Models

Introduction to Automated Science

SLAS 2023

Why models?

Models trained on measured responses can predict the response to untested treatments.

Automated Science relies on models that are

- **Faster** than real-life experiments
- **Cheaper** than real-life experiments
- Aware of the **uncertainty** around untested treatments

Why models?

Models trained on measured responses can predict the response to untested treatments.

Automated Science relies on models that are

- **Faster** than real-life experiments
- **Cheaper** than real-life experiments
- Aware of the **uncertainty** around untested treatments

Models don't need to be "exact".

- The goal of modeling is to design future experiments.
- All model predictions are tested experimentally.

Why models?

Models trained on measured responses can predict the response to untested treatments.

Automated Science relies on models that are

- **Faster** than real-life experiments
- **Cheaper** than real-life experiments
- Aware of the **uncertainty** around untested treatments

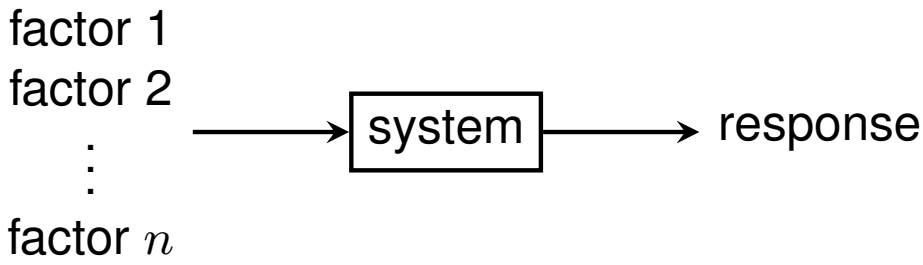
Models don't need to be "exact".

- The goal of modeling is to design future experiments.
- All model predictions are tested experimentally.

However, models need to be *calibrated*—correct in their uncertainty.

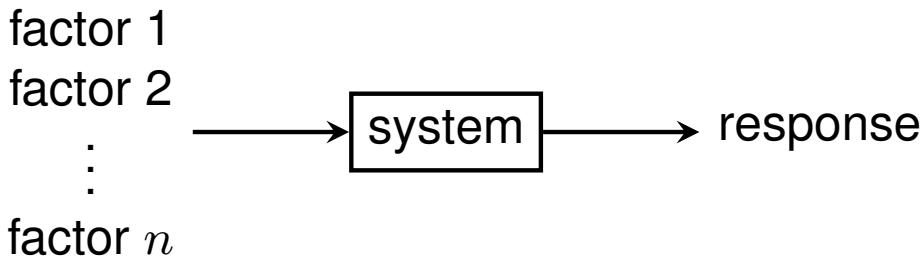
The anatomy of a model

Real world systems

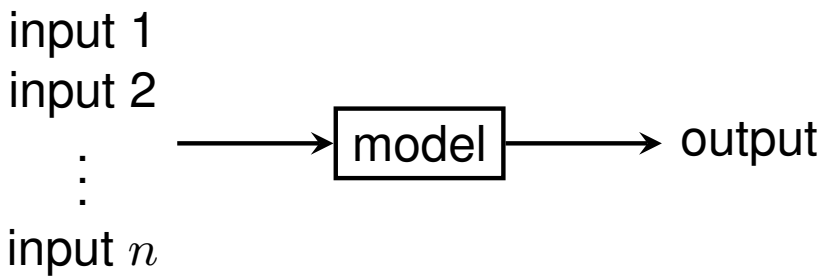


The anatomy of a model

Real world systems

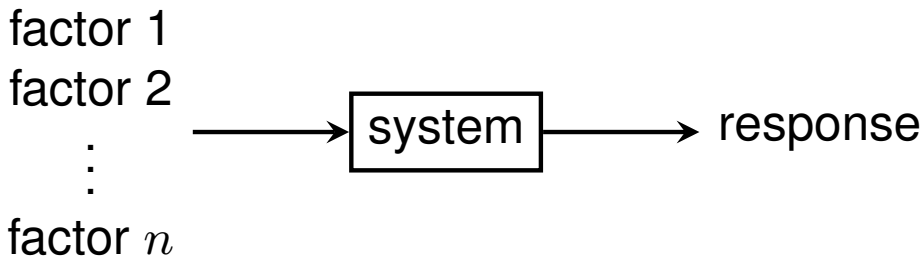


Surrogate model

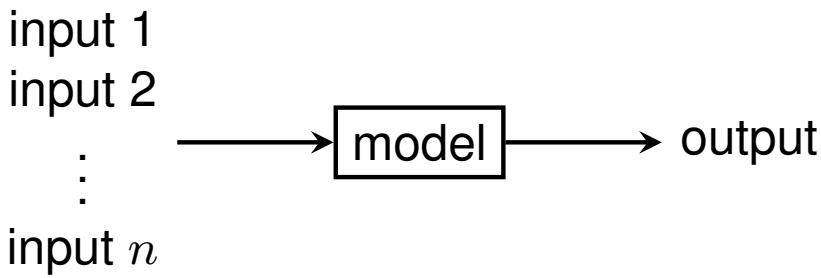


The anatomy of a model

Real world systems



Surrogate model



Considerations when choosing a model

- Mechanistic or black-box
- Fixed parameter or Bayesian (uncertainty aware)
- Efficiency or flexibility

Mechanistic vs. black-box models

Mechanistic model

$$y = \beta_1[\text{serum}] + \beta_2[\text{abx}]$$

- Direct inference of inputs that give desired outputs.
- Limited to known relationships between inputs and outputs.
- Requires less data.

Black-box model

$$y = f(\text{serum}, \text{abx})$$

- Desired outputs must be found by searching over all inputs.
- Can capture any relationship in the data.
- Requires more data.

Fixed parameter vs. Bayesian models

Model: $y = \beta_1[\text{serum}] + \beta_2[\text{abx}]$

What is y when serum = 1, abx = 0.5?

Fixed parameter

$$\beta_1 = 0.32, \quad \beta_2 = -0.12$$

$$y = 0.32(1) - 0.12(0.5) = 0.24$$

Bayesian

$$\beta_1 \sim \text{Normal}(0.32, 0.1)$$
$$\beta_2 \sim \text{Normal}(-0.12, 0.03)$$

$$\begin{aligned} y &= 0.24(1) - 0.10(0.5) &&= 0.19 \\ y &= 0.32(1) - 0.08(0.5) &&= 0.27 \\ y &= 0.27(1) - 0.12(0.5) &&= 0.21 \\ y &= 0.23(1) - 0.14(0.5) &&= 0.16 \end{aligned}$$

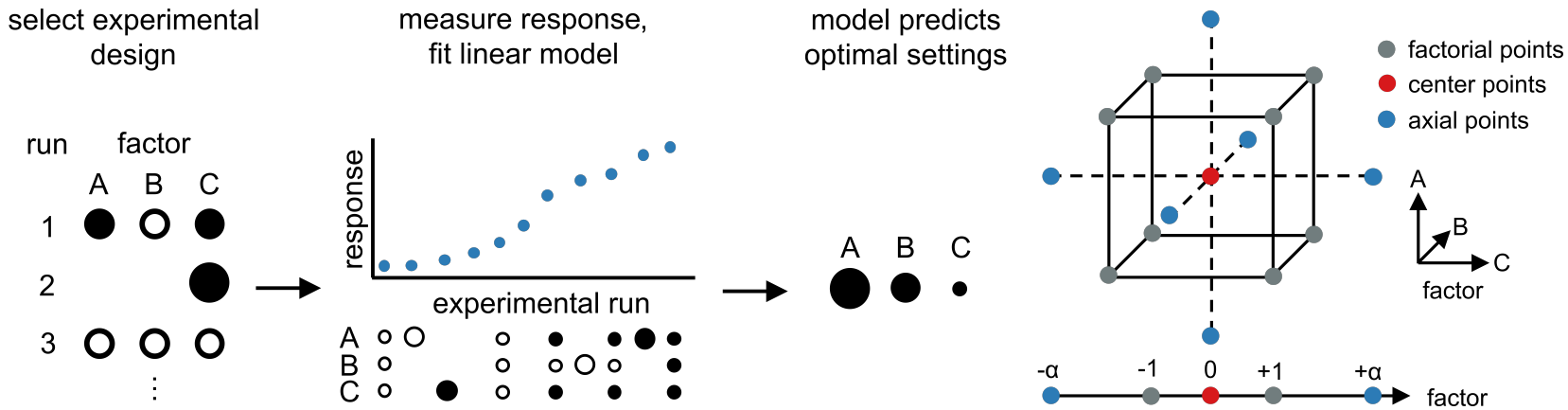
$$y = 0.21 \pm 0.05$$

Parametric Linear Models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$

- Easy to fit and interpret; very data efficient.
- Second-order models can approximate **local** optima.
- Widely used for screening and Response Surface Methodology (RSM).

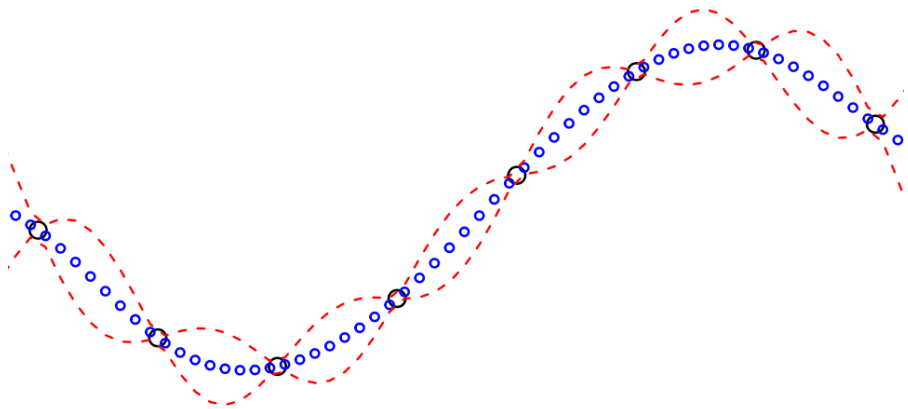
RSM searches for local improvements to a process with linear models.



Gaussian Process Regression (GPR) Models

- Nonparametric interpolators based on multivariate normal distributions.
- Flexible, yet relatively data efficient.
- Analytic estimates of uncertainty.
- Most widely used models for sequential optimization.

GPR models interpolate between data and can estimate the uncertainty of predictions.

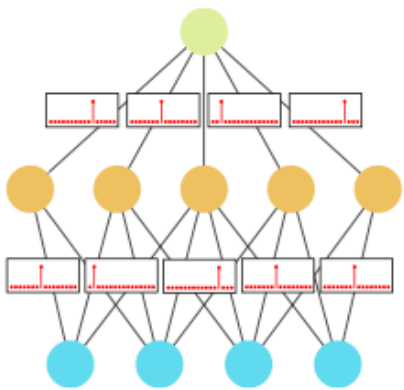


Bayesian Neural Networks

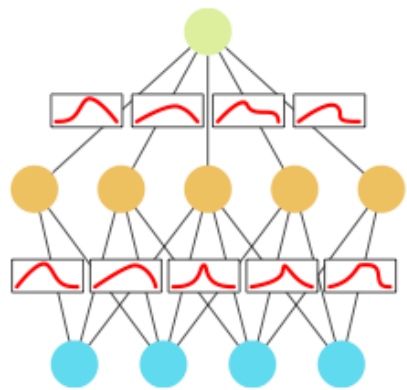
- Neural networks are *universal approximators*.
- Can be difficult to train and require extensive data.
- Must be trained as Bayesian models for planning.

Bayesian neural networks use statistical distributions to describe the network weights. They are ensembles of infinitely many models that fit the data.

Pointwise Neural Network



Bayesian Neural Network



Jospin, *et al.*, (2022). https://arxiv.org/pdf/2007.06823.pdf

Which model should I use?

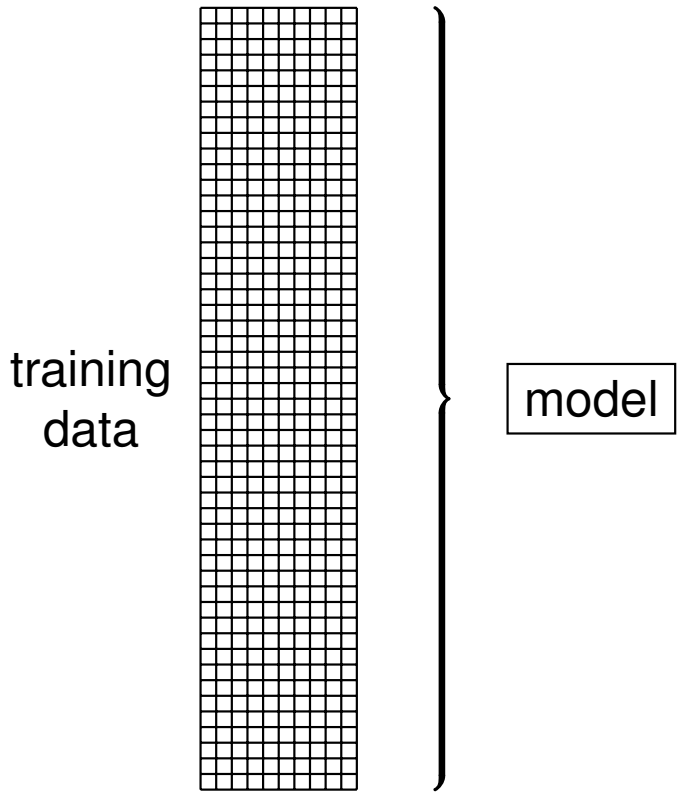
Picking a model requires balancing two factors:

1. How nonlinear is the process (i.e. how flexible must the model be)?
2. How much data is available?

Model	**Samples Needed**	**Flexibility**
Linear Models	10's	Least flexible
Gaussian Processes	10–100's	Flexible
Neural Networks	100–1000+	Most flexible

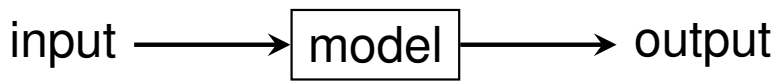
What if I already have a model?

Fixed parameter models can be made Bayesian by **bagging**.



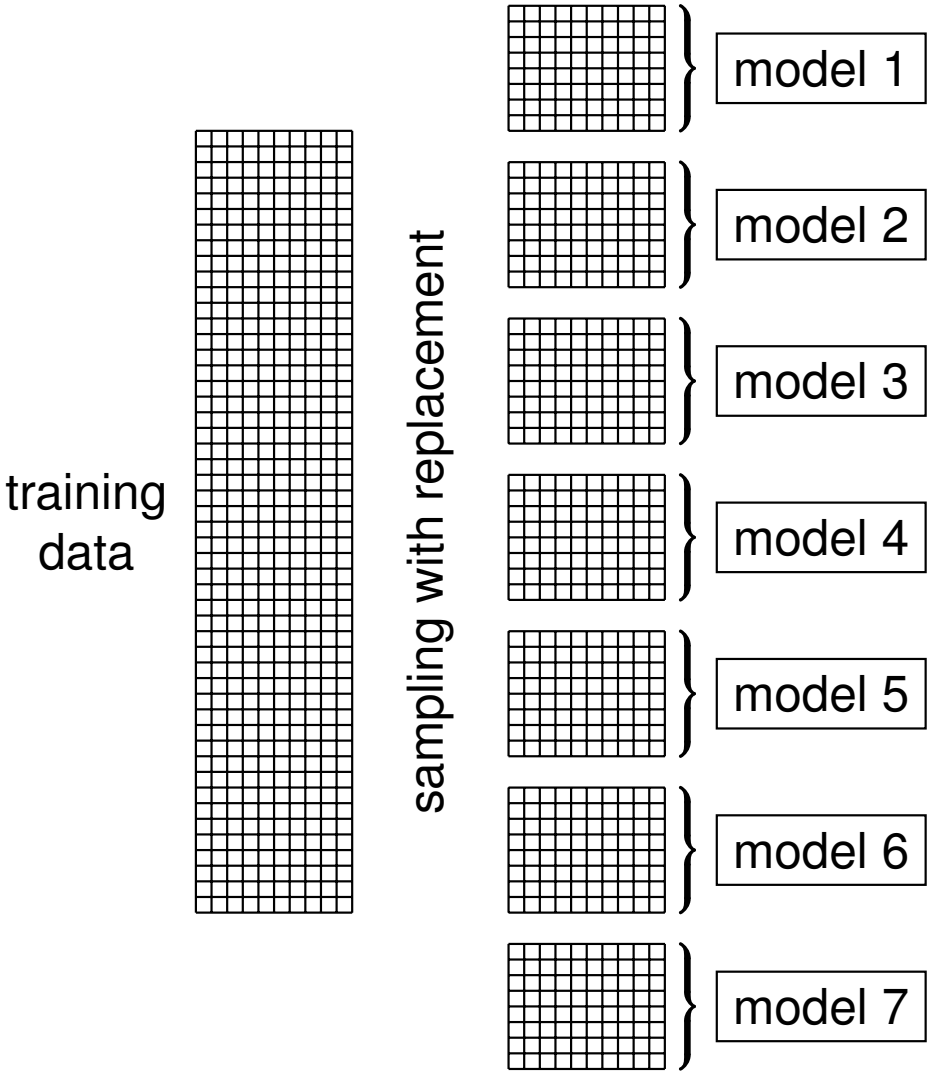
What if I already have a model?

Fixed parameter models can be made Bayesian by **bagging**.



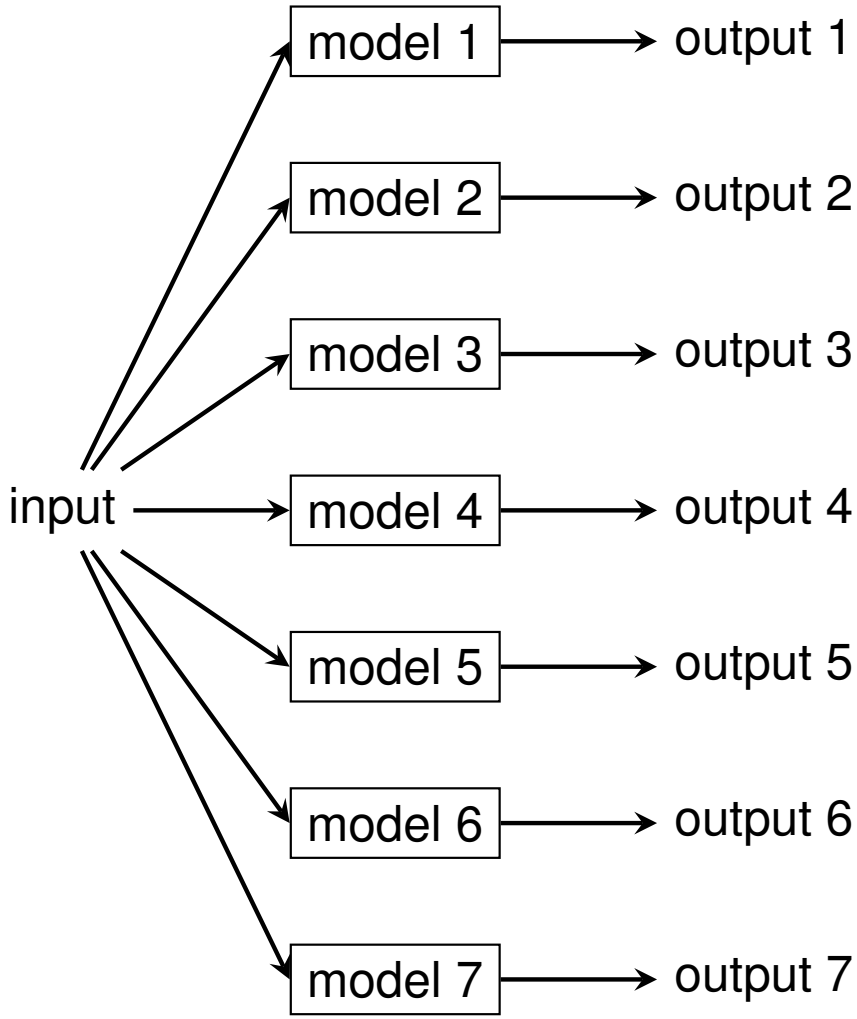
What if I already have a model?

Fixed parameter models can be made Bayesian by **bagging**.



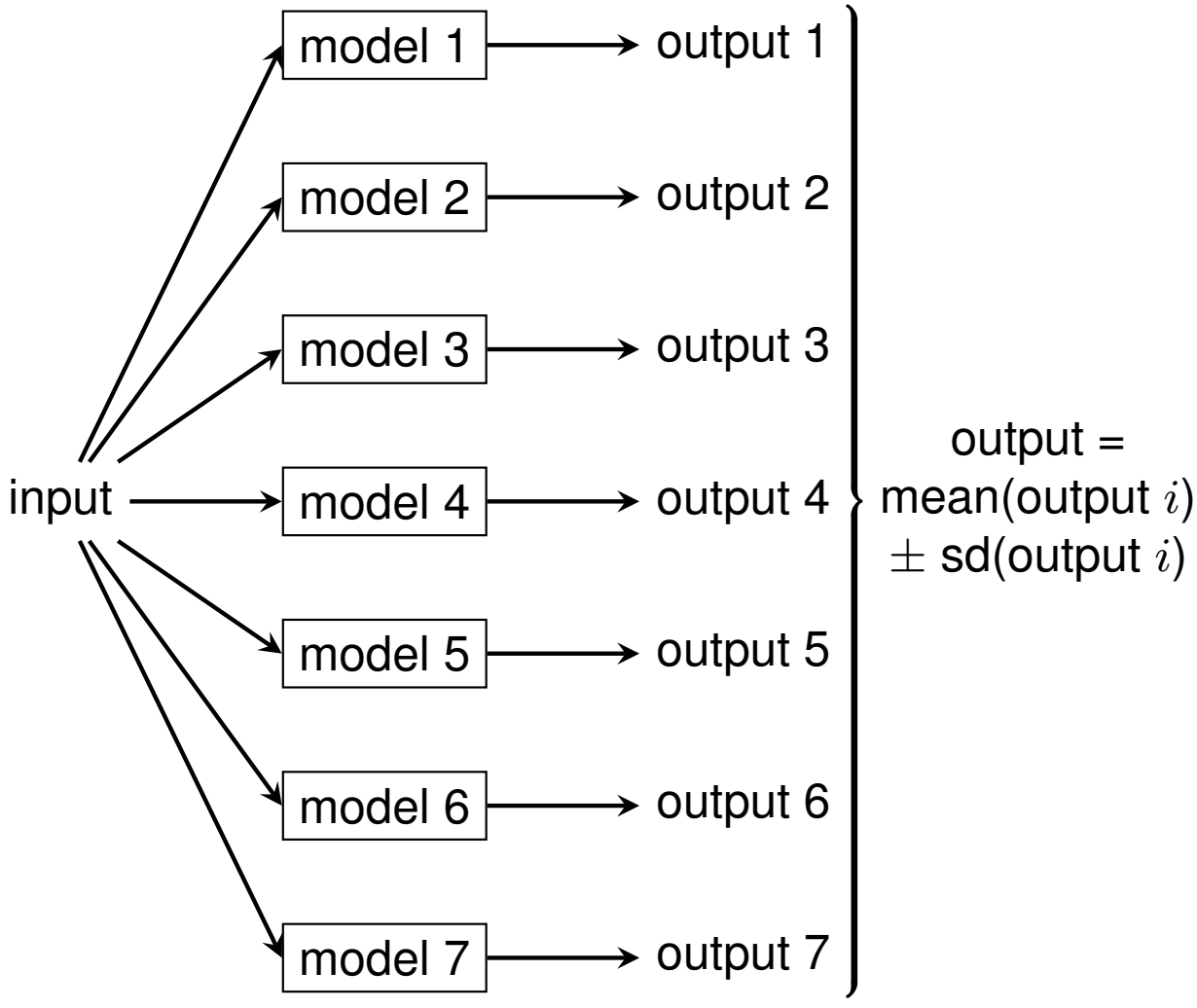
What if I already have a model?

Fixed parameter models can be made Bayesian by **bagging**.



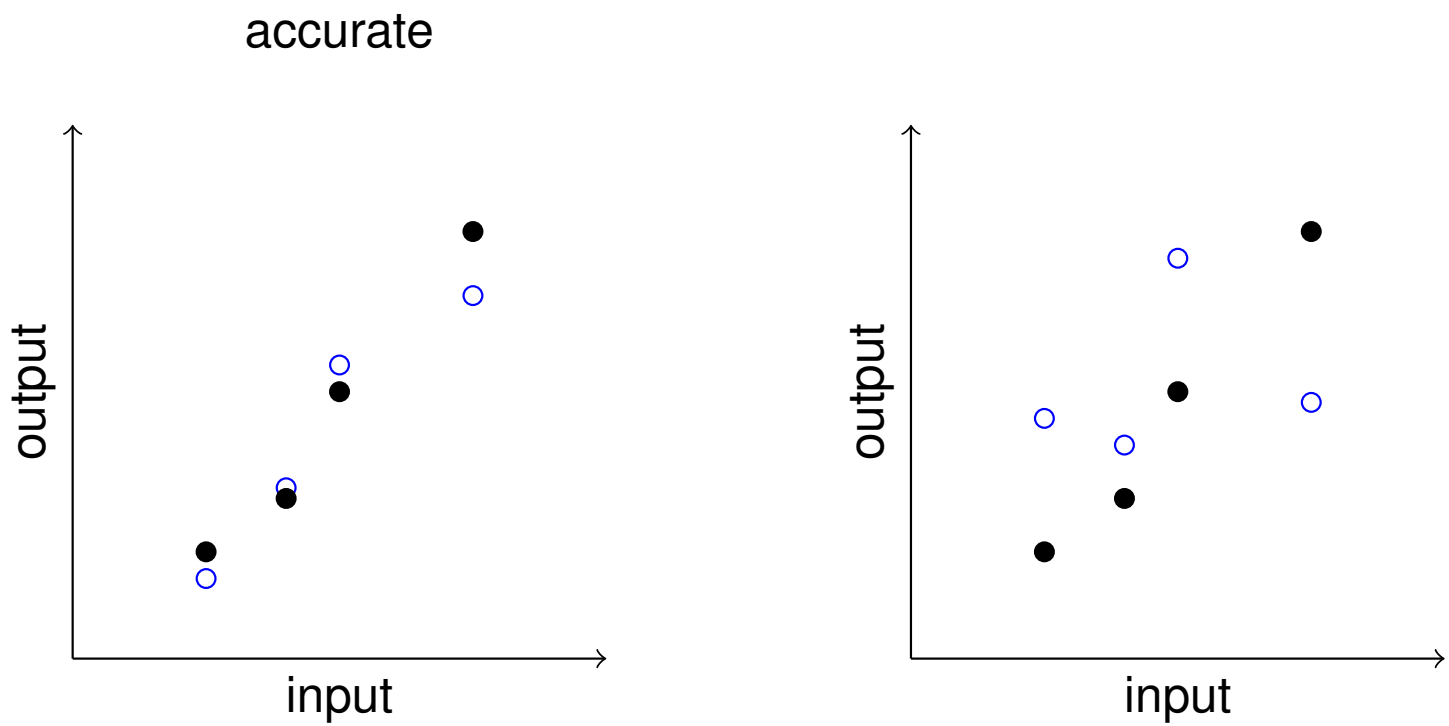
What if I already have a model?

Fixed parameter models can be made Bayesian by **bagging**.



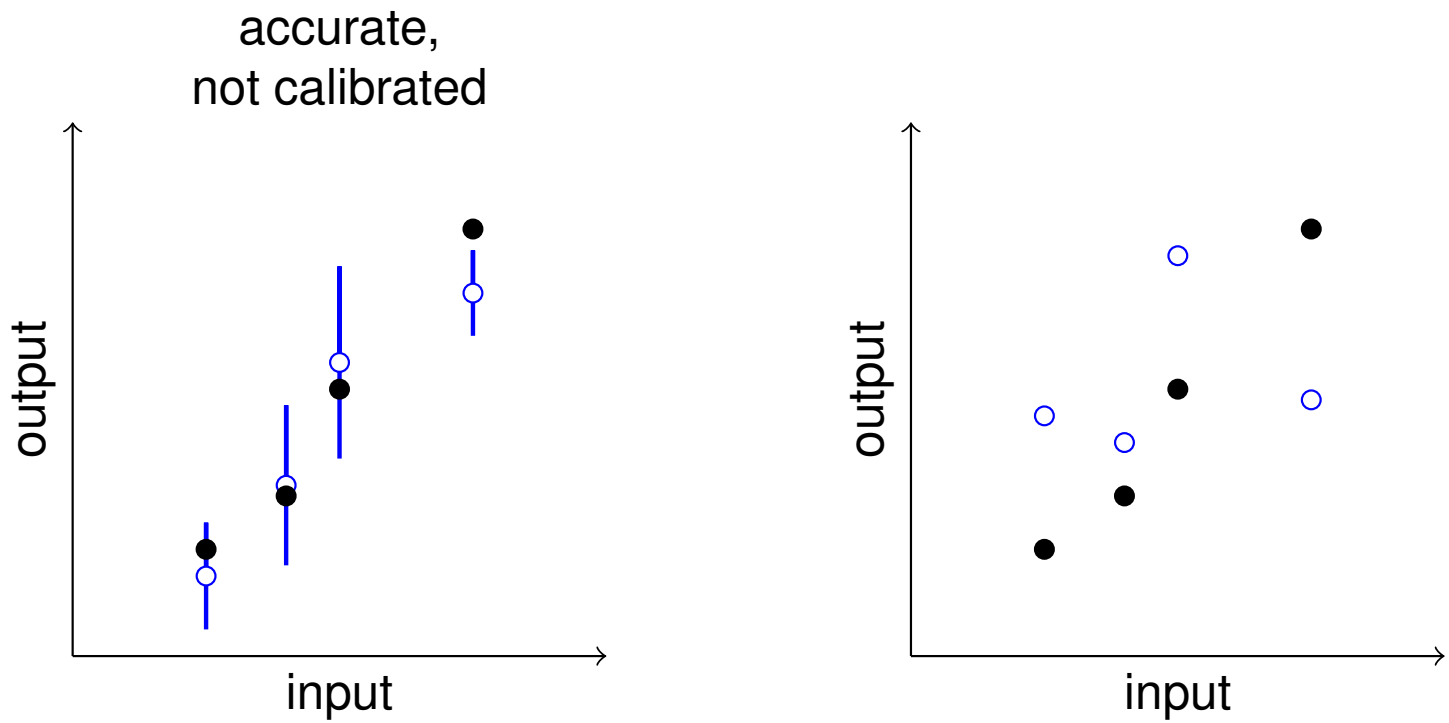
Calibration vs. Accuracy

- *Exploiting* requires an **accurate** model to find optimal responses.
- *Exploring* requires a **calibrated** model to find uncertain responses.



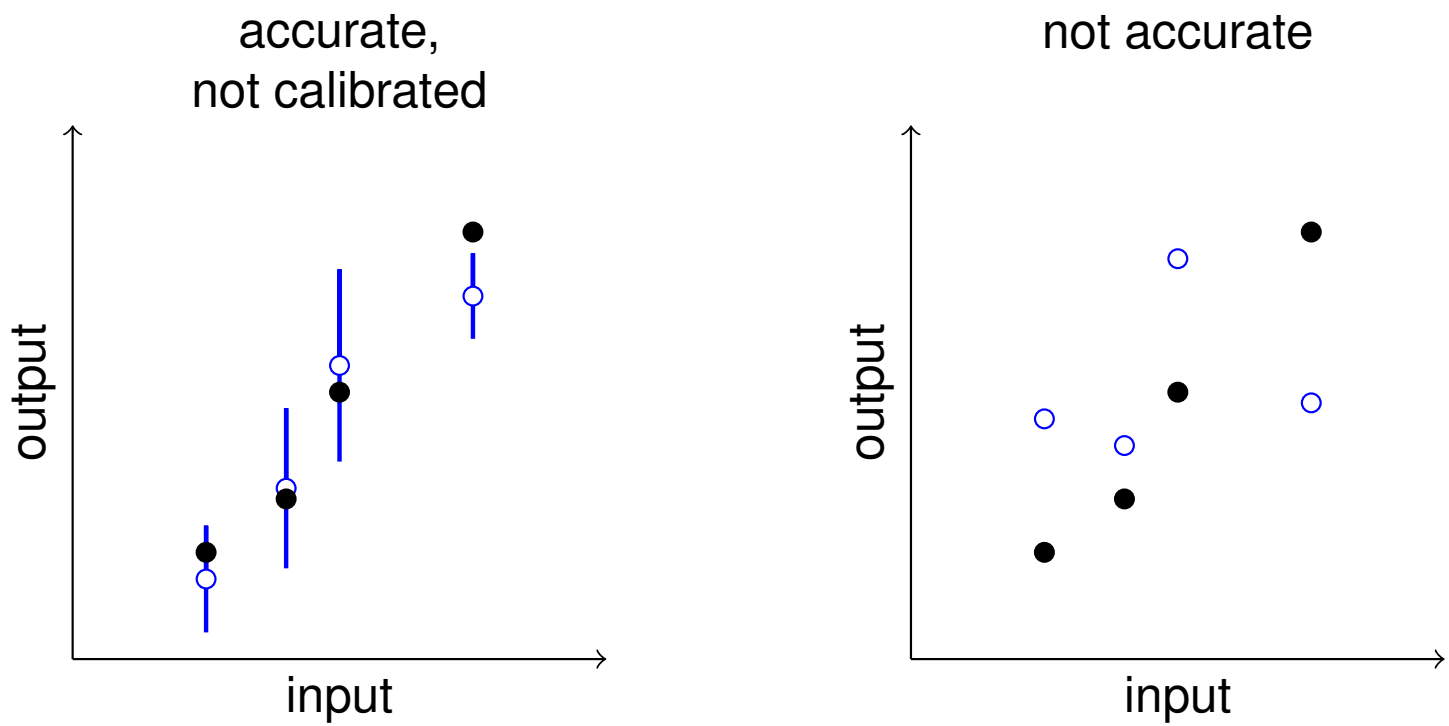
Calibration vs. Accuracy

- *Exploiting* requires an **accurate** model to find optimal responses.
- *Exploring* requires a **calibrated** model to find uncertain responses.



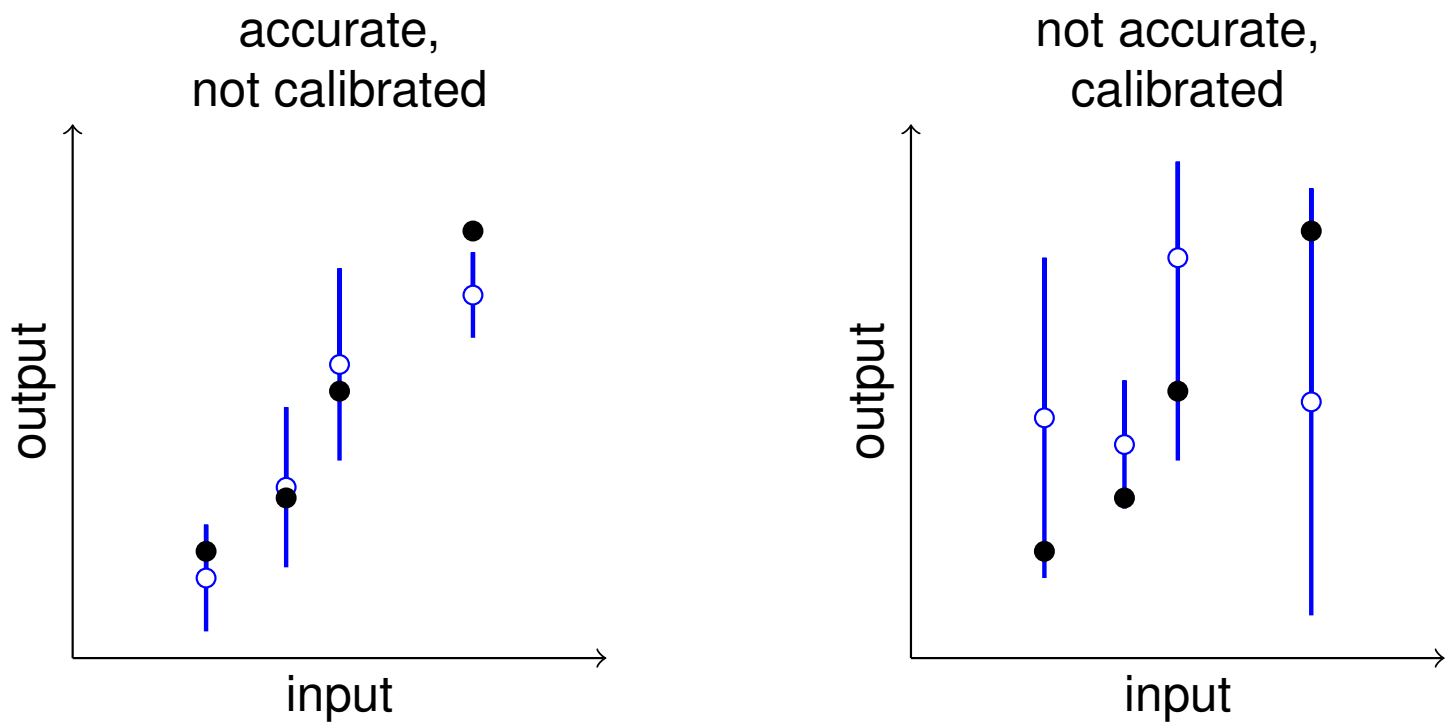
Calibration vs. Accuracy

- *Exploiting* requires an **accurate** model to find optimal responses.
- *Exploring* requires a **calibrated** model to find uncertain responses.



Calibration vs. Accuracy

- *Exploiting* requires an **accurate** model to find optimal responses.
- *Exploring* requires a **calibrated** model to find uncertain responses.



Summary

- Models **summarize** current knowledge.
- Planning with models is **faster** and **cheaper** than searching with real-world experiments.
- Bayesian models return the **uncertainty** for every input.
- A good model is both **accurate** and **calibrated**.