High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach
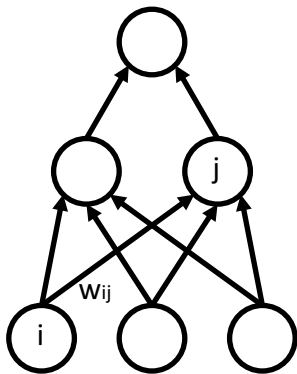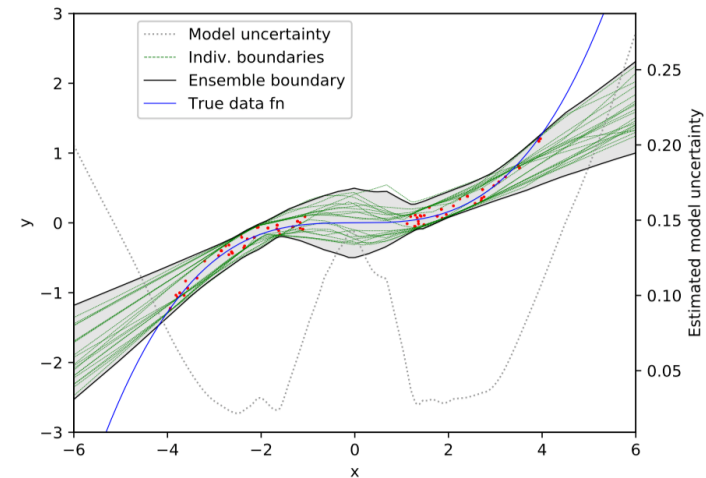
Tim Pearce

June 2018

UNIVERSITY OF CAMBRIDGE
Cambridge Service Alliance

The Alan Turing Institute

tp424@cam.ac.uk

Andy Neely | Mohamed Zaki | Alexandra Brintrup

Contents

- Introduction to deep learning

- Motivating uncertainty

- Uncertainty philosophy

- Existing approaches

- High-Quality Prediction Intervals
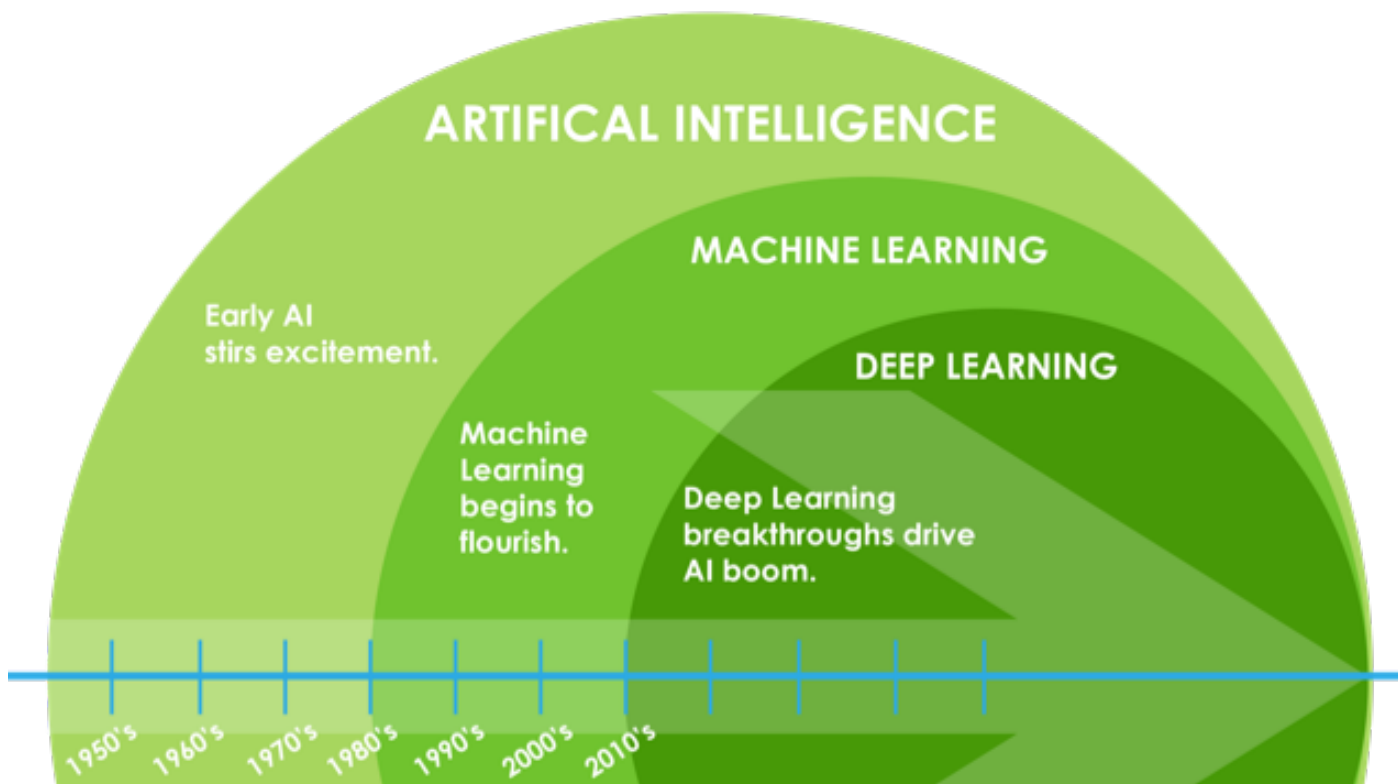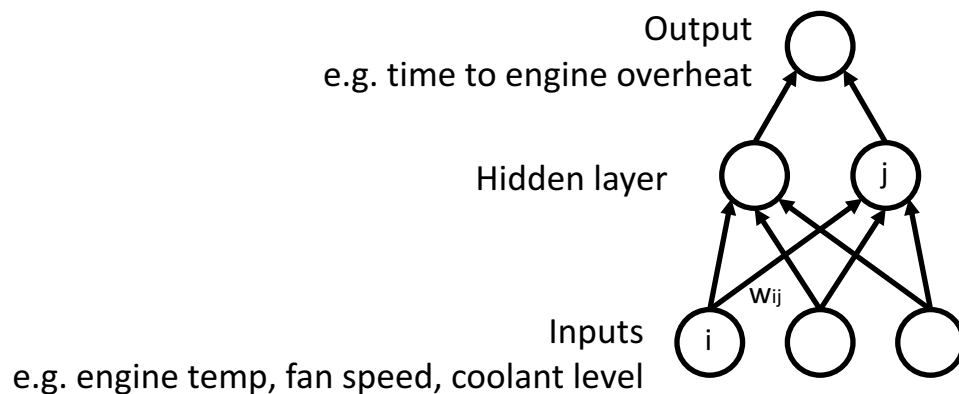
- Results

- Real-world application
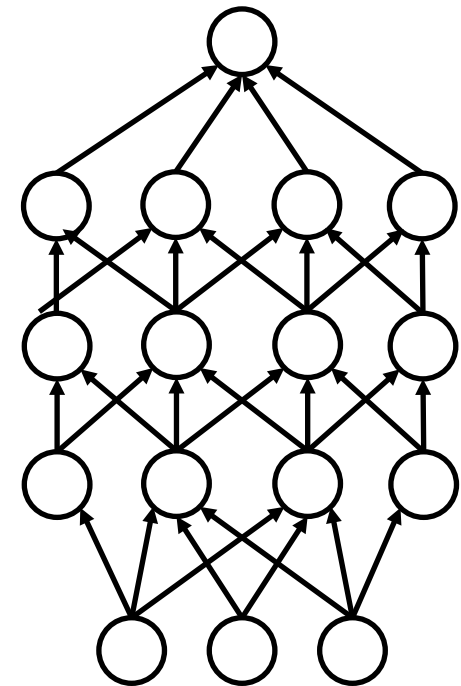
A Brief Introduction to Deep Learning

Deep Learning

Deep Learning ≈ Neural Networks

- NNs with many hidden layers

- Powerful function approximation ability – universal approximation theorem

- Step-change performance improvement in range of applications – computer vision, speech, NLP, recommendation systems...
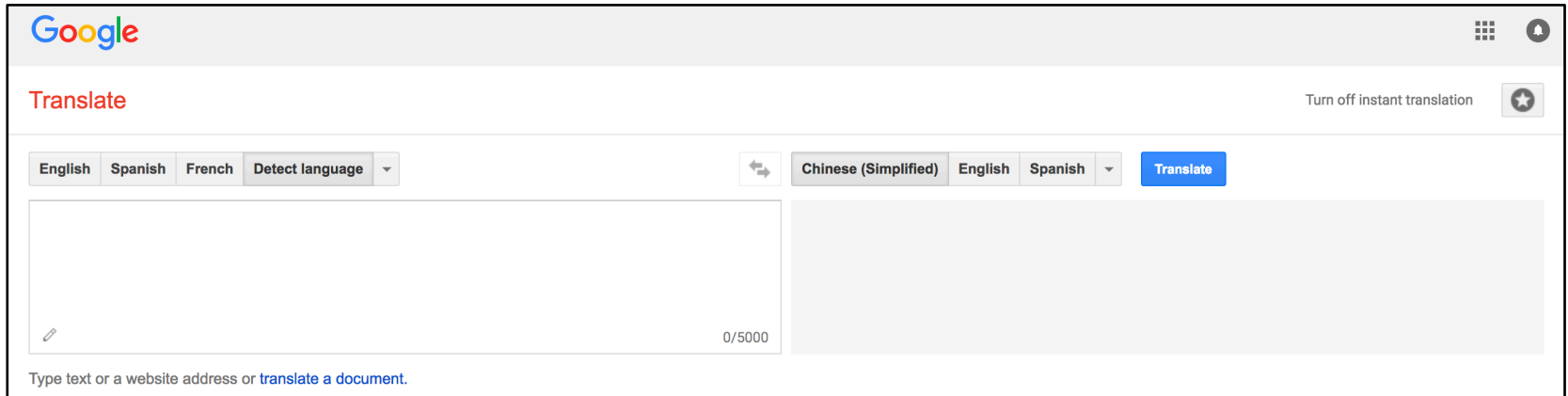
Output
e.g. time to engine overheat

Hidden layer

w_{ij}

Inputs
e.g. engine temp, fan speed, coolant level

Add hidden layers

1980's
Neural Networks

2010's
'Deep Learning'

Deep Learning

"Google shrinks language translation code from 500,000 to 500 lines with AI"
Import AI: #63
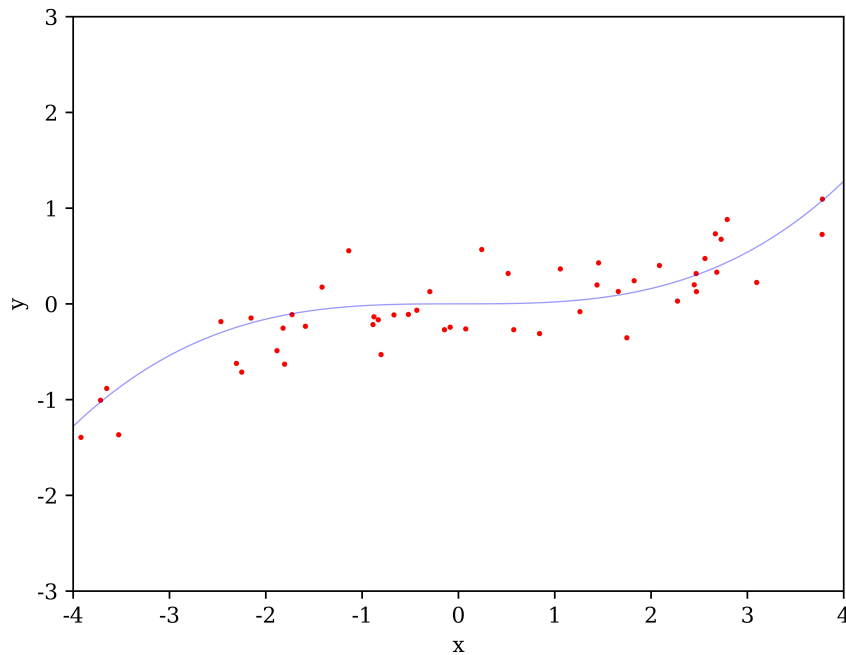
Motivating Uncertainty

Motivating Uncertainty

1. A factory manager is notified by an algorithm that a machine will fail in **60 days**… How do they schedule maintenance on this information? Does it need repair tomorrow, or can it be run for 59 days?

2. A factory manager is notified by an algorithm that a machine will fail in **between 45-65 days with 99% probability**… timing of a repair is easily scheduled.

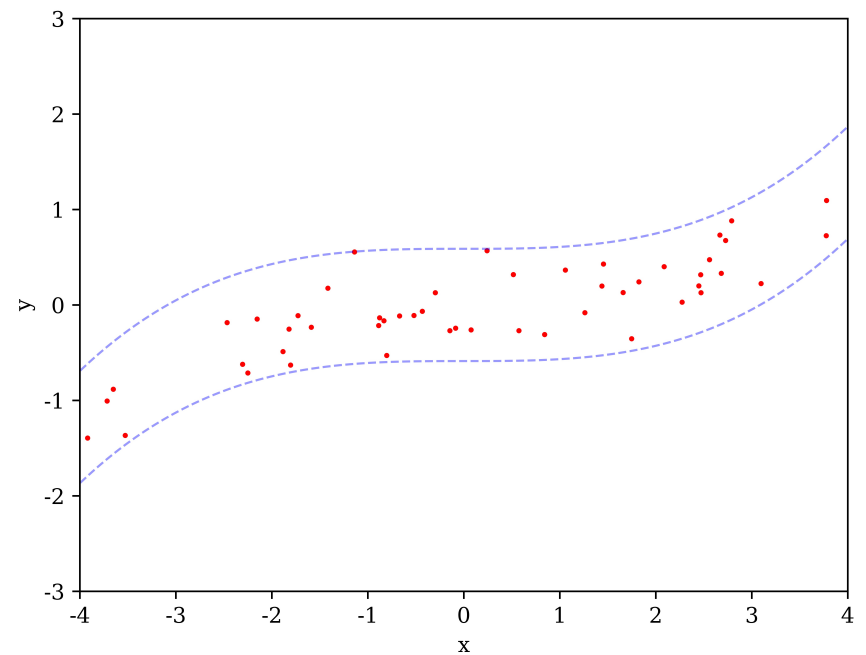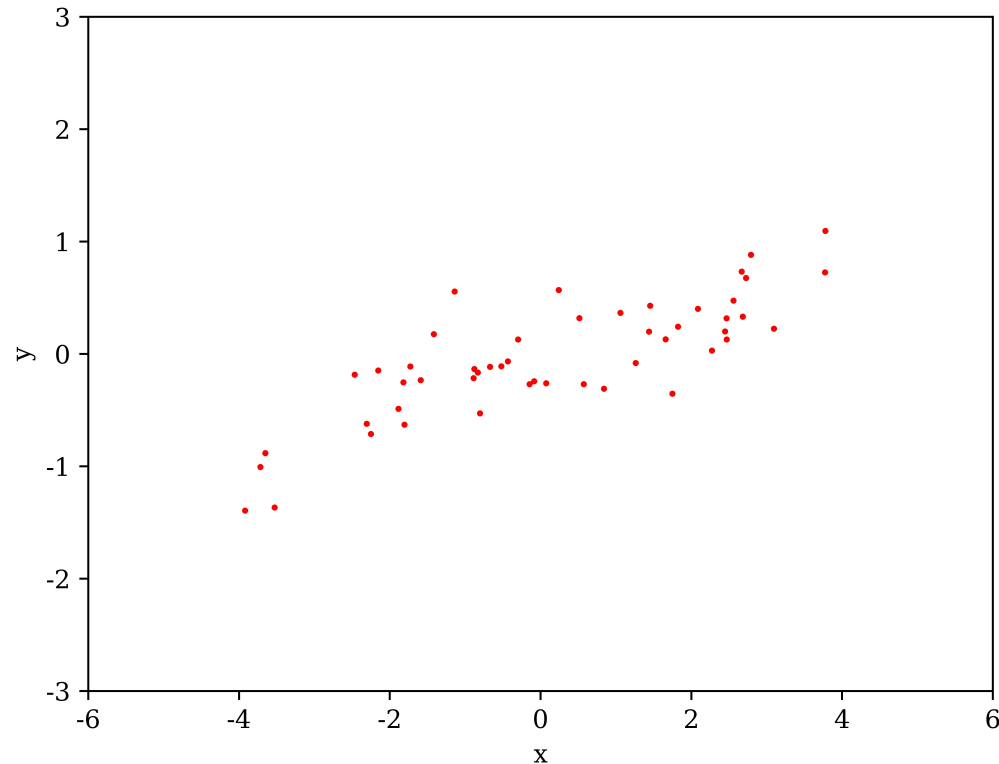How can we make our algorithms do this?

Motivating Uncertainty

Point estimate ➡ Prediction Interval (PI)

Motivating Uncertainty

- Given x, what is the expected value of y? $E(y|x)$
- How **certain** are we of this estimate? $Var(y|x)$
- Model predictive distribution, $Pr(y|x)$
- Model Prediction Interval (PI), y_upper, y_lower

Uncertainty Philosophy

Uncertainty Philosophy
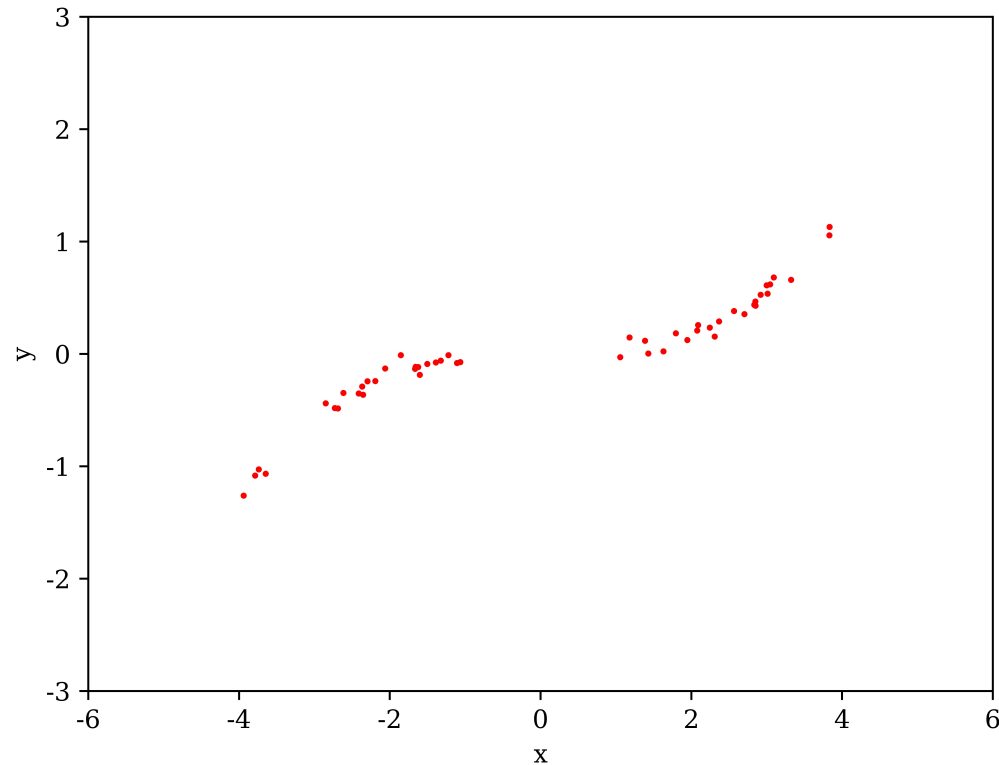
- Regression philosophy

$$y = f(\mathbf{x}) + \epsilon$$

- Uncertainty of y

$$\sigma_y^2 = \sigma_{model}^2 + \sigma_{noise}^2$$

- Input covariates, x
- Outputs, y
- Data generating function, f(.)
- Noise, e

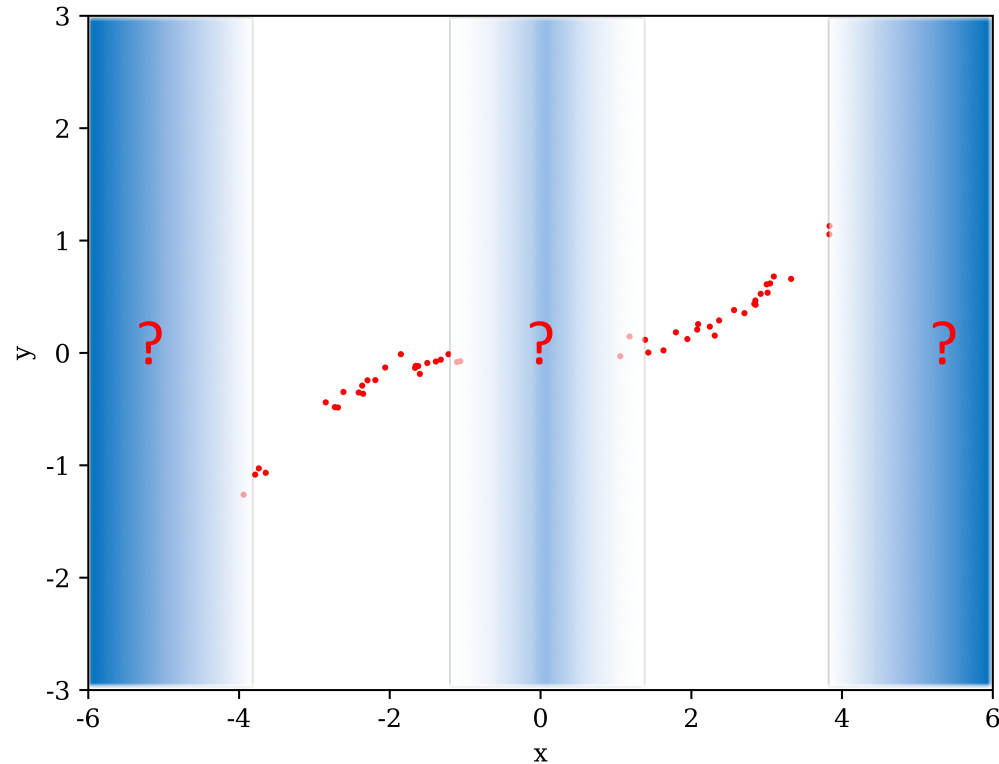# Uncertainty Philosophy

- What's the certainty of our prediction when x=2?
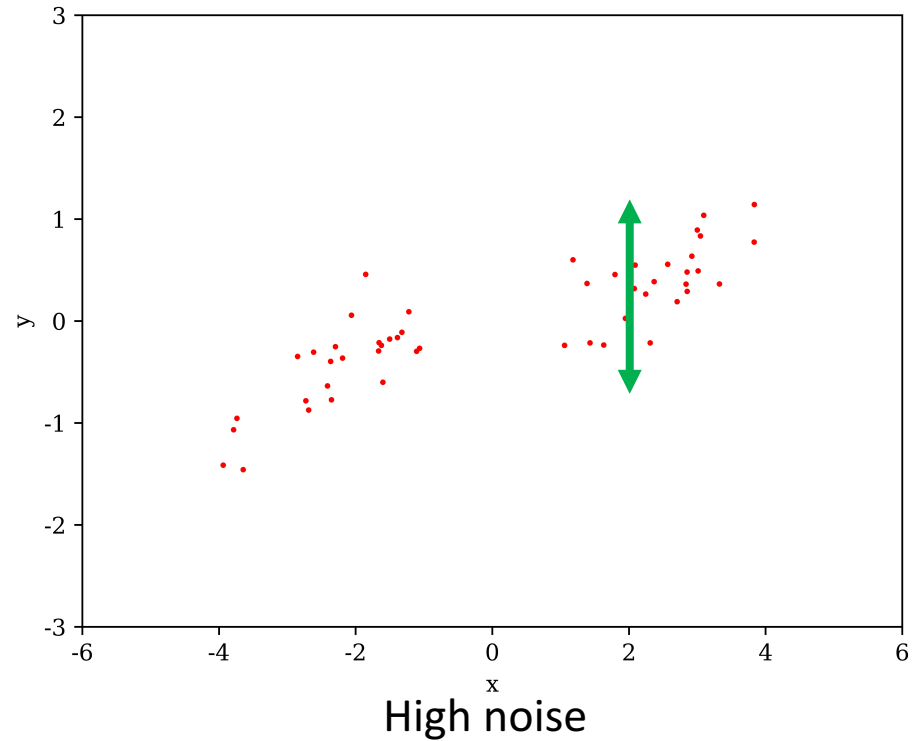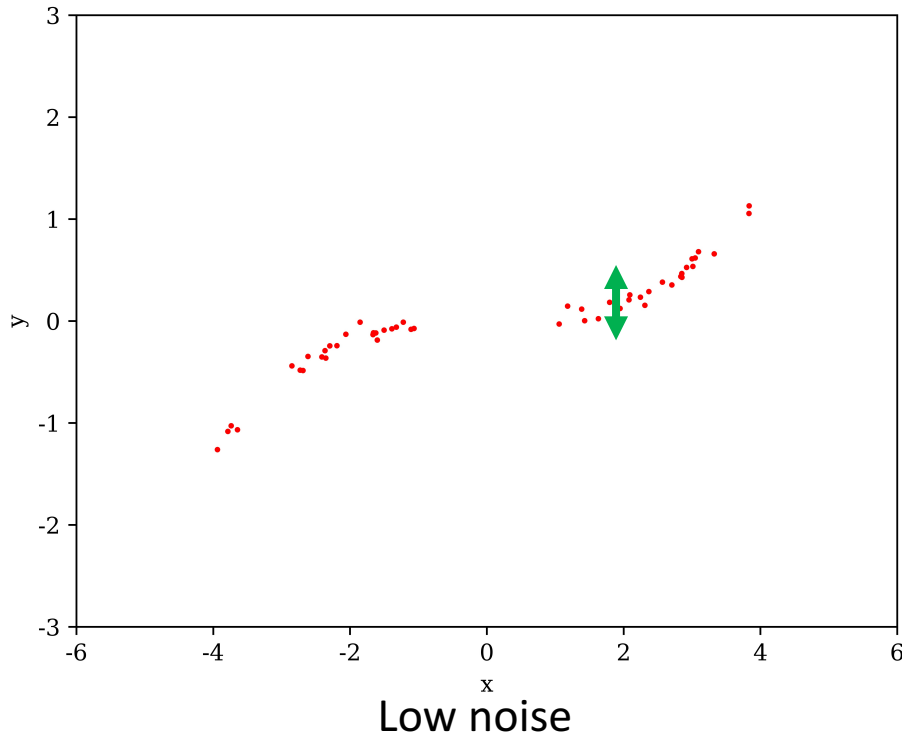- How about x=0?
- How about x= 10?

Uncertainty Philosophy

- Uncertainty grows in regions we don't have data
- Model uncertainty (*epistemic*)

Uncertainty Philosophy

Low noise

High noise

- Data noise variance (*aleatoric*)

Uncertainty Philosophy

- Regression philosophy

$$y = f(\mathbf{x}) + \epsilon$$

- Uncertainty of y

$$\sigma_y^2 = \boxed{\sigma_{model}^2} + \boxed{\sigma_{noise}^2}$$

Model Uncertainty Data Noise Uncertainty

Uncertainty in Deep Learning

Existing Approaches

Uncertainty in Deep Learning

- Generally a Neural Network (NN) outputs a single point estimate
- Existing methods to output uncertainty

Model uncertainty:
- Bayesian Neural Networks
- Dropout as Variational Inference
- Ensembling
- Conformal Prediction

Data noise uncertainty
- Mean Variance Estimation
- Lower Upper Bound Estimation

High-Quality Prediction Intervals for Deep Learning

A new method to capture data variance
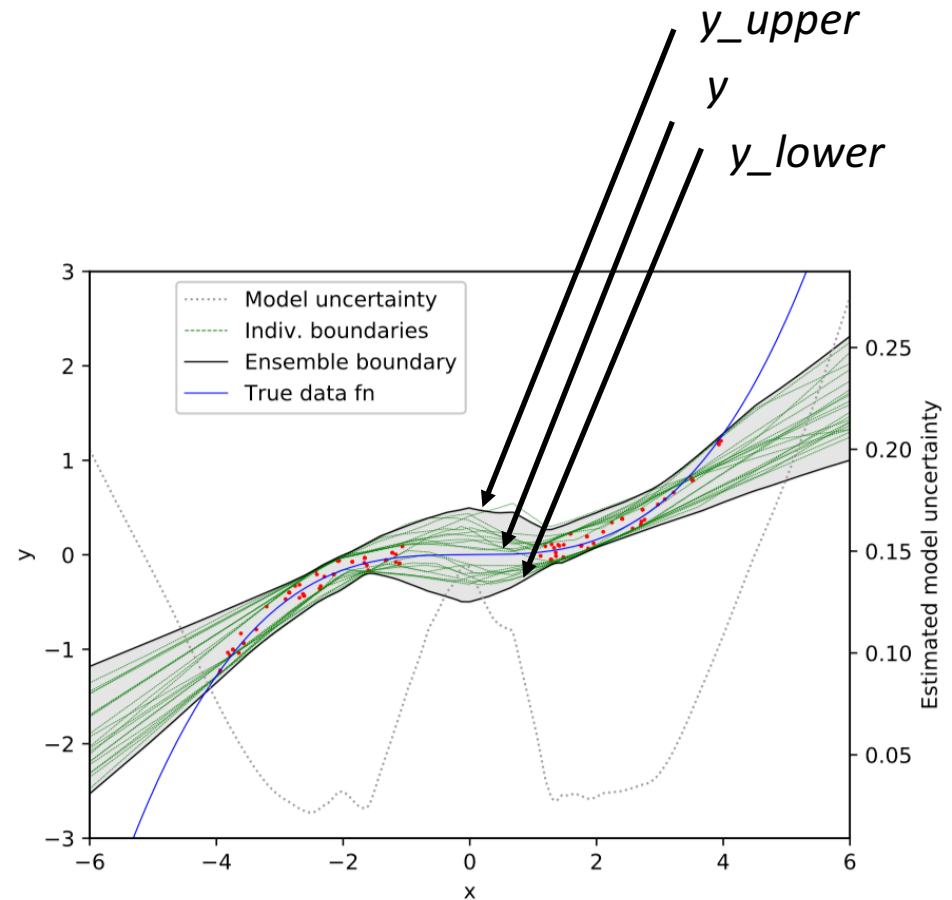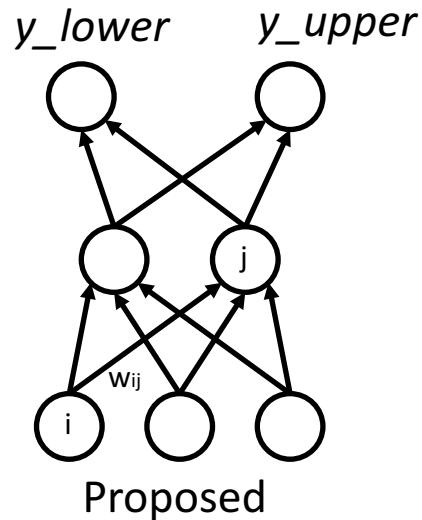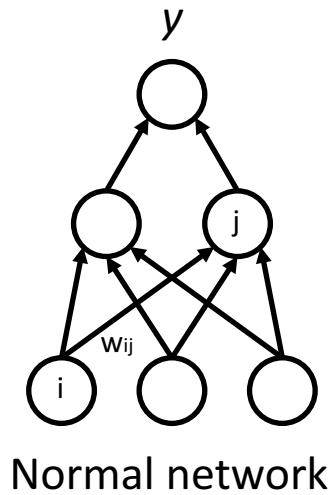
- The objective of Prediction Intervals
  - Narrow width – not 0-1000 days
  - Whilst capturing some predefined proportion

So why not use this directly as the objective function?

Prediction Intervals and Neural Networks

Prediction Intervals and Neural Networks

Let the set of input covariates and target observations be $\mathbf{X}$ and $\mathbf{y}$, for n data points, and with $\mathbf{x}_i \in \mathbb{R}^D$ denoting the ith D dimensional input corresponding to y_i, for $1 \leq i \leq n$. The predicted lower and upper PI bounds are $\hat{\mathbf{y}}_{\mathbf{L}}, \hat{\mathbf{y}}_{\mathbf{U}}$. A PI should capture some desired proportion of the observations, $(1 - \alpha)$, common choices of α being 0.01 or 0.05,

A vector, $\mathbf{k}$, of length n represents whether each data point has been captured by the estimated PIs, with each element $k_i \in \{0, 1\}$ given by,

$$k_i = \begin{cases} 1, & \text{if } y_{Li} \leq y_i \leq y_{Ui} \\ 0, & \text{else.} \end{cases} \tag{4}$$

We define the total number of data points captured as c,

$$c := \sum_{i=1}^{n} k_i. \tag{5}$$

Prediction Intervals and Neural Networks

- Mean Prediction Interval Width (MPIW)

$$MPIW := \frac{1}{n} \sum_{i=1}^{n} \hat{y}_{Ui} - \hat{y}_{Li}.$$

- Prediction Interval Coverage Proportion (PICP)

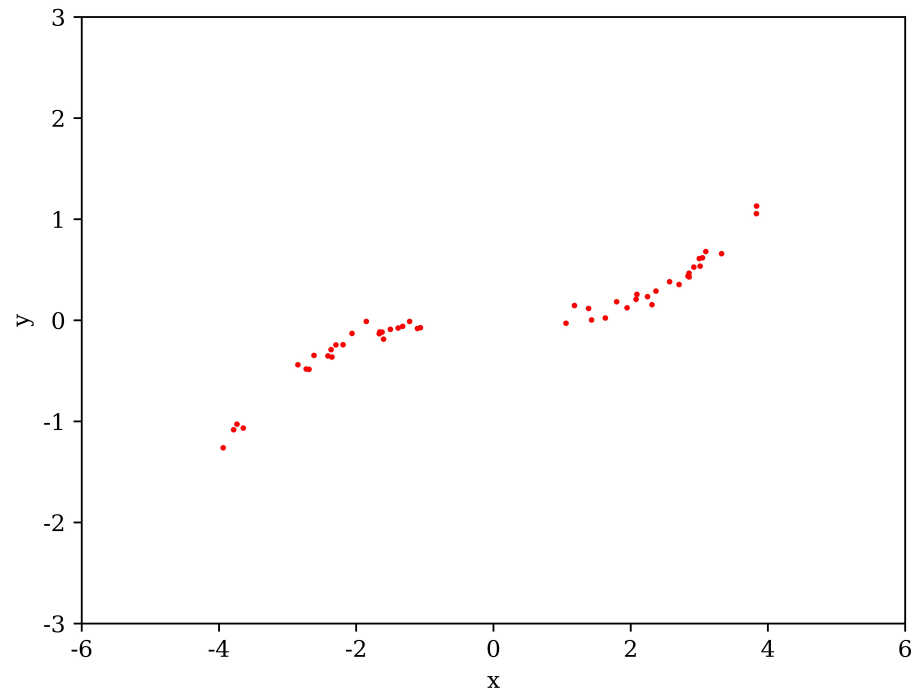$$PICP := \frac{c}{n},$$

Prediction Intervals and Neural Networks

- *[skipped the maths in between...]*

- The objective function...

$$Loss_{QD} = MPIW_{capt.} + \lambda \frac{n}{\alpha(1-\alpha)} \max(0, (1-\alpha) - PICP)^2$$
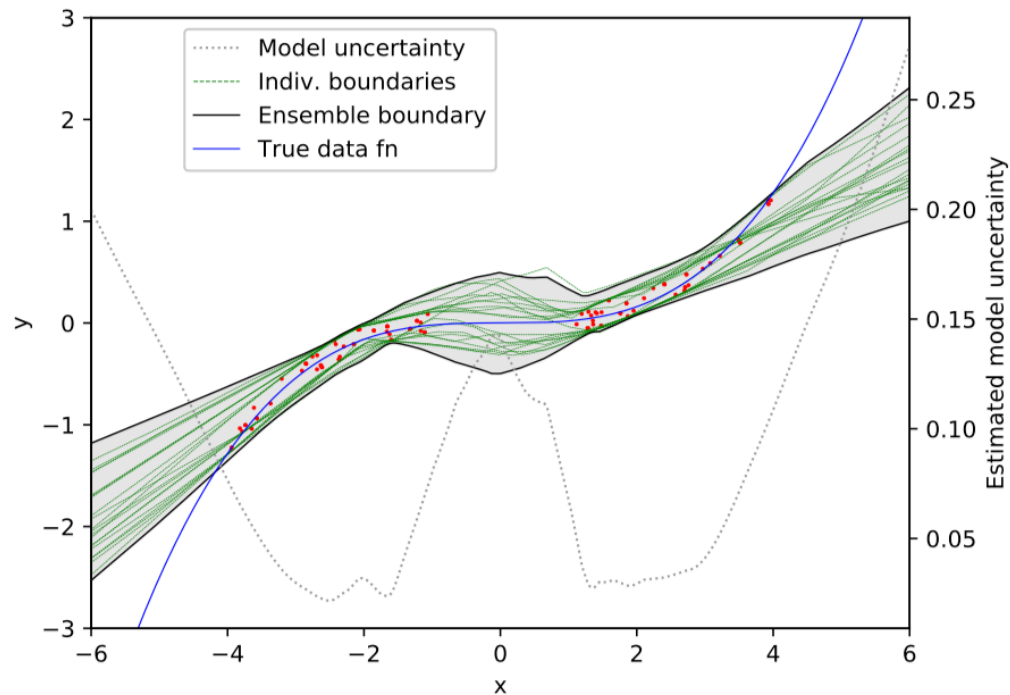
- Showed how to use it with gradient descent

- Also doesn't capture 'model uncertainty', proposed ensembling for this

- Named the **Quality-Driven (QD) method**
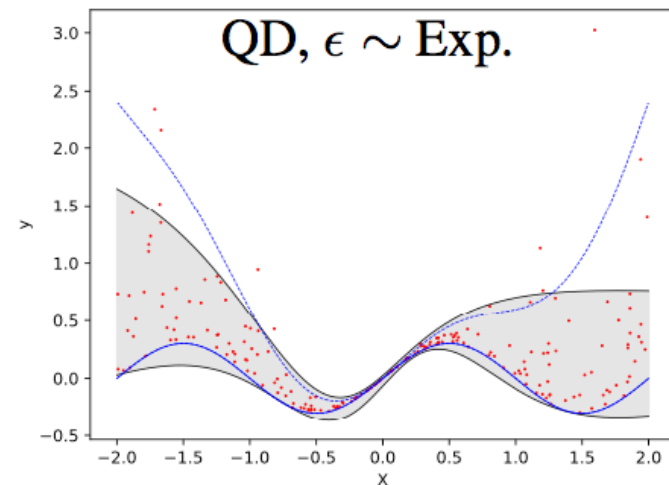
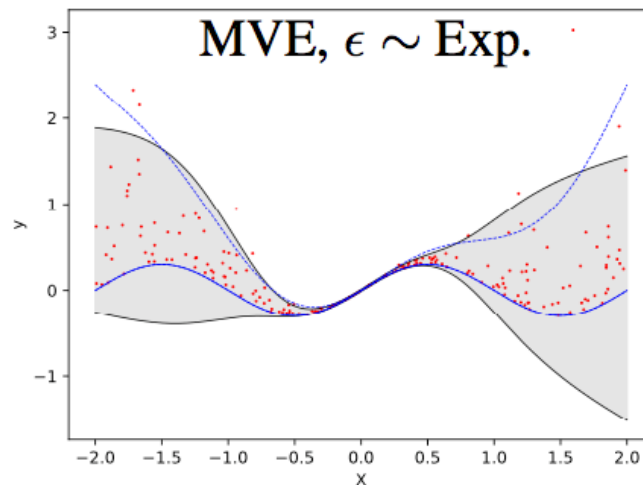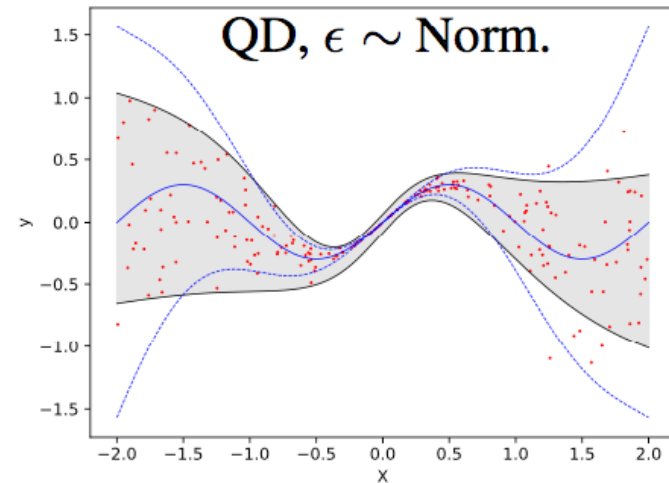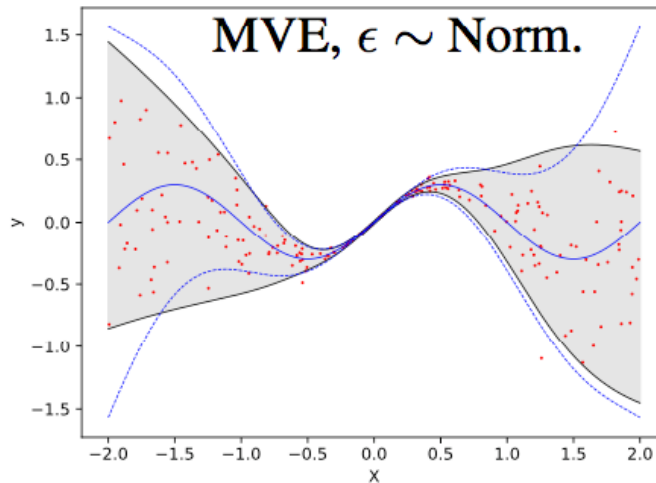# Prediction Intervals and Neural Networks

Prediction Intervals and Neural Networks

Prediction Intervals and Neural Networks

- Comparing to model with distributional assumption

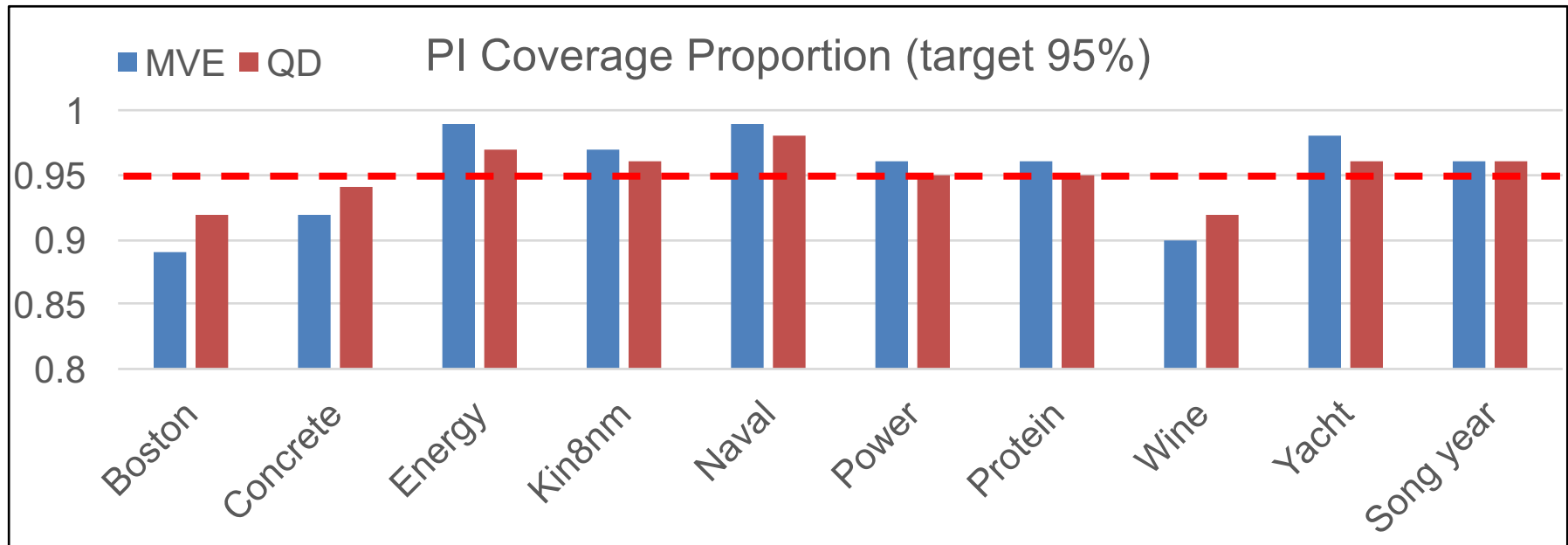$$y = 0.3\sin(x) + 0.2\epsilon$$

Results

Results

- Ten open-access benchmarking datasets

Results

- Ten open-access benchmarking datasets

Mean PI Width (smaller better)
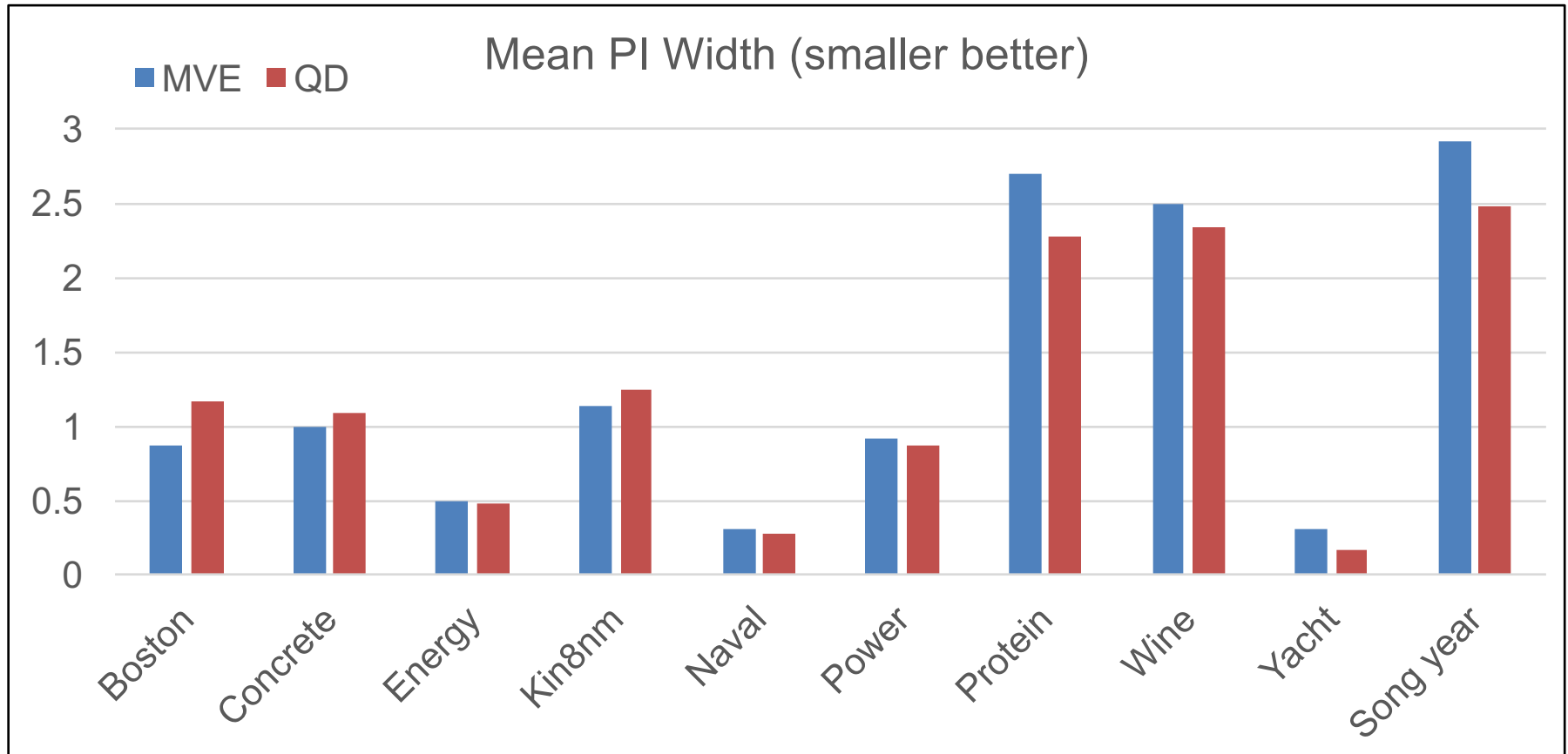
Results

- Prediction of house prices
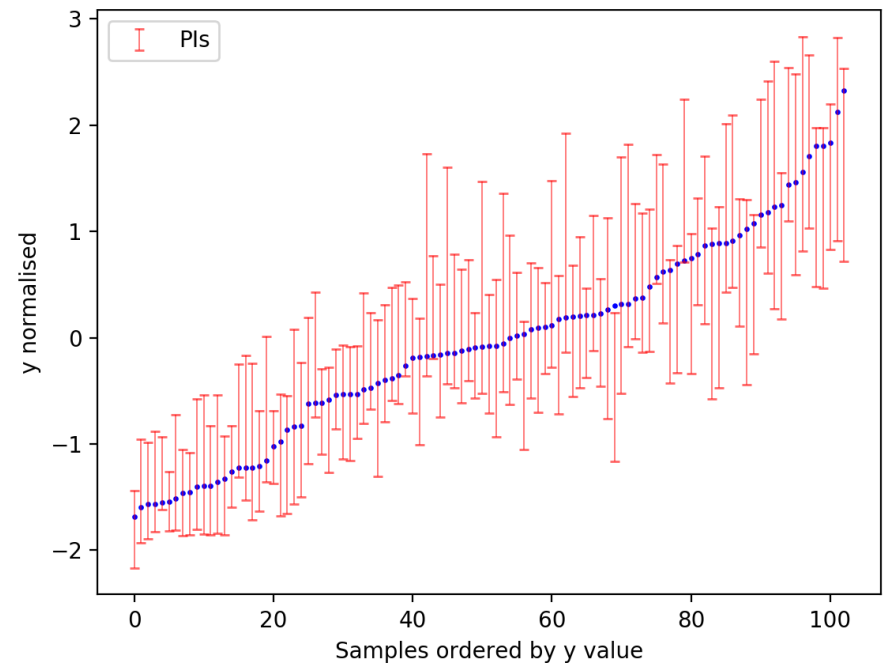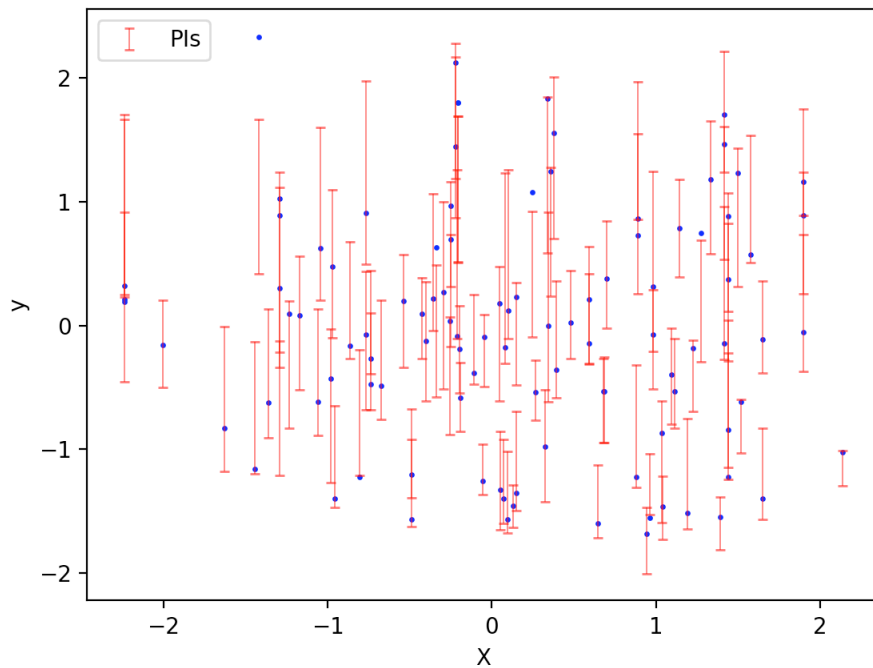
High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach

Tim Pearce [1 2]   Mohamed Zaki [1]   Andy Neely [1]   Alexandra Brintrup [1]

Abstract

Deep Neural Networks are a powerful technique for learning complex functions from data. However, their appeal in real-world applications can be hindered by an inability to quantify the uncertainty of predictions. In this paper, the generation of prediction intervals (PI) for quantifying uncertainty in regression tasks is considered.

It is axiomatic that high-quality PIs should be as narrow as possible, whilst capturing a specified

is a large downside to an incorrect prediction: Examples can be found in prognostics, manufacturing, finance, weather, traffic and energy networks. There is therefore interest in how NNs can be modified to meet this requirement.

In this work the output of Prediction Intervals (PIs) in regression tasks is considered. Whilst NNs by default output point estimates, PIs directly communicate uncertainty, offering a lower and upper bound for a prediction and assurance that with some high probability (e.g. 95% or 99%), the realised data point will fall between these bounds. Having this information allows for better informed decisions and

Available at: https://arxiv.org/abs/1802.07167

Real-World Application

Real-World Application

- Application to warranty data – use to screen incoming claims for consistency

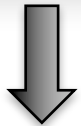| Claim No.	Product	Region	Comment	Part
123	Turbo range	UK	Customer complained about noise during operation. Fan was found to be faulty and was replaced.	Fan-13A
124	Eco series	UK	Door sticking. Hinge appeared to have been forced. Repaired fine.	Hinge-112

1. Estimate PIs

Expected amount
$80 to $200
$20 to $120

2. Compare to actual

Actual amount
$123
$9,876

3. Action

Review?
NO
YES

Thanks for listening

tp424@cam.ac.uk

Andy Neely | Mohamed Zaki | Alexandra Brintrup