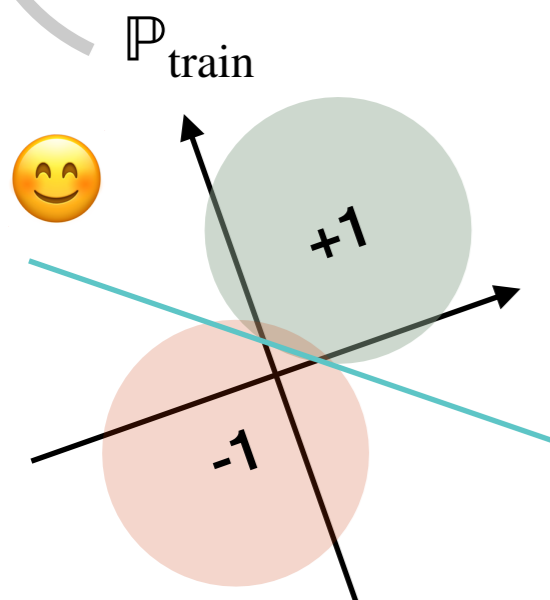
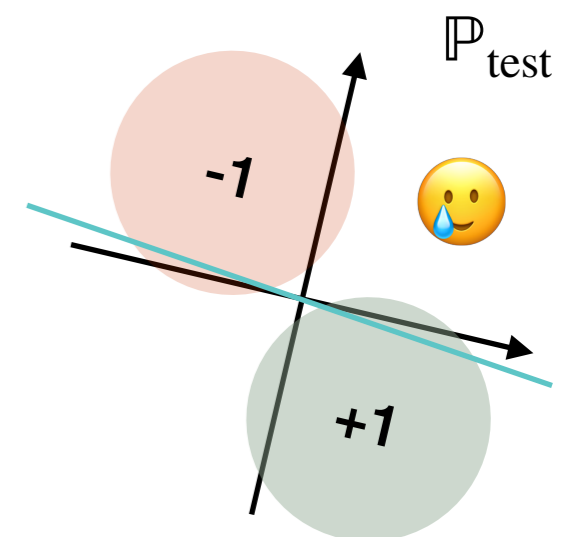


Out-of-distribution generalization and transfer learning: pitfalls and opportunities



Guest lecture by Julia Kostin
25 November 2025
Guarantees for Machine Learning Fall 2025



So far in our lecture:

Supervised learning: given training data $\{(X_i, Y_i)\}_{i \in [n]}$, where $(X, Y) \sim \mathbb{P}$, find a model $f \in \mathcal{F}$ with low excess risk on test data $(X, Y) \sim \mathbb{P}$:

$$\mathcal{E}(f) = \mathcal{R}(f, \mathbb{P}) - \mathcal{R}(f^*, \mathbb{P}),$$

where $f^* = \arg \min_{f \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[\ell(f(X), Y)] =: \arg \min_{f \in \mathcal{F}} \mathcal{R}(f, \mathbb{P})$.

Crucial assumption: training and test data are sampled from the same distribution \mathbb{P} !

This is true, for instance, if one **randomly** splits a pre-existing dataset into **train-val-test**:

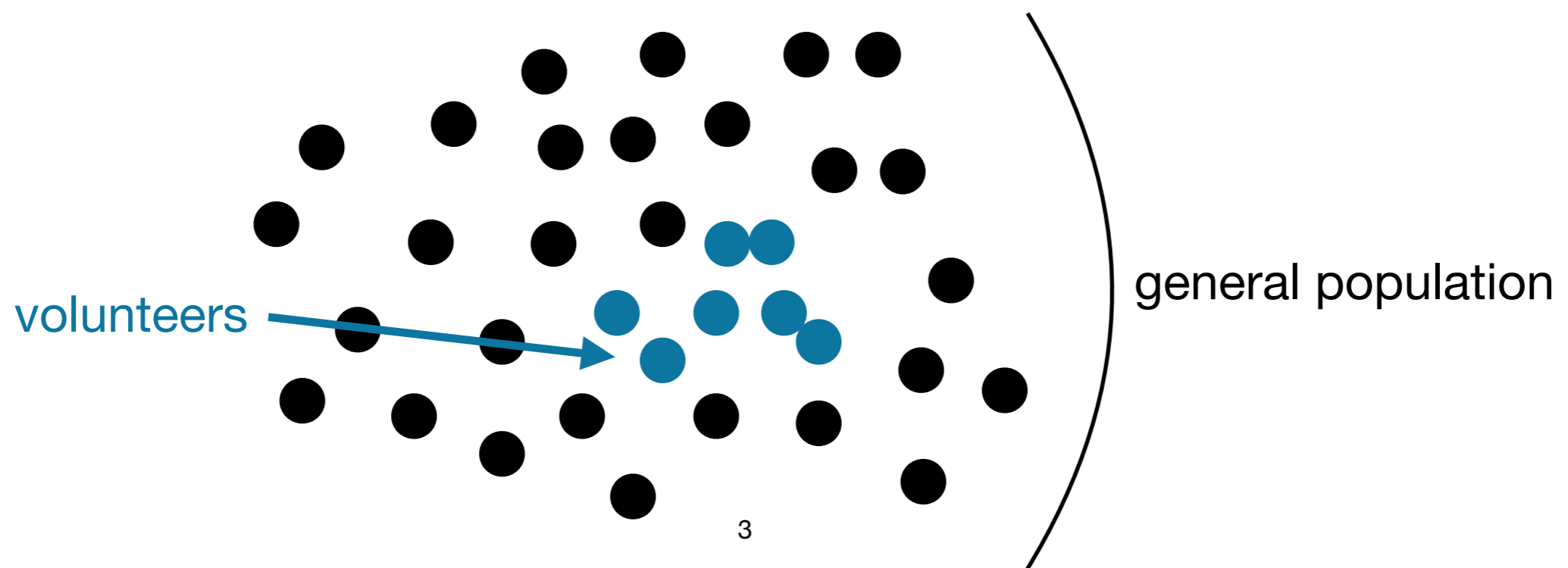


"Dataset" and its corresponding split

I.I.D. assumption vs. real world

However, in practice, training data can be collected:

- **way before** the test data (e.g. time series models trained before COVID);
- at a different geographical location (e.g. personalized ads in different countries);
- only from a certain subgroup / subpopulation (e.g. medical studies on volunteers).



I.I.D. assumption -> OOD assumption

Instead, we need to solve the following **out-of-distribution generalization** problem:

Given training data $\{(X_i, Y_i)\}_{i \in [n]}$, where $(X, Y) \sim \mathbb{P}_{\text{train}}$, find a model $f \in \mathcal{F}$ with low excess risk on test data $(X, Y) \sim \mathbb{P}_{\text{test}}$:

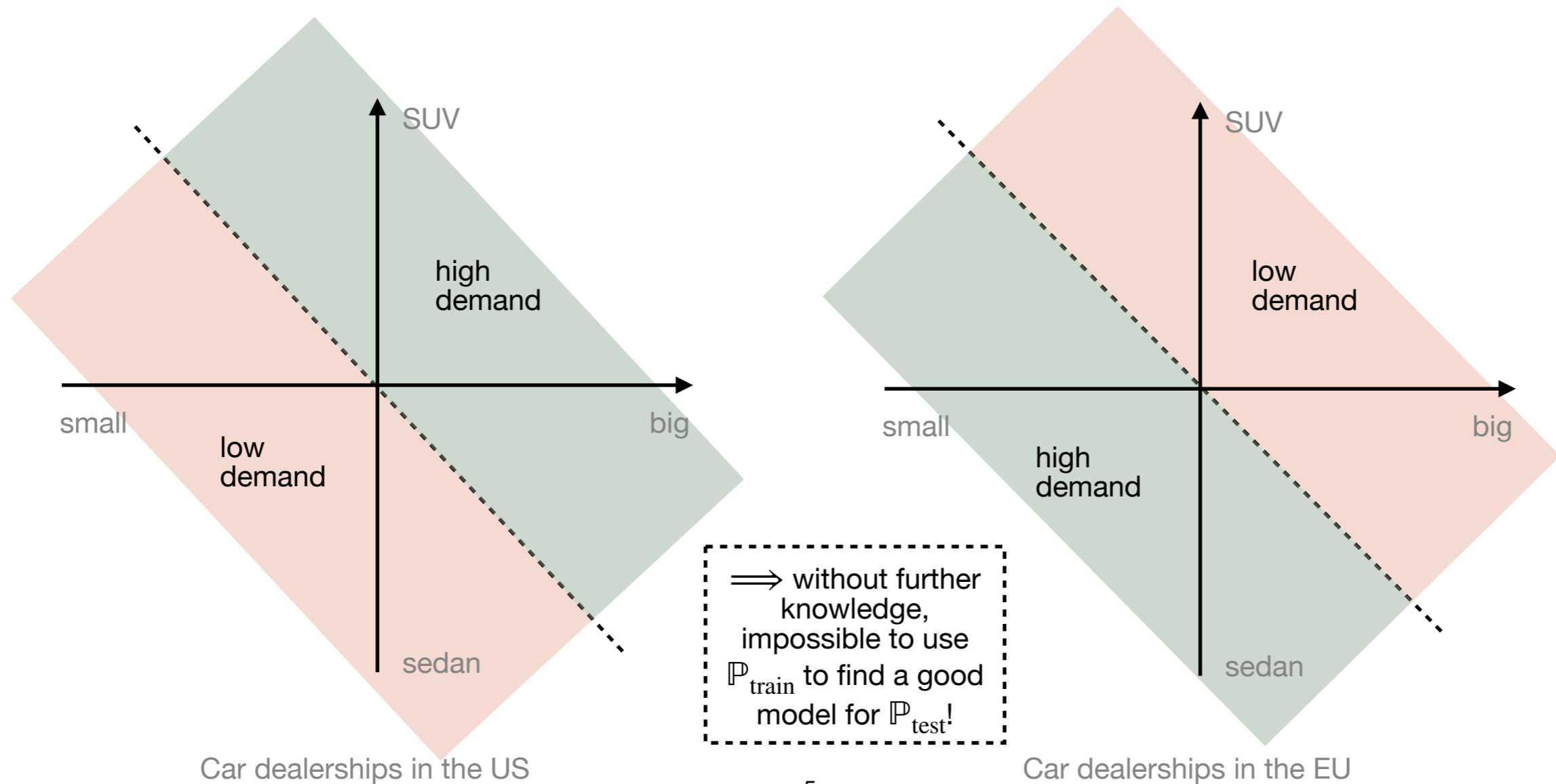
$$\mathcal{E}_{\text{test}}(f) = \mathcal{R}_{\text{test}}(f) - \mathcal{R}_{\text{test}}(f^\star),$$

where $\mathcal{R}_{\text{test}}(f) = \mathbb{E}_{\mathbb{P}_{\text{test}}}[\ell(f(X), Y)]$.

If we can find such an f , we say we can generalize out-of-distribution.

Impossibility of OOD generalization

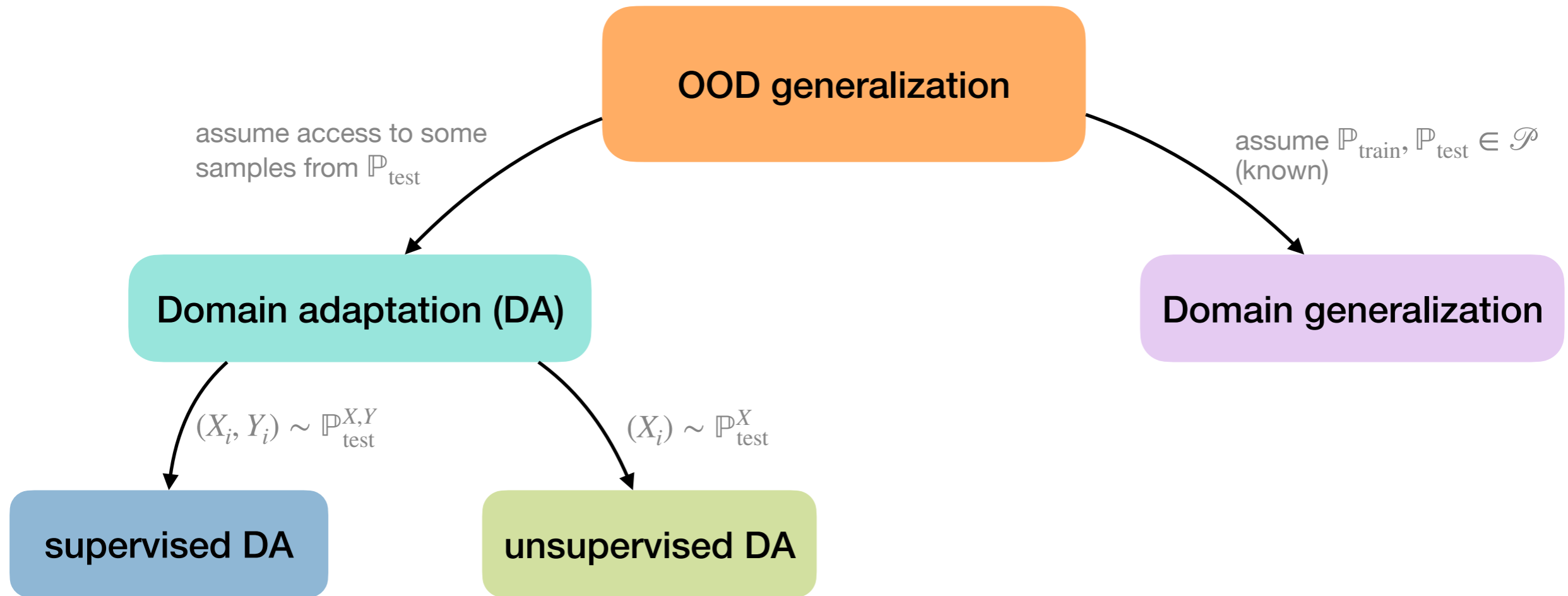
However, even in the infinite-sample limit $n \rightarrow \infty$ (or, given $\mathbb{P}_{\text{train}}$), OOD generalization is, in general, impossible:



Assumptions for OOD generalization

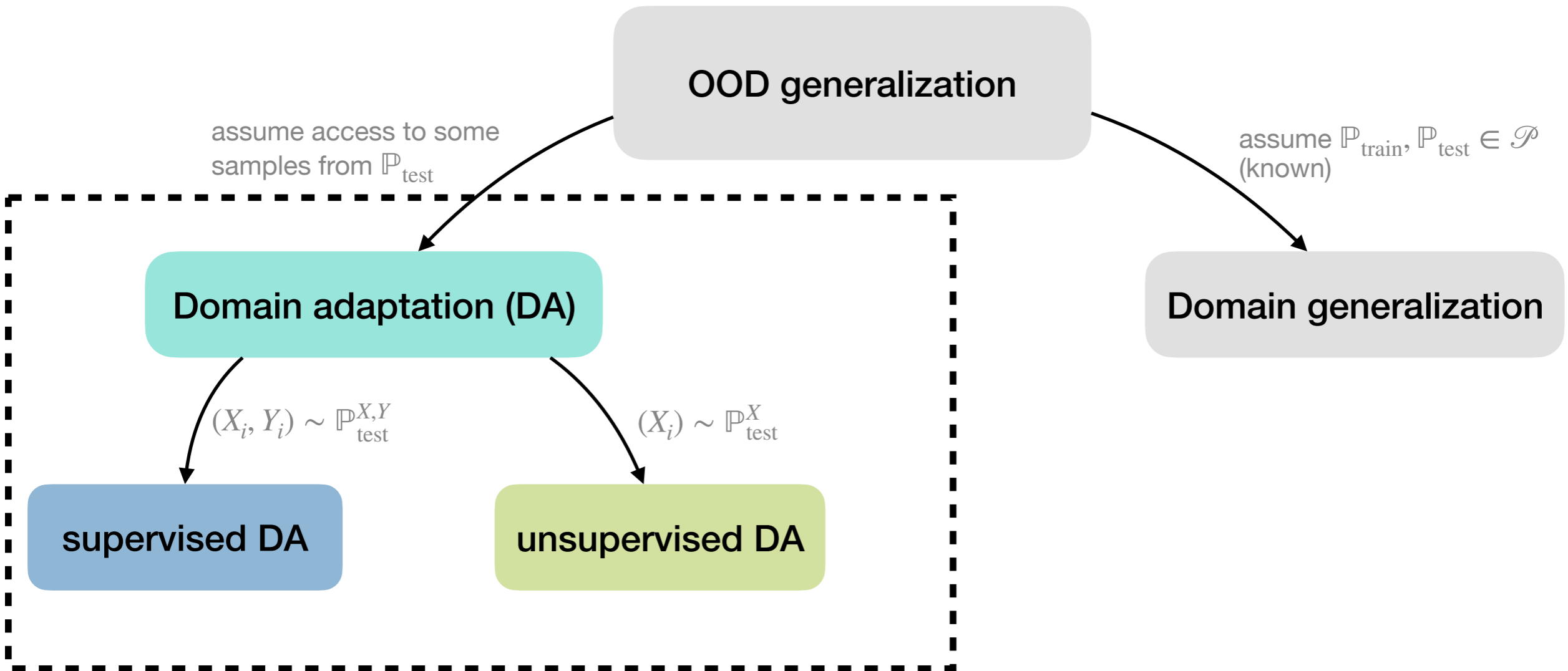
Question: which additional knowledge can be brought into the problem and how can it be motivated?

Taxonomy of OOD generalization

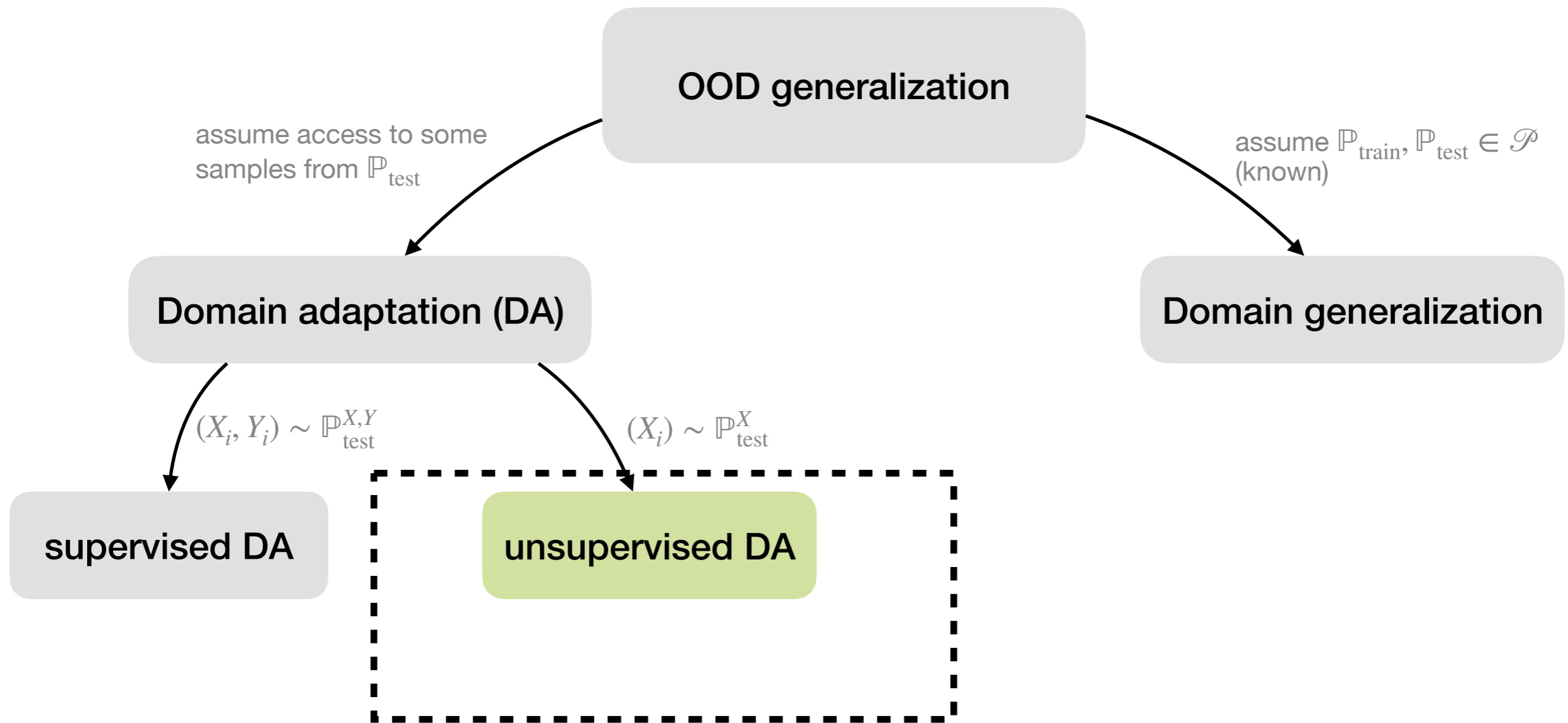


Note: domain adaptation often aims to provide finite-sample guarantees in both **source** data from $\mathbb{P}_{\text{train}}$ and **target** data from \mathbb{P}_{test} . In contrast, domain generalization often operates in the infinite-sample regime.

Taxonomy of OOD generalization



Taxonomy of OOD generalization



Domain adaptation: classical results

We consider the following **classification** setting due to Ben-David:

Source domain: $\langle P_X, f_P \rangle$; Target domain: $\langle Q_X, f_Q \rangle$ (\neq domain of a function!)

where $f_D : \mathcal{X} \rightarrow [0,1]$ a (non-deterministic) labelling function.

Given a (binary) classifier $h : \mathcal{X} \rightarrow \{0,1\}$, define the risk on domain $D \in \{P, Q\}$:

$$\mathcal{R}_D(h) = \mathbb{E}_{D_X} [|h(X) - f(X)|]$$

(corresponds to 0-1 loss if f_D deterministic). Empirical risk: $\hat{\mathcal{R}}(h)$.

L1 divergence and the first bound

Intuitively: if the distributions P_X, Q_X and the labelling functions f_P, f_Q are "far apart", source data is not useful.

⇒ need appropriate measures of divergence of distributions and labellers.

Definition. For distributions P_X, Q_X over \mathcal{X} , the L_1 divergence (or absolute variation distance) is defined as

$$d_1(P_X, Q_X) = 2 \sup_{B \in \mathcal{B}} |P_X(B) - Q_X(B)|,$$

where \mathcal{B} are sets measurable under P_X and Q_X .

L1 divergence and the first bound

Definition. For distributions P_X, Q_X , the L_1 divergence (or absolute variation distance) is defined as

$$d_1(P_X, Q_X) = 2 \sup_{B \in \mathcal{B}} |P_X(B) - Q_X(B)|,$$

where \mathcal{B} are sets measurable under P_X and Q_X .

Theorem [Ben-David 2009]. For a hypothesis h , it holds

$$\mathcal{R}_Q(h) \leq \mathcal{R}_P(h) + d_1(P_X, Q_X) + \min\{\mathbb{E}_{P_X}[|f_P(X) - f_Q(X)|], \mathbb{E}_{Q_X}[|f_P(X) - f_Q(X)|]\}.$$

T1: source risk T2: covariate shift T3: difference of labelling functions

L1 divergence and the first bound

Theorem [Ben-David 2010]. *For a hypothesis h , it holds*

$$\mathcal{R}_Q(h) \leq \mathcal{R}_P(h) + d_1(P_X, Q_X) + \min\{\mathbb{E}_{P_X}[|f_P(X) - f_Q(X)|], \mathbb{E}_{Q_X}[|f_P(X) - f_Q(X)|]\}.$$

T1: source risk **T2: covariate shift** **T3: difference of labelling functions**

T1: minimized by ERM on source, "ideal case"

T3: difference of labelling functions, small if tasks similar

T2: want to upper bound d_1 , however:

- hard to estimate in finite samples
- overly strict measure -- not all measurable subsets are relevant in practice, only ones on which hypotheses of interest can make mistakes!

Refinement: the \mathcal{H} -divergence

Definition. Let P_X, Q_X be two distributions over \mathcal{X} and \mathcal{H} a hypothesis class over \mathcal{X} . Denote $I(h) := \{x \in \mathcal{X} : h(x) = 1\}$. The \mathcal{H} -divergence is defined as

$$d_{\mathcal{H}}(P_X, Q_X) := 2 \sup_{h \in \mathcal{H}} |P_X(I(h)) - Q_X(I(h))|.$$

- \mathcal{H} -divergence is never larger than the L_1 -divergence, can be much smaller if \mathcal{H} has finite VC dimension.
- \mathcal{H} -divergence can be estimated from finite samples:

$$d_{\mathcal{H}}(P_X, Q_X) \leq \hat{d}_{\mathcal{H}}(\mathcal{D}_P, \mathcal{D}_Q) + 4 \sqrt{\frac{VC(\mathcal{H}) \log(2m) + \log(2/\delta)}{m}},$$

where $\mathcal{D}_P = \{X_i\}_{i=1}^m$ and $\mathcal{D}_Q = \{X_j\}_{j=1}^m$.

Refinement: the \mathcal{H} -divergence

\mathcal{H} -divergence can be estimated from finite samples:

$$d_{\mathcal{H}}(P_X, Q_X) \leq \hat{d}_{\mathcal{H}}(\mathcal{D}_P, \mathcal{D}_Q) + 4\sqrt{\frac{VC(\mathcal{H})\log(2m) + \log(2/\delta)}{m}},$$

where $\mathcal{D}_P = \{X_i\}_{i=1}^m$ and $\mathcal{D}_Q = \{X_j\}_{j=1}^m$.

- The empirical \mathcal{H} -divergence can be computed via the minimum error on a surrogate "source vs. target" classification task using $h \in \mathcal{H}$.

Domain adaptation bounds using \mathcal{H} -divergence

Definition. *The ideal joint hypothesis is the hypothesis which minimizes the combined source and target risk:*

$$h^* := \arg \min_{h \in \mathcal{H}} [\mathcal{R}_P(h) + \mathcal{R}_Q(h)].$$

We denote its combined risk by $\lambda := \mathcal{R}_P(h^) + \mathcal{R}_Q(h^*)$*

Intuition: if the "domains", or "tasks" are similar and \mathcal{H} is expressive enough, they admit a classifier which is simultaneously good on both tasks.

Domain adaptation bounds using \mathcal{H} -divergence

Define the **symmetric difference hypothesis class**

$$\mathcal{H} \Delta \mathcal{H} := \{h \oplus h' : h, h' \in \mathcal{H}\},$$

where \oplus is the XOR operation. "Set of disagreements" between two hypotheses in \mathcal{H} .

The $\mathcal{H} \Delta \mathcal{H}$ -divergence only takes the supremum of $|P_X(B) - Q_X(B)|$ on sets B on which two hypotheses from \mathcal{H} can disagree!

Domain adaptation bounds using \mathcal{H} -divergence

Theorem. Let \mathcal{H} be a hypothesis class of VC dimension d . Let $\mathcal{D}_P, \mathcal{D}_Q$ be unlabelled datasets from source and target distributions of size m' . Then with probability at least $1 - \delta$, for every $h \in \mathcal{H}$:

$$\mathcal{R}_Q(h) \leq \mathcal{R}_P(h) + \frac{1}{2} \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_P, \mathcal{D}_Q) + 4 \sqrt{\frac{2d \log(2m') + \log(2/\delta)}{m'}} + \lambda$$

T1: source risk

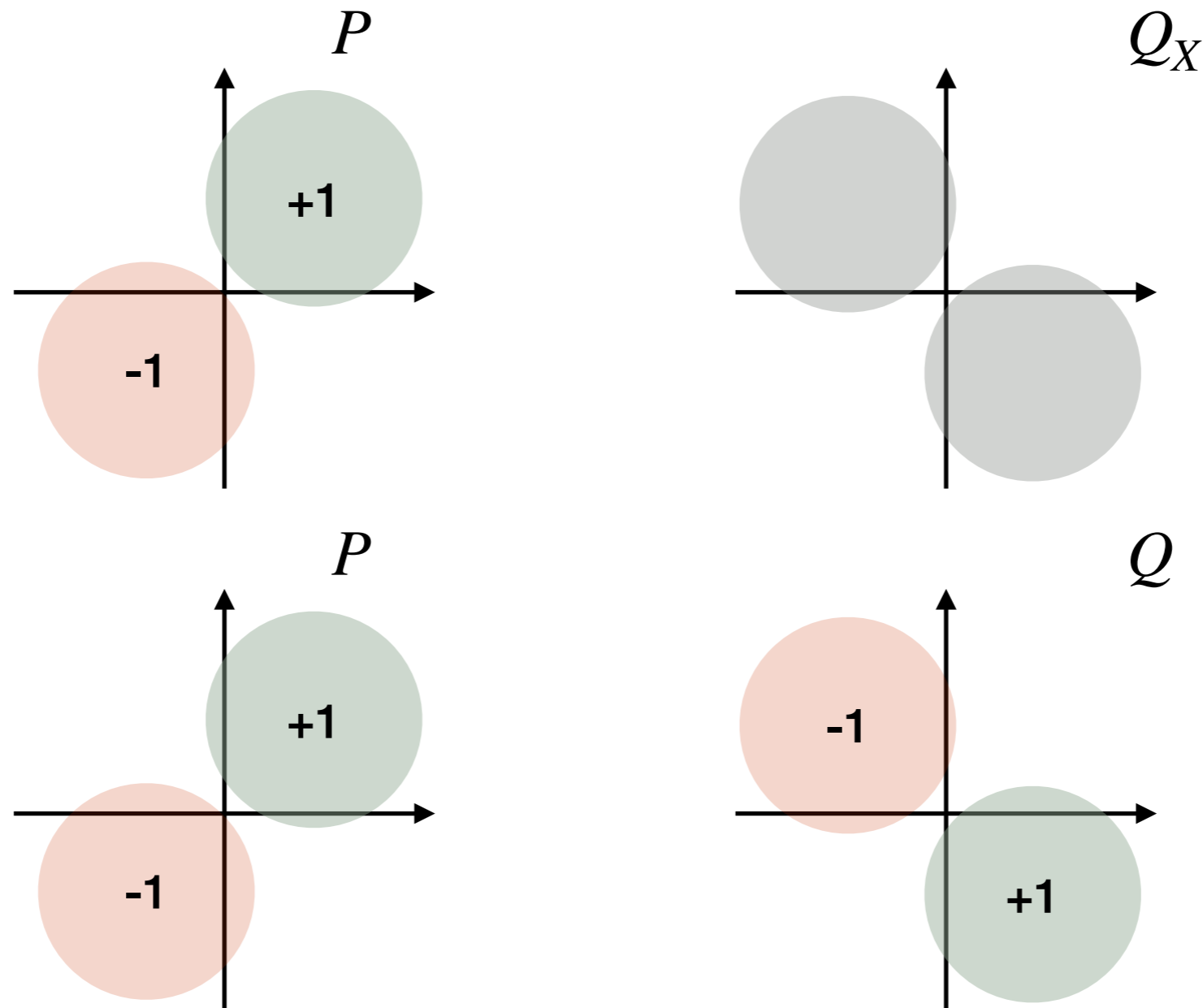
T2: covariate shift + finite sample error

T3: joint hypothesis risk

- Refined measure of divergence which can be approximated from data
- Term λ depends on the loss function, expresses how well the tasks can be jointly solved
- Without further assumptions, all terms are necessary - cf. lower bounds in Ben David et al., Impossibility Theorems for Domain Adaptation.

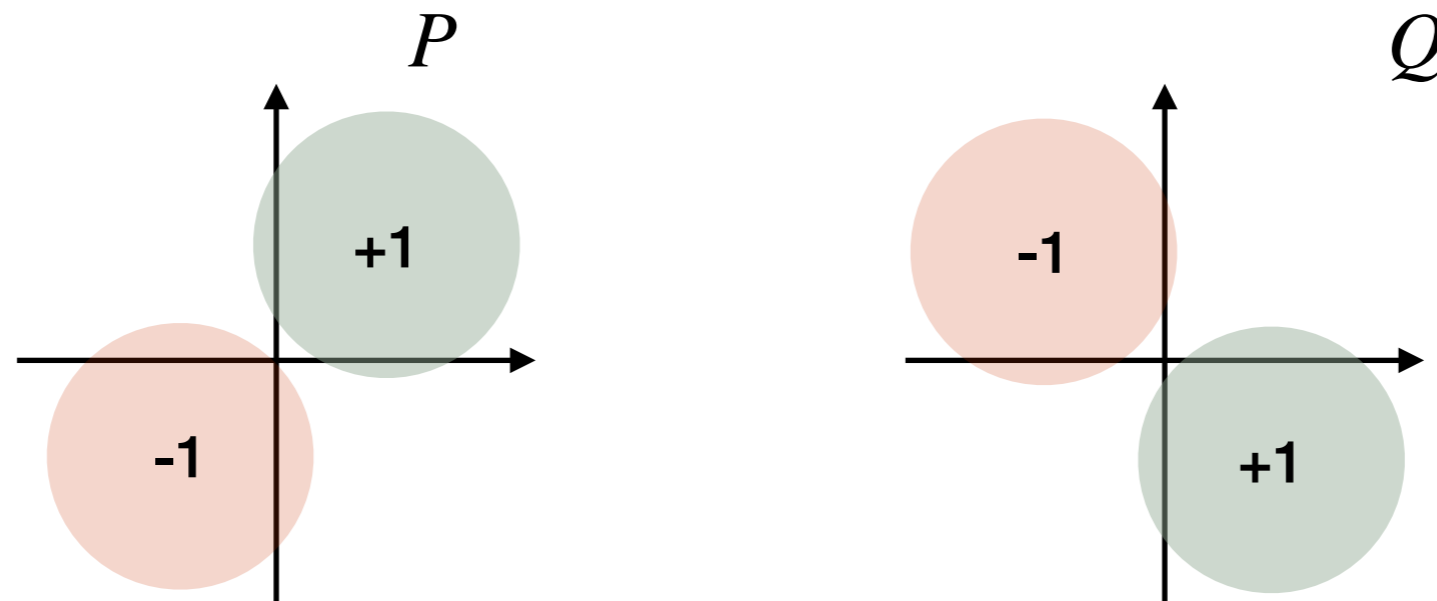
Divergence-based bounds: discussion

- Inspired a series of "domain-adversarial" methods, where the goal is to learn a representation $\Phi(X)$ of the feature space such that $d_{\mathcal{H}\Delta\mathcal{H}}(P_X^\Phi, Q_X^\Phi)$ small.
- However:



Divergence-based bounds: discussion

- Inspired a series of "domain-adversarial" methods, where the goal is to learn a representation $\Phi(X)$ of the feature space such that $d_{\mathcal{H}\Delta\mathcal{H}}(P_X^\Phi, Q_X^\Phi)$ small.
- However: learning such a representation using solely unlabelled target data can result in label-flipping (spurious) features!



Divergence-based bounds: discussion

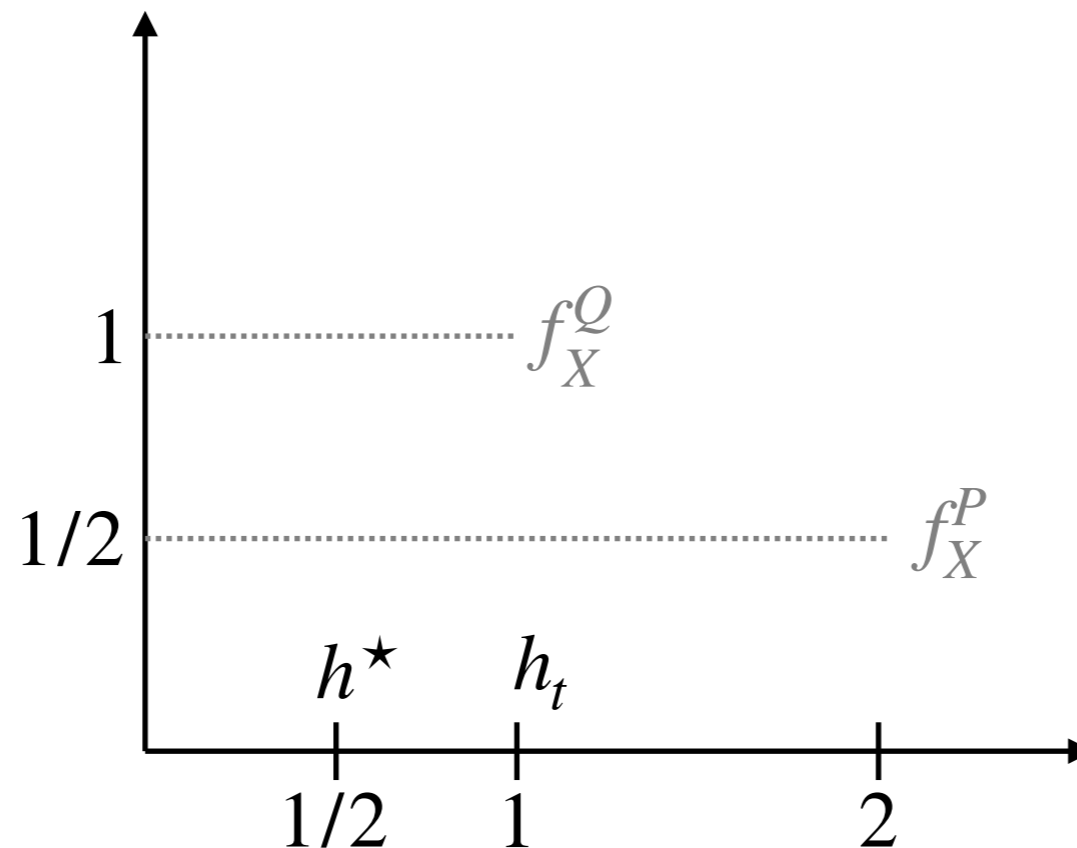
- $\mathcal{H} \Delta \mathcal{H}$ -divergence is symmetric in P and Q , however, transfer of information from one distribution to another is **not symmetric**
- $\mathcal{H} \Delta \mathcal{H}$ -divergence can be constant when transfer from P and Q is possible

Divergence-based bounds: discussion

- Example: \mathcal{H} = threshold functions, $P_X = \text{Unif}(0,2)$, $Q_X = \text{Unif}(0,1)$. Y deterministic given X with $h^*(X) = \text{sign}(x \leq 1/2)$.
- Here $d_{\mathcal{H}\Delta\mathcal{H}}(P_X, Q_X) = 1/2$, but

$$\mathcal{E}_Q(h) = Q(h \neq h^*) \leq 2P(h \neq h^*) = \mathcal{E}_P(h)$$

- In particular, the target risk decreases if the source risk decreases!



Divergence-based bounds: discussion

- Example: \mathcal{H} = threshold functions, $P_X = \text{Unif}(0,2)$, $Q_X = \text{Unif}(0,1)$. Y deterministic given X with $h^\star(X) = \text{sign}(x \leq 1/2)$.
- Here $d_{\mathcal{H} \Delta \mathcal{H}}(P_X, Q_X) = 1/2$, but

$$\mathcal{E}_Q(h) = Q(h \neq h^\star) \leq 2P(h \neq h^\star) = \mathcal{E}_P(h)$$

- In particular, the target risk decreases if the source risk decreases!

This motivates introducing notions of complexity of domain adaptation **beyond divergences**

Beyond divergence

Definition. We call $\rho > 0$ a *transfer exponent* from P to Q w.r.t. \mathcal{H} , if

$$\mathcal{E}_Q^\rho(h) \leq C_\rho \mathcal{E}_P(h), \forall h \in \mathcal{H}$$

for some universal constant C_ρ .

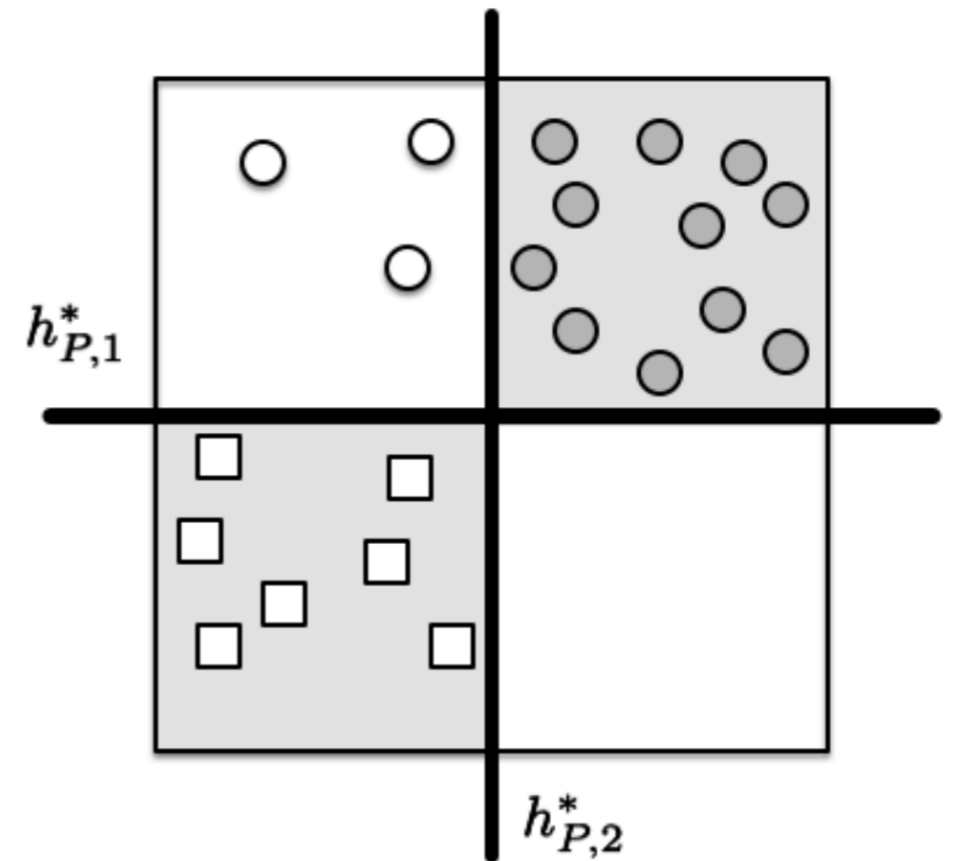
Intuition: equivalently $\mathcal{E}_Q(h) \leq C_\rho^{1/\rho} \mathcal{E}_P^{1/\rho}(h)$, i.e. the target risk of any hypothesis is upper-bounded by an exponent of its source risk. Smaller ρ - more information to transfer. If $\rho \leq 1$, "super-transfer". If $\rho = \infty$, no transfer.

- The notion of transfer exponent allows for adaptive procedures which achieve zero excess risk with growing number of source **or** target samples;
- Allows to establish bounds for unsupervised or supervised DA
- All provided such an exponent $\rho < \infty$ exists!

Non-monotonic risks

Consider the following example:

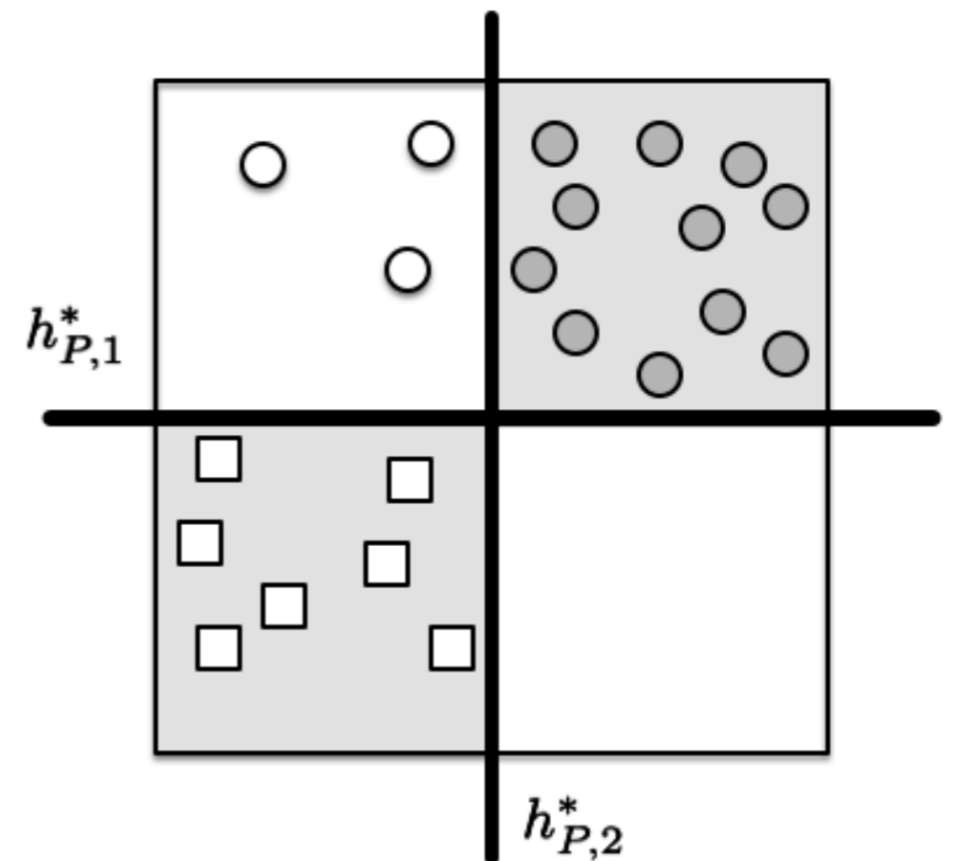
- **Task:** classify circle from square.
- Source distribution P_X is supported in the darker region
- Target distribution Q_X is supported in the lighter region
- $\mathcal{H} = \{v : v = ce_1 \text{ or } v = ce_2, c \in \mathbb{R}\}$
(classifiers along one of the coordinates).
- We have $\mathcal{R}_P(h_{P,1}^*) = \mathcal{R}_P(h_{P,2}^*) = 0$,
however $\mathcal{R}_Q(h_{P,1}^*) = 0; \mathcal{R}_Q(h_{P,2}^*) = 1$.
- With just few **labeled** target samples, we could easily detect the model with low risk.



- Transfer exponents etc. are too pessimistic in this scenario!

Non-monotonic risks

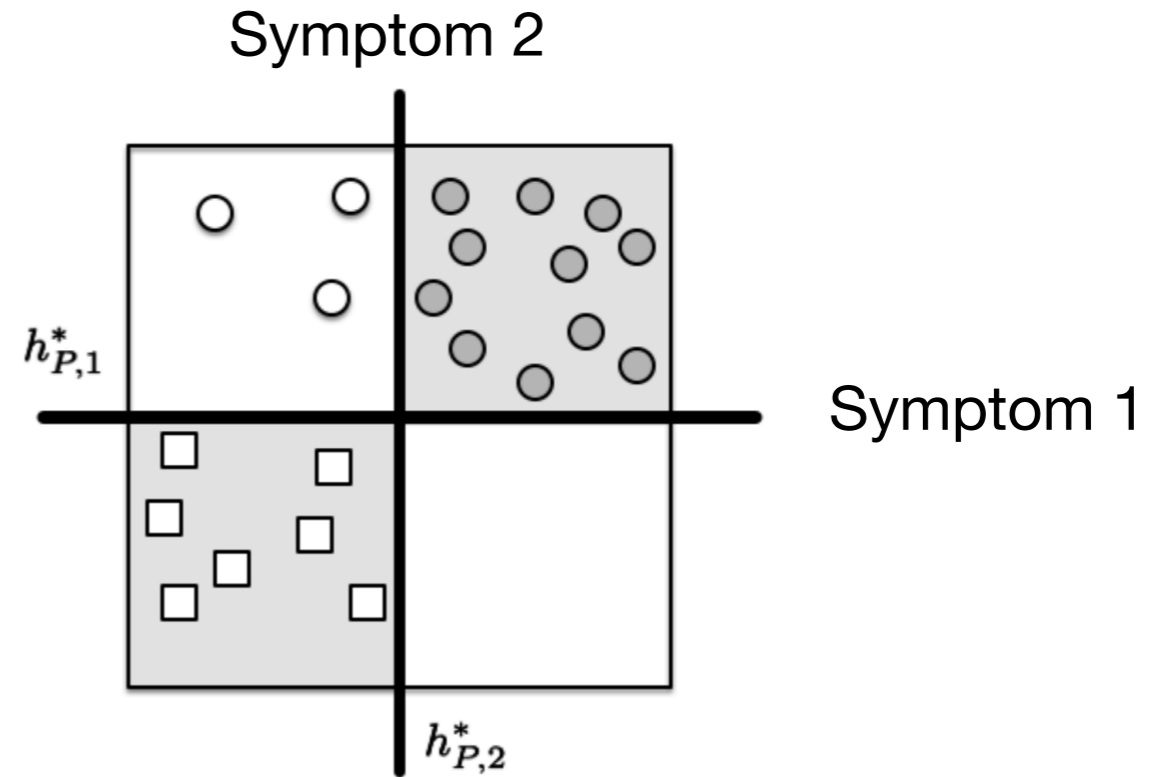
- In this example, we have utilised "belief" that **at least one** of the good models $h_{P,i}^*$ will have a low target risk, as opposed to any good model on P .
- Such situations do occur in practice!
- But to formalize such "belief", we need to assume *additional structure* of what "can change" between source and target distributions.



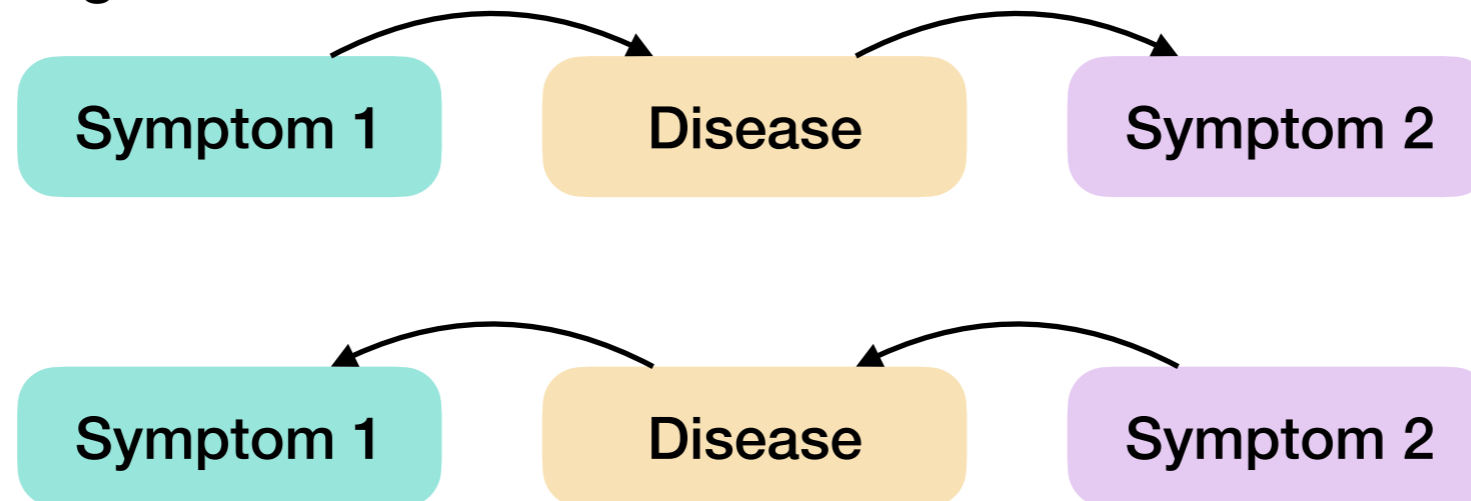
- Transfer exponents etc. are too pessimistic in this scenario!

A "medical" example

- **Example:** predicting whether a patient has a certain disease based on observations of two symptoms.



Domain knowledge: either

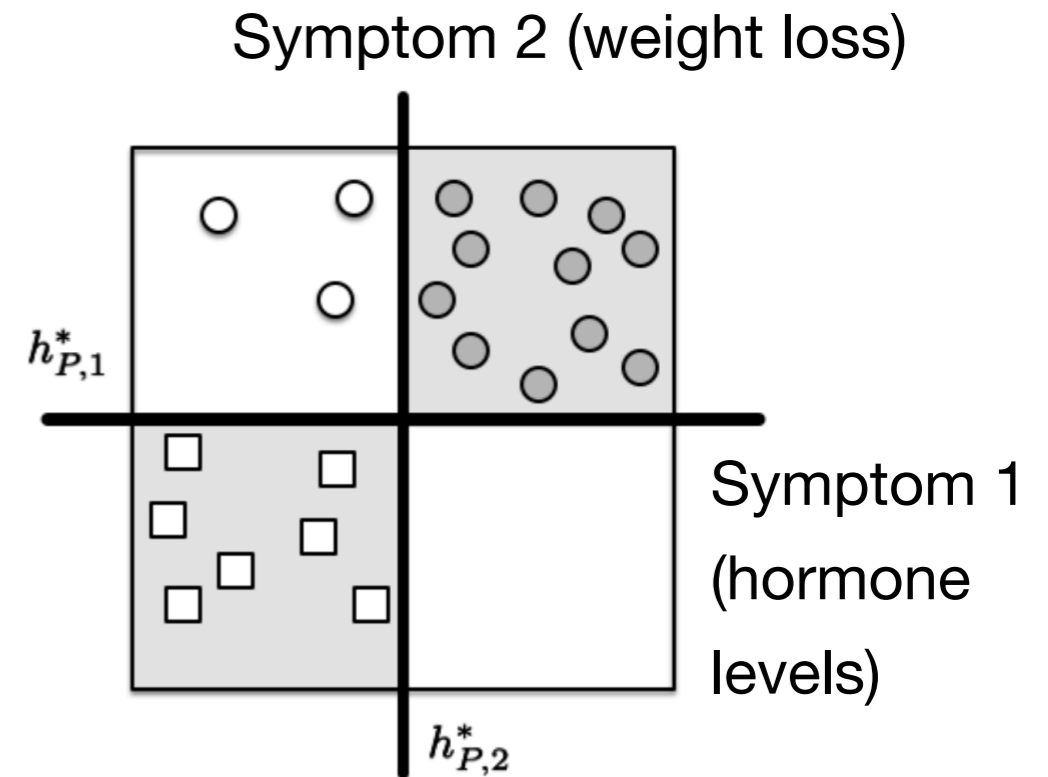


A "medical" example

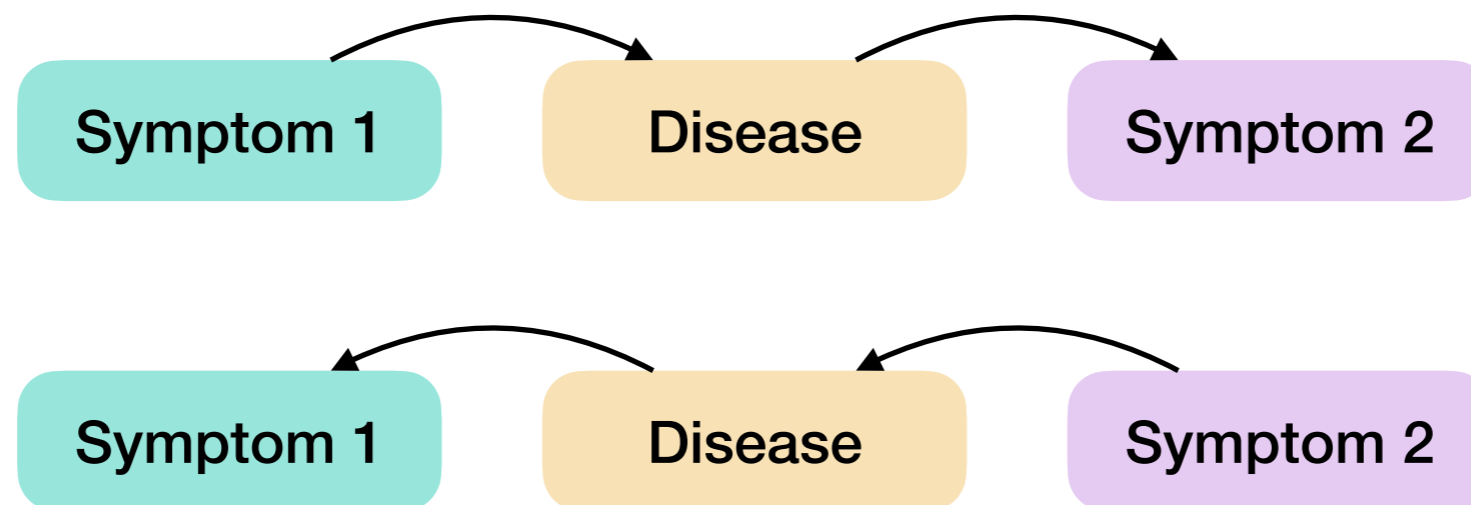
Population A (source) shows Symptom 2 only if they have the disease.

Population B (target) shows Symptom 2 only if they do not have the disease.

Using only the symptom which **causes** the disease will yield a model which performs reasonably well **on both source and target domains**.



Domain knowledge: either



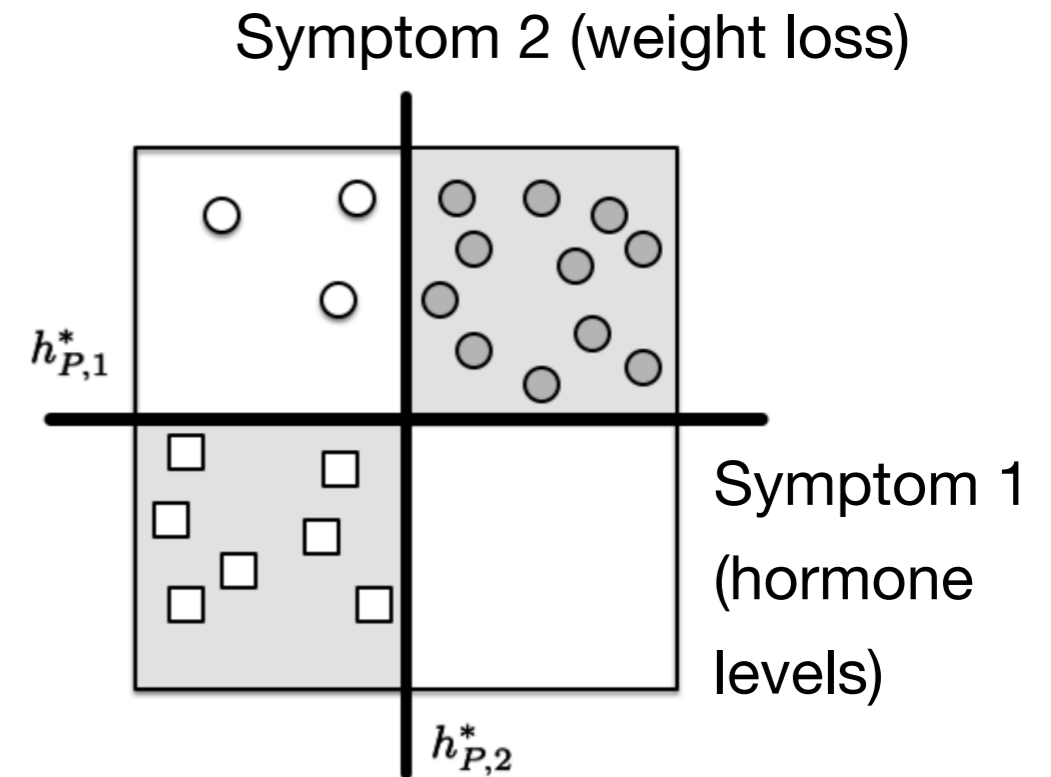
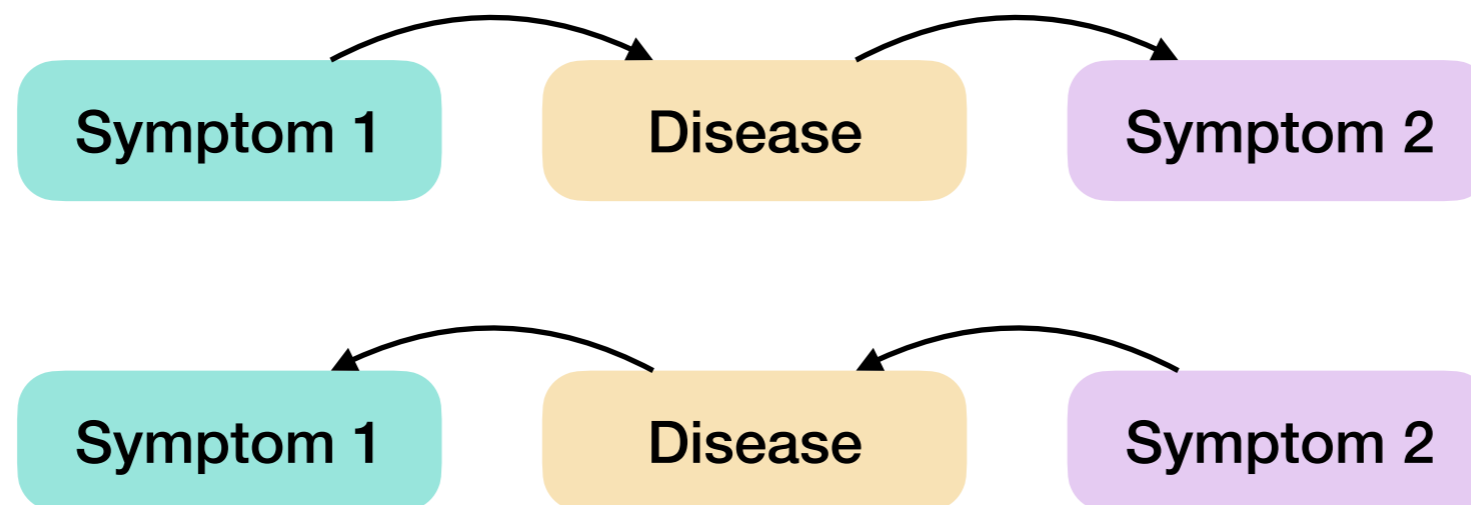
A "medical" example

In this example, we have assumed:

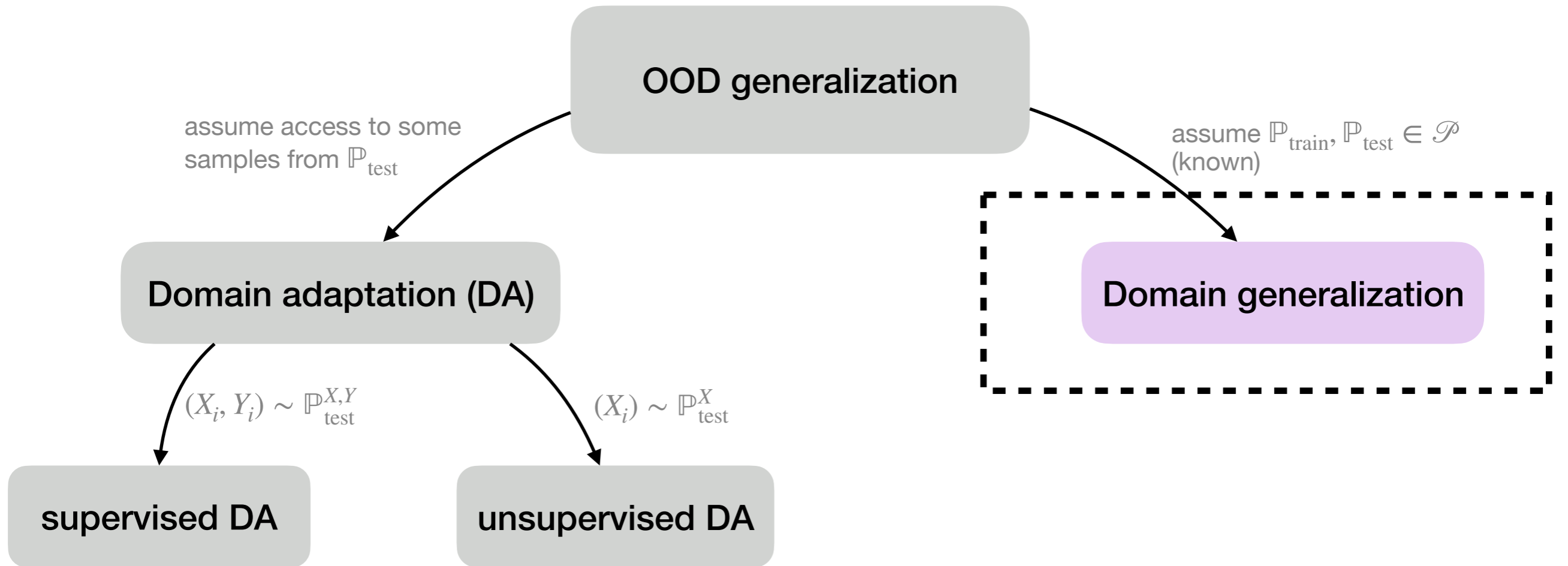
although we do not know which symptom causes the disease, the **biological mechanism** of the disease is **invariant** and remains the same across source and target. However, **spurious correlations** of the disease with other symptoms or the way the disease manifests **can vary**.

We will formalize this assumption:

- 1) in the abstract framework of worst-case **domain generalization**;
- 2) more concretely through **causality**



Taxonomy of OOD generalization



Small excursion into the world of domain generalization

Worst-case domain generalization

Given: (data from) the source distribution P and belief that an (*unseen*) target distribution can only exhibit certain shifts from P , i.e. $Q \in \mathcal{Q}_P$. Here, \mathcal{Q}_P is a (potentially infinite) collection of possible target distributions Q .

Goal: find model $h \in \mathcal{H}$ such that

$$h \in \arg \min_{h \in \mathcal{H}} \max_{Q \in \mathcal{Q}_P} \mathcal{R}_Q(h)$$

(game: we pick the model, then an adversary picks the target distribution)

Finding such a model h = achieving *distributional robustness w.r.t. target shift*.

Solving this min-max problem is intractable, i.a. since \mathcal{Q}_P is unknown.

However, one can try to utilize **multiple source domains** $\{P_e : e \in \mathcal{E}_{\text{train}}\}$.

Variability across these domains can help approximate \mathcal{Q}_P .

Achieving worst-case domain generalization

Many **multi-environment** methods have been proposed to obtain distributionally robust models, including:

- (group) distributionally robust optimization (**groupDRO**):

$$\min \max_{e \in \mathcal{E}_{\text{train}}} \mathcal{R}_{P_e}(h)$$

Sagawa et al. (2019).
Distributionally robust neural networks for group shifts

- variance-of-risks penalization (**vREx**):

$$\min_h \left[\beta \text{Var}(\mathcal{R}_{P_1}(h), \dots, \mathcal{R}_{P_m}(h)) + \sum_e \mathcal{R}_{P_e}(h) \right]$$

Krueger et al. (2021).
Out-of-distribution generalization via risk extrapolation (rex).

- invariant risk minimization (**IRM**):

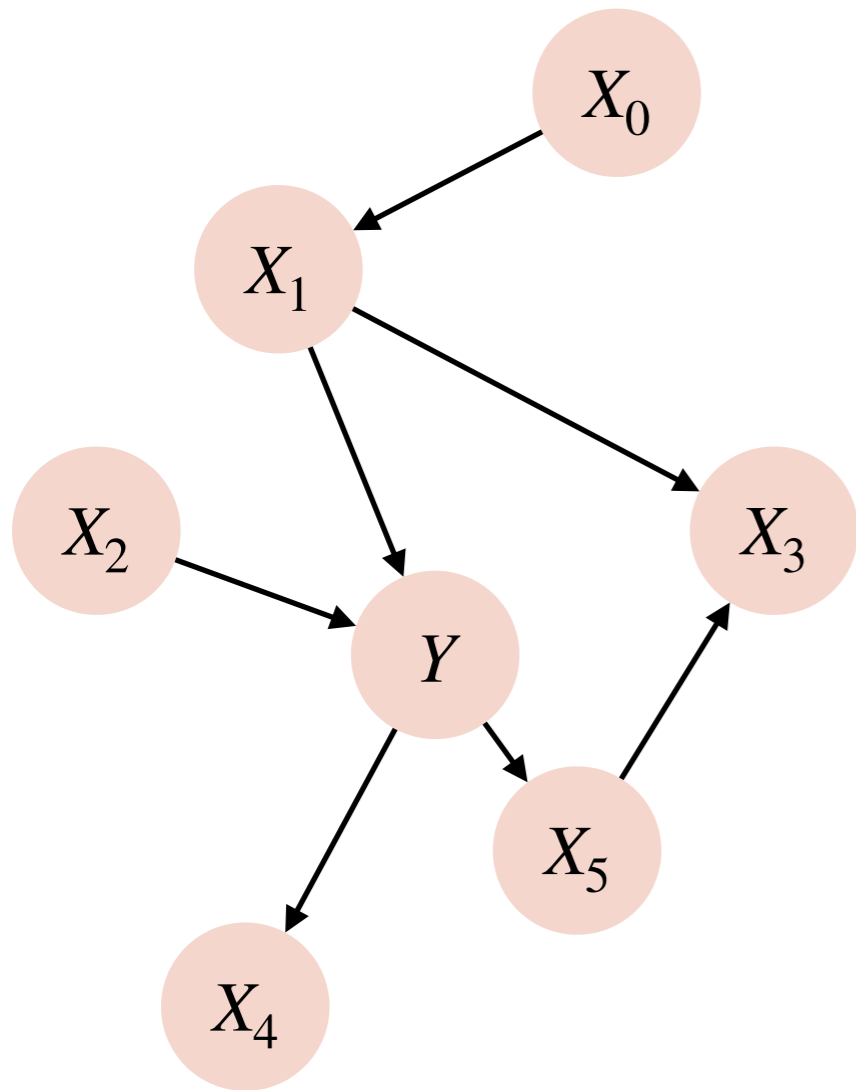
$$\min_{h, \Phi} \sum_e \mathcal{R}_{P_e}(h \circ \Phi) \text{ s.t. } h \in \arg \min_{h'} \mathcal{R}_{P_e}(h' \circ \Phi) \forall e$$

Arjovsky et al. (2019).
Invariant risk minimization

- and many more.

When can we hope to recover a well-generalizing model with these objectives?

DAGs and Structural Causal Models

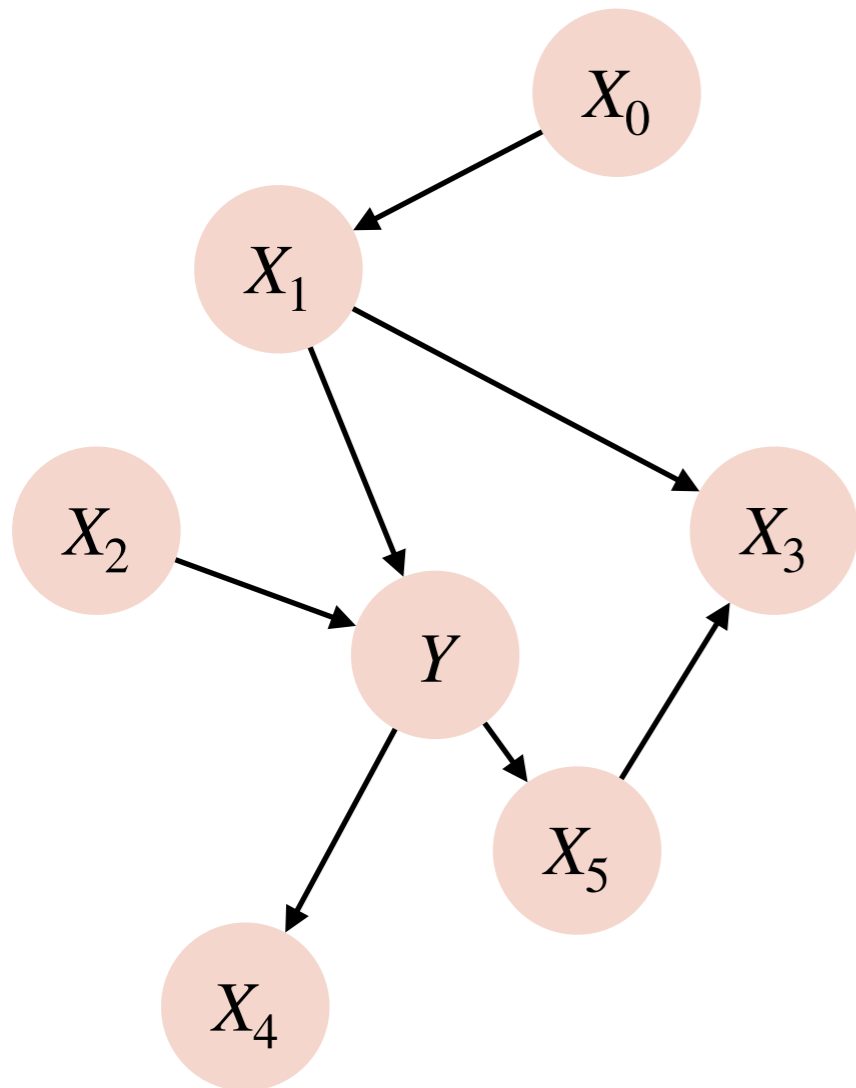


- Causal relationships between variables can be graphically represented via a directed acyclic graph (**DAG**) $\mathcal{G} = (V, E)$
- $X_i \rightarrow X_j$ means X_i is a **parent** (cause) of X_j , X_j is a **child** of X_i .
- A structural causal model (**SCM**) is a collection (for each **node** X_i) of **causal mechanisms** f_i and **exogenous noise** (random) variables U_i such that

$$X_i = f_i(\text{Pa}(X_i), U_i)$$

and all U_i are independent.

DAGs and Structural Causal Models



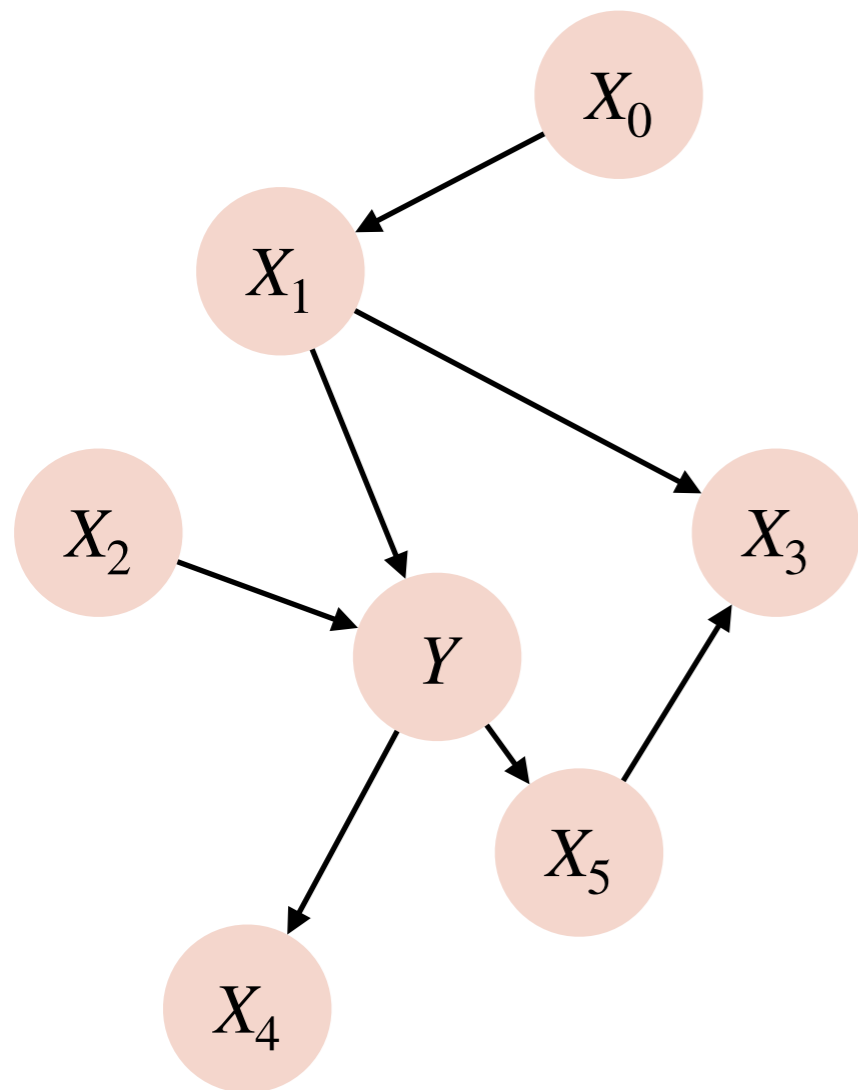
- Causal relationships between variables can be graphically represented via a directed acyclic graph (**DAG**) $\mathcal{G} = (V, E)$
- $X_i \rightarrow X_j$ means X_i is a **parent** (cause) of X_j , X_j is a **child** of X_i .
- A structural causal model (**SCM**) is a collection (for each **node** X_i) of **causal mechanisms** f_i and **exogenous noise** (random) variables U_i such that

$$X_i = f_i(\text{Pa}(X_i), U_i)$$

and all U_i are independent.

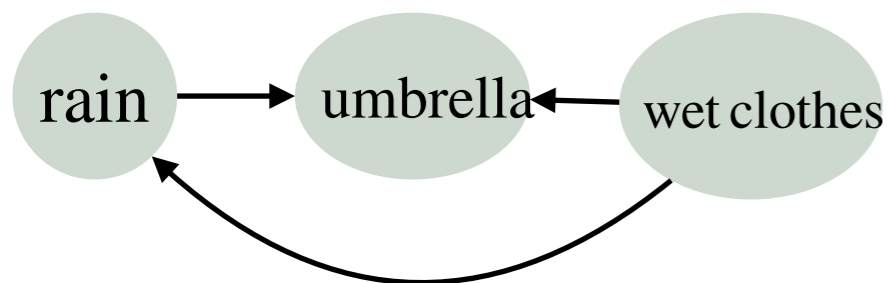
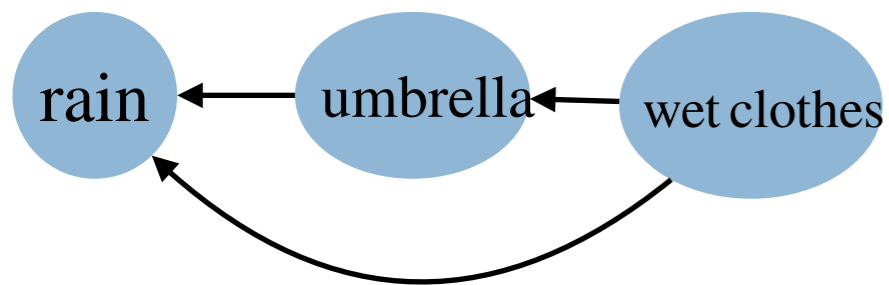
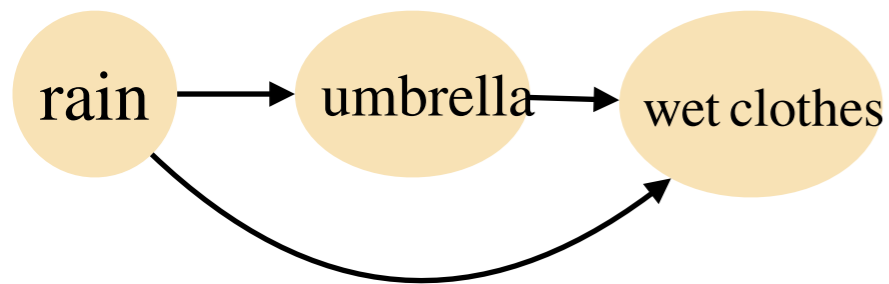
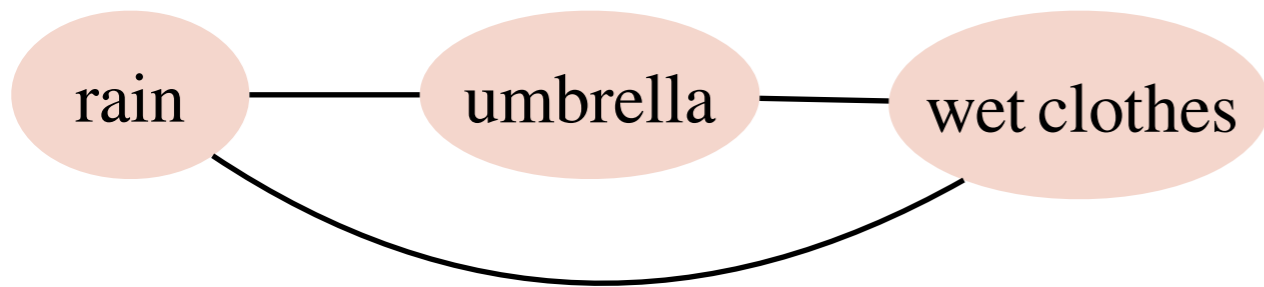
Natural (and necessary) assumption: **causal mechanism of Y remains invariant** across domains, while the causal mechanisms of other variables might change.

DAGs, interventions



- A DAG encodes conditional independence relationship in form of **d-separation**.
- Example: $Y \perp X_0 \mid X_1$
- For a given distribution P , there might be multiple DAGs which encode the correct conditional independence relationships
- Thus, from just a single distribution, one in general cannot identify the causal structure
- Instead, one can identify its **skeleton**, whose possible orientations give rise to the **Markov Equivalence Class (MEC)** of possible DAGs

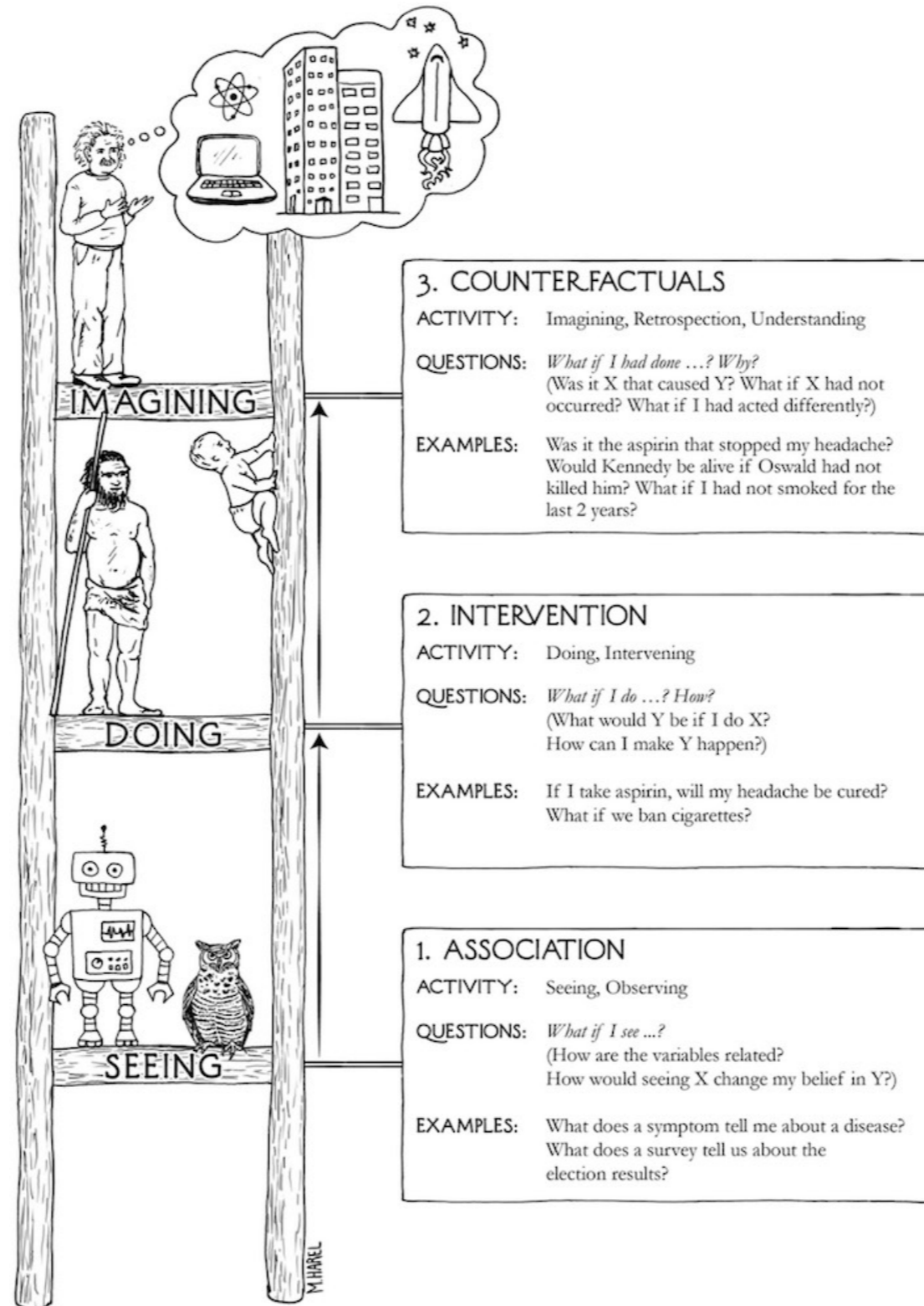
Identifiability of DAGs: example



and more... in total 8 candidate graphs!

- Consider the 3-node causal graph
- Rain is correlated with umbrella usage, wet clothes – but what is causing what?
- Making it rain increases umbrella usage and wet clothes
- But making everyone use umbrellas does not make it rain
- The above are examples of **interventions**:
- Intervention on X_j forces $X_j \sim P_{int}$, regardless of causal mechanism
- Intervention then propagates through the graph

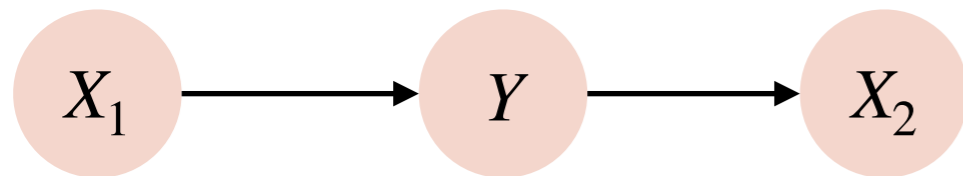
Ladder of causation



Causality for domain generalization: example

Key observation: shared causal structure results in models with stable risk!

For instance, consider the following linear Gaussian SCM:



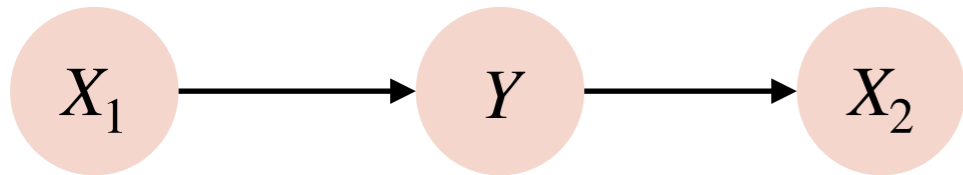
$$X_1 = U_1 \sim \mathcal{N}(\mu_1, 1)$$

$$Y = X_1 + U_Y, U_Y \sim \mathcal{N}(0, 1)$$

$$X_2 = Y + U_2, U_2 \sim \mathcal{N}(\mu_2, 1)$$

Assumption: the **source distribution** P is induced by the SCM with $\mu_1 = 0, \mu_2 = 0$; the **target distribution** Q is induced by the same SCM but with non-zero μ_1, μ_2 .

Causality for domain generalization: example



Linear regression setting with squared loss:

$$\mathcal{R}_P(h) = \mathbb{E}_P[(Y - h(X))^2]$$

Family of target distributions:

$$\mathcal{Q}_P = \{Q : \text{follows the same SCM as } P, (\mu_1, \mu_2) \in \mathbb{R}^2\}$$

Objective: find linear model $h \in \mathbb{R}^2$ s.t.

$$h \in \arg \min \max_{Q \in \mathcal{Q}_P} \mathcal{R}_Q(h)$$

Causality for domain generalization: example

We explicitly compute the models and their risks:

$$h_1 = \arg \min_{h=ce_1} \mathcal{R}_P(h); \quad h_2 = \arg \min_{h=ce_2} \mathcal{R}_P(h); \quad h_{1,2} = \arg \min_{h \in \mathbb{R}^2} \mathcal{R}_P(h)$$

1. $h_1 = (1,0), \quad \mathcal{R}_P(h_1) = 1, \quad \mathcal{R}_Q(h_1) = 1$ **causal model**

2. $h_2 = (0,2/3), \quad \mathcal{R}_P(h_1) = 2/3, \quad \mathcal{R}_Q(h_1) = 2/3 + \frac{(\mu_1 - 2\mu_2)^2}{9}$ **anticausal model**

3. $h_{1,2} = (1/2,1/2), \quad \mathcal{R}_P(h_{1,2}) = 1/2, \quad \mathcal{R}_Q(h_{1,2}) = 1/2 + \mu_2^2/4$

**OLS model
(unconstrained)**

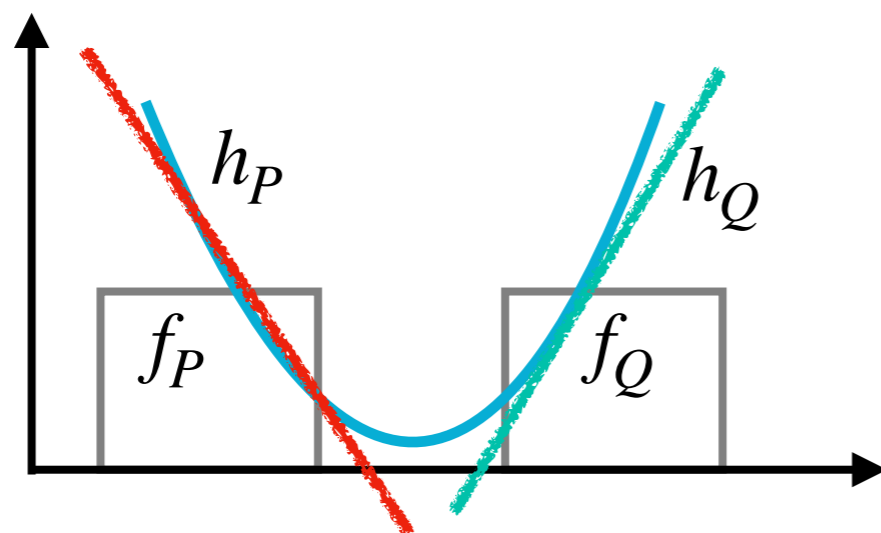
Causal models result in bounded target risks for a wide variety of shifts!

Causality for domain generalization

Many modern domain generalization works focus on identifying and learning the invariant causal predictor / model.

However, the framework runs into multiple **limitations**:

1. Identifying the parent set of the target variable is **hard**, requires strong assumptions – if they are not satisfied, wrong set can be discovered
2. Some variables might be **unobserved**, resulting in no or partial invariances
3. **Finite-sample** and **learnability** aspects are completely ignored: recall first part of the lecture! Even if one correctly identifies the parent set $\text{Pa}(Y)$, no guarantee that the resulting causal model trained on finitely many samples / for a misspecified function class will have a bounded target risk!



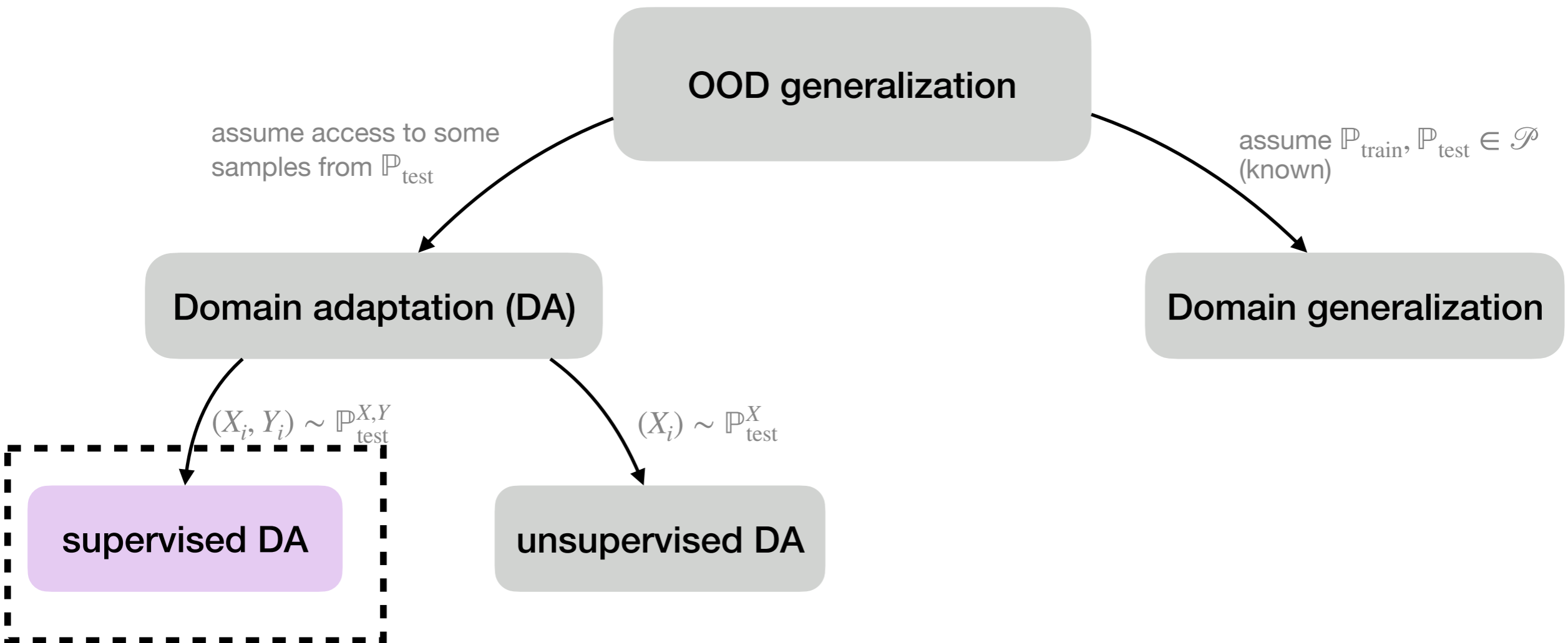
Here, the ground truth is not learnable by \mathcal{H} (linear functions) and the supports of P_X and Q_X are disjoint.

Causality for domain generalization

However, also has advantages over unstructured domain adaptation:

1. Offers a **realistic**, structured view on **what can shift** between source and target
2. Provides **methods to find models**, of which at least one often will have stable target risk

Taxonomy of OOD generalization



Can causality improve (supervised) domain adaptation?

- **Setting:** Linear regression setting with $X \in \mathcal{X} \subset \mathbb{R}^d$, $Y \in \mathbb{R}$, squared loss.
- Fix $\mathcal{H} = \{h \in \mathbb{R}^d : \|h\|_2 \leq B\}$ and for any $I \subset [d]$, define $\mathcal{H}_I = \{h \in \mathcal{H} : h_{I^c} = 0\}$
- Let P denote the source and Q the target distribution
- For $D \in \{P, Q\}$ let $h_{I,D} \in \arg \min_{h \in \mathcal{H}_I} \mathcal{R}_D(h)$ be the **best model on D only using features X_I** .

Can causality improve (supervised) domain adaptation?

- We are given a **source dataset** $\mathcal{D}_P = \{(X_i, Y_i)\}_{i \in [n_P]}$ and a (small) **target dataset** $\mathcal{D}_Q = \{(X_i, Y_i)\}_{i \in [n_Q]}$
- Further, we are given a set \mathcal{I} of feature sets I , i.e. $\mathcal{I} \subset 2^{[d]}$ and the corresponding **set of candidate models**

$$\{h_{I,P} : I \in \mathcal{I}\}$$

- For simplicity, let us assume an idealized setting in which $n_P = \infty$
- Further notation: $\Sigma_D := \mathbb{E}_D[XX^\top]$ for $D \in \{P, Q\}$
- "best-generalizing source model from collection":

$$h_{(1),P} = \arg \min_{h_{I,P}: I \in \mathcal{I}} \mathcal{R}_Q(h_{I,P})$$

Theorem [KJBKY 25']. Assume for $D \in \{P, Q\}$ X sub-Gaussian with $\|X\|_{\psi_2} \leq C_0$, moreover $\|Y - \mathbb{E}_D[Y | X_I]\|_{\psi_2} \leq \sigma_Y^2$ for all $I \subset [d]$ and $\lambda_{\min}(\Sigma_D) \geq \lambda_0$. Let $\delta > 0$. Define

$$\mathcal{F}_{\text{acc}} := \left\{ h_{I,P} : I \in \mathcal{F}, \|h_{I,P} - \hat{h}_Q\|_{\hat{\Sigma}_Q}^2 \leq C \frac{d + 1/\delta}{n_Q} \right\}. \text{ Let } M = |\mathcal{F}_{\text{acc}}|.$$

Assume that $\mathcal{E}_Q(h_{(1),P}) \leq C \frac{\log(M/\delta)}{n_Q}$.

Define $\mathcal{F}_{\text{good}} := \left\{ h_{I,P} : \Delta_I := \mathcal{E}_Q(h_{I,P}) - \mathcal{E}_Q(h_{(1),P}) \leq C \frac{\log(M/\delta)}{n_Q} \right\}$.

Then there exists an adaptive procedure \tilde{h} , which depends only on $\mathcal{D}_P, \mathcal{D}_Q, \mathcal{F}$ and δ , for which with probability at least $1 - \delta$

$$\mathcal{E}_Q(\tilde{h}) \leq \min \left\{ \sup_{I \in \mathcal{F}_{\text{good}}} \mathcal{E}_Q(h_{I,P}), C \frac{d + \log(1/\delta)}{n_Q} \right\}.$$

Procedure

Algorithm 2 Iterative localized aggregation

Require: n_P samples from P ; n_Q samples from Q ; empirical target risk minimizer \hat{h}_Q ; collection of models $\{\hat{h}_{I,P} : I \in \mathcal{J}\}$; constants $C_1, C_2 > 0$; dimension d .

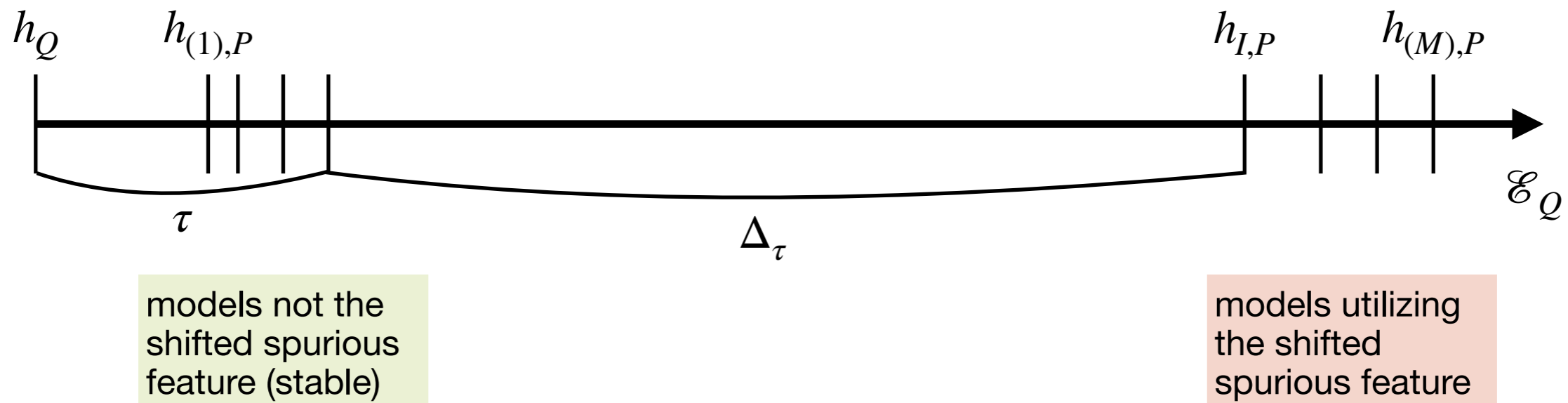
- 1: **Step 1 (preventing negative transfer).**
- 2: Define the initial candidate set

$$\mathcal{J}^{\text{acc}} \leftarrow \{ \hat{h}_{I,P} : I \in \mathcal{J}, \|\hat{h}_{I,P} - \hat{h}_Q\|_{\hat{\Sigma}_Q}^2 \leq C_1 d/n_Q \}.$$

- 3: **Step 2 (iterative aggregation and refinement).**
 - 4: **while** true **do**
 - 5: Compute aggregation estimator \bar{h} over \mathcal{J}^{acc} .
 - 6: $M \leftarrow |\mathcal{J}^{\text{acc}}|$.
 - 7: $\mathcal{J}^{\text{rej}} \leftarrow \{ \hat{h}_{I,P} \in \mathcal{J}^{\text{acc}} : \|\hat{h}_{I,P} - \bar{h}\|_{\hat{\Sigma}_Q}^2 \geq C_2 \frac{\log M}{n_Q} \}$.
 - 8: **if** $\mathcal{J}^{\text{acc}} \setminus \mathcal{J}^{\text{rej}} = \emptyset$ **or** $\mathcal{J}^{\text{rej}} = \emptyset$ **then**
 - 9: **return** \bar{h}
 - 10: **end if**
 - 11: $\mathcal{J}^{\text{acc}} \leftarrow \mathcal{J}^{\text{acc}} \setminus \mathcal{J}^{\text{rej}}$
 - 12: **end while**
-

Discussion

- The first step helps remove models which have worse rate than the retrained target estimator. In practice, this greatly reduces number of models
- The second step is an iterative refinement step aimed to further improve target risk in a favorable "gap" situation:



- Then if $n_Q \gtrsim \frac{\log(M/\delta)}{\Delta_\tau}$, we have w.h.p.

$$\mathcal{E}_Q(\tilde{h}) \leq \min \left\{ \tau, C \frac{d + \log(1/\delta)}{n_Q} \right\}$$

Outlook

- Real-world data often exhibits causal structure, however
 - it may be impossible to fully identify
 - it might not be enough for domain adaptation in finite samples
- To avoid making mistakes or being too conservative on target environments, one has to have access to some labelled target data
- How to use this data most efficiently in presence of limited causal knowledge?
- Can one find good candidate sets so that they are guaranteed to contain at least one well-generalizing model?

