The Role of *Uncertainty* in Machine Learning

Andreas Damianou - Alexa Shopping Science

Accelerate Science Winter School, Cambridge, 2 Feb. 2021

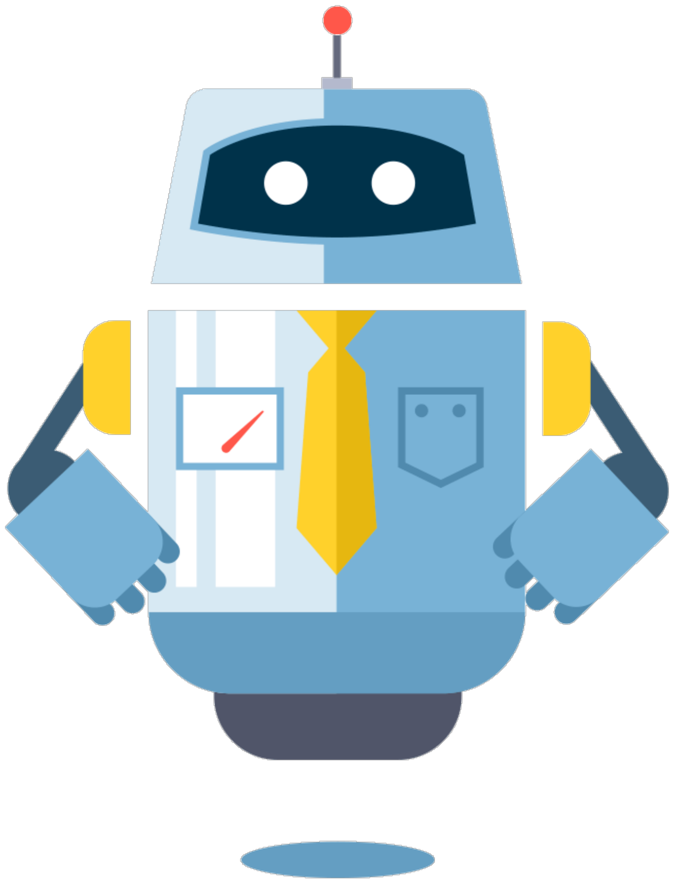
Give me $50,000 to invest for you..
I ***predict*** that GameStop stock will
rise dramatically

Give me $50,000 to invest for you..
I ***predict*** that GameStop stock will
rise dramatically

Are you ***sure***??

Give me $50,000 to invest for you..
I ***predict*** that GameStop stock will rise dramatically

Are you ***sure***??

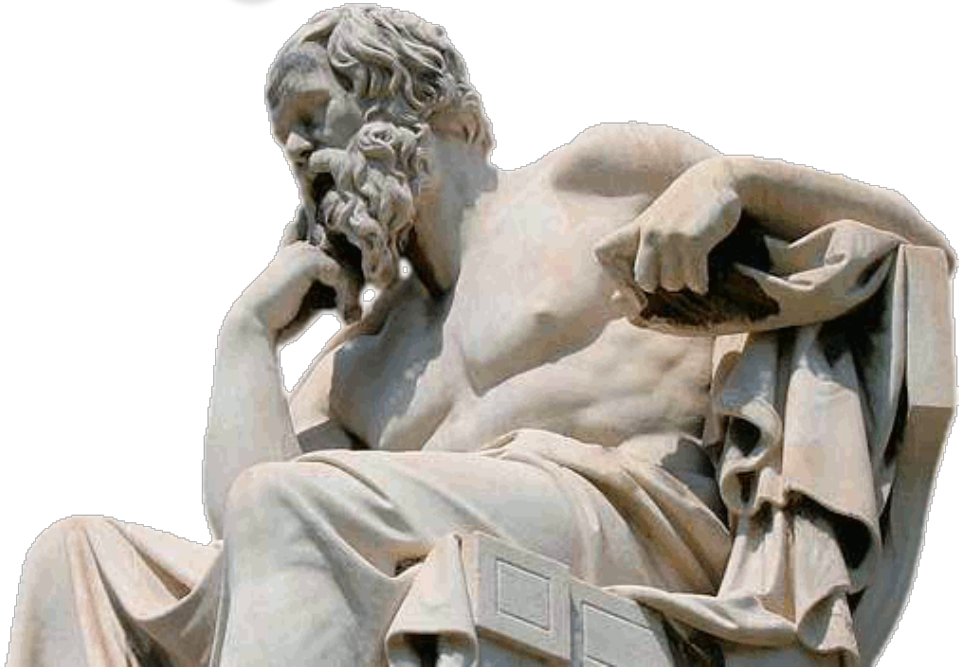


Machine Learning

Motivation: what is uncertainty and how is it introduced in ML

What is uncertainty?

In general, uncertainty is **lack of knowledge**.

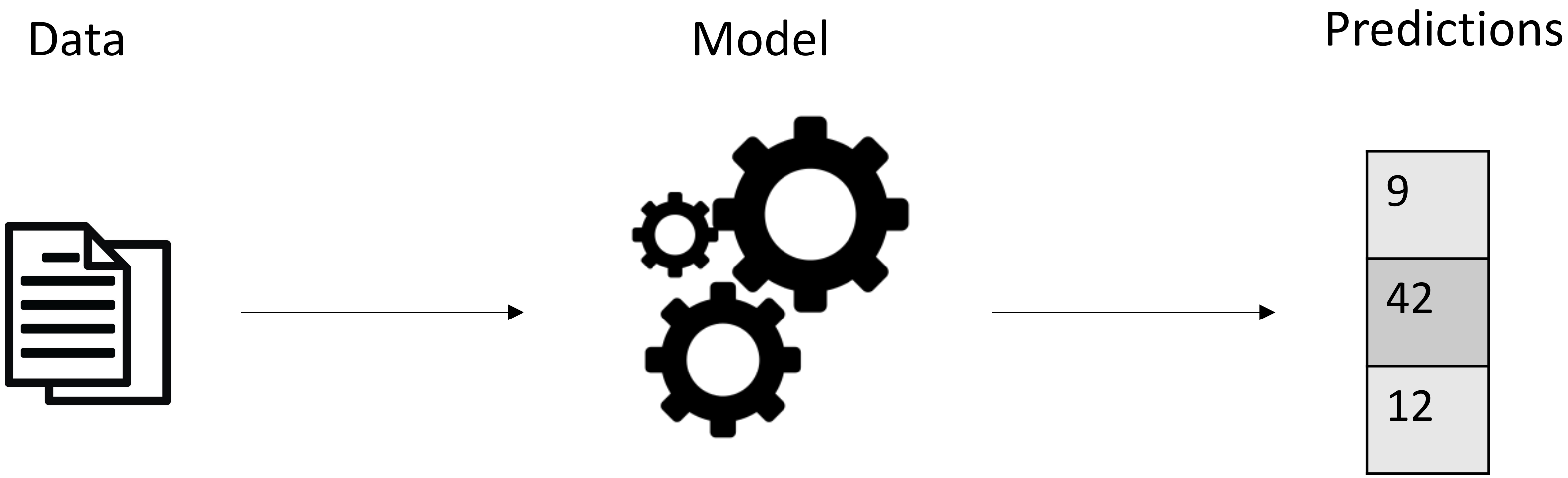


What is uncertainty?

In general, uncertainty is **lack of knowledge**.

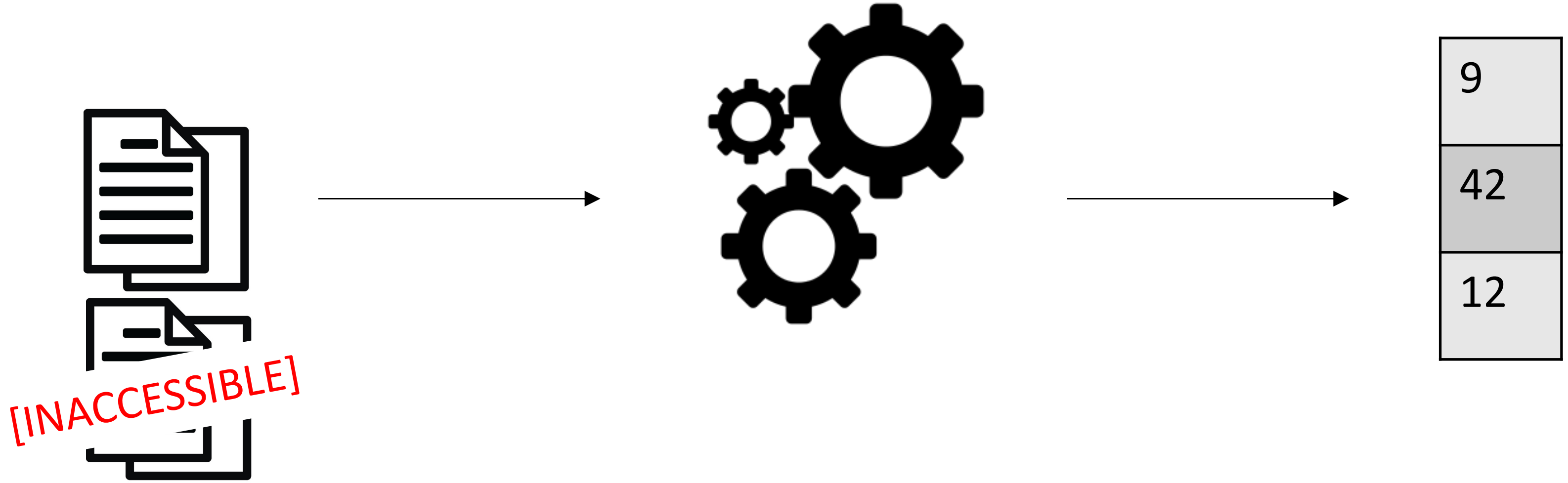
See: Neil's and Carl Henrik's talk!

Machine Learning in a nutshell (in theory)

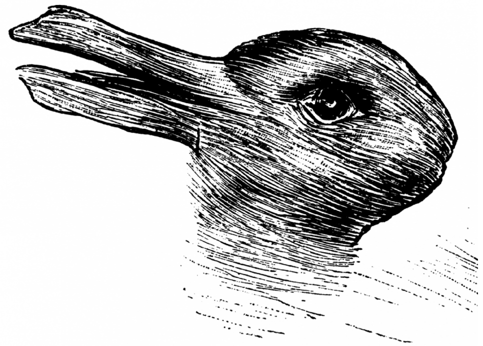


Machine Learning in a nutshell (in reality)

Data Model Predictions

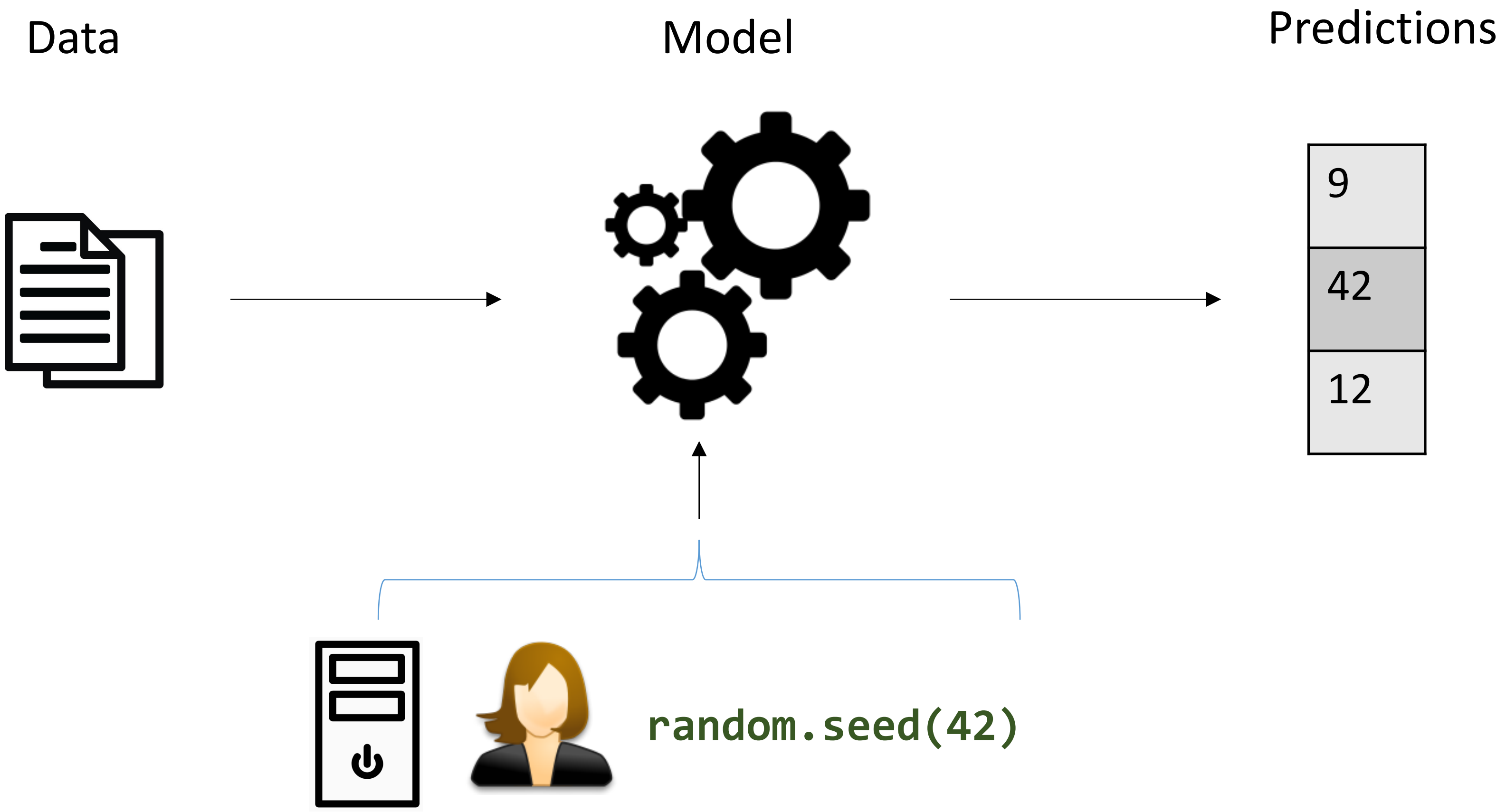


Our model only sees **partial** and **noisy data.**



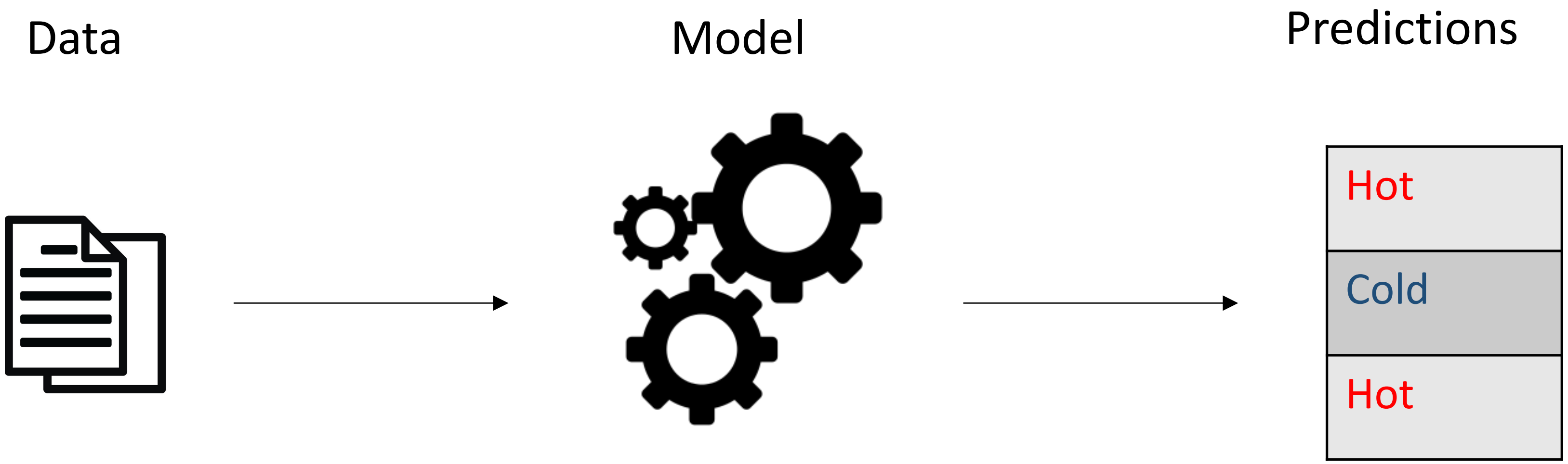
Duck or Rabbit?

Machine Learning in a nutshell (in reality)



Our model is **imperfect,** possibly **biased** and often inherently **random**.

Machine Learning in a nutshell (in reality)



Interpretation of concepts can be **fuzzy**.

Computers are deterministic, but our world isn’t

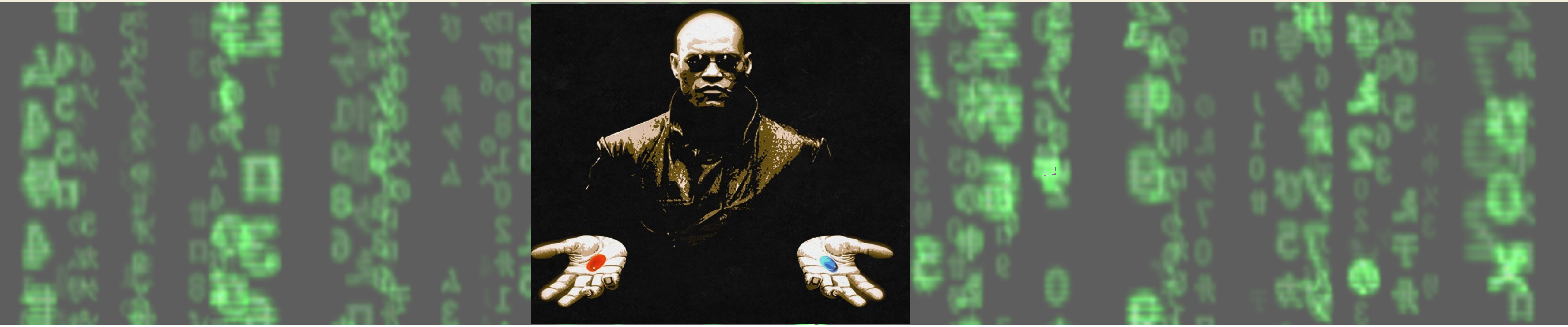
- Data, assumptions and model imperfections come from the real world. These induce uncertainty in our ML modeling

- Uncertainty is hidden in any modeling scenario whether we want it or not: we can never have complete knowledge (otherwise we wouldn’t resort to modeling).

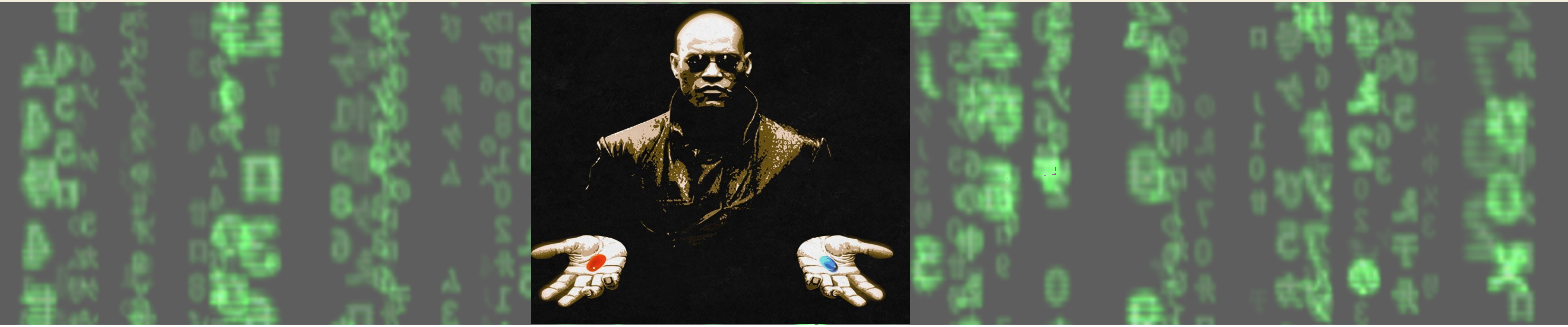
- Using uncertainty in our models is natural for humans: we plan our lives and actions using uncertainty and risk; *we acknowledge we don’t know everything*

Photo by Alessandro Erbetta

Uncertainty is inevitable in modeling



Uncertainty is inevitable in modeling



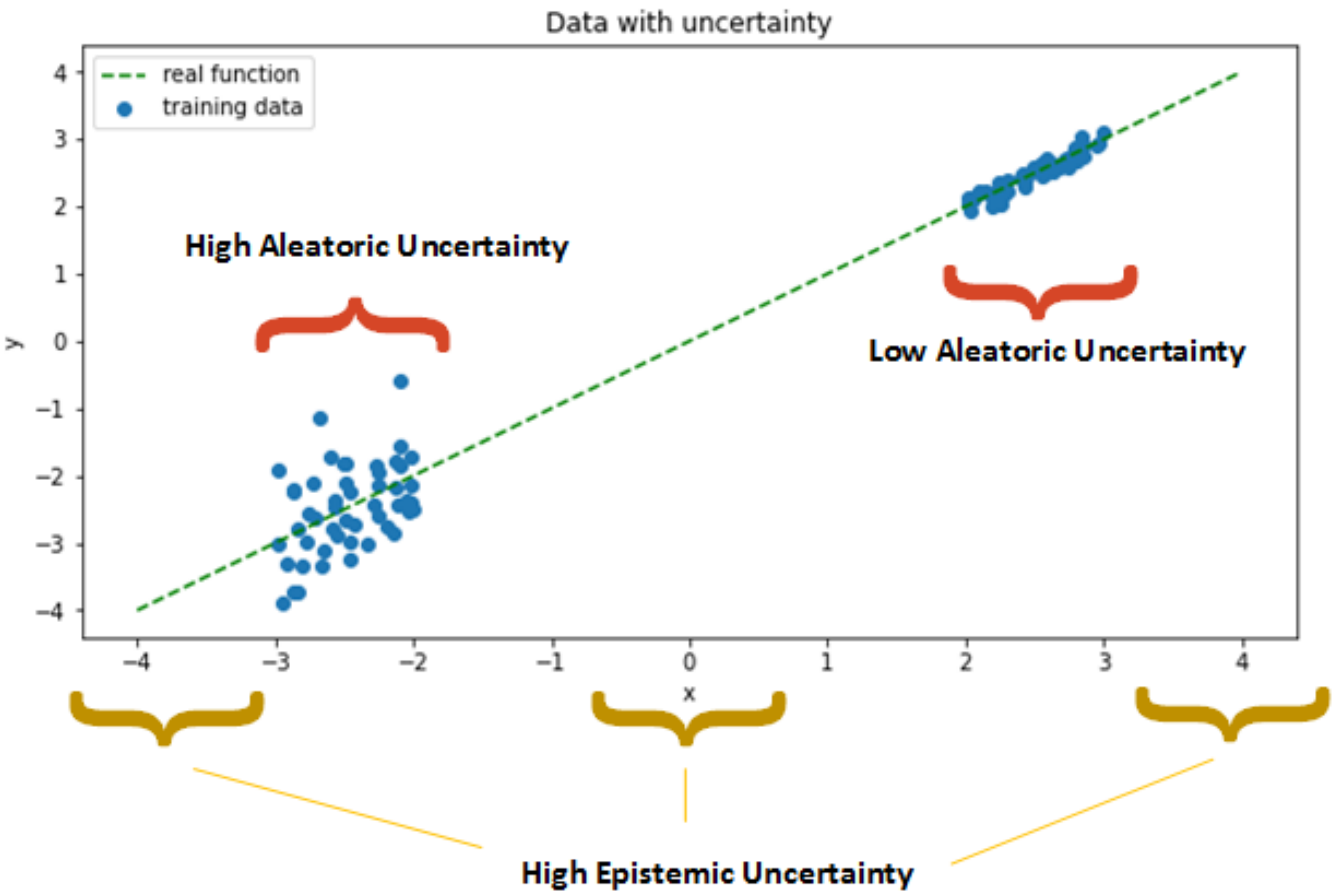
Epistemic uncertainty

Neo doesn't know that he lives in a simulation (Ignorance about the correct model that generated the data e.g. Matrix glitches)

Aleatoric uncertainty

Neo knows that he lives in a simulation but the simulation's complexity introduces *inherent* uncertainty (not enough capacity to perfectly observe the world)

Uncertainty is inevitable in modeling



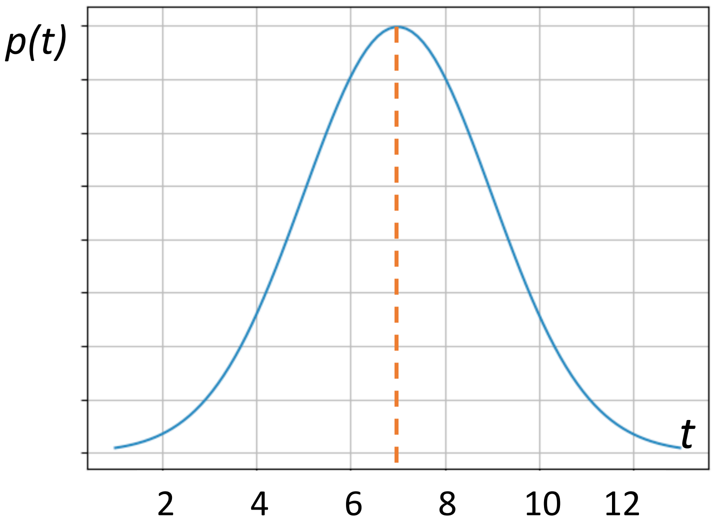
 (Image by Michel Kana)

Applications where uncertainty matters

Predictive uncertainty

- **Classification**: *"I am 92% certain that this stock is a buy"*

- **Regression**: *"The temperature tomorrow will follow* $\mathcal{N}(7, 4)$*"*

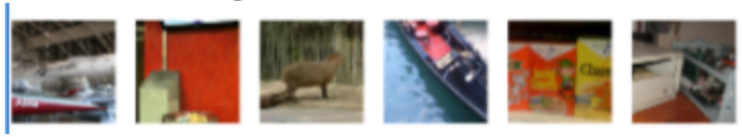


Uncertainty in decision making

- Uncertainty can be used to guide decisions
- Same prediction can lead to different decisions depending on degree of confidence

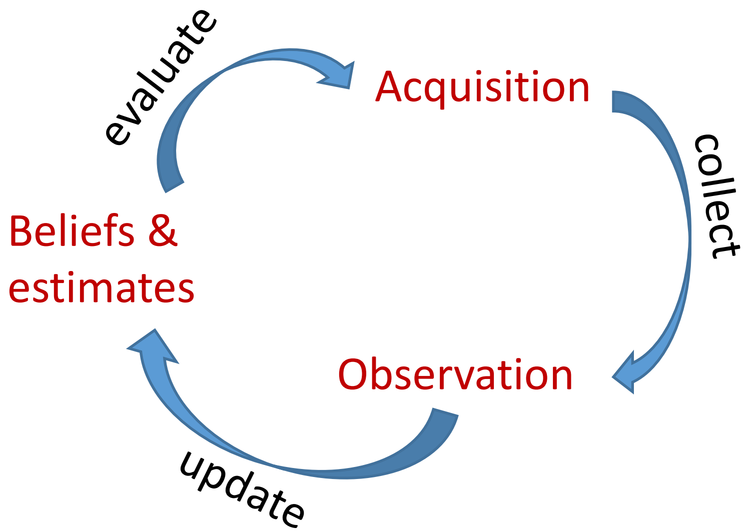
Sequential decision making

- **Active Learning:**
  - Select images to label such that expected accuracy is maximized



- **Bayesian Optimization:**
  - Find the minimum of a function *f*

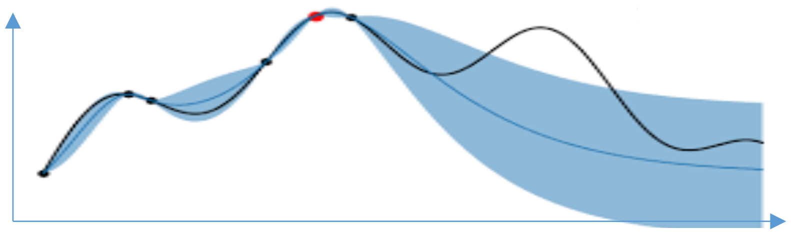
- **Reinforcement Learning:**
  - Take *K* actions to collect maximum combined reward

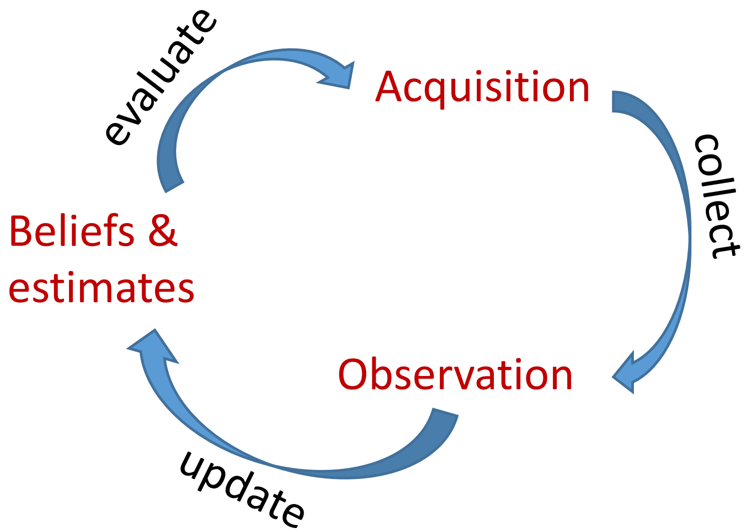


Sequential decision making

- **Active Learning:**
  - Select images to label such that expected accuracy is maximized

- **Bayesian Optimization:**
  - Find the minimum of a function *f*



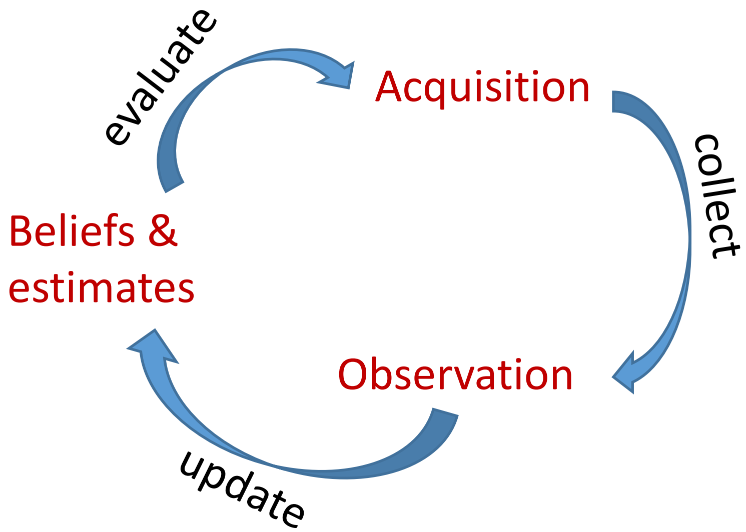


- **Reinforcement Learning:**
  - Take *K* actions to collect maximum combined reward

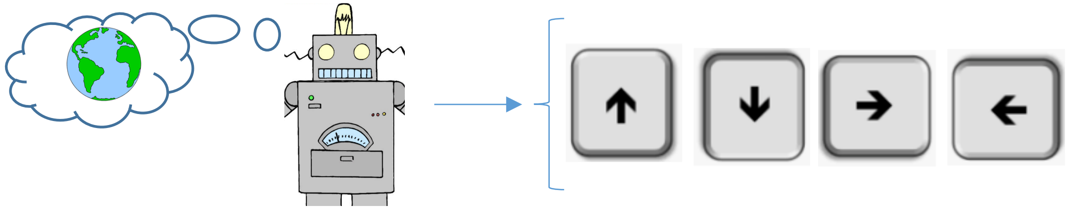
Sequential decision making

- **Active Learning:**
  - Select images to label such that expected accuracy is maximized

- **Bayesian Optimization:**
  - Find the minimum of a function f



- **Reinforcement Learning:**
  - Take K actions to collect maximum combined reward



Probabilistic numerics (aka uncertainty *everywhere*)

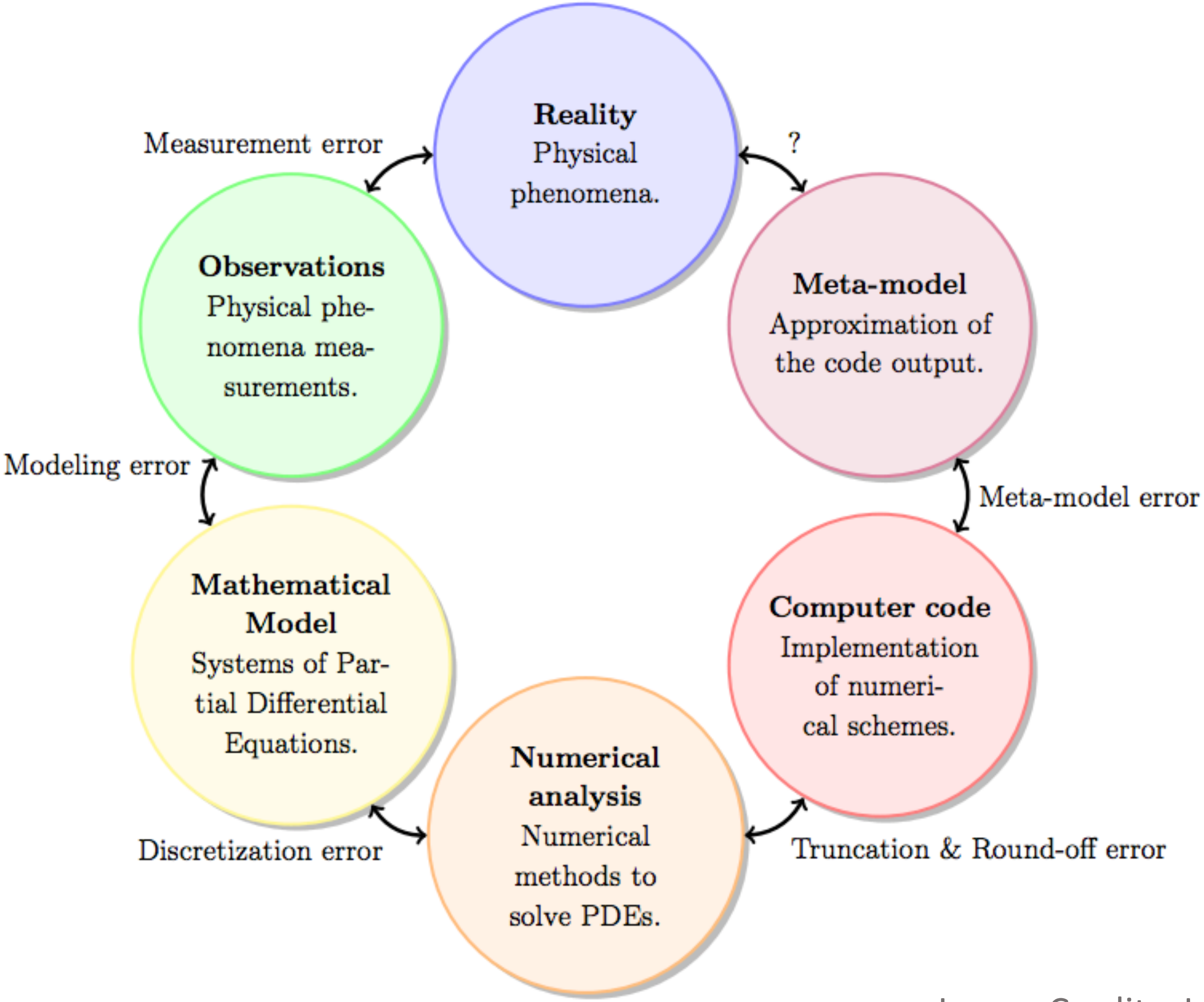


Image Credits: L. Le Gratiet 2013

Quantifying and Auditing Uncertainty

Calibration

- We use uncertainty because we don’t trust predictions (i.e. estimates of dependent variable).

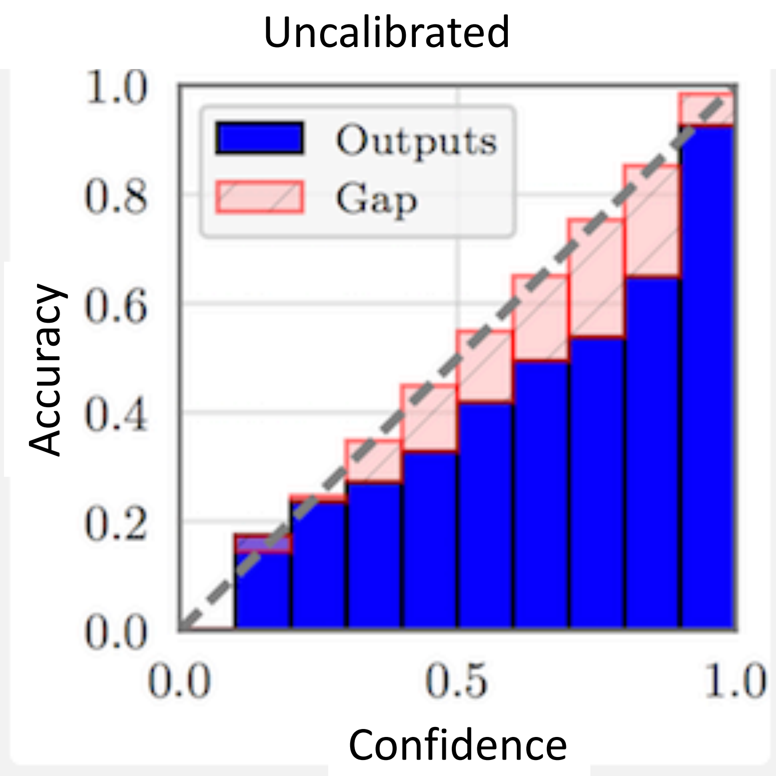
- But why should we trust estimates of uncertainty?

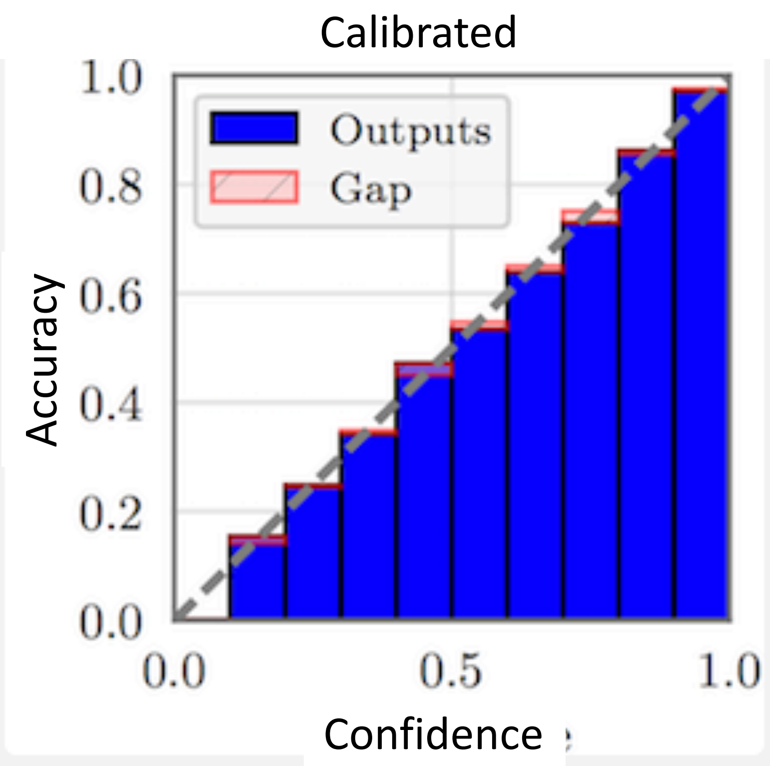
Miscalibrated model predictions

Alcorn et al. 2019

Calibration plots

[Credit: Geoff Pleiss]

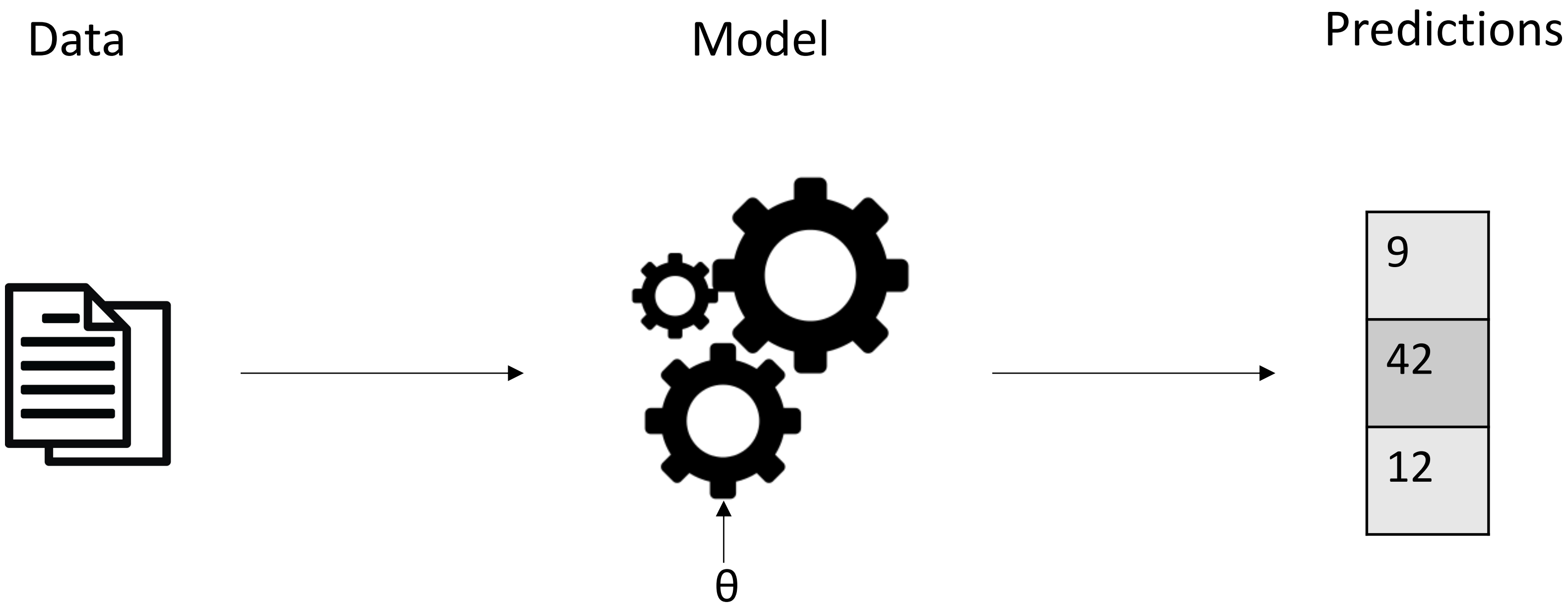


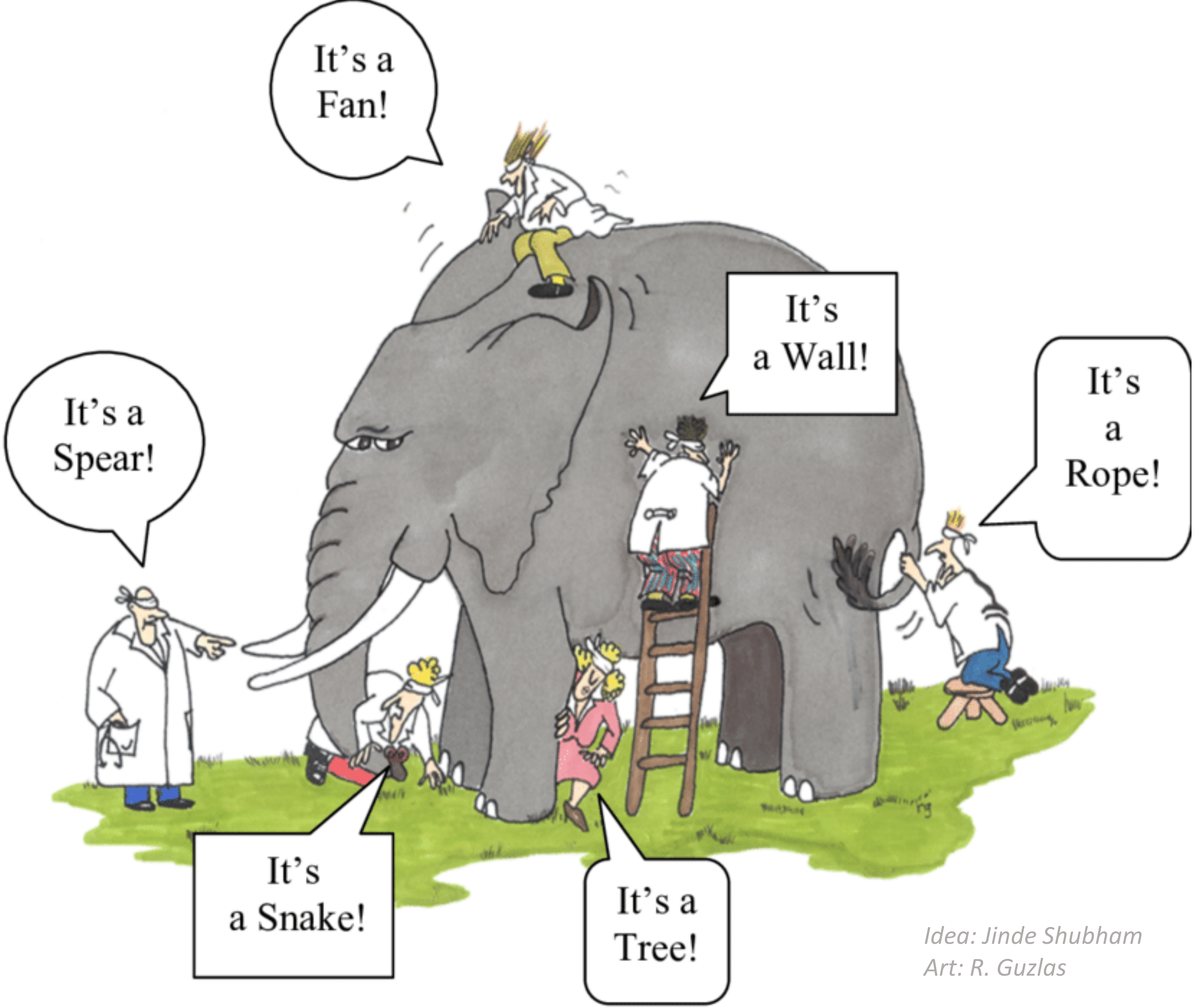


*"If, for example, each of 100 predictions have confidence 80%, then we'd expect that 80% of those are actually correct. If this is the case, we say the model is **calibrated**."*

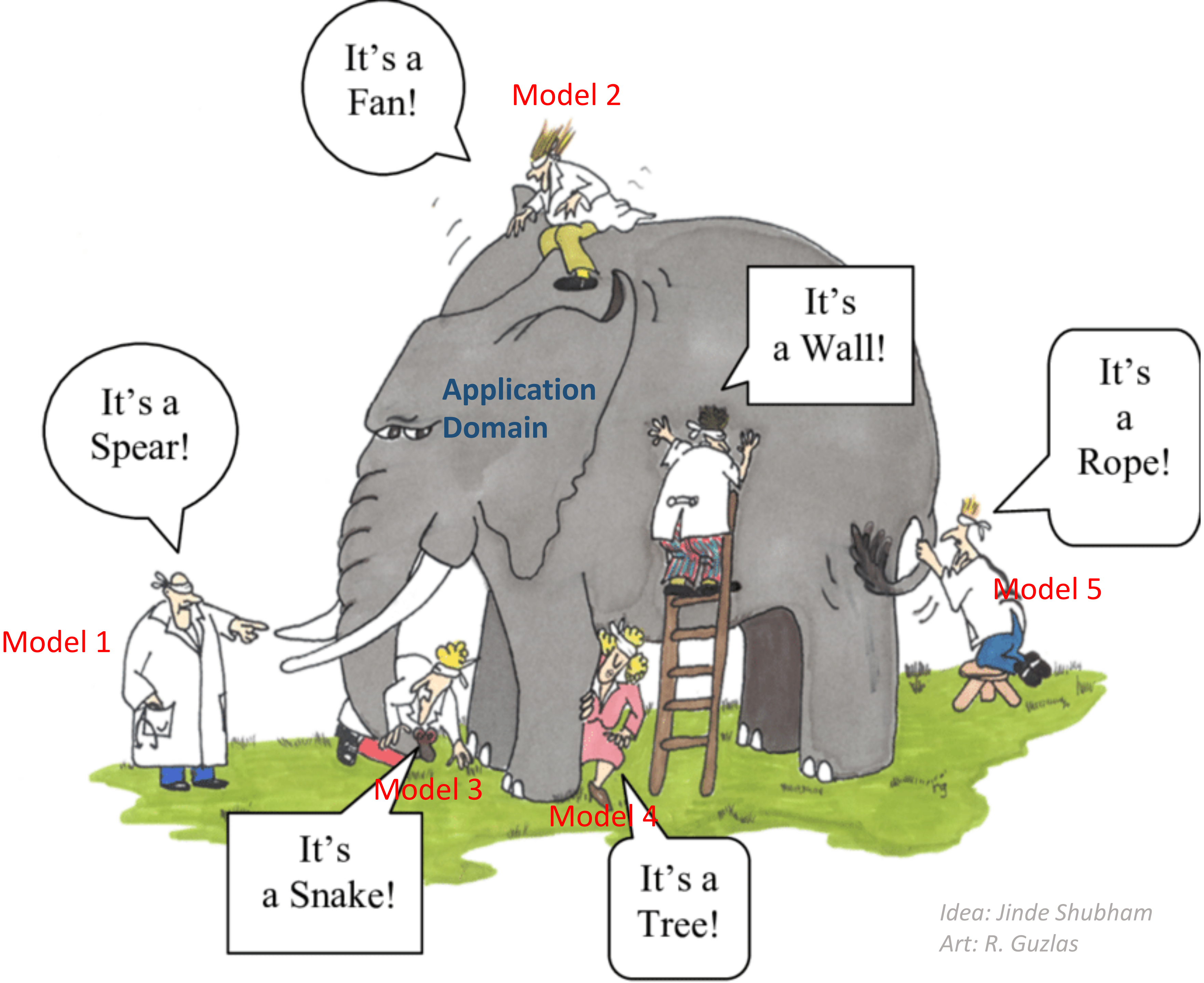
Plot: Split predictions in bins. Then average accuracy and average confidence per bin should match.

Machine Learning modeling





Idea: Jinde Shubham
Art: R. Guzlas

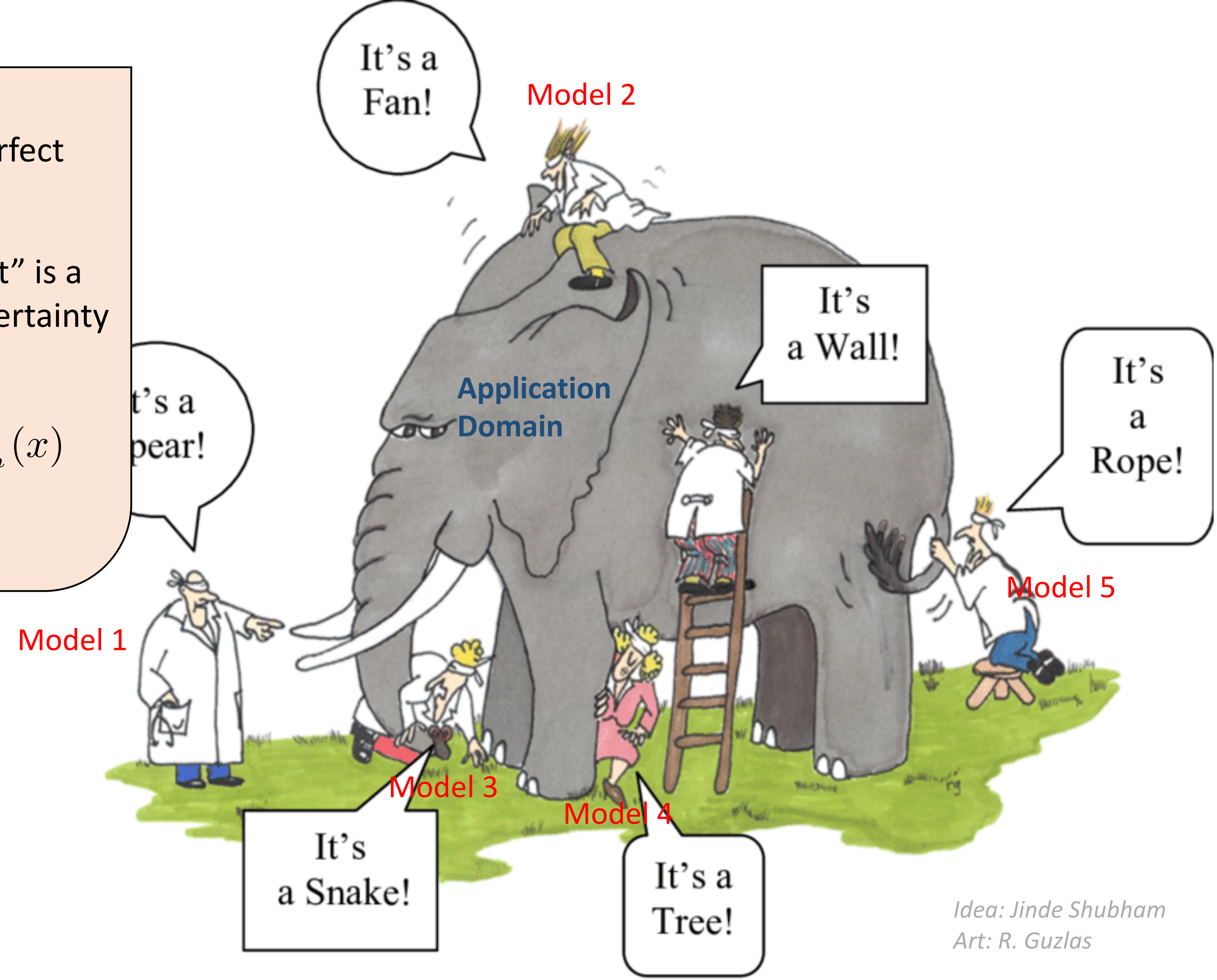


Idea: Jinde Shubham
Art: R. Guzlas

- Every trained model is imperfect
- Let's **combine Models**
- "Modeling the disagreement" is a crude way of improving uncertainty quantification

$$\text{prediction} = \frac{1}{M}\sum_{m} p_{\theta_m}(x)$$

[Lakshminarayanan et al. 2017]



Idea: Jinde Shubham
Art: R. Guzlas

Bayesian model averaging

Model Combination:

$$p(x) = \frac{1}{M}\sum_m p_{\theta_m}(x)$$

(Bayesian)
Model Averaging:

$$p(x) = \int_\theta p_\theta(x)\boxed{p(\theta|\mathrm{data})}$$

Well calibrated
model uncertainty!

Bayes Rule:

$$\boxed{p(\theta|\mathrm{data})} = \frac{p(\mathrm{data}|\theta)p(\theta)}{p(\mathrm{data})}$$

Bayesian model averaging

Model Combination:

$$p(x) = \frac{1}{M}\sum_m p_{\theta_m}(x)$$

(Bayesian)
Model Averaging:

$$p(x) = \int_\theta p_\theta(x)\boxed{p(\theta|\text{data})}$$

Given a test case, Bayesian modeling allows to ***propagate*** all uncertainties (data, prior, model…) into the final predictive distribution.

Bayes Rule:

$$\boxed{p(\theta|\text{data})} = \frac{p(\text{data}|\theta)p(\theta)}{p(\text{data})}$$

(Bayesian) Linear Regression

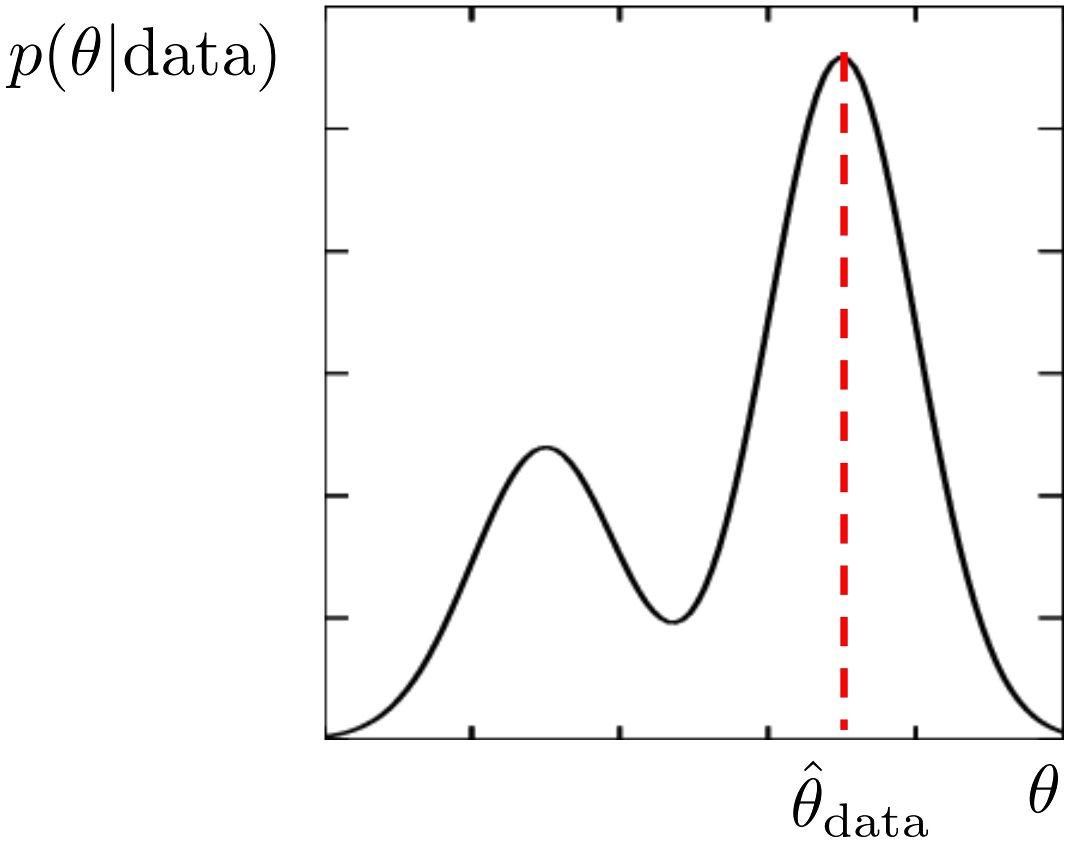
Error minimization formulation	**Probabilistic formulation**
$y = \theta^T x$	$p(y\|x;\theta) = \mathcal{N}(\theta^T x, \boldsymbol{\sigma}^2)$
$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_i (y_i - \theta^T x_i)^2$	$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(y\|x;\theta)$

Bayesian formulation

$$p(\theta|x,y) = \frac{p(x,y|\theta)p(\theta)}{p(x,y)}$$

Single point vs full posterior

Model Fitting



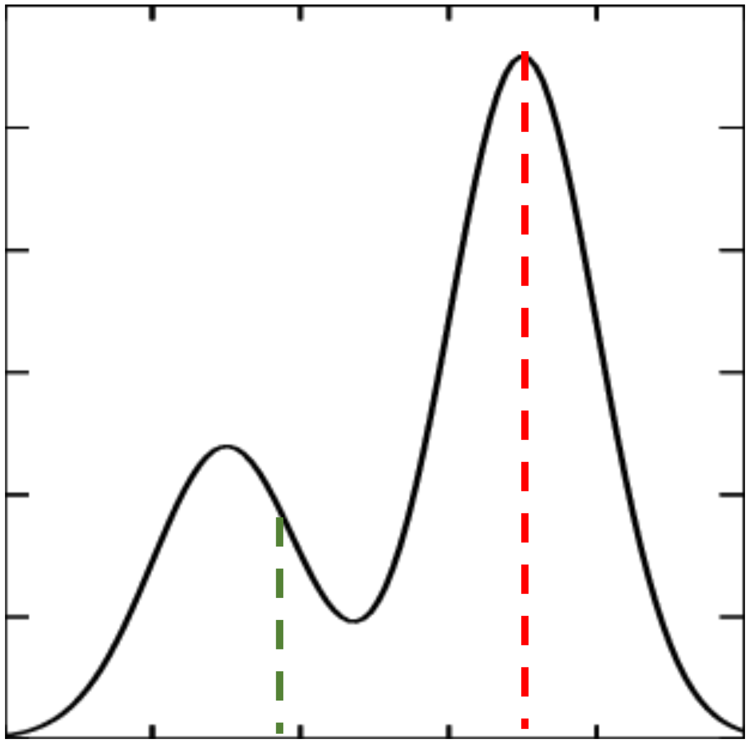
Predictions

$$p(x) = \int_\theta p_\theta(x) p(\theta|\text{data})$$

Single point vs full posterior

Model Fitting

$p(\theta|\text{data})$



$\hat{\theta}_{\text{data}^*}$ $\hat{\theta}_{\text{data}}$ θ

Predictions

$$p(x) = \int_\theta p_\theta(x) p(\theta|\text{data})$$

Out-of-sample data

$$y \sim p(Y; \theta)$$

$$y \sim p(Y; \tilde{\theta})$$

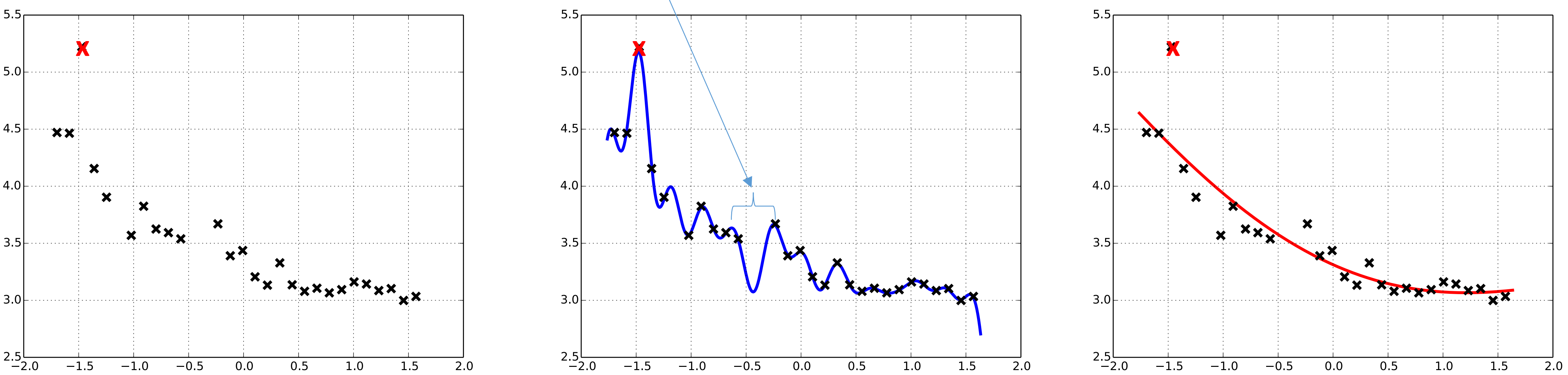
Photo: Karen Plimmer, Terry Chuang; Boredpanda

Generalization

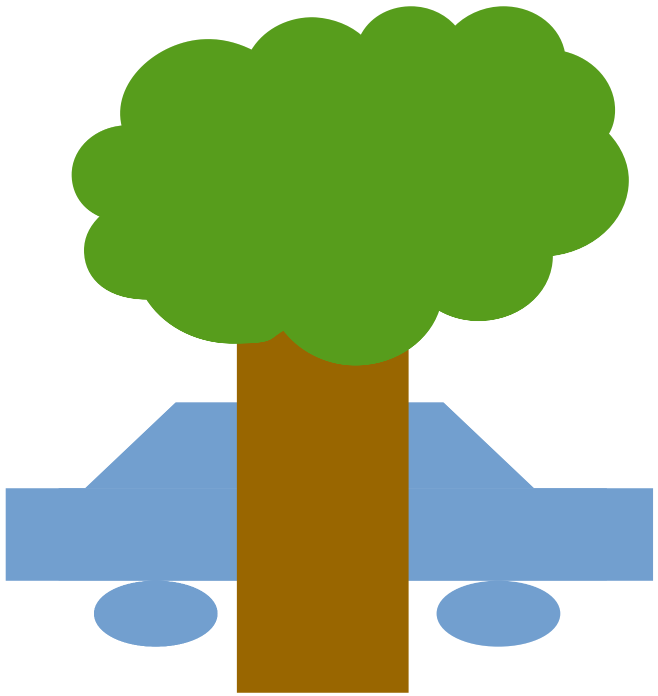
- In a stretch, every point not in our training set can be thought of as out-of-sample.

- Knowing what you don't know helps to not be over-confident. i.e *generalize* well

  Example: recognizing *epistemic uncertainty* helps being regularized

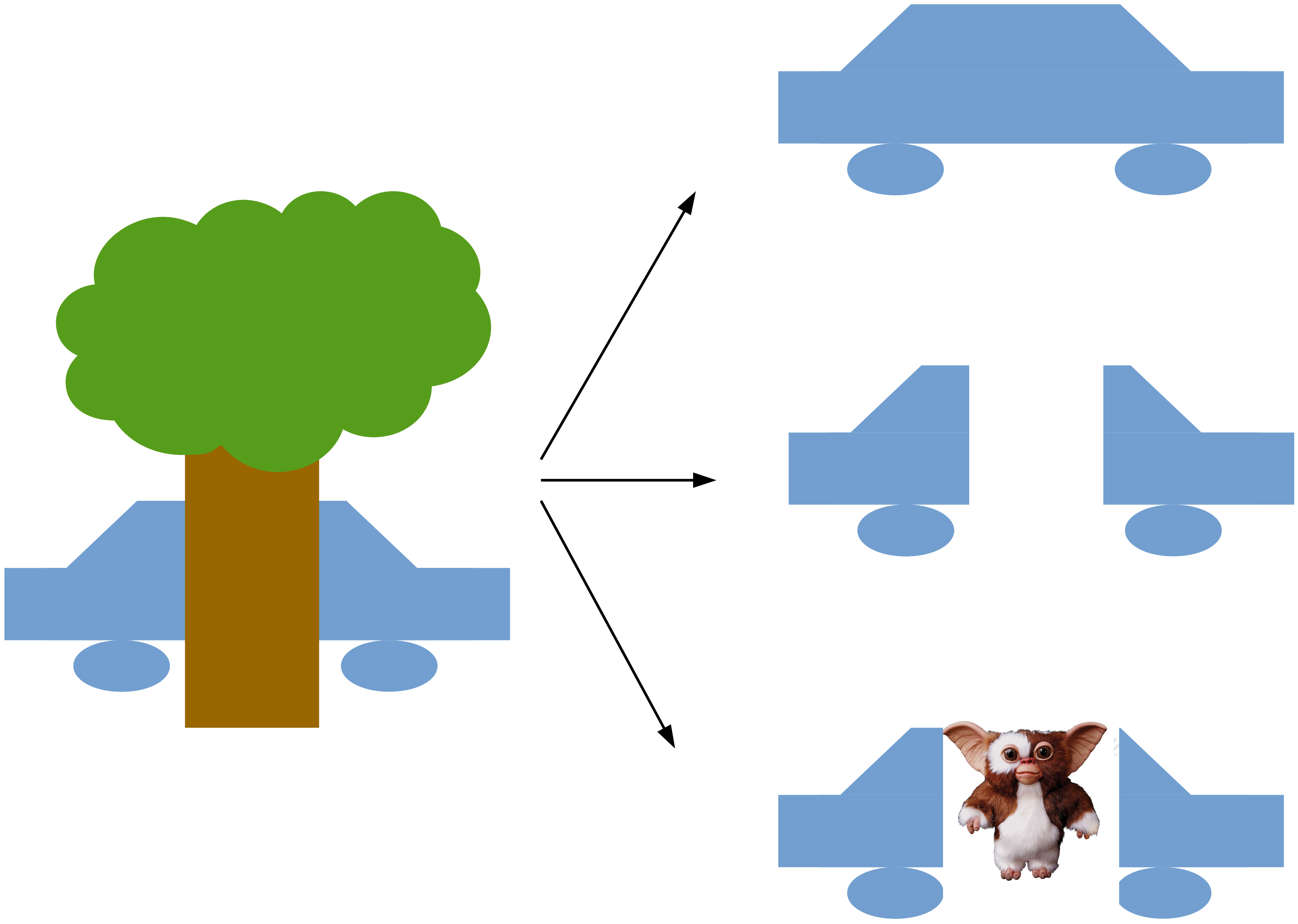


Occam’s Razor



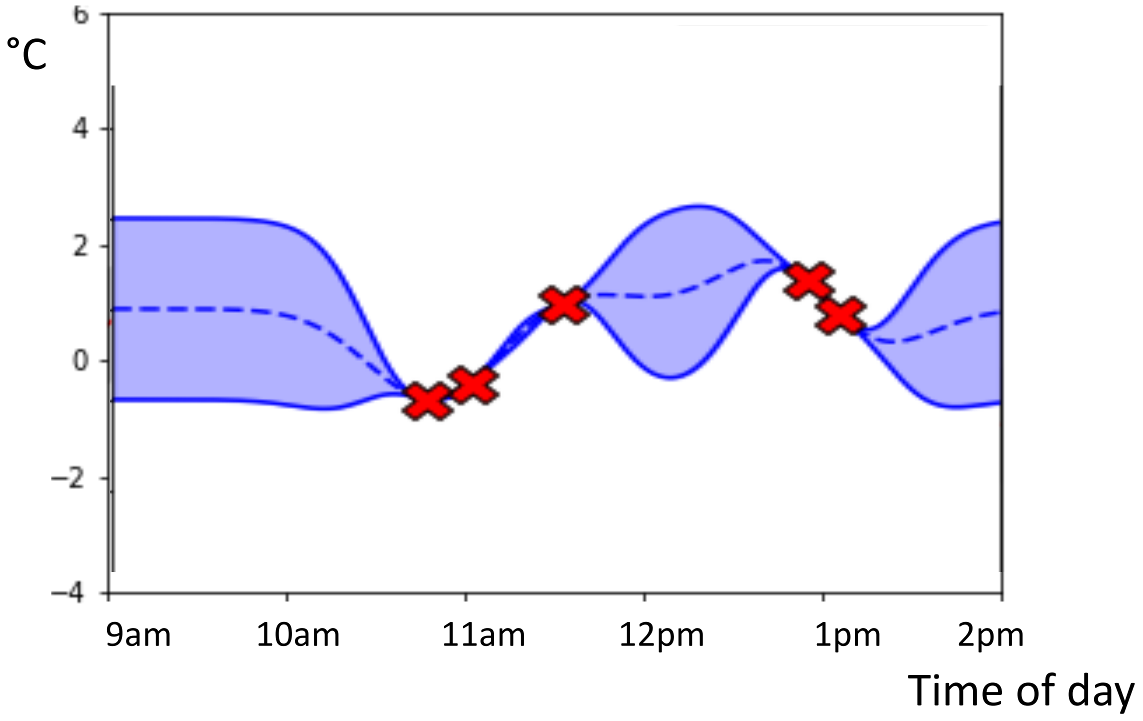
Occam’s Razor

- Which inference is more *probable*?
- Which is *simpler*?



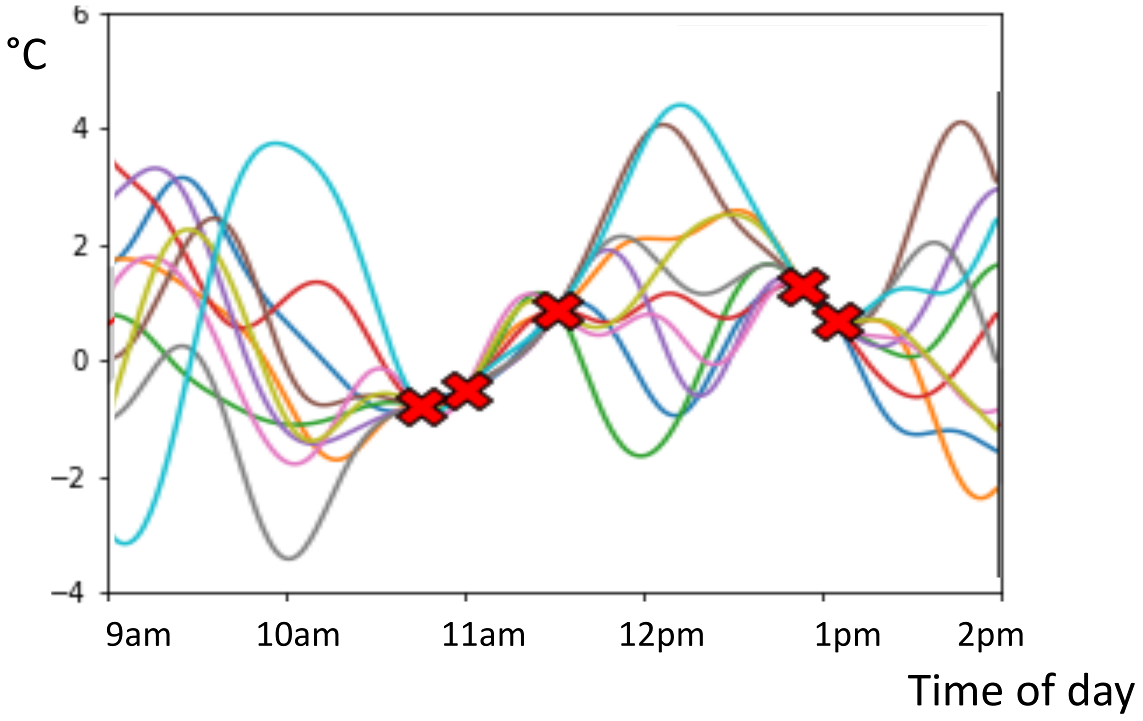
It's not magic: we combine **data** with **assumptions**

Uncertainty in Regression: Balancing *some notion* of simplicity,
with *some notion* of smoothness & prior knowledge



It’s not magic: we combine **data** with **assumptions**

Uncertainty in Regression: Balancing *some notion* of simplicity,
with *some notion* of smoothness & prior knowledge



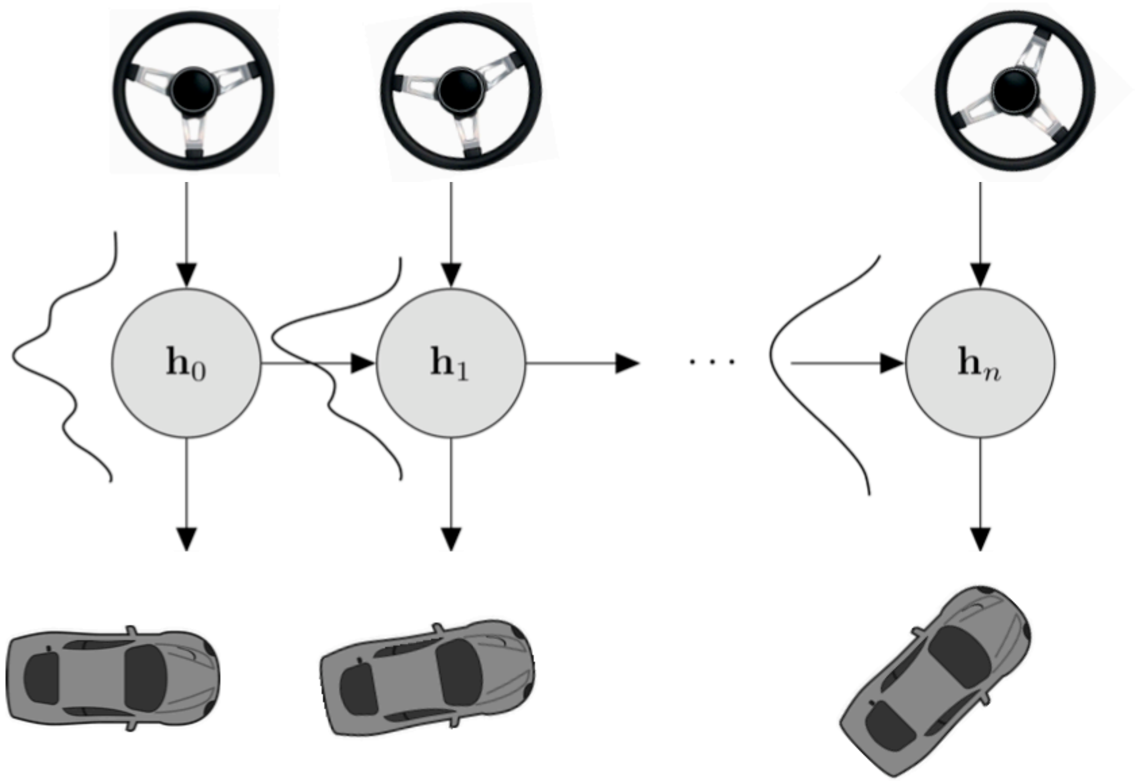
Propagation of uncertainty

Example: A ML model is trying to estimate the internal state of a car-system relying on noisy sensors.

Control:

System Model:

Outcome:



Caveats

- Uncertainty modeling is computationally expensive
- Uncertainty propagation is even harder
- Approximations must often be used
- Calibration is not always guaranteed
- Mis-uses of uncertainty (e.g. mis-interpretations)

Thanks!

Questions?

Do you want to share your story?

Do you have an application where predictions alone aren’t sufficient?
..or where uncertainty needs to be propagated across the scientific pipeline?