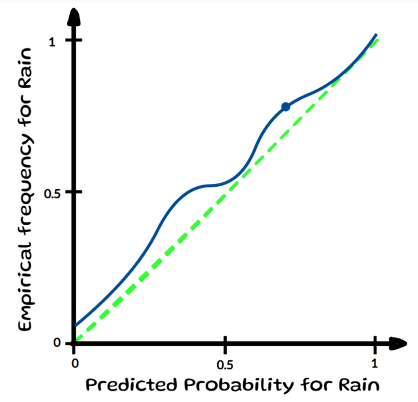
Beyond Accuracy:
Understanding Model Calibration – A gentle introduction

Maja Pavlovic

Why care about calibration?

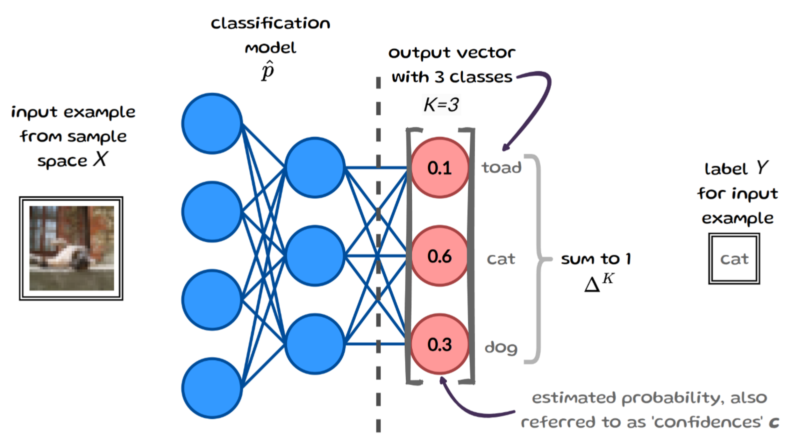
Calibration makes sure that a model's estimated probabilities match real-world likelihoods. For example, if a weather forecasting model predicts a 70% chance of rain on several days, then roughly 70% of those days should actually be rainy for the model to be considered well calibrated. This makes model predictions ***more reliable*** and ***trustworthy***, which makes calibration *relevant for many applications across various domains*.



What calibration means more precisely depends on the specific definition being considered...

But first, ***a bit of notation***:

We consider a classification task with K possible classes, with labels $Y \in \{1, \ldots, K\}$ and a classification model $\hat{p} : \mathscr{X} \rightarrow \Delta^K$, that takes inputs in $\mathscr{X}$ (e.g. an image or text) and returns a probability vector as its output. Δ^K refers to the K-simplex, which just means that the elements of the output vector must sum to 1 and that each estimated probability in the vector is between 0 & 1.
These individual probabilities (*or confidences*) indicate how likely an input belongs to each of the K classes.

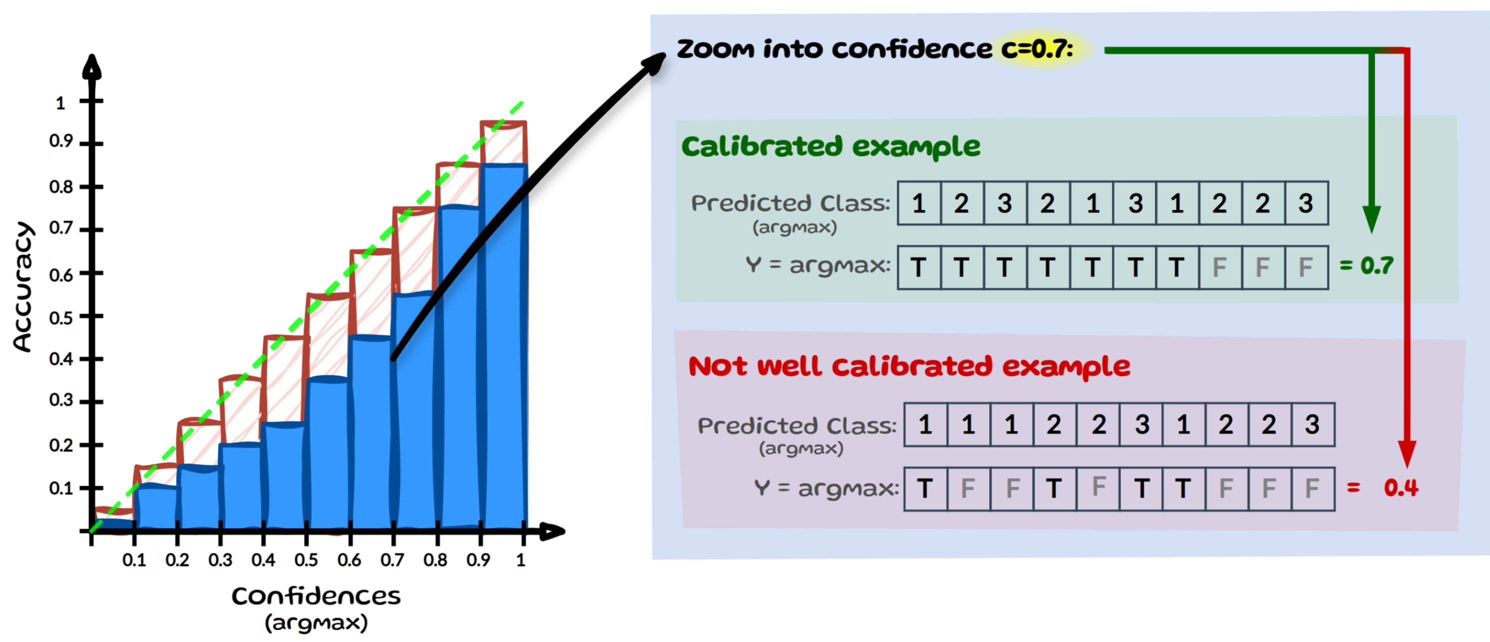


Confidence Calibrated

A model is considered confidence-calibrated if, for all confidences c the model is correct c proportion of the time:

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

where (X, Y) is a datapoint and $\hat{p} : \mathscr{X} \rightarrow \Delta^K$ returns a probability vector as its output.



Evaluating Calibration –> Expected Calibration Error (ECE)

One widely used evaluation measure for confidence calibration is the Expected Calibration Error (ECE). ECE measures how well a model's estimated probabilities match the observed probabilities by taking a *weighted average* over the ***absolute difference*** between ***average accuracy*** (acc) and ***average confidence*** (conf):

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|, \quad \text{where:} \quad acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \;\; \& \;\; conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}(x_i)$$

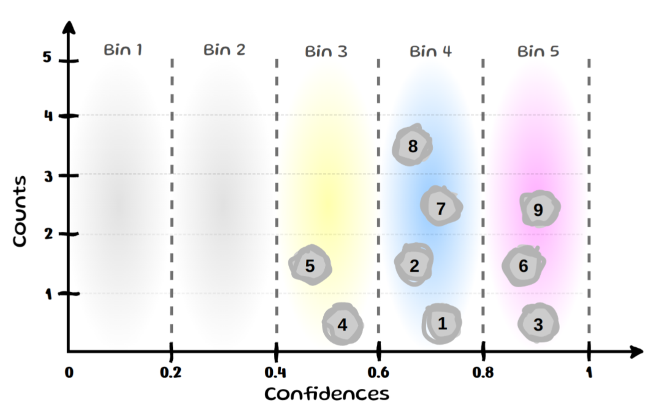
Simple step-by-step example:

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

The table we has **9** samples indexed by ***i*** with estimated probabilities $\hat{p}(x_i)$ (simplified as $\hat{p}_i$) for class **cat (C)**, **dog (D)** or **toad (T)**.

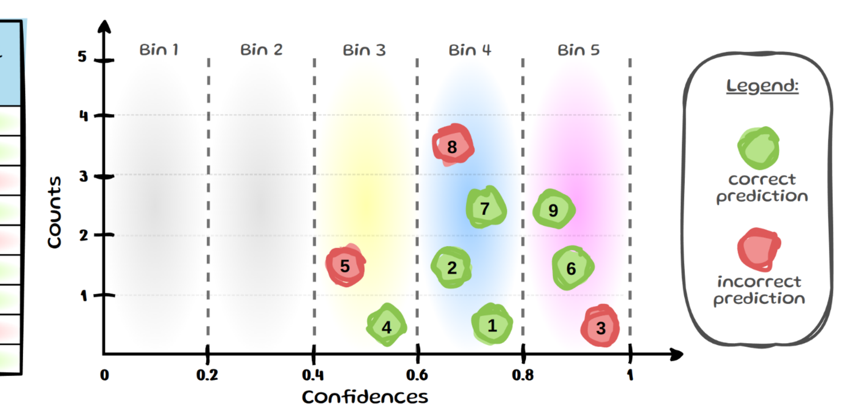
Only the maximum probabilities, which determine the predicted label are used in ECE, so we only keep those in the next table & bin samples based on the max. probabilities across classes.

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



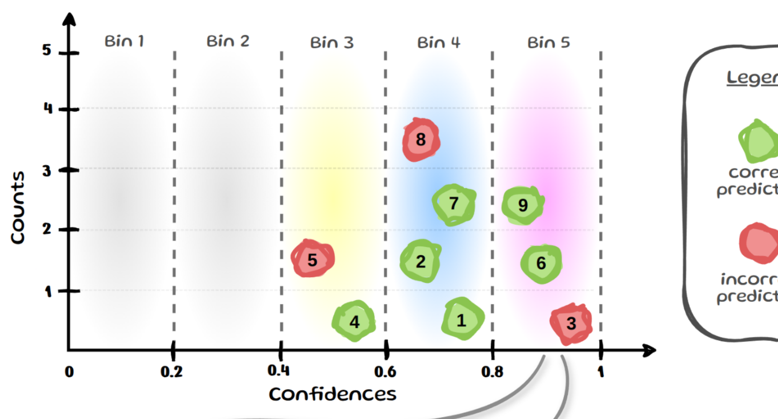
If the model predicts the class correctly, the prediction is highlighted in green; incorrect predictions are marked in red:

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



Below briefly runs through how to calculate the values for *bin 5* **(B5)**. The other bins then simply follow the same process:

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



(3) $conf(B_5) = \frac{0.92 + 0.85 + 0.83}{3} \approx 0.8667$

(1) $\frac{|B_5|}{n} = \frac{3}{9}$

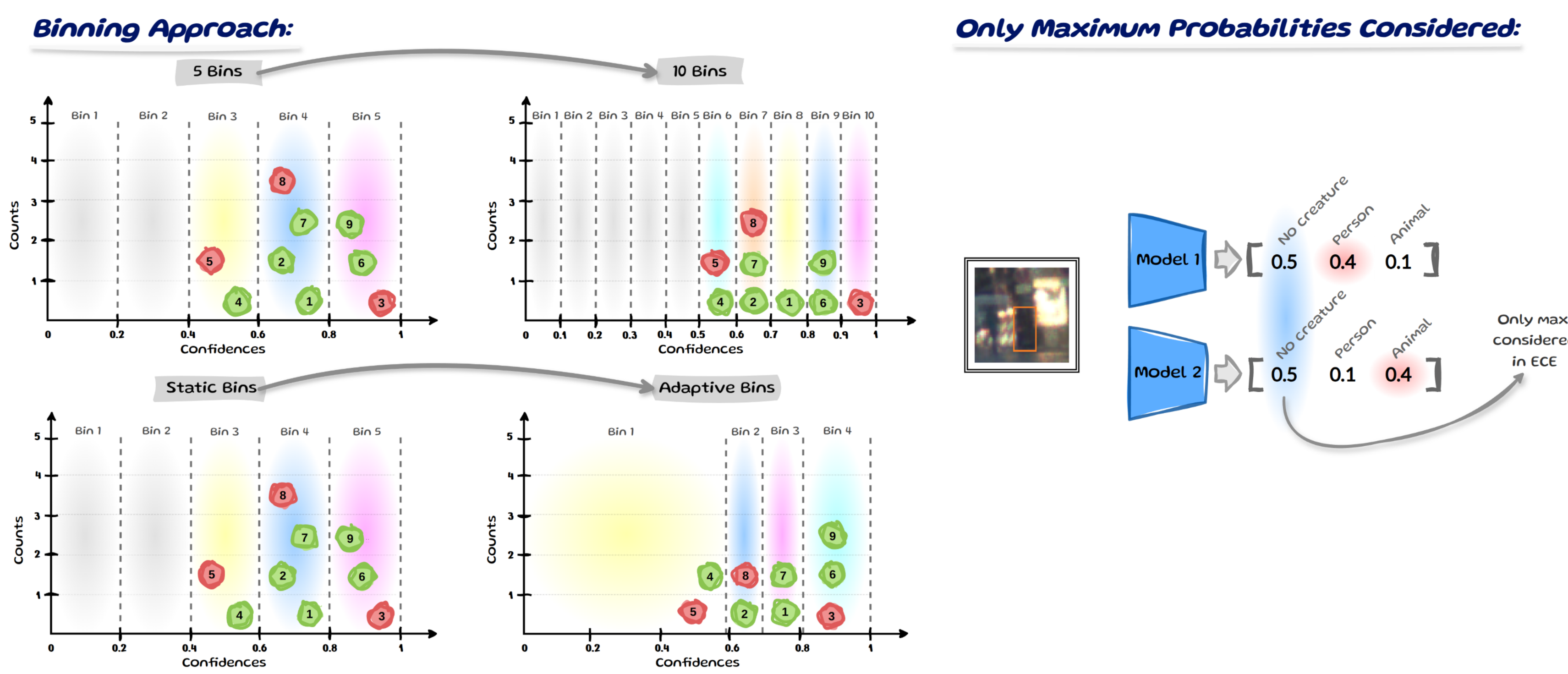
(2) $acc(B_5) = \frac{2}{3}$

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

Frequently Mentioned Drawbacks of ECE

Binning Approach:

Only Maximum Probabilities Considered:

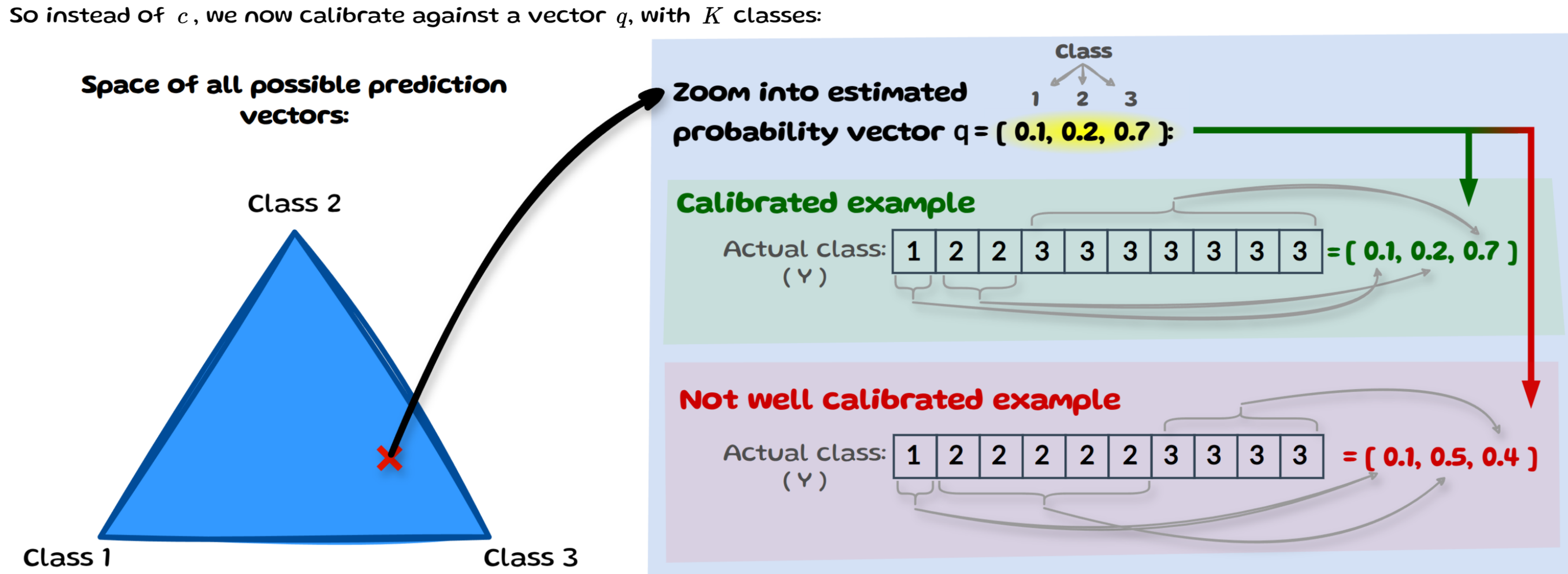


Multi-class Calibrated

A model is considered multi-class calibrated if, for any prediction vector $q = (q_1, \ldots, q_K) \in \Delta^K$, the class proportions among all values of X for which a model outputs the same prediction $\hat{p}(X)$ match the values in the prediction vector q .

$$\mathbb{P}(Y = k \mid \hat{p}(X) = q) = q_k \qquad \forall k \in \{1, \ldots, K\}, \; \forall q \in \Delta^K$$

So instead of c , we now calibrate against a vector q, with K classes:

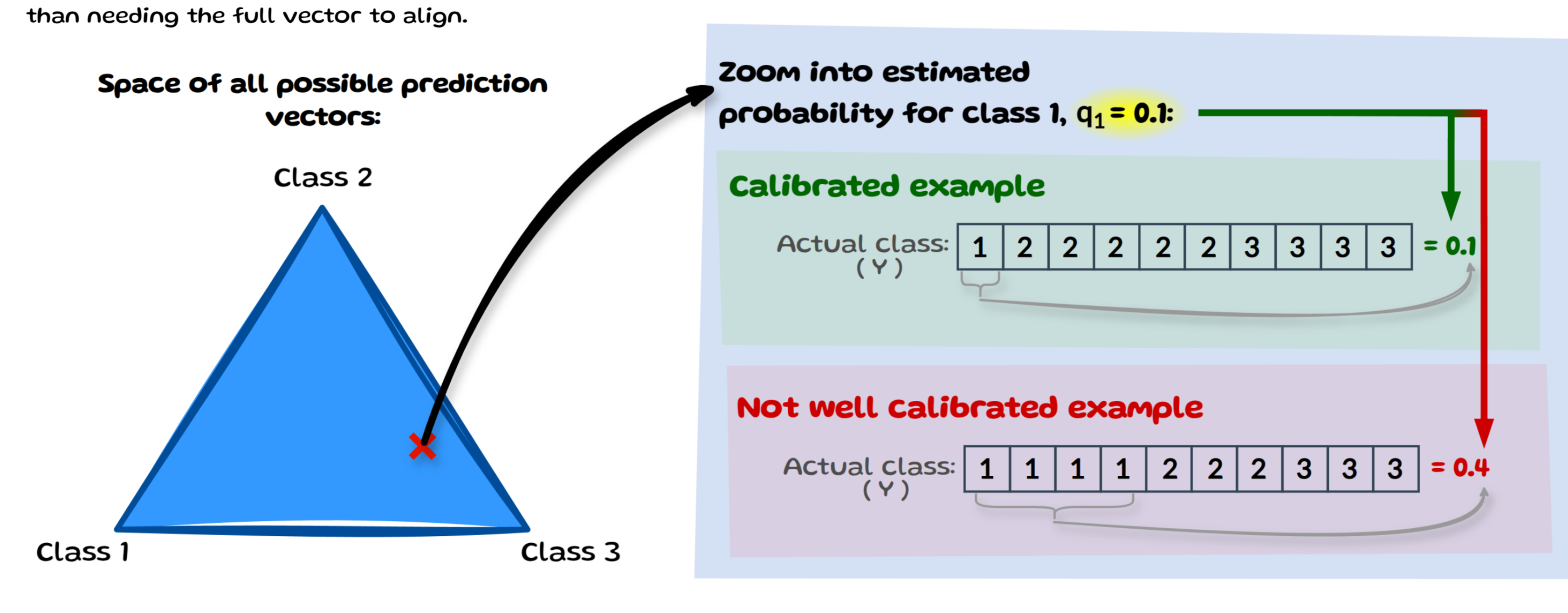


Class-wise Calibrated

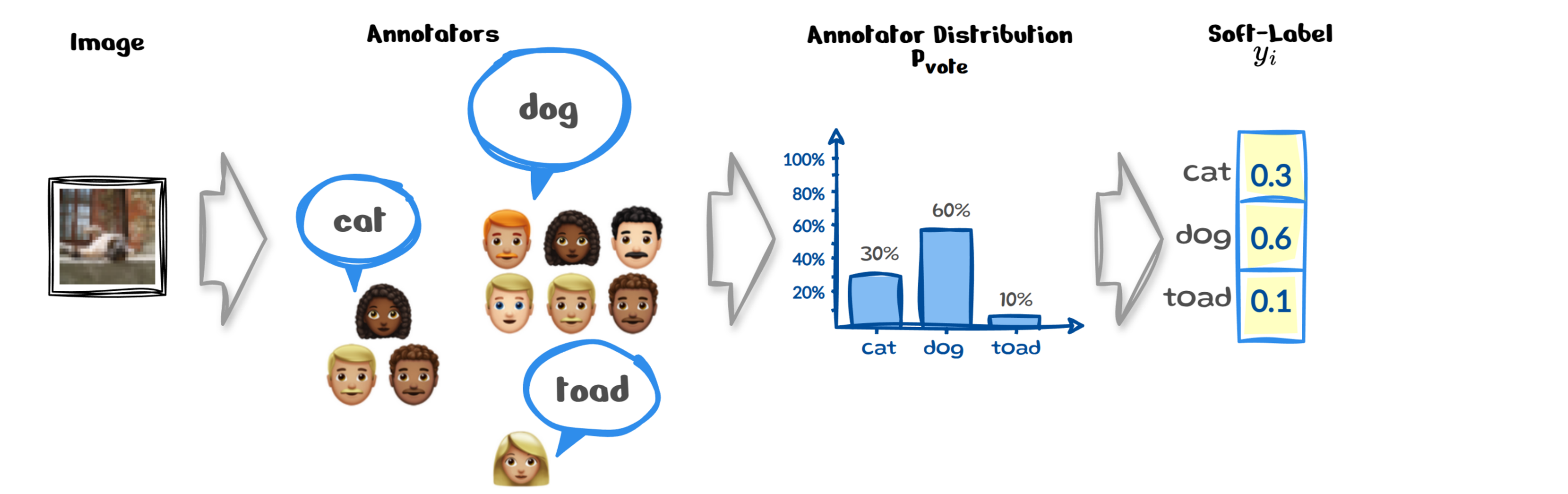
A model is considered class-wise calibrated if, for each class k, all inputs that share an estimated probability $\hat{p}_k(X)$ align with the true frequency of class k when considered on its own:

$$\mathbb{P}(Y = k \mid \hat{p}_k(X) = q_k) = q_k \qquad \forall k \in \{1, \ldots, K\}$$

Class-wise calibration is a ***weaker*** definition than ***multi-class calibration*** as it considers each class probability in ***isolation*** rather than needing the full vector to align.



*All approaches mentioned so far share a key assumption: **ground-truth labels are available,** BUT annotators might unresolvably and justifiably disagree on the real label*

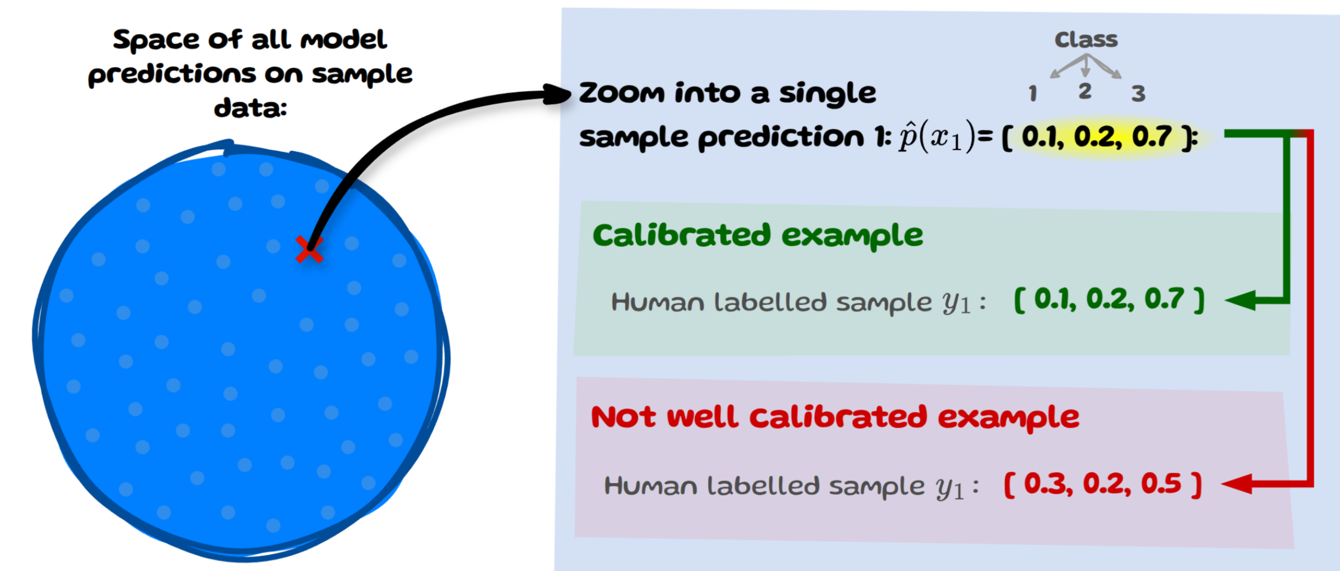


Human Uncertainty Calibrated

A model is considered human-uncertainty calibrated if, for each specific sample x , the predicted probability for each class k matches the '*actual*' probability P_{vote} of that class being correct.

$$\mathbb{P}_{vote}(Y = k \mid X = x) = \hat{p}_k(x) \qquad \forall k \in \{1, \ldots, K\}$$

This definition of calibration is more granular and strict than the previous ones as it applies directly at the level of individual predictions rather than being averaged or assessed over a set of samples.



Takeaways

- A model might have high accuracy but is it calibrated?
- Different notions of Model Calibration exist – determining which notion of calibration best fits a specific context and how to evaluate it should help avoid misleading results
- Despite several works arguing against the use of ECE for evaluating calibration, it remains widely used. Is ECE simply so easy, intuitive and just good enough for most applications that it is here to stay?

Find the blogpost or ArXiv:

ICLR Blogs

arXiv PDF

Follow up questions / contact:

@majapavlo.bsky.social

@majapavlo