# **Imprecise Generalisation:** From Invariance to Heterogeneity

Krikamol Muandet

Rational Intelligence Lab — CISPA Helmholtz Center for Information Security

Saarbrücken, Germany

The Inaugural Workshop on Frontiers in Statistical Machine Learning (FSML) — 2 August 2025

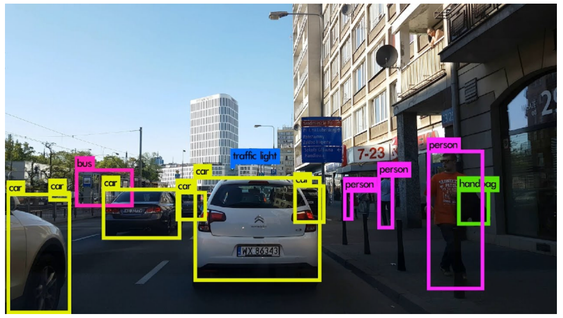
# Generalisation

2, 4, 6, 8, 10, 12, 14, 16, 18, 20, ... → $2x$ → $y = 22$

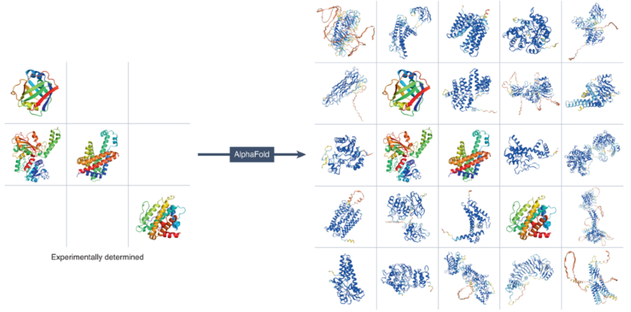
3, 4, 4, 10, 9, 14, 13, 17, 16, 22, ... → ? → $y = ?$

Binge ... on | - | and | of | is
Binge drinking ... is | and | had | in | was
Binge drinking may ... be | also | have | not | increase
Binge drinking may not ... be | have | cause | always | help
Binge drinking may not necessarily ... be | lead | cause | results | have
Binge drinking may not necessarily kill ... you | the | a | people | your
Binge drinking may not necessarily kill or ... even | injure | kill | cause | prevent
Binge drinking may not necessarily kill or even ... kill | prevent | cause | reduce | injure
Binge drinking may not necessarily kill or even damage ... your | the | a | you | someone
Binge drinking may not necessarily kill or even damage brain ... cells | functions | tissue | neurons
Binge drinking may not necessarily kill or even damage brain cells, ... some | it | the | is | long

Cevoli et al. (2022)




Sagarkar et al. (2020)




Jumper et al. 2021

# Empirical Risk Minimisation (ERM)

- Observe sample of size $n$ from some **unknown but fixed** probability distribution $P(X, Y)$

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \overset{IID}{\sim} P(X, Y), \quad (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$$

- Find the best hypothesis $h^*$ from a hypothesis space $H$ of functions $h : \mathcal{X} \to \mathcal{Y}$

- **Popular recipe**: Minimise an empirical error on the observed data:

$$\hat{h} = \arg\min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(X_i)), \qquad h^* = \arg\min_{h} \mathbb{E}_{\cdot} [\ell(Y, h(X))]$$

**Data Uncertainty**

- $R(\hat{h}) - R(h^*) < B\sqrt{\dfrac{2\log(2|H|) + 2\log(1/\delta)}{n}}$ with probability at least $1 - \delta$

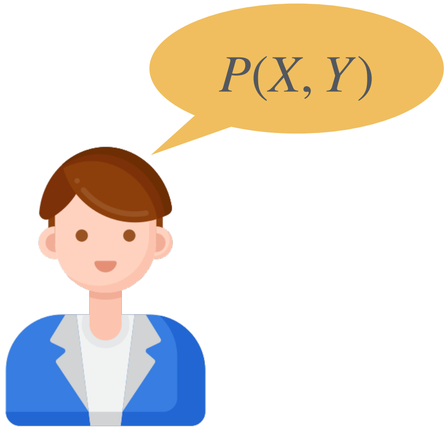
*Vapnik, Vladimir. Principles of Risk Minimization for Learning Theory, NeurIPS 1991*

# Empirical Risk Minimisation (ERM)

- The learner calibrates their **belief probability** to the **physical probability** $P(X, Y)$:

$$h^* = \arg\min_{h \in H} \mathbb{E}_{(X,Y)\sim P(X,Y)}[\ell(Y, h(X))]$$

- An **expected utility maximiser** (von Neumann and Morgenstern, 1944)
- Focuses of machine learning research:
  - Powerful hypothesis spaces (e.g., RKHS, CNN, transformer)
  - Efficient optimisation algorithms (e.g., SGD, Adam, L-BFGS)
  - Scalable data and compute (e.g., MapReduce, GPUs, TPUs)

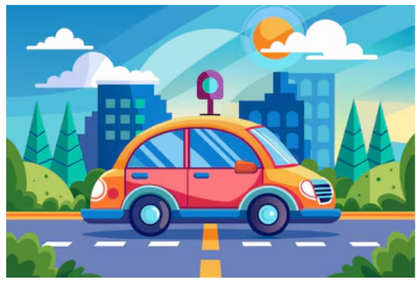



Precise Learner
(Jane)

*Neumann, John von, Oskar Morgenstern. Theory of Games and Economic Behavior. Princeton University Press, 1944.*

# Distribution Shifts

- Train and test data may not be **independent and identically distributed (IID)**


- Two sources of uncertainties:

    1. **Data uncertainty**: Jane only observes a finite data
    2. **Distribution uncertainty**: Jane is uncertain about the data distribution

- Out-of-distribution (OOD) generalisation

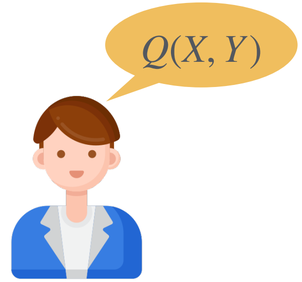
*Quionero-Candela, Sugiyama, Schwaighofer, and Lawrence. Dataset Shift in Machine Learning. The MIT Press, 2009.*

1. What is OOD generalisation?
2. How to perform statistical learning from heterogeneous sources?

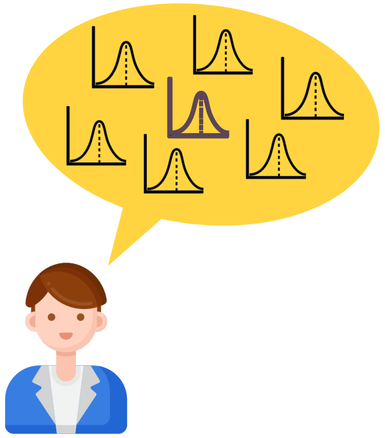
$$P_1(X, Y), P_2(X, Y), \ldots, P_n(X, Y)$$

# Precise Generalisation

- **Domain adaptation** (DA): $(X_{tr}, Y_{tr}) \overset{IID}{\sim} P(X, Y)$ and $(X_{te}, Y_{te}) \overset{IID}{\sim} Q(X, Y)$:




- **Covariate shift**: $P(X, Y) = P(Y|X)P(X)$ and $Q(X, Y) = P(Y|X)Q(X)$
  - Other scenarios: Label shift $P(Y) \neq Q(Y)$, conditional shift $P(X|Y) \neq Q(X|Y)$, concept shift $P(Y|X) \neq Q(Y|X)$, and confounding shift $P(X, Y) \neq Q(X, Y)$

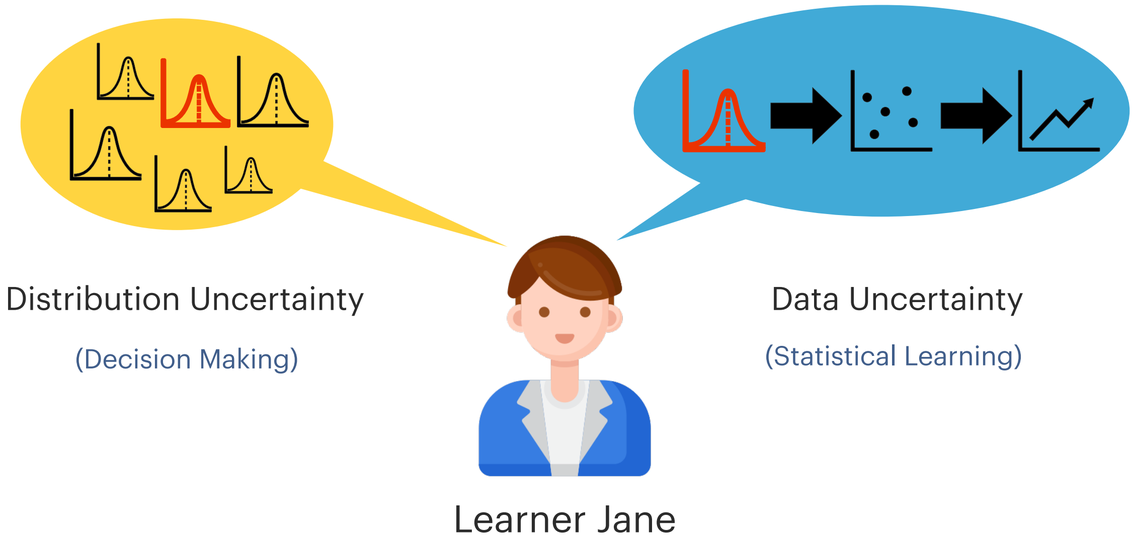


- **Domain generalisation (DG)**: $P_1(X, Y), P_2(X, Y), \ldots, P_N(X, Y) \rightarrow P_{N+1}(X, Y)$
- Domain-Adversarial Training of Neural Networks **[Ganin et al. 2016]**; Causal Invariant Prediction (CIP) **[Peters et al., 2016; Heinze-Deml et al., 2018]**; Invariant Risk Minimisation (IRM) **[Arjovsky et al., 2019]**; Distributional Robust Optimisation (DRO) **[Sagawa et al., 2020]**; Probable Domain Generalisation **[Eastwood et al., 2022]**

*Quionero-Candela, Sugiyama, Schwaighofer, and Lawrence. Dataset Shift in Machine Learning. The MIT Press, 2009; Ben-David et al. (2010)*
*Shimodaira, 2000; Ben-David et al. (2010); Sugiyama, Krauledat, Müller (2007)*
*Blanchard et al. NeurIPS (2011); Muandet et al. ICML (2013)*

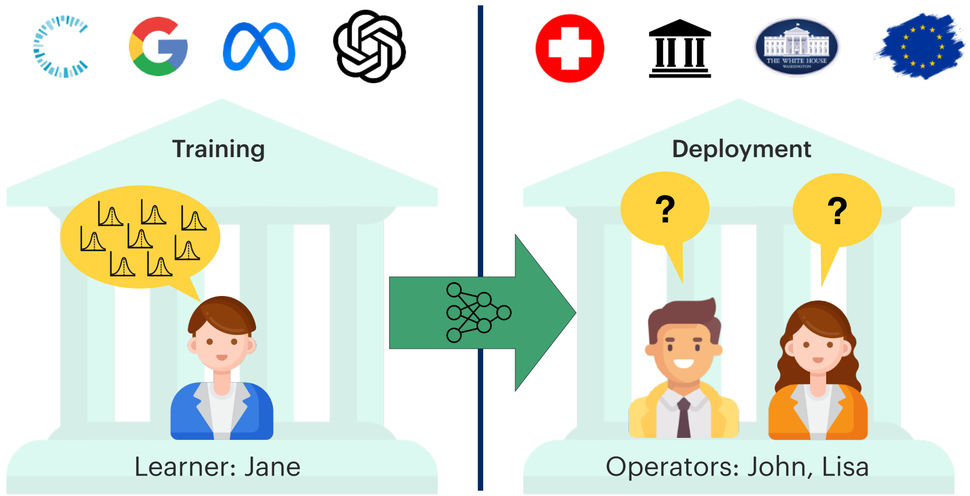
# (Im)precise Generalisation

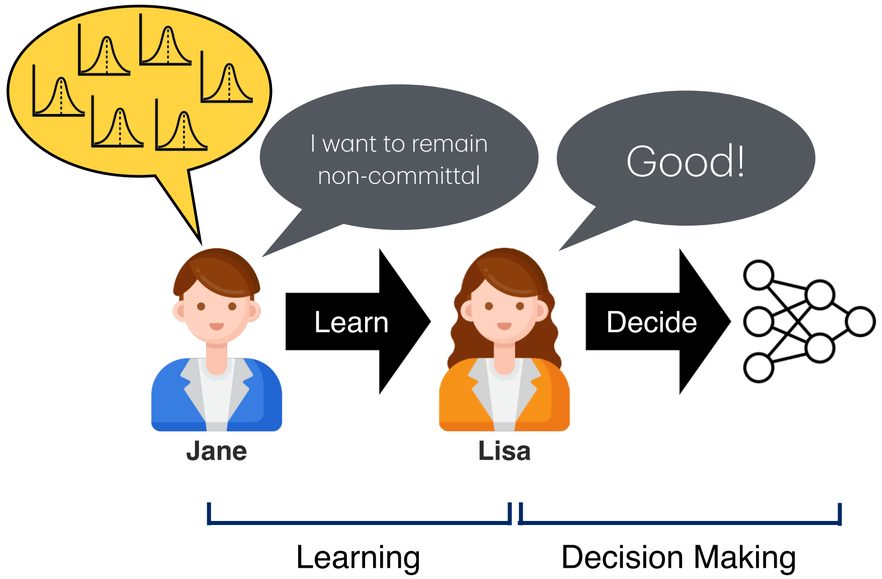
- **Precise learner, Jane,** must deal with two sources of uncertainties simultaneously.
  1. Jane **chooses the notion of generalisation** (pick a specific distribution)
  2. Jane then conducts statistical learning to **choose the best hypothesis**



# Institutional Separation

- **Precise learner** deals with two sources of uncertainties simultaneously.
  1. **The learner decides the notion of generalisation (pick a specific distribution)**
  2. The learner then conducts statistical learning to choose the best hypothesis




*Manski, Charles F. Statistical Inference for Statistical Decisions. 2019*

# Domain Generalisation via Imprecise Learning


**Anurag Singh**
CISPA

**Siu Lun Chau**
CISPA

**Shahine Bouabid**
MIT

**Krikamol Muandet**
CISPA


---



**Domain Generalisation via Imprecise Learning**


**Anurag Singh** [1] **Siu Lun Chau** [1] **Shahine Bouabid** [2] **Krikamol Muandet** [1]



## Abstract

Out-of-distribution (OOD) generalisation is challenging because it involves not only learning from empirical data, but also deciding among various notions of generalisation, e.g., optimising the average-case risk, worst-case risk, or interpolations thereof. While this choice should in prin-

(LLM) that surpass human-level generalisation capabilities in specific domains.

Despite notable achievements, these systems may catastrophically fail when operated on out-of-domain (OOD) data because theoretical guarantees for their generalisation hinge on the assumption of independent and identically distributed (IID) training and deployment data, with empirical

# Problem Formulation


- A **risk profile** on $N$ observed environments $P_1(X, Y), P_2(X, Y), \ldots, P_N(X, Y)$

$$\mathbf{R}(f) := (R_1(f), \ldots, R_N(f)), \quad f \in H$$

- An **aggregation function** $\rho_\lambda : L_2^N(H) \to L_2(H)$ for some $\lambda \in \Lambda$

- For a fixed $\lambda \in \Lambda$, we can learn from $H$ by minimising an aggregated risk

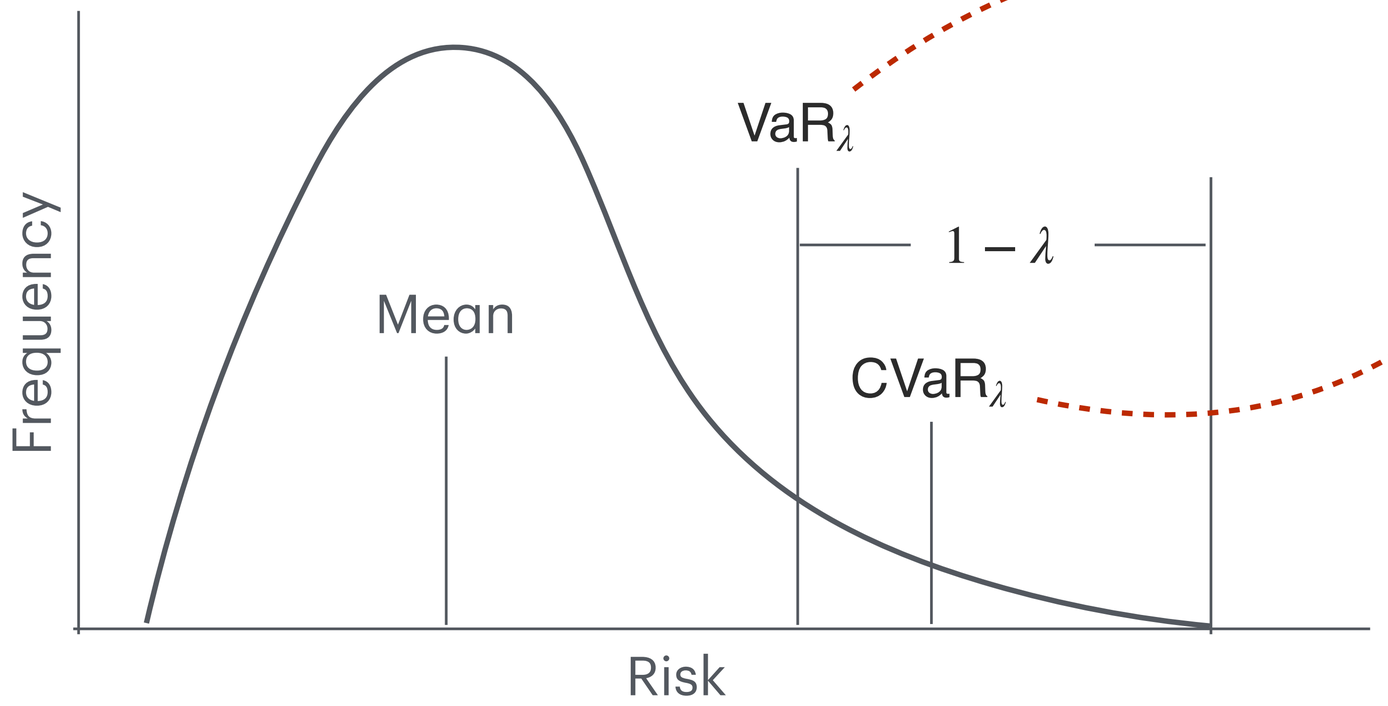
$$f_\lambda^* = \arg\min_{f \in H} \rho_\lambda[\mathbf{R}](f), \quad \lambda \in \Lambda$$

- Learn an **augmented hypothesis** $h_\theta : H \times \Lambda \to \mathcal{Y}$ such that

$$h_\theta^*(\cdot, \lambda) = f_\lambda^* = \arg\min_{f \in H} \rho_\lambda[\mathbf{R}](f), \quad \lambda \in \Lambda$$

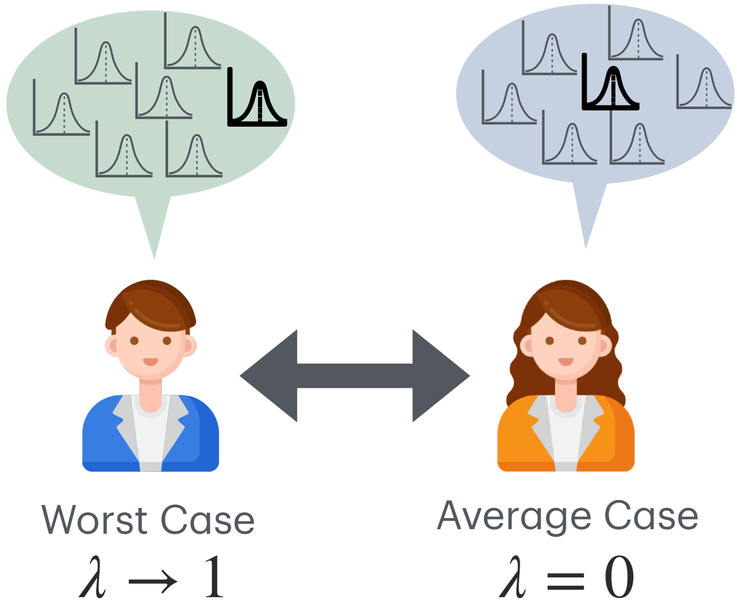
Traverse credal set

# Conditional Value at Risk (CVaR)




$$\text{VaR}_\lambda(R) = \min\{r \mid F_R(r) \geq \lambda\}$$

$$\text{CVaR}_\lambda(R) = \mathbb{E}\left[R \mid R > \text{VaR}_\lambda(R)\right]$$




- **Interpretation**: $\lambda$ is the level of risk aversion
  (Robey et al., 2022; Eastwood et al., 2022a; Li et al., 2023)

# C-Pareto Optimality

C-Pareto Optimality

$h_\theta$ dominates $h'_\theta$ if for all $\lambda \in \Lambda$,

$$\rho_\lambda[\mathbf{R}](h_\theta(\,\cdot\,,\lambda)) \leq \rho_\lambda[\mathbf{R}](h'_\theta(\,\cdot\,,\lambda))$$

Scalarised Objective

For $Q \in \Delta(\Lambda)$ with **full support**,

$$J_Q(h_\theta) := \mathbb{E}_{\lambda\sim Q}\left[\rho_\lambda[\mathbf{R}](h_\theta(\,\cdot\,,\lambda))\right]$$

- We pick $Q$ such that a parameter update makes **C-Pareto improvement**: $\theta_t \leftarrow \theta_{t-1} - \eta\,\nabla_\theta \hat{J}_{Q_t}(h_\theta)$ :

$$Q_t \in \arg\min_{Q\in\Delta(\Lambda)} \left\| \nabla_{\theta_{t-1}} \hat{J}_Q\left(h_{\theta_{t-1}}\right) \right\|_2, \quad \hat{J}_Q(h_\theta) := \frac{1}{N}\sum_{i=1}^{N} \rho_{\lambda_i}[\mathbf{R}](h_\theta(\,\cdot\,,\lambda_i))$$

- Similar to the **multiple-gradient descent algorithm (MGDA)** (Desideri, 2012).

# Theoretical Guarantee

**Proposition**: Under some technical assumptions, there exists $q \in (0,1)$ such that if

$$\hat{g} \in \arg\min_{\bar{g} \in H_\Lambda} \frac{1}{m} \sum_{i=1}^{m} \rho_{\lambda_i}[\hat{\mathbf{R}}](\bar{g}(\cdot, \lambda_i))$$

Where $\lambda_1, \ldots, \lambda_m \sim Q \in \Delta(\Lambda)$, then for any $\delta > q^m$, this holds with prob $1 - \delta$:

$$|\rho_{\lambda_{op}}[\mathbf{R}](\hat{g}(\cdot, \lambda_{op})) - \rho_{\lambda_{op}}[\mathbf{R}](h^*(\cdot, \lambda_{op}))| \leq 2M \left( \sqrt{\frac{\log(6/\eta_\delta)}{2n}} + \sqrt{\frac{\log(6/\eta_\delta)}{2m(1-q)(1-q^m)}} \right)$$

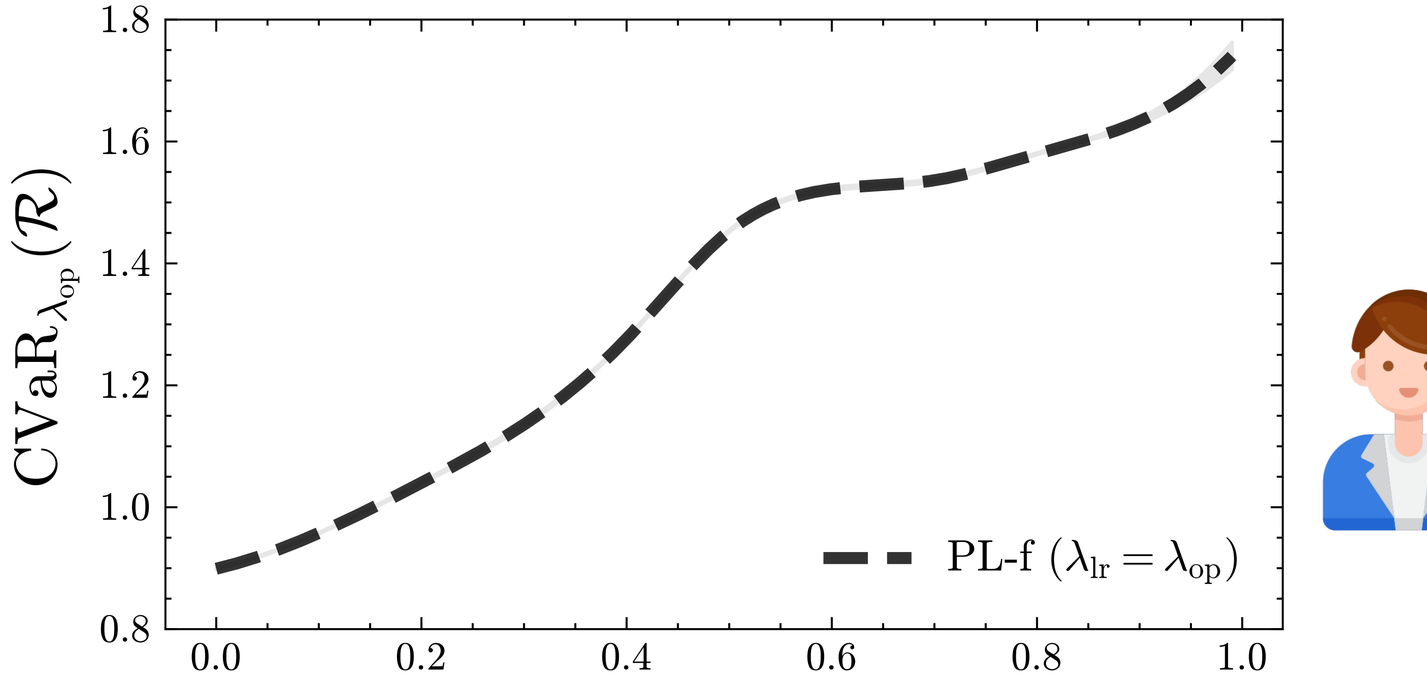
Where $\eta_\delta = (\delta - q^m)/(1 - q^m)$.

$$O(n^{-1/2} + m^{-1/2})$$

# Precise vs Imprecise Learning

$$Y_d = \theta_d X + \varepsilon,\ X \sim \mathcal{N}(1,0.5),\ \epsilon \sim \mathcal{N}(0,0.1),\ \theta_d \sim \mathcal{U}(1,1.1) \text{ or } \mathcal{U}(-1.1,-1)$$

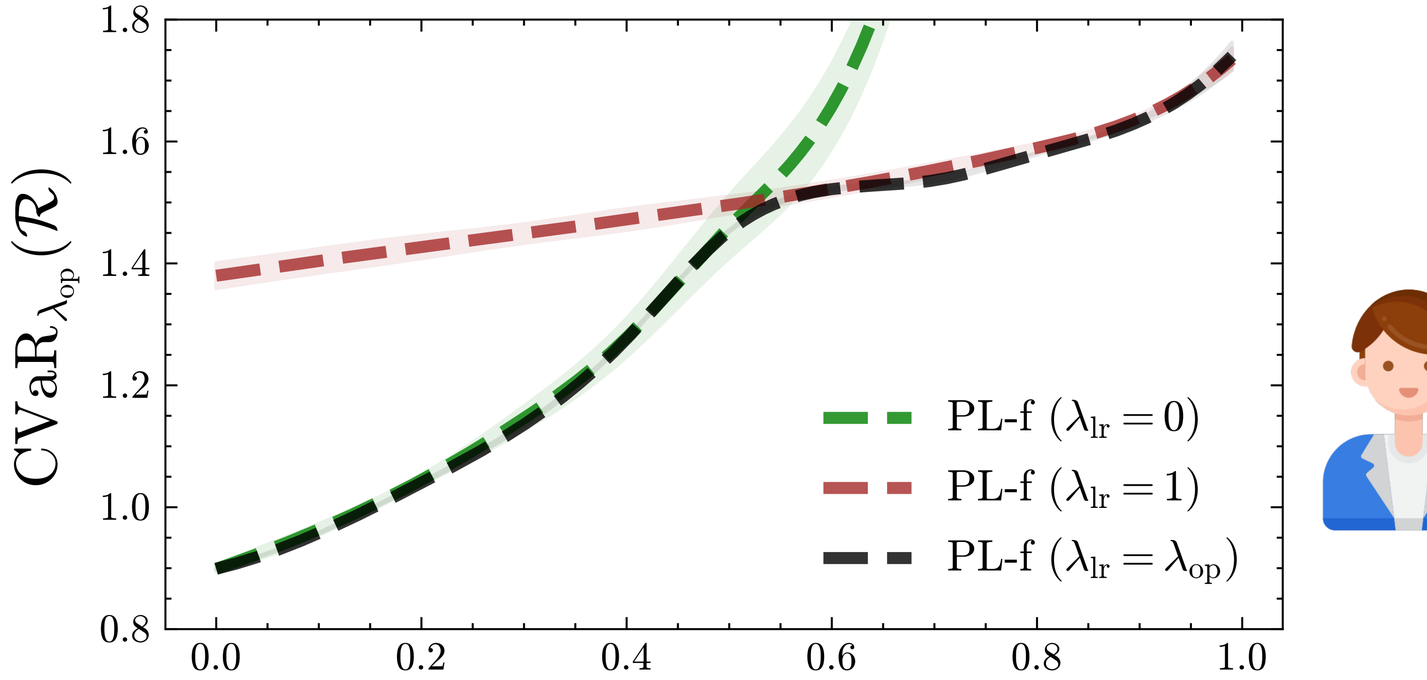
$n_{\text{train}}$ = 250, $n_{\text{test}}$ = 250, sample size = 100 from each domain




# Precise vs Imprecise Learning

$$Y_d = \theta_d X + \varepsilon,\ X \sim \mathcal{N}(1,0.5),\ \epsilon \sim \mathcal{N}(0,0.1),\ \theta_d \sim \mathcal{U}(1,1.1) \text{ or } \mathcal{U}(-1.1,-1)$$

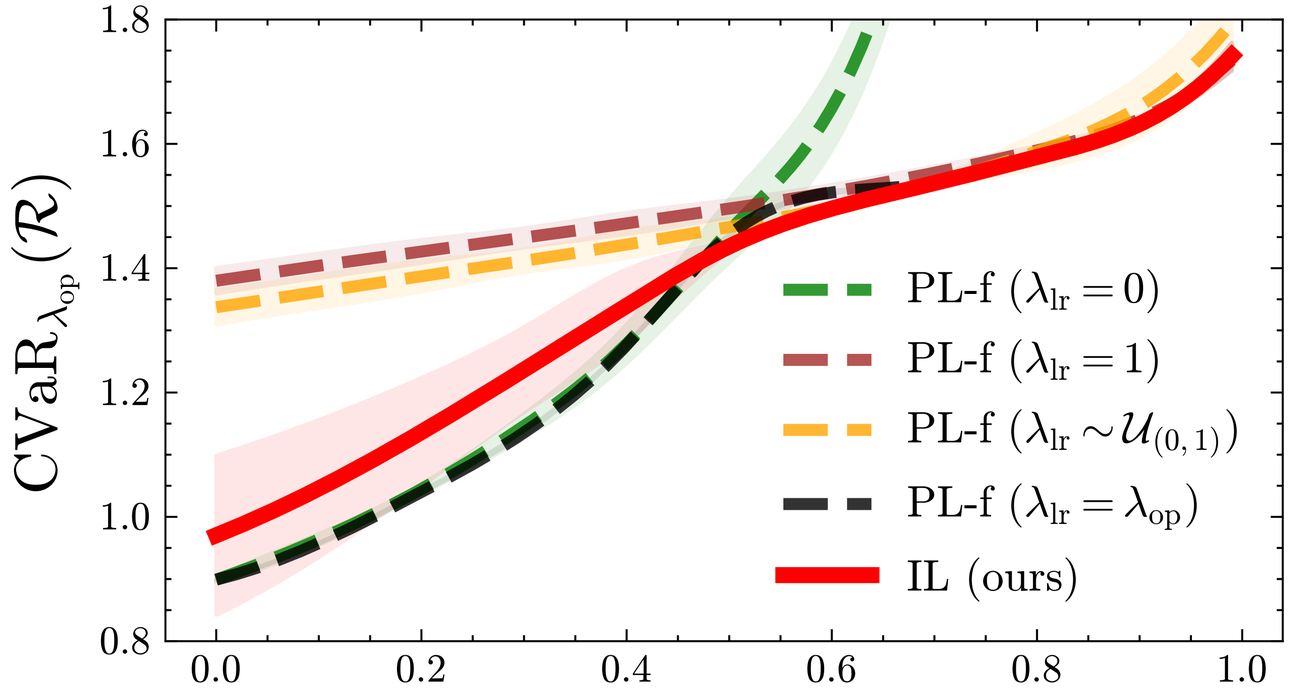
$n_{\text{train}}$ = 250, $n_{\text{test}}$ = 250, sample size = 100 from each domain



# Precise vs Imprecise Learning

$$Y_d = \theta_d X + \varepsilon,\ X \sim \mathcal{N}(1,0.5),\ \epsilon \sim \mathcal{N}(0,0.1),\ \theta_d \sim \mathcal{U}(1,1.1) \text{ or } \mathcal{U}(-1.1,-1)$$

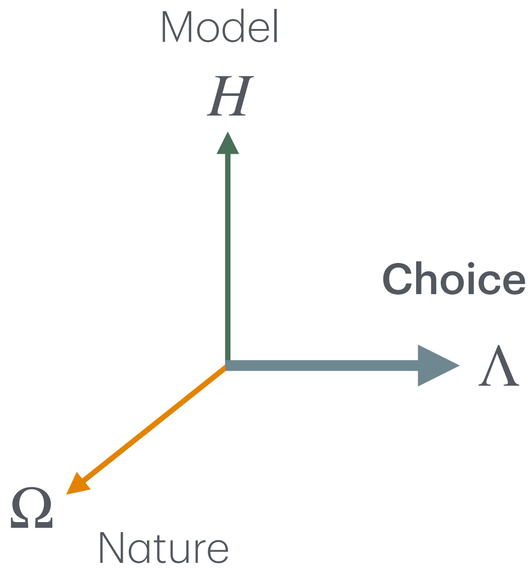
$n_{\text{train}}$ = 250, $n_{\text{test}}$ = 250, sample size = 100 from each domain




Risk neutral ← $\lambda_{\text{op}}$ → Risk averse

# Limitations and Outlook

- No assumption on the operators, except on the choice of $\rho_\lambda$
- Ignore **heterogeneity** of learning objectives across the choice of $\lambda$
- Assume that the operators know the **true choice** $\lambda^*$
- **Institutional separation** changes the training pipeline
  - Principal-agent model (Holmström, 1979)
  - Incomplete information games (Bergemann and Morris, 2019)

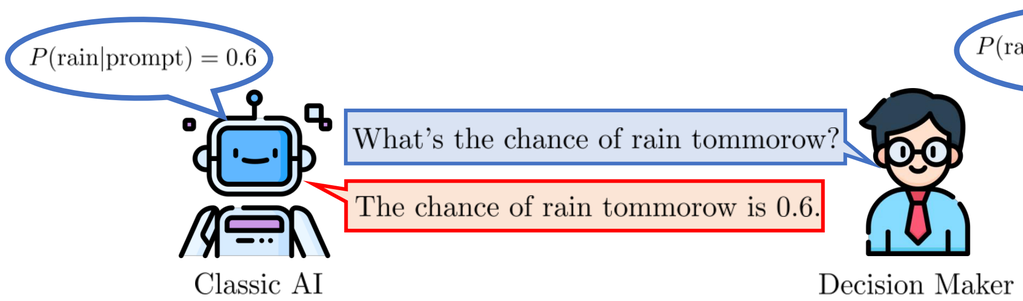



*Holmström, "Moral Hazard and Observability," The Bell Journal of Economics, 1979.*
*Bergemann and Morris, "Information Design: A Unified Perspective," Journal of Economic Literature, 2019.*

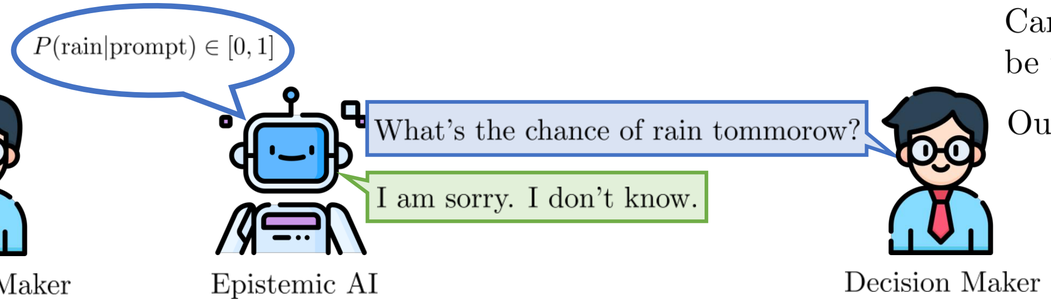
# Truthful Elicitation of Imprecise Forecast

UAI 2025 (Oral)

## Precise Forecast




## Imprecise Forecast




Can such epistemic AI be trained with ERM?

Our theory says,


✓ Strictly Proper Scoring Rules

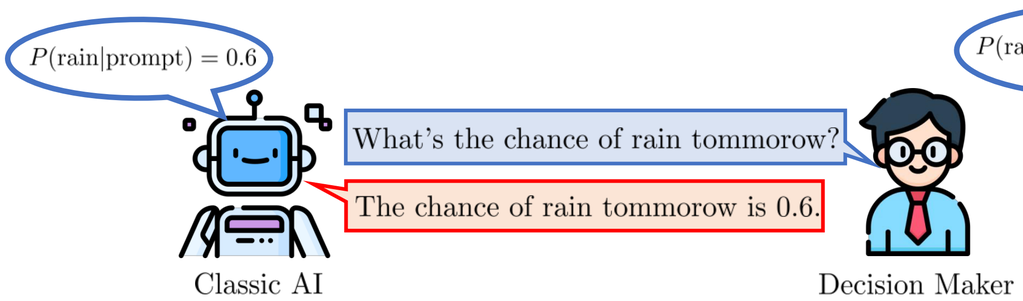
✗ Strictly Proper Scoring Rules

Impossibility results on IP scoring rules
Seidenfeld 2012; Mayo-Wilson 2015; Schoenfield 2017

# Truthful Elicitation of Imprecise Forecast

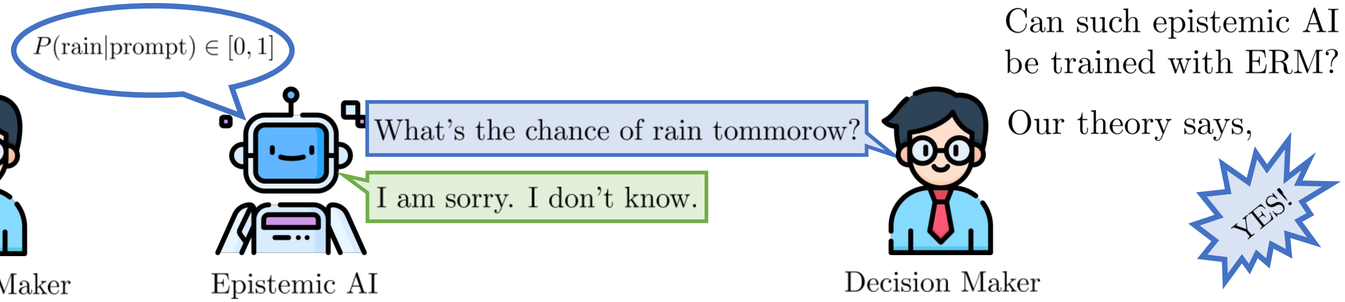
UAI 2025 (Oral)

## Precise Forecast




✓ Strictly Proper Scoring Rules

## Imprecise Forecast




✓ Strictly Proper Scoring Rules


Impossibility results on IP scoring rules
Seidenfeld 2012; Mayo-Wilson 2015; Schoenfield 2017

# (Im)possibility of Collective Intelligence

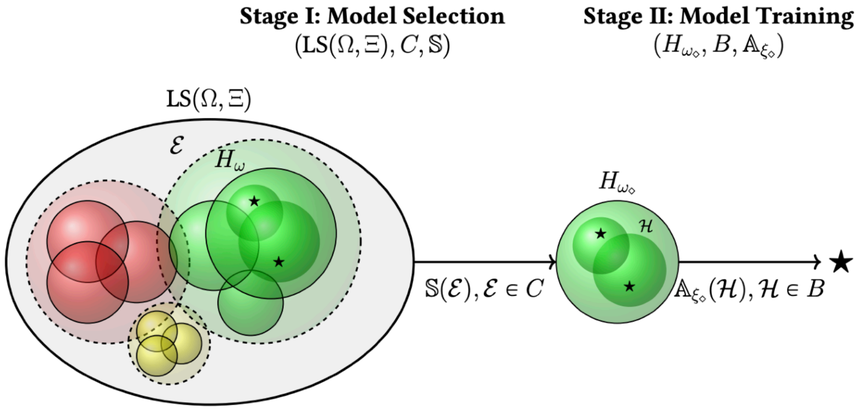
Krikamol Muandet*
CISPA Helmholtz Center for Information Security
Stuhlsatzenhaus 5, 66123 Saarbrücken, Germany
muandet@cispa.de, km@cifer.ai

May 20, 2025

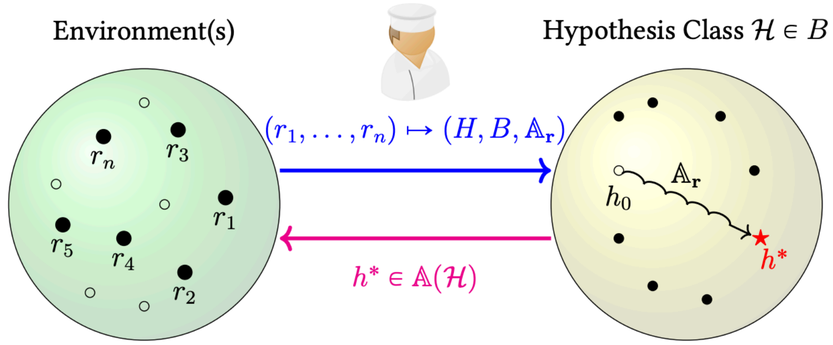
**Abstract**

Modern applications of AI involve training and deploying machine learning models across heterogeneous and potentially massive environments. Emerging diversity of data not only brings about new possibilities to advance AI systems, but also restricts the extent to which information can be shared across environments due to pressing concerns such as privacy, security, and equity. Based on a novel characterization of learning algorithms as choice correspondences on a hypothesis space, this work provides a minimum requirement in terms of intuitive and reasonable axioms under which the only rational learning algorithm in heterogeneous environments is an empirical risk minimization (ERM) that unilaterally learns from a single environment without information sharing across environments. Our (im)possibility result underscores the fundamental trade-off that any algorithms will face in order to achieve Collective Intelligence (CI), i.e., the ability to learn across heterogeneous environments. Ultimately, collective learning in heterogeneous environments are inherently hard because, in critical areas of machine learning such as out-of-distribution generalization, federated/collaborative learning, algorithmic fairness, and multi-modal learning, it can be infeasible to make meaningful comparisons of model predictive performance across environments.

**Keywords.** Democratization of AI, social choice theory, OOD generalization, federated learning, algorithmic fairness, multi-modal learning, collaborative learning



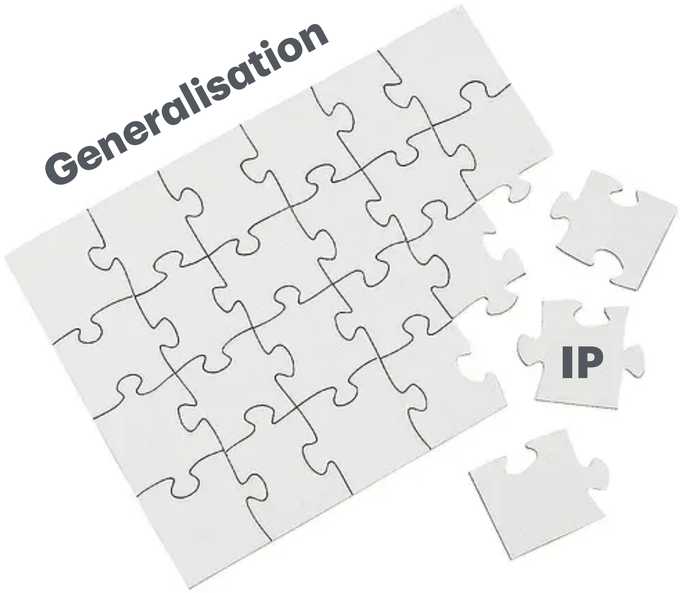

Learning and Model Selection




Collective Learning

# Takeaway

- Classical generalisation can be achieved via precise learning (ERM)
- Previous work in DA, CS, and DG addressed the distribution shifts by precise learning
- OOD generalisation involves both **decision-making** and **statistical learning** problems.
- An **institutional separation** hinders a precise learning
- Imprecise learning enables the learner to be
  less committal to specific notion of generalisation,
  allowing the operator to make informed decisions.
- Subjectivity in fairness, interpretability, robustness,
  trustworthiness, and privacy creates learning ambiguity.



# References


- Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. ***Domain generalization via invariant feature representation***. In Proceedings of the 30th International Conference on International Conference on Machine Learning (ICML'13), 2013.

- Anurag Singh, Siu Lun Chau, Shahine Bouabid, and Krikamol Muandet. ***Domain generalisation via imprecise learning***. In Proceedings of the 41st International Conference on Machine Learning (ICML'24), 2024.

- Siu Lun Chau, Antonin Schrab, Arthur Gretton, Dino Sejdinovic, Krikamol Muandet. ***Credal Two-Sample Tests of Epistemic Uncertainty*.** In Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (AISTATS'25), 2025.

- Michele Caprio, Maryam Sultana, Eleni G. Elia, Fabio Cuzzolin. ***Credal Learning Theory***. Advances in Neural Information Processing Systems (NeurIPS'24), 2024.

- Anurag Singh, Siu Lun Chau, Krikamol Muandet. ***Truthful Elicitation of Imprecise Forecast.*** Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence, 2025.

# Rational Intelligence Lab @ CISPA

**Krikamol Muandet**
PI

**Siu Lun Chau**
Postdoc (—> NTU)

**Gowtham Reddy**
Postdoc

**Julian Rodemann**
Postdoc (Sep 2025)

**Anurag Singh**
PhD student

**Kiet Vo**
PhD student

**Amine M'Charrak**
Visiting Student (Oxford)

**Majeed Mohammadi**
Visiting Postdoc (VU)

**Obaid Ur Rehman**
Master Student

**Cheng Song**
Research Assistant

**Open Position**

**Open Position**

**Krikamol Muandet**
CISPA Helmholtz Center for Information Security
muandet@cispa.de

RATIONAL INTELLIGENCE

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY