Why a highly accurate model could be not the best[1]?

Mohamed Maher

Products Group*(SWDC)*, Giza Systems
https://www.menti.com/alxf2i73h12j (4786 8330)

Silva Filho, Telmo, et al. "Classifier calibration: a survey on how to assess and improve predicted class probabilities." *Machine Learning* 112.9 (2023): 3211-3260.

Background: Probabilistic Classification

Cats

Dogs

Background: Probabilistic Classification

ML Model is trained to separate between dogs and cats by creating a **<u>decision boundary</u>**

99.5% Accuracy

Cats

Dogs

Background: Probabilistic Classification

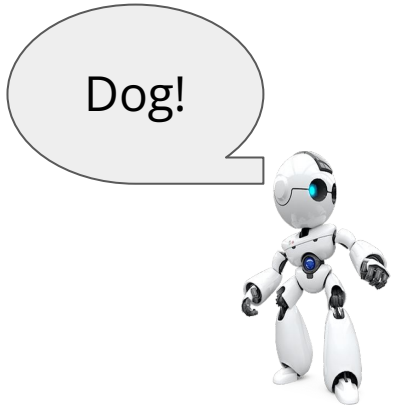
Input Image

ML Model

Cats

Dogs

Background: Probabilistic Classification



ML Model

Cats

Dogs

Background: Probabilistic Classification

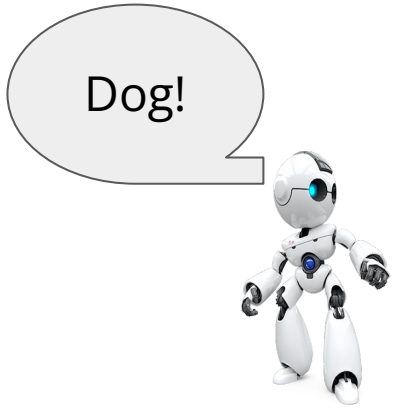
Input Image

ML Model

Cats

Dogs

Background: Probabilistic Classification

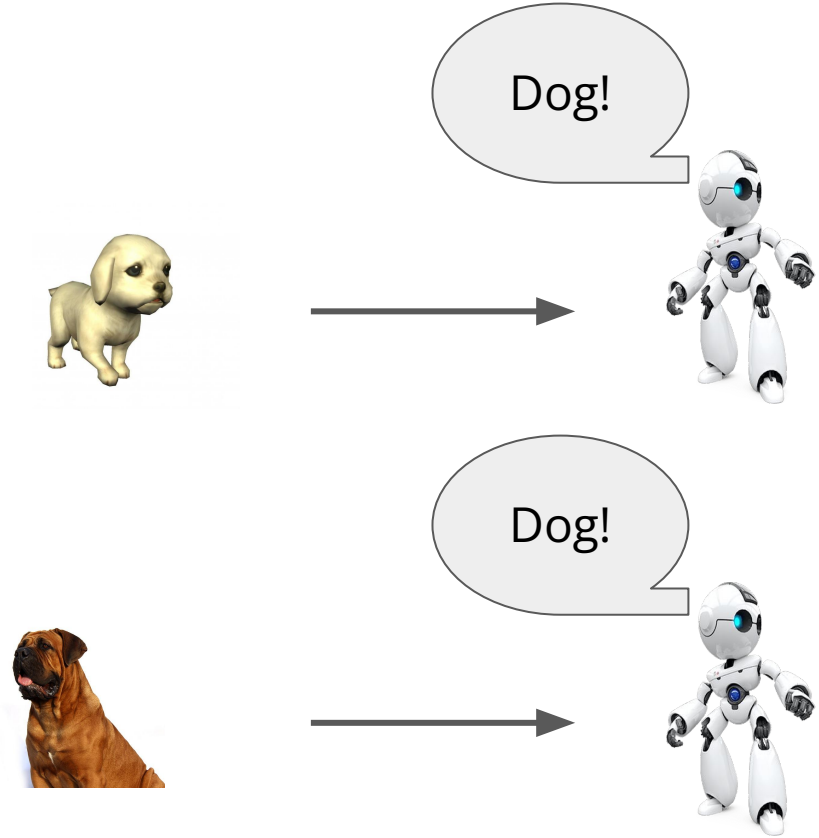


ML Model

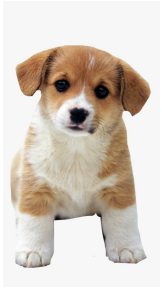
Cats

Dogs

Background: Probabilistic Classification



Background: Probabilistic Classification



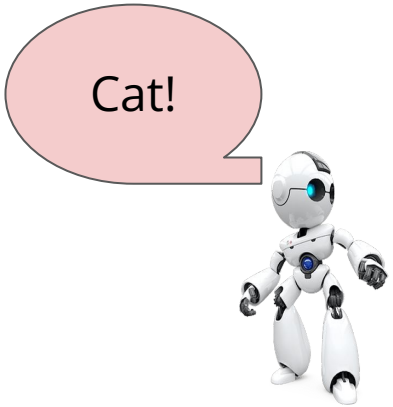
Input Image

ML Model

Cats

Dogs

Background: Probabilistic Classification

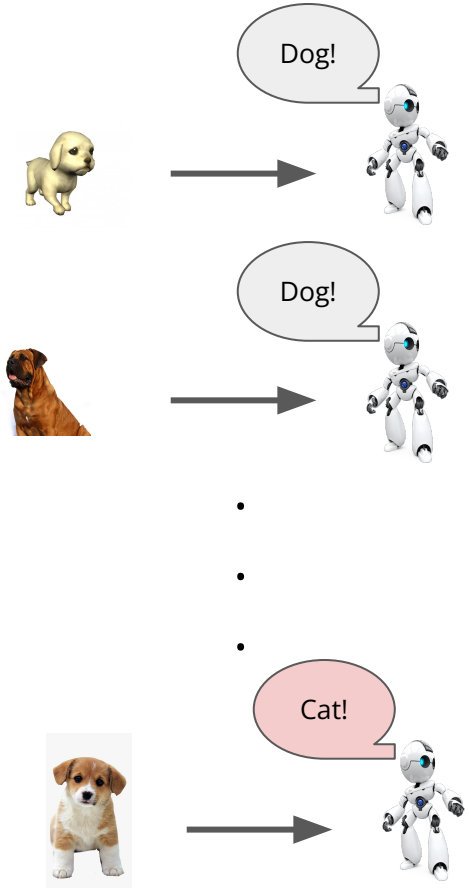


ML Model

Cats

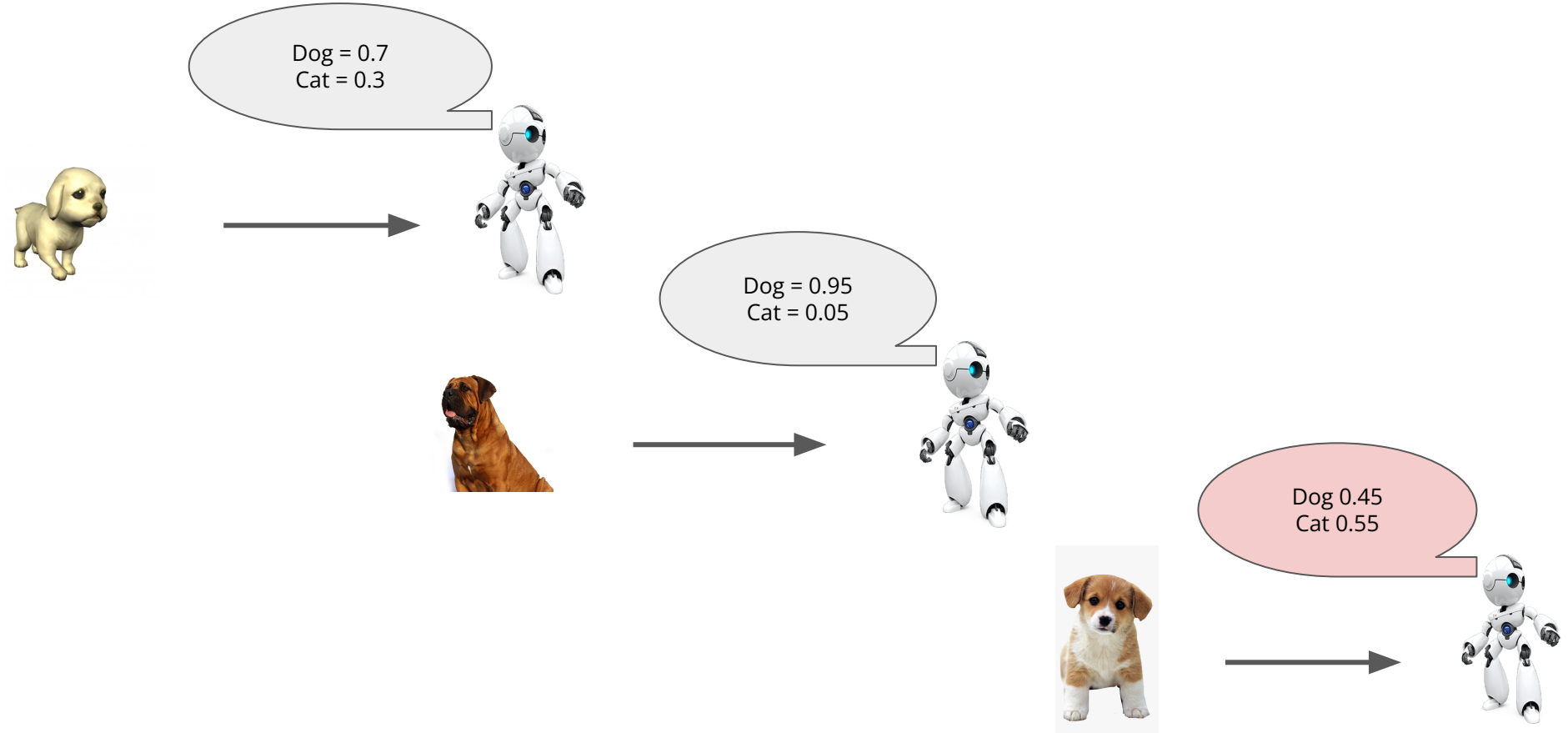
Dogs

Background: Probabilistic Classification



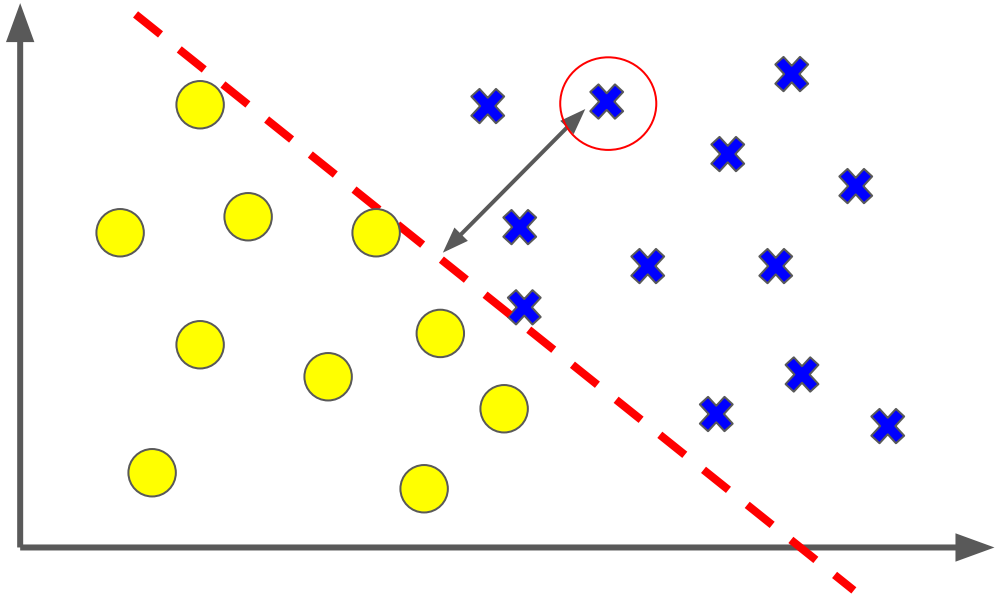
- Still 99.5% Accurate.
- In sensitive domains, I can not tolerate many wrong decisions especially when data distribution changed.
- Probabilistic Classification is needed.

Background: Probabilistic Classification



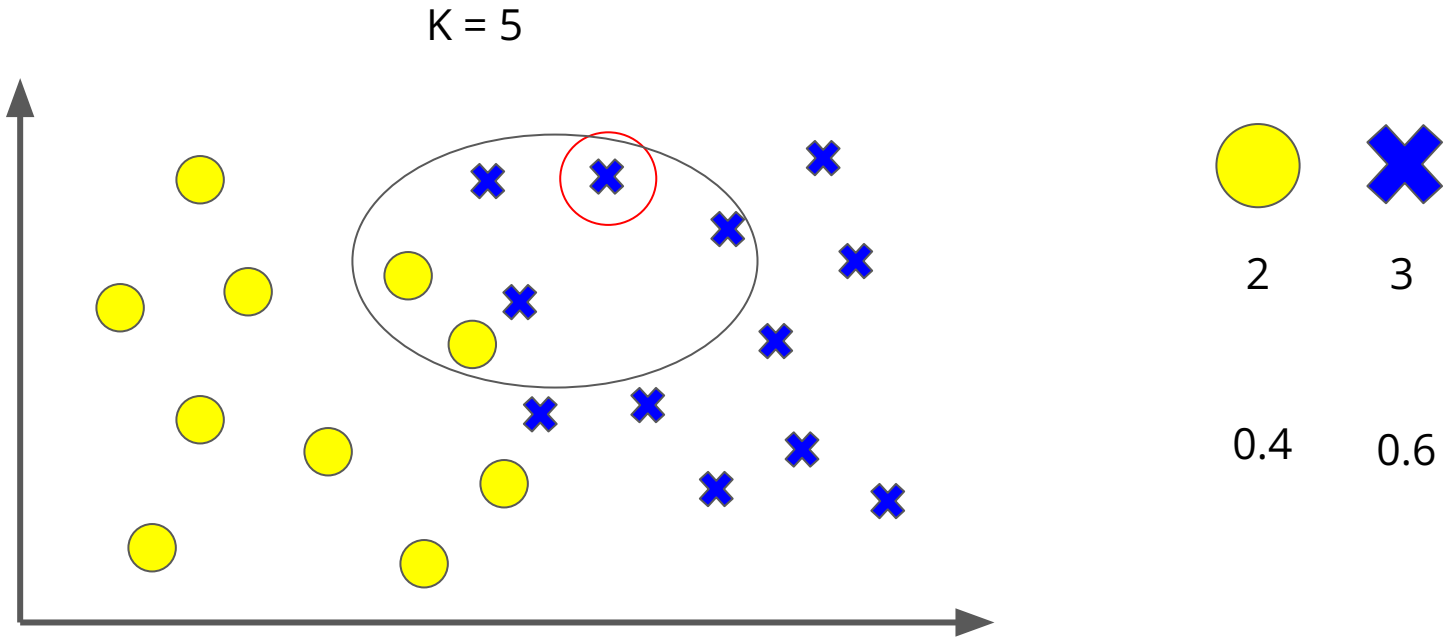
Background

- Can the model predict probabilities instead of labels only?
  - Linear models: distance to the decision boundary



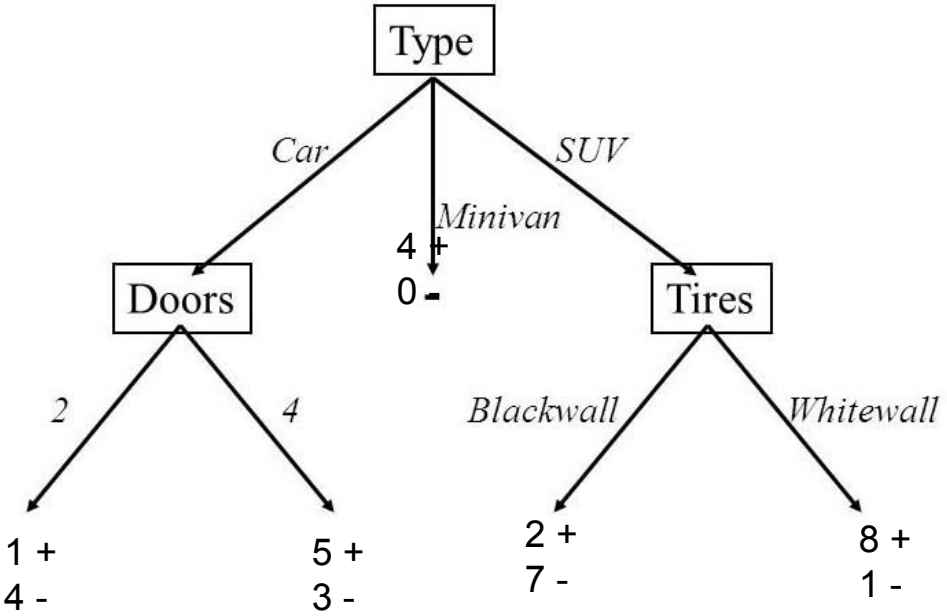
Background

- Can the model predict probabilities instead of labels only?
  - KNN: number of neighbors from one class / total number of neighbors



Background

- Can the model predict probabilities instead of labels only?
  - Decision Tree



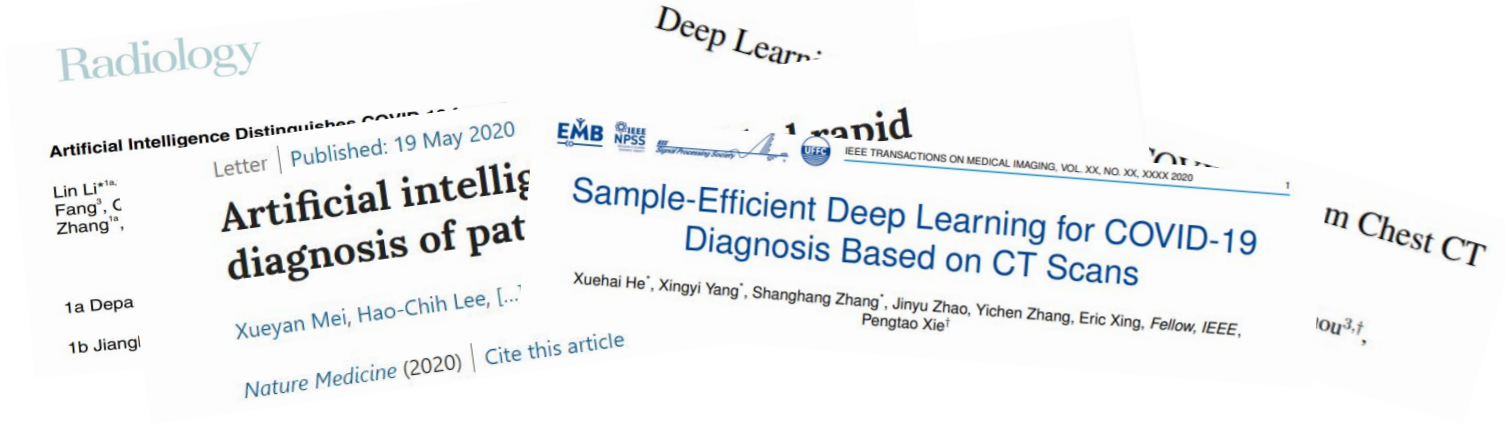
SUV with Blackwall tires has probability 2/(2+7) = 0.22 to be a taxi!

Motivation

AI-based diagnostic tools for COVID-19

- More than 36 published studies are proposing different machine learning models to detect the viral infection based on tomography scans [1].
- Can not be used in practice because of 1) Interpretability 2) Calibration

[1] Wynants Laure, et al. "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal." *bmj* 369 (2020).



Motivation

Giza Systems

Let's use AI to diagnose COVID-19 faster

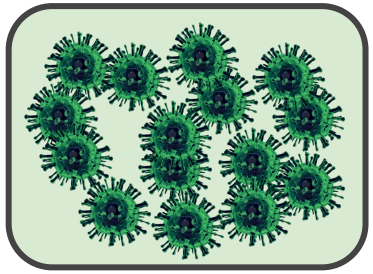
Example: Two models are used to diagnose **100** infected similar samples with COVID-19.

Model 1

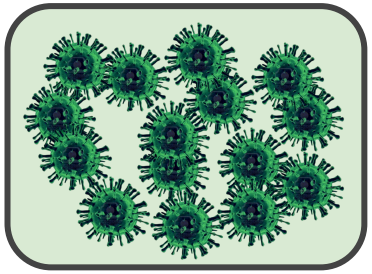
Accuracy = **60%**

Confidence = **0.6**, **0.4**

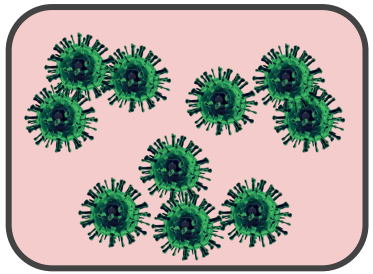
True Positives(60)



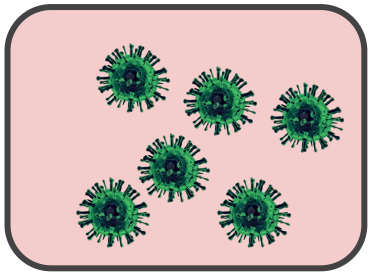
True Positives(75)



False Negatives(40)



False Negatives(25)



Confidence = **0.8**, **0.2**

Model 2

Accuracy = **75%**

Motivation

Giza Systems

Let's use AI to diagnose COVID-19 faster

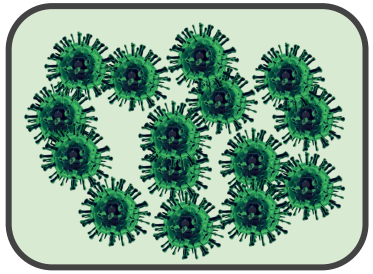
Example: Two models are used to diagnose **another 100** infected similar samples

Model 1

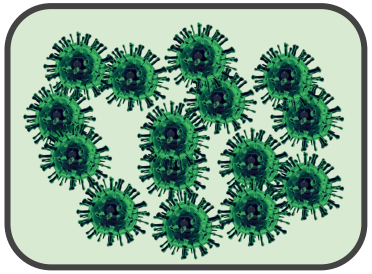
Accuracy = **99%**

Confidence = **0.99**

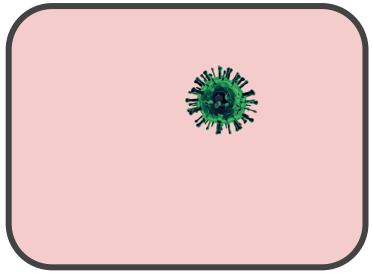
True Positives(99)



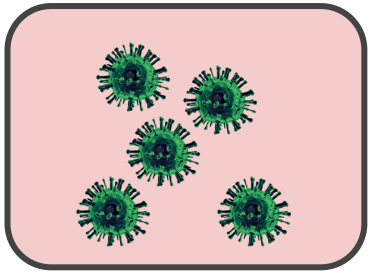
True Positives(95)



False Negatives(1)



False Negatives(5)



Confidence = **0.99**

Model 2

Accuracy = **95%**

Motivation

Which model to trust ?

Model **1** has lower accuracy but if it predicts with high confidence, I would **trust** it.

Overall Accuracy
79.5 %

Model 1

- Less accuracy
- Calibrated

If model **1** is confident of its predictions, it is almost correct.
If it is not, we can verify the results using slower standard diagnostic tests.

Model 2

- Higher accuracy
- Overconfident

If model **2** is confident of its predictions, few false negatives can lead to ***disastrous spread of the virus***

Overall Accuracy
85 %

Motivation: Calibrated Classifier

Perfect Calibrated Classifier

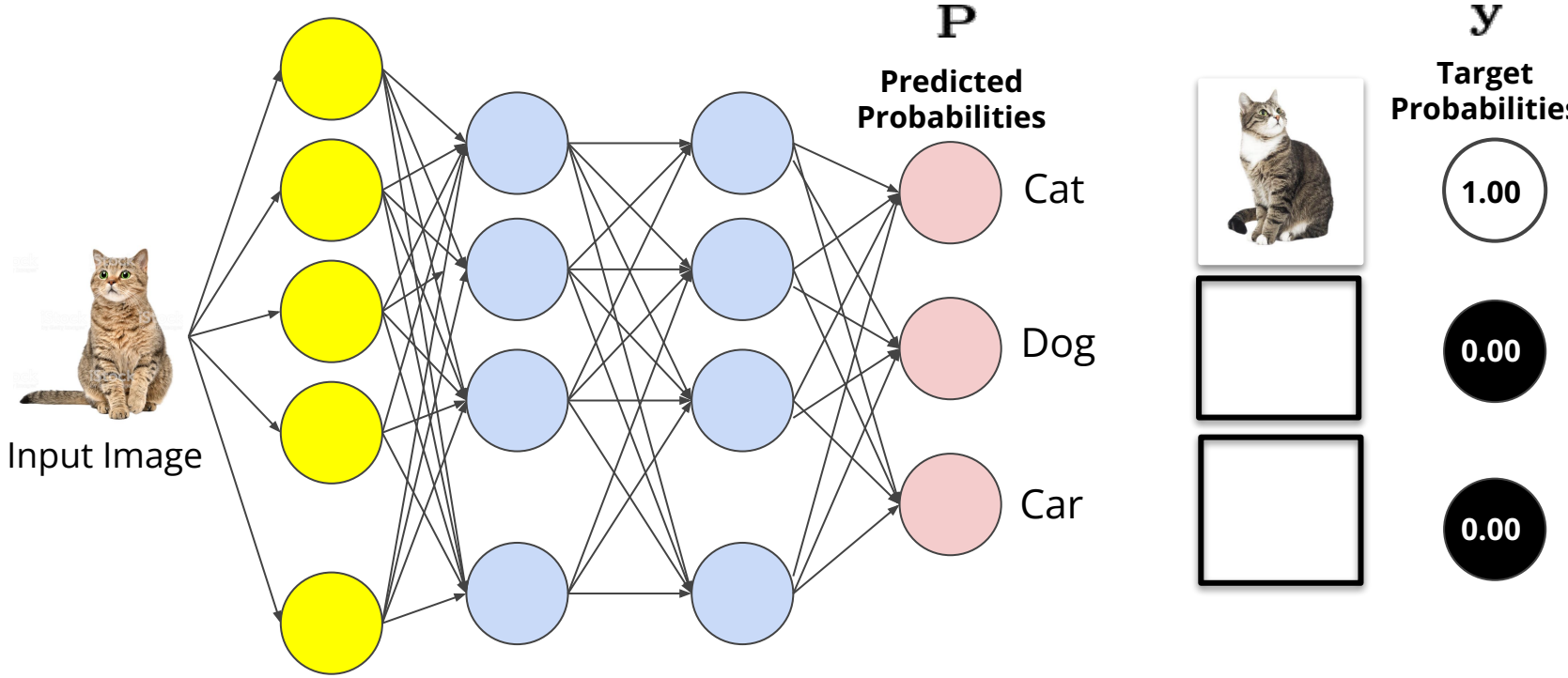
A model is *perfectly calibrated* if, for any probability value **p**, a prediction of a class with confidence **p** is correct **100*p %** of the time.

Model 1 is the **WINNER!**

Motivation: E.g. OverConfidence in Deep Learning

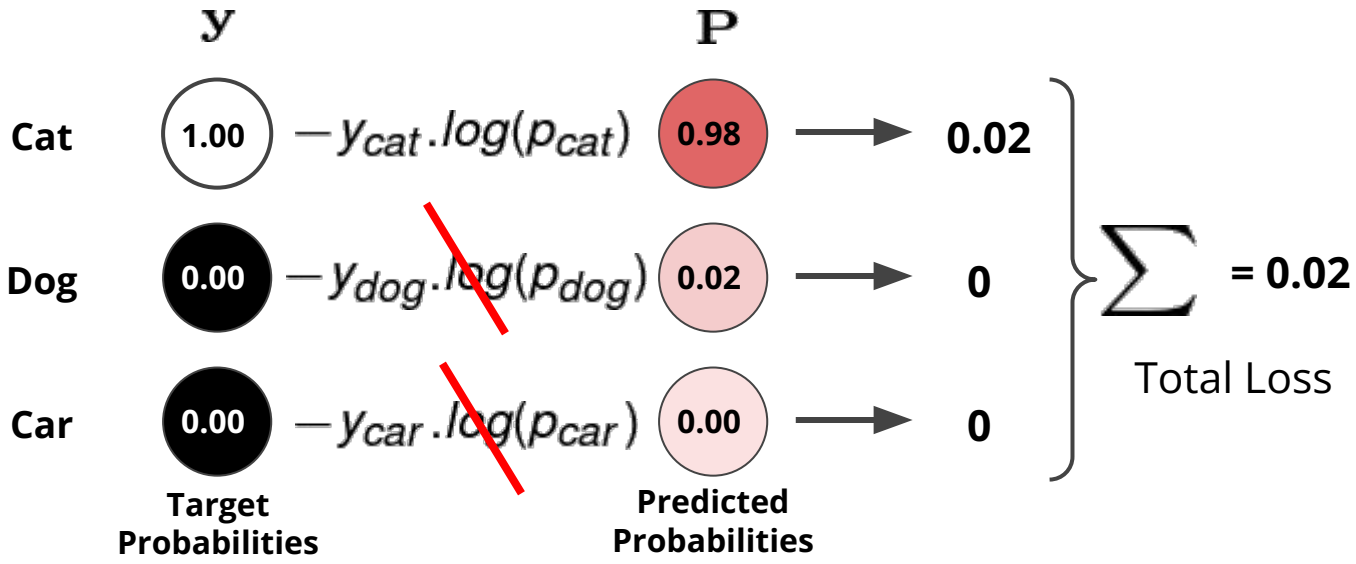
Giza Systems

Teach to predict a cat with full confidence.



Motivation: E.g. OverConfidence in Deep Learning

- **Overconfident predictions**



Motivation: E.g. OverConfidence in Deep Learning

CIFAR10 → Model Training

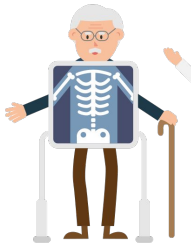
SVHN → Model Evaluation

Neural Networks yield overconfident predictions on unrelated tasks!

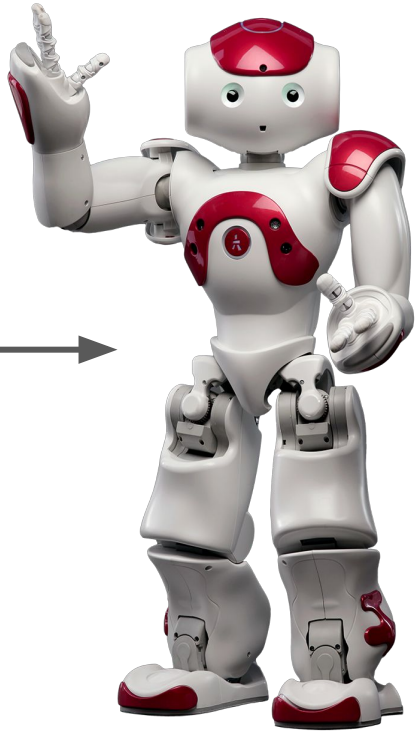
Hein, Matthias, Maksym Andriushchenko, and Julian Bitterwolf. "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

How to measure Calibration? Reliability Diagram

Input

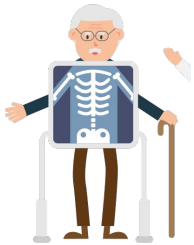


1000 patients

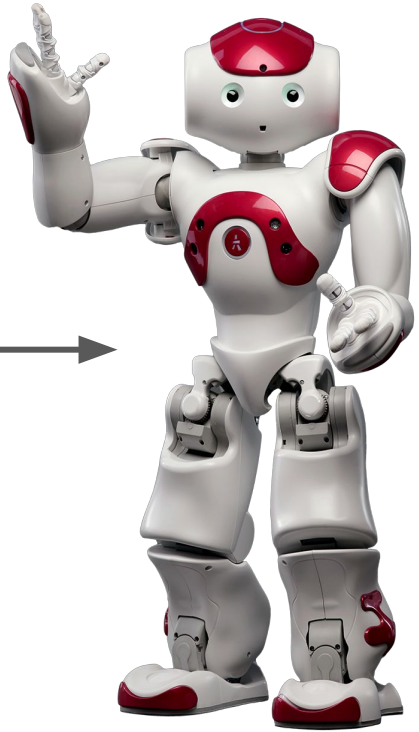


How to measure Calibration? Reliability Diagram

Input



1000 patients

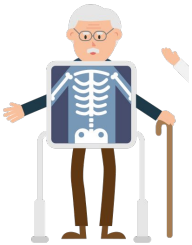


- 90 patients with predicted probability from 0 to 0.1 (average 0.05)
- 2% of those patients were actually Covid patients

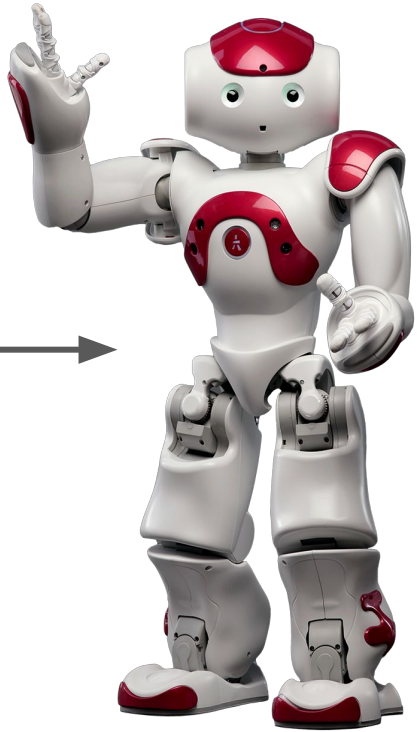
How to measure Calibration? Reliability Diagram

Giza Systems

Input



1000 patients

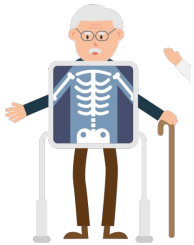


0.05 → 2% (90)

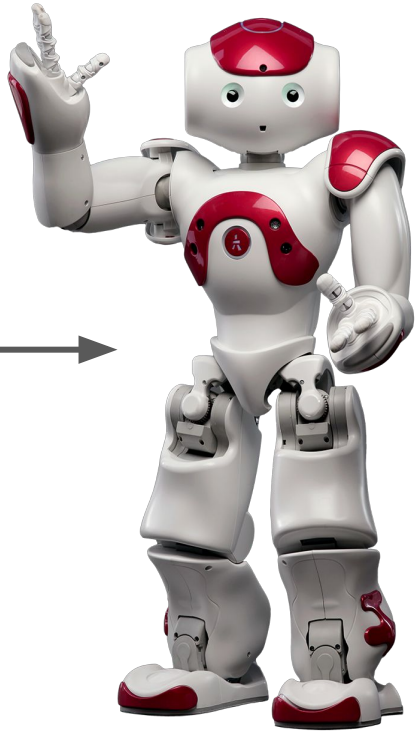
How to measure Calibration? Reliability Diagram

Giza Systems

Input



1000 patients



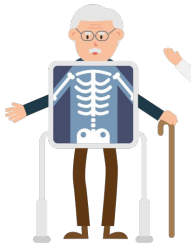
0.05 → 2% (90)

- 110 patients with predicted probability from 0.1 to 0.2 (average 0.17)
- 13% of those patients were actually Covid patients

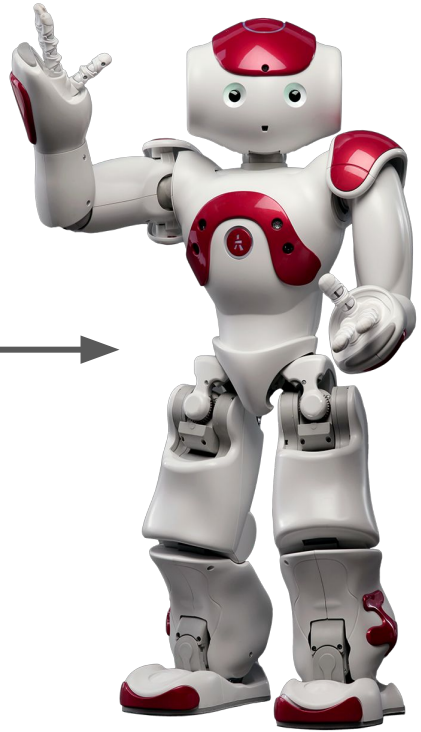
How to measure Calibration? Reliability Diagram

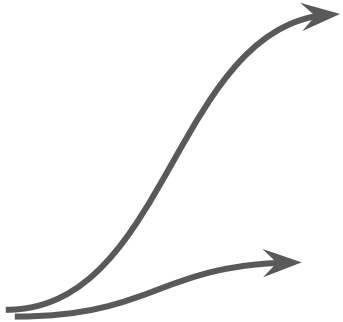
Giza Systems

Input



1000 patients





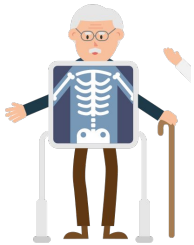
0.05 → 2% (90)

0.17 → 13% (110)

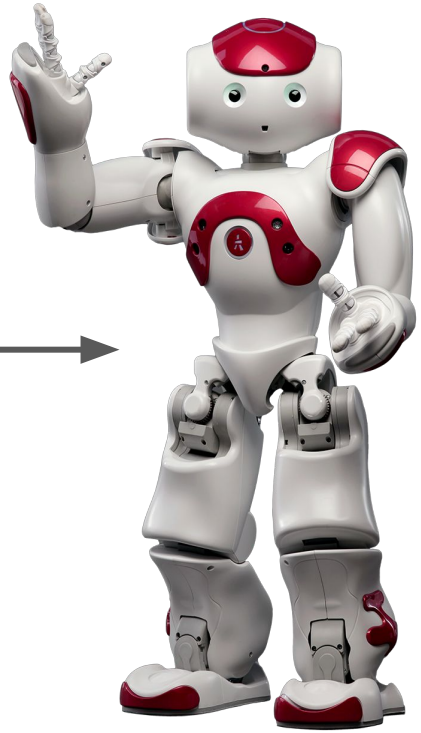
How to measure Calibration? Reliability Diagram

Giza Systems

Input



1000 patients



0.05 → 2% (90)

0.17 → 13% (110)

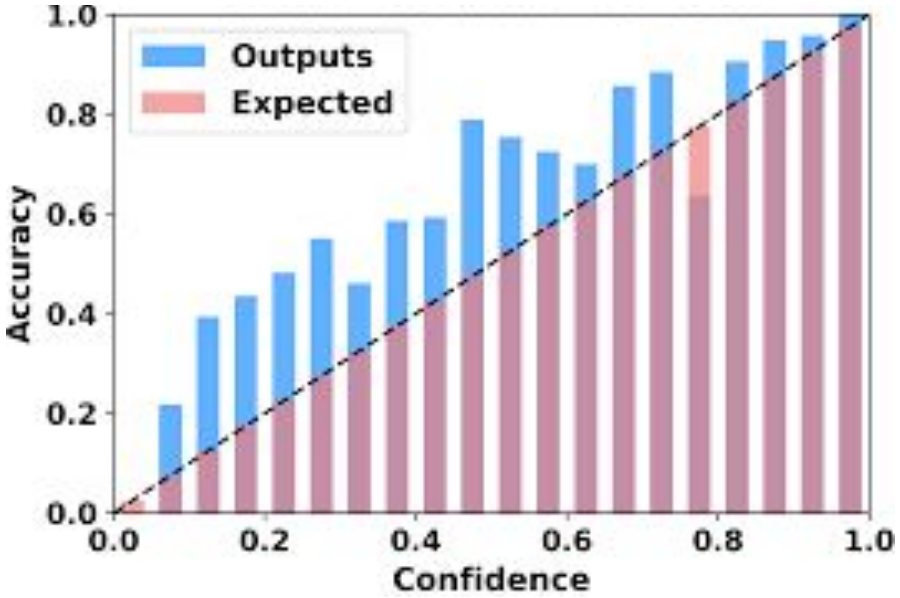
•
•
•

0.98 → 92% (290)

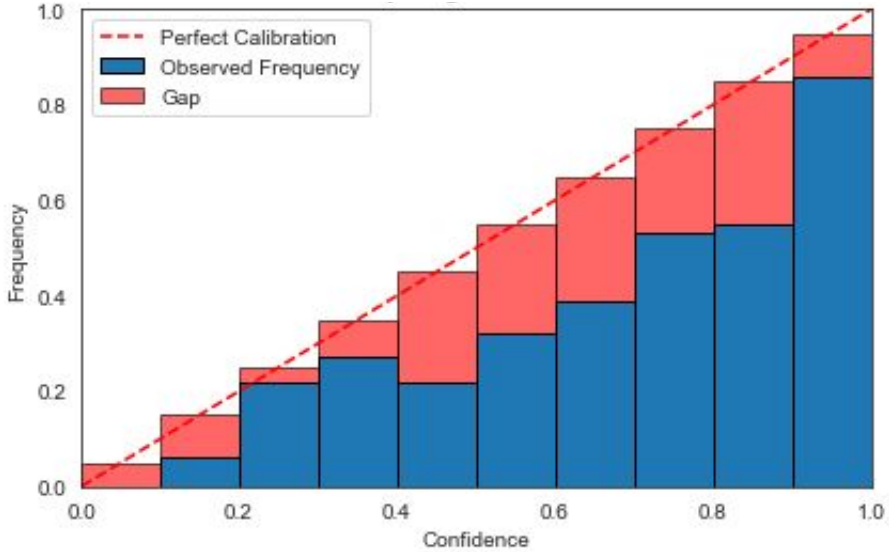
How to measure Calibration? Reliability Diagram

Giza Systems

Underconfident Uncalibrated Model



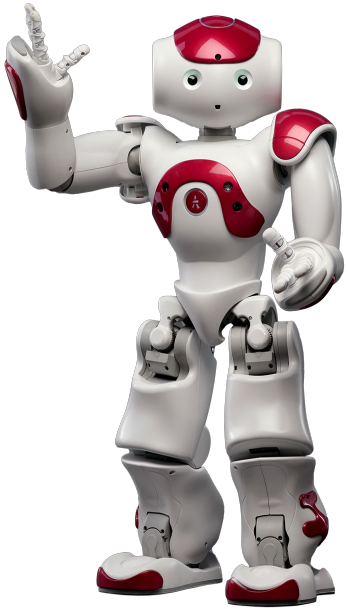
Overconfident Uncalibrated Model



How to measure Calibration? ECE

- Expected Calibration Error

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



0.05 → 2% (90)

0.17 → 13% (110)

•
•
•

0.98 → 92% (290)

90/1000 * |0.05-0.02|

+

110/1000 * |0.17-0.13|

+

•
•
•

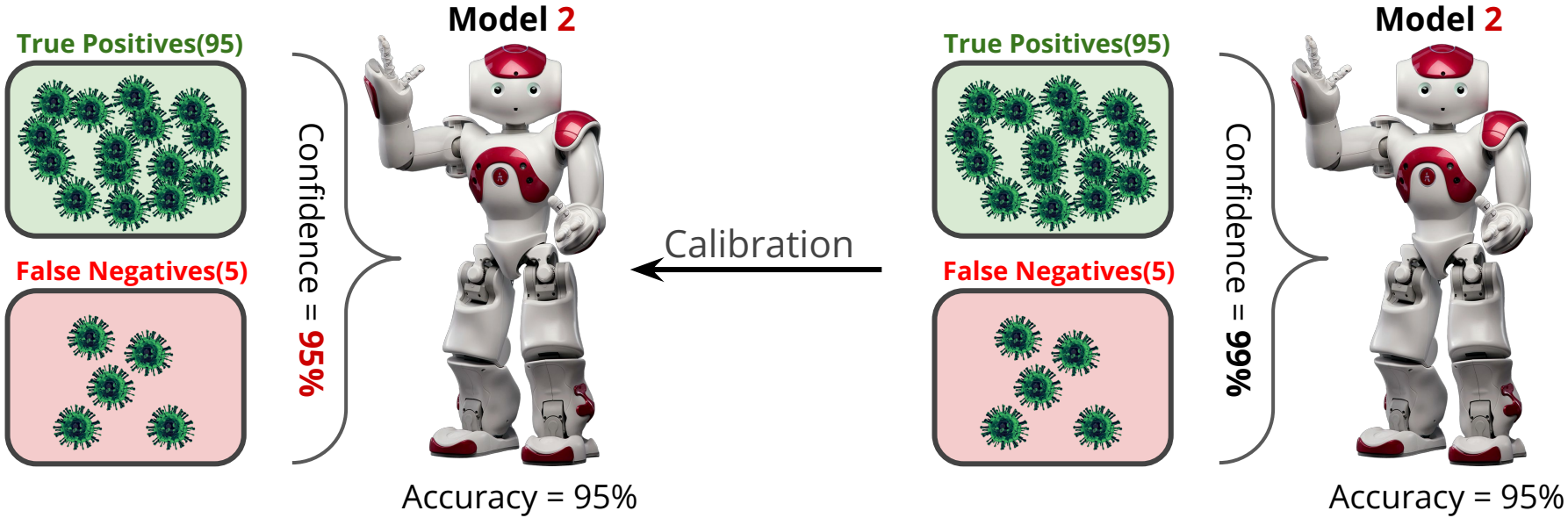
290/1000 * |0.98-0.92|

Lower ECE is better!

Solution: Calibration Process

Calibration of a classifier

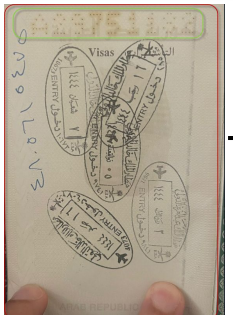
A post-processing step where we take a trained model to improve its predicted confidence levels to match the accuracy of these predictions.



Use Case in Real-Life Scenario

Problem Statement

- A customer asked Giza Systems to develop a feature in the customs mobile app, where the user captures an image of his passport with the entry stamp to Egypt.
- A manual effort is done to verify all customs exemption requests.
- Can we reduce efforts by excluding irrelevant images?



Backend Mobile App
(Passport Paper or not?)

No but confident

DENIED

No but not confident

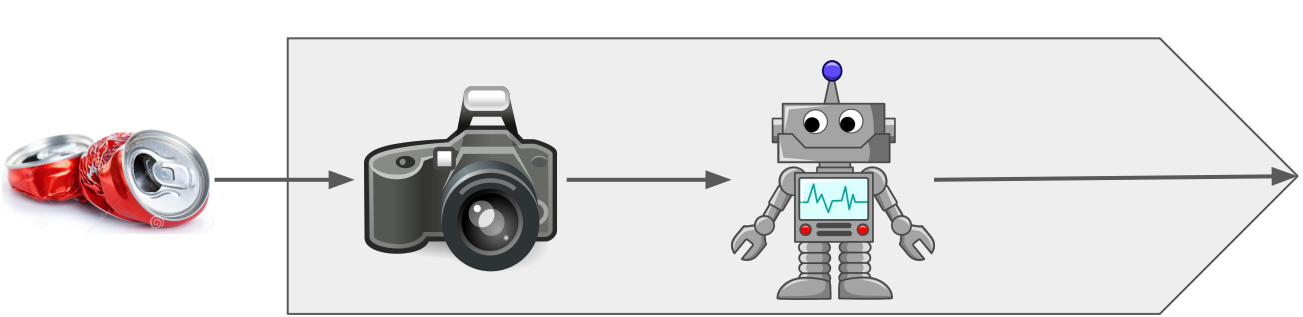
Yes

Use Case in Real-Life Scenario

Problem Statement

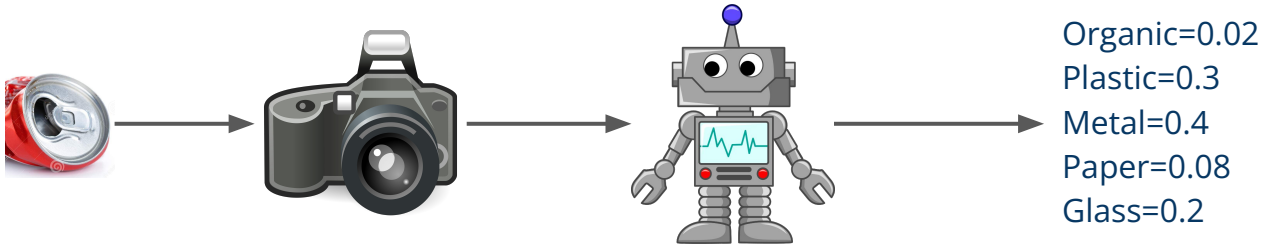
- A customer asked Giza Systems to develop a smart waste management system. When the user throw a trash in the bin hole, a model classifies its type (Organic, Plastic, Meta, Paper, Glass) and the systems directs it automatically towards the correct waste type pipeline.

- Accurate classification could save much manual effort (time and money)



Use Case in Real-Life Scenario

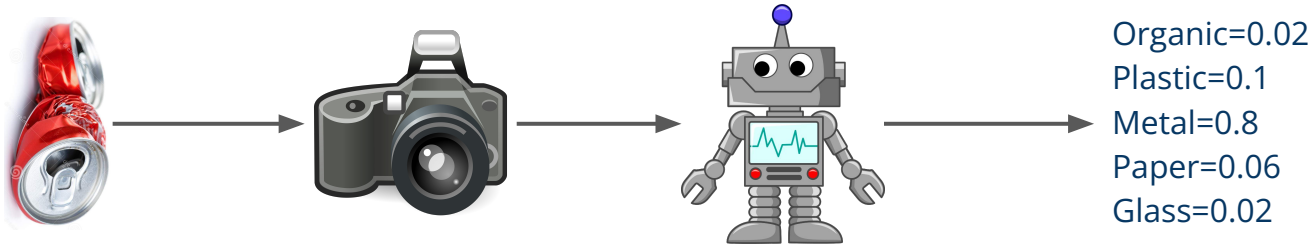
1. Collect dataset for images of trashes of all the needed classes.
2. Train a deep learning model to classify the trashes.
3. Check the calibration performance of your model
   a. Calibrated → Proceed
   b. Uncalibrated → Calibrate it OR Re-train with more regularization imposed
4. Is your model confident of the prediction?



40% Confidence = Not Confident

Use Case in Real-Life Scenario

1. Collect dataset for images of trashes of all the needed classes.
2. Train a deep learning model to classify the trashes.
3. Check the calibration performance of your model
   a. Calibrated → Proceed
   b. Uncalibrated → Calibrate it OR Re-train with more regularization imposed
4. Is your model confident of the prediction?



80% Confidence = Confident

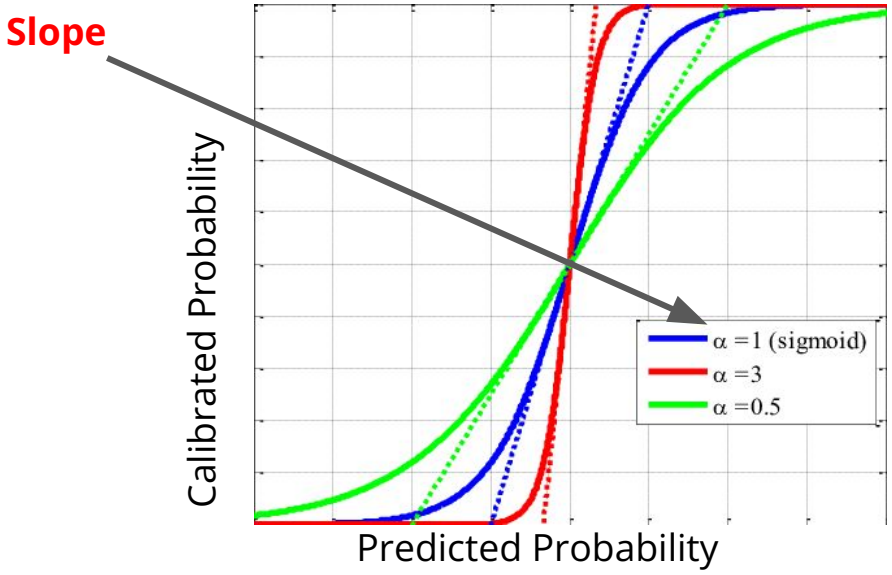
Calibrators

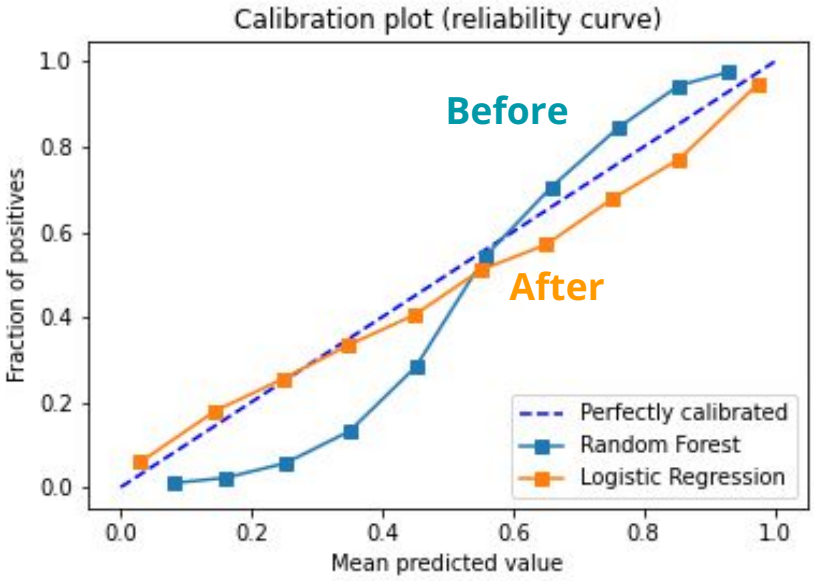
Task	Method	Use Case
Binary Classification	**Logistic Calibration [url]**	**Small datasets**
	Beta Calibration [url]	**Small to Midsize datasets**
	Isotonic Calibration [url]	Large Datasets > 1000 samples to be used in calibration
Multi-Class Classification	**Temperature Scaling [url]**	**Independent of dataset size (Good for Neural Networks)**
	Dirichlet Calibration [url]	Needs very large datasets
	Label Smoothing [url]	**Occurs during training (No post processing)**
	MixUp [url]	**Occurs during training (improves generalization and reduce over-confidence)**

Examples of Calibrators: Logistic Calibration

Logistic Calibration

Fit the best fitting logistic sigmoid function that maps the model predicted probabilities into calibrated probabilities for a binary class dataset.





Examples of Calibrators: Beta Calibration

Beta Calibration

Fit the best fitting beta function that maps the model predicted probabilities into calibrated probabilities for a binary class dataset.

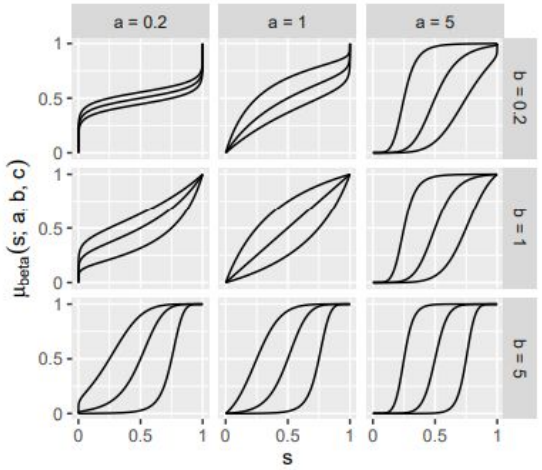
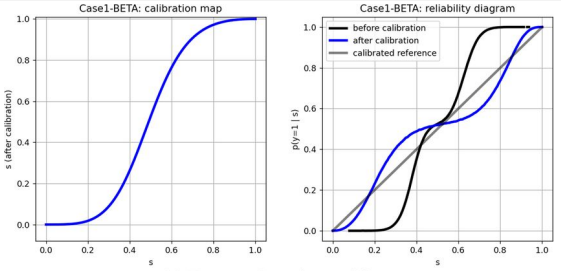
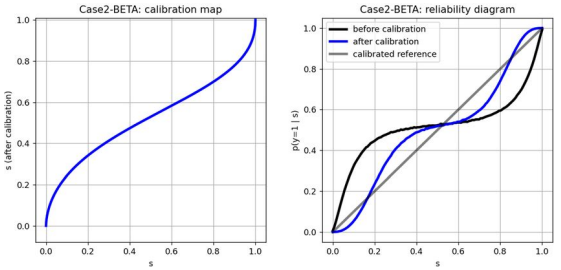


Figure 3: Examples of beta curves with parameters $a, b \in \{0.2, 1, 5\}$, $m \in \{0.25, 0.5, 0.75\}$ and $c = b\ln(1-m) - a\ln m$.

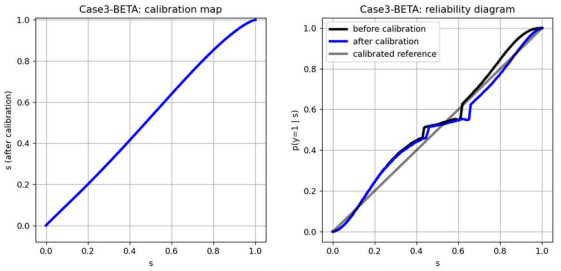
$$\mu_{beta}(s;a,b,c) = \frac{1}{1+1\Big/\left(e^{c}\frac{s^{a}}{(1-s)^{b}}\right)}$$



(a) Test case 1: under-confidence

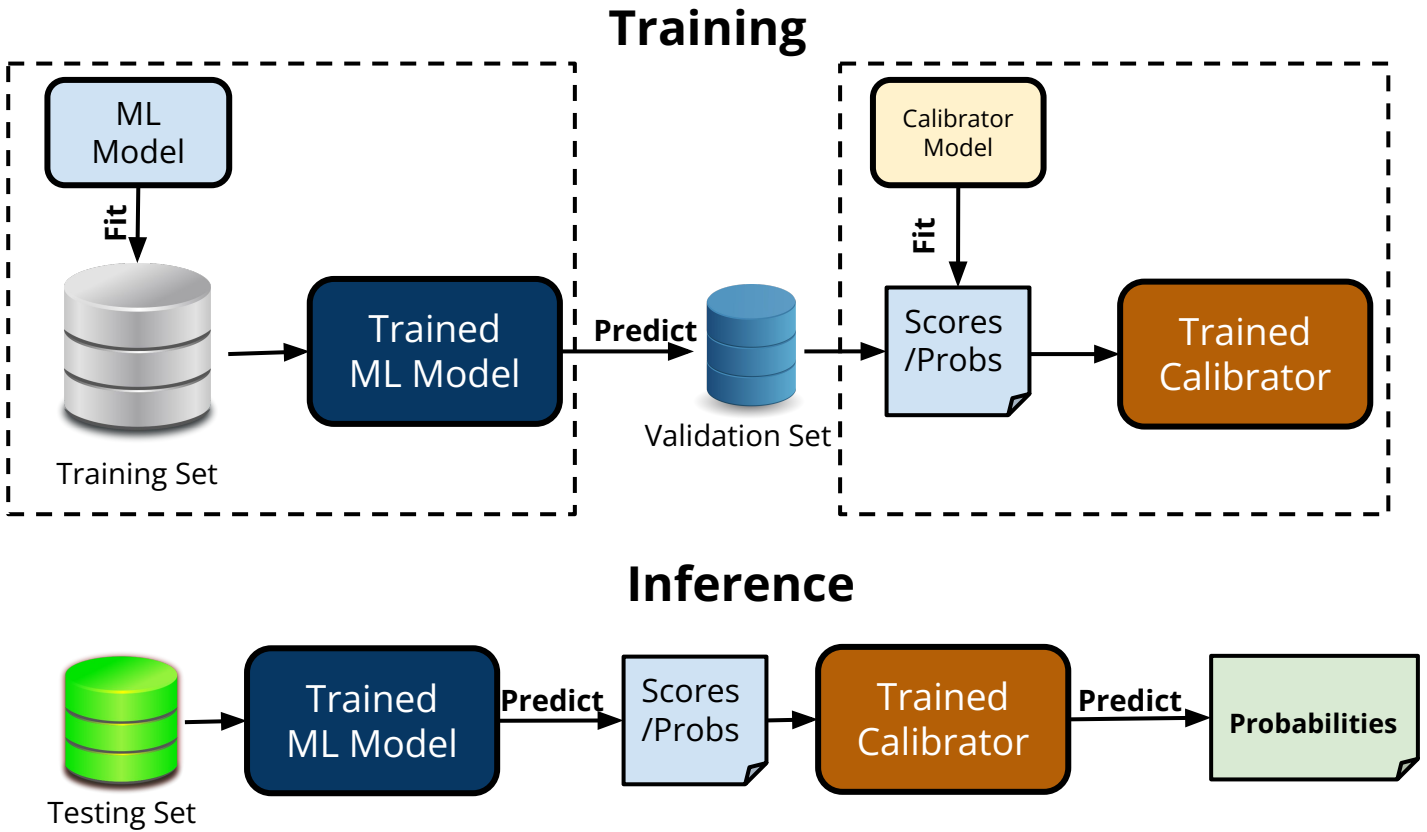


(b) Test case 2: over-confidence



(c) Test case 3: MLP classifier

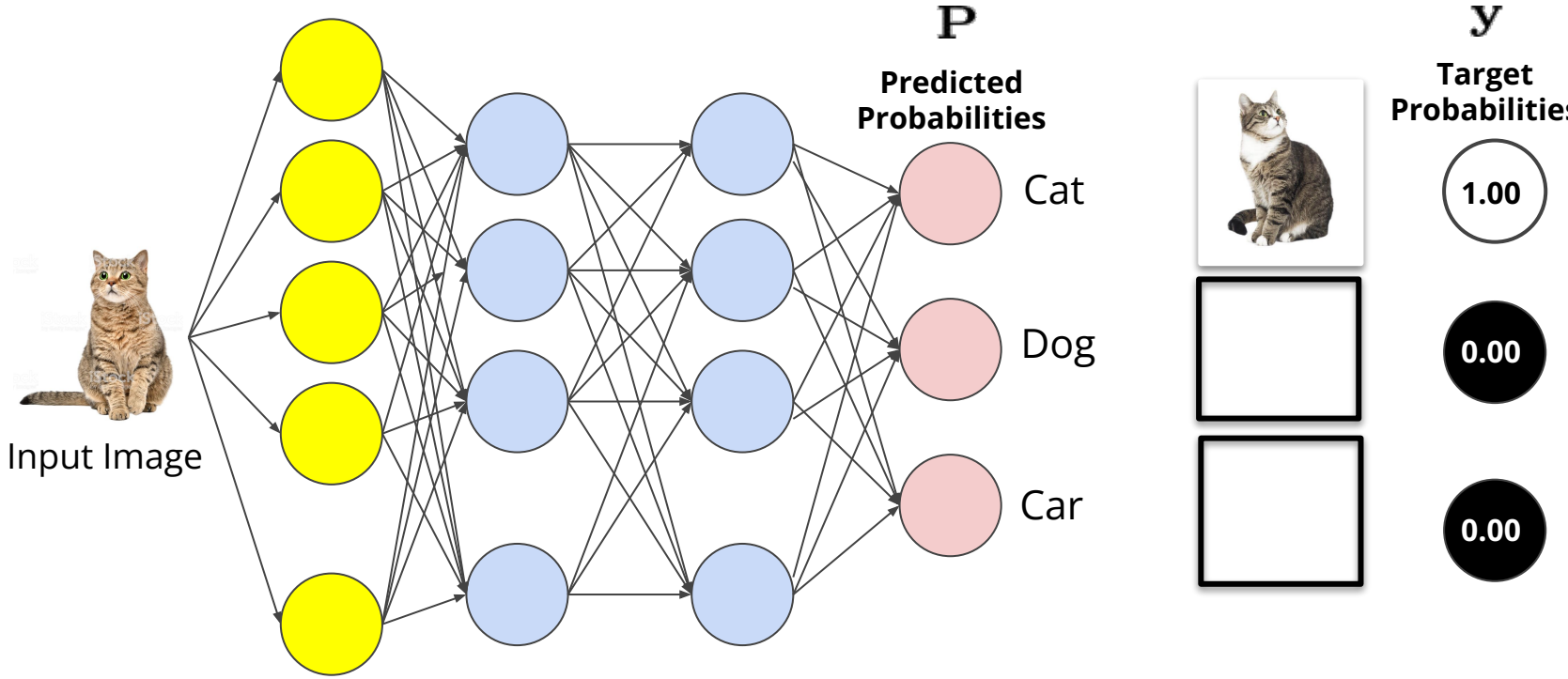
Examples of Calibrators:



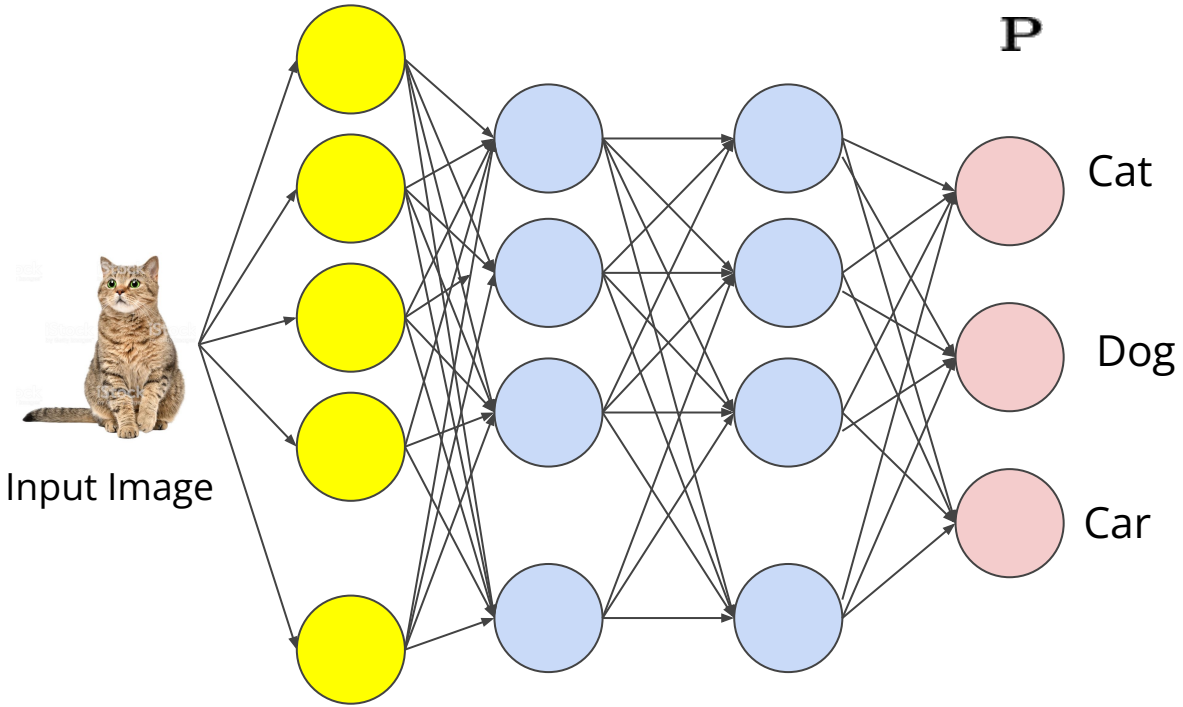
Examples of Calibrators: Label Smoothing

Giza Systems

Teach to predict a cat with full confidence.

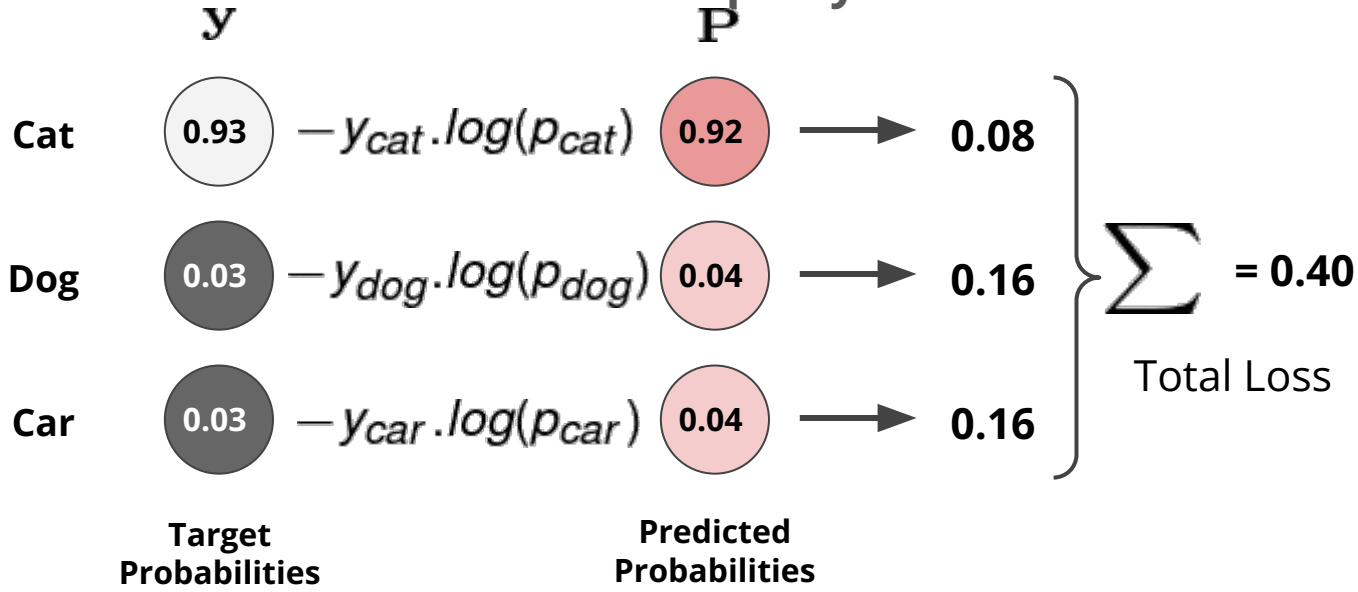


Examples of Calibrators: Label Smoothing



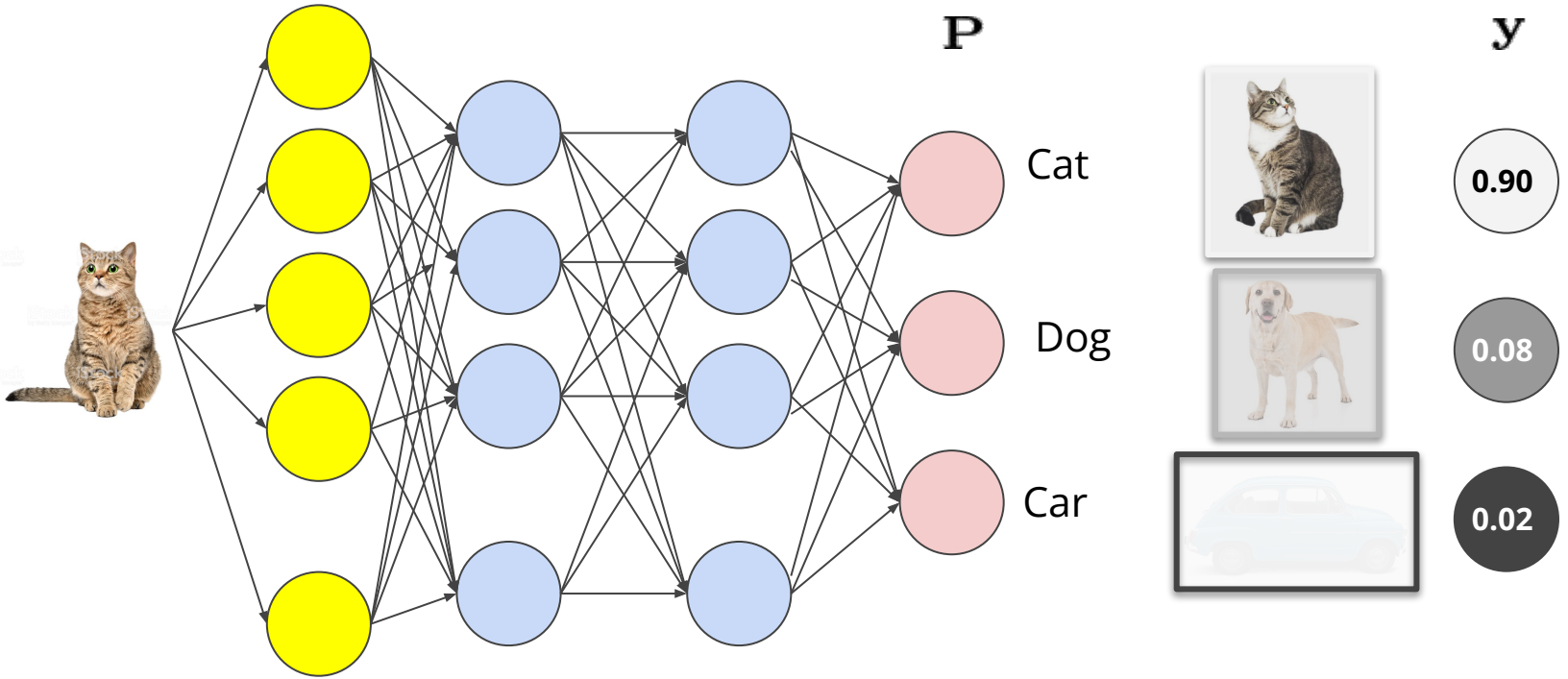
Examples of Calibrators: Label Smoothing

- Less confident predictions
- Enforce to learn all incorrect classes as **equally similar**.



https://proceedings.neurips.cc/paper_files/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf

Instance-based label smoothing

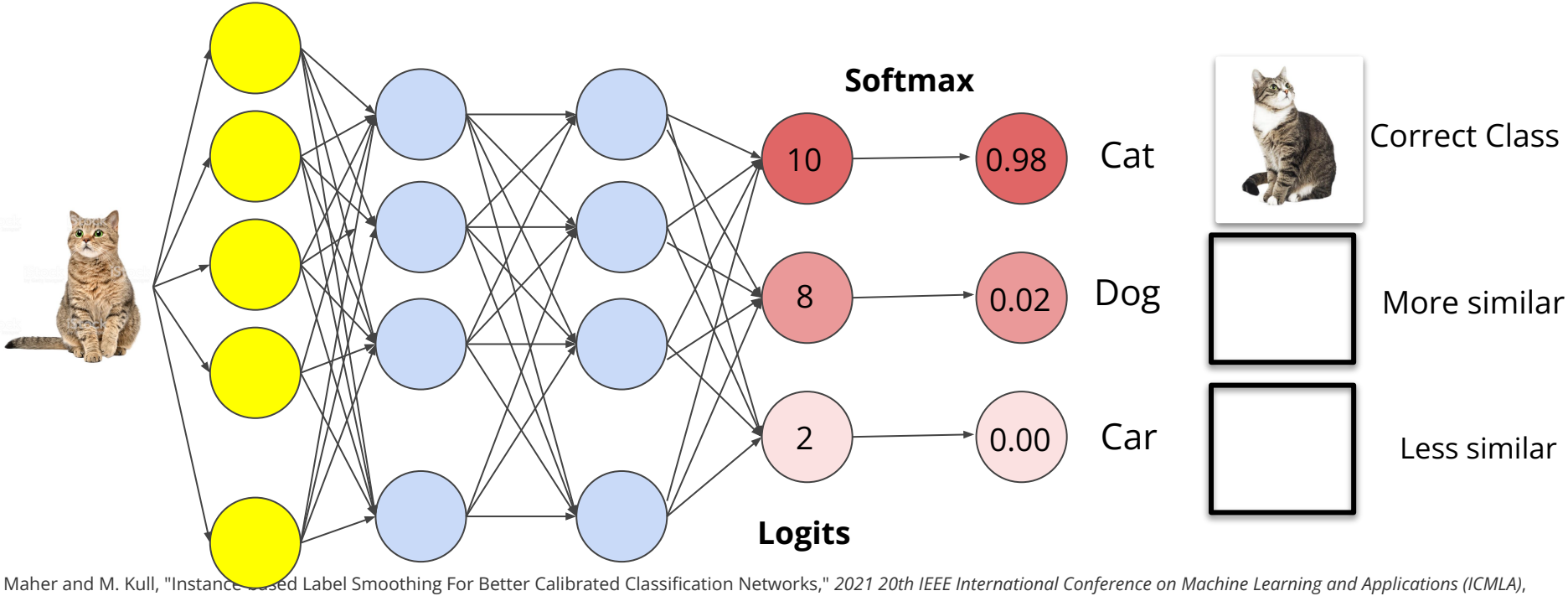


M. Maher and M. Kull, "Instance-based Label Smoothing For Better Calibrated Classification Networks," *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Pasadena, CA, USA, 2021, pp. 746-753, doi: 10.1109/ICMLA52953.2021.00124.

Remember: Training without label smoothing

Giza Systems

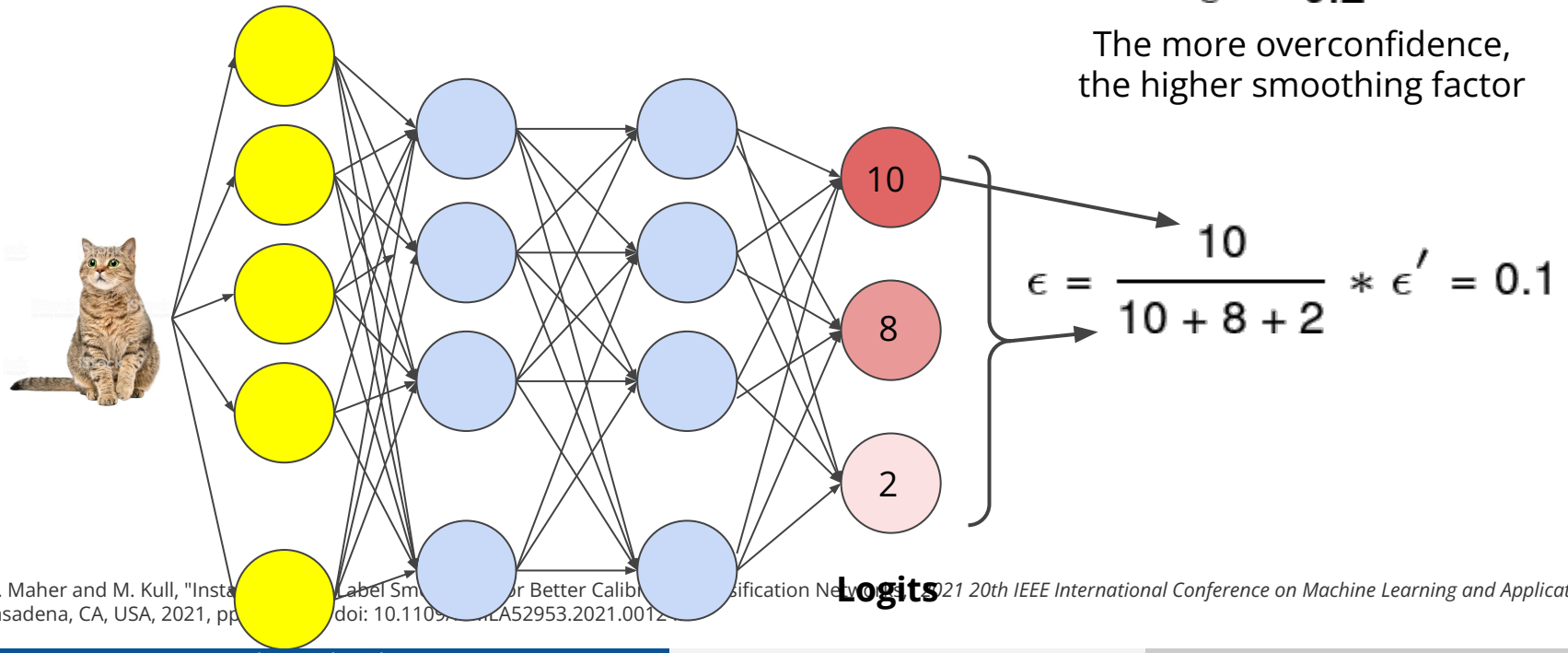
- No enforcement to learn a specific distribution for incorrect classes.



M. Maher and M. Kull, "Instance-based Label Smoothing For Better Calibrated Classification Networks," *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Pasadena, CA, USA, 2021, pp. 746-753, doi: 10.1109/ICMLA52953.2021.00124.

Choosing the smoothing factor

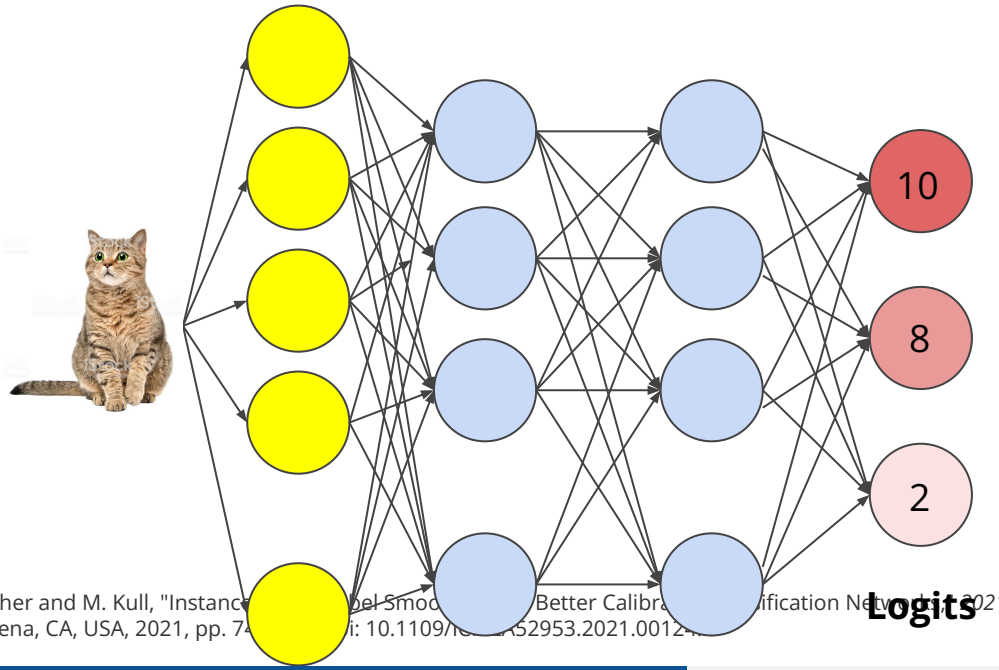
- **Less confident predictions.**



M. Maher and M. Kull, "Insta… Label Sm… or Better Calib… sification Ne… 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Pasadena, CA, USA, 2021, pp… doi: 10.1109/… ILA52953.2021.00012…

Instance-based label smoothing

- Less confident predictions.
- **Enforce to learn classes' similarity of the original dataset.**



$$\epsilon = 0.1$$

$$y_{cat} = 1 - \epsilon = 0.9$$

$$y_{dog} = \frac{8}{8+2}\epsilon = 0.08$$

$$y_{car} = \frac{2}{8+2}\epsilon = 0.02$$

0.90

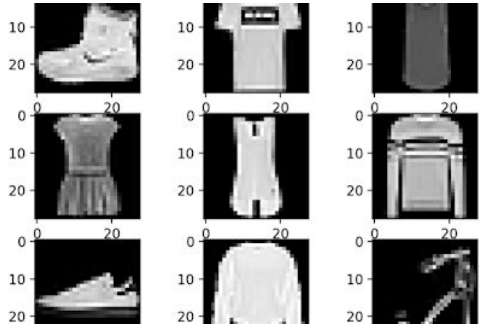
0.08

0.02

M. Maher and M. Kull, "Instance-... Better Calibra... ification Networks," *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Pasadena, CA, USA, 2021, pp. 7... i: 10.1109/IC...S2953.2021.00124.

Experimental Results

Fashion-MNIST

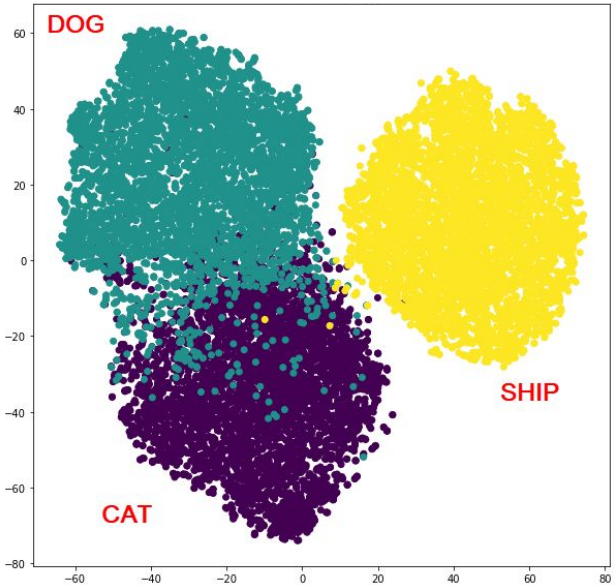


Cifar-10

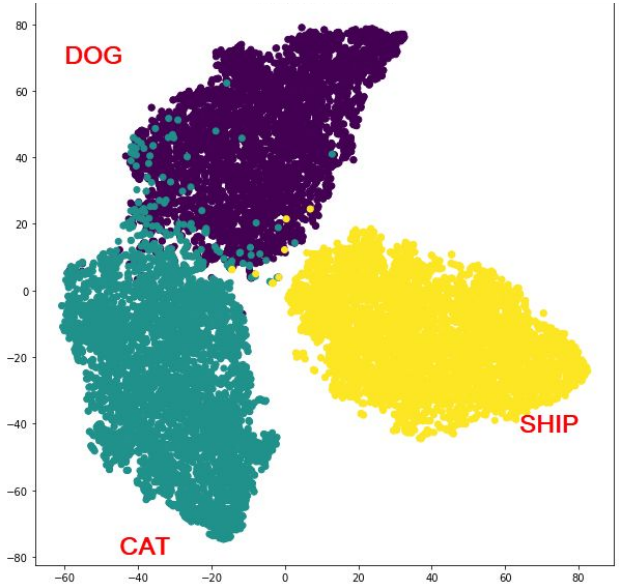
Cifar-100

Experimental Results

No label smoothing



Standard label smoothing



Instance-based label smoothing

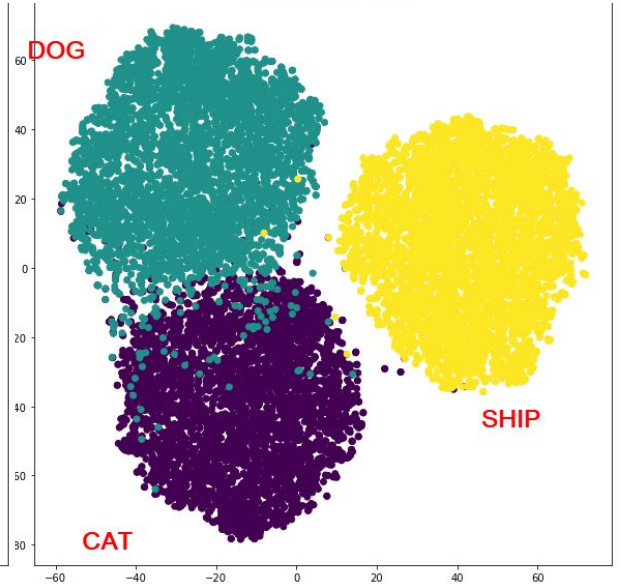


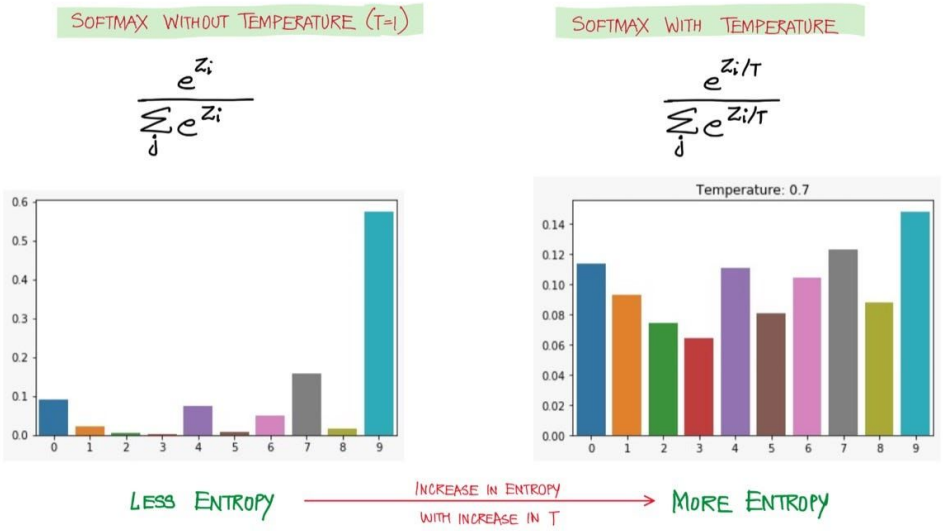
Figure: T-SNE visualization of the output logits for 3 different classes of Cifar10 dataset.

Temperature Scaling

- A 3-class dataset (cats, dogs and cars)

Logits: `[10.0, 6.0, 1.0]` → Softmax → **Probabilities**: `[0.982, 0.018, 0.0001]`

Logits: `[10/3, 6/3, 1/3]`
Temperature: 3
→ Temp Scaling + Softmax → **Probabilities**: `[0.761, 0.201, 0.038]`

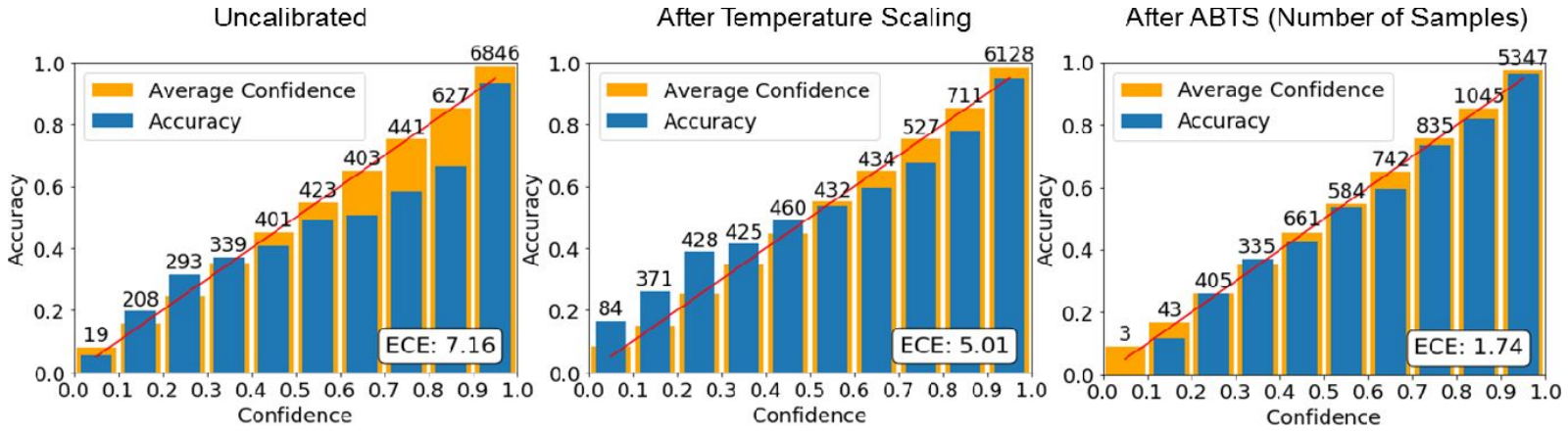


Temperature Scaling

- A 3-class dataset (cats, dogs and cars)

Logits: `[10.0, 6.0, 1.0]` → Softmax → **Probabilities**: `[0.982, 0.018, 0.0001]`

Logits: `[10/3, 6/3, 1/3]`
Temperature: 3
→ Temp Scaling + Softmax → **Probabilities**: `[0.761, 0.201, 0.038]`



MixUp: Reduce Neural Networks Over-confidence

Giza Systems

1. **Data Interpolation**: Given two training examples **(x_i,y_i)** and **(x_j,y_j)**, mixup generates a new training example **($\tilde{x}$,$\tilde{y}$)**

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$$

   where λ is a mixing coefficient sampled from a Beta distribution, typically λ~Beta(α,α) for some parameter α>0.
   A common choice is α=0.2.

2. **Label Interpolation**: The labels are also linearly interpolated, so instead of assigning a hard class label, the model learns from a *soft label*, representing a mix of the two classes.

For instance, if y_i is the label for a cat image (e.g., one-hot encoded as [1,0]) and y_j is the label for a dog image [0,1], then the synthetic label might be something like [0.8,0.3], indicating 80% "cat" and 20% "dog."

Zhang, Hongyi. "mixup: Beyond empirical risk minimization." *arXiv preprint arXiv:1710.09412* (2017). → Published in ICLR 2018

Takeaways

- Do not get fooled with “highly accurate models”.
  - Metrics like Accuracy, F-Score, Precision, Recall cares only about the class label
- Probabilistic Classification is important in cost sensitive domains.
  - I can discard unconfident model predictions.
  - I can use another model to improve the prediction confidence.
- Machine Learning Classifiers can be used to predict scores/probabilities
- These probabilities are usually uncalibrated and need further calibration methods.
- Calibration performance is measured through Expected Calibration Error (ECE) and Reliability diagram.
- Calibration can be done using methods like logistic calibration and isotonic calibration for binary classification, or temperature scaling and label smoothing for multi-class classification.

Takeaways

- Can I use binary classification calibration methods in multi-classification task?
  - ***Yes, 1 Vs All***

- Good Research Point:
  - Automation of the Calibration / Over-confidence Reduction?

Measuring Uncertainty in Regression Models: Bootstrapping

- Jack-Knife+:
  - **leave-one-out** method consists of training a model for each data point in our dataset, training it on the entire dataset but removing one sample at a time.

F1	F2	F3	Tar
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..

Model 1

Kim, Byol, Chen Xu, and Rina Barber. "Predictive inference is free with the jackknife+-after-bootstrap." *Advances in Neural Information Processing Systems* 33 (2020): 4138-4149.

Measuring Uncertainty in Regression Models: Bootstrapping

Giza Systems

- Jack-Knife+:
  - **leave-one-out** method consists of training a model for each data point in our dataset, training it on the entire dataset but removing one sample at a time.

F1	F2	F3	Tar
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..

Model 1

Model 2

Kim, Byol, Chen Xu, and Rina Barber. "Predictive inference is free with the jackknife+-after-bootstrap." *Advances in Neural Information Processing Systems* 33 (2020): 4138-4149.

Measuring Uncertainty in Regression Models: Bootstrapping

- Jack-Knife+:
  - **leave-one-out** method consists of training a model for each data point in our dataset, training it on the entire dataset but removing one sample at a time.

F1	F2	F3	Tar
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..

Model 1

Model 2

Model 3

Kim, Byol, Chen Xu, and Rina Barber. "Predictive inference is free with the jackknife+-after-bootstrap." *Advances in Neural Information Processing Systems* 33 (2020): 4138-4149.

Measuring Uncertainty in Regression Models: Bootstrapping

Giza Systems

- Jack-Knife++:
  - **leave-one-out** method consists of training a model for each data point in our dataset, training it on the entire dataset but removing one sample at a time.

F1	F2	F3	Tar
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..

Model 1

Model 2

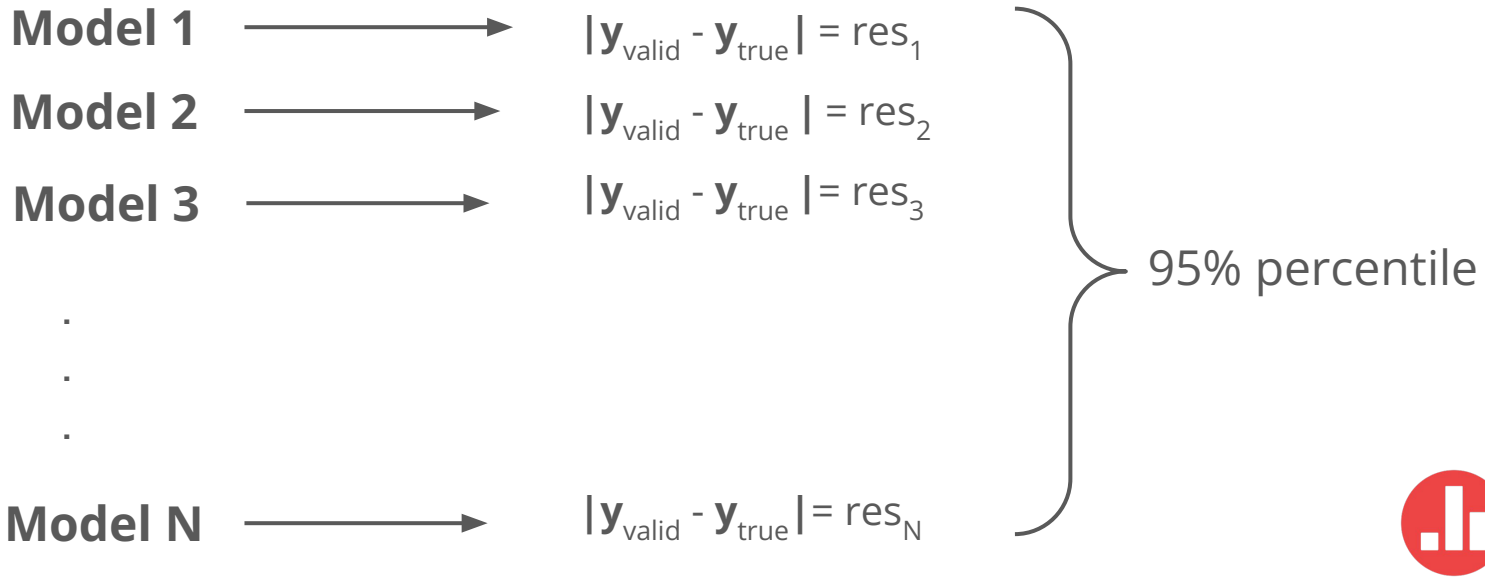
Model 3

Model N

Kim, Byol, Chen Xu, and Rina Barber. "Predictive inference is free with the jackknife+-after-bootstrap." *Advances in Neural Information Processing Systems* 33 (2020): 4138-4149.

Measuring Uncertainty in Regression Models: Bootstrapping

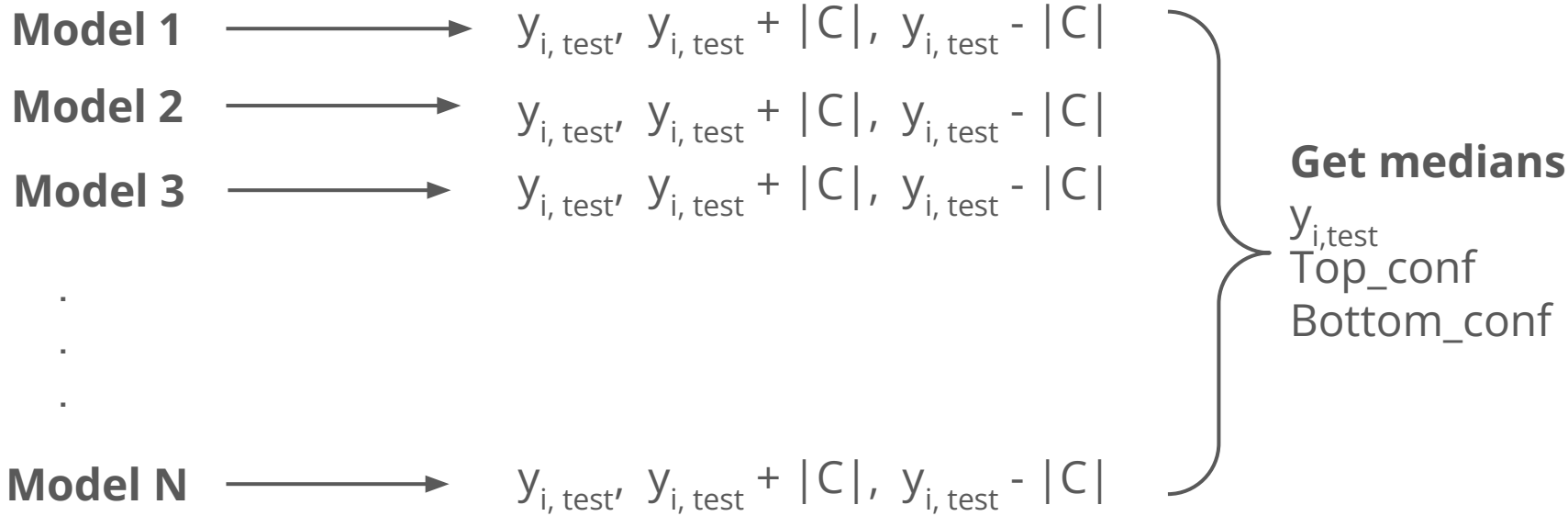
- Jack-Knife+:
  - Get residuals from each model on the validation set.
  - Compute the confidence interval thresholds for the N-percentile.



Kim, Byol, Chen Xu, and Rina Barber. "Predictive inference is free with the jackknife+-after-bootstrap." *Advances in Neural Information Processing Systems* 33 (2020): 4138-4149.

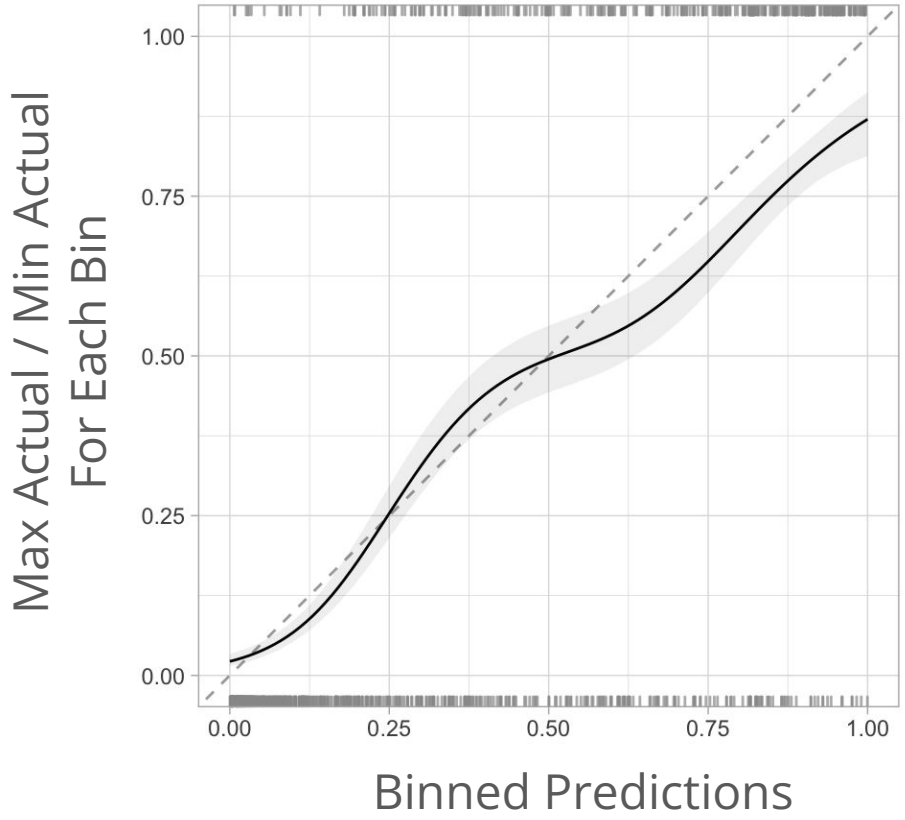
Measuring Uncertainty in Regression Models: Bootstrapping

- Jack-Knife+:
  - Compute the prediction on the test set instance
  - Add the confidence interval threshold



Kim, Byol, Chen Xu, and Rina Barber. "Predictive inference is free with the jackknife+-after-bootstrap." *Advances in Neural Information Processing Systems* 33 (2020): 4138-4149.

Calibration Plot for Regression Models



Hands-On Task (New Task)

- **Data Team Details:**
  - Individual Submissions
  - Due date: **11 November (Data)**
  - Delivery via email to mohamed.maher@gizasystems.com & headway_2024_data_testing@gizasystems.com
    - Title: **“Calibration Task Delivery”**
    - Content:
      - Model Training Code files (notebook/.py),
      - Describe briefly your approach and selected models,
      - Pipeline evaluation file (.py) to be deployed as an endpoint when executed.
    - Endpoint
      - Body:
        - {0: {‘meta-feature1_name’: value, ‘meta-feature2_name’: value, ….},
          1: {‘meta-feature1_name’: value, ‘meta-feature2_name’: value, ….}, …}
      - Response:
        - {0: {‘XGBRegressor’: 0.73, ‘LinearSVR’: 0.14, ….},
          1: {‘LASSO’: 0.63, ‘LinearSVR’: 0.26, …}, …}

Hands-On Task (New Task)

- **Data Testing Team Details**
  - Work in pairs
  - Due date: **17 November (Data Testing)**
  - Delivery via email to mohamed.maher@gizasystems.com
    - Title: **“Calibration Task Evaluation Delivery”**
    - Content:
      - The script used to evaluate the models
      - Report
        - Team members and who worked on which models?
        - Comparison:
          - Measuring the accuracy (doesn’t mean accuracy metric) performance of the models.
            - F1, AUC, Micro Recall, …?
            - Justify your selection
          - Measuring the reliability of the models.
            - Reliability diagram
            - ECE