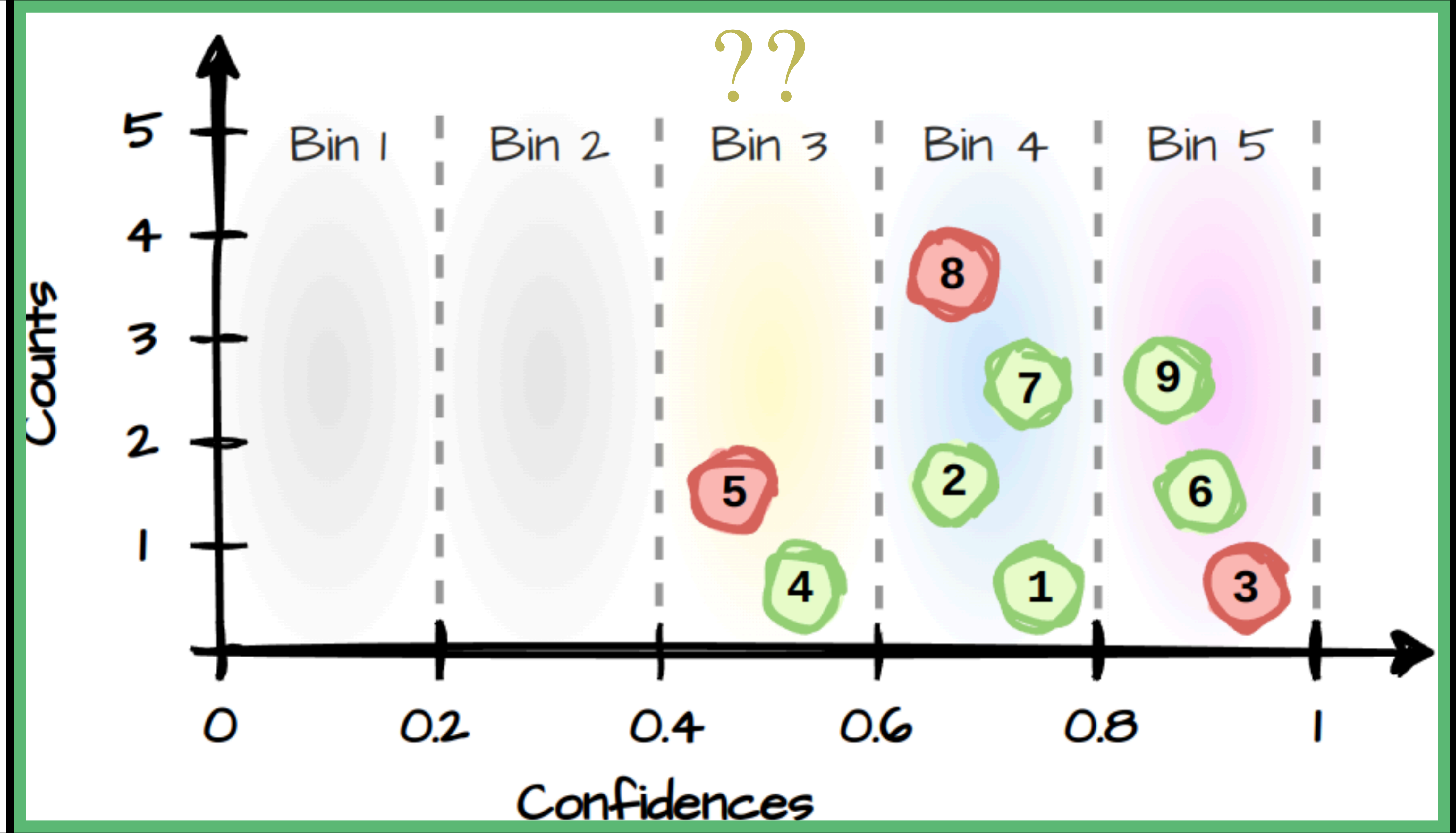


Expected Calibration Error

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}(x_i)$$

Step 3: Confidence is simply average of maximum estimated probabilities $\hat{\boldsymbol{p}}_{\boldsymbol{i}}$ per bin

Sample (i)	Max estimated probabilities ($\hat{p}_i$)	Predicted Label ($\hat{y}_i$)	True Label (y_i)
1	0.78	C	C
2	0.64	D	D
3	0.92	T	D
4	0.58	C	C
5	0.51	D	C
6	0.85	C	C
7	0.7	D	D
8	0.63	C	T
9	0.83	T	T

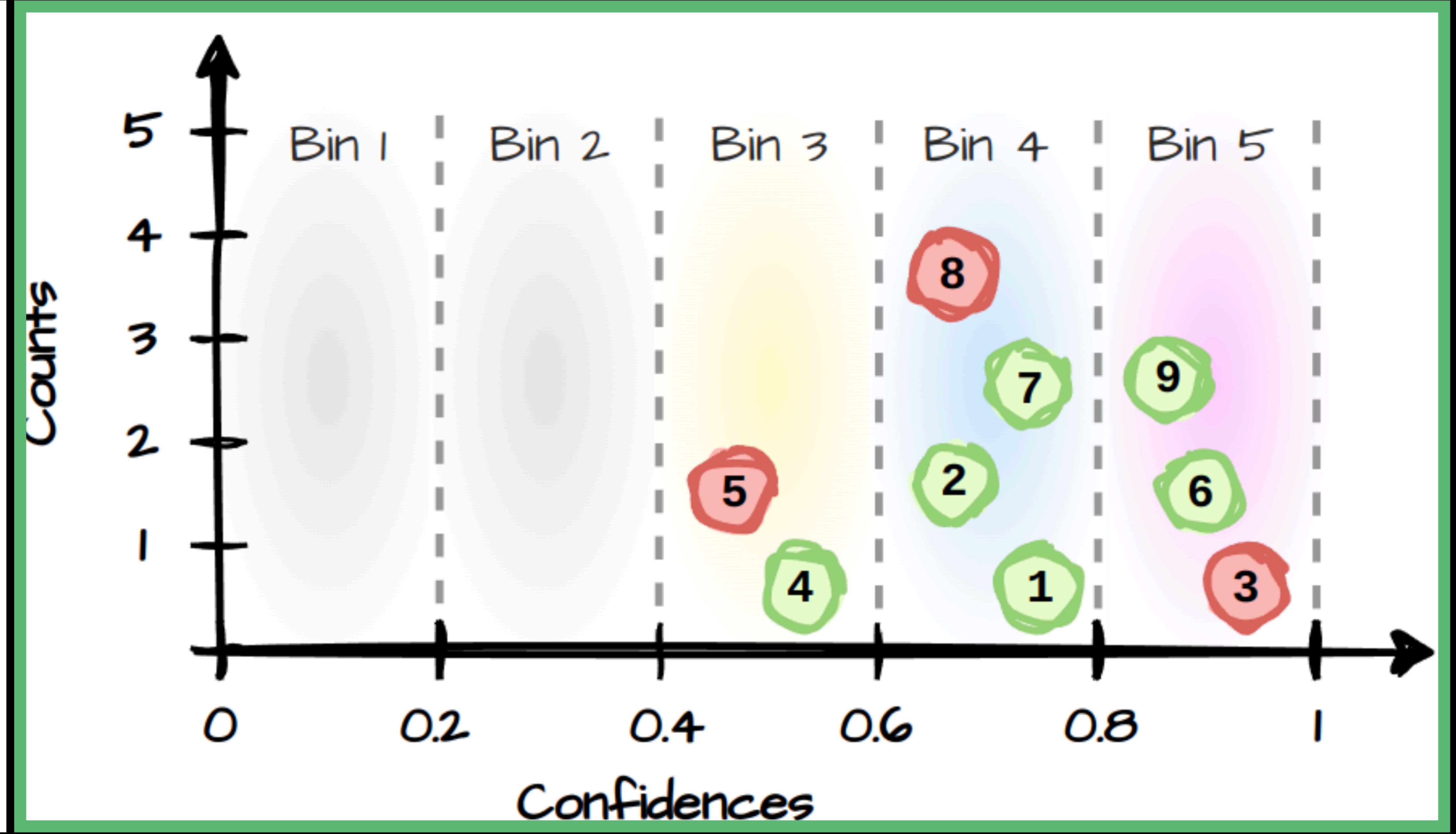


Expected Calibration Error

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \text{acc}(B_m) - \text{conf}(B_m) \right|$$

Substituting the values

Sample (i)	Max estimated probabilities ($\hat{p}_i$)	Predicted Label ($\hat{y}_i$)	True Label (y_i)
1	0.78	C	C
2	0.64	D	D
3	0.92	T	D
4	0.58	C	C
5	0.51	D	C
6	0.85	C	C
7	0.7	D	D
8	0.63	C	T
9	0.83	T	T



$$\text{ECE} = 0 + 0 + \frac{2}{9} \cdot \left| \frac{1}{2} - 0.545 \right| + \frac{4}{9} \cdot \left| \frac{3}{4} - 0.6875 \right| + \frac{3}{9} \cdot \left| \frac{2}{3} - 0.8667 \right| \approx 0.10445$$