Imprecise Generalisation

Krikamol Muandet

Rational Intelligence Lab — CISPA Helmholtz Center for Information Security

Saarbrücken, Germany

SIPTA Seminar — 24 April 2025

Empirical Inference Symposium
In honour of the 75th birthday anniversary of Prof. Vladimir Vapnik
December 8-10, 2011

Generalisation

Generalisation

2, 4, 6, 8, 10, 12, 14, 16, 18, 20, …

Generalisation

2, 4, 6, 8, 10, 12, 14, 16, 18, 20, … → $2x$

Generalisation

2, 4, 6, 8, 10, 12, 14, 16, 18, 20, … $\longrightarrow$ $\boxed{2x}$ $\longrightarrow$ $y = 22$

Generalisation

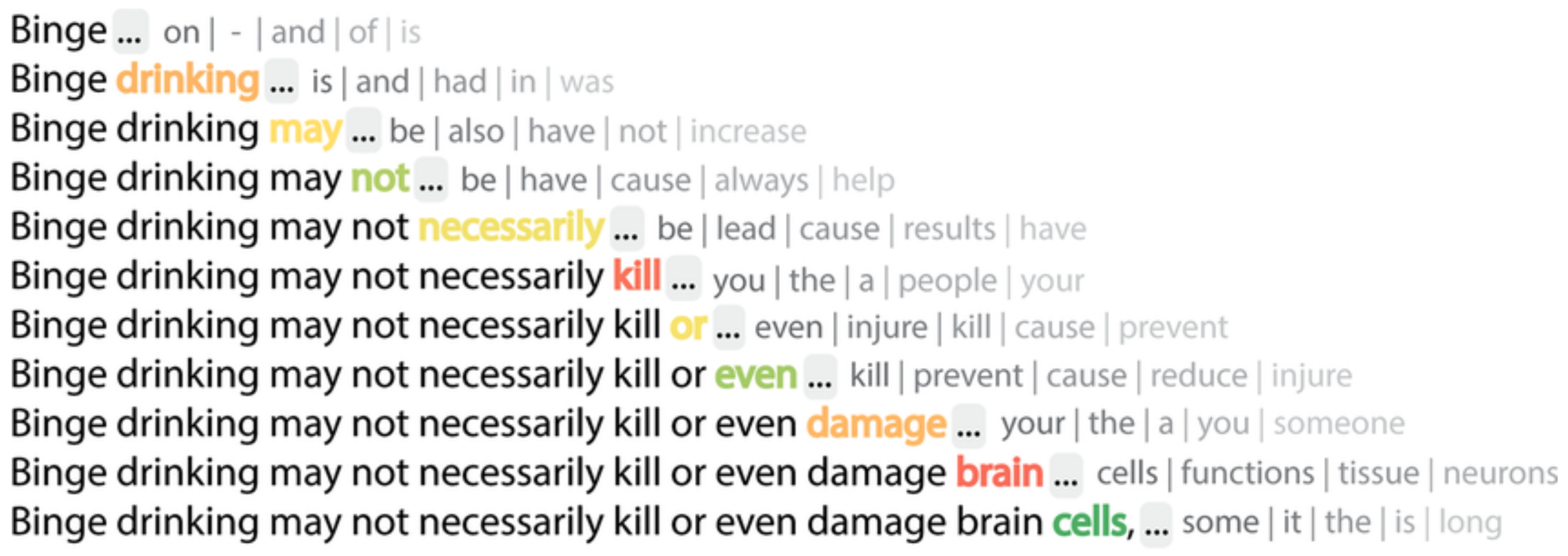
2, 4, 6, 8, 10, 12, 14, 16, 18, 20, … $\longrightarrow$ $\boxed{2x}$ $\longrightarrow$ $y = 22$

3, 4, 4, 10, 9, 14, 13, 17, 16, 22, … $\longrightarrow$ $\boxed{?}$ $\longrightarrow$ $y = ?$

Generalisation

$$2, 4, 6, 8, 10, 12, 14, 16, 18, 20, \ldots \longrightarrow \boxed{2x} \longrightarrow y = 22$$

$$3, 4, 4, 10, 9, 14, 13, 17, 16, 22, \ldots \longrightarrow \boxed{?} \longrightarrow y = ?$$



Cevoli et al. (2022)

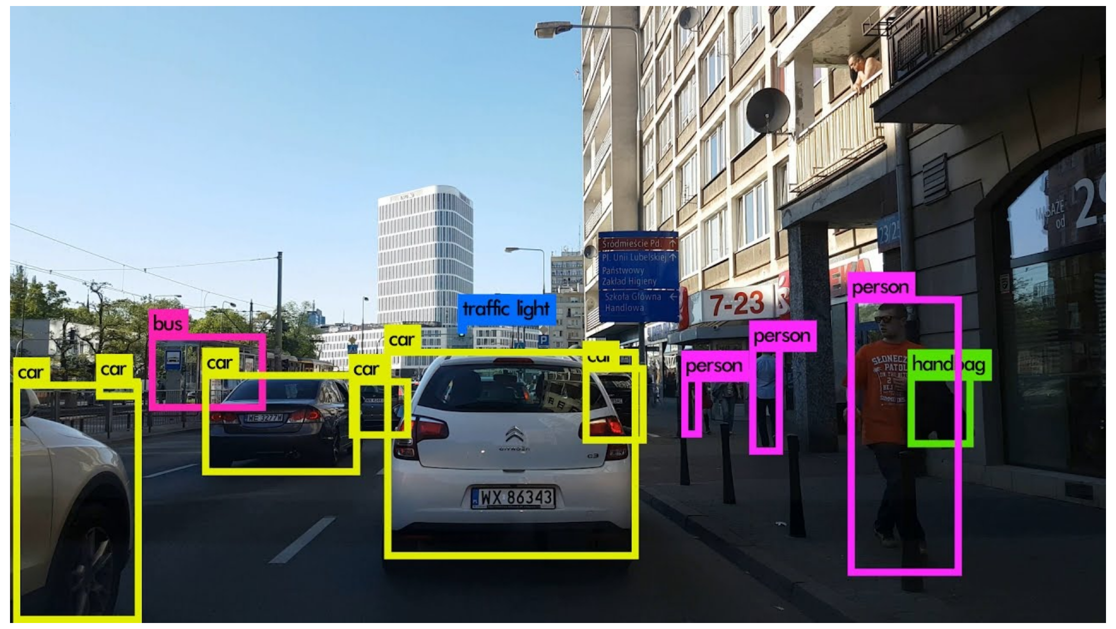
Generalisation

2, 4, 6, 8, 10, 12, 14, 16, 18, 20, … $\longrightarrow$ $2x$ $\longrightarrow$ $y = 22$

3, 4, 4, 10, 9, 14, 13, 17, 16, 22, … $\longrightarrow$? $\longrightarrow$ $y = ?$

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)

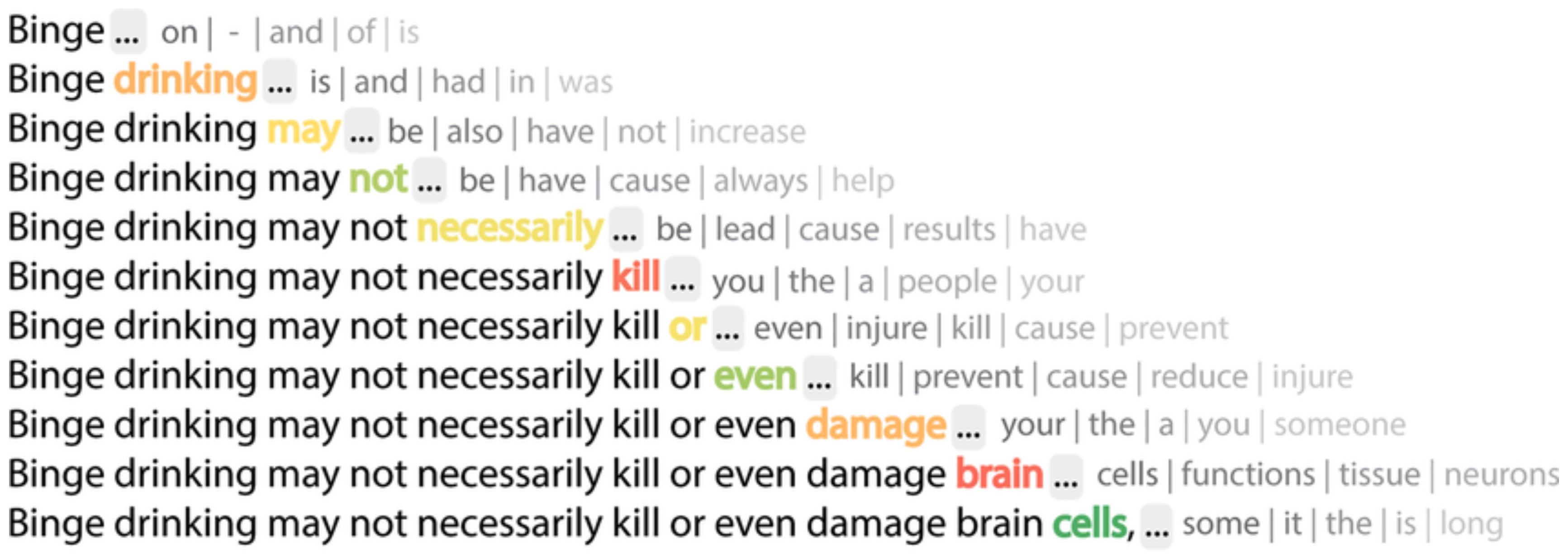


Sagarkar et al. (2020)

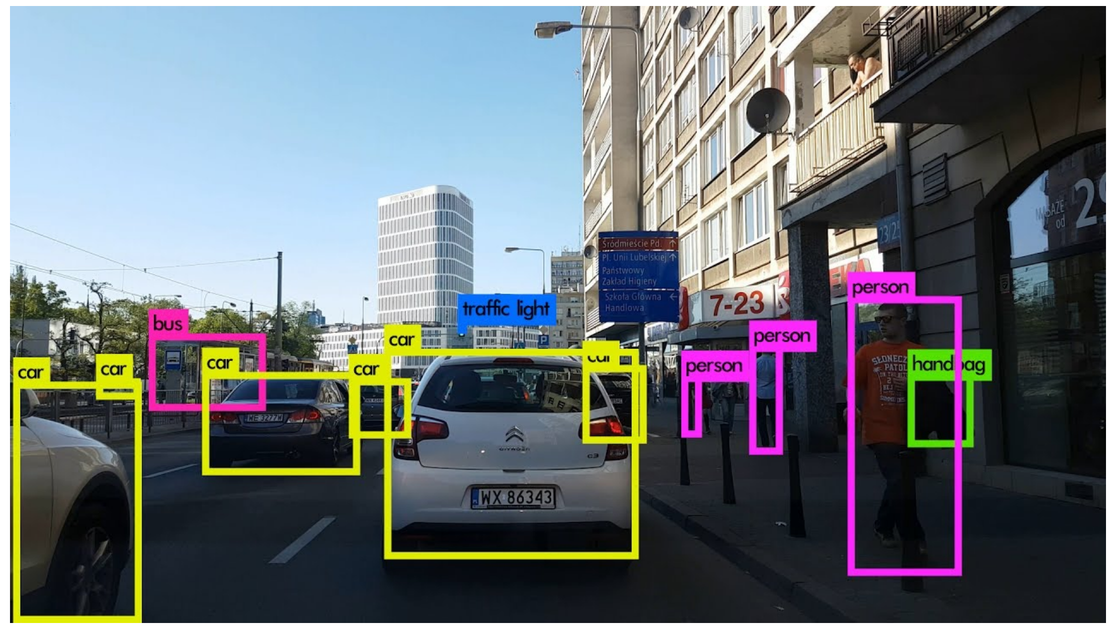
Generalisation

2, 4, 6, 8, 10, 12, 14, 16, 18, 20, … $\longrightarrow$ $\boxed{2x}$ $\longrightarrow$ $y = 22$

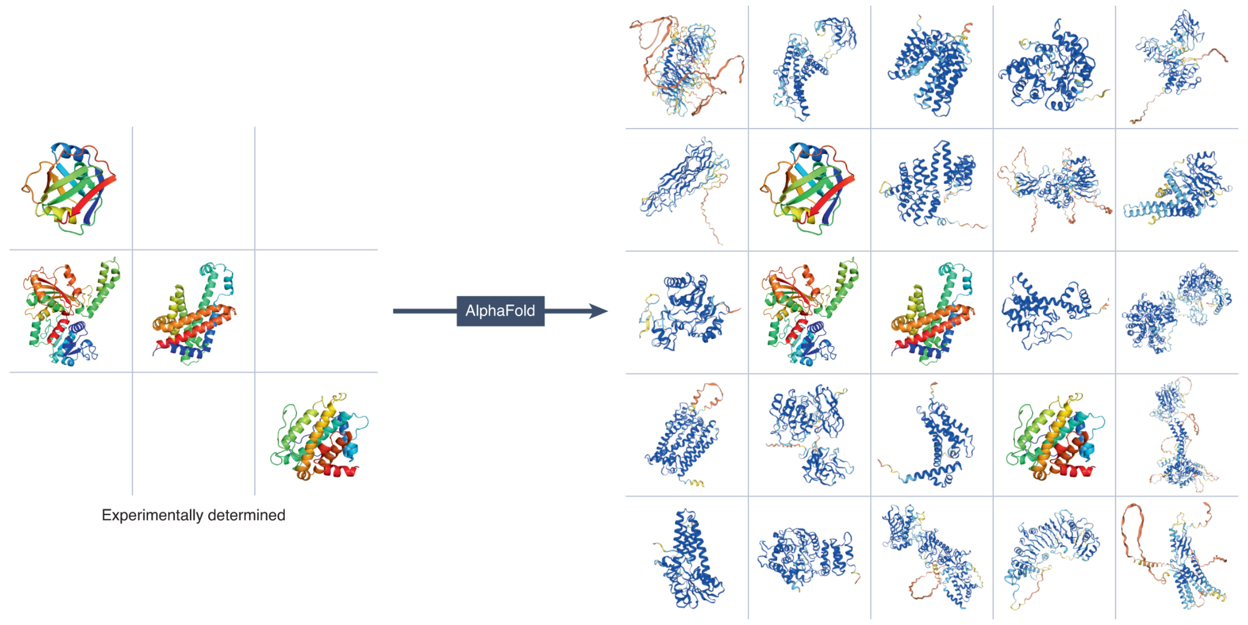
3, 4, 4, 10, 9, 14, 13, 17, 16, 22, … $\longrightarrow$ $\boxed{?}$ $\longrightarrow$ $y = ?$



Cevoli et al. (2022)



Sagarkar et al. (2020)



Jumper et al. 2021

Empirical Risk Minimisation (ERM)

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

Empirical Risk Minimisation (ERM)

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

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

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

Empirical Risk Minimisation (ERM)

- Observe sample of size n from some fixed but unknown 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 $h : \mathcal{X} \rightarrow \mathcal{Y}$

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

Empirical Risk Minimisation (ERM)

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

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

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

- **Recipe**: Minimise an empirical error on the observed data:

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

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

Empirical Risk Minimisation (ERM)

- Observe sample of size n from some fixed but unknown 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 $h : \mathcal{X} \to \mathcal{Y}$

- **Recipe**: Minimise an empirical error on the observed data:

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

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

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

Empirical Risk Minimisation (ERM)

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

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

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

- **Recipe**: Minimise an empirical error on the observed data:

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

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 $P(X, Y)$:

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

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

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)

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

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:

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

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)

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

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)

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

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)

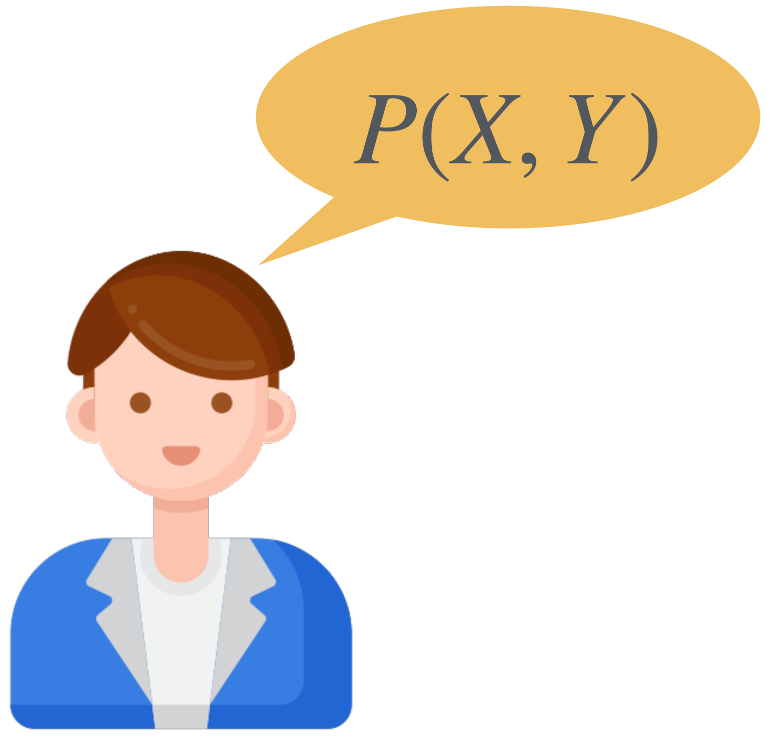
Neumann, John von, Oskar Morgenstern. Theory of Games and Economic Behavior. Princeton University Press, 1944.

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

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

Distribution Shifts

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

Distribution Shifts

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

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

Distribution Shifts

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

- Two sources of uncertainties:

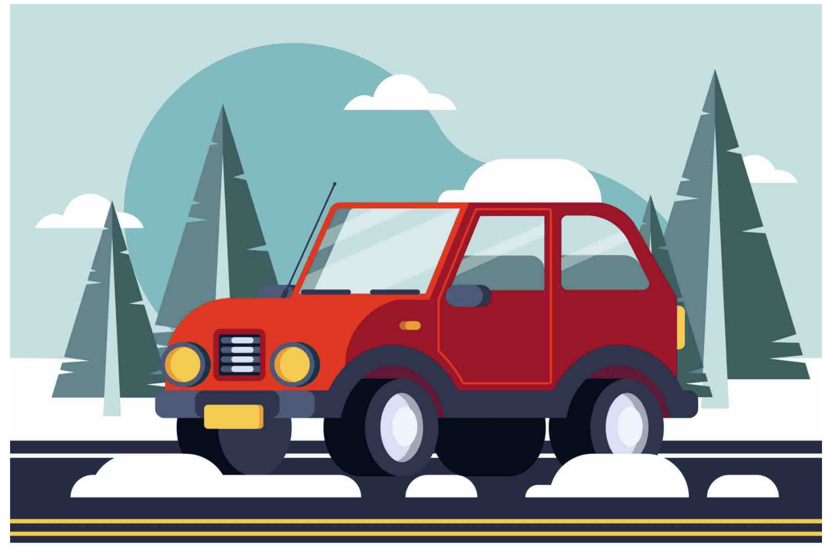
  1. Data uncertainty: The learner only observes a finite data
  2. Distribution uncertainty: The learner is uncertain about the data distribution

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

Distribution Shifts

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





- Two sources of uncertainties:
  1. Data uncertainty: The learner only observes a finite data
  2. Distribution uncertainty: The learner 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.

Domain Adaptation (DA)

Quionero-Candela, Sugiyama, Schwaighofer, and Lawrence. Dataset Shift in Machine Learning. The MIT Press, 2009; Ben-David et al. (2010)

Domain Adaptation (DA)

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

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

Quionero-Candela, Sugiyama, Schwaighofer, and Lawrence. Dataset Shift in Machine Learning. The MIT Press, 2009; Ben-David et al. (2010)

Domain Adaptation (DA)

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

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

- We can rewrite the expected loss under $Q(X, Y)$ as

$$R_Q(h) = \mathbb{E}_{(X,Y)\sim Q(X,Y)}[\ell(Y, h(X))] = \mathbb{E}_{(X,Y)\sim P(X,Y)}\left[\frac{Q(X, Y)}{P(X, Y)}\ell(Y, h(X))\right]$$

Quionero-Candela, Sugiyama, Schwaighofer, and Lawrence. Dataset Shift in Machine Learning. The MIT Press, 2009; Ben-David et al. (2010)

Domain Adaptation (DA)

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

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

- We can rewrite the expected loss under $Q(X, Y)$ as

$$R_Q(h) = \mathbb{E}_{(X,Y)\sim Q(X,Y)}[\ell(Y, h(X))] = \mathbb{E}_{(X,Y)\sim P(X,Y)}\left[\frac{Q(X,Y)}{P(X,Y)}\ell(Y, h(X))\right]$$

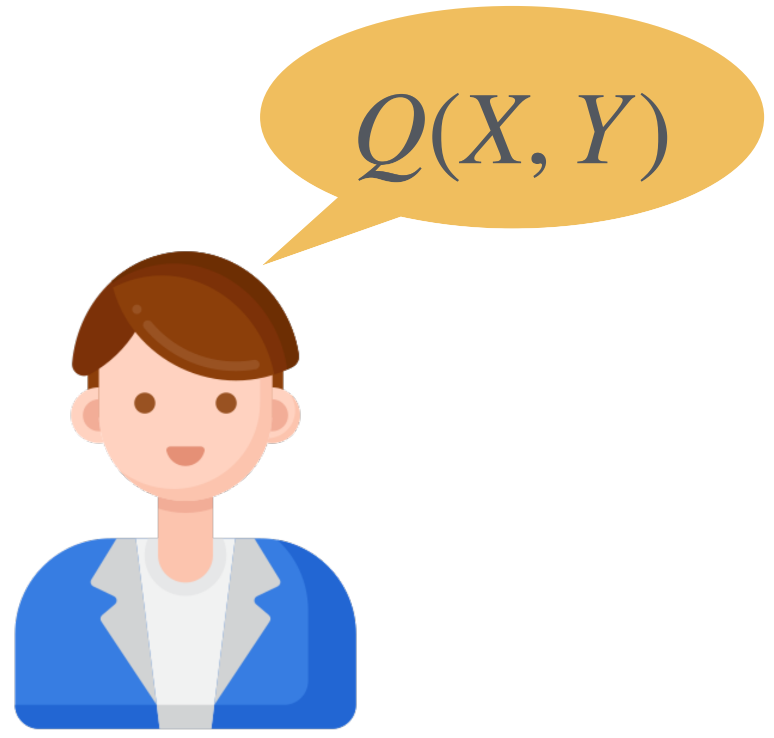
- We can learn h^* from (X_{tr}, Y_{tr}) given the density ratio $\omega(x, y) = Q(x, y)/P(x, y)$

Quionero-Candela, Sugiyama, Schwaighofer, and Lawrence. Dataset Shift in Machine Learning. The MIT Press, 2009; Ben-David et al. (2010)

Domain Adaptation (DA)

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

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



Precise Learner

- We can rewrite the expected loss under $Q(X,Y)$ as

$$R_Q(h) = \mathbb{E}_{(X,Y)\sim Q(X,Y)}[\ell(Y, h(X))] = \mathbb{E}_{(X,Y)\sim P(X,Y)}\left[\frac{Q(X,Y)}{P(X,Y)}\ell(Y, h(X))\right]$$

- We can learn h^* from (X_{tr}, Y_{tr}) given the density ratio $\omega(x,y) = Q(x,y)/P(x,y)$

Quionero-Candela, Sugiyama, Schwaighofer, and Lawrence. Dataset Shift in Machine Learning. The MIT Press, 2009; Ben-David et al. (2010)

Covariate Shift

- Covariate shift: $P(X, Y) = P(Y|X)P(X)$ and $Q(X, Y) = P(Y|X)Q(X)$

$$R_Q(h) = \mathbb{E}_{(X,Y)\sim Q(X,Y)}[\ell(Y, h(X))] = \mathbb{E}_{(X,Y)\sim P(X,Y)}\left[\frac{Q(X)}{P(X)}\ell(Y, h(X))\right]$$

Shimodaira, 2000; Ben-David et al. (2010); Sugiyama, Krauledat, Müller (2007)

Covariate Shift

- Covariate shift: $P(X, Y) = P(Y|X)P(X)$ and $Q(X, Y) = P(Y|X)Q(X)$

$$R_Q(h) = \mathbb{E}_{(X,Y)\sim Q(X,Y)}[\ell(Y, h(X))] = \mathbb{E}_{(X,Y)\sim P(X,Y)}\left[\frac{Q(X)}{P(X)}\ell(Y, h(X))\right]$$

- The density ratio $\omega(x) = Q(x)/P(x)$ is estimable from unlabelled data $X_{te} \sim Q(X)$

Shimodaira, 2000; Ben-David et al. (2010); Sugiyama, Krauledat, Müller (2007)

Covariate Shift

- Covariate shift: $P(X, Y) = P(Y|X)P(X)$ and $Q(X, Y) = P(Y|X)Q(X)$

$$R_Q(h) = \mathbb{E}_{(X,Y)\sim Q(X,Y)}[\ell(Y, h(X))] = \mathbb{E}_{(X,Y)\sim P(X,Y)}\left[\frac{Q(X)}{P(X)}\ell(Y, h(X))\right]$$

- The density ratio $\omega(x) = Q(x)/P(x)$ is estimable from unlabelled data $X_{te} \sim Q(X)$
- **Assumption**: $P(X)$ is absolutely continuous with respect to $Q(X)$

Shimodaira, 2000; Ben-David et al. (2010); Sugiyama, Krauledat, Müller (2007)

Covariate Shift

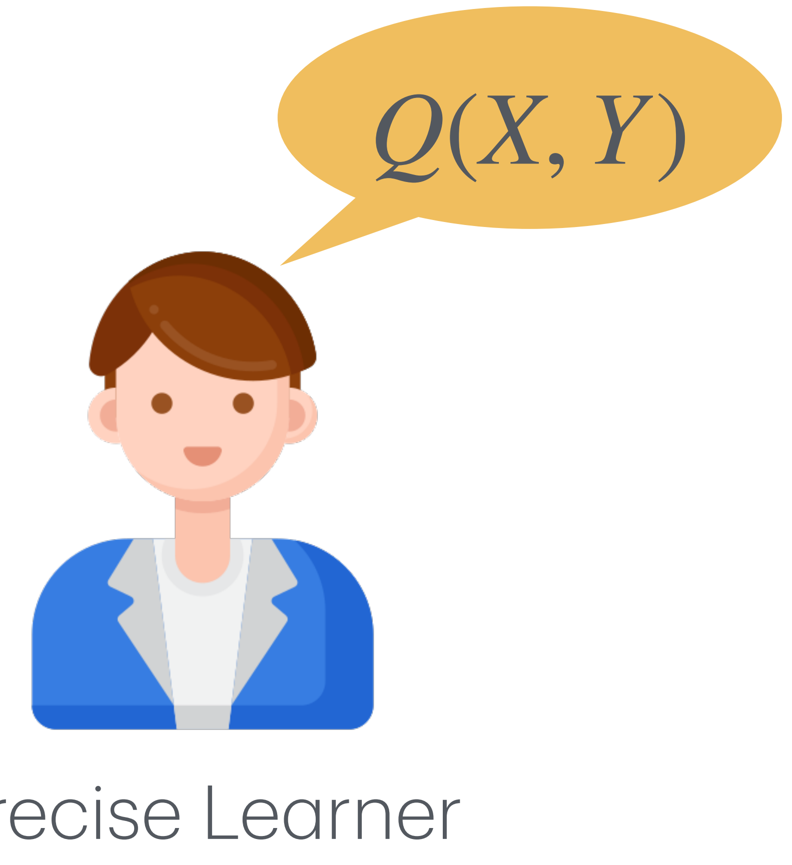
- Covariate shift: $P(X, Y) = P(Y|X)P(X)$ and $Q(X, Y) = P(Y|X)Q(X)$

$$R_Q(h) = \mathbb{E}_{(X,Y)\sim Q(X,Y)}[\ell(Y, h(X))] = \mathbb{E}_{(X,Y)\sim P(X,Y)}\left[\frac{Q(X)}{P(X)}\ell(Y, h(X))\right]$$

- The density ratio $\omega(x) = Q(x)/P(x)$ is estimable from unlabelled data $X_{te} \sim Q(X)$
- **Assumption:** $P(X)$ is absolutely continuous with respect to $Q(X)$
- 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* $P(X, Y) \neq Q(X, Y)$

Shimodaira, 2000; Ben-David et al. (2010); Sugiyama, Krauledat, Müller (2007)

Covariate Shift



- Covariate shift: $P(X, Y) = P(Y|X)P(X)$ and $Q(X, Y) = P(Y|X)Q(X)$

$$R_Q(h) = \mathbb{E}_{(X,Y)\sim Q(X,Y)}[\ell(Y, h(X))] = \mathbb{E}_{(X,Y)\sim P(X,Y)}\left[\frac{Q(X)}{P(X)}\ell(Y, h(X))\right]$$

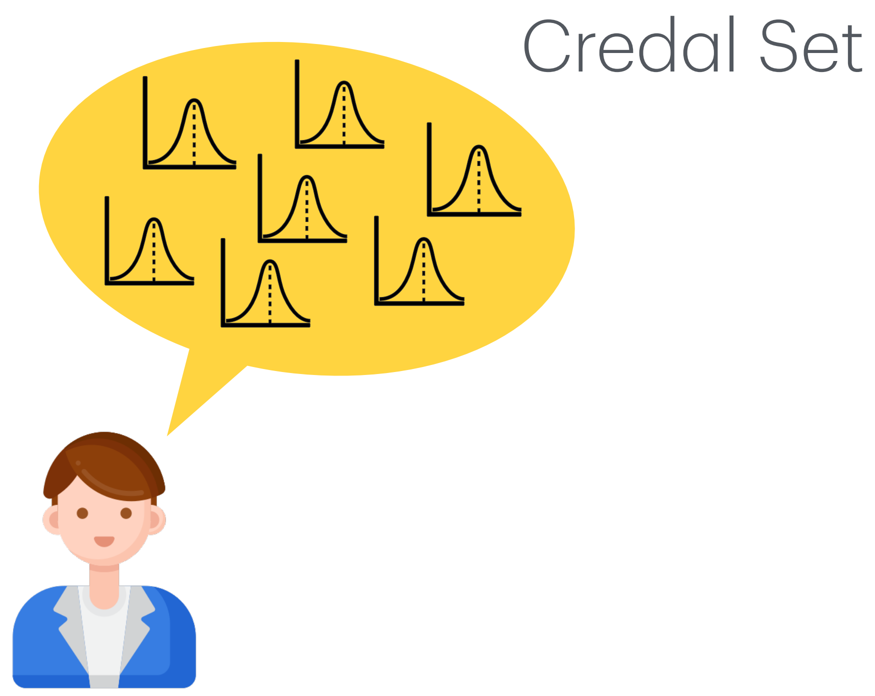
- The density ratio $\omega(x) = Q(x)/P(x)$ is estimable from unlabelled data $X_{te} \sim Q(X)$

- **Assumption**: $P(X)$ is absolutely continuous with respect to $Q(X)$

- 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* $P(X, Y) \neq Q(X, Y)$

Shimodaira, 2000; Ben-David et al. (2010); Sugiyama, Krauledat, Müller (2007)

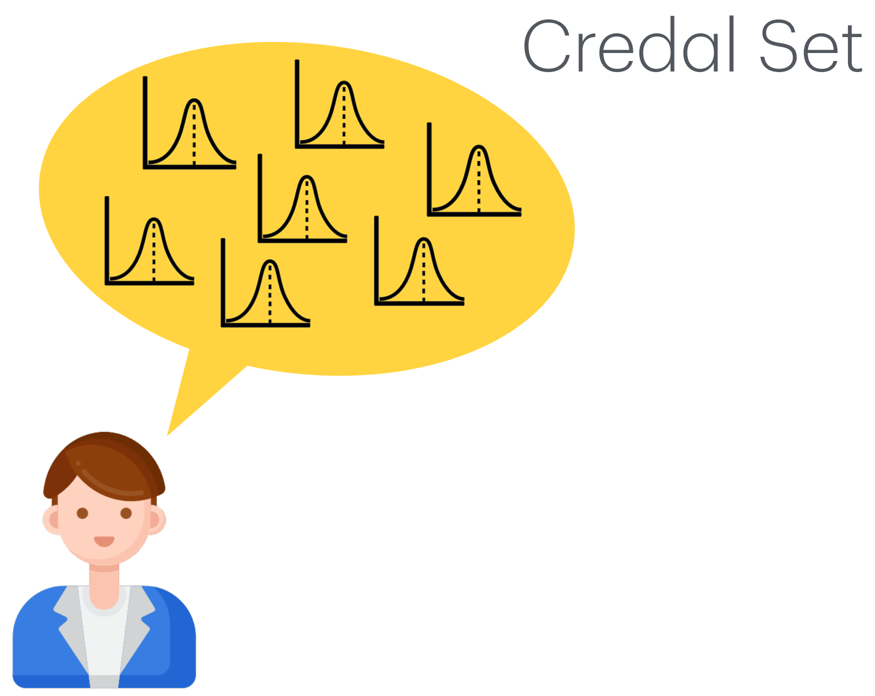
Domain Generalisation (DG)



- A collection of training distributions: $P_1(X, Y), P_2(X, Y), \ldots, P_N(X, Y)$

Blanchard et al. NeurIPS (2011); Muandet et al. ICML (2013)

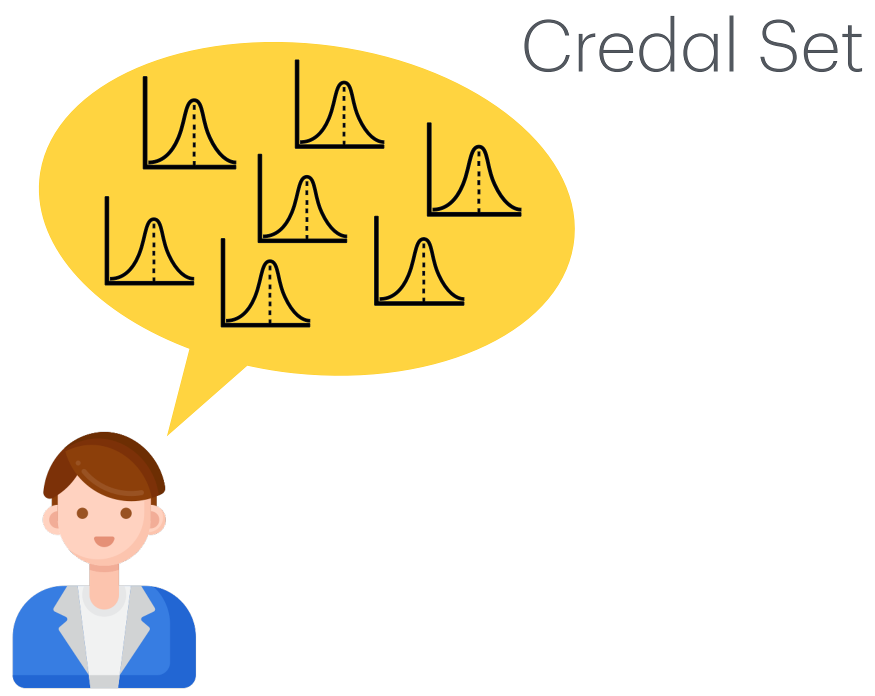
Domain Generalisation (DG)



- A collection of training distributions: $P_1(X, Y), P_2(X, Y), \ldots, P_N(X, Y)$
- *How to take knowledge acquired from an arbitrary number of related domains and apply it to previously **unseen** domains?*

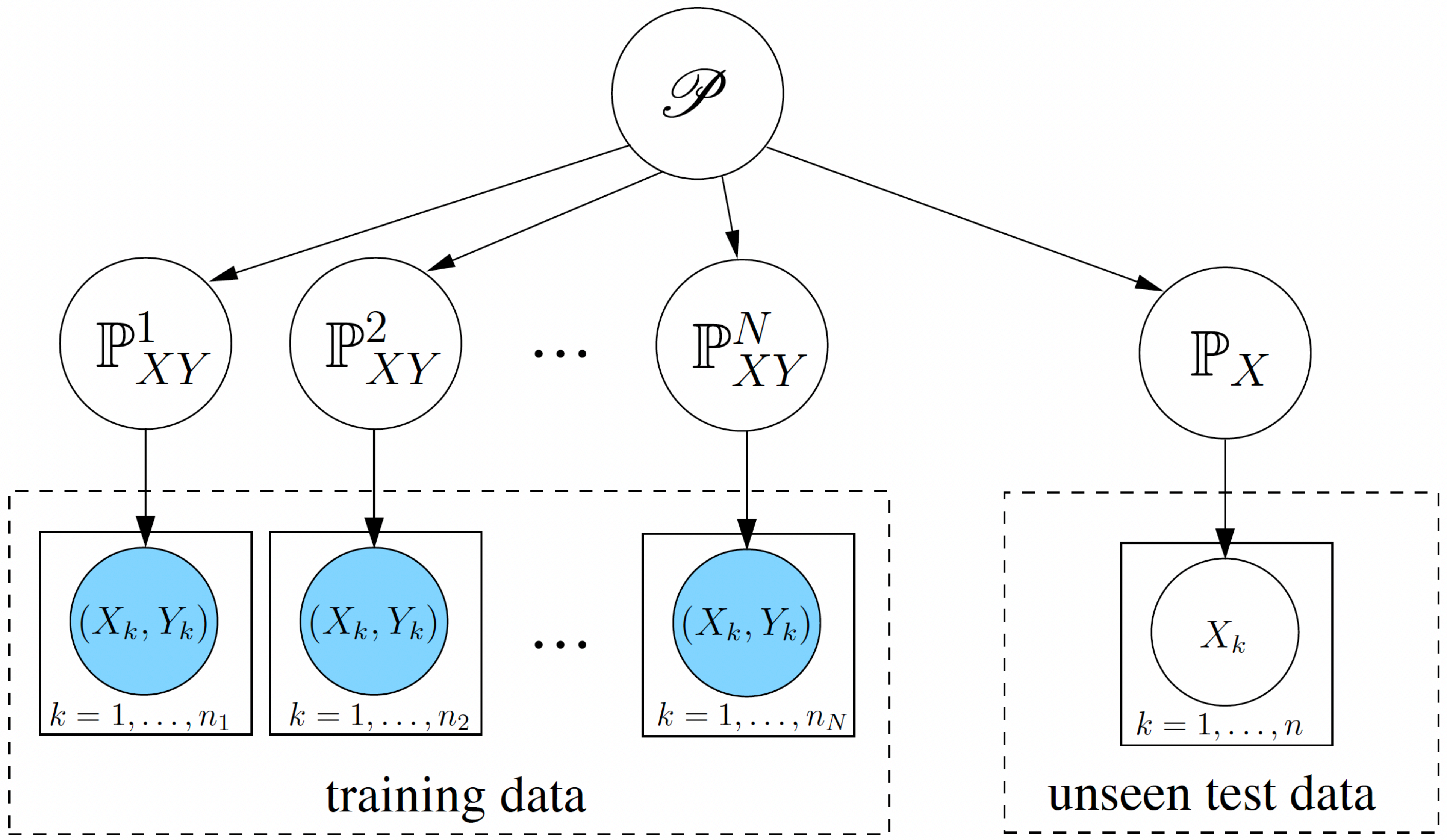
Blanchard et al. NeurIPS (2011); Muandet et al. ICML (2013)

Domain Generalisation (DG)



- A collection of training distributions: $P_1(X, Y), P_2(X, Y), \ldots, P_N(X, Y)$

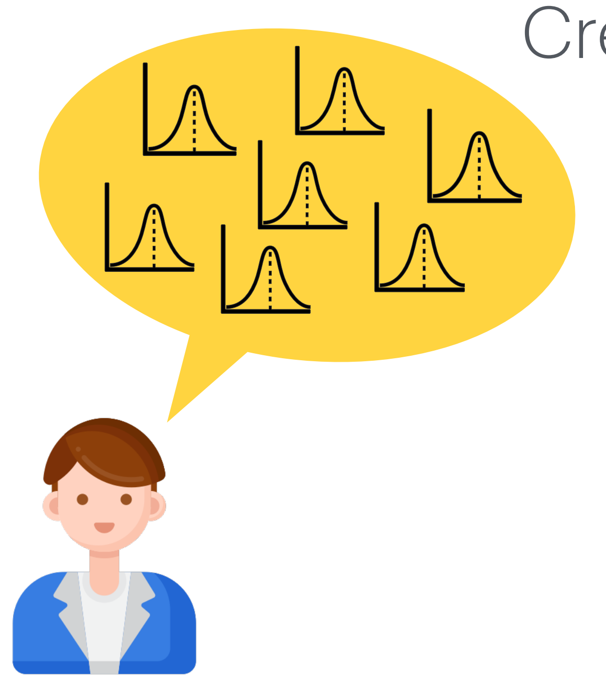
- *How to take knowledge acquired from an arbitrary number of related domains and apply it to previously* ***unseen*** *domains?*



Blanchard et al. NeurIPS (2011); Muandet et al. ICML (2013)

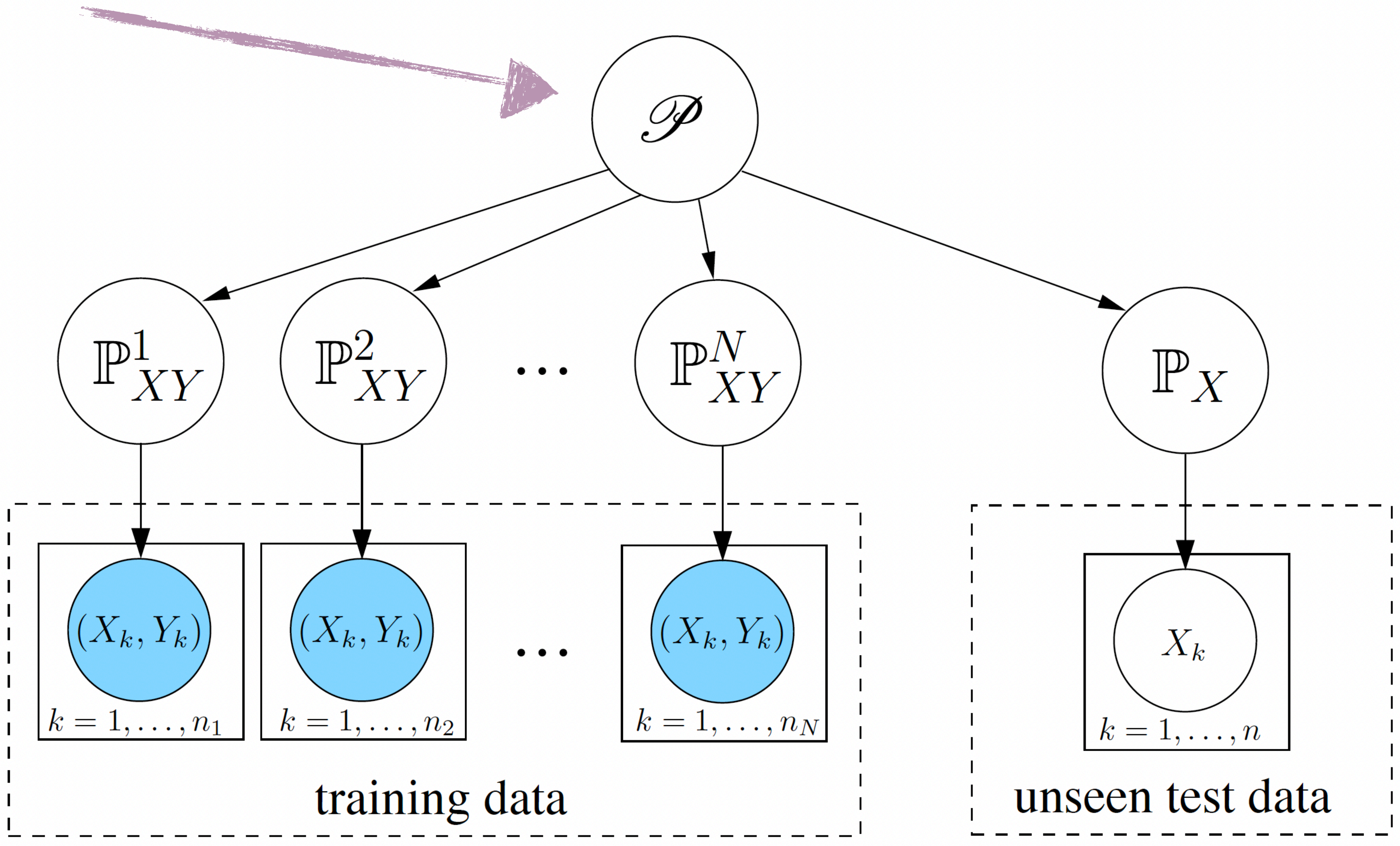
Domain Generalisation (DG)

Credal Set



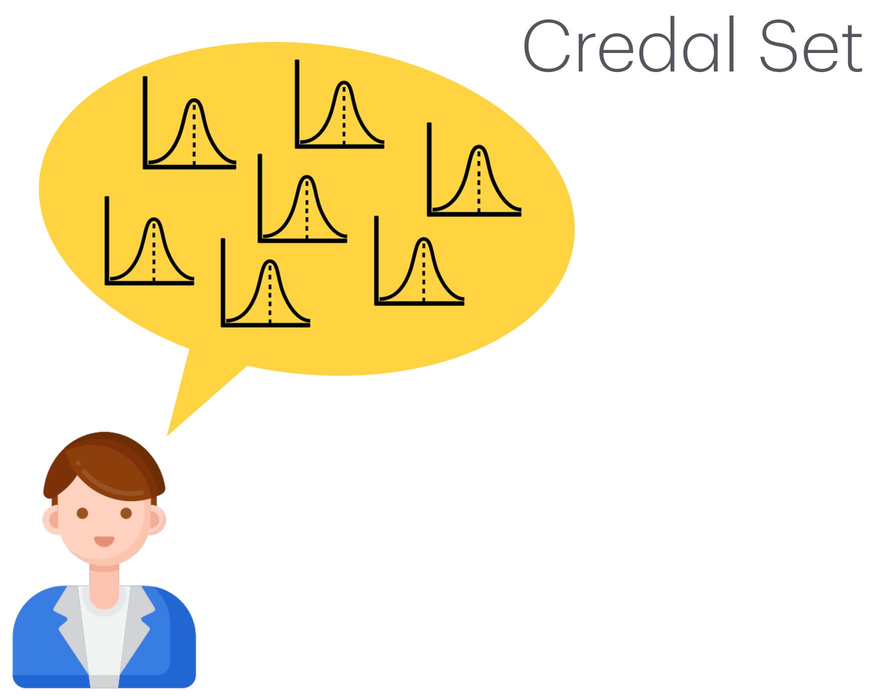
- A collection of training distributions: $P_1(X, Y), P_2(X, Y), \ldots, P_N(X, Y)$
- *How to take knowledge acquired from an arbitrary number of related domains and apply it to previously **unseen** domains?*

Meta-distribution



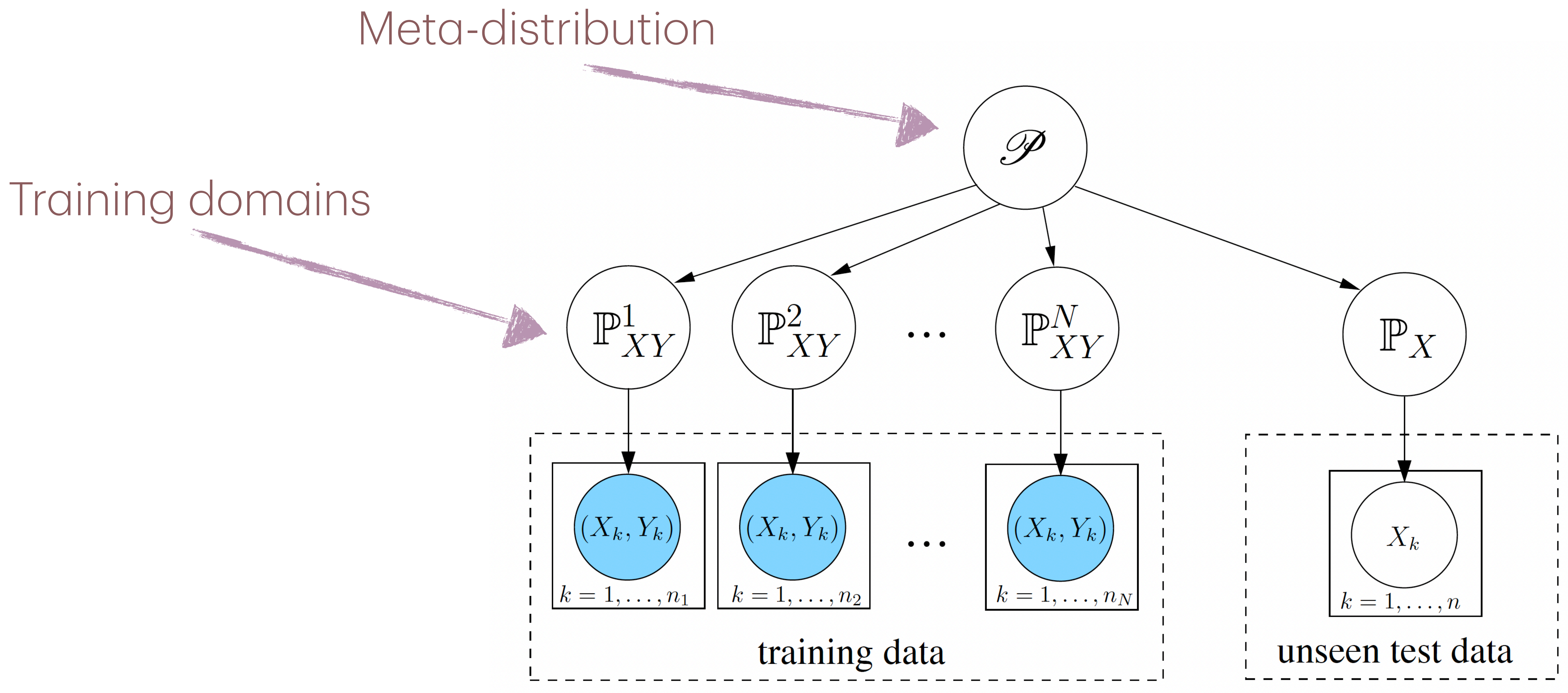
Blanchard et al. NeurIPS (2011); Muandet et al. ICML (2013)

Domain Generalisation (DG)



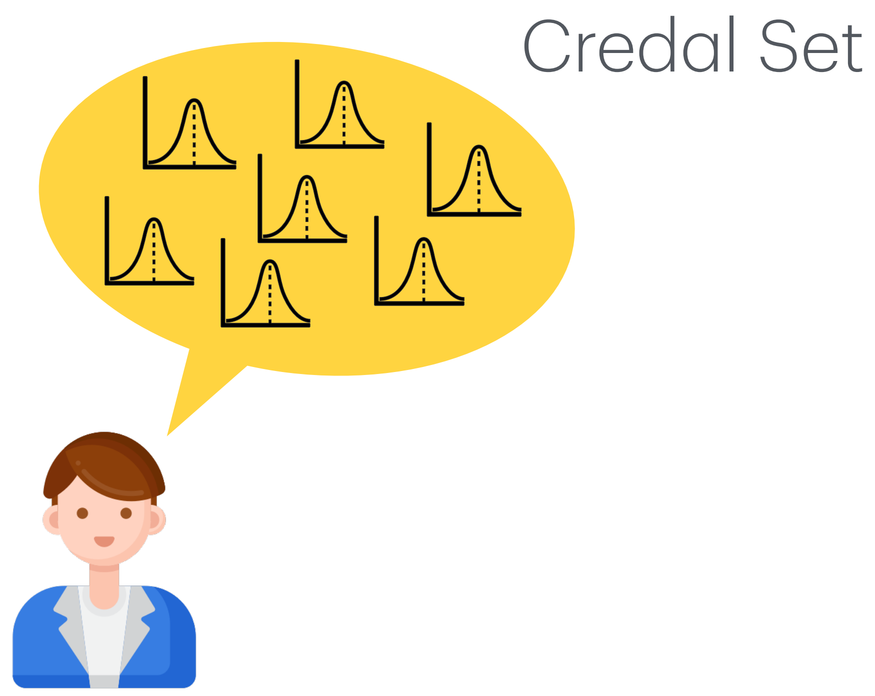
- A collection of training distributions: $P_1(X, Y), P_2(X, Y), \ldots, P_N(X, Y)$

- *How to take knowledge acquired from an arbitrary number of related domains and apply it to previously **unseen** domains?*

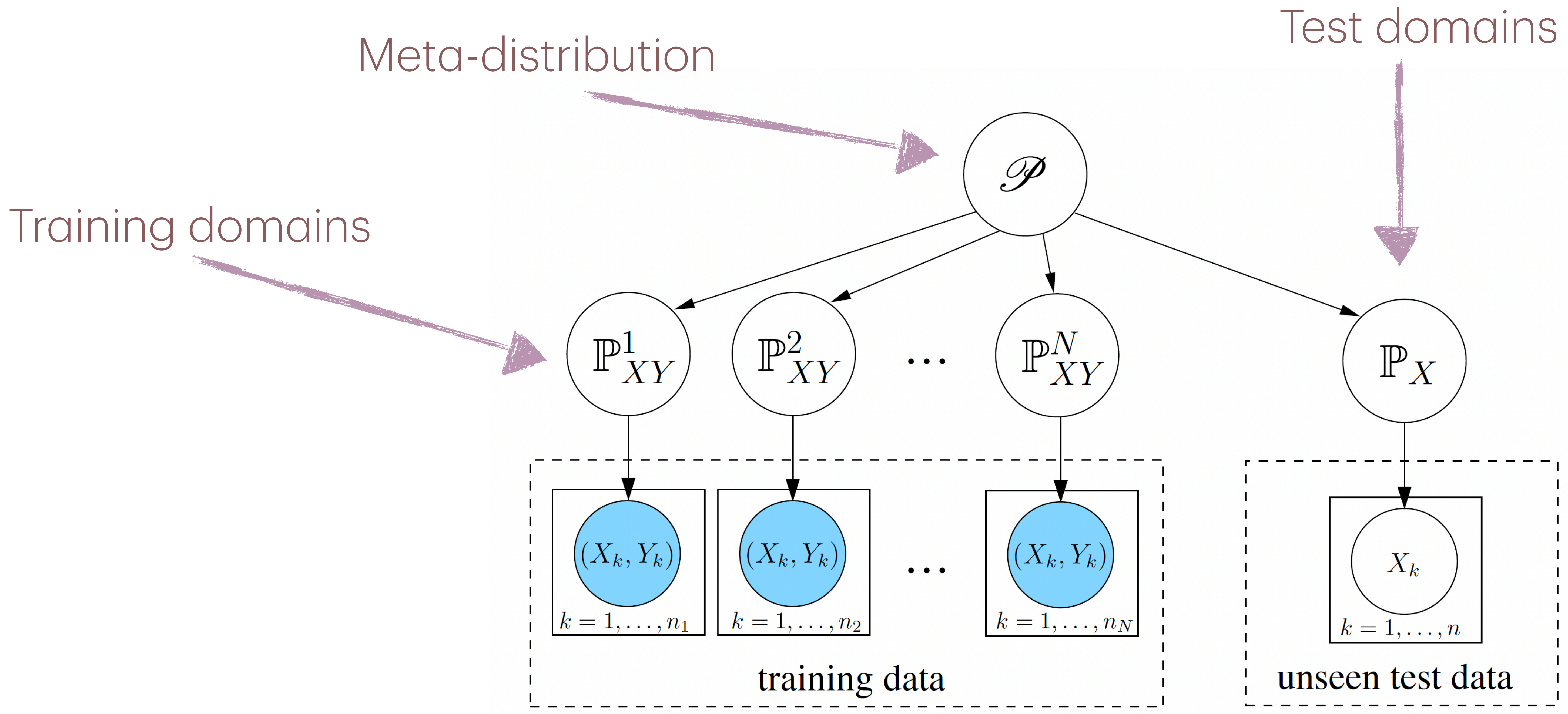


Blanchard et al. NeurIPS (2011); Muandet et al. ICML (2013)

Domain Generalisation (DG)



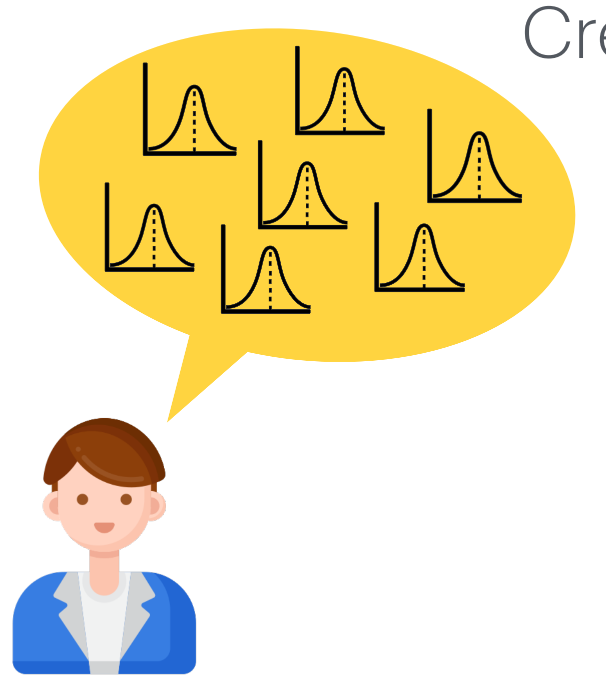
- A collection of training distributions: $P_1(X, Y), P_2(X, Y), \ldots, P_N(X, Y)$
- *How to take knowledge acquired from an arbitrary number of related domains and apply it to previously **unseen** domains?*



Blanchard et al. NeurIPS (2011); Muandet et al. ICML (2013)

Domain Generalisation (DG)

Credal Set



- A collection of training distributions: $P_1(X, Y), P_2(X, Y), \ldots, P_N(X, Y)$

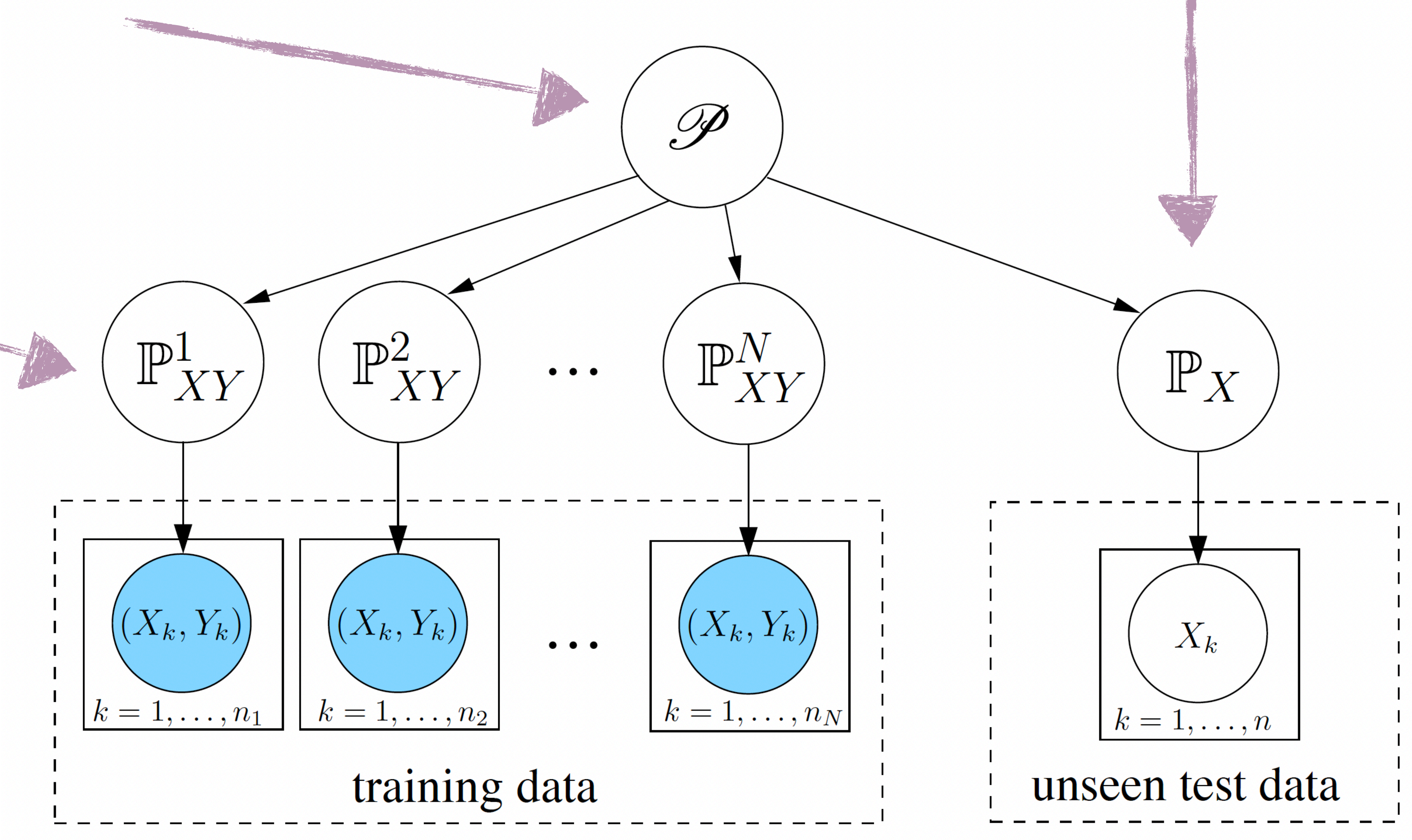
- *How to take knowledge acquired from an arbitrary number of related domains and apply it to previously **unseen** domains?*

Meta-distribution

Test domains

- The goal is to learn

Training domains



$$h : \mathscr{P}_{\mathscr{X}} \times \mathscr{X} \to \mathscr{Y}$$

where $\mathscr{P}_{\mathscr{X}}$ is the set of distributions on $\mathscr{X}$

Blanchard et al. NeurIPS (2011); Muandet et al. ICML (2013)

Invariance Principle

Muandet, Balduzzi, and Schölkopf. Domain Generalisation via Invariant Feature Representation, ICML (2013)

Invariance Principle

- Muandet et al. (2013) proposes to learn a feature representation $\phi : \mathscr{X} \to \mathscr{F}$ that

    1. minimises the **distributional variance** of $P(\phi(X))$ between domains

    2. preserves the **functional relationship** between $\phi(X)$ and Y

Muandet, Balduzzi, and Schölkopf. Domain Generalisation via Invariant Feature Representation, ICML (2013)

Invariance Principle

- Muandet et al. (2013) proposes to learn a feature representation $\phi : \mathscr{X} \to \mathscr{F}$ that
    1. minimises the **distributional variance** of $P(\phi(X))$ between domains
    2. preserves the **functional relationship** between $\phi(X)$ and Y
- Domain-Invariant Component Analysis (DICA):

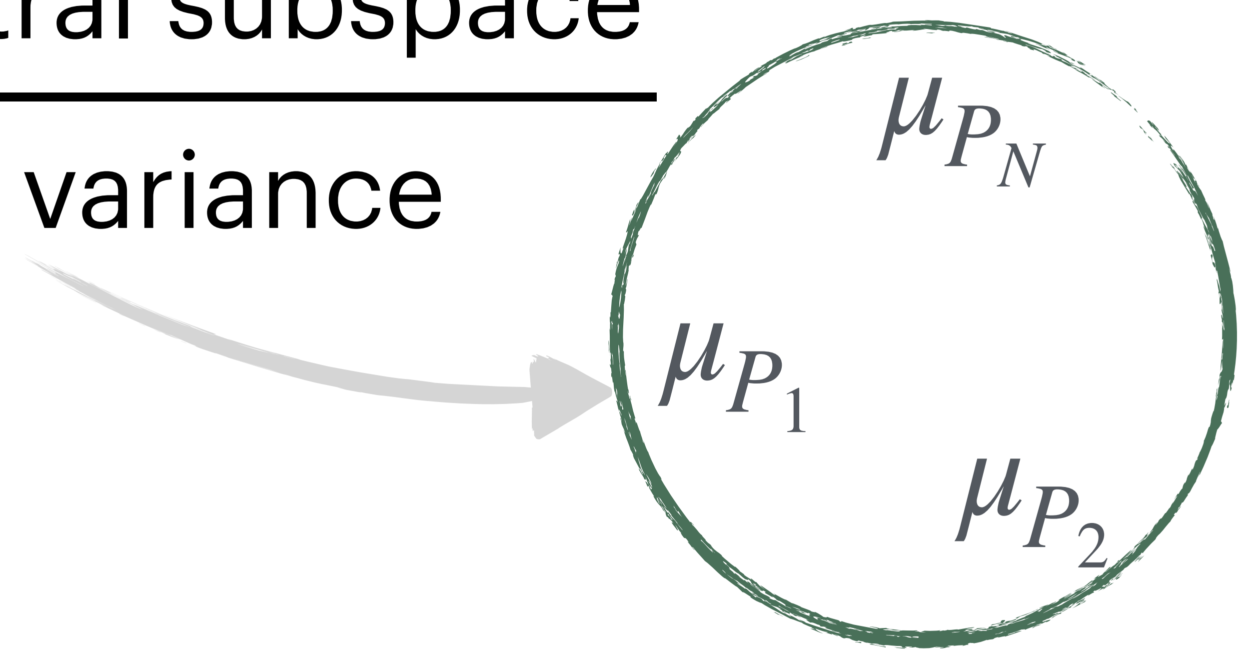
$$\max_{\phi} \frac{\text{preserve the central subspace}}{\text{distributional variance}}$$

Muandet, Balduzzi, and Schölkopf. Domain Generalisation via Invariant Feature Representation, ICML (2013)

Invariance Principle

- Muandet et al. (2013) proposes to learn a feature representation $\phi : \mathscr{X} \to \mathscr{F}$ that
    1. minimises the **distributional variance** of $P(\phi(X))$ between domains
    2. preserves the **functional relationship** between $\phi(X)$ and Y
- Domain-Invariant Component Analysis (DICA):

$$\max_{\phi} \frac{\text{preserve the central subspace}}{\text{distributional variance}}$$

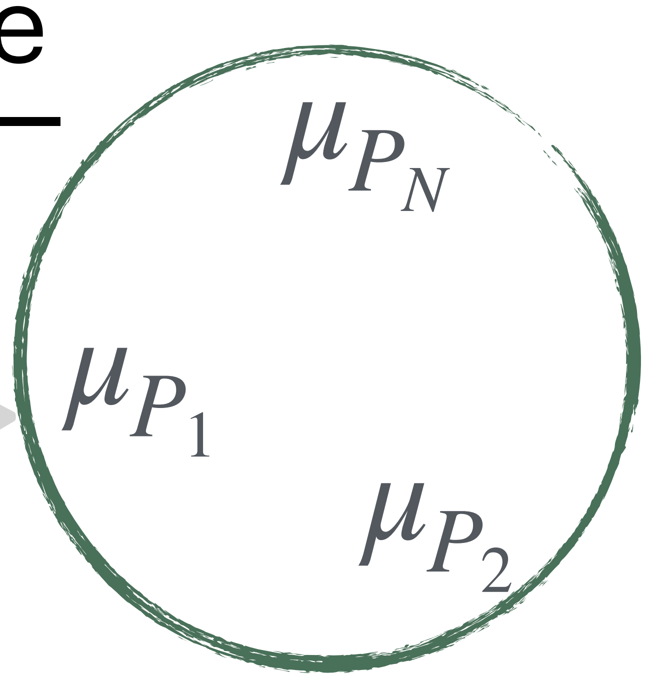


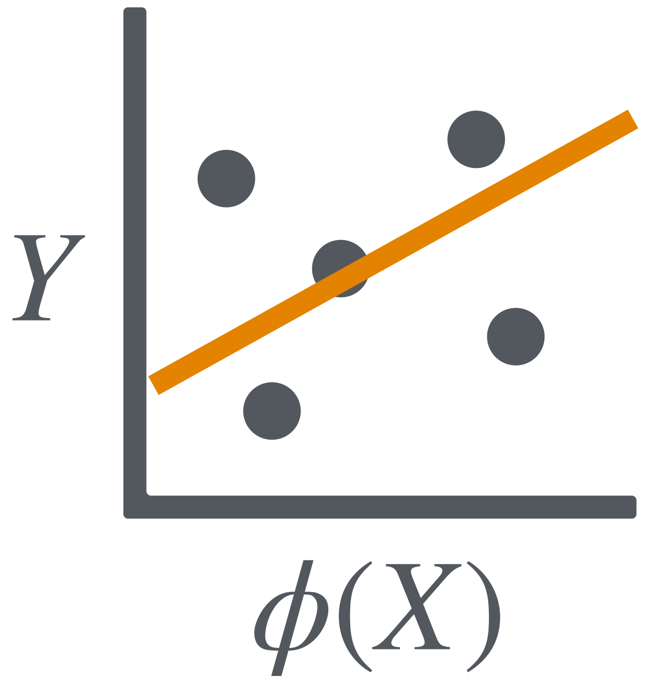
Muandet, Balduzzi, and Schölkopf. Domain Generalisation via Invariant Feature Representation, ICML (2013)

Invariance Principle

- Muandet et al. (2013) proposes to learn a feature representation $\phi : \mathscr{X} \to \mathscr{F}$ that
  1. minimises the **distributional variance** of $P(\phi(X))$ between domains
  2. preserves the **functional relationship** between $\phi(X)$ and Y
- Domain-Invariant Component Analysis (DICA):

$$\max_{\phi} \frac{\text{preserve the central subspace}}{\text{distributional variance}}$$



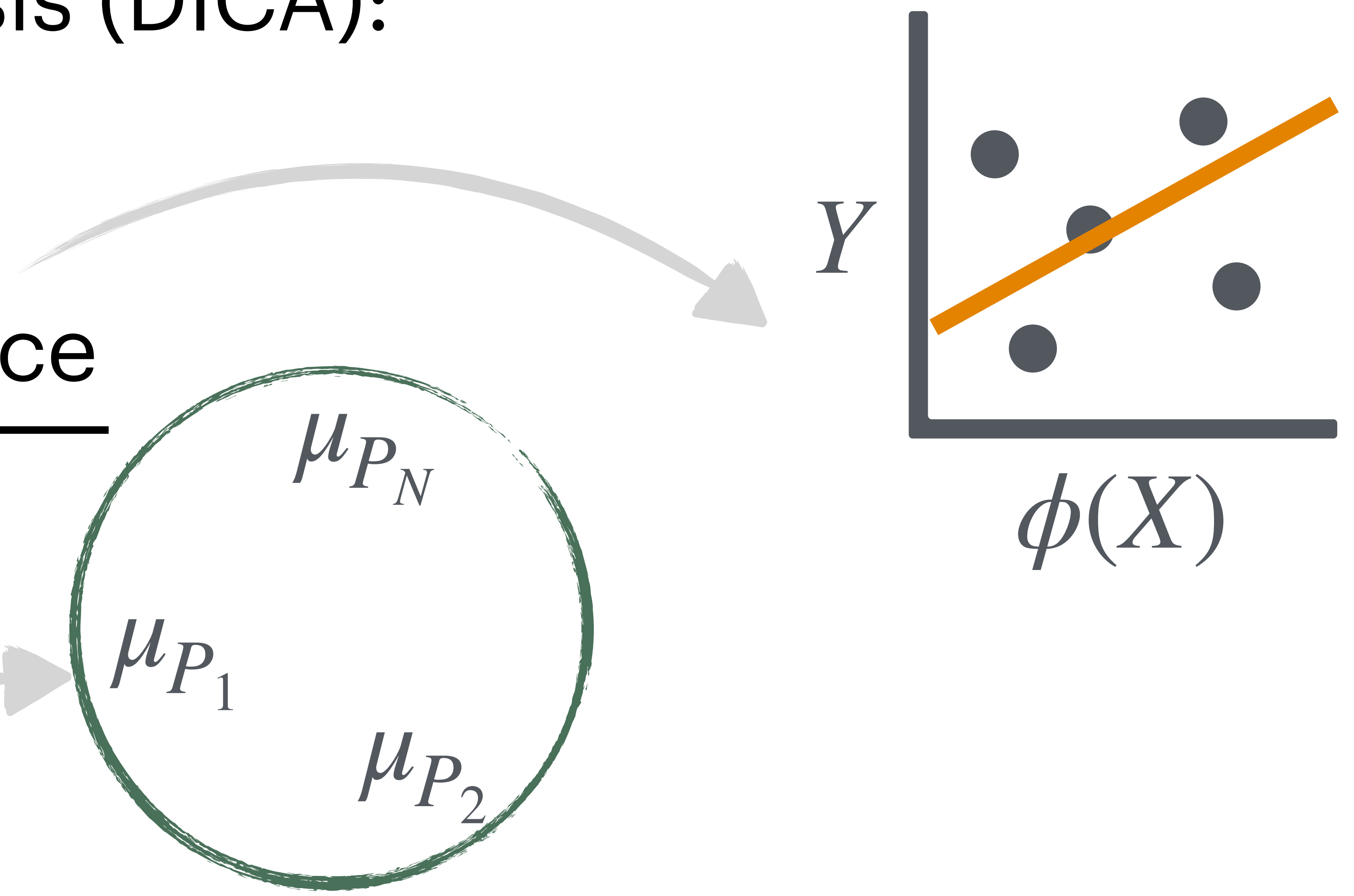


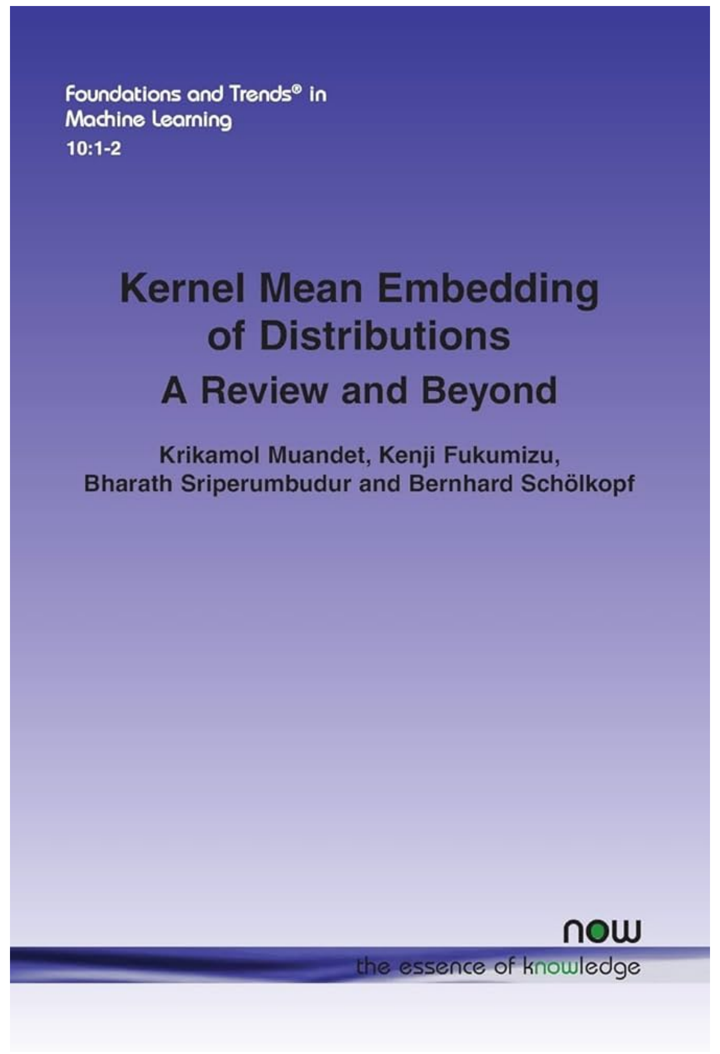
Muandet, Balduzzi, and Schölkopf. Domain Generalisation via Invariant Feature Representation, ICML (2013)

Invariance Principle

- Muandet et al. (2013) proposes to learn a feature representation $\phi : \mathscr{X} \rightarrow \mathscr{F}$ that
  1. minimises the **distributional variance** of $P(\phi(X))$ between domains
  2. preserves the **functional relationship** between $\phi(X)$ and Y
- Domain-Invariant Component Analysis (DICA):

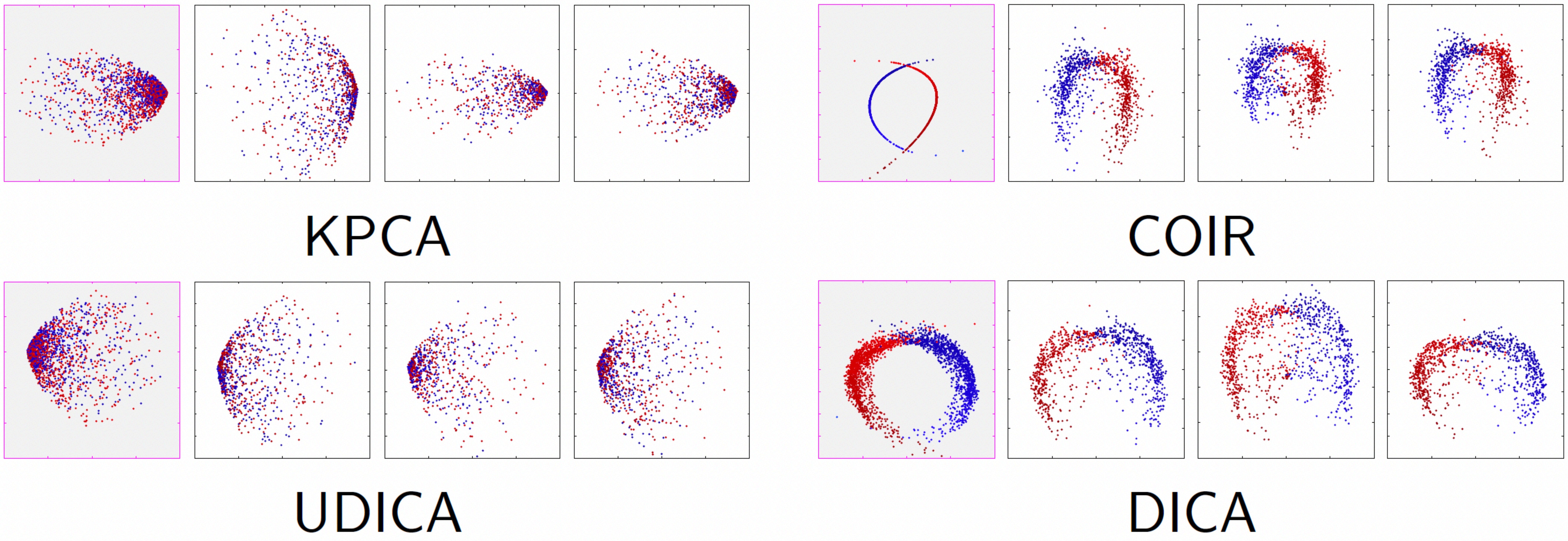
$$\max_{\phi} \frac{\text{preserve the central subspace}}{\text{distributional variance}}$$





Muandet, Balduzzi, and Schölkopf. Domain Generalisation via Invariant Feature Representation, ICML (2013)

Domain-Invariant Component Analysis



Muandet, Balduzzi, and Schölkopf. Domain Generalisation via Invariant Feature Representation, ICML (2013)

Learning-Theoretic Bound

- After learning representation, we minimise **average accuracy** across domains
- With probability at least $1-\delta$,

$$\sup_{\|f\|_{\mathcal{H}}\leq 1} \left| \mathbb{E}^{*}_{\mathscr{P}} \mathbb{E}_{\mathbb{P}} \ell(f(\tilde{X}_{ij}), Y_i) - \mathbb{E}_{\hat{\mathbb{P}}} \ell(f(\tilde{X}_{ij}), Y_i) \right|^2$$

$$\leq c_1 \cdot \underbrace{\mathbb{V}_{\mathcal{H}}(\mathbb{P}^1, \mathbb{P}^2, \ldots, \mathbb{P}^N)}_{distributional\ variance} + \underbrace{c_2 \frac{N \cdot (\log \delta^{-1} + 2 \log N)}{n} + c_3 \frac{\log \delta^{-1}}{N} + \frac{c_4}{N}}_{vanish\ as\ N, n \to \infty}$$

Muandet, Balduzzi, and Schölkopf. Domain Generalisation via Invariant Feature Representation, ICML (2013)

Well-Known Methods for DG

- 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]

$$h^* = \min_{h \in H} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{(X,Y)\sim P_i}[\ell(Y, h(X))]$$

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

Well-Known Methods for DG

- Domain-Adversarial Training of Neural Networks [Ganin et al. 2
- Causal Invariant Prediction (CIP) [Peters et al., 2016; H 018]
- Invariant Risk Minimisation (IRM) [Arjovsky e
- Distributional Robust Optimisatio 2020]
- Probable Domain Genera ., 2022]

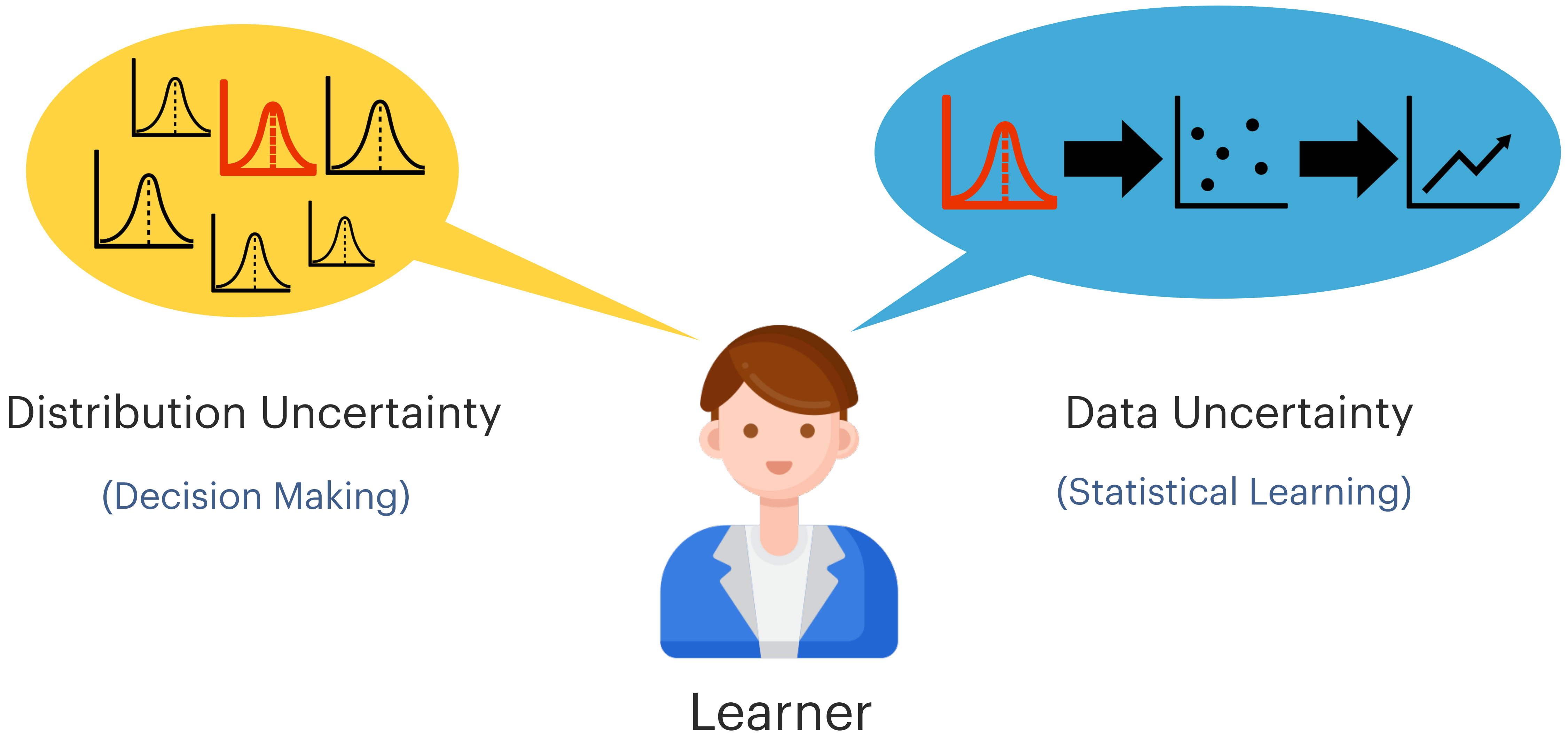
Precise Learning

$$h^* = \min_{h \in H} \frac{1}{N} \sum_{i=1}^{N} \ldots (Y, h(X))]$$

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

(Im)precise Generalisation

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



Institutional Separation

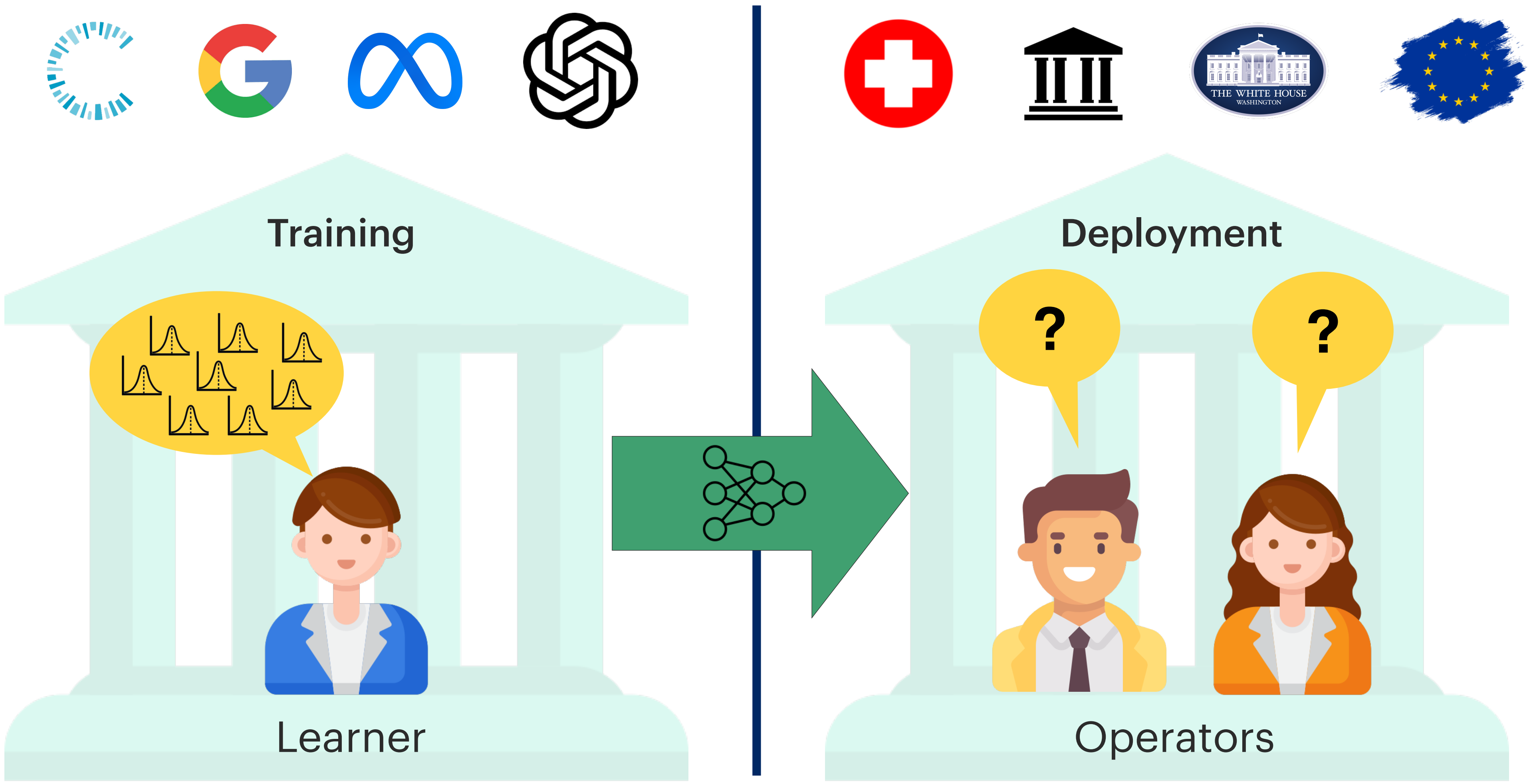
- **Precise learner** deals with two sources of uncertainties simultaneously.
    1. **The learner decides the notion of generalisation (pick a specific distribution)**
    2. The learner then conducts statistical learning to choose the best hypothesis

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

Institutional Separation

- **Precise learner** deals with two sources of uncertainties <u>simultaneously</u>.

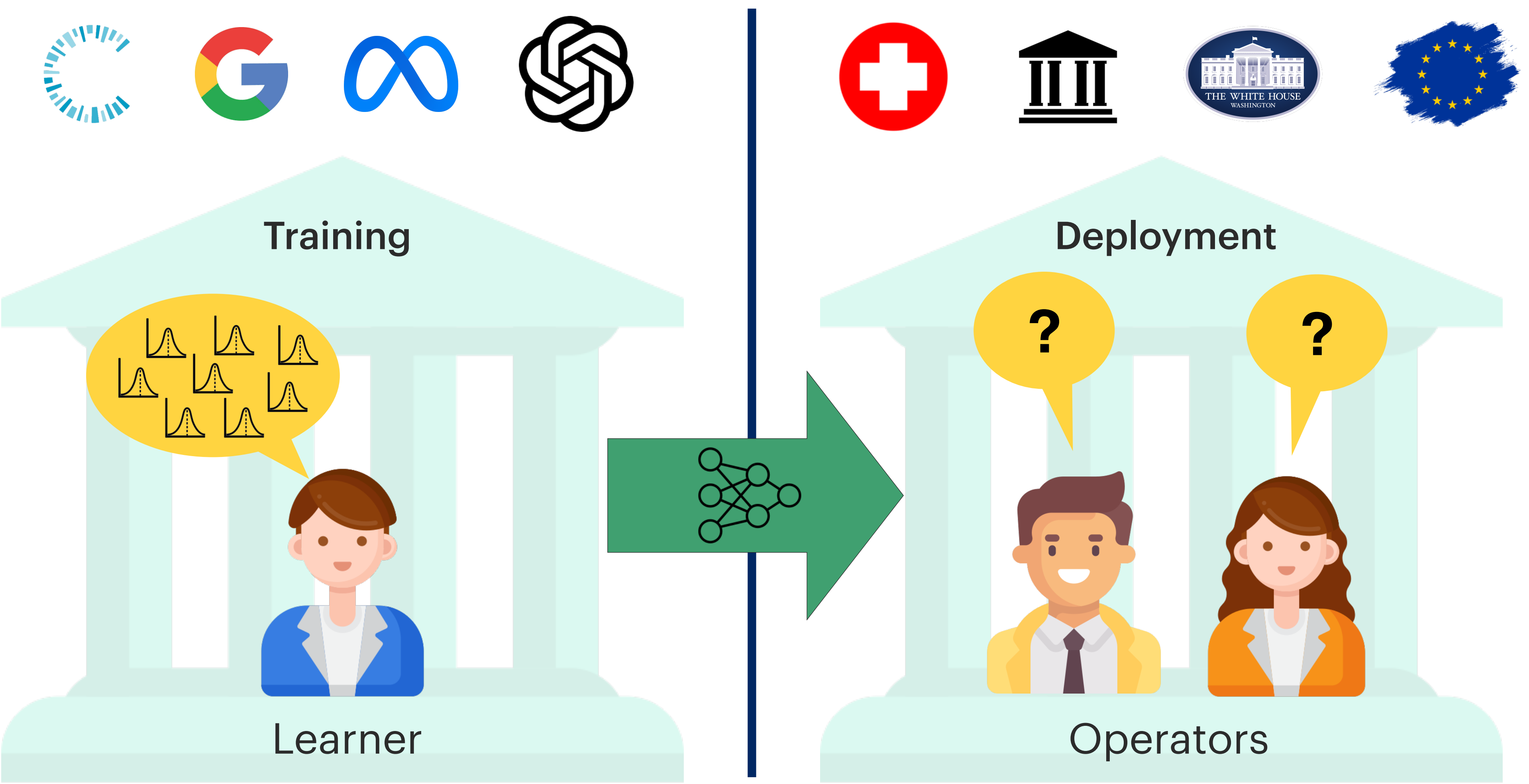
  1. **The learner decides the notion of generalisation (pick a specific distribution)**
  2. The learner then conducts statistical learning to choose the best hypothesis

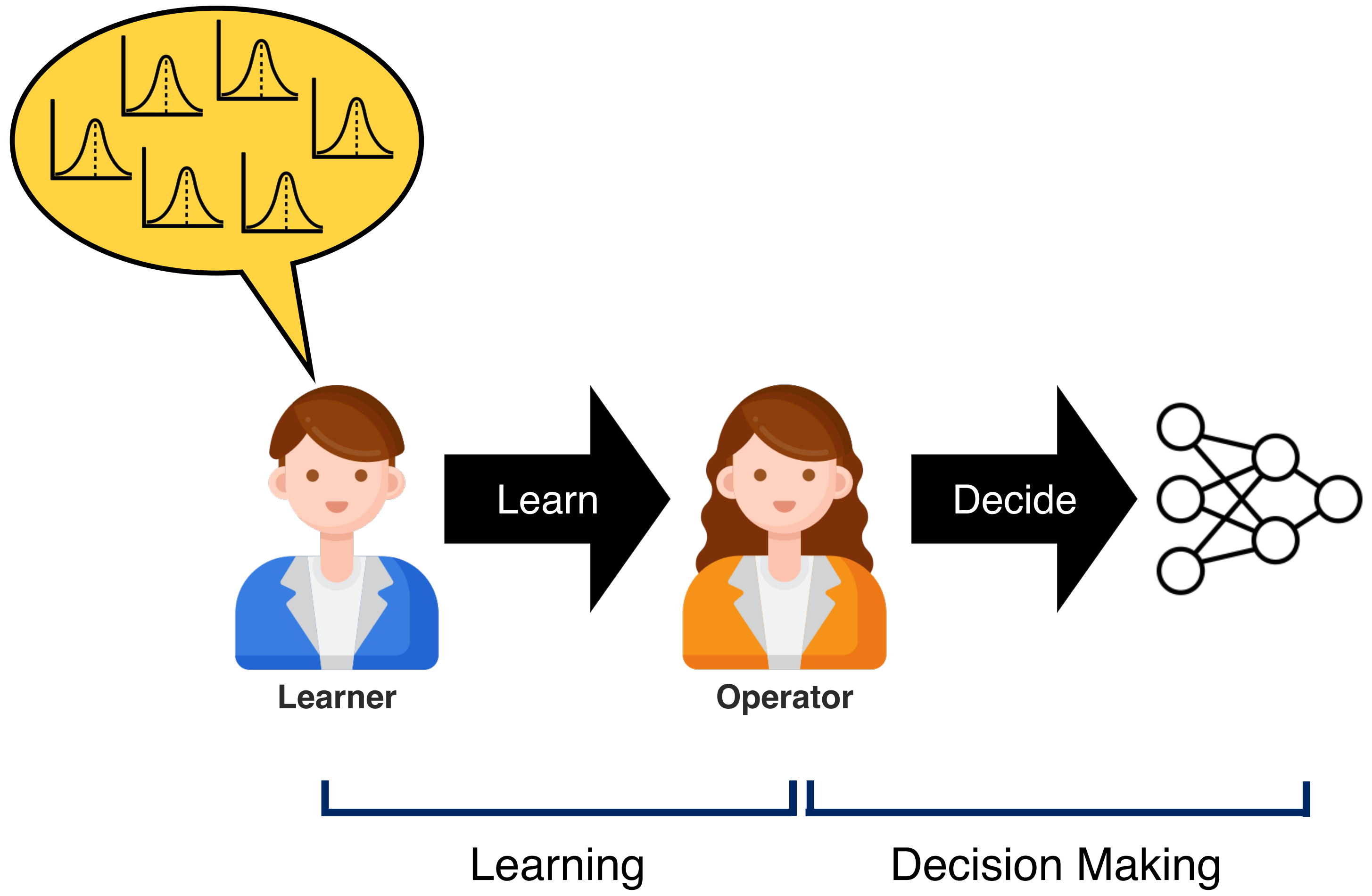


Manski, Charles F. Statistical Inference for Statistical Decisions. 2019

Institutional Separation

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



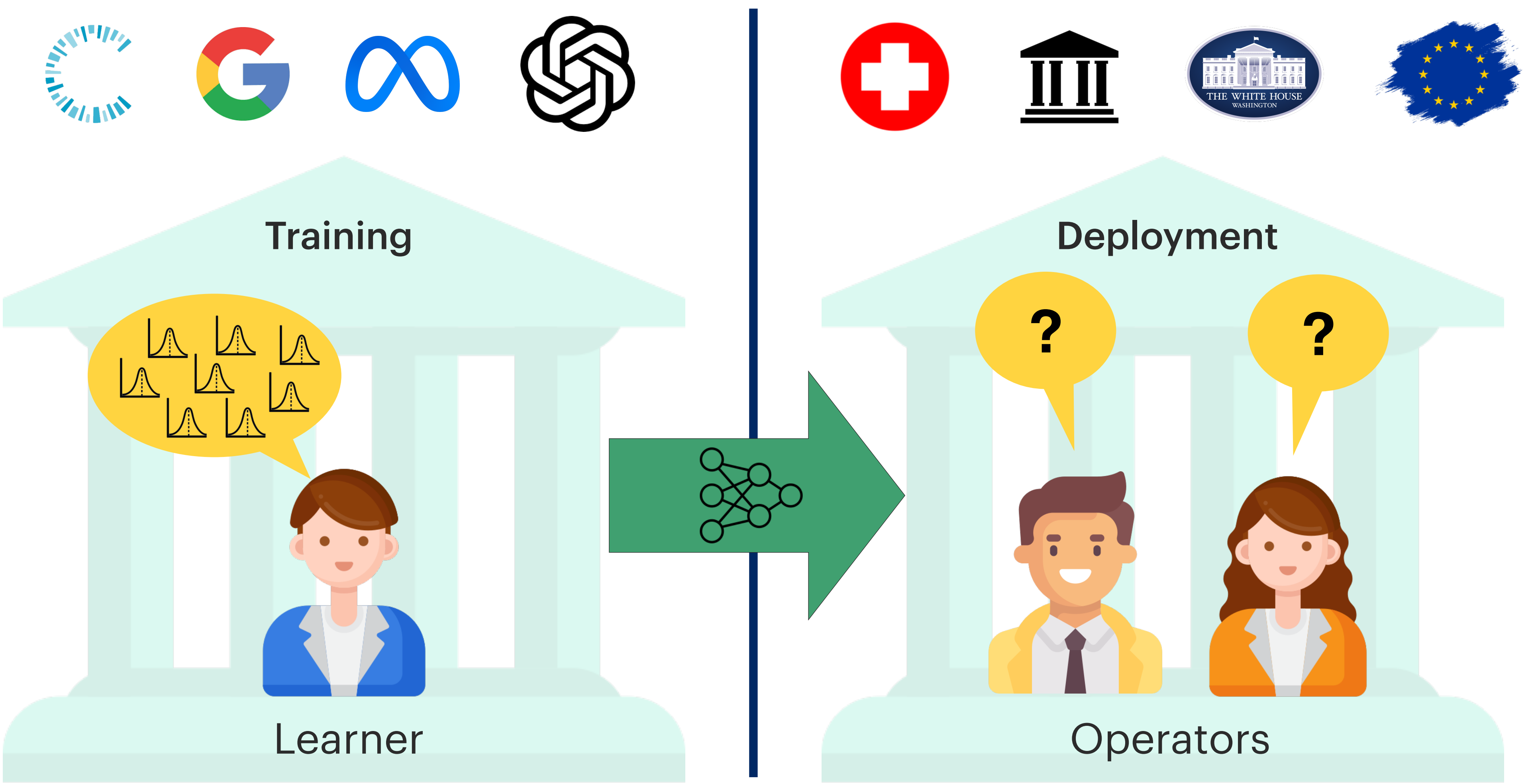


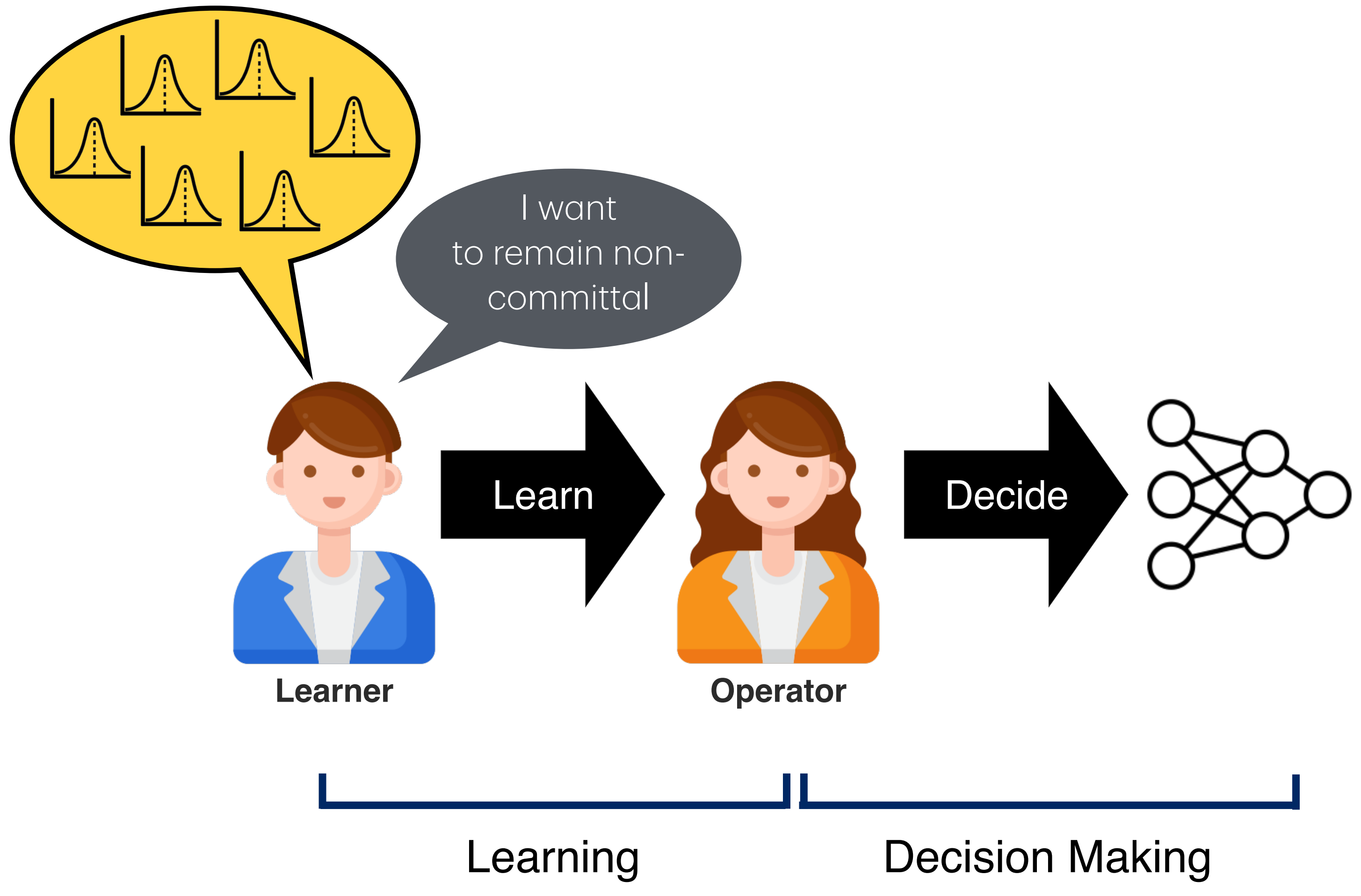
Manski, Charles F. Statistical Inference for Statistical Decisions. 2019

Institutional Separation

- **Precise learner** deals with two sources of uncertainties <u>simultaneously</u>.

  1. **The learner decides the notion of generalisation (pick a specific distribution)**
  2. The learner then conducts statistical learning to choose the best hypothesis

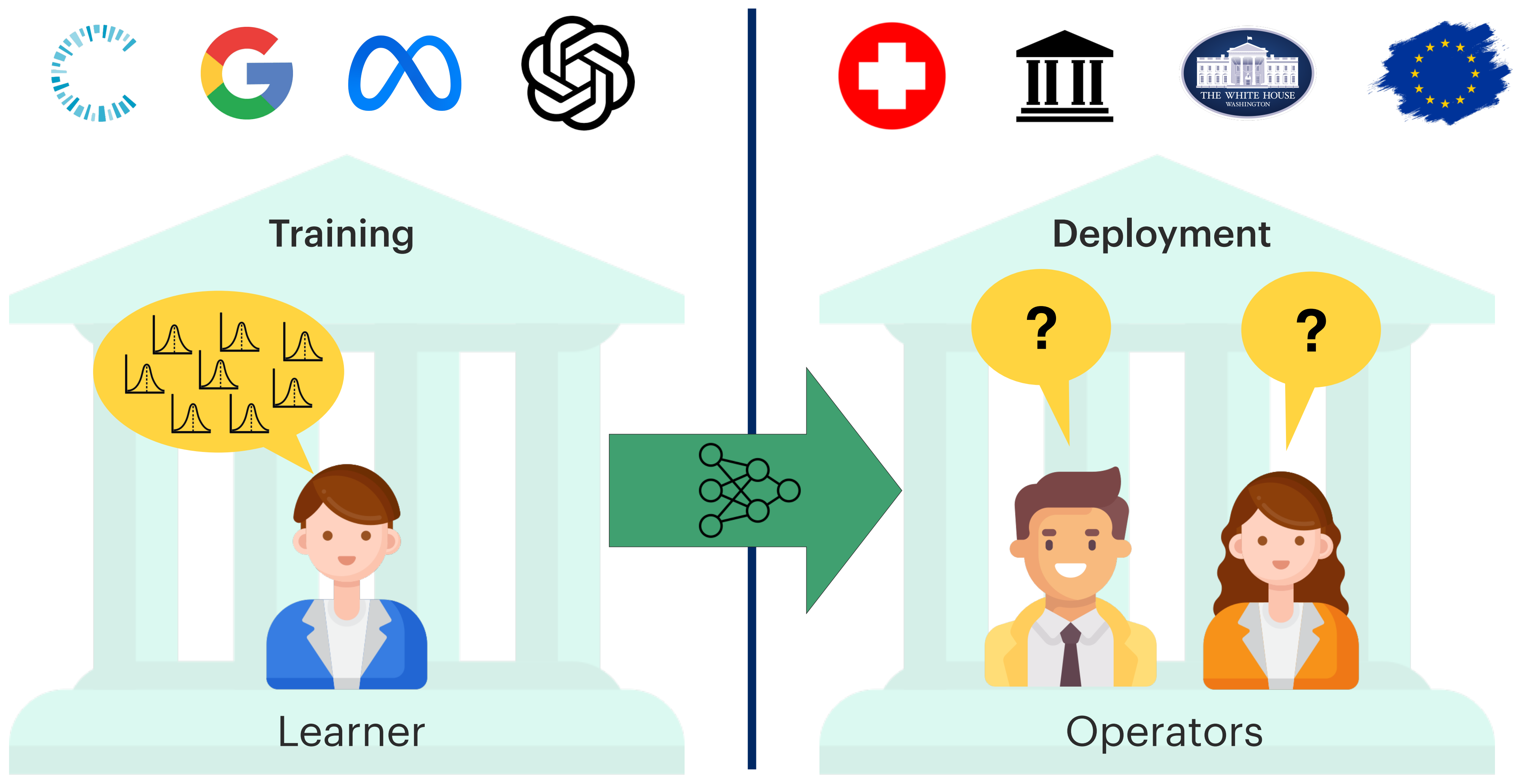


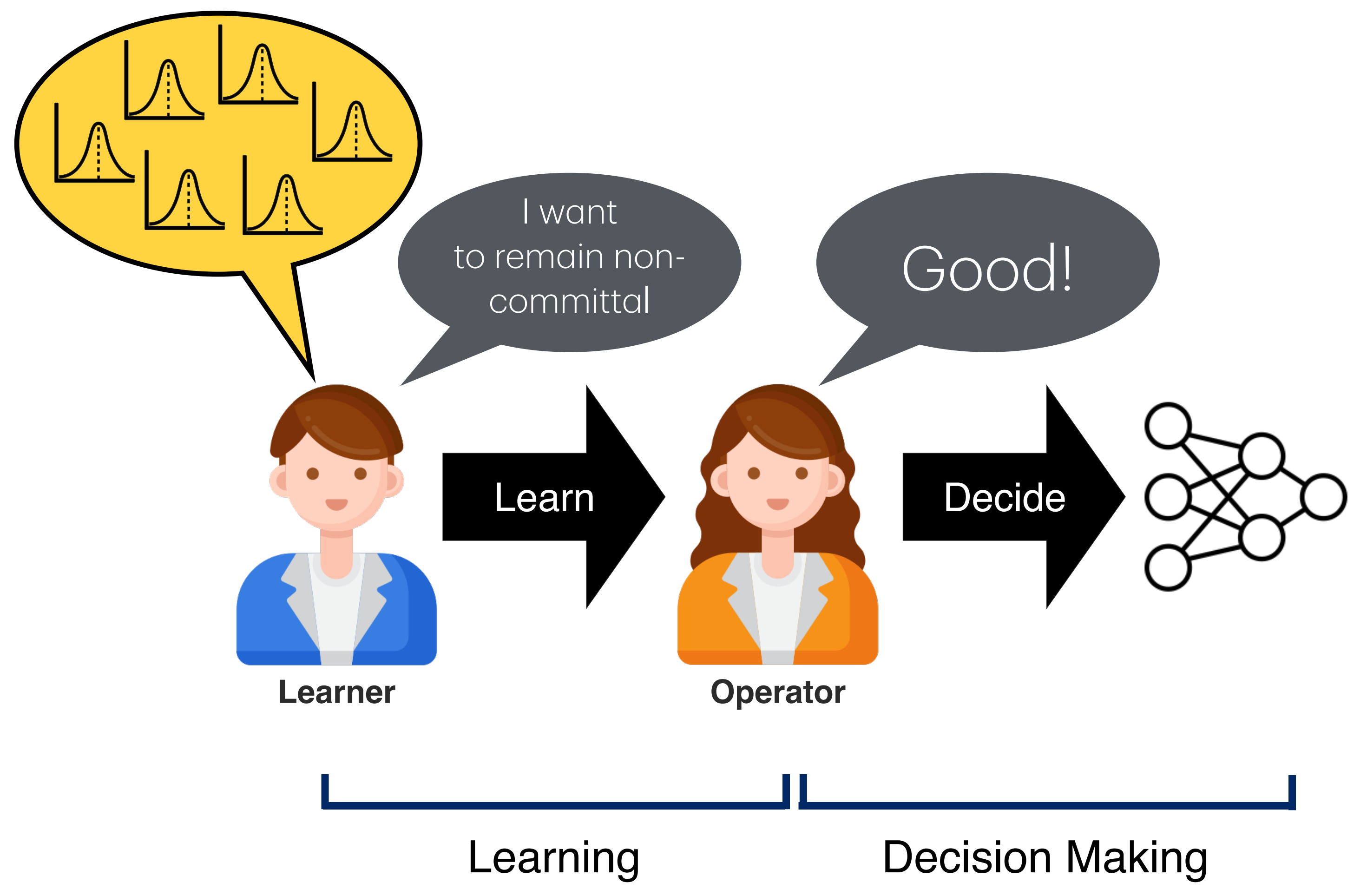


Manski, Charles F. Statistical Inference for Statistical Decisions. 2019

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

ICML International Conference On Machine Learning

Spotlight

Domain Generalisation via Imprecise Learning

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

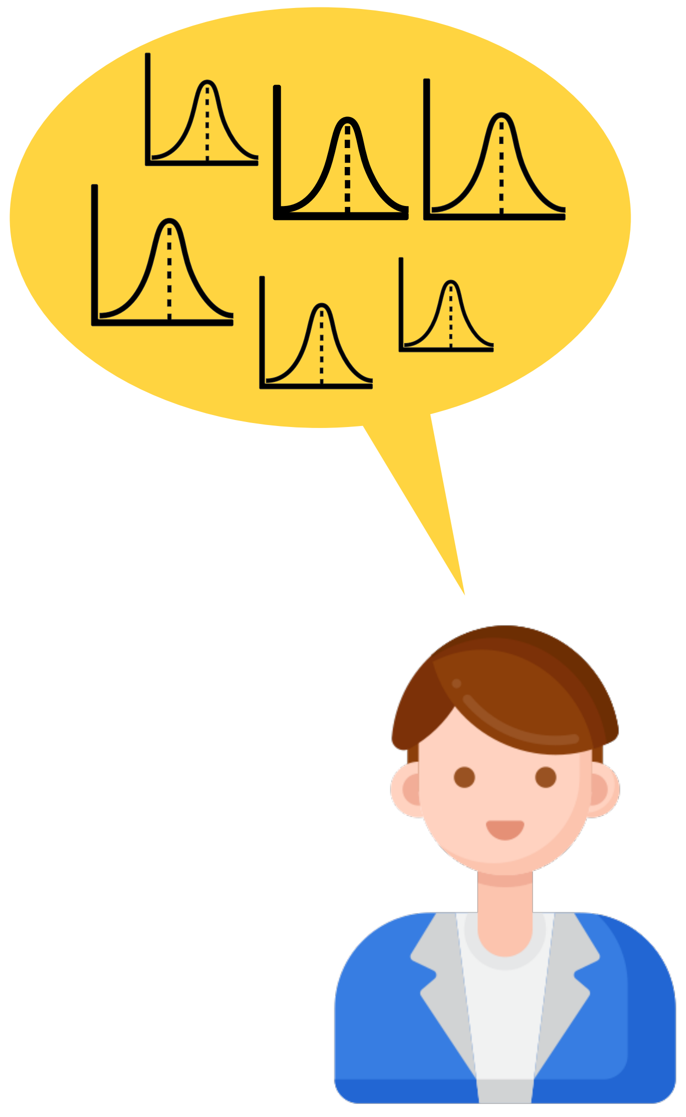
Abstract

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.

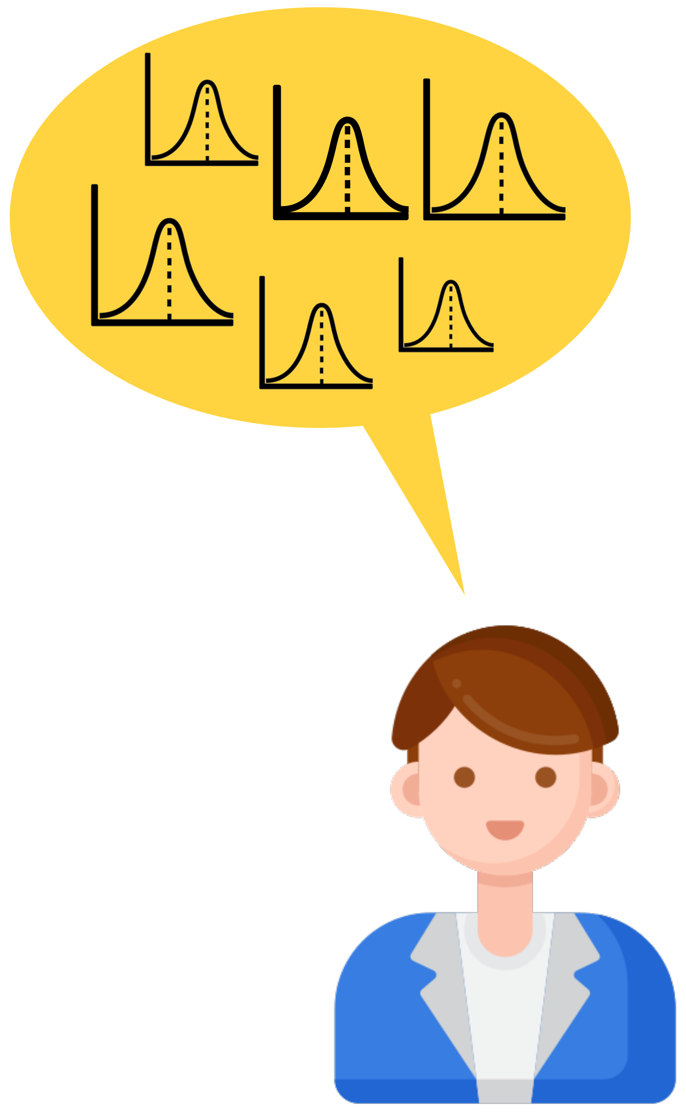
Despite notable achievements, these systems may catastrophically fail when operated on out-of-domain (OOD) data because theoretical guarantees for their generalisation hinge on the assumption of independent and identically distributed (IID) training and deployment data, with empirical

Problem Formulation



Imprecise Learner

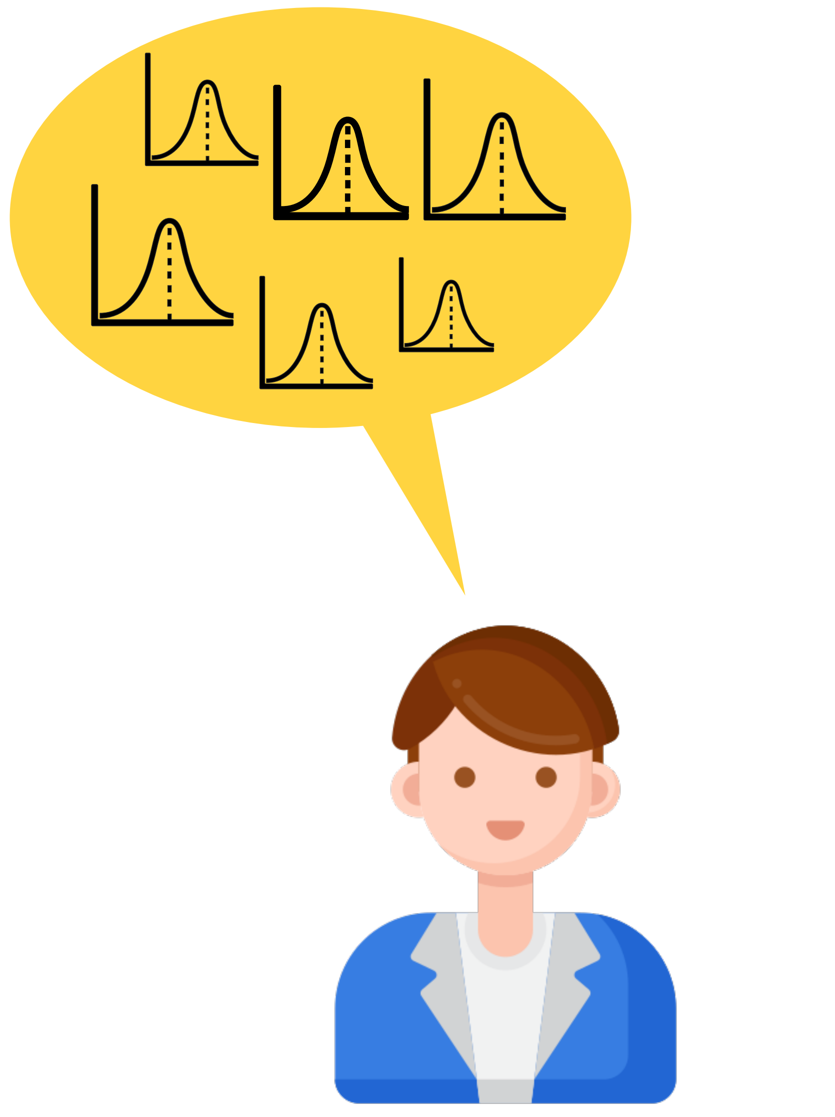
Problem Formulation



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

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

Problem Formulation

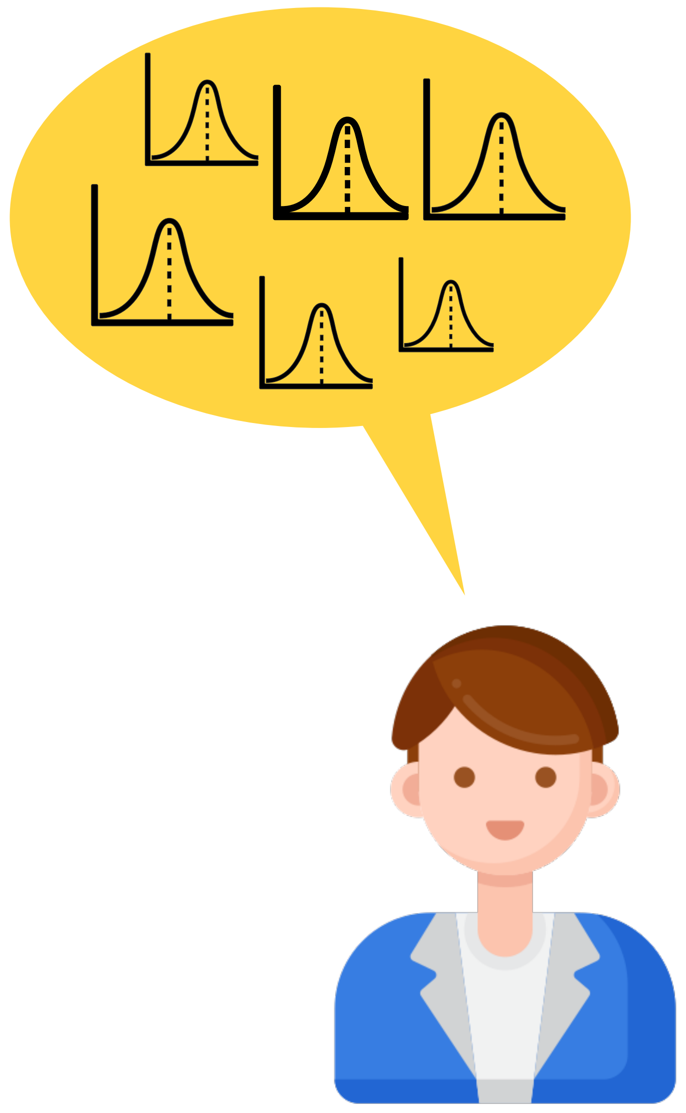


- A **risk profile** on n observed environments $P_1(X, Y), P_2(X, Y), \ldots, P_N(X, Y)$

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

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

Problem Formulation



Imprecise Learner

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

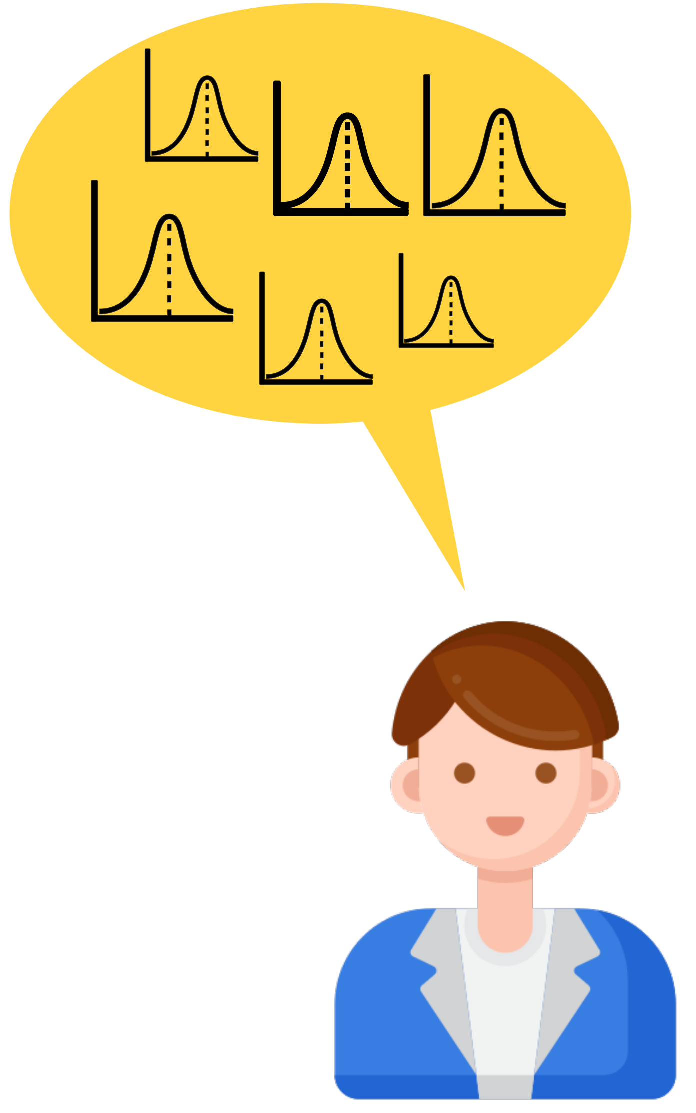
$$\mathbf{R}(f) := (R_1(f), \ldots, R_N(f)), \quad f \in H$$

- 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

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

Problem Formulation



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

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

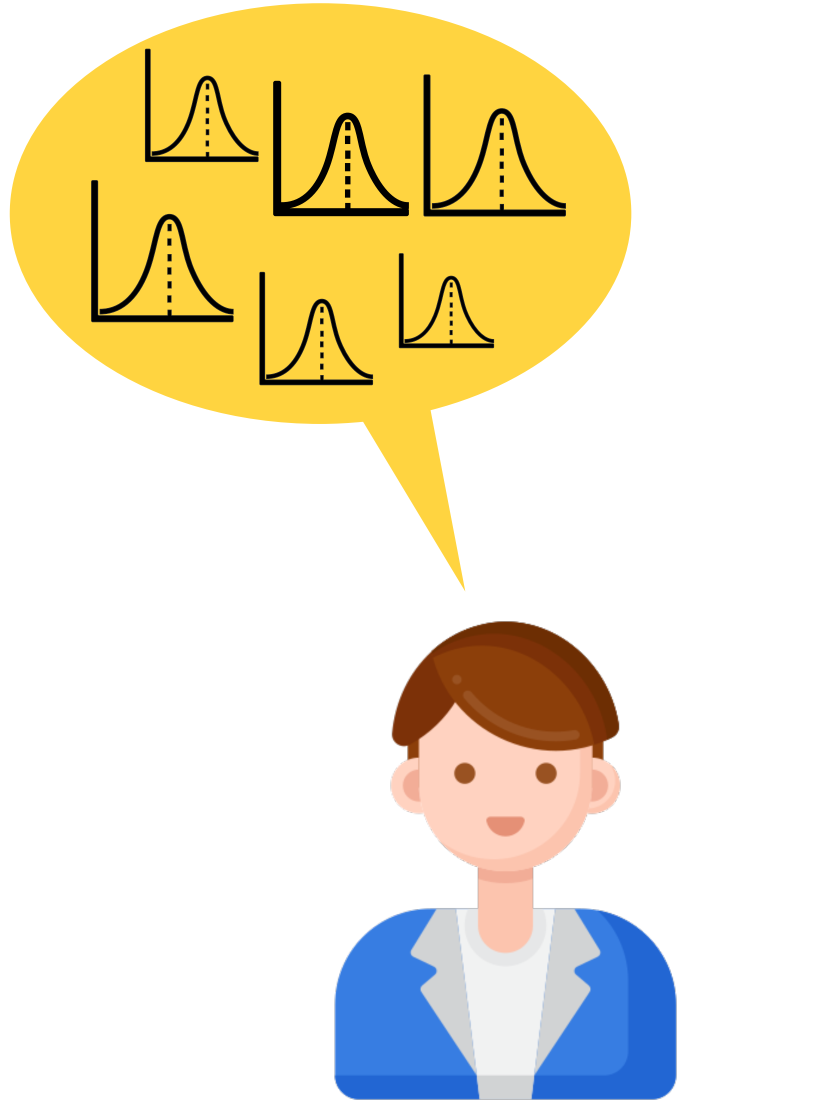
- An **aggregation function** $\rho_\lambda : L_2^N(H) \to L_2(H)$ for some $\lambda \in \Lambda$

- 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

Problem Formulation



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

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

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

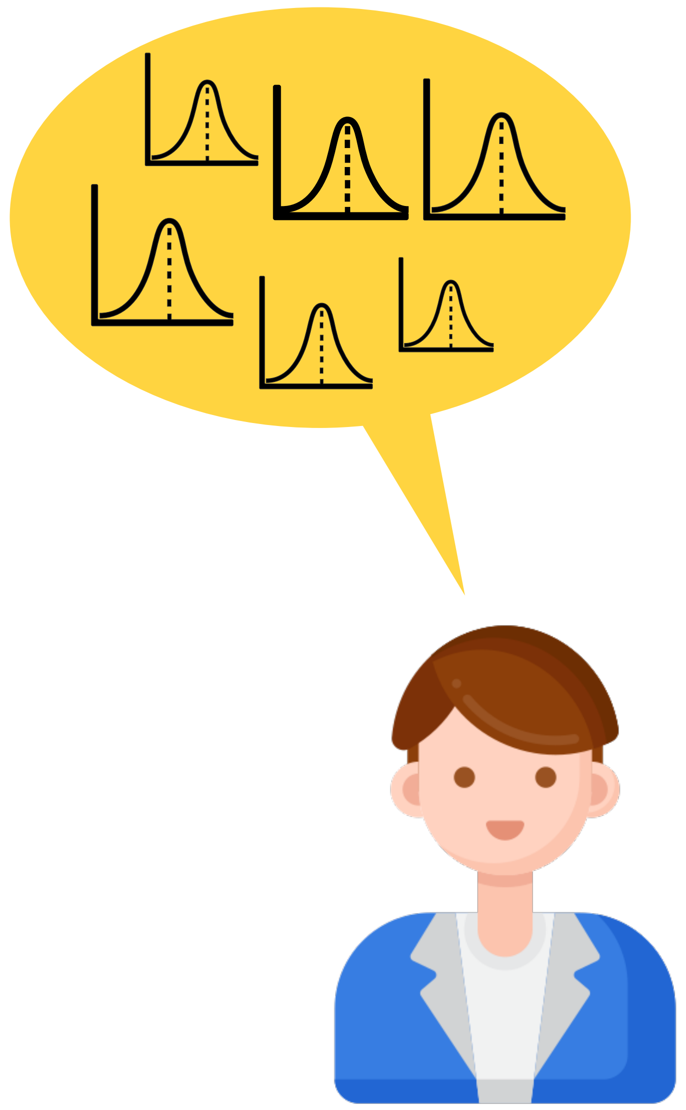
- For a fixed $\lambda \in \Lambda$, we can learn from H by minimising an aggregated risk

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

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

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

Problem Formulation



Imprecise Learner

- 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

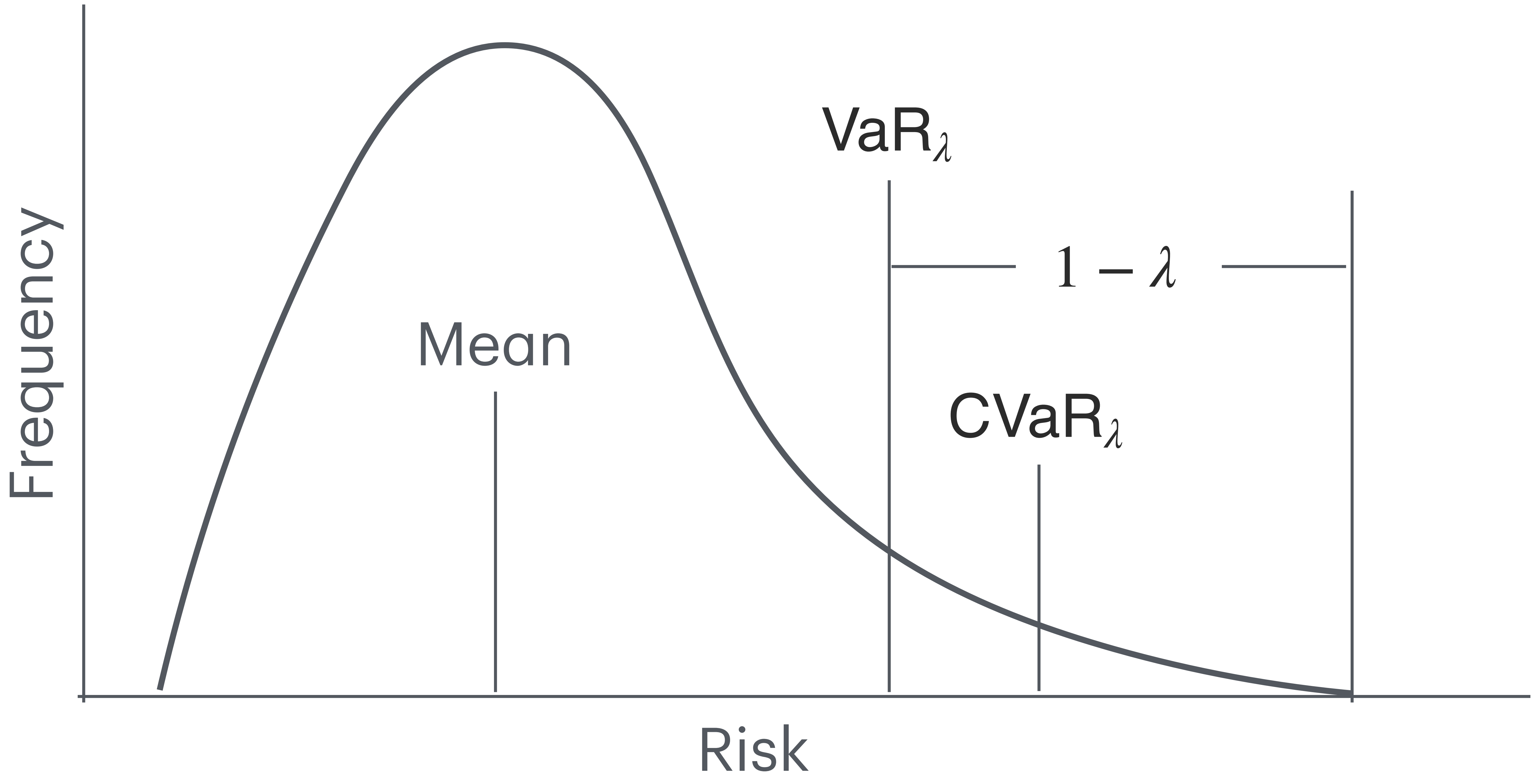
$$f_\lambda^* = \arg\min_{f \in H} \rho_\lambda[\mathbf{R}](f), \quad \lambda \in \Lambda$$

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

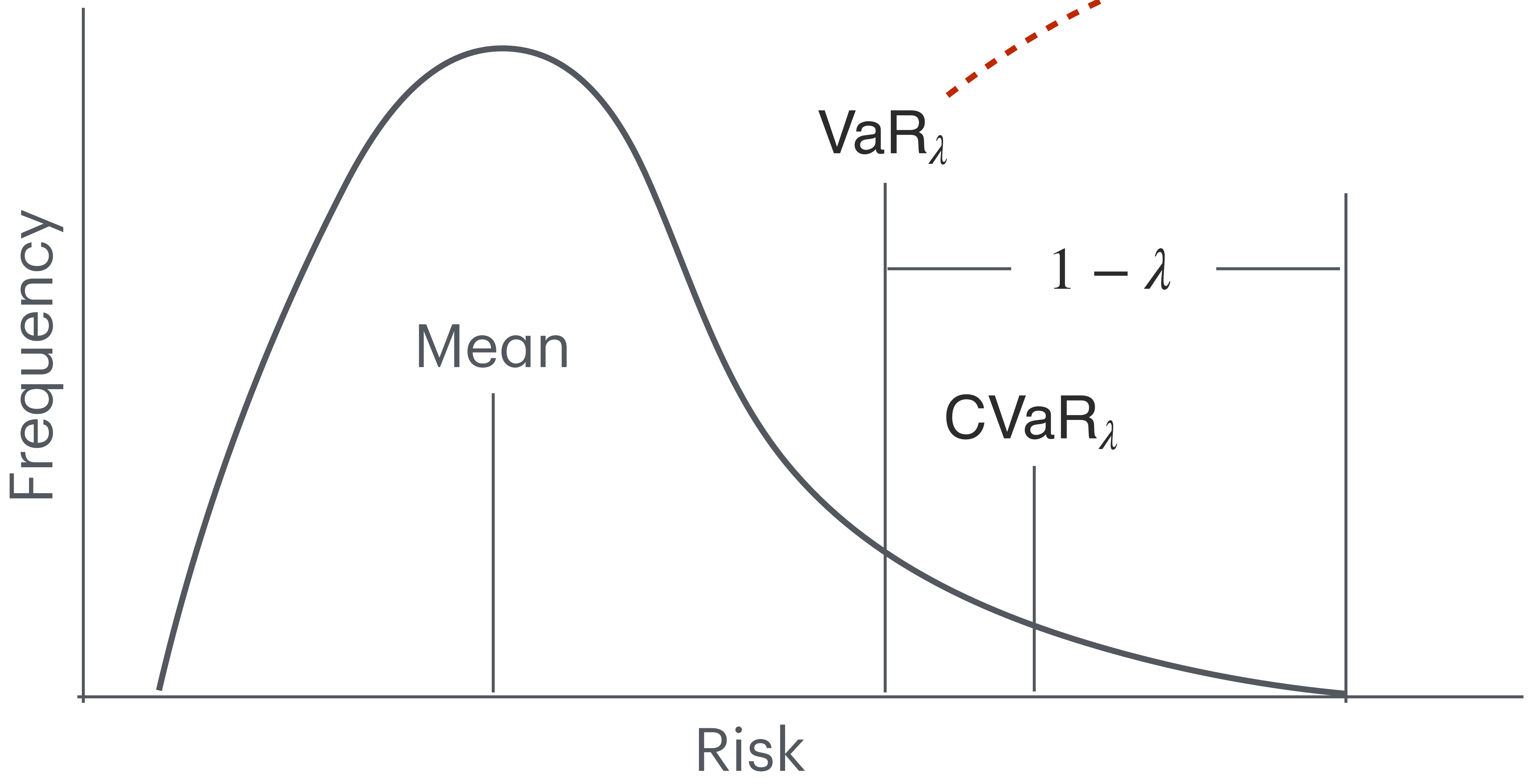
$$h_\theta^*(\cdot, \lambda) = f_\lambda^* = \arg\min_{f \in H} \rho_\lambda[\mathbf{R}](f), \quad \lambda \in \Lambda$$

Traverse credal set

Conditional Value at Risk (CVaR)

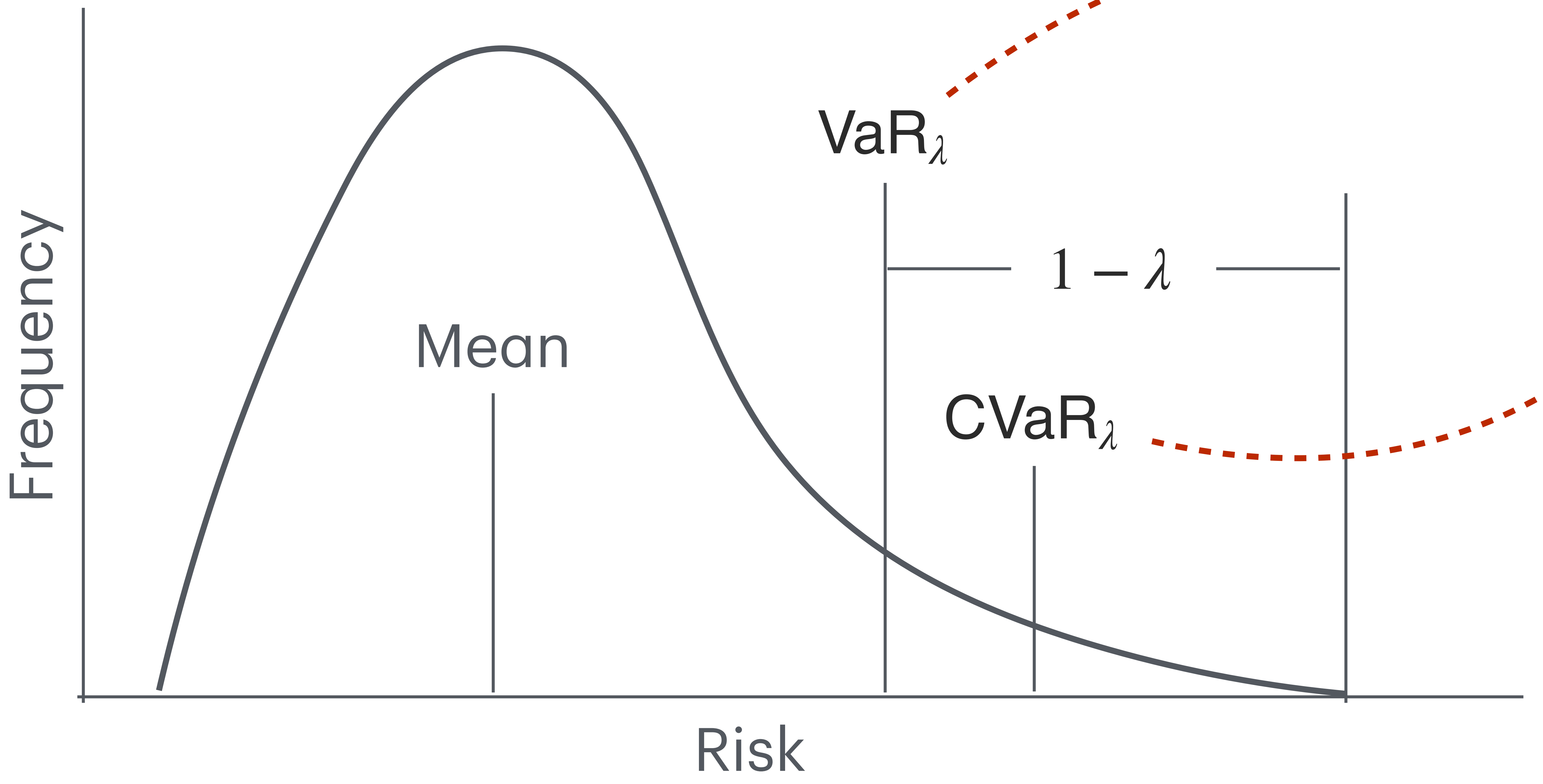


Conditional Value at Risk (CVaR)



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

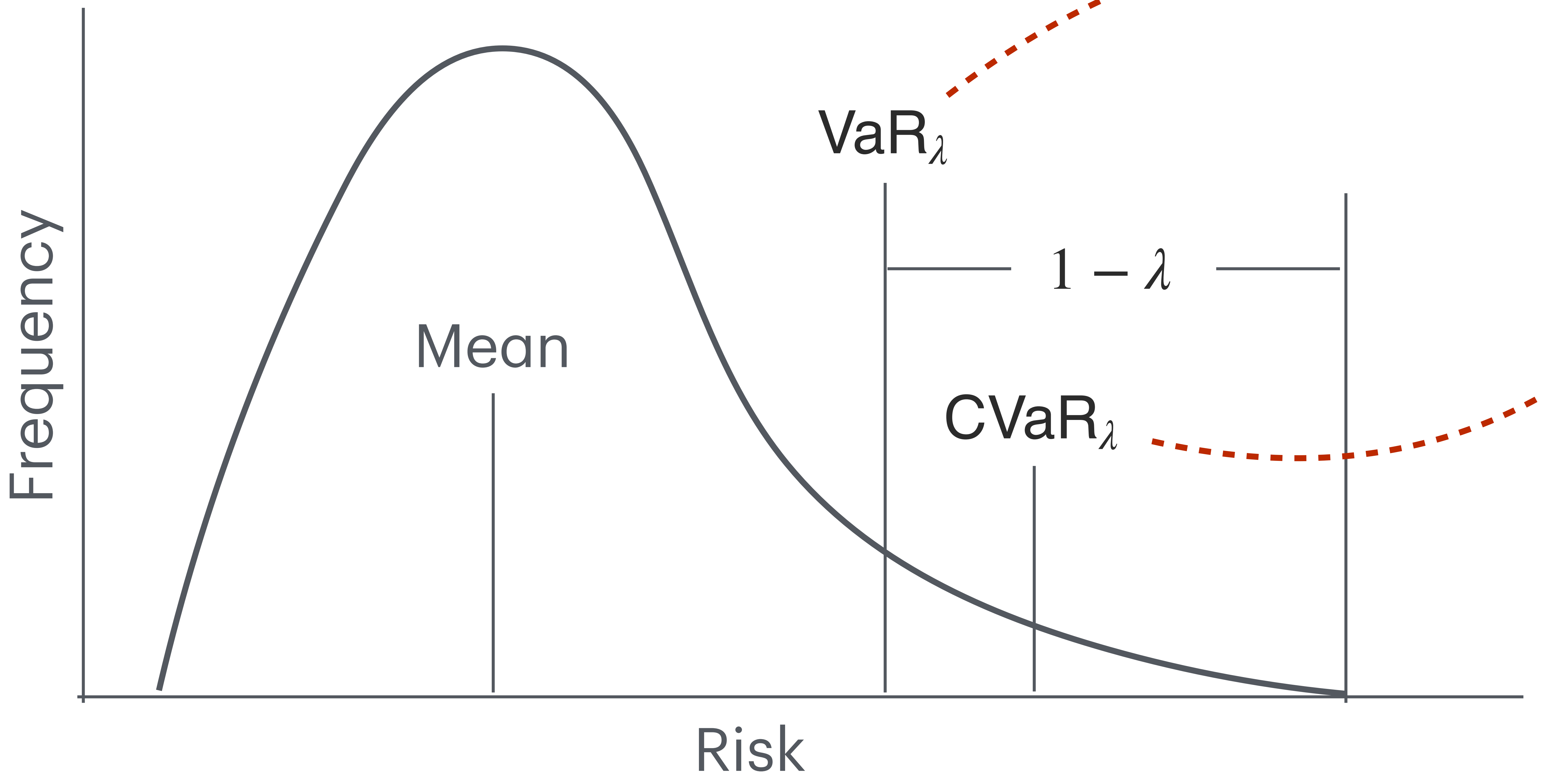
Conditional Value at Risk (CVaR)



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

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

Conditional Value at Risk (CVaR)

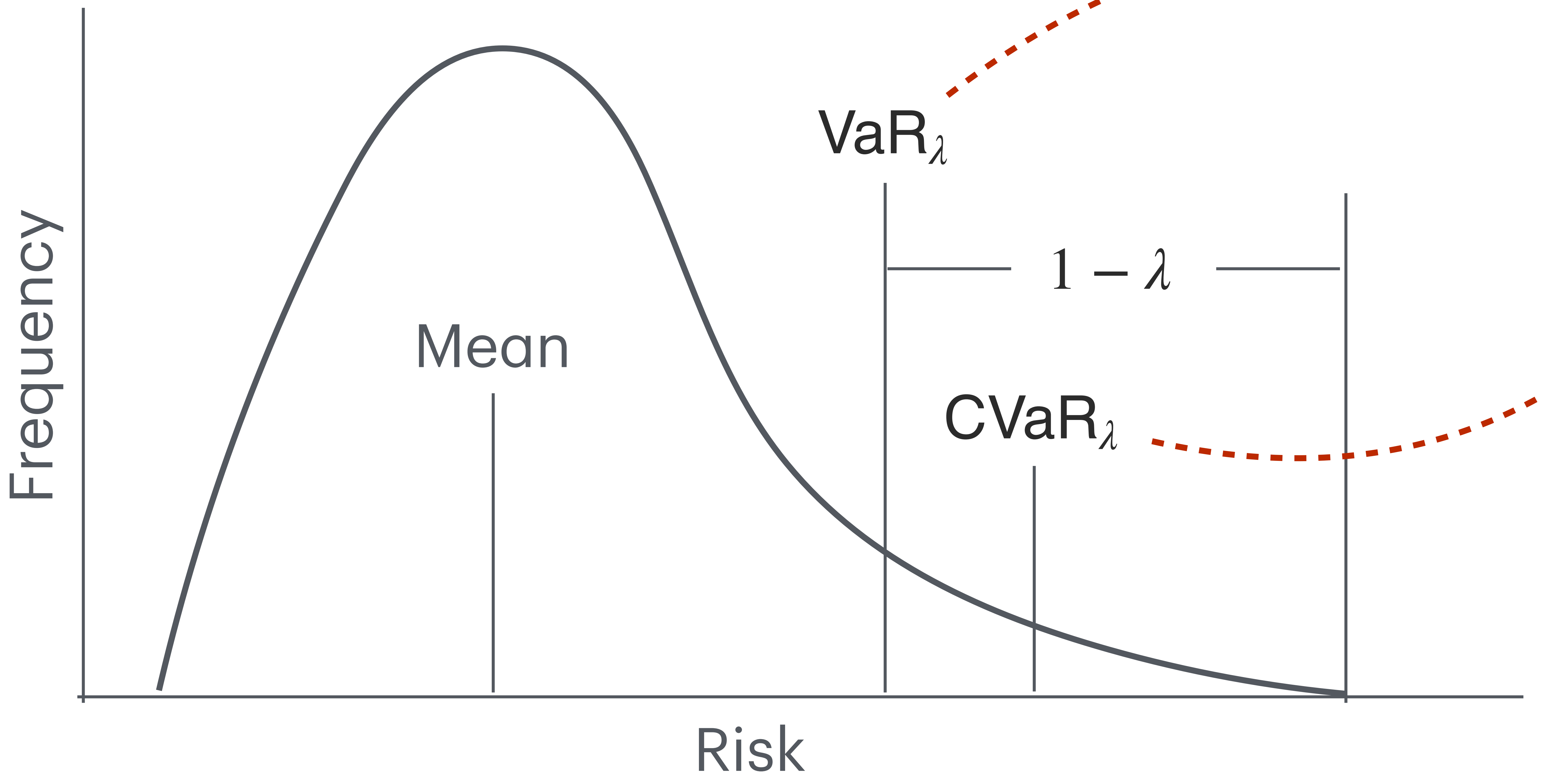


$$\mathrm{VaR}_\lambda(R) = \min\{r \mid F_R(r) \geq \lambda\}$$

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

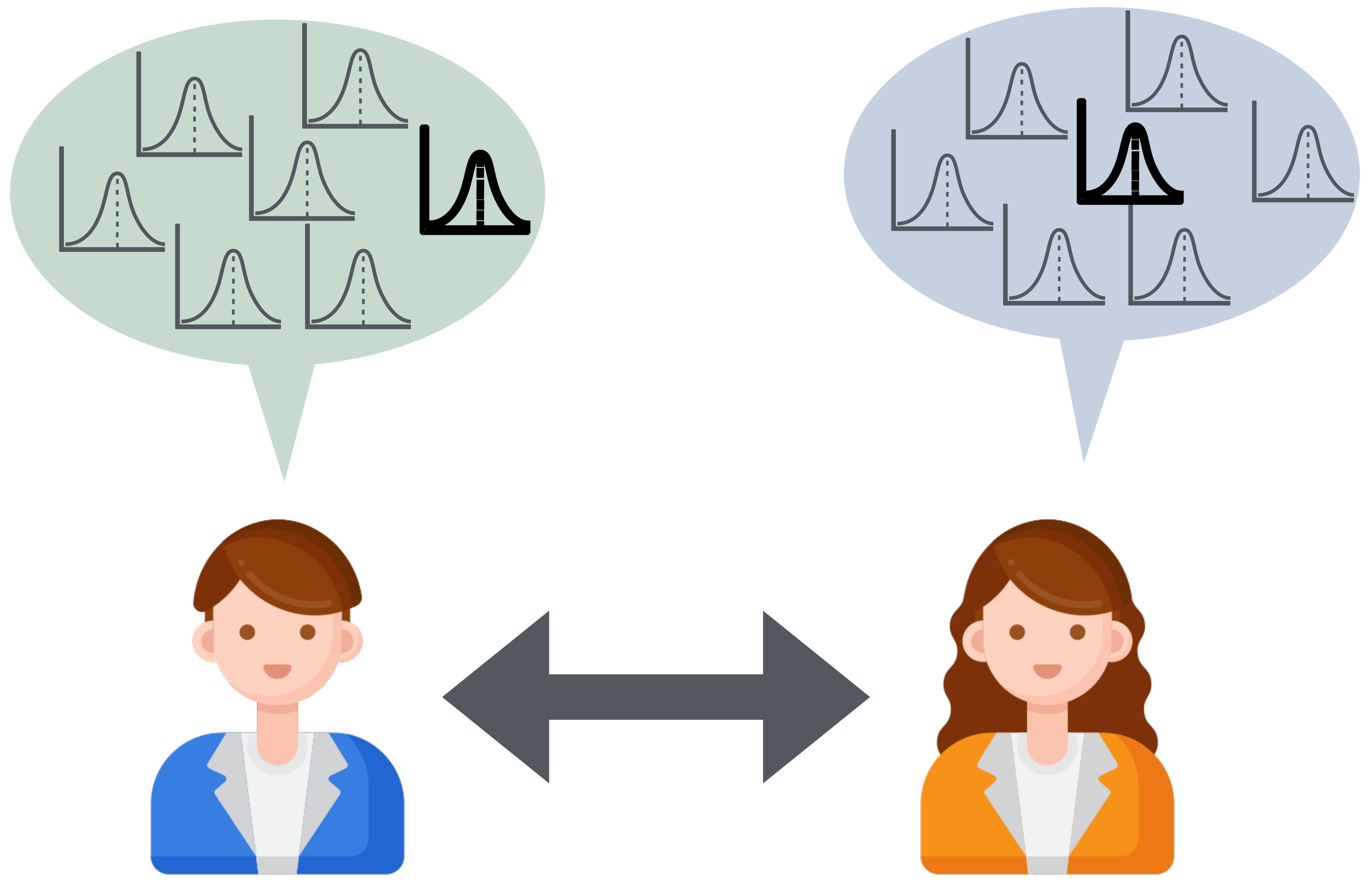
- **Interpretation:** λ is the level of risk aversion (Robey et al., 2022; Eastwood et al., 2022a; Li et al., 2023)

Conditional Value at Risk (CVaR)



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

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



Worst Case $\lambda \to 1$

Average Case $\lambda = 0$

- **Interpretation:** λ 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_θ dominates h'_θ if for all $\lambda \in \Lambda$,

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

C-Pareto Optimality

C-Pareto Optimality

h_θ dominates h'_θ if for all $\lambda \in \Lambda$,

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

$\Longleftrightarrow$

Scalarised Objective

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

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

C-Pareto Optimality

C-Pareto Optimality

h_θ dominates h'_θ 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**: $\theta_t \leftarrow \theta_{t-1} - \eta \nabla_\theta \hat{J}_{Q_t}(h_\theta)$:

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

C-Pareto Optimality

C-Pareto Optimality

h_θ dominates h'_θ 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**: $\theta_t \leftarrow \theta_{t-1} - \eta \nabla_\theta \hat{J}_{Q_t}(h_\theta)$:

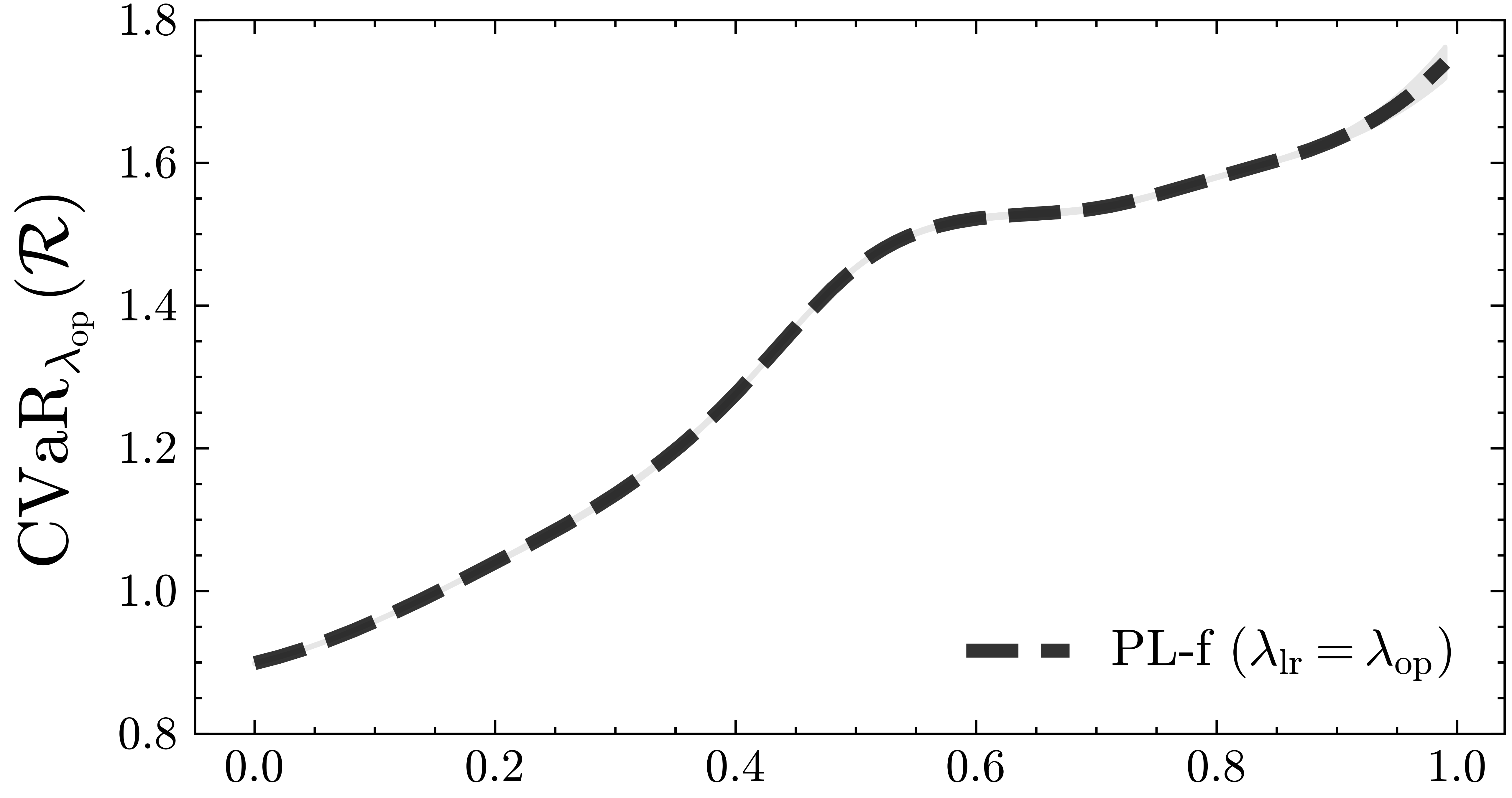
$$Q_t \in \arg\min_{Q \in \Delta(\Lambda)} \left\| \nabla_{\theta_{t-1}} \hat{J}_Q\left(h_{\theta_{t-1}}\right) \right\|_2, \quad \hat{J}_Q(h_\theta) := \frac{1}{N}\sum_{i=1}^{N} \rho_{\lambda_i}[\mathbf{R}](h_\theta(\,\cdot\,,\lambda_i))$$

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

Precise vs Imprecise Learning

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

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

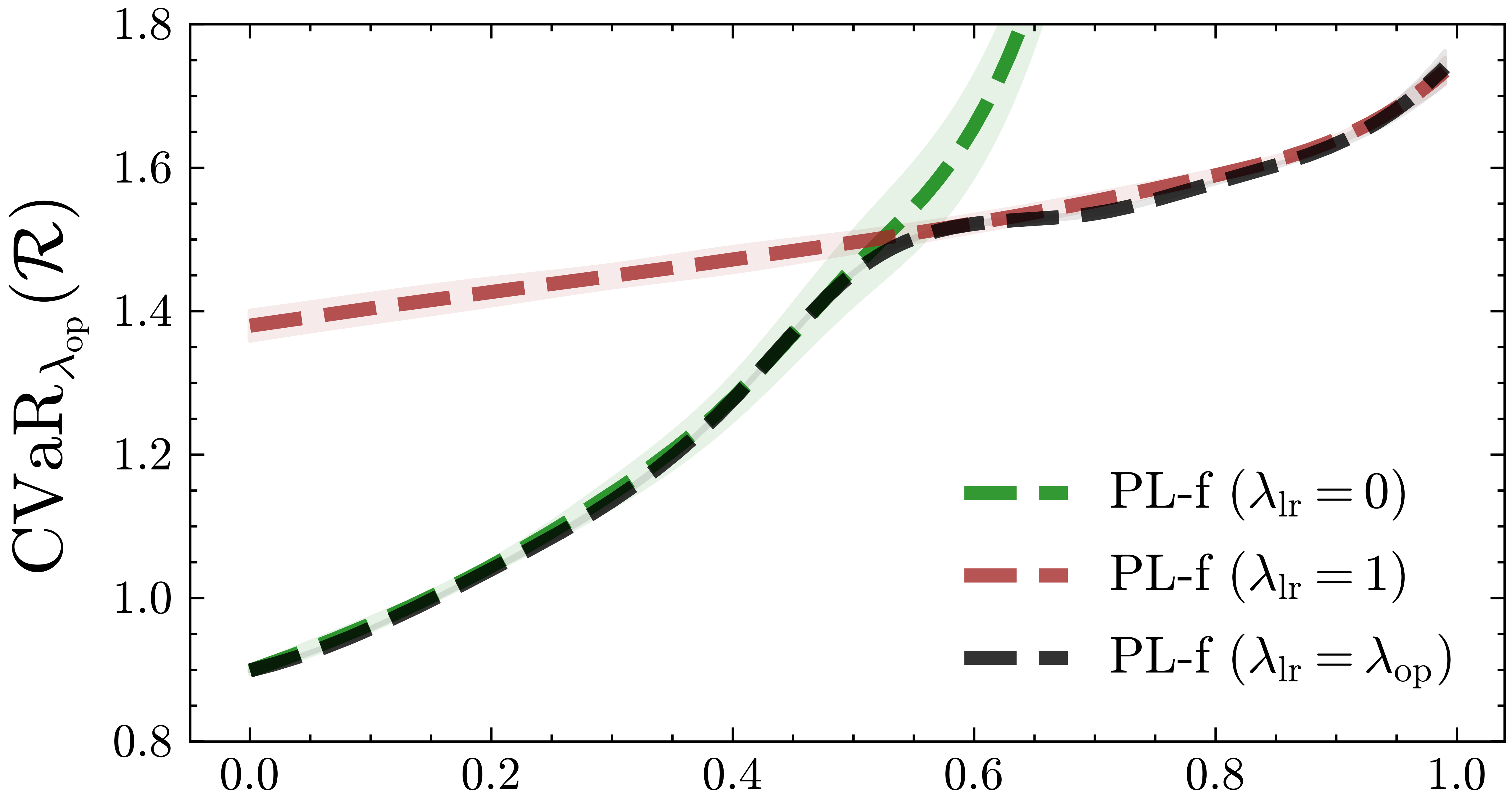


Risk neutral ← λ_{op} → Risk averse

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_{train} = 250, n_{test} = 250, sample size = 100 from each domain

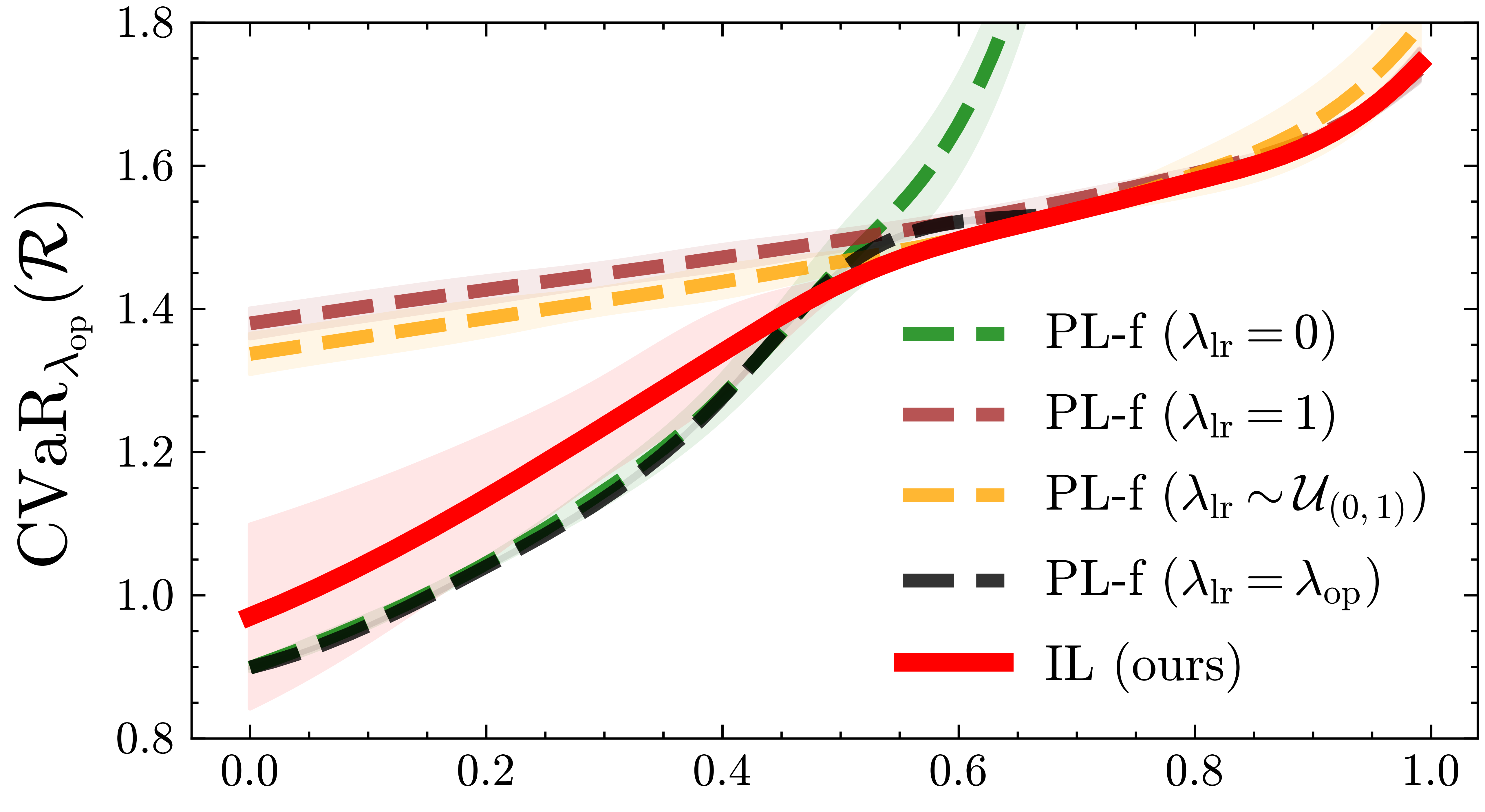


Risk neutral ← λ_{op} → Risk averse

Precise vs Imprecise Learning

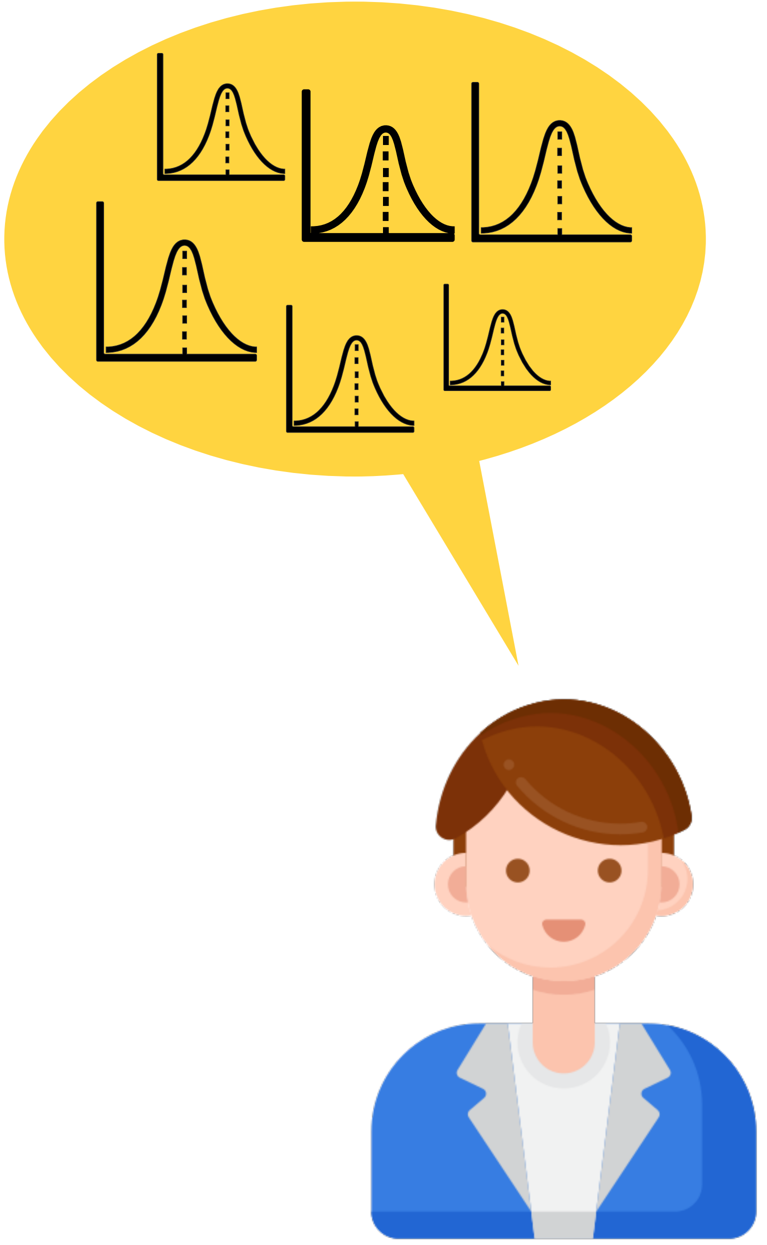
$$Y_d = \theta_d X + \varepsilon,\ X \sim \mathcal{N}(1{,}0.5),\ \epsilon \sim \mathcal{N}(0{,}0.1),\ \theta_d \sim \mathcal{U}(1{,}1.1) \text{ or } \mathcal{U}(-1.1, -1)$$

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



Risk neutral ← λ_{op} → Risk averse

Related Work



Imprecise Learner

Credal Learning Theory

Michele Caprio
Department of Computer Science
University of Manchester, Manchester, UK
michele.caprio@manchester.ac.uk

Maryam Sultana **Eleni G. Elia** **Fabio Cuzzolin**
School of Engineering Computing & Mathematics
Oxford Brookes University, Oxford, UK
{msultana,eelia,fabio.cuzzolin}@brookes.ac.uk

Abstract

Statistical learning theory is the foundation of machine learning, providing theoretical bounds for the risk of models learned from a (single) training set, assumed to issue from an unknown probability distribution. In actual deployment, however, the data distribution may (and often does) vary, causing domain adaptation/generalization issues. In this paper we lay the foundations for a 'credal' theory of learning, using convex sets of probabilities (credal sets) to model the variability in the data-generating distribution. Such credal sets, we argue, may be inferred from a finite sample of training sets. Bounds are derived for the case of finite hypotheses spaces (both assuming realizability or not), as well as infinite model spaces, which directly generalize classical results.

NeurIPS 2024

Learning with Ambiguity

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

Learning with Ambiguity

- OOD generalisation is learning with ambiguity

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

Learning with Ambiguity

- OOD generalisation is learning with ambiguity
- Subjectivity in fairness, interpretability, robustness, trustworthiness, and privacy creates learning ambiguity.

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

Learning with Ambiguity

- OOD generalisation is learning with ambiguity
- Subjectivity in fairness, interpretability, robustness, trustworthiness, and privacy creates learning ambiguity.
- **Institutional separation** changes the training pipeline
    - Foundation model: pre-training + fine-tuning
    - LLM alignment

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

Learning with Ambiguity

- OOD generalisation is learning with ambiguity
- Subjectivity in fairness, interpretability, robustness, trustworthiness, and privacy creates learning ambiguity.
- **Institutional separation** changes the training pipeline
    - Foundation model: pre-training + fine-tuning
    - LLM alignment
- Principal-agent model (Holmström, 1979)

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

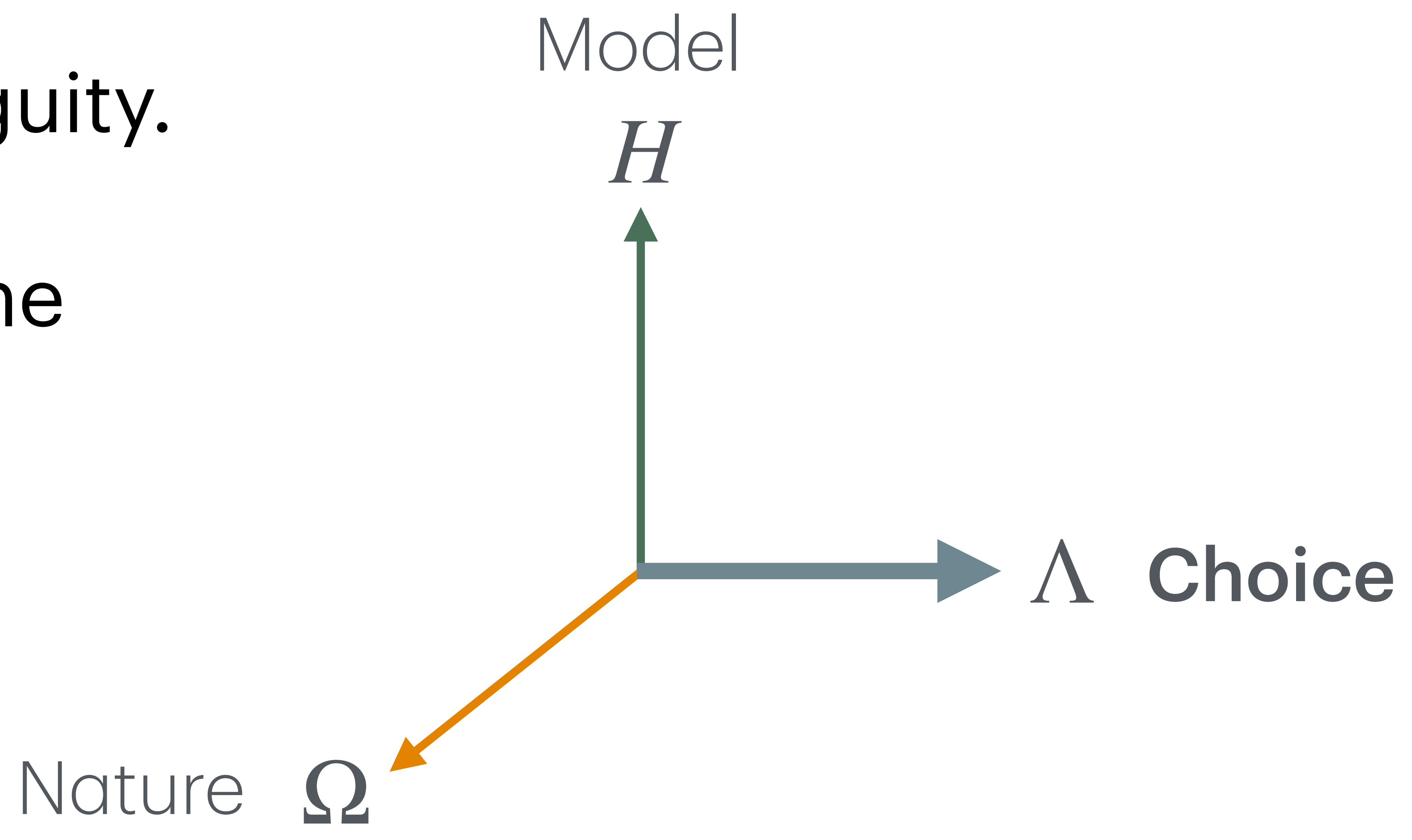
Learning with Ambiguity

- OOD generalisation is learning with ambiguity
- Subjectivity in fairness, interpretability, robustness, trustworthiness, and privacy creates learning ambiguity.
- **Institutional separation** changes the training pipeline
    - Foundation model: pre-training + fine-tuning
    - LLM alignment
- 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.

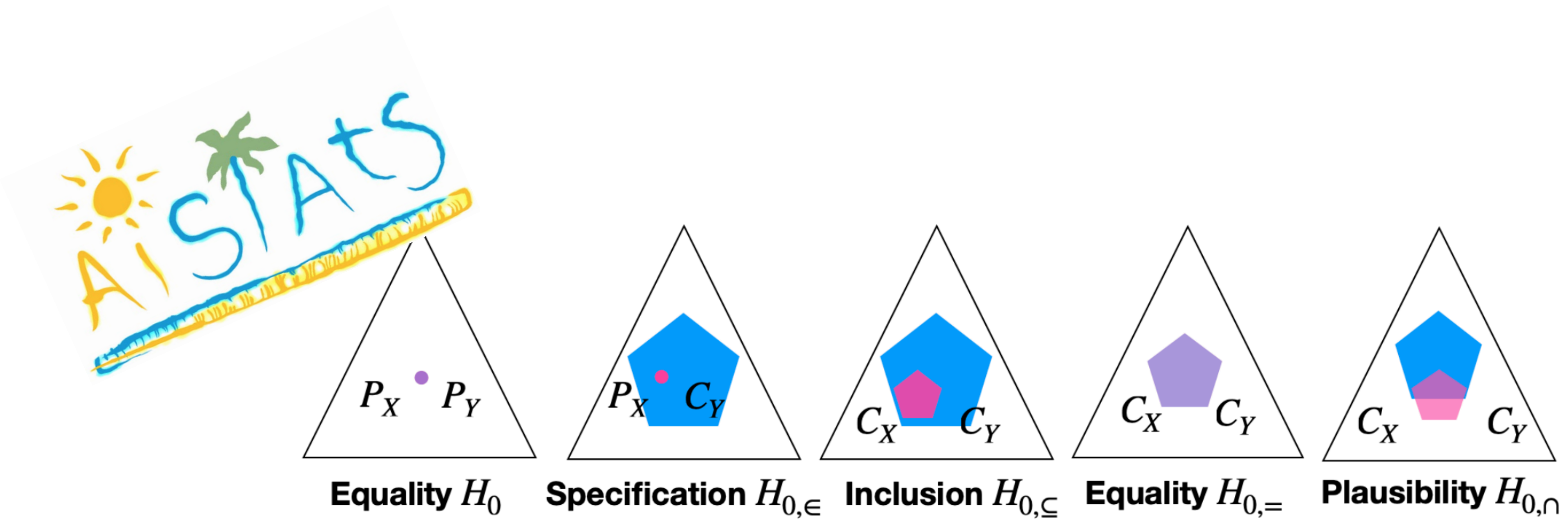
Learning with Ambiguity

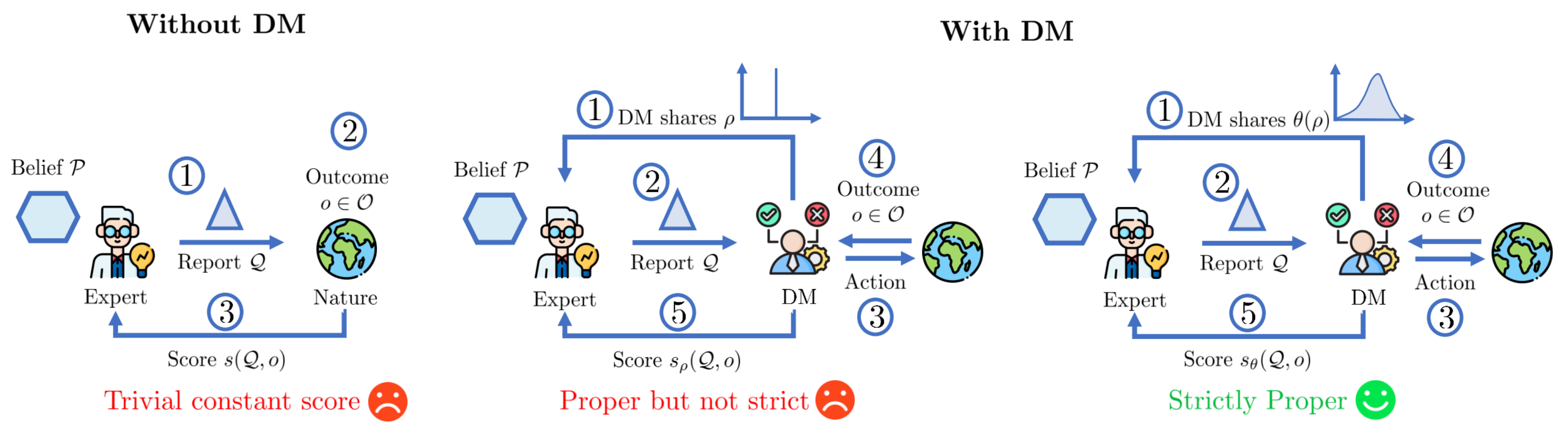
- OOD generalisation is learning with ambiguity
- Subjectivity in fairness, interpretability, robustness, trustworthiness, and privacy creates learning ambiguity.
- **Institutional separation** changes the training pipeline
  - Foundation model: pre-training + fine-tuning
  - LLM alignment
- **Principal-agent 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.

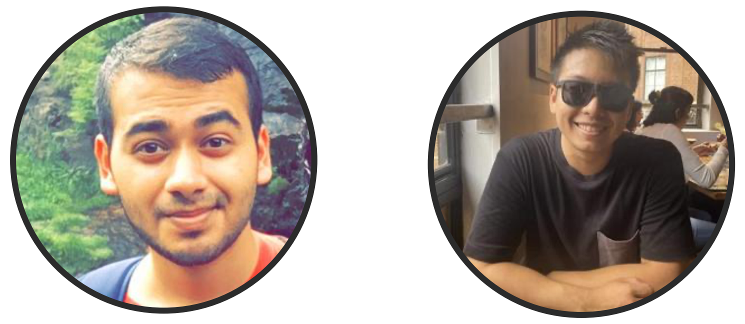
Recent Work and Future Directions





Credal Two-Sample Tests of Epistemic Uncertainty (AISTATS 2025)

Truthful Elicitation of Imprecise Forecast (Under Review)

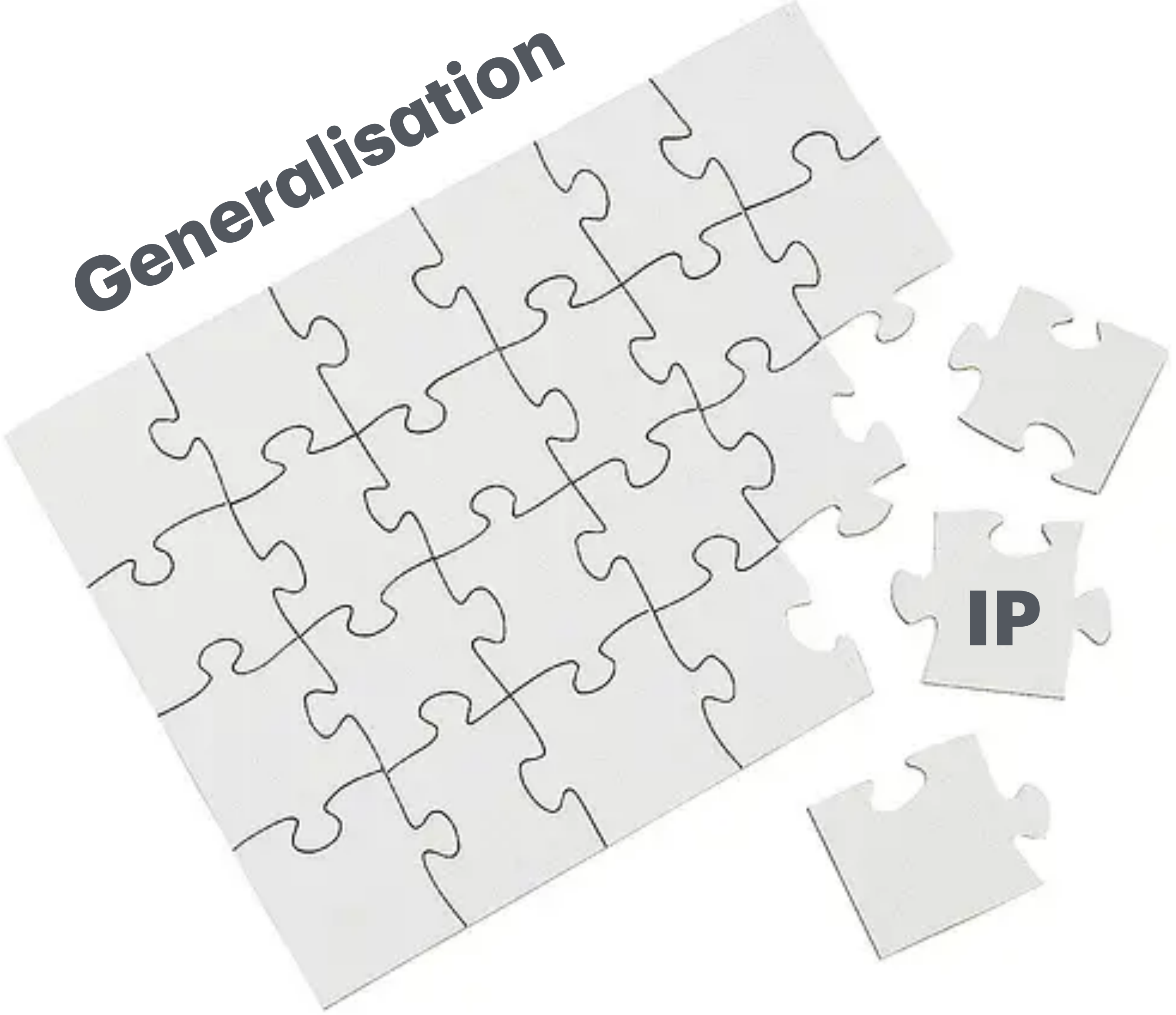


Inspired by “*Scoring Rules and Calibration for Imprecise Probabilities*” from Christian Fröhlich and Robert C. Williamson

Multi-calibration, Decision Making, Elicitation of Causal and Counterfactual Distributions

Conclusion

- 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.



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

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