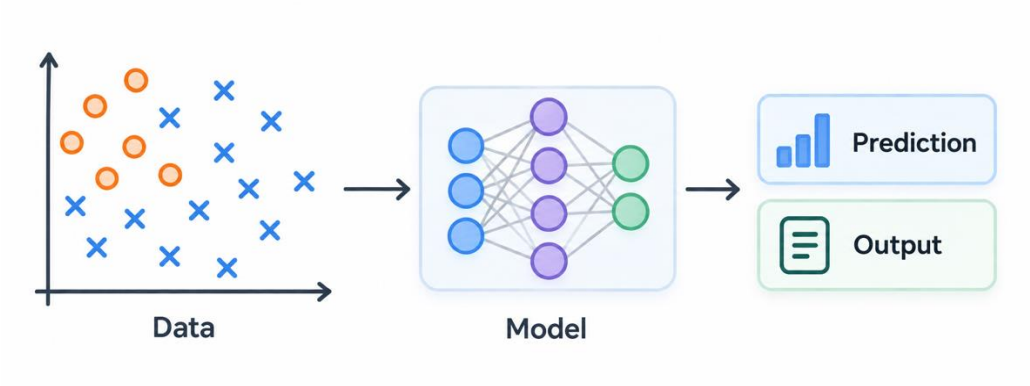
Lecture 1: Introduction

Krikamol Muandet

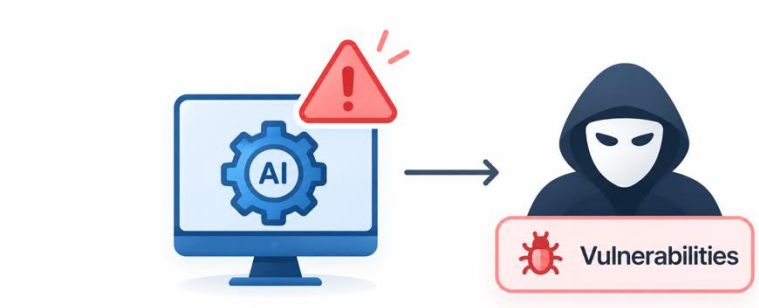
16 October 2026



Outline



- **Machine Learning**
  - Supervised learning
  - Unsupervised learning
  - Generative modeling
  - Reinforcement learning



- **Pitfalls of Machine Learning**
  - Overconfident predictions
  - OOD generalisation
  - Algorithmic bias and unfairness
  - Adversarial vulnerability
  - Reliability and trustworthiness
  - AI safety and misalignment



Machine Learning

The Achilles of modern-day computation

Machine Learning

- A **hypothesis space** $\mathcal{H}$, a data set $\mathcal{D}$ of finite **observations**, and a **loss function** ℓ.
- A **statistical learning** aims to find the "best" hypothesis in $\mathcal{H}$ based on the observations in $\mathcal{D}$:

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \ell_h(\mathcal{D})$$

- How does $\hat{h}$ compare with the **optimal hypothesis** h^* as the number of observations grows?

Machine Learning

- An **input space** $\mathcal{X}$ and **target space** $\mathcal{Y}$
- Consider a finite hypothesis class $\mathcal{H}_{\mathrm{m}} = \{h_1, h_2, \ldots, h_m\}, h_i : \mathcal{X} \to \mathcal{Y}$
- A data set $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}, (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$
- A real-valued loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to R_+$

$$\hat{h} \in \arg\min_{h \in \mathcal{H}_{\mathrm{m}}} \widehat{\boldsymbol{R}}(h) =: \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, h(x_i)\big)$$

$$h^* = \arg\min_{h \in \mathcal{H}_{\mathrm{m}}} \boldsymbol{R}(h) =: \int \ell\big(Y, h(X)\big)\, dP(X, Y)$$

Learning here amounts to searching over the hypothesis class and selecting the hypothesis with the lowest average loss.

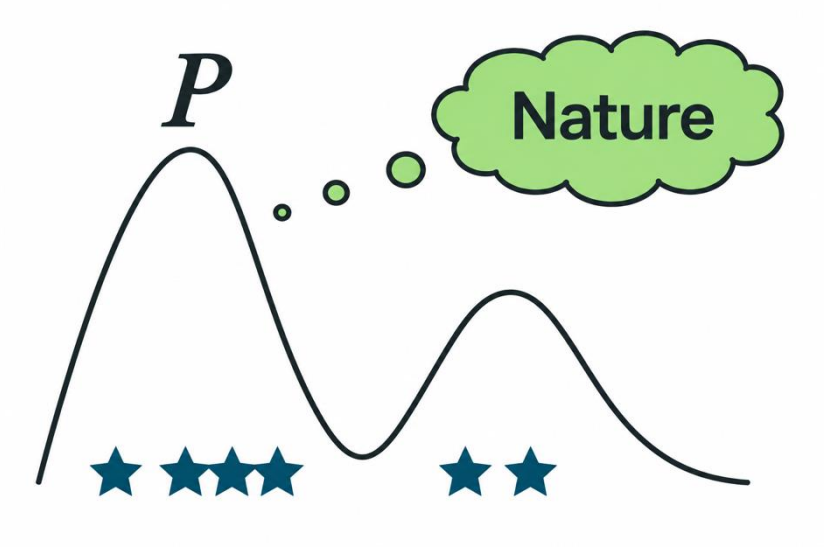
Machine Learning

Hypothesis Space



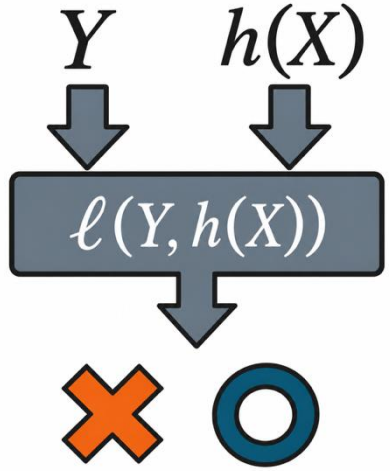
A **feasible** hypothesis space $\mathcal{H}$ is a subset of all **conceivable** hypotheses $\mathcal{F}$

Data



Data are drawn **IID** from some unknown but fixed distribution $(X, Y) \sim P(X, Y)$

Loss Function



The loss function ℓ is **task-specific** and **quantifies the correctness** of the prediction

Probably Approximately Correct (PAC)

- Let $c: \mathcal{X} \to \{0,1\}$ be **target concept** to learn and $\mathcal{C}$ a **concept class** comprising target concepts c.
- A **learning algorithm** $\mathcal{A}$ receives sample S and selects a hypothesis h_S from $\mathcal{H}$ approximating the target c.

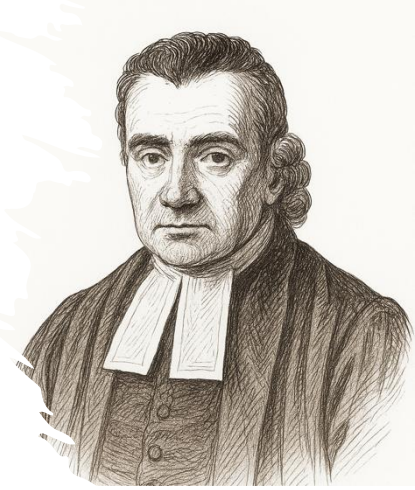
A hypothesis class $\mathcal{H}$ is **PAC-learnable** if there exists a learning algorithm $\mathcal{A}$ such that for all concepts $c \in \mathcal{C}, \epsilon > 0, \delta > 0$, and all distributions P,

$$\Pr_{S \sim P^{\mathrm{m}}}[R(h_S) \leq \epsilon] \geq 1 - \delta$$

for sample S of size $\mathrm{m} = \mathrm{poly}(1/\epsilon, 1/\delta)$ for a fixed polynomial.

Valiant (1984)

Learning as Belief Update

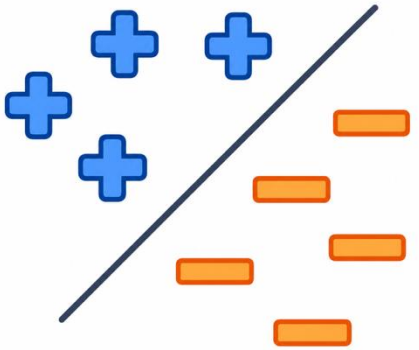


- Let Θ be a **parameter** space and $\mathcal{D}$ an observed **evidence**
  - $P(\theta)$ – A **prior** probability over the parameter space Θ
  - $P(\mathcal{D}|\ \theta)$– A **likelihood** function of $\mathcal{D}$ given θ
  - $P(\theta\ |\mathcal{D})$ – A **posterior** probability of θ given $\mathcal{D}$
- The **Bayes rule** relates *posterior* to *likelihood* and *prior*:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D}|\ \theta)P(\theta)$$

- The most popular update rule in the precise setting, known for its many desirable properties, though not the only possible choice in the imprecise setting.

Classification Problems



- $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)\},\ \mathcal{X} \subset R^d$ and $\mathcal{Y} = \{+1, -1\}$
- A **0-1** loss function $\ell(y, \hat{y}) = \mathbb{1}_{\{y \neq \hat{y}\}}$ (misclassification loss)

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i \neq h(\boldsymbol{x}_i)\}}, \quad h_{\boldsymbol{w},\boldsymbol{b}}(\boldsymbol{x}) = \begin{cases} +1, \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b > 0 \\ -1, \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b \leq 0 \end{cases}$$

- For K classes, a **cross-entropy** loss $\ell(y, \hat{y}) = -\sum_{k=1}^{K} y^{(k)} \log \hat{y}^{(k)}$ where y is a one-hot encoding and $\hat{y}$ is a probability score:

$$y = (0, 0, \mathbf{1}, 0), \qquad \hat{y} = (0.02, 0.1, \mathbf{0.7}, 0.18)$$

Kullback-Leibler (KL) Divergence

- The **Shannon entropy** $H(P) = -E_{x\sim P}[\log P(x)]$.
- For two probability distribution $P(x)$ and $Q(x)$:

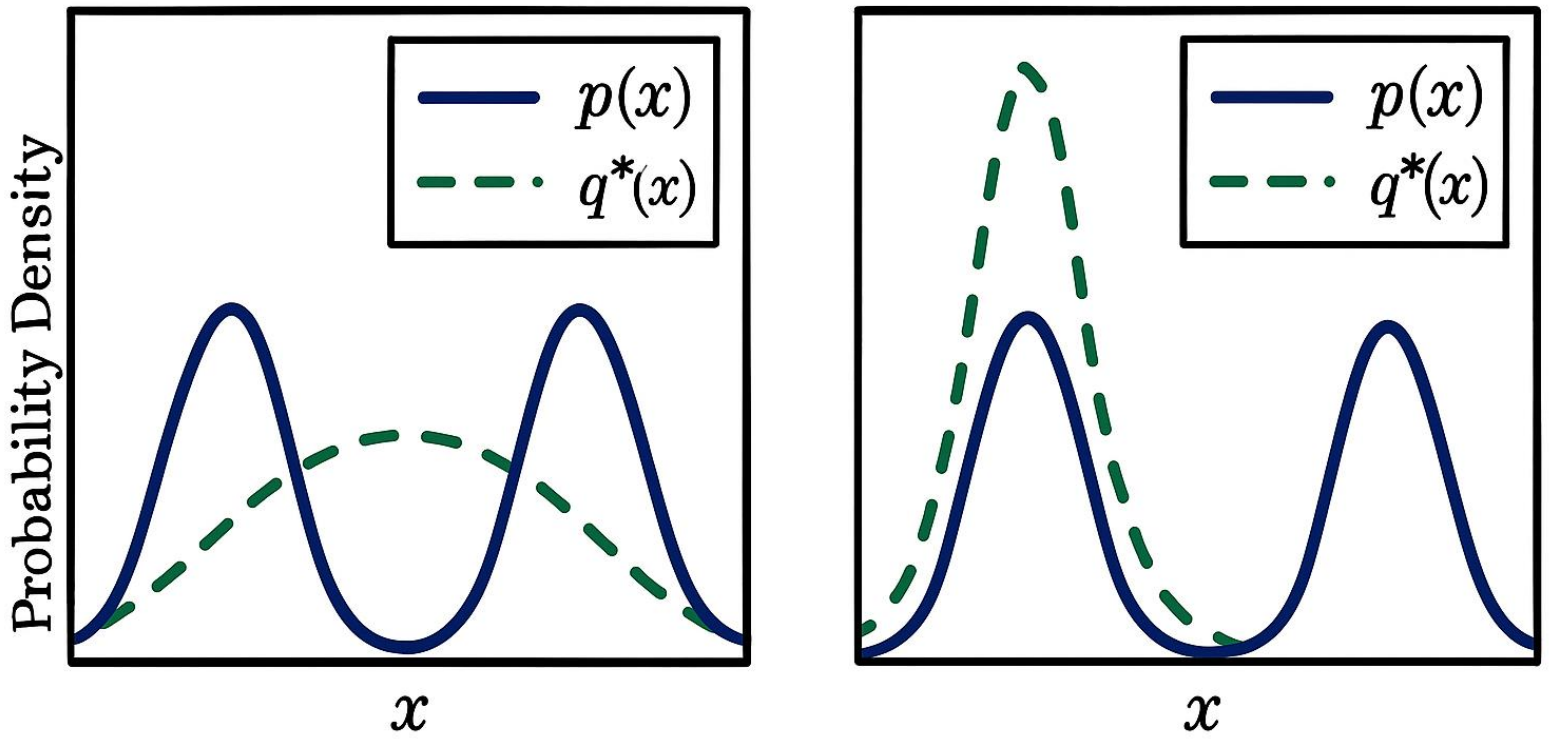
$$D_{KL}(P \,||\, Q) = E_{x\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = E_{x\sim P}[\log P(x) - \log Q(x)]$$

- **Important properties**: (1) $D_{KL}(P \,||\, Q) \geq 0$ (2) $D_{KL}(P \,||\, P) = 0$ (3) $D_{KL}(P \,||\, Q) \neq D_{KL}(Q \,||\, P)$
- The cross-entropy can be expressed as

$$H(P, Q) = H(P) + D_{KL}(P \,||\, Q) = -E_{x\sim P}[\log Q(x)]$$

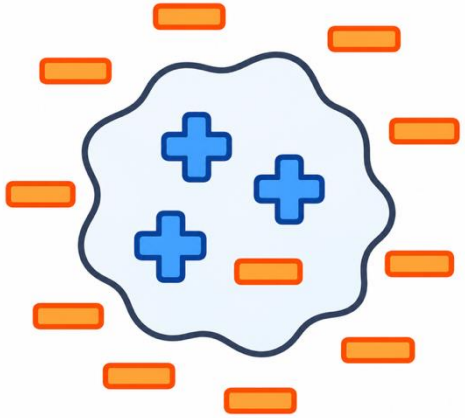
Kullback-Leibler (KL) Divergence

$$q^* = \arg\min_q D_{KL}(p \,||\, q) \qquad q^* = \arg\min_q D_{KL}(q \,||\, p)$$



mode-covering vs. **mode-seeking**

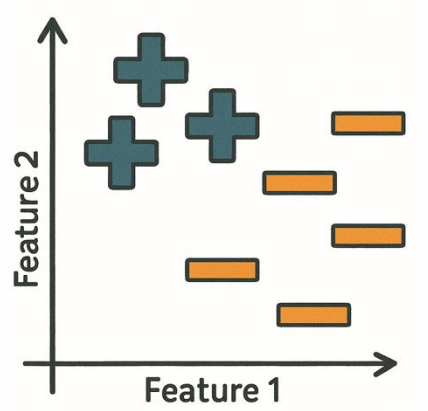
Classification Problems



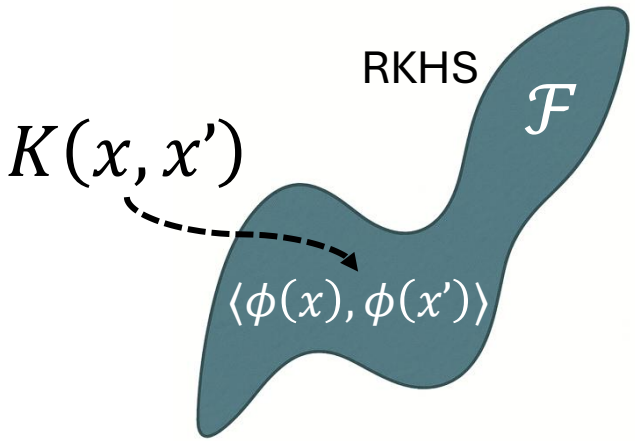
- $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}, \mathcal{X} \subset R^d$ and $\mathcal{Y} = \{+1, -1\}$
- A hypothesis class

$$\mathcal{H} = \{h(x) : h(x) = \langle w, \phi(x) \rangle_{\mathcal{F}} + b\}$$

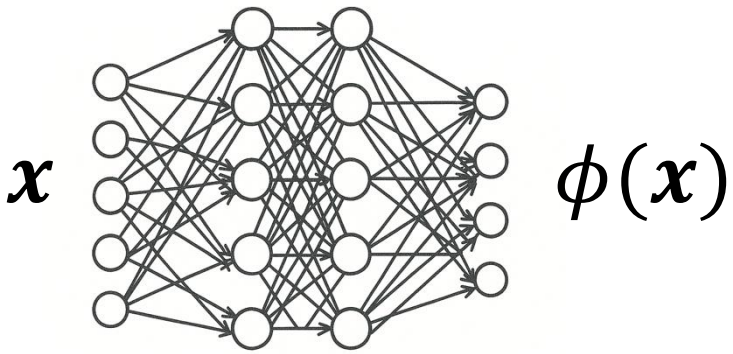
for some **non-linear feature map** ϕ: $\mathcal{X} \to \mathcal{F}$.



Hand-crafted Features



Implicit Features



Explicit Features

Regression Problems

- $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\},\ \mathcal{X} \subset R^d, \mathcal{Y} \subset R$
- A **square loss** $\ell(y, \hat{y}) = (y - \hat{y})^2$
- A hypothesis class

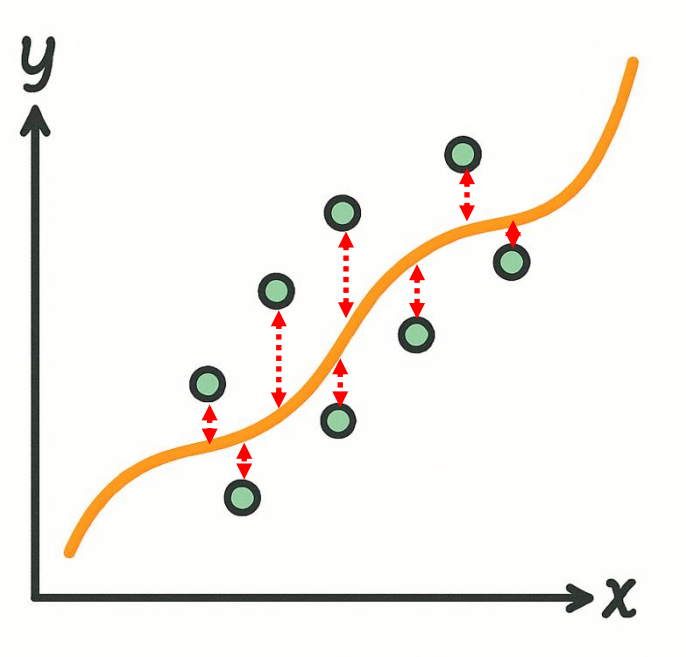
$$\mathcal{H} = \{h(x) : h(x) = \langle w, \phi(x) \rangle_{\mathcal{F}} + b\}$$

- **Least Square (LS)**: $h_{\boldsymbol{\beta}}(\boldsymbol{x}) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}$

$$\widehat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

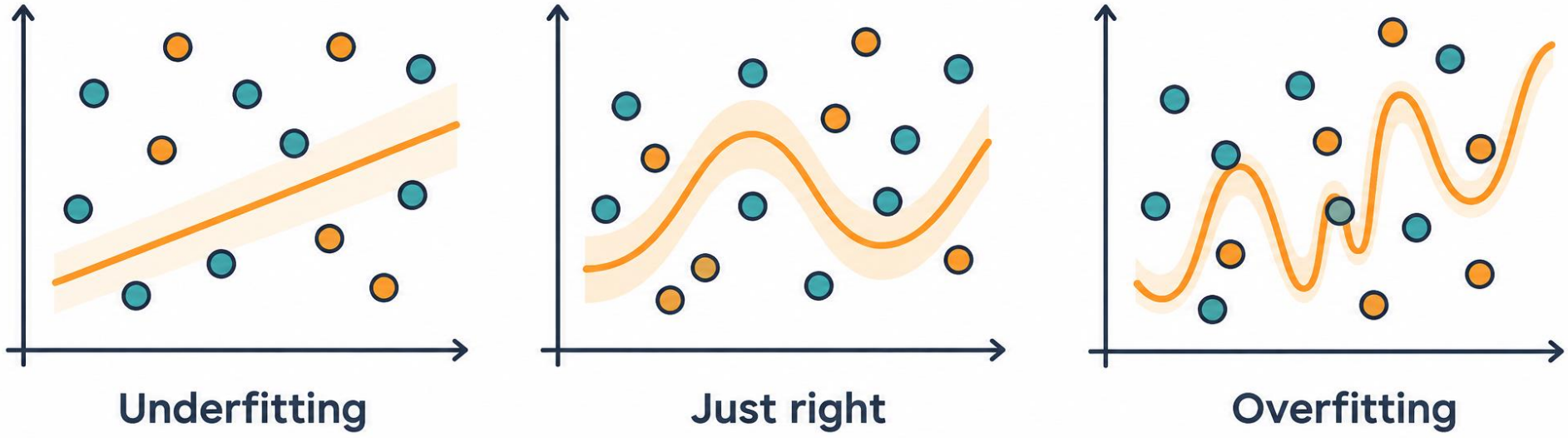
where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}, \boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \sum_{\mathrm{i}=1}^{\mathrm{n}} \left| y_i - \sum_{\mathrm{j}=1}^{\mathrm{d}} \boldsymbol{x}_{ij}\beta_j \right|^2$$



Underfitting vs Overfitting

$$\hat{h}_\lambda \in \arg\min_{h \in \mathcal{H}} \widehat{\boldsymbol{R}}_\lambda(h) =: \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, h(x_i)\big) + \underbrace{\boldsymbol{\lambda \Omega}(||\boldsymbol{h}||^2)}_{\textbf{Regularisation}}$$



over-regularisation

optimal regularisation

under-regularisation

Unsupervised Learning

- In unsupervised setting, we observe a set of **unlabeled** data:

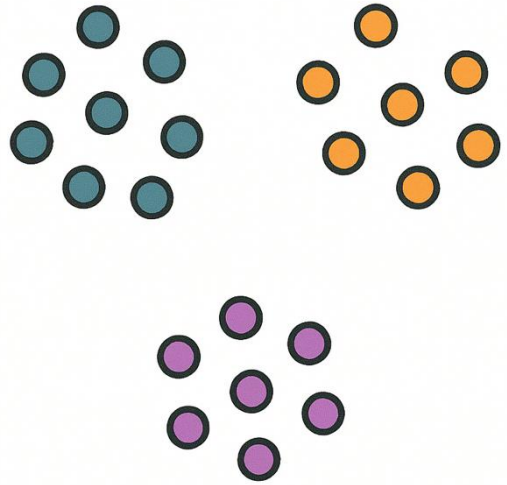
$$\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\},\ \mathcal{X} \subset R^d, \qquad x_i \sim P_{\theta_0}(X),\ \theta_0 \in \Theta$$

  where Θ is a **parameter space**.

- The goal is to infer about the **true parameter** θ_0.
- Maximum likelihood estimation (MLE):

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}_n(\theta; \boldsymbol{x})$$

  where $\mathcal{L}_n(\theta; \boldsymbol{x})$ is the **likelihood function**.



Unsupervised Learning

- For a parametric family $\{\boldsymbol{f}(\cdot\,;\theta) \mid \theta \in \Theta\}$ and independent RV:

$$\mathcal{L}_{\mathcal{n}}(\theta;\boldsymbol{x}) = \boldsymbol{f}_{\mathrm{n}}(\boldsymbol{x};\theta) = \prod_{\mathrm{k}=1}^{\mathrm{n}} f_k(x_k;\theta)$$

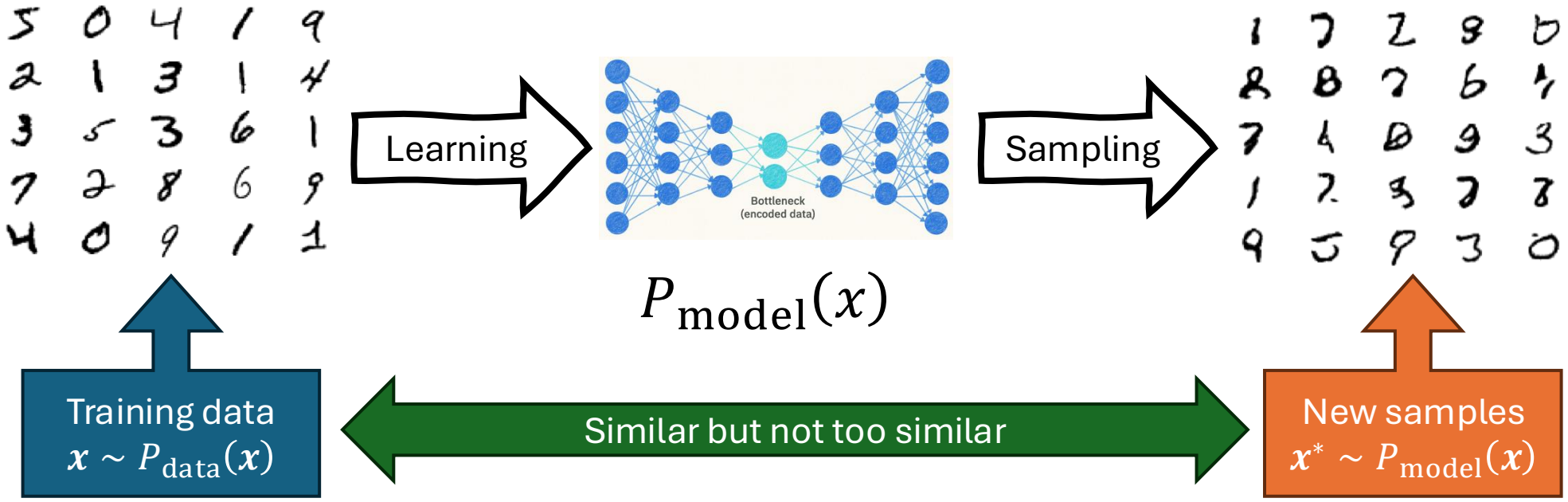
- **Example**: For a family of Gaussian distributions:

$$\{\boldsymbol{\mathcal{N}}(\cdot\,;\,\mu,1) \mid \mu \in R\}, \qquad \boldsymbol{\mathcal{N}}(x\,;\,\mu,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$$

Find the MLE of μ from the observations $\boldsymbol{x} = \{x_1, x_2, \ldots, x_{\mathrm{n}}\}$

Generative Modeling

- Given training data $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, generate **new samples**.

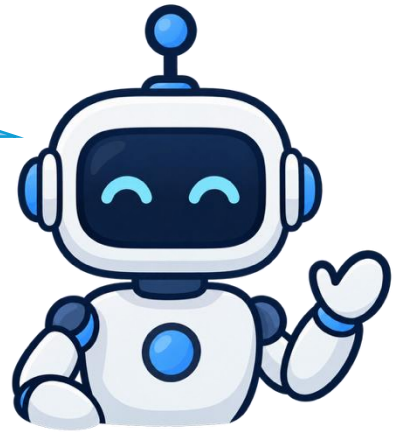


1. **Explicit**: Autoregressive models, variational autoencoders (VAEs)
2. **Implicit**: Generative adversarial networks (GANs) and diffusion models

Reinforcement Learning

What should I do?

$a_1, a_2, \dots, a_n$



Action a_t

State s_{t+1}
Reward r_{t+1}



What will happen?

$$(s_t, a_t) \begin{cases} s_{t+1}^{(1)}, r_{t+1}^{(1)} \\ s_{t+1}^{(2)}, r_{t+1}^{(2)} \\ s_{t+1}^{(3)}, r_{t+1}^{(3)} \end{cases}$$

Uncertain knowledge / policy

$$a_t \sim \pi(\cdot|\, s_t), \qquad \pi \in \prod$$

Uncertain outcomes / dynamic

$$(s_{t+1}, r_{t+1}) \sim P(\cdot,\cdot|\, s_1, a_t)$$

Learning here amounts to learning how to **act through interaction** with the environment to maximize cumulative reward.

Various Sources of Uncertainty

Data D

- Sampling uncertainty
- Label noise / annotation error
- Measurement noise
- Missing data / censoring
- Latent confounding
- Class imbalance
- Rare events

Aleatoric (irreducible)

Models M

- Model misspecification
- Parameter uncertainty
- Hyperparameter uncertainty
- Approximation error
- Training stochasticity
- Overfitting / underfitting
- Representation uncertainty

Epistemic (reducible)

Environments E

- Distribution shift
- Concept drifts
- Intervention / policy shift
- Adversarial perturbations
- Hardware / system noise
- Social / contextual uncertainty
- Task redefinition

Out-of-distribution / Structural

Multifaceted Nature of Uncertainty

Aleatoric Uncertainty – the irreducible part



The coin's bias might be known, but every coin flip could still yield a different outcome.



The reading of the patient's sample may vary each time the doctor collects it.



There's always a chance of rain tomorrow despite a weather forecast.

Epistemic Uncertainty – the reducible part



The uncertainty about the true bias of a coin can be reduced through experimentation.



If the training set contains only images of "a cow on a field" and "a camel in a desert," the model will be epistemically uncertain about unseen images of "a cow in a desert" and "a camel on a field."

Multifaceted Nature of Uncertainty

Risk and **uncertainty** **are different.**



Risk *is a quantifiable form of uncertainty. For example, the chance of rain tomorrow can be estimated from historical data.*

Uncertainty *is an immeasurable form of risk. For example, when moving to an entirely new country, the outcome is so unpredictable that it's impossible to quantify the risk of what might happen.*



Risk

Measurable uncertainty

Uncertainty

Unmeasurable risk

Known outcomes and known probabilities

Known or unknown outcomes and unknown probabilities

Unknown outcomes and unknown probabilities

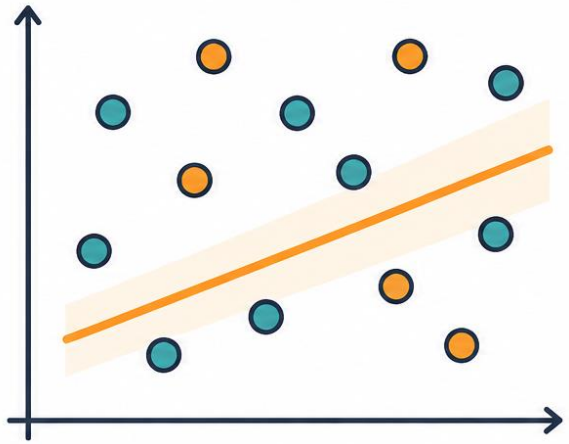


Pitfalls of Machine Learning

The Achilles' heel

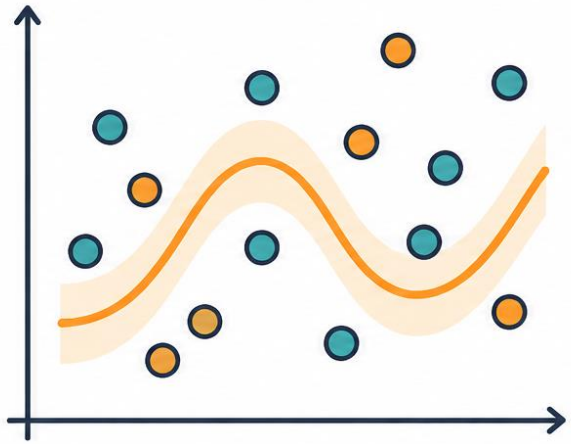
Underfitting and Overfitting

$$\hat{h}_\lambda \in \arg\min_{h \in \mathcal{H}} \widehat{\boldsymbol{R}}_\lambda(h) =: \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, h(x_i)\big) + \underbrace{\boldsymbol{\lambda \Omega}(||\boldsymbol{h}||^2)}_{\textbf{Regularisation}}$$



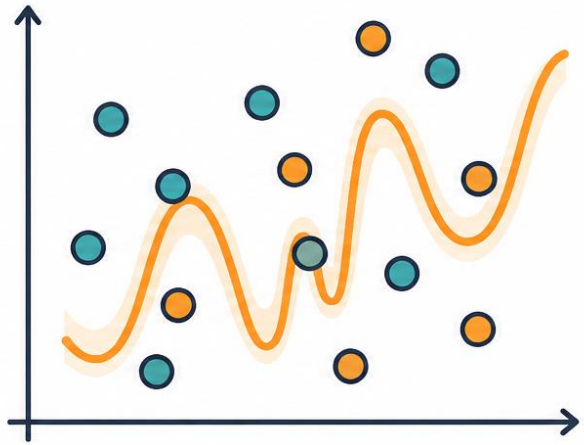
Underfitting

over-regularisation



Just right

optimal regularisation



Overfitting

under-regularisation

Overconfident Predictions

Consider a binary classification problem (cat vs dog) using SOTA deep neural network (e.g., transformer)



$P(cat \mid x) = 0.9$
$P(dog \mid x) = 0.1$

$P(cat \mid x) = 0.05$
$P(dog \mid x) = 0.95$



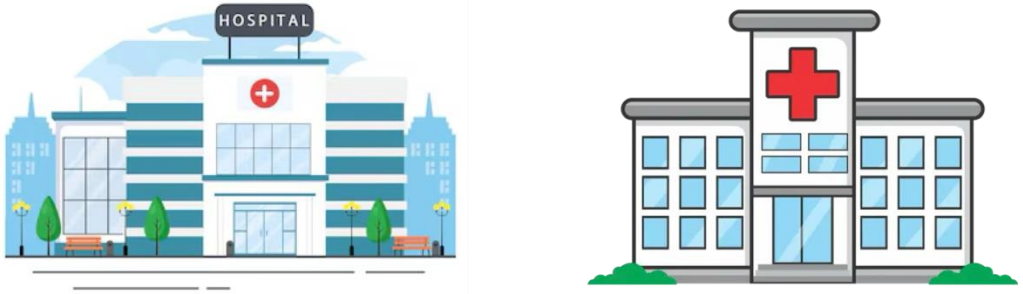
$P(cat \mid x) = 0.1$
$P(dog \mid x) = ?$

False Confidence:

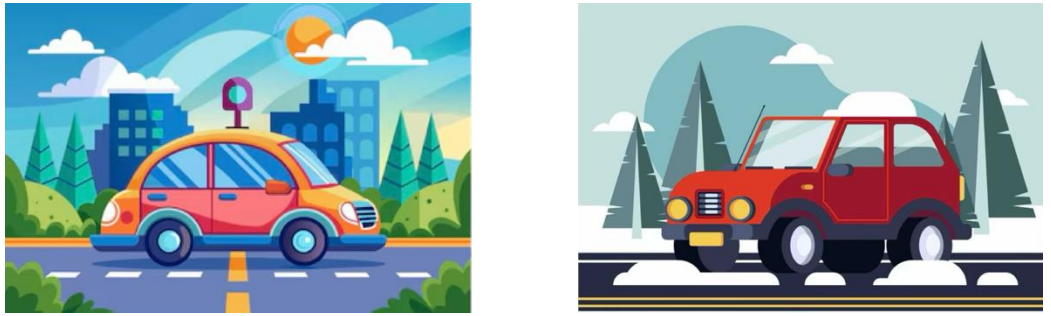
The model is overly confident that I'm a dog simply because the probability of me being a cat is low.

Out-of-Distribution Generalisation

- Under the **IID assumption**, training and test data are drawn from the same distribution, i.e., $P_{\text{train}} = P_{\text{test}}$.
- Highly unrealistic in practical applications i.e., $P_{\text{train}} \neq P_{\text{test}}$.



A medical model trained on data from large urban hospitals (where data are abundant) may fail catastrophically when deployed in rural hospitals (where data are scarce) due to distribution shift.



An advanced autonomous car trained primarily in urban areas near the company's headquarters may perform unreliably or unsafely when deployed in unfamiliar environments.

- Covariate shift, concept shift, label shift, etc.

Singh et al. (ICML 2023)

Algorithmic Bias and Unfairness

*Algorithmic bias reflects a model's failure to recognize the broader societal context—an instance of an **unknown unknown**.*

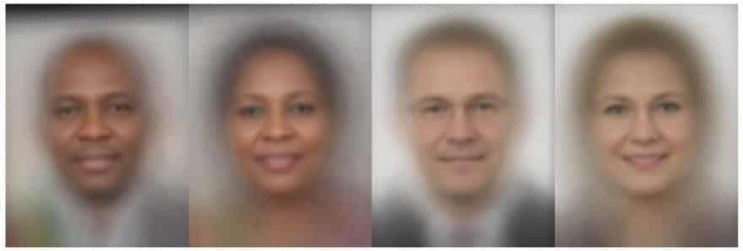


Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Raji et al. (2020). **Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing**. In AIES '20.

https://doi.org/10.1145/3375627.3375820

Algorithmic Bias and Unfairness



*Algorithmic bias reflects a model's failure to recognize the broader societal context—an instance of an **unknown unknown**.*



Amazon reportedly scraps AI recruiting tool that was biased against women

Reuter, MIT TR, The Verge, BBC



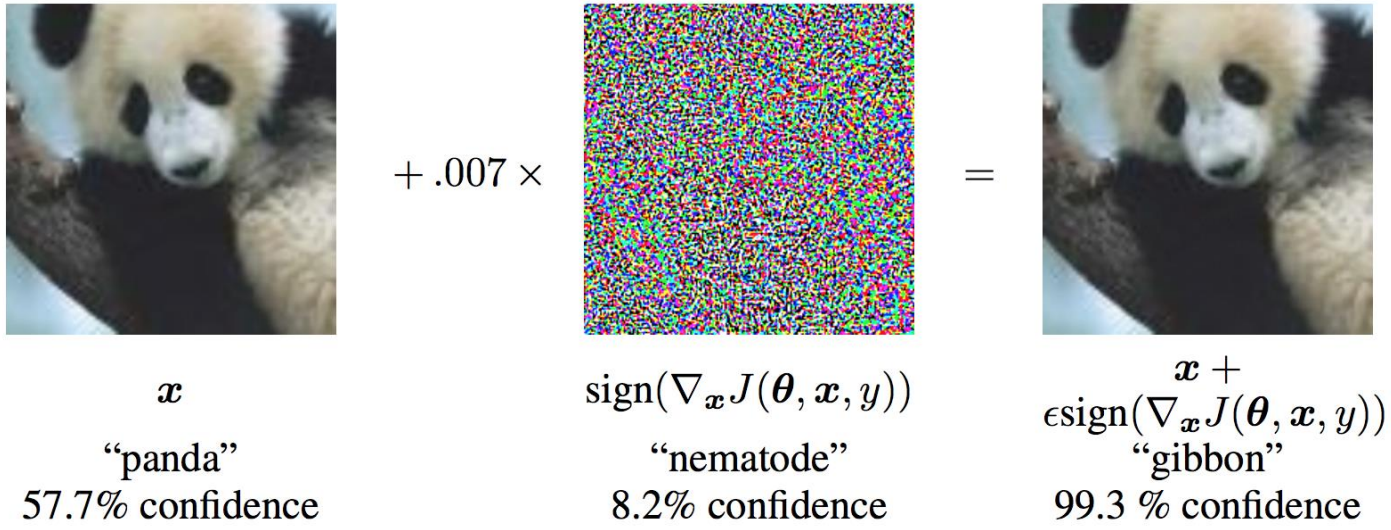
Stable Diffusion Exhibited Biases for Prompts Featuring Professions

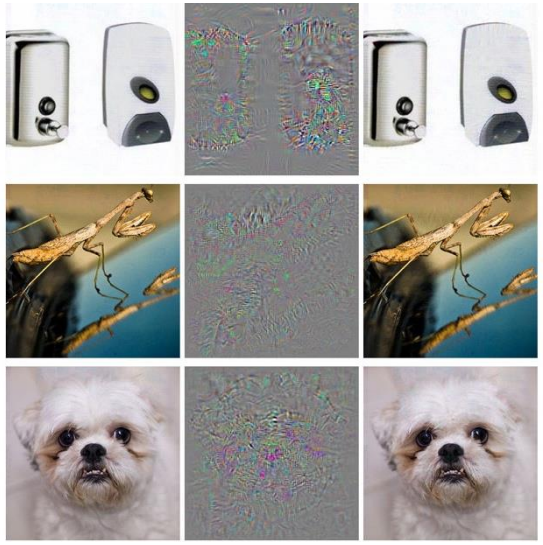
AI Incident Database, Wu et al. (2023), Luccioni et al. (2023)

Adversarial Vulnerability



Adversarial vulnerability exposes how current ML models conflate ***precision*** *with* ***confidence****. A high-confidence decision boundary that fits the data well often extends that confidence into regions lacking epistemic support.*





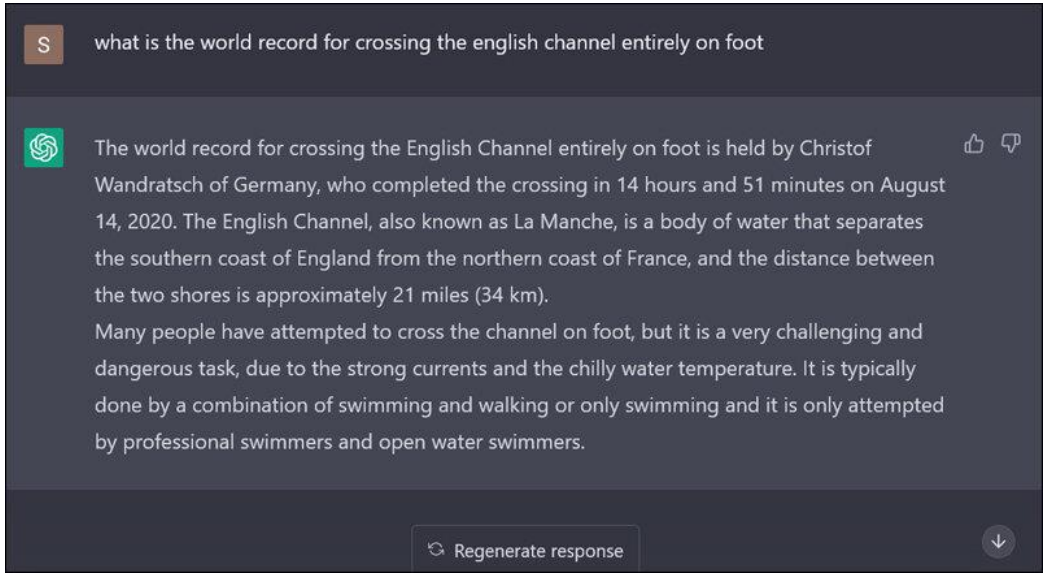
Goodfellow et al. (2014): **Explaining and Harnessing Adversarial Examples (**🔗 https://arxiv.org/abs/1412.6572)
Szegedy et al. (2013): **Intriguing properties of neural networks (**🔗 https://arxiv.org/abs/1312.6199)

Reliability and Trustworthiness

Hallucination is an epistemic overconfidence problem — the system behaves as if uncertainty were resolved when it is not. These are not cases of noisy data (aleatoric), but rather of the model over-committing where its epistemic uncertainty should be high.



Deepfake



Hallucination

AI Misalignment

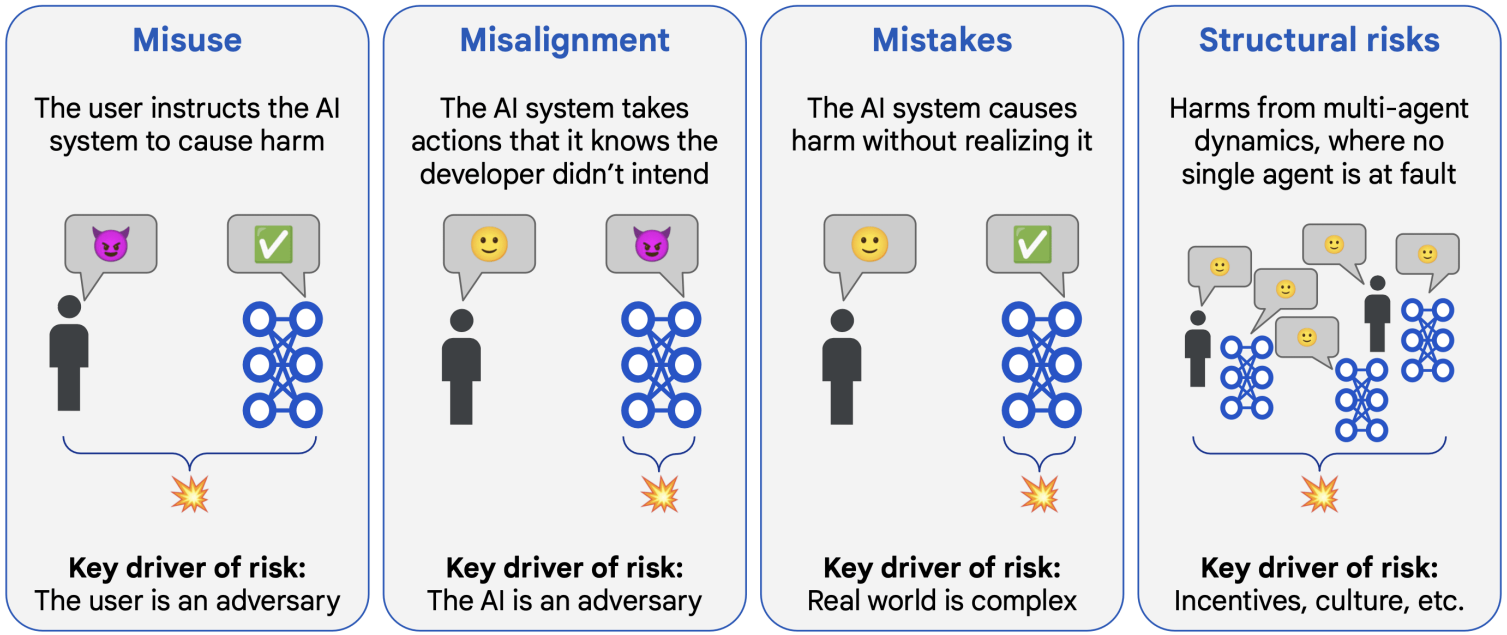


Figure 1 | **Overview of risk areas.** We group risks based on factors that drive differences in mitigation approaches. For example, misuse and misalignment differ based on which actor has bad intent, because mitigations to handle bad human actors vary significantly from mitigations to handle bad AI actors.

AI Alignment Forum: www.alignmentforum.org

AI Safety



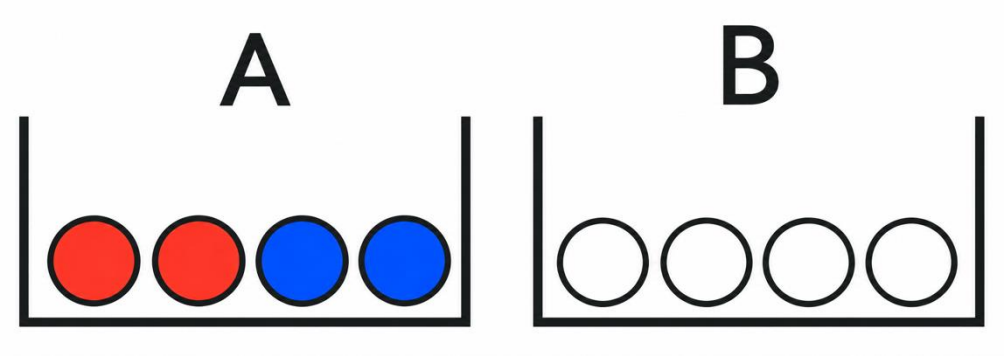


"There are ***known knowns***. There are ***known unknowns***. But there are also ***unknown unknowns***—things we do not yet realize we do not know."—**Donald Rumsfeld (2002)**

Ellsberg Paradox

The Pitfall of Additive Probability

Ellsberg Paradox



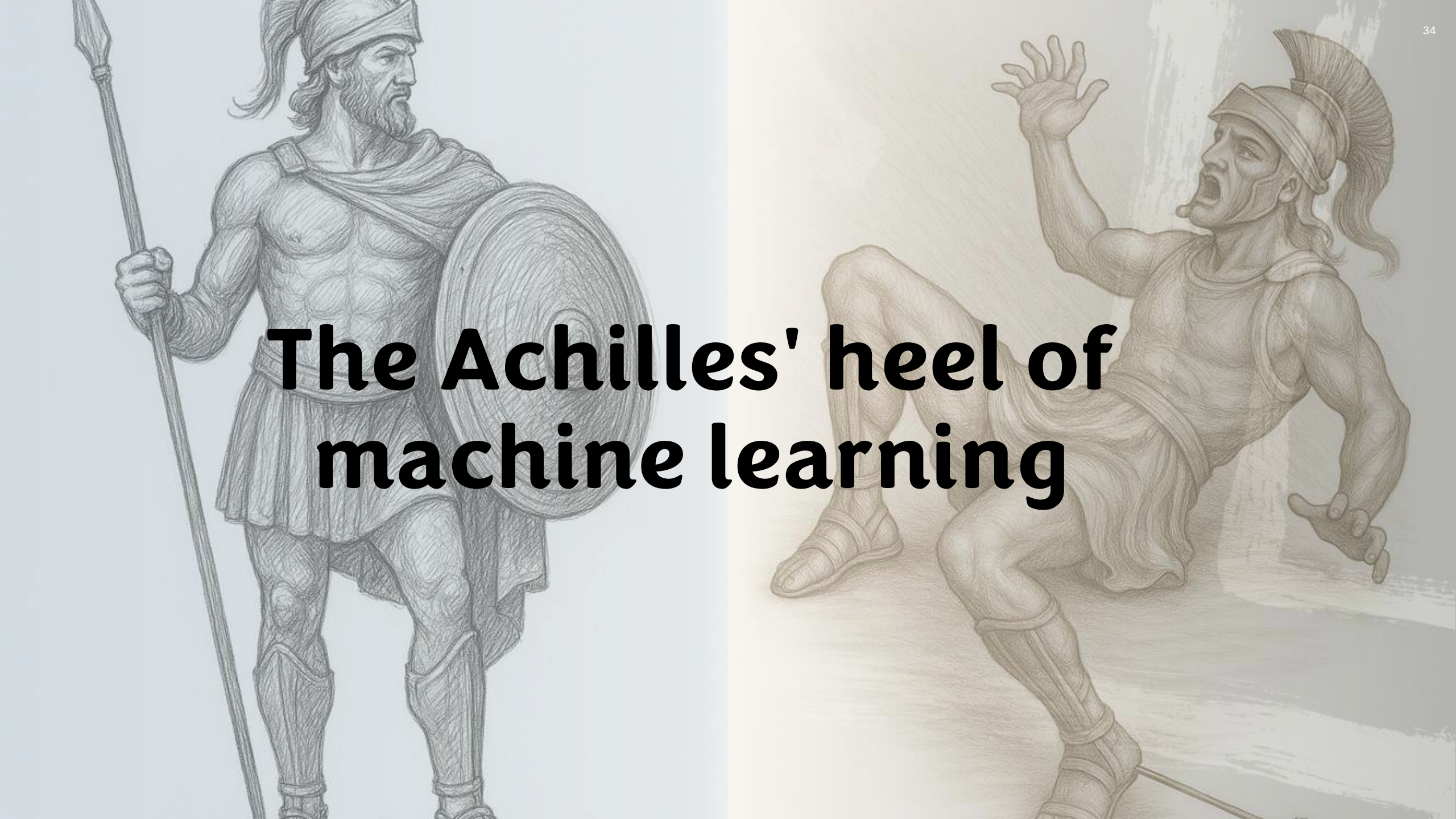
A person is shown two urns, A and B. In urn A, there are 50 red balls and 50 blue balls. There are red and blue balls in urn B with unknown proportion.

One ball is drawn at random from each urn:

1. Bet on **Ar** or **Ab** (indifferent)
2. Bet on **Br** or **Bb** (indifferent)
3. Bet on **Ar** or **Br** **(Ar** $\succ$ **Br** $\Longrightarrow p_{\mathrm{Br}} < p_{\mathrm{Ar}}$**)**
4. Bet on **Ab** or **Bb** **(Ab** $\succ$ **Bb** $\Longrightarrow p_{\mathrm{Bb}} < p_{\mathrm{Ab}} \Leftrightarrow 1 - p_{\mathrm{Br}} < 1 - p_{\mathrm{Ar}} \Leftrightarrow p_{\mathrm{Br}} > p_{\mathrm{Ar}}$**)**

Contradiction!

Daniel Ellsberg, **Risk, Ambiguity, and the Savage Axioms**, *The Quarterly Journal of Economics*, Volume 75, Issue 4, November 1961, Pages 643–669, https://doi.org/10.2307/1884324



The Achilles' heel of machine learning