Calibrated Uncertainty in Deep Learning

Uncertainty in Deep Learning Workshop @ UAI18

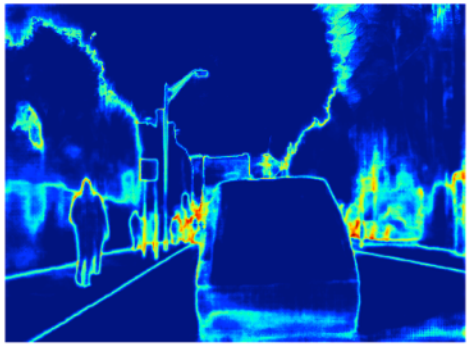
Volodymyr Kuleshov

August 10, 2018

Estimating Uncertainty is Crucial in Many Applications

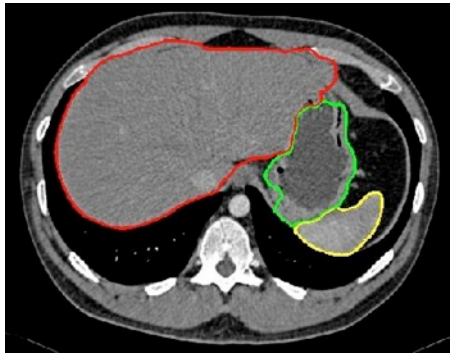
Assessing uncertainty can be as important as obtaining high accuracy.

Computer Vision



Smith & Cheeseman (1986)
McAllister et al. (2017)

Healthcare



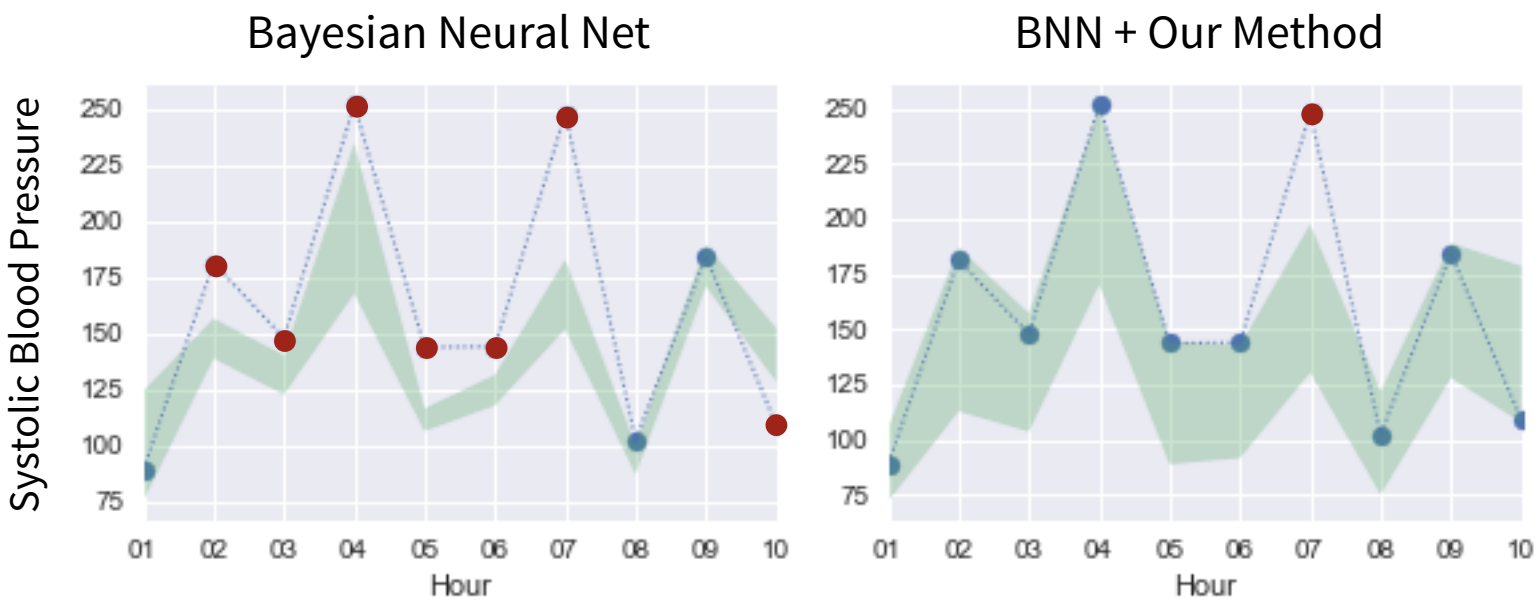
Heckerman et al. (1989)
Kohl et al. (2018)

Robotics

Deisenroth & Rasmussen (2011)
Kahn et al. (2018)

Uncertainties in Bayesian Models Are Often Inaccurate

Blood Pressure Forecasts Over Time For Woman With Heart Disease



Most points outside 90% interval!

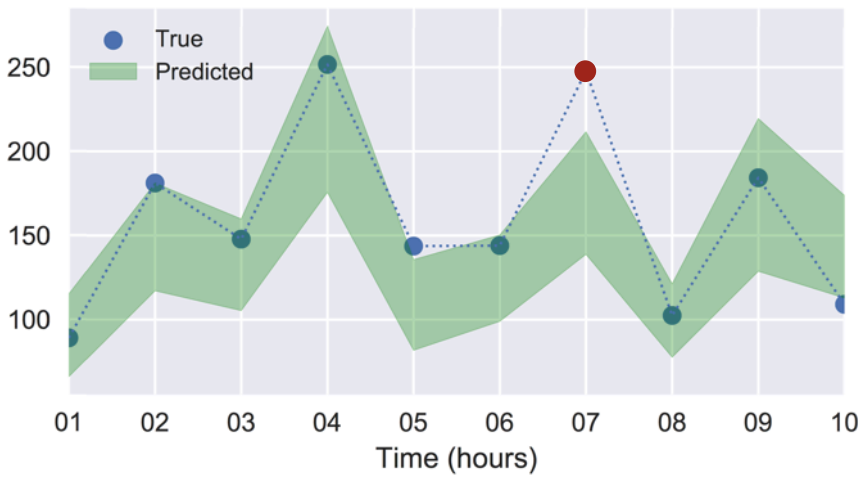
90% interval now has 9/10 points

Ground Truth Blood Pressure Measurements
90% Confidence Predictions (Bayesian Neural Net)

Guo et al. (2017)
Lakshminarayanan et al. (2017)

Ideally, Bayesian Uncertainties Should Be Calibrated

Forecasts are calibrated when predicted and empirical probabilities match.



Predicted probability for credible interval

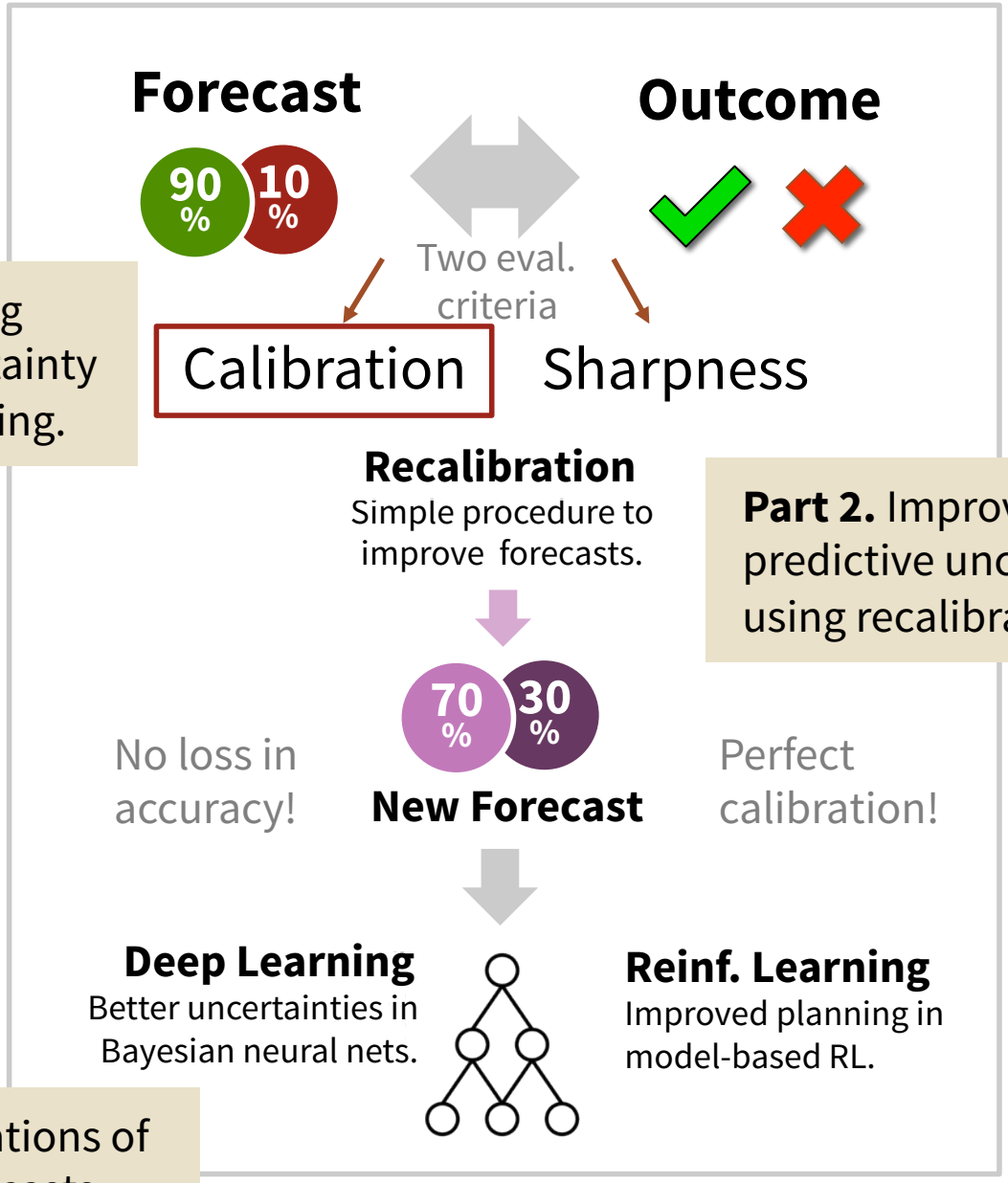
=

True probability of Y falling in the interval

This work

1. A simple way to improve predictive model uncertainty, ensuring calibration, without affecting accuracy.
2. Discuss the applications of this method to DL and RL.

Part 1. Evaluating predictive uncertainty in machine learning.

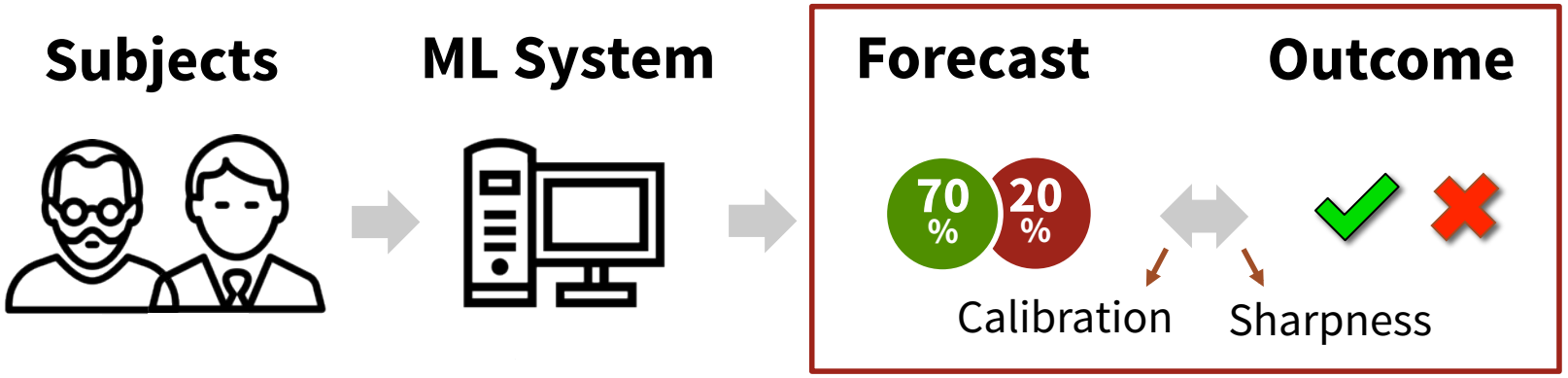


Part 2. Improving predictive uncertainty using recalibration.

Part 3. Applications of calibrated forecasts.

Running Example: Medical Risk Prediction

Our running example will be the task of predicting medical risk.



$X \sim \mathbb{P}$

Features

Genomes, medical records, etc.

$H : \mathcal{X} \to (\mathcal{Y} \to [0, 1])$

Forecaster

Outputs probability distribution over Y

$F : \mathcal{Y} \to [0, 1]$

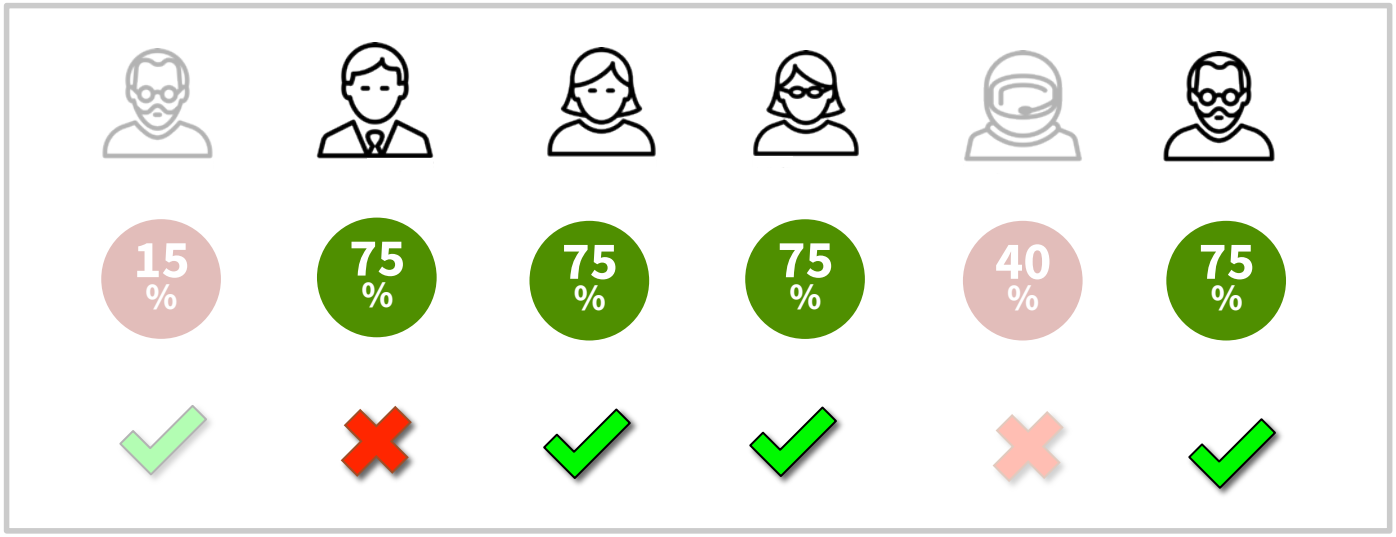
$Y \sim \mathbb{P}(Y \mid X)$

True Label

Outcome can be binary or continuous

Calibration of Predictive Uncertainty

We want predictions that marth the empirical rates of success.



Calibration: observed and predicted rates should match.

$$\underbrace{p}_{\text{predicted probability}} = \underbrace{\mathbb{P}(Y \mid F(X) = p)}_{\text{true probability of outcome}}$$

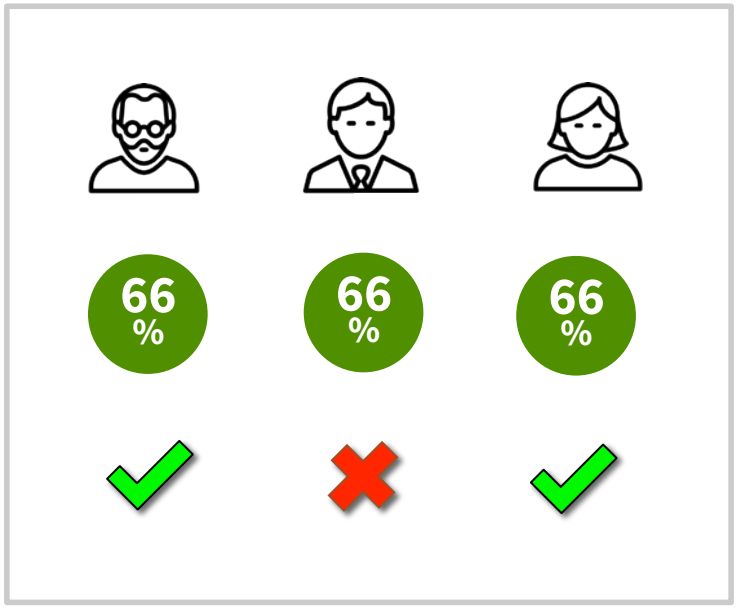
75 % predicted probability

true probability of outcome ✓.

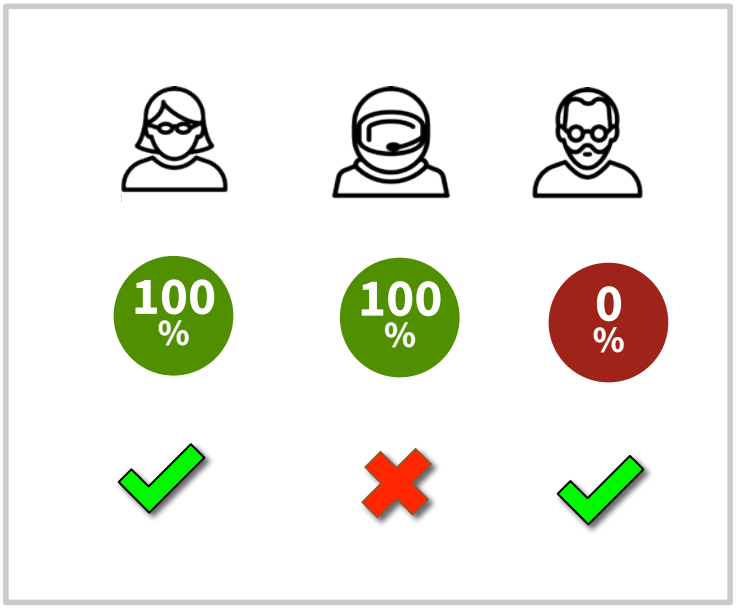
3 ✓ out of 4 trials

Calibration By Itself Is Not Enough: We Need Sharpness

Ideal predictions should also be maximally certain about the outcome.



Calibrated, but not sharp



Sharp, but not calibrated

Sharpness: Predictions should be close to one or zero.

The Calibration/Sharpness Decomposition

Most loses decompose precisely into calibration and sharpness.

Lemma (informal)**.** [Brier, 1950; Brocker 2007]. Any proper loss function decomposes into the sum of a calibration and a sharpness term.

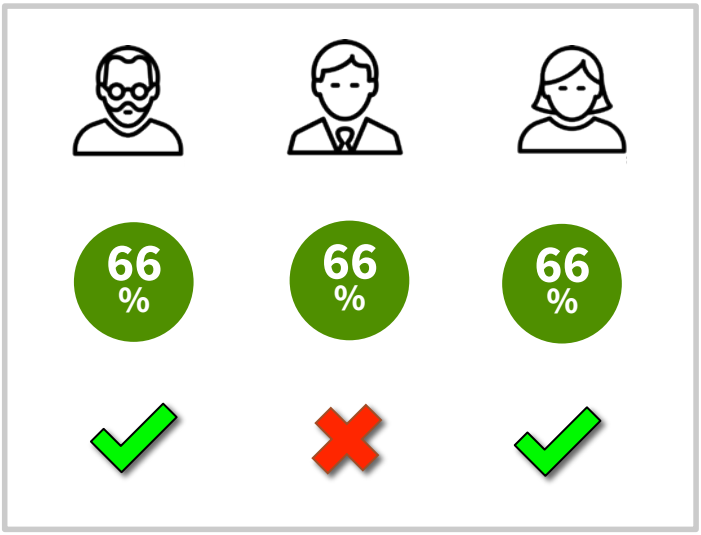
Proper Loss = **Calibration** + **Sharpness** + **Uncertainty**

Proper Loss: If we had to forecast only one probability, it would be the empirical success %.

$$\underbrace{\arg\min_{q} \left[\operatorname*{avg}_{t=1,\ldots,T} L(q, y_t) \right]}_{\text{minimizer of average loss}} = \underbrace{r_T}_{\text{empirical success rate}}$$

66 %

2 ✓
1 ✗



The Calibration/Sharpness Decomposition

Most loses decompose precisely into calibration and sharpness.

Lemma (informal)**.** [Brier, 1950; Brocker 2007]. Any proper loss function decomposes into the sum of a calibration and a sharpness term.

Proper Loss **=** **Calibration** **+** **Sharpness** **+** **Uncertainty**

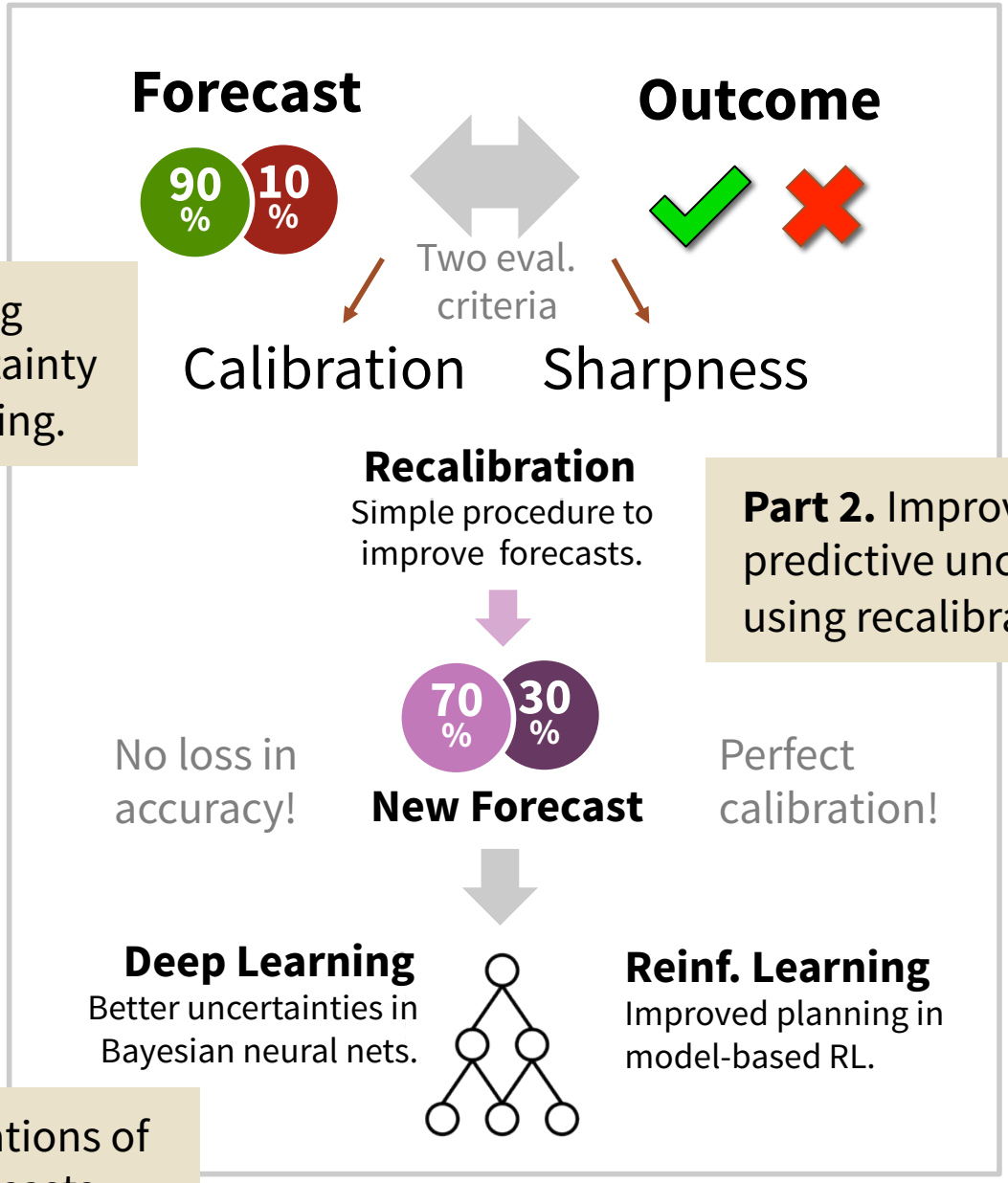
An example:

$$\underbrace{\mathbb{E}\left[Y - F(X)\right]^2}_{\text{L2 Loss}} = \underbrace{\mathbb{E}\left[p - T(p)\right]^2}_{\text{Calibration Error}} - \underbrace{\text{Var}[T(p)]}_{\text{Variance}} + \underbrace{\text{Var}[Y]}_{\text{Irreducible Error}}$$

where $T(p) = \mathbb{P}(Y \mid F(X) = p)$ is outcome probability given forecast of p

Other proper losses: log-loss, exponential loss, continuous probability rank score.

Part 1. Evaluating predictive uncertainty in machine learning.

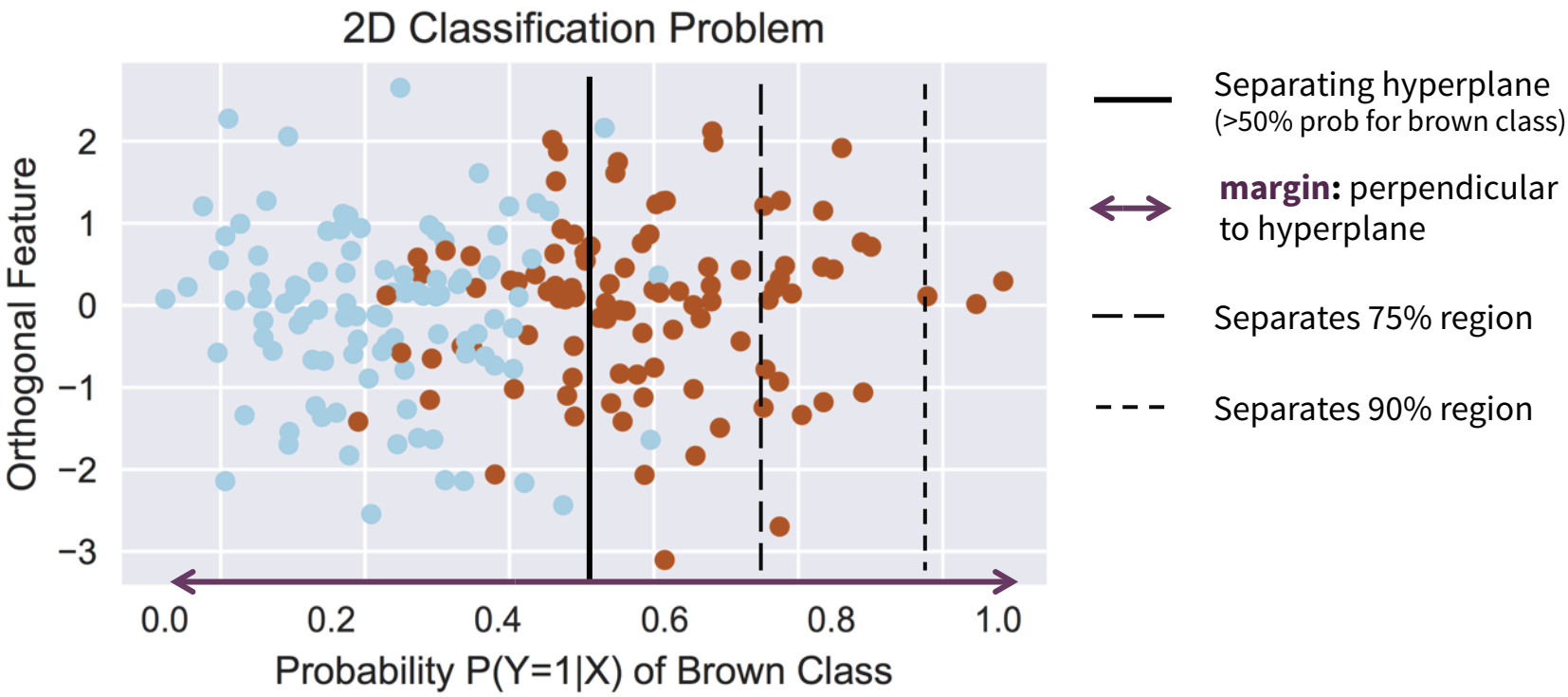


Part 2. Improving predictive uncertainty using recalibration.

Part 3. Applications of calibrated forecasts.

Why Are All Models Not Calibrated Out Of The Box?

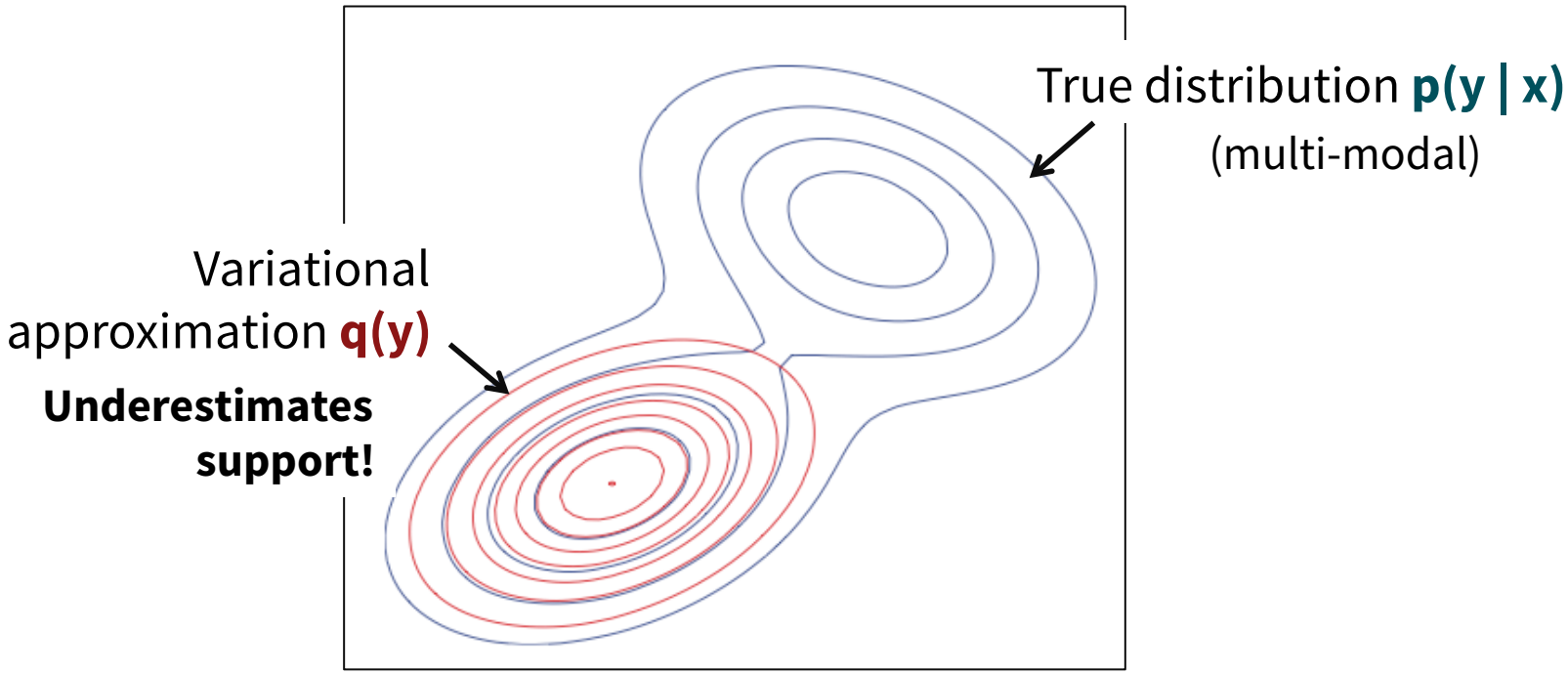
One reason is model *mis-specification*: model is not sufficiently expressive.



A linear classifier cannot accurately represent all the credible regions!

Why Are All Models Not Calibrated Out Of The Box?

The second reason is the use of computational approximations.

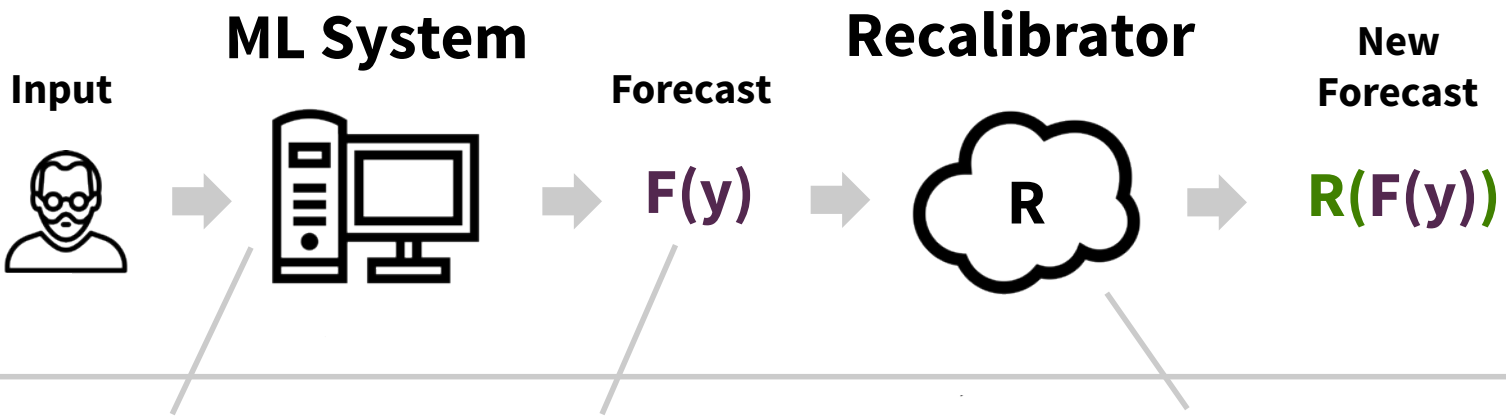


Variational inference algorithms find an approximation **q** to **p** by minimizing **KL(q||p)**, which is mode-seeking and is often over-confident [Bishop, 2006].

Recalibration

Technique that **transforms** predictions into ones that are **calibrated.**

Platt, 1999; Zadrozny and Elkan, 2002



$H : \mathcal{X} \to (\mathcal{Y} \to [0, 1])$

Can be any model (treated as black box)

$F : \mathcal{Y} \to [0, 1]$

Uncalibrated forecast

$R : [0, 1] \to [0, 1]$

Transforms probabilities coming out of **F(y)**

Definition of calibration:

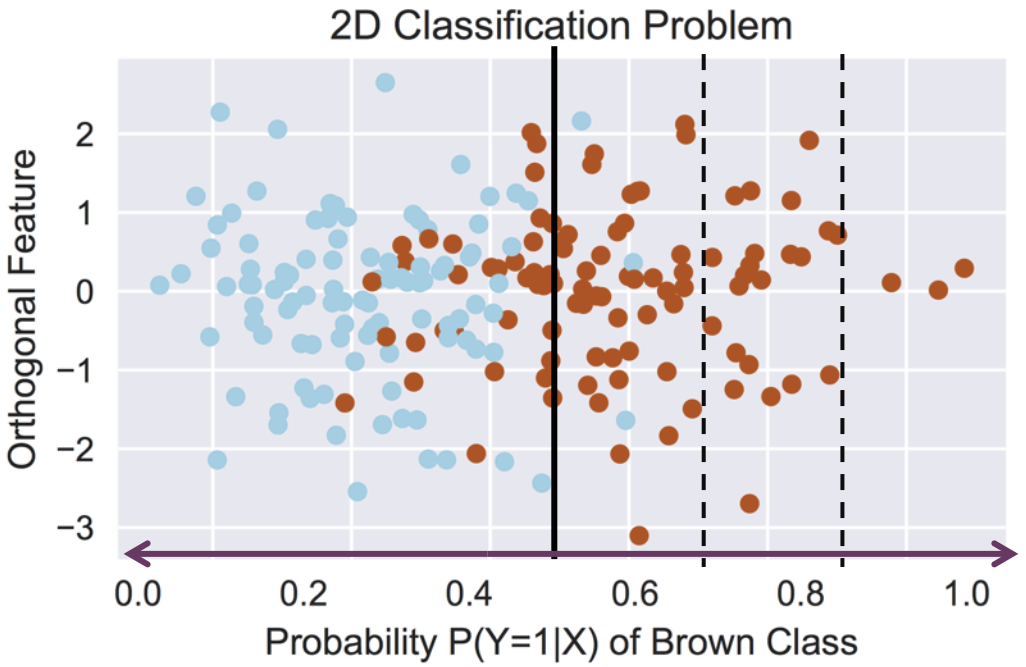
$$p = \mathbb{P}(Y = 1 \mid F(x) = p)$$

Fact: Ideal recalibrator is

$$R(p) = \mathbb{P}(Y = 1 \mid F(x) = p)$$

Explaining How Recalibration Works: An Example

Consider a logistic regression classifier on a two-dimensional dataset.



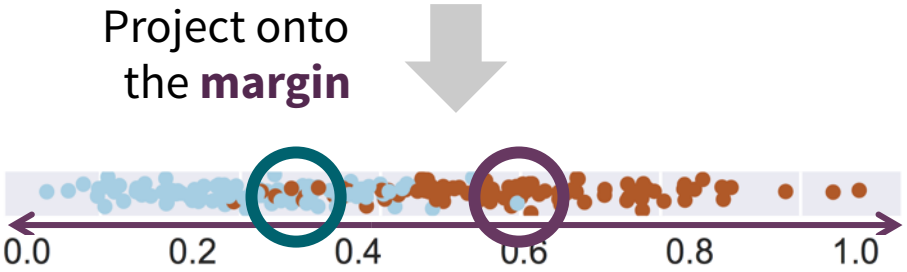
Separating hyperplane
(>50% prob for brown class)

Defines credible intervals

margin: perpendicular to hyperplane

Set of points for which F(x)=30%

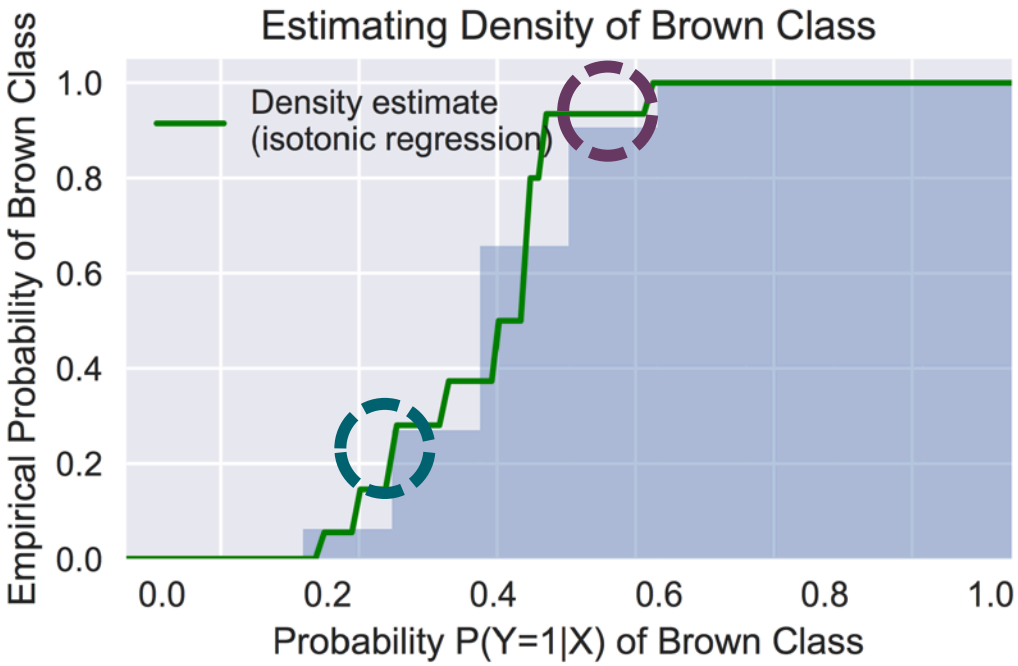
Set of points for which F(x)=60%

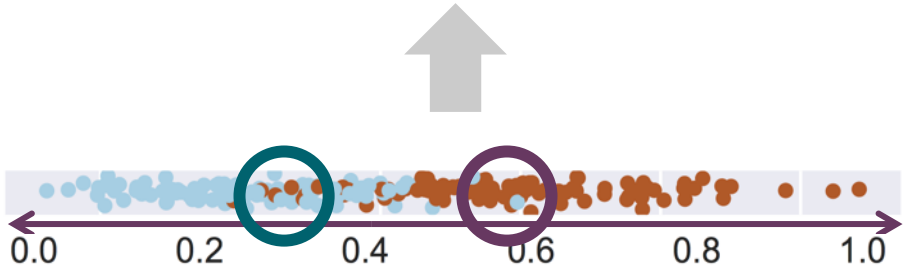


Note: The ideal recalibrator $R(p) = \mathbb{P}(Y = 1 \mid F(x) = p)$ is the density of class Y = 1 after projecting onto margin.

Explaining How Recalibration Works: Constructing R(p)

The recalibrator is the Bayes optimal classifier after projection on margin.





Points for which F(x)=30%

Points for which F(x)=90%

Actual empirical success rate when predicting 30% is 20%.

Actual empirical success rate when predicting 60% is 90%.

Note: The ideal recalibrator $R(p) = \mathbb{P}(Y = 1 \mid F(x) = p)$ is the density of class Y = 1 after projecting onto margin.

We Can Ensure Calibration Without Loss of Accuracy

We can drive calibration error down to zero without increasing initial loss.

Theorem (informal). [Kuleshov and Ermon, 2017]

The recalibrated forecasts **R(F(y))** will have the following properties:

Calibration loss of the **R(F(y))** → 0

L_2 loss **after** recalibration

also: any proper loss

$\leq$

L_2 loss **before** recalibration

as time → ∞

Calibration = Free Lunch!

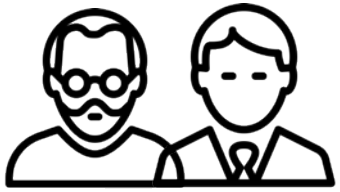
Also true for streaming data that is non i.i.d. (can be adversarial!)

(requires a different recalibrator)

Calibration for Continuous Outputs [Kuleshov et al., ICML18]

Predicted and empirical confidence intervals should match.

Subjects



Features

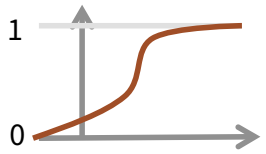
$X \sim \mathbb{P}$

ML System

Outputs prob. dist.

$H : \mathcal{X} \to (\mathcal{Y} \to [0, 1])$

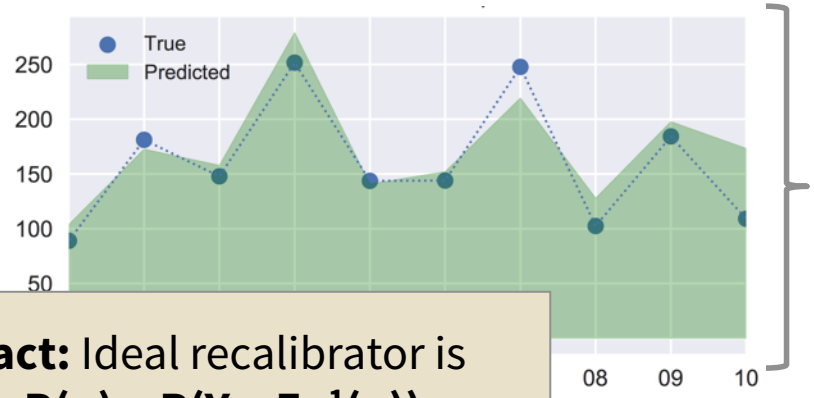
Forecast



Cumulative dist. func.

$F : \mathcal{Y} \to [0, 1]$

Calibration for Continuous Y



90 %

Predicted probability of p-th quantile

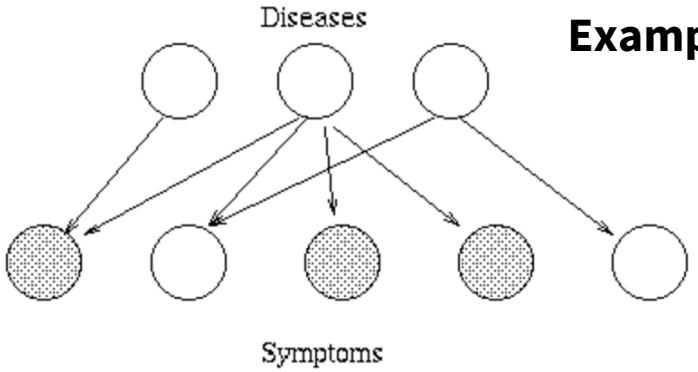
$$p = \mathbb{P}(Y \leq F_X^{-1}(p))$$

True probability of Y below p-th quantile.

Fact: Ideal recalibrator is $\mathbf{R(p) = P(Y \leq F_X^{-1}(p))}$.

Calibrated Structured Prediction [Kuleshov and Liang, NIPS15]

Sometimes, $\mathbf{y_t}$ is a vector encoding multiple correlated outcomes.



Example: Quick Medical Reference [Shwe et al.]

P(600 diseases | 4000 syptoms**)**

- ❑ Influenza
- ❑ Malaria
- ❑ Poisoning
- ❑ Cough
- ❑ Fever
- ❑ Headache

P(malaria | symptoms**)**

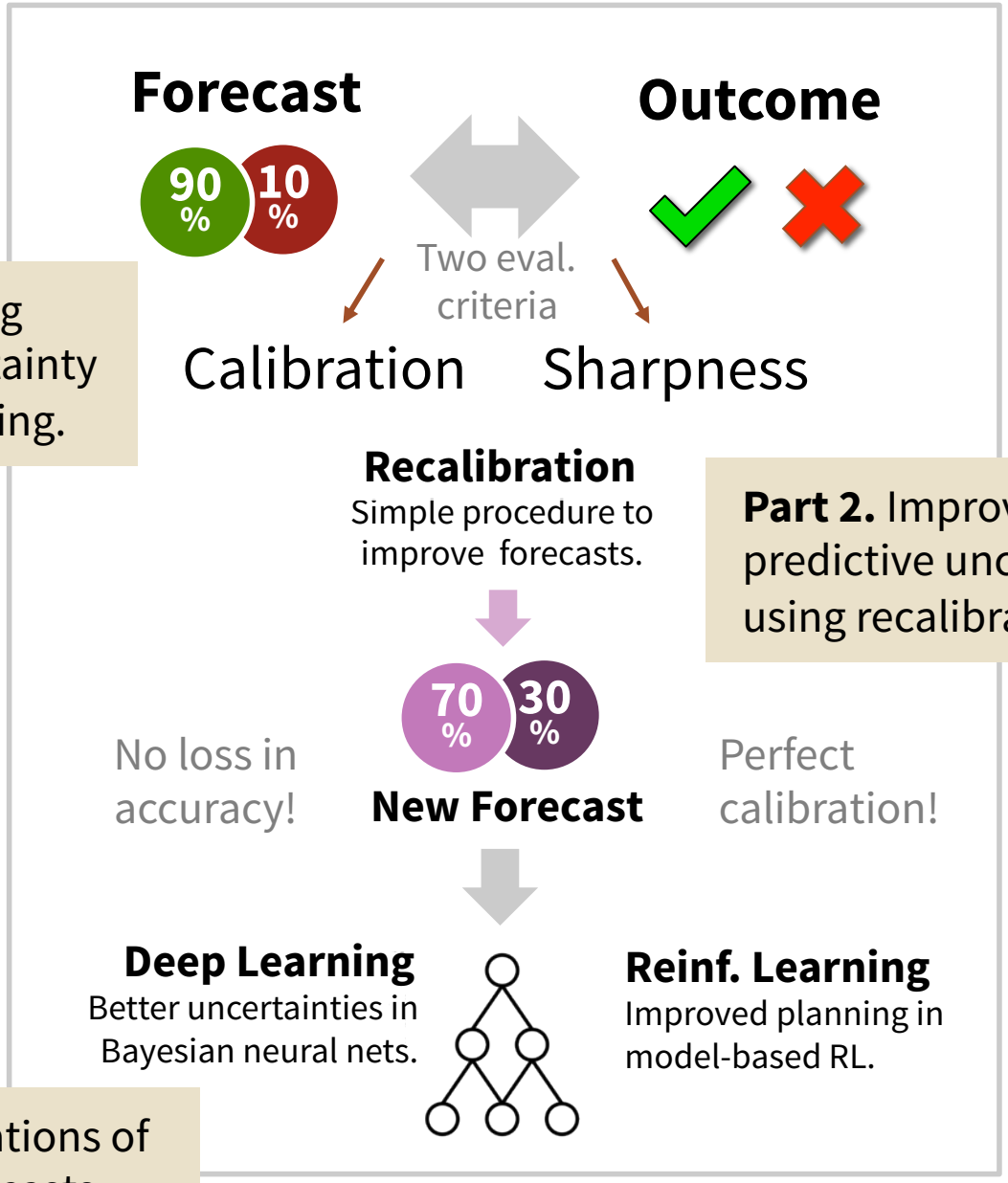
Marginal probability of any given disease

P(disease profile | symptoms**)**

Joint probability of most likely disease profile

Near-Perfect Calibration Without Loss of Accuracy

Part 1. Evaluating predictive uncertainty in machine learning.

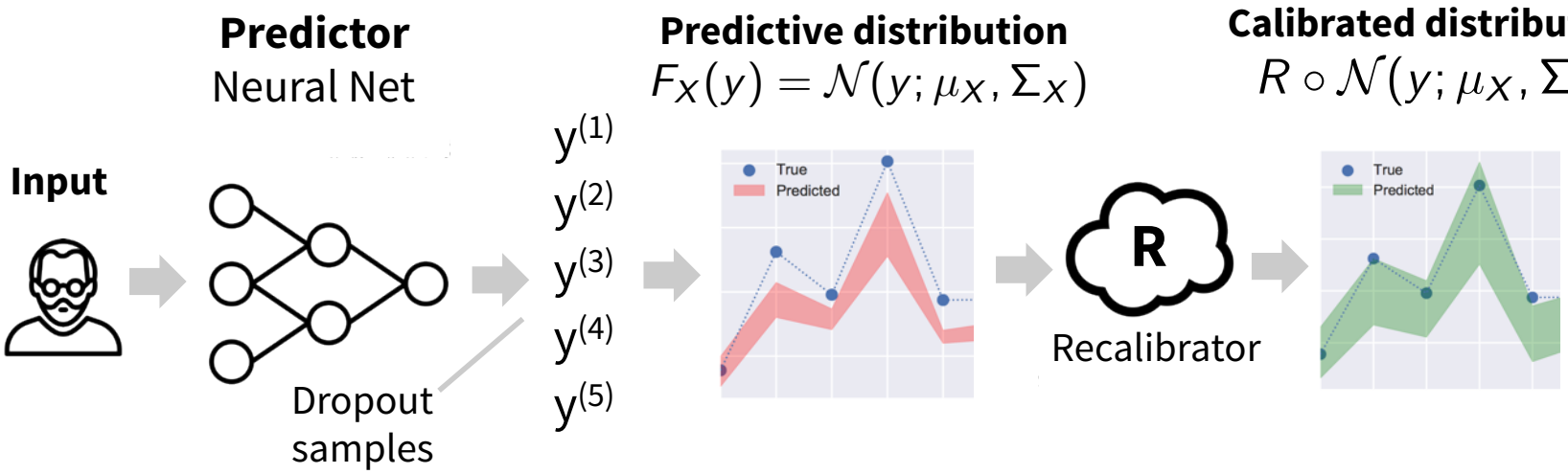


Part 2. Improving predictive uncertainty using recalibration.

Part 3. Applications of calibrated forecasts.

An Application to Deep Learning

We use our method to provide accurate uncertainties for Bayesian DL.



Our method consistently outputs well-calibrated uncertainties without any loss in predictive accuracy on multiple tasks.

Datasets for Regression
Lakshminarayanan et al. (2017)

Time Series Forecasting
Kaggle Datasets

Depth Estimation
Kendall and Gal (2017)

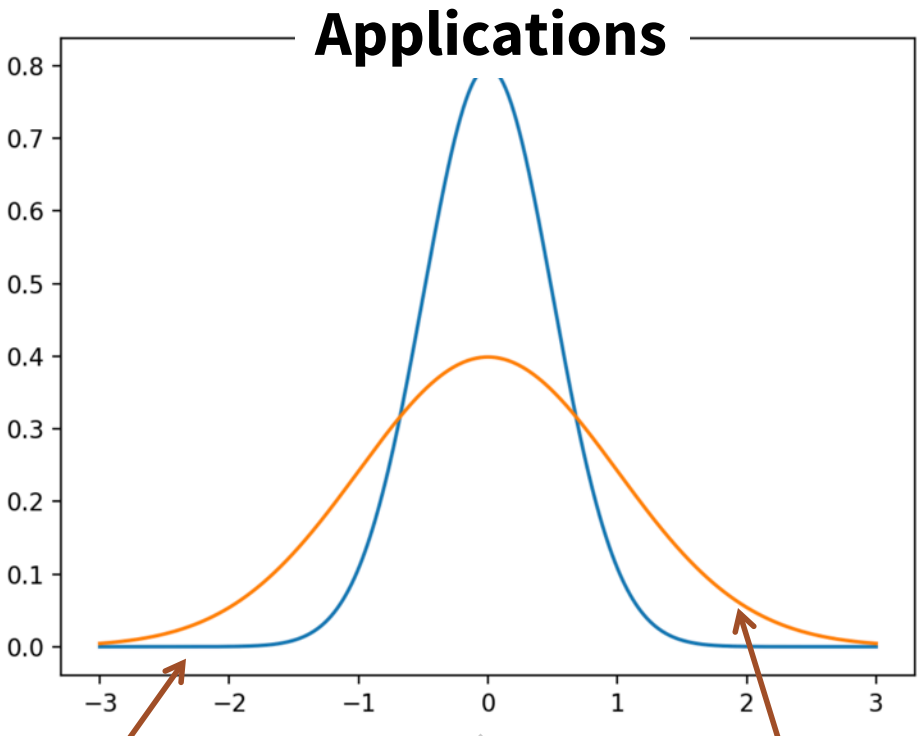
Gal and Gharamani (2016)

Calibration Is Also Important For Decision Making

Calibrated uncertainty is important for real-world decision-making tasks.

Safety

Maximizing Reward



Exploration/ Exploitation

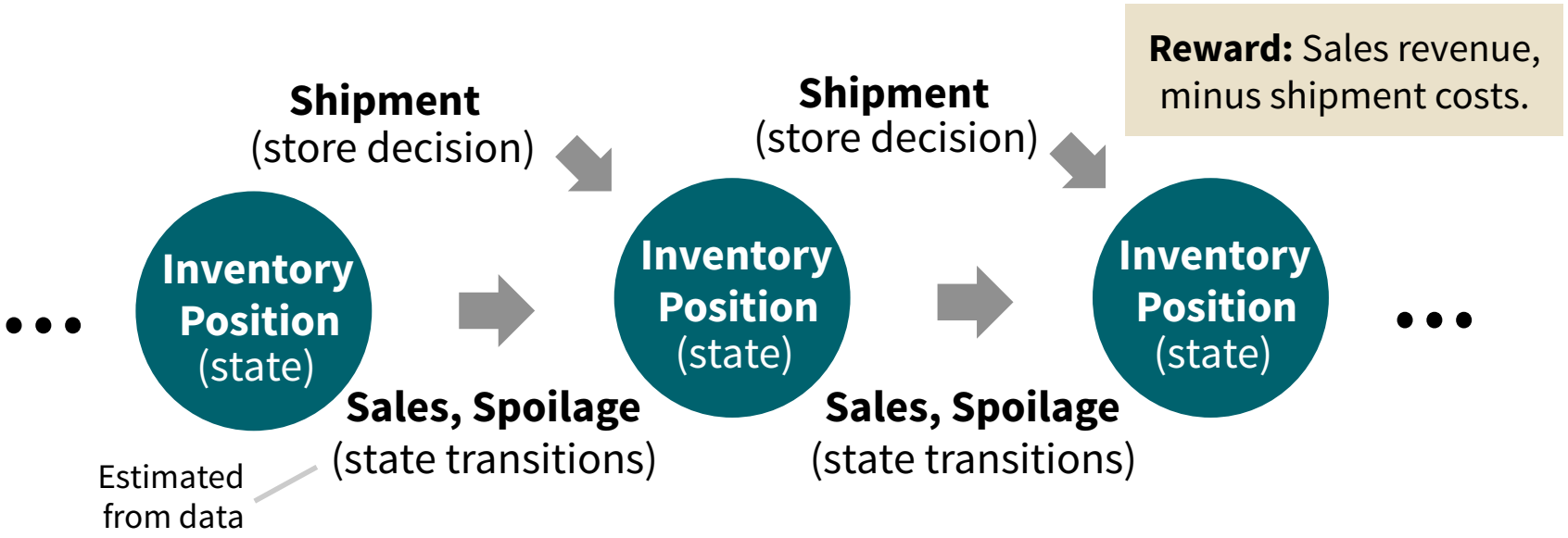
Interpretable predictions

Probability distribution of where the agent **thinks** it will go.

Probability distribution of where the agent will **actually** go.

Calibrated Model-Based Reinforcement Learning

Learning a calibrated RL model improves planning and cumulative reward: an example from inventory management (Van Roy et al., 1999).



Calibrated Transition Model Results in Higher Cumulative Reward

Concrete Dropout	Deep Ensemble	Ours
$60,082	$60,894	$61,690

Philosophical Musings: Model Criticism and Box's Loop

Recalibration combines Bayesian and frequentist techniques.

A Bayesian would have included calibrated models in the set of all valid models, and integrated over them.

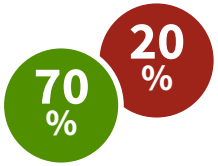
Recalibration within Box's Loop

Build Model

Fit forecaster **H**

$H : \mathcal{X} \to (\mathcal{Y} \to [0, 1])$

Apply Model



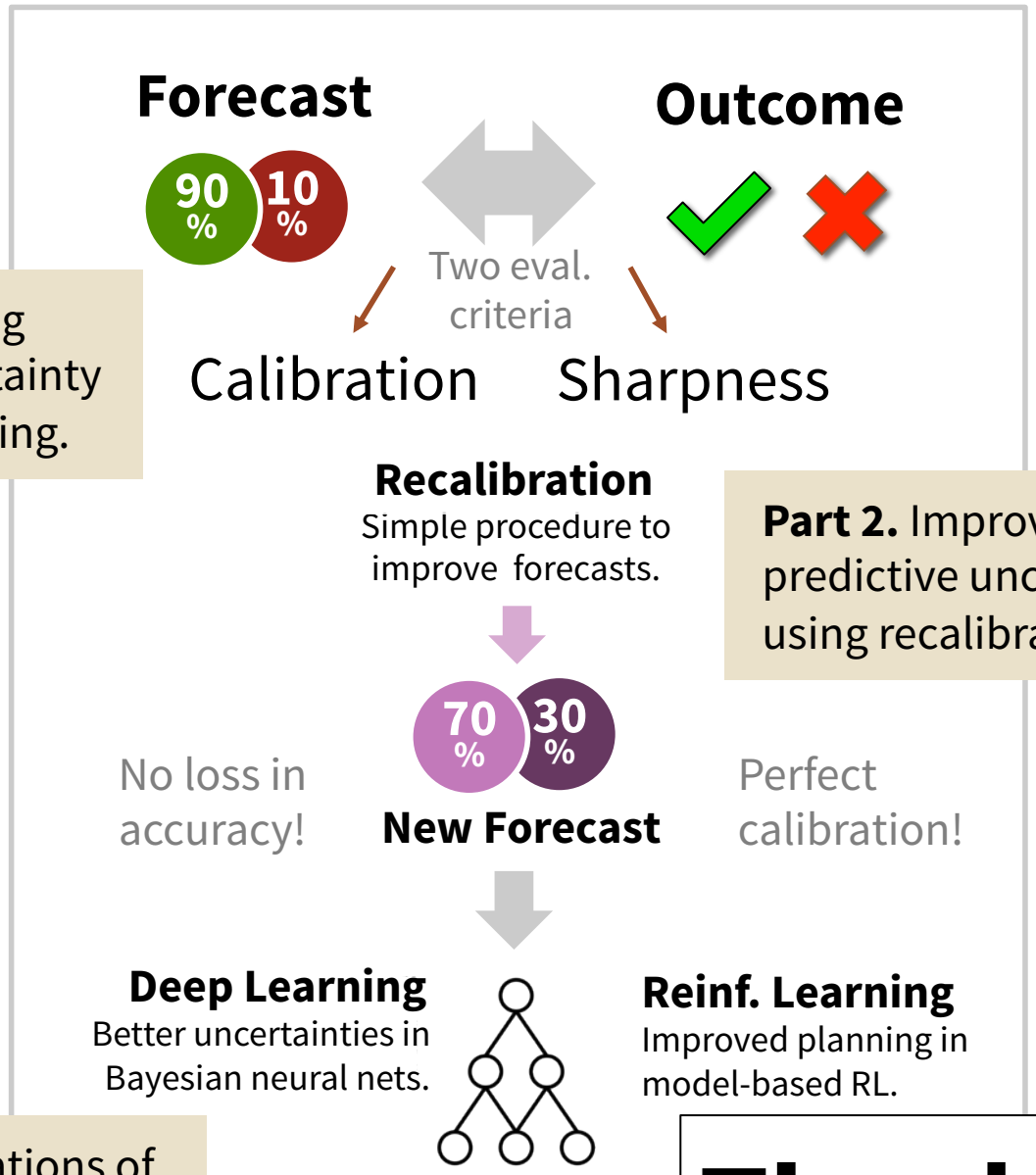
Make predictions

$F : \mathcal{Y} \to [0, 1]$

Criticize

Recalibrate

$R \circ \mathcal{N}(y; \mu_X, \Sigma_X)$



Part 1. Evaluating predictive uncertainty in machine learning.

Part 2. Improving predictive uncertainty using recalibration.

Part 3. Applications of calibrated forecasts.

Thank you!