What Uncertainty in ML Can and Cannot Tell Us

Christian Henning

Talk at the Institute of Computer Research, University of Alicante

23 October 2026

© 2026 Christian Henning

About me

ML researcher & engineer

2017–2022

PhD, ETH Zurich

Bayesian deep learning, continual learning, uncertainty estimation

2022–2026

Ethon

First employee; led computer vision for quality inspection and built R&D

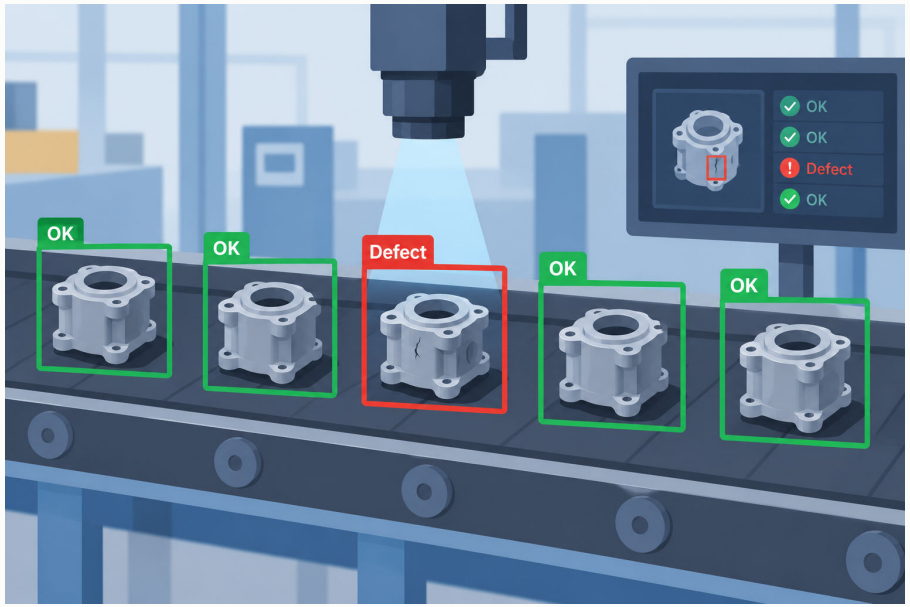


Image generated with ChatGPT

since 2026

AWS

AI for healthcare & life sciences

© 2026 Christian Henning

> Essentially, all models are wrong, but some are useful.
>
> George Box

Uncertainty is part of the model, so it is wrong too.

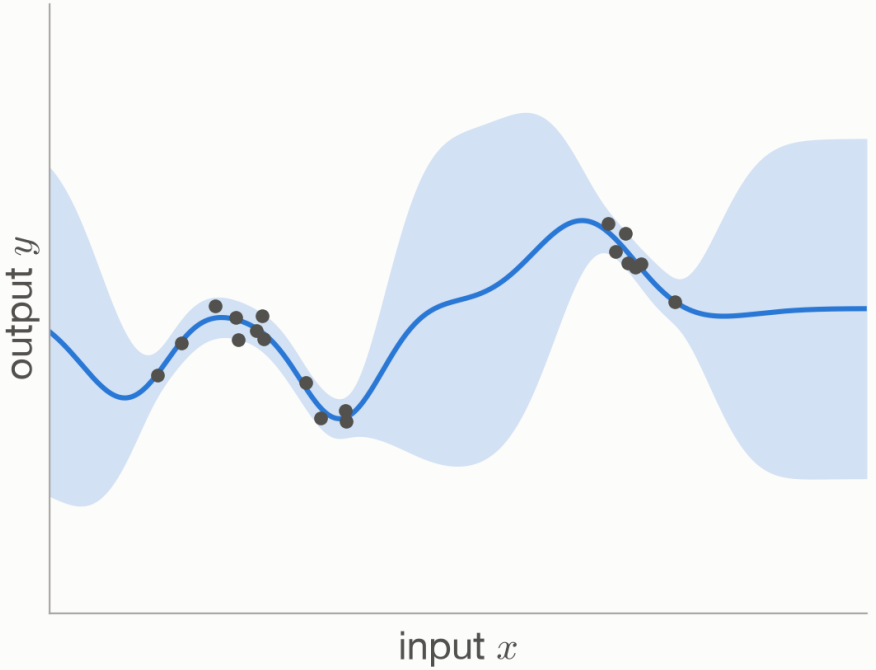
How can we make it useful?

© 2026 Christian Henning · Box & Draper (1987), Empirical Model-Building and Response Surfaces

Uncertainty is incomplete knowledge

What counts as random depends on what we know: a coin flip follows classical mechanics, so it becomes predictable once we measure how it was thrown

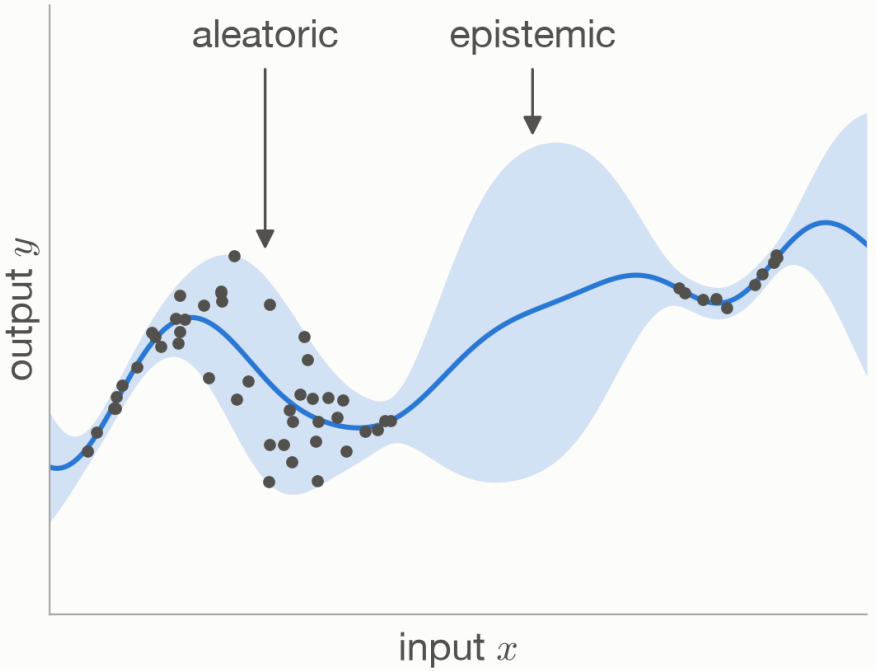
- Probability is the language we use to represent uncertainty over outcomes



© 2026 Christian Henning

What we haven't learned vs. what we can't observe

- **Epistemic:** we don't know the true relationship between inputs and outputs. More data reduces it
- **Aleatoric:** what we observe doesn't determine the outcome. More data of the same kind won't remove it



Richer inputs can reduce aleatoric uncertainty but leave more to learn.
Whether that pays off depends on how much data we have

© 2026 Christian Henning

Why care? The prediction isn't the decision

- **A human decides:** the model doesn't know what's at stake. "20% rain" means one thing for a wedding, another for a walk

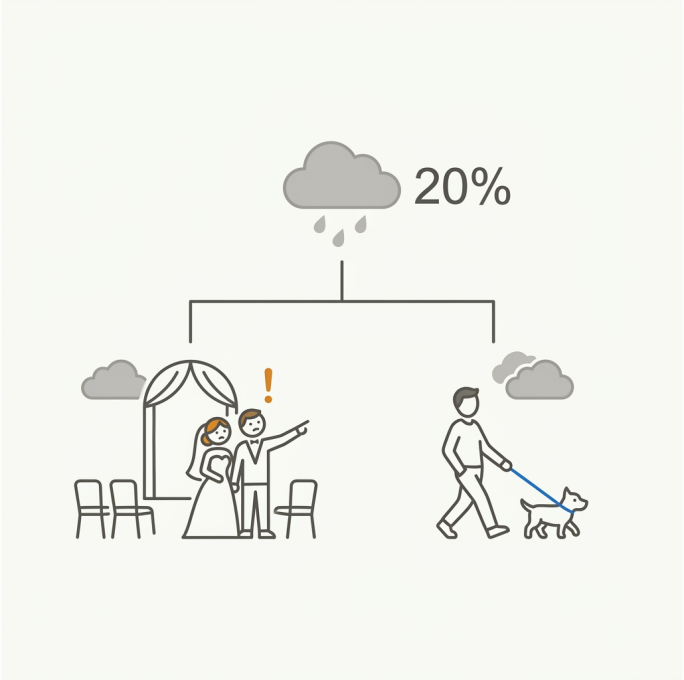


Image generated with Gemini

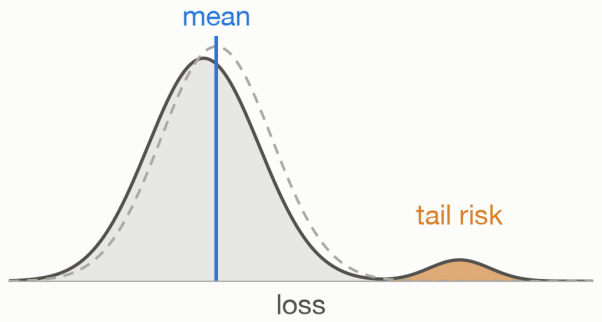
© 2026 Christian Henning

Why care? The prediction isn't the decision

- **A human decides**
- **The mean isn't enough:** we want to capture distinct choices or quantify tail risks



By United States Department of Agriculture, Public Domain, via Wikimedia Commons



© 2026 Christian Henning

Why care? The prediction isn't the decision

- **A human decides**
- **The mean isn't enough**
- **We can choose not to decide yet:** selective automation hands unsure cases to a human

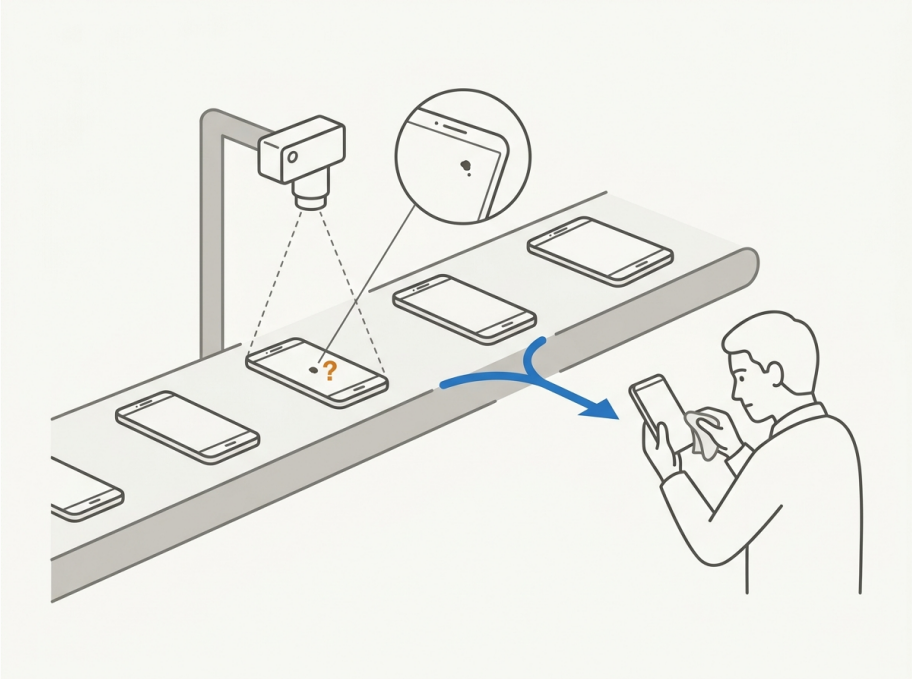
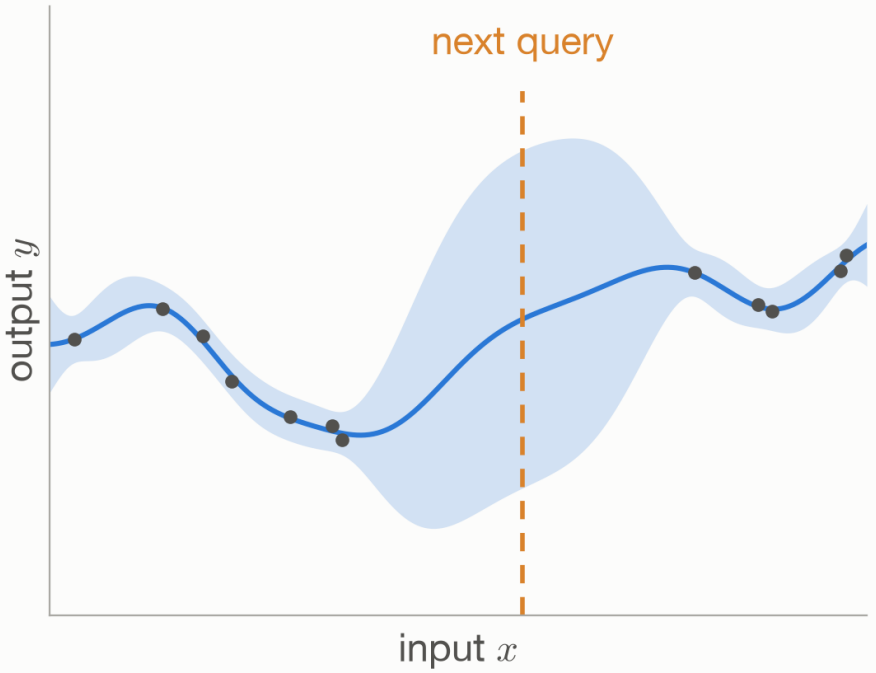


Image generated with Gemini

© 2026 Christian Henning

Why care? The prediction isn't the decision

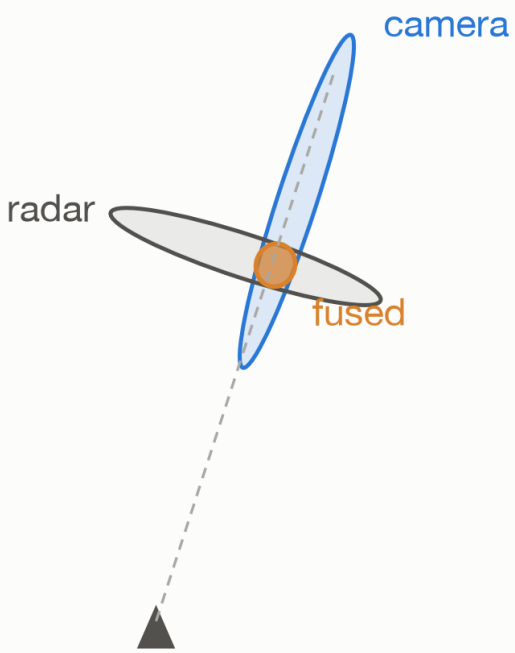
- **A human decides**
- **The mean isn't enough**
- **We can choose not to decide yet**
- **Learning where it pays:** active learning and Bayesian optimisation query where the model knows least



© 2026 Christian Henning

Why care? The prediction isn't the decision

- **A human decides**
- **The mean isn't enough**
- **We can choose not to decide yet**
- **Learning where it pays**
- **The uncertainty feeds further computation:** sensor fusion and Kalman filters use it to weigh each input



© 2026 Christian Henning

Why care? The prediction isn't the decision

- **A human decides:** the model doesn't know what's at stake. "20% rain" means one thing for a wedding, another for a walk
- **The mean isn't enough:** we want to capture distinct choices or quantify tail risks
- **We can choose not to decide yet:** selective automation hands unsure cases to a human
- **Learning where it pays:** active learning and Bayesian optimisation query where the model knows least
- **The uncertainty feeds further computation:** sensor fusion and Kalman filters use it to weigh each input

All of these need the uncertainty to be useful. When is it?

© 2026 Christian Henning · Examples, not an exhaustive list

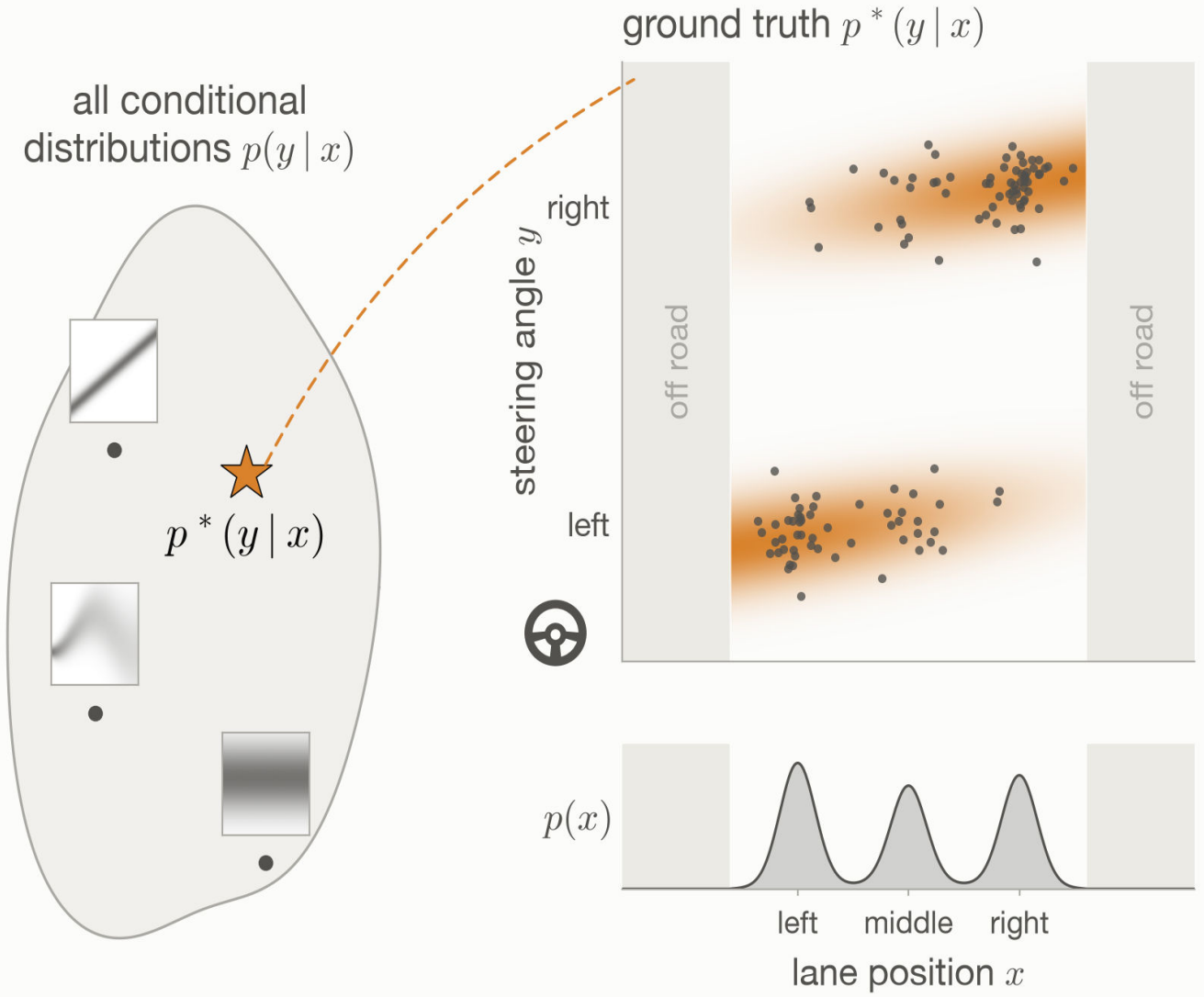
Searching for the truth

Models, hypotheses and learning from noisy data

© 2026 Christian Henning · Details in chapters 2 and 3 of Henning (2022), Knowledge uncertainty and lifelong learning in neural systems, PhD thesis



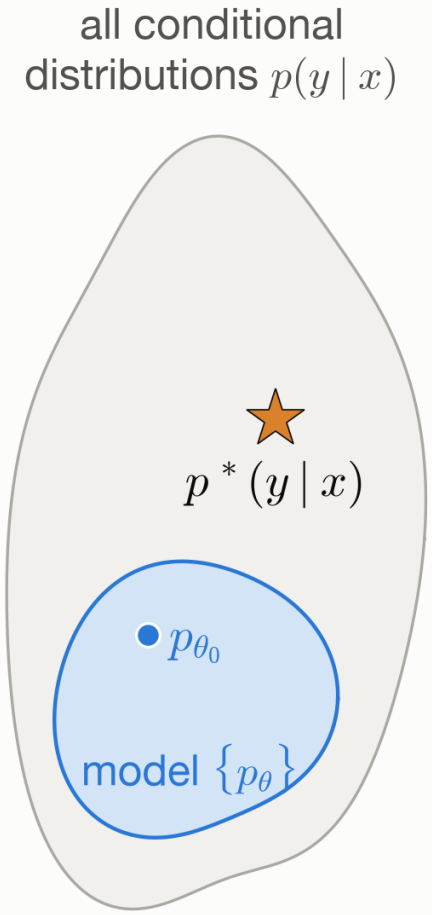
The space of all hypotheses

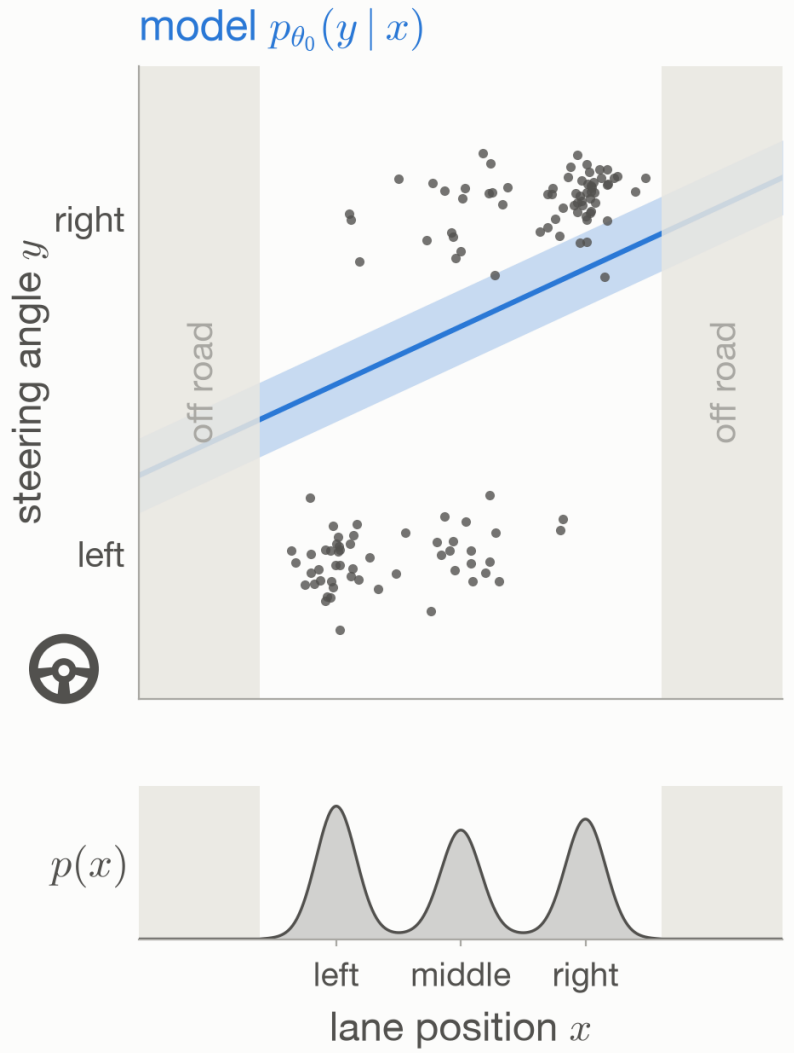


© 2026 Christian Henning

A model is a subset of hypotheses

- A **model** is a family of hypotheses, indexed by parameters θ
- Here: a Gaussian with learned mean and fixed variance
  $p_\theta(y \mid x) = \mathcal{N}\left(y;\ f_\theta(x),\ \sigma^2\right)$
- The **truth** need not lie inside the model

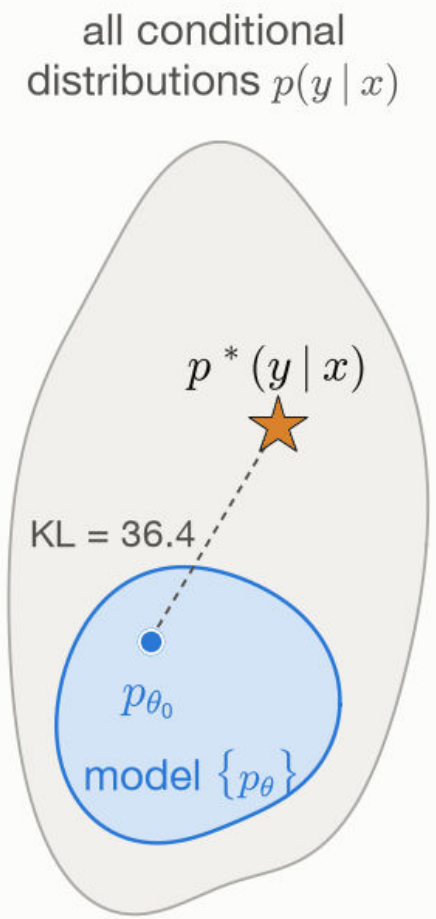


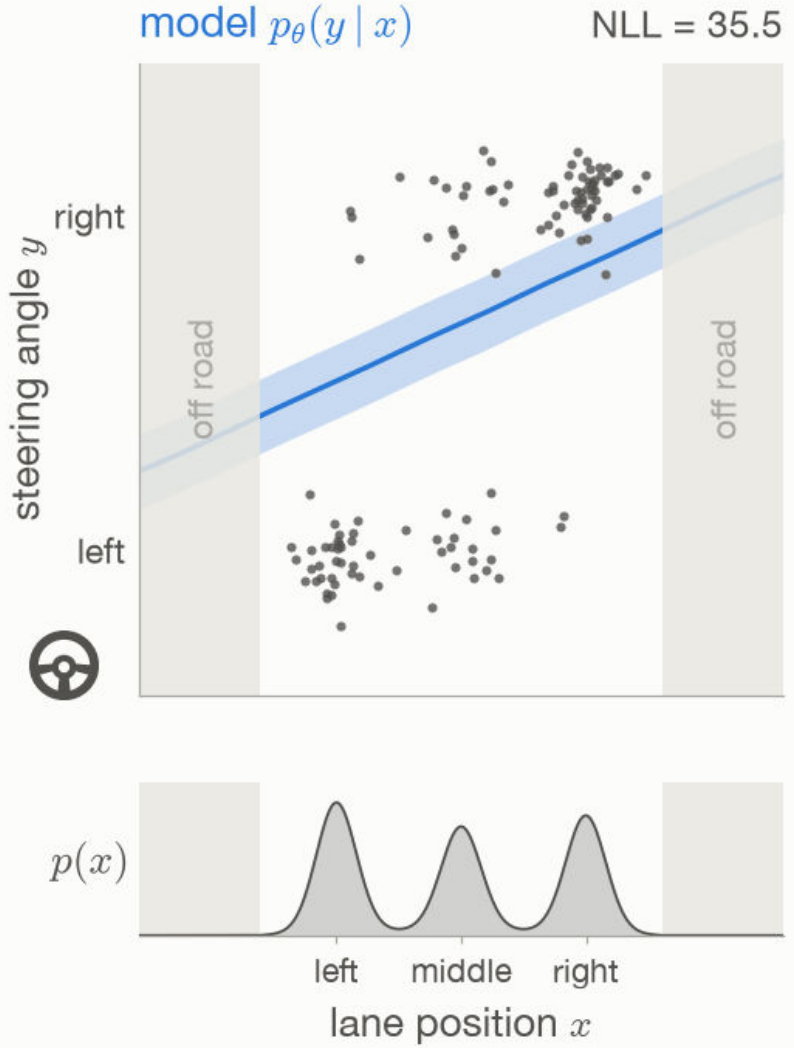


© 2026 Christian Henning

Training moves us closer to the truth

- We minimize the **negative log-likelihood** (NLL) on data
- For a Gaussian with fixed variance, this is the familiar **mean squared error**:
  $\mathrm{NLL} = \frac{1}{2\sigma^2}\,\mathrm{MSE} + \mathrm{const}$
- We can only estimate the NLL, which doesn't tell us how wrong we are:





$$\mathbb{E}_{x,y}[\mathrm{NLL}] = \mathbb{E}_x\Big[\mathrm{KL}\big(p^* \,\|\, p_\theta\big)\Big] + \mathbb{E}_x\big[H(p^*)\big]$$

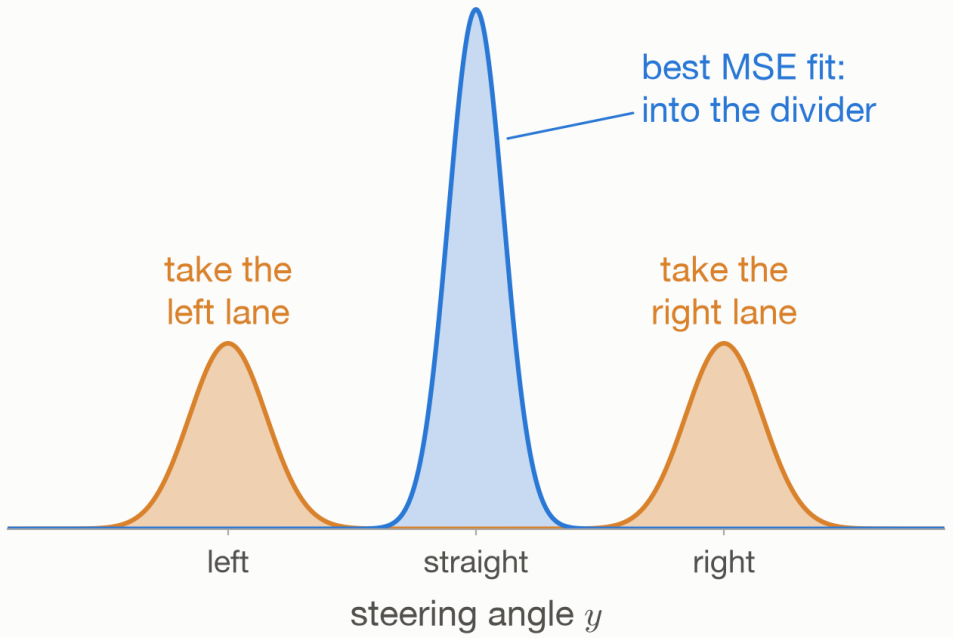
© 2026 Christian Henning



The best fit steers into the divider

- Drivers take the **left** or the **right** lane
- Any f_θ in $\mathcal{N}\left(y;\ f_\theta(x),\ \sigma^2\right)$ averages the two

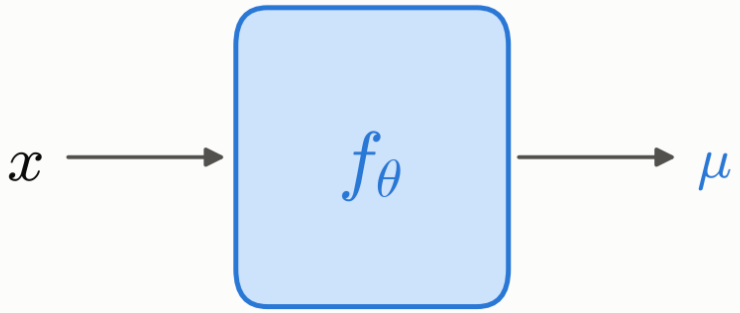
middle lane: $p(y \mid x = \text{middle})$



© 2026 Christian Henning · Cervera … Henning (2021), Uncertainty estimation under model misspecification · Henning (2025), Why an MSE loss might make your self-driving car crash

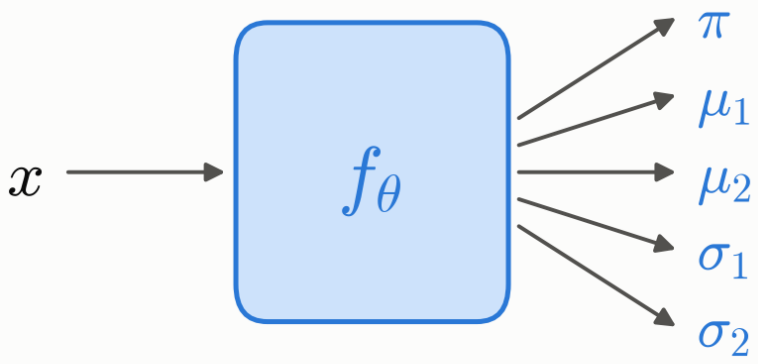
Reducing model misspecification

one Gaussian, fixed variance



$$p_\theta(y \mid x) = \mathcal{N}(y;\ \mu,\ \sigma^2)$$

mixture of two Gaussians

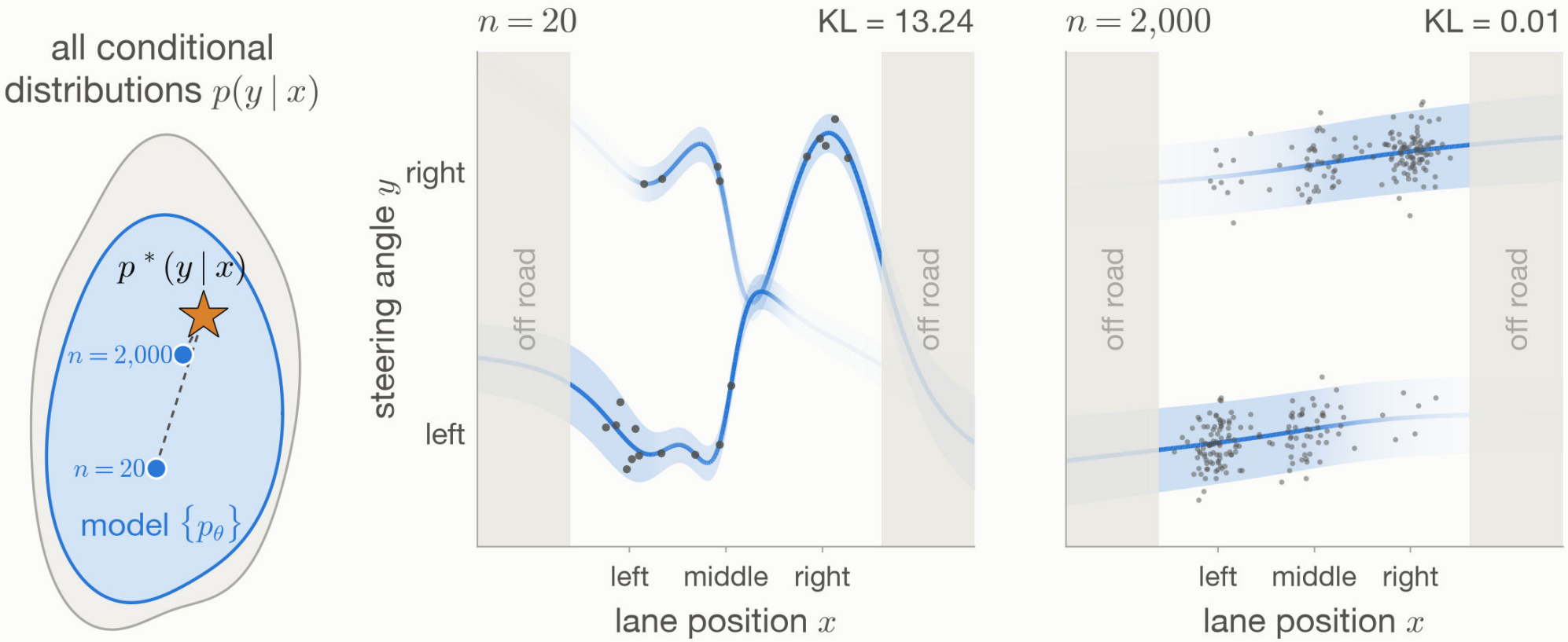


$$p_\theta(y \mid x) = \pi\,\mathcal{N}(y;\ \mu_1,\ \sigma_1^2) + (1-\pi)\,\mathcal{N}(y;\ \mu_2,\ \sigma_2^2)$$

As modellers, it's our job to understand the problem well enough to make good assumptions

© 2026 Christian Henning

A richer model contains the truth, but needs data

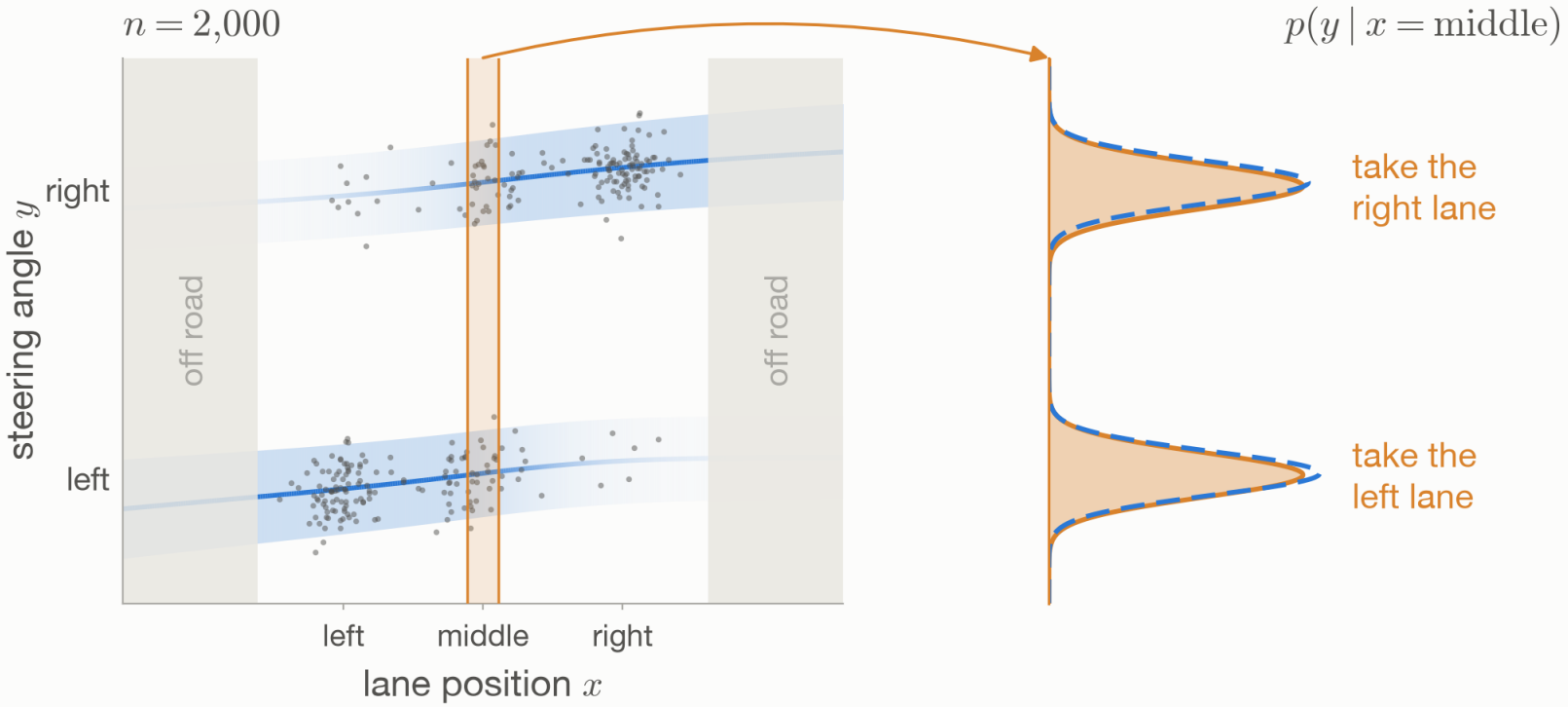


More hypotheses leave more room for ambiguity: without helpful inductive biases, little data can't single out the truth

© 2026 Christian Henning

A good fit captures aleatoric uncertainty





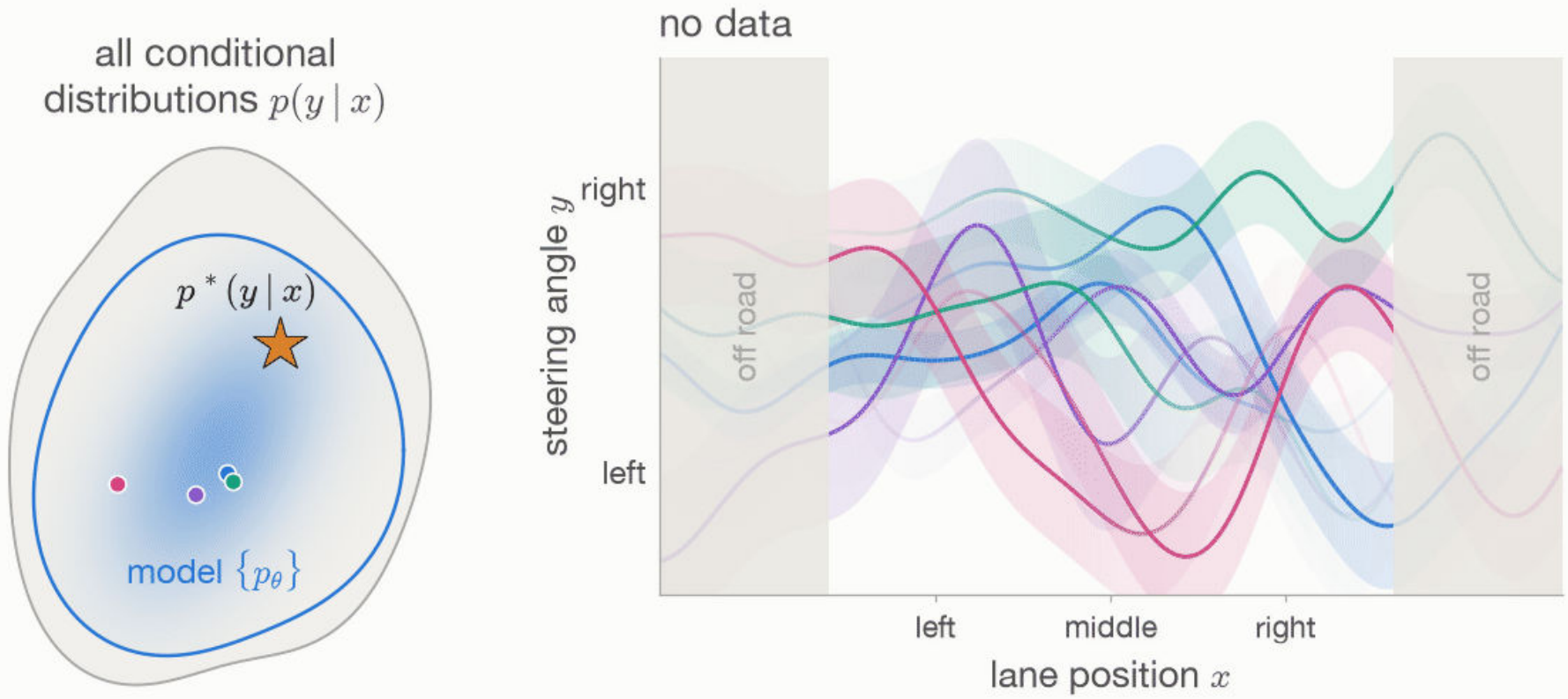
Aleatoric: even a perfect model can't say which lane the driver will take

---

But how do we know our fit is good?

© 2026 Christian Henning

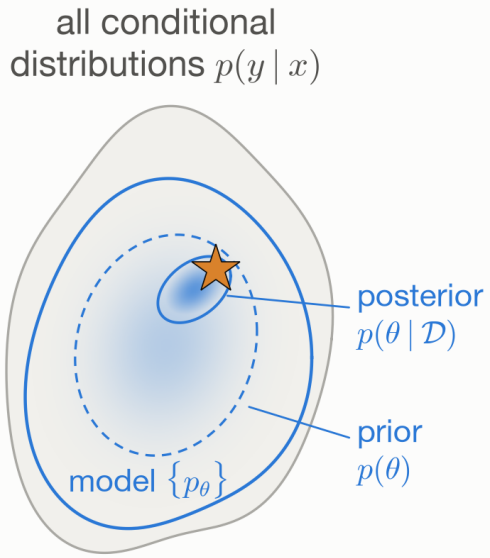
Instead of one fit, keep every plausible fit



- Before any data, we put a soft weighting on all hypotheses in the model: the **prior**
- Data moves the weight to the hypotheses that explain it: the **posterior**

© 2026 Christian Henning

Bayes: weigh all hypotheses, don't commit to one



$$\underbrace{p(\theta \mid \mathcal{D})}_{\text{posterior}} = \frac{\overbrace{\prod_i p_\theta(y_i \mid x_i)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{p(\mathcal{D})}$$

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ is our data; the likelihood measures how well hypothesis θ agrees with it

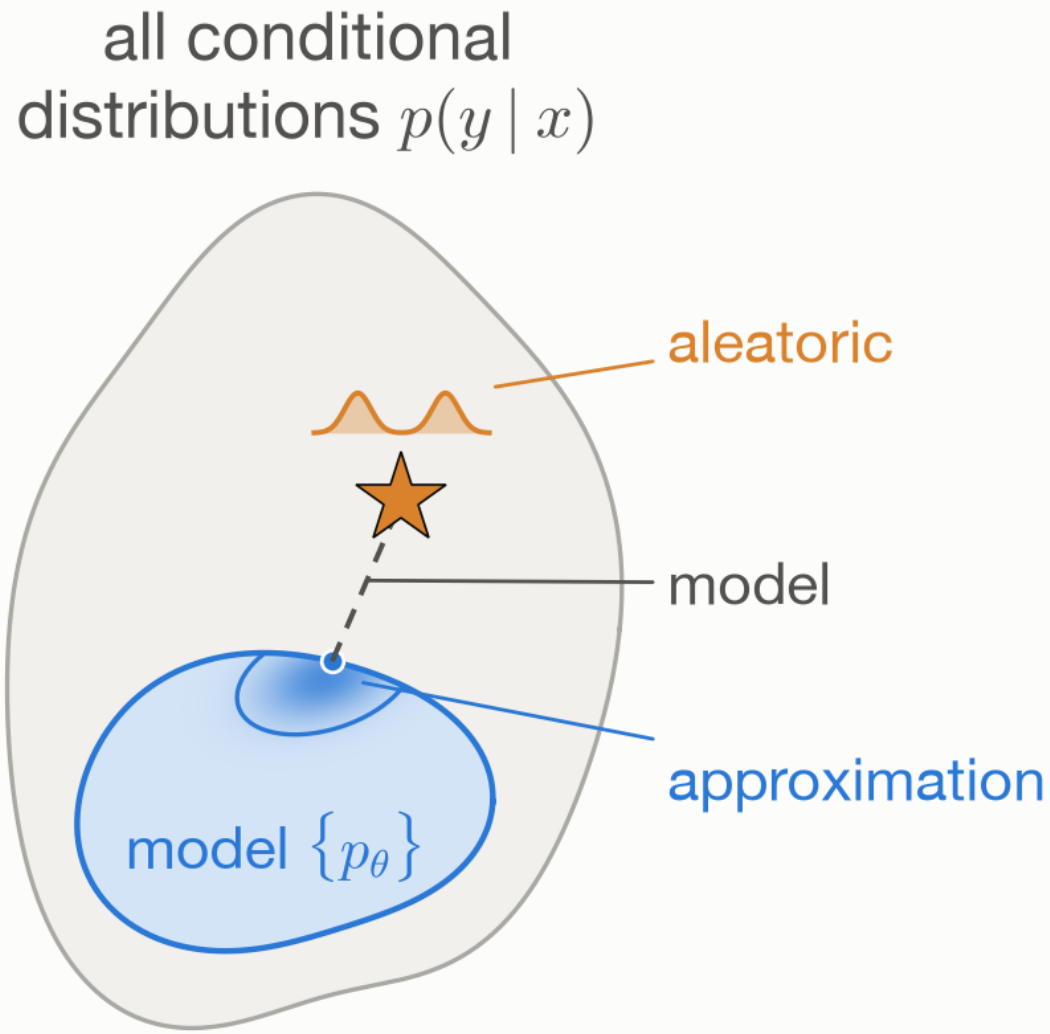
To predict, we average all hypotheses, weighted by the posterior: the **posterior predictive**

$$p(y \mid x, \mathcal{D}) = \int p_\theta(y \mid x)\, p(\theta \mid \mathcal{D})\, \mathrm{d}\theta \xrightarrow{n \to \infty} p_{\hat{\theta}_{\mathrm{MLE}}}(y \mid x)$$

© 2026 Christian Henning

Sources of uncertainty

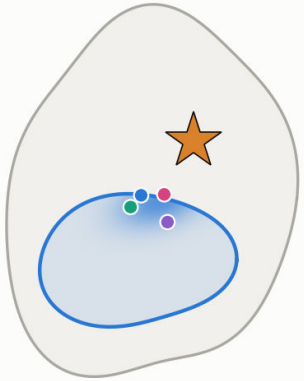
- **Aleatoric:** the truth itself is noisy
- **Epistemic:** what we haven't learned
  - **Approximation:** which hypothesis in the model? The posterior captures it
  - **Model:** is the truth in the model at all? The posterior can't see it

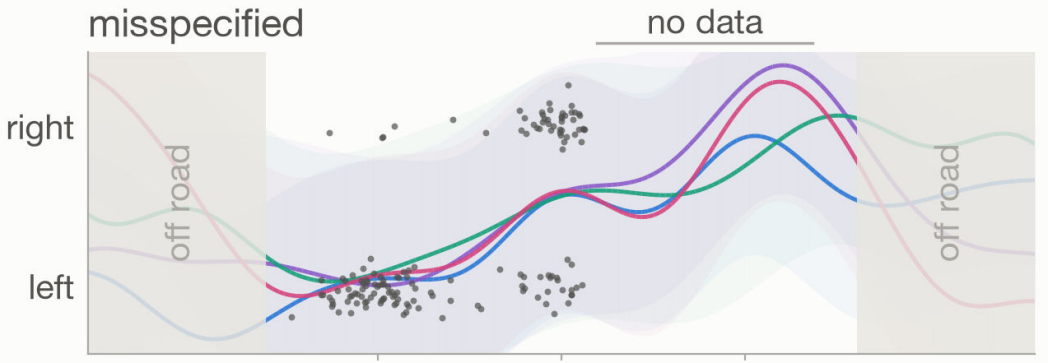


© 2026 Christian Henning · Terms from Hüllermeier & Waegeman (2021), Aleatoric and epistemic uncertainty in machine learning

Bayes tells us where we lack data, not whether our model is wrong

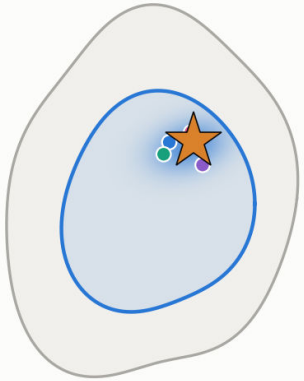
$\mathcal{N}(y;\, \mu,\, \sigma^2)$

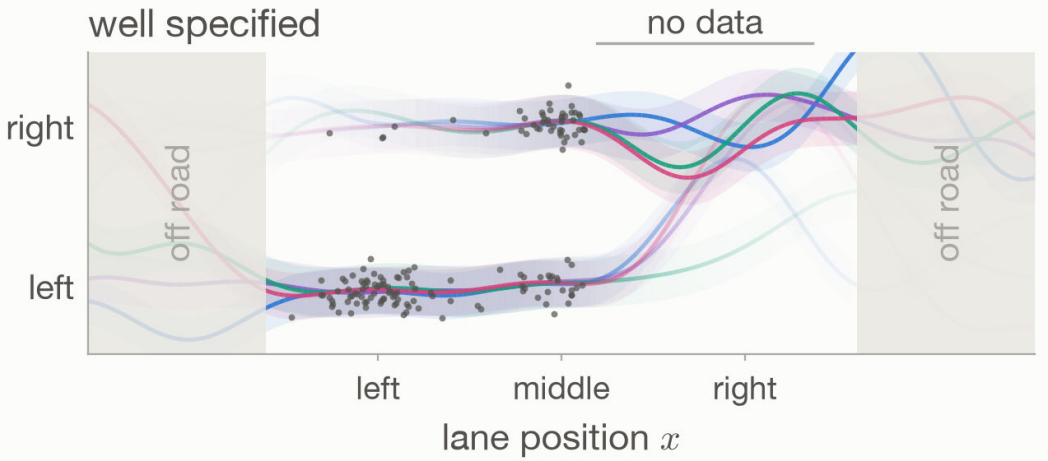




$\log p(\mathcal{D} \mid \mathcal{M}) = -159$

$\pi\,\mathcal{N}(y;\, \mu_1,\, \sigma_1^2)$
$+\,(1-\pi)\,\mathcal{N}(y;\, \mu_2,\, \sigma_2^2)$





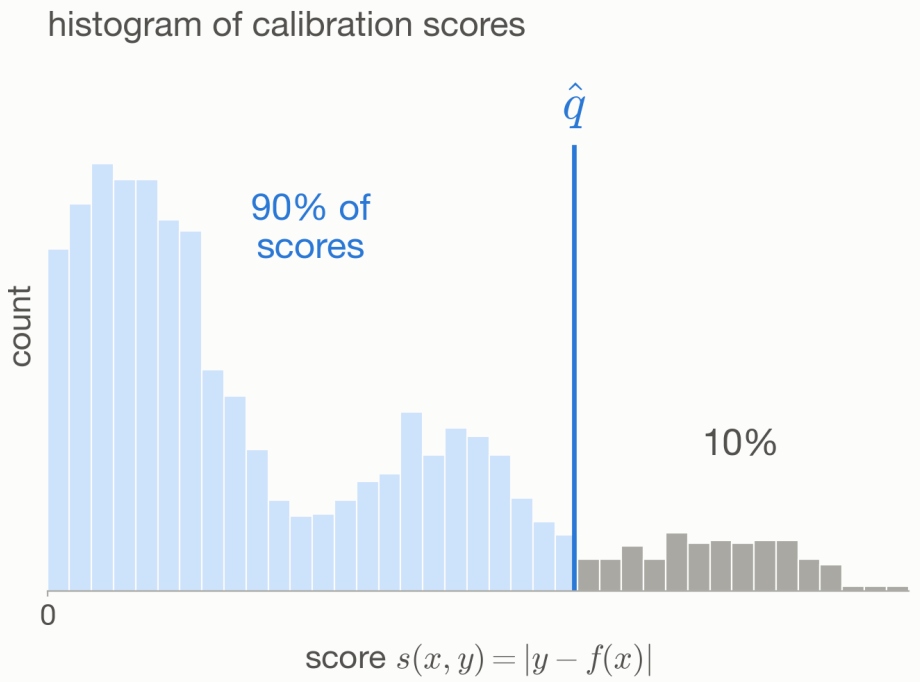
$\log p(\mathcal{D} \mid \mathcal{M}) = 295$

- **Bayesian Occam's razor:** the evidence $p(\mathcal{D} \mid \mathcal{M}) = \int p(\mathcal{D} \mid \theta)\, p(\theta \mid \mathcal{M})\, \mathrm{d}\theta$
- It ranks models, but only those we try

© 2026 Christian Henning

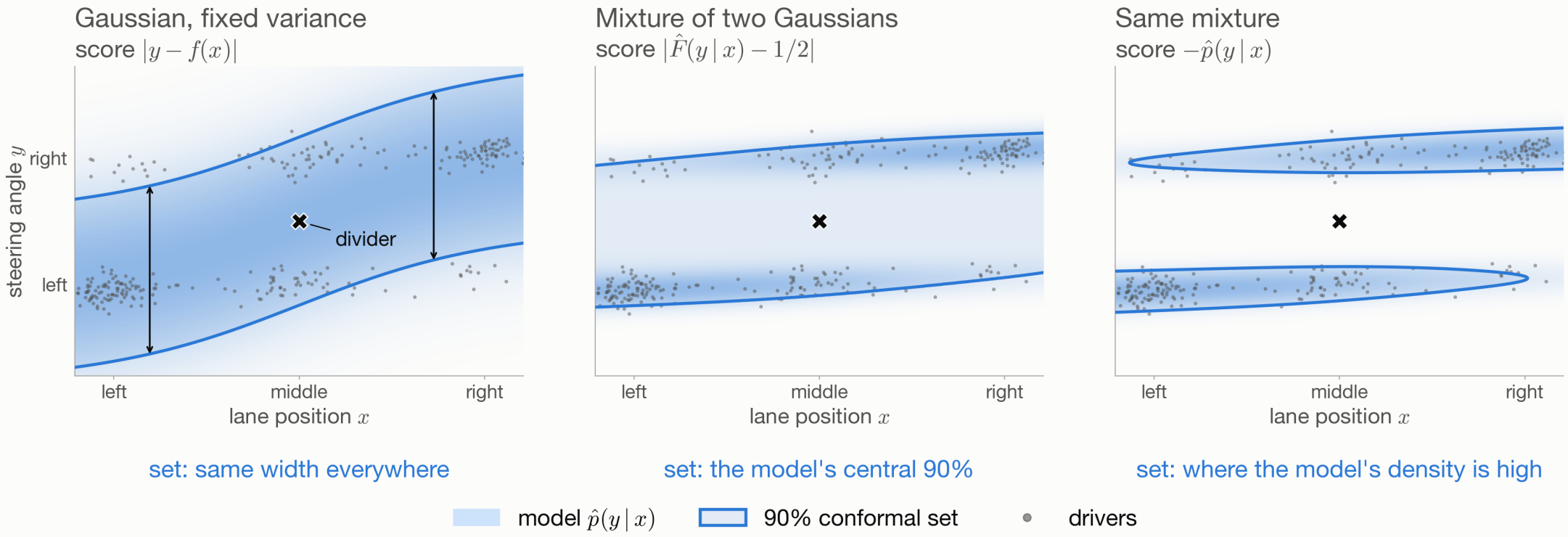
How to get a guarantee: conformal prediction

- Hold out a **calibration set**
- Pick a **score** for how wrong the model is
- Take the 90% quantile $\hat{q}$ of the calibration scores
- **Prediction set** for a new x: every y with $s(x, y) \leq \hat{q}$



Guarantee: for any model, the set contains the true y 90% of the time, on average over inputs

The guarantee is free, a useful set is not



- **Choose the model:** a model with one lane gives a set with one lane
- **Choose the score:** a good model can still give a set through the divider

© 2026 Christian Henning · Middle score: Chernozhukov, Wüthrich & Zhu (2021), Distributional conformal prediction · Henning (2026), Uncertainty is not distribution-free

From uncertainty estimation to decision making

© 2026 Christian Henning

Decision theory: pick the action with the lowest expected loss

- The **model** provides the belief $p(y \mid x)$
- Whoever acts provides the **loss** $L(a, y)$: the cost of action a if the outcome is y
- We act to minimize the expected loss:

$$a^* = \arg\min_a \ \mathbb{E}_{y \sim p(y|x)} \left[L(a, y) \right]$$

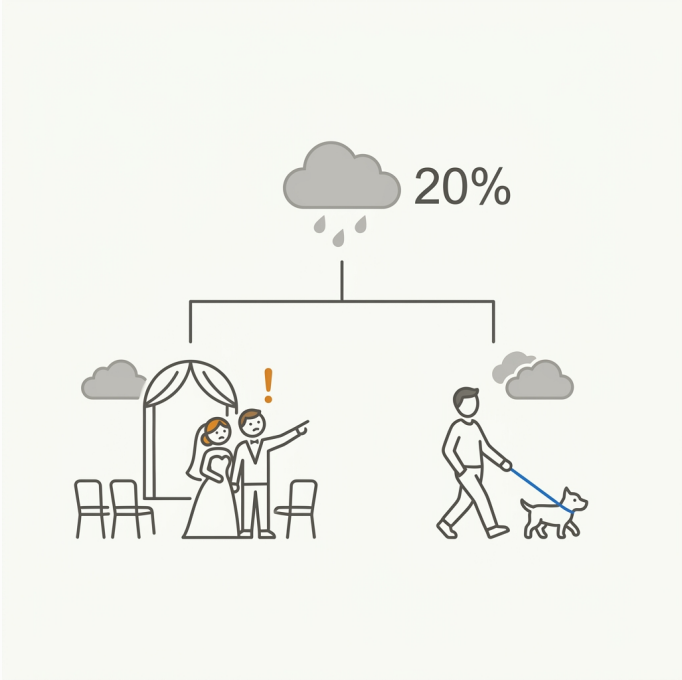


Image generated with Gemini

© 2026 Christian Henning

Example: visual quality inspection

- The **model** scores each part with $p(\text{defect} \mid x)$, and we reject above a threshold
- Does this score need to be faithful?
  - **Calibrated:** of all parts with $p = 0.1$, 10% are defective
  - **Guaranteed:** at most 1% of defects escape

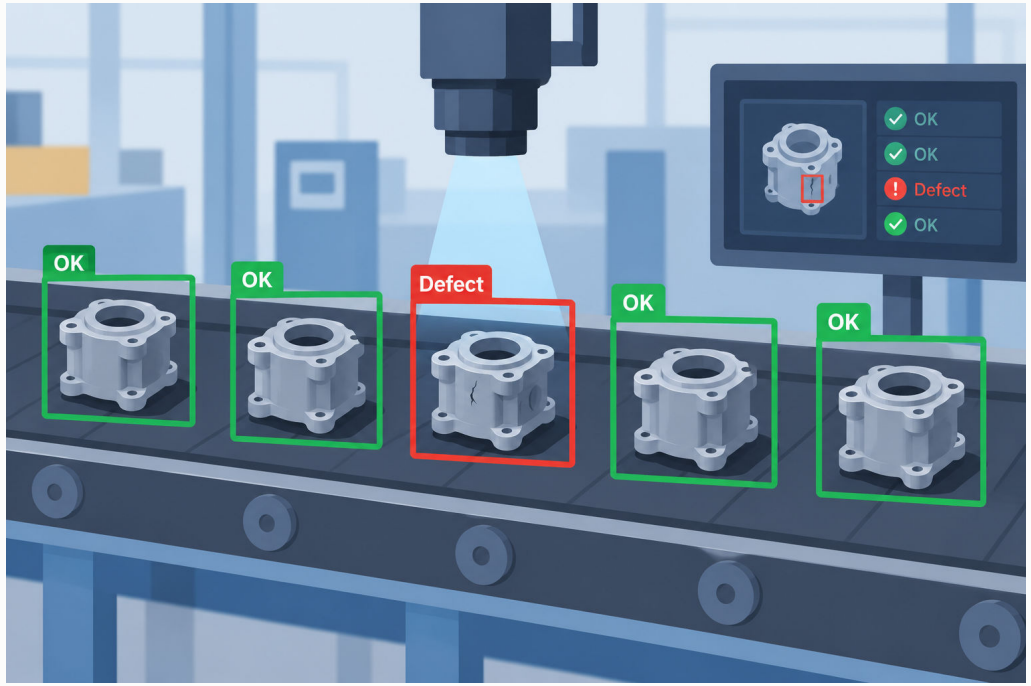


Image generated with ChatGPT

© 2026 Christian Henning

Calibration? Not needed

- Textbook rule: reject if

$$p(\text{defect} \mid x) > \frac{c_{FP}}{c_{FP} + c_{FN}}$$

  c_{FP}: cost of rejecting a good part
  c_{FN}: cost of shipping a bad one

- Calibrated against whom? Inspectors disagree on borderline parts
- c_{FN} (recalls, lost customers, reputation) is rarely known, so the threshold is a guess anyway

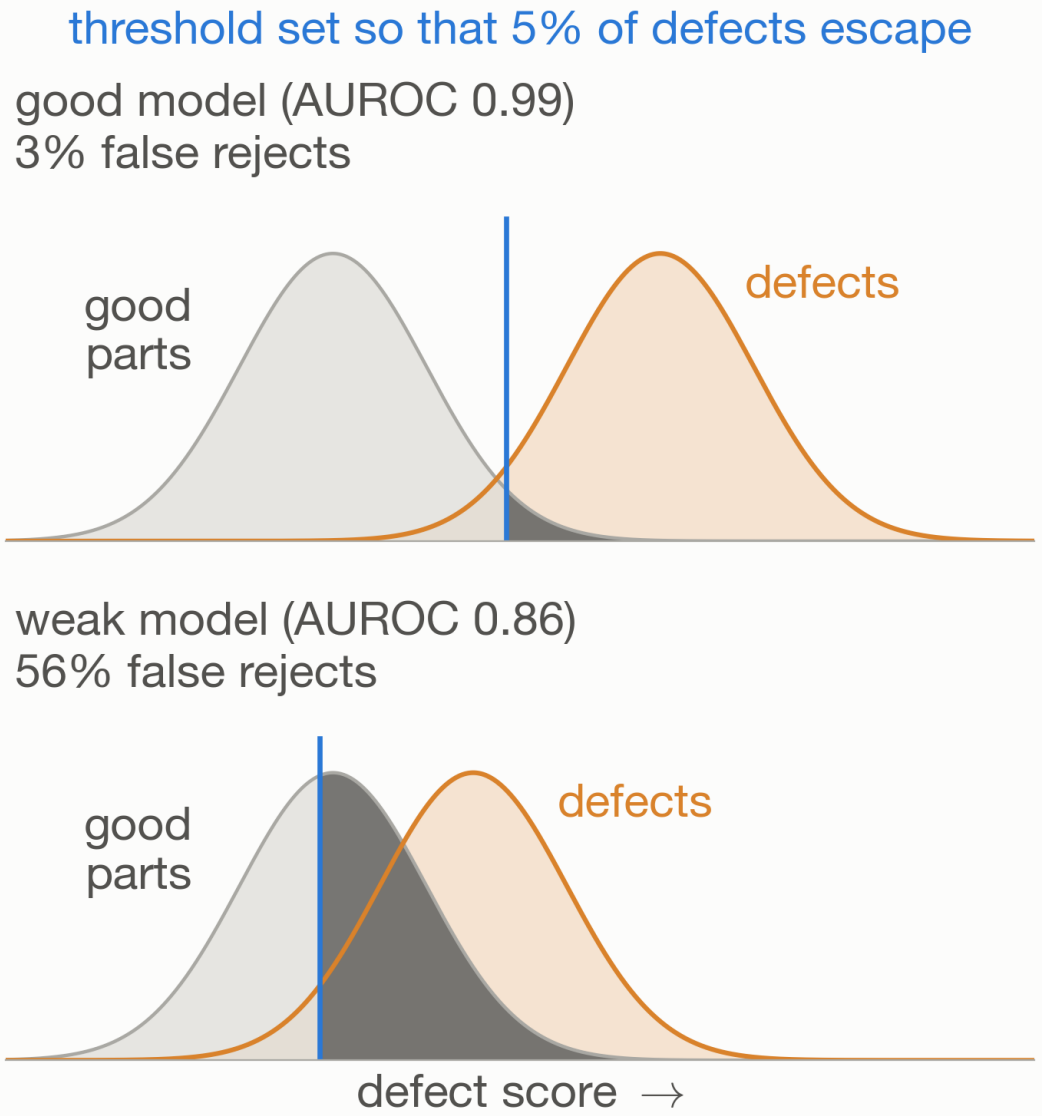


Image generated with ChatGPT

© 2026 Christian Henning

Guarantees? At a price

- Class-conditional conformal prediction, calibrated on defects only, sets the threshold so that at most α of them escape
- Needs about $1/\alpha$ labelled defects, and the next defect may look like none of them
- The guarantee fixes the threshold, however many good parts it rejects
- Too many? We raise it, and are back to a cost we can't measure



© 2026 Christian Henning · Vovk (2012), Conditional validity of inductive conformal predictors

Not faithful, but useful

- Small scratches are rarely called defects, deep cracks almost always
- In between, inspectors disagree, and there is no objective truth
- A useful score mirrors that: less severe, lower score
- Its exact values don't matter, only the order
- In the gray area, the operator sets the threshold

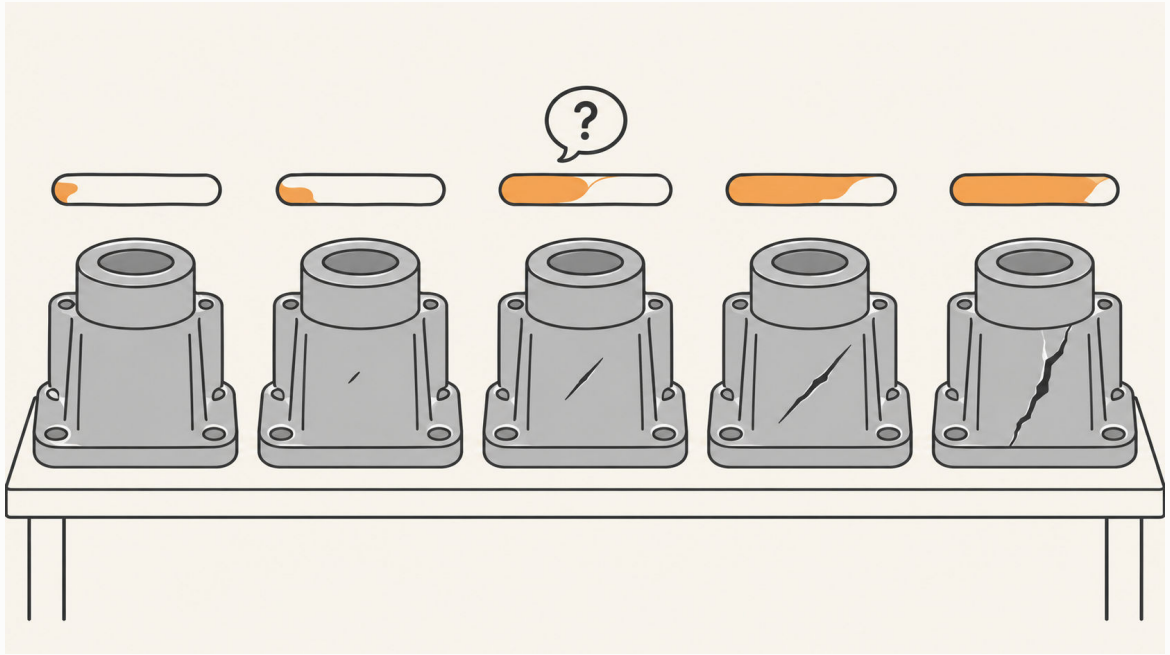


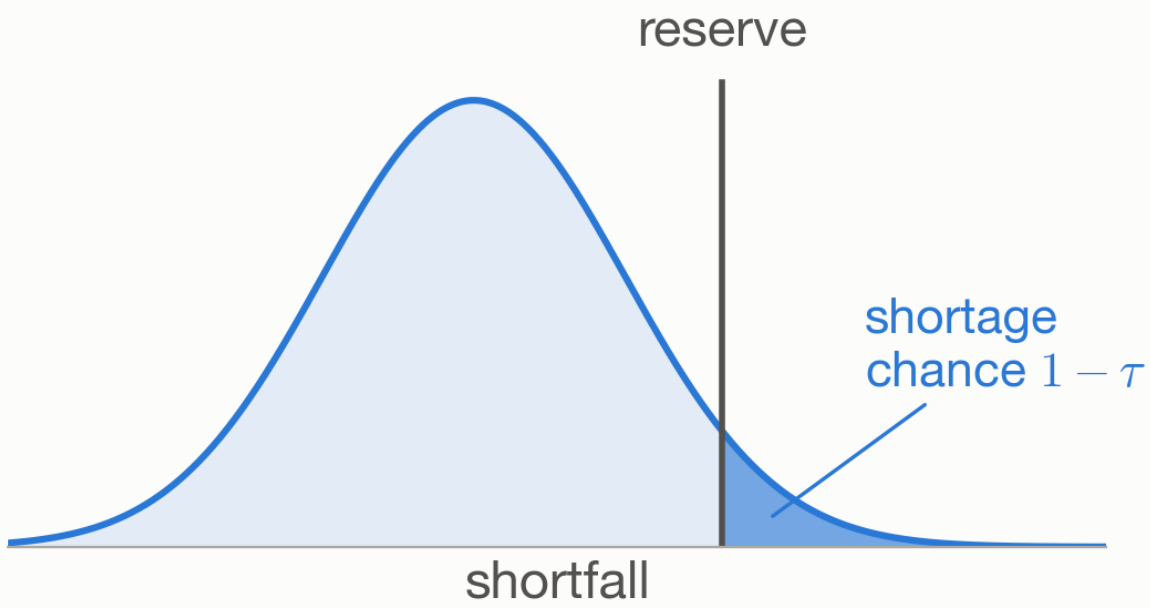
Image generated with ChatGPT

© 2026 Christian Henning

Example: holding reserve for wind power

- The **model** predicts the shortfall: how much less wind power arrives than planned
- The operator keeps backup power on standby, the reserve. If the shortfall exceeds it, each missing unit costs c_{short}. Each unused unit costs c_{idle}
- The best reserve is the τ-quantile of the prediction:

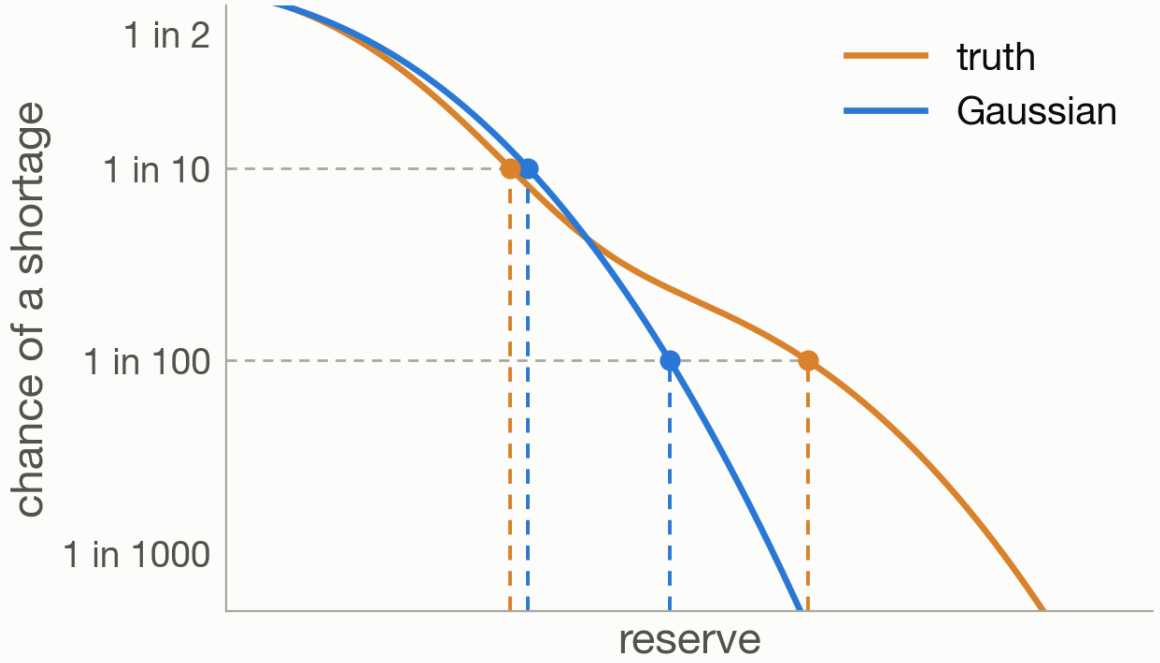
$$\tau = \frac{c_{\text{short}}}{c_{\text{short}} + c_{\text{idle}}}$$



© 2026 Christian Henning

The costs pick the quantile

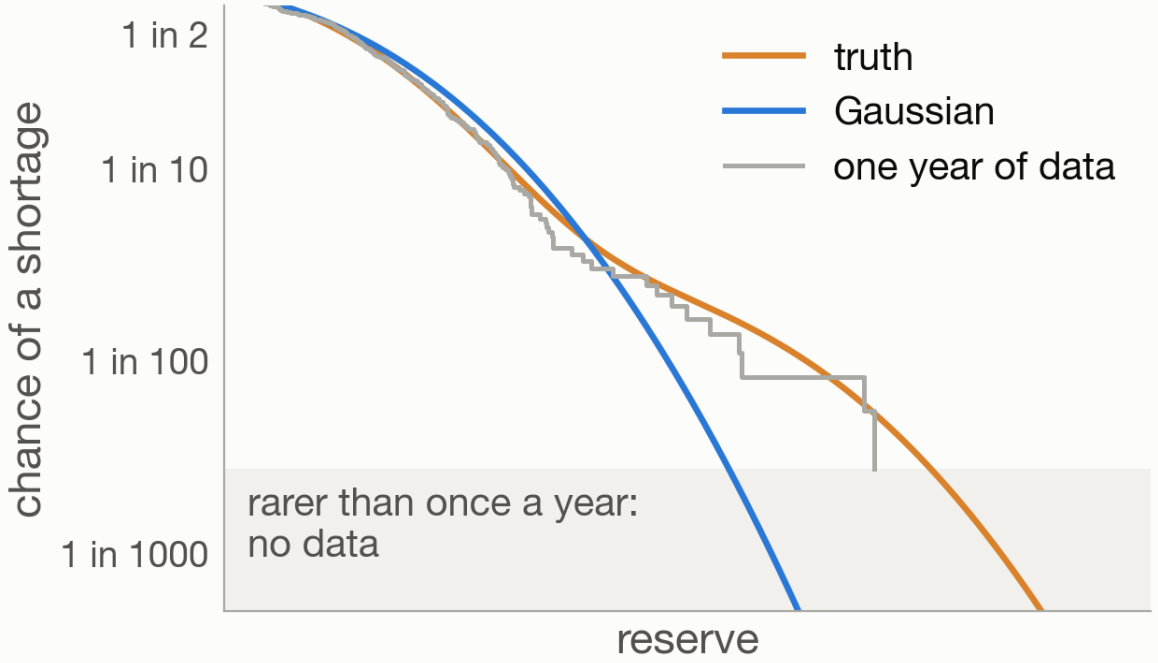
- Prices change by the hour, and with them τ
- The **Gaussian** gets the reserve right at 90%
- At 99% it holds too little, and **shortages** come more than twice as often as planned
- With a changing τ, every quantile has to be right: the whole distribution must be faithful



© 2026 Christian Henning

The tail is where data runs out

- A year of data already shows the Gaussian is wrong at 1 in 100
- For anything rarer than once a year, there's no data: the reserve rests on the shape we assume
- The data can't show us that shape. As modellers, we have to build it into the model



© 2026 Christian Henning

Make uncertainty ~~faithful~~ useful

Humans misjudge probabilities,
so **does uncertainty need to be faithful?**

It takes real modelling work,
so **ask first what you gain from it**

© 2026 Christian Henning · Tversky & Kahneman (1974), Judgment under uncertainty: heuristics and biases

Thank you!

Questions?

You can find more on these topics on **chrhenning.com**



© 2026 Christian Henning

Backup: why the τ-quantile?

- Add one more unit of reserve
- With probability $1 - F(r)$ the shortfall exceeds the reserve: the unit saves c_{short}
- With probability $F(r)$ it doesn't: the unit sits idle and costs c_{idle}
- Keep adding reserve while the expected saving beats the expected cost:

$$c_{\text{short}} \big(1 - F(r)\big) > c_{\text{idle}}\, F(r)$$

- $F(r)$ grows with r, so stop where both are equal:

$$F(r^*) = \frac{c_{\text{short}}}{c_{\text{short}} + c_{\text{idle}}} = \tau$$

F is the CDF of the predicted shortfall, so r^* is its τ-quantile

© 2026 Christian Henning