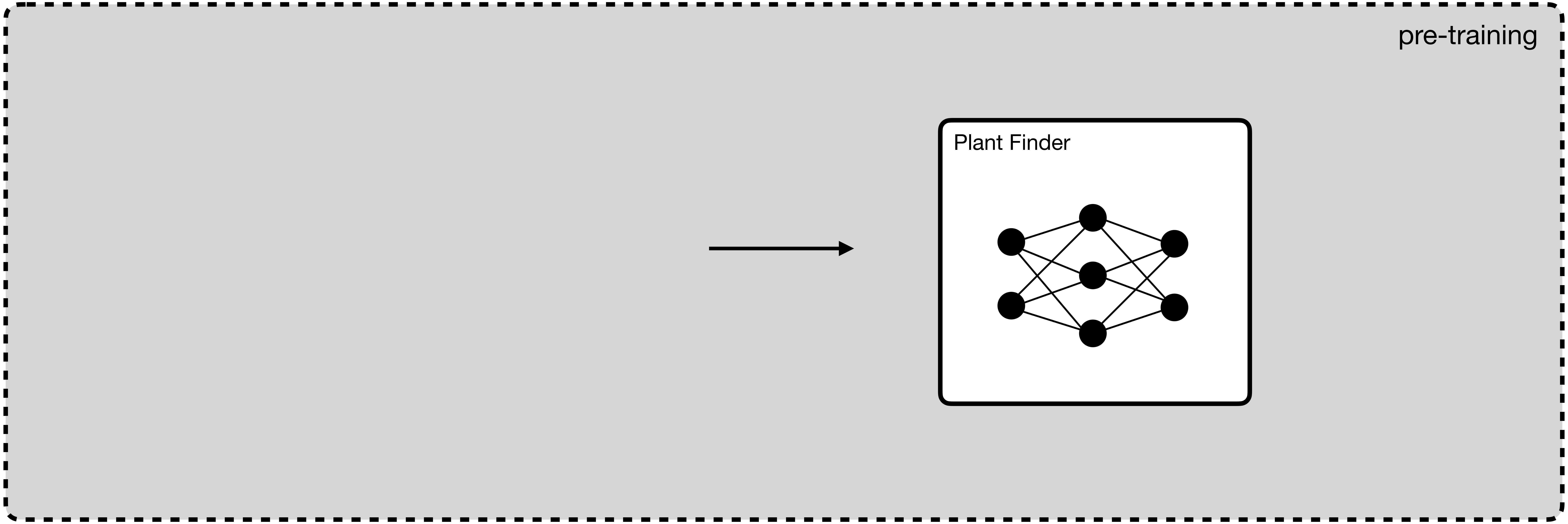
Transductive Active Learning: Theory and Applications

Jonas Hübotter, Bhavya Sukhija, Lenart Treven,
Yarden As, Andreas Krause

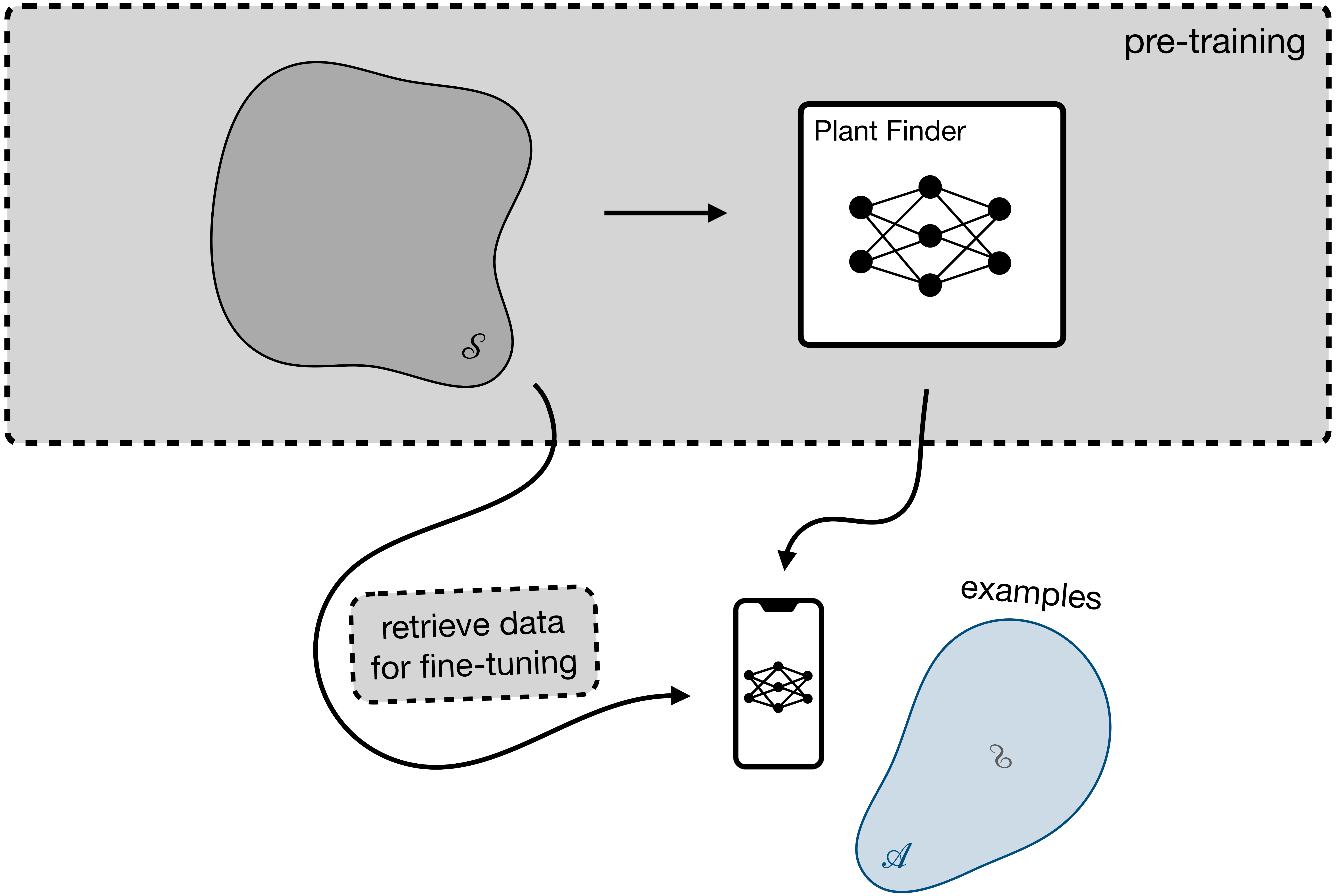
ETH zürich




Motivation



Motivation



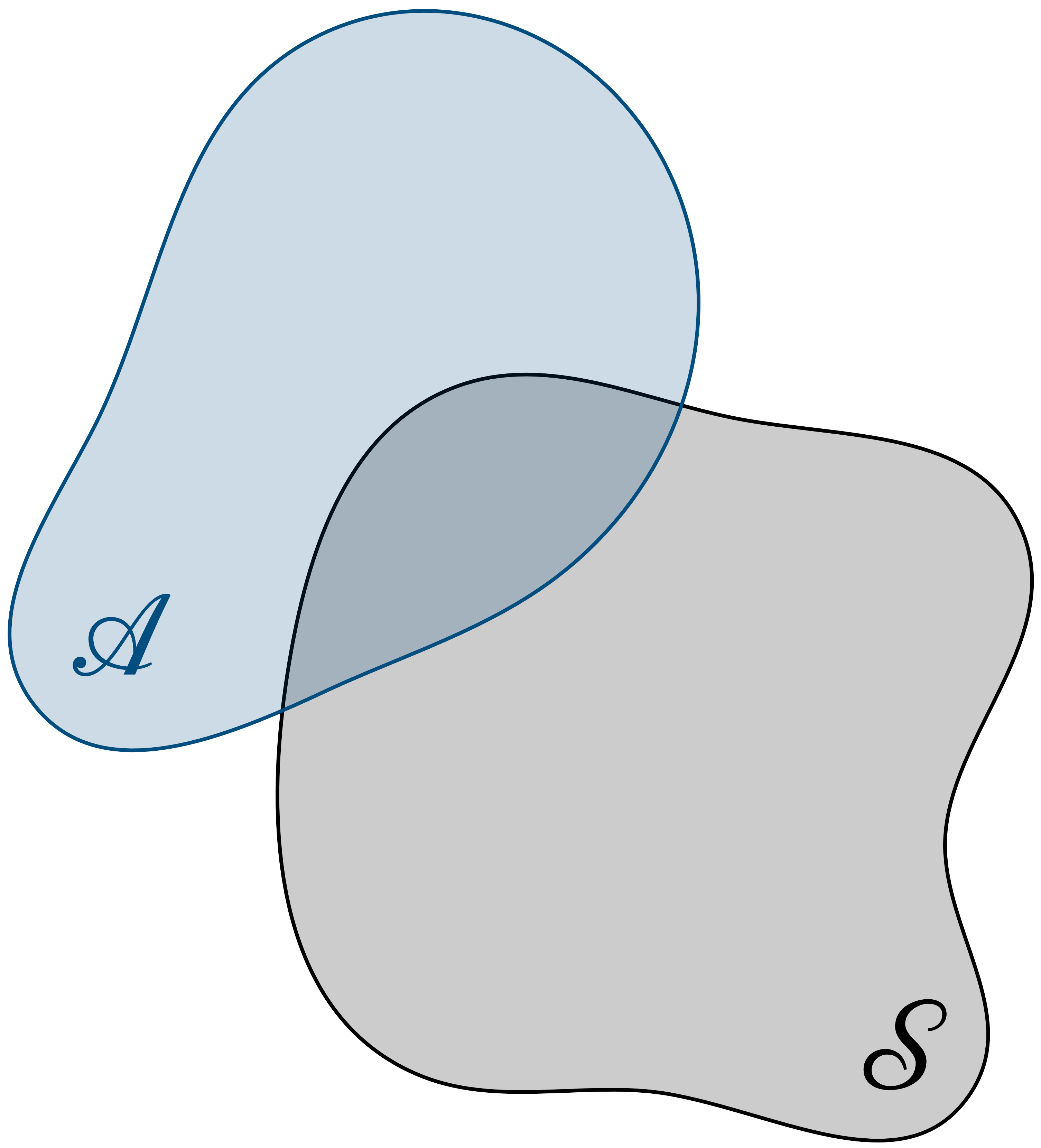
pre-training
Plant Finder
$\mathcal{S}$
retrieve data
for fine-tuning
examples
$\mathcal{A}$

Setting

- $\mathscr{S} \subseteq \mathscr{X}$ - sample space
- $\mathscr{A} \subseteq \mathscr{X}$ - target space
- Unknown function f over $\mathscr{X}$

Goal: Learn f within $\mathscr{A}$ by sampling from $\mathscr{S}$

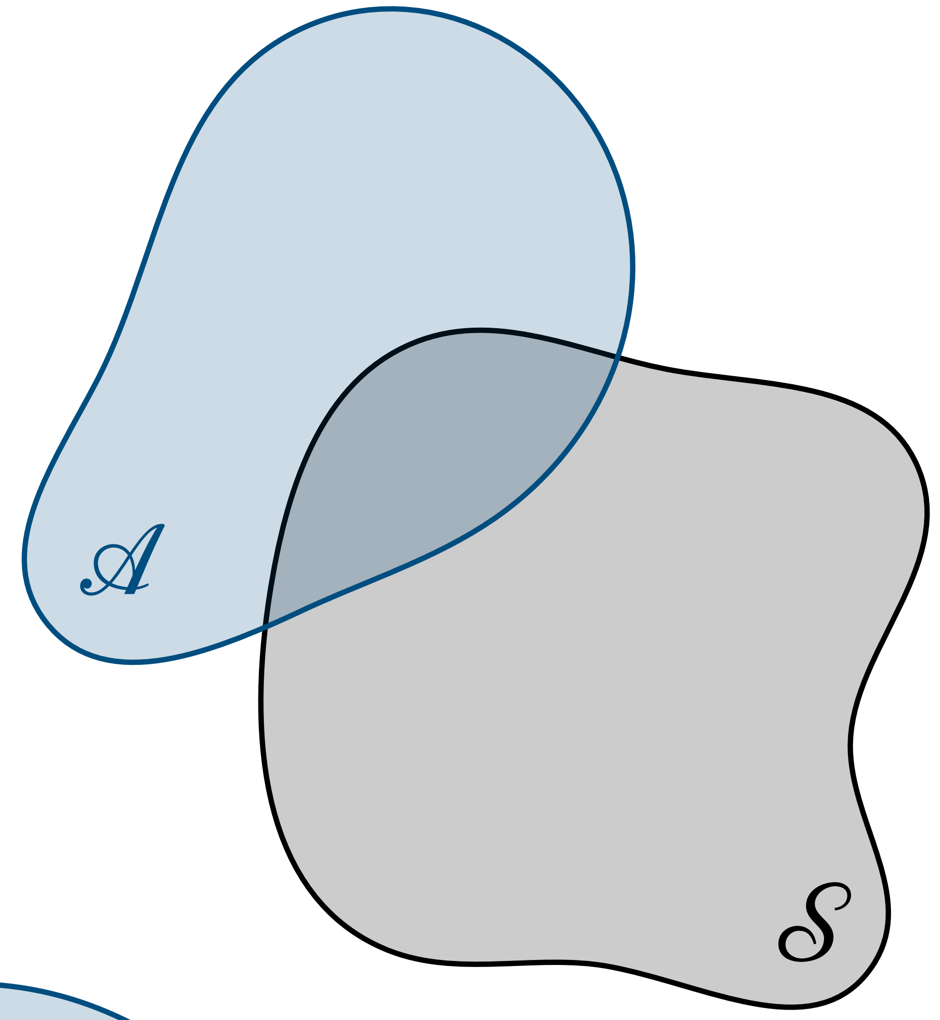
We call this **Transductive Active Learning**



Transductive Active Learning

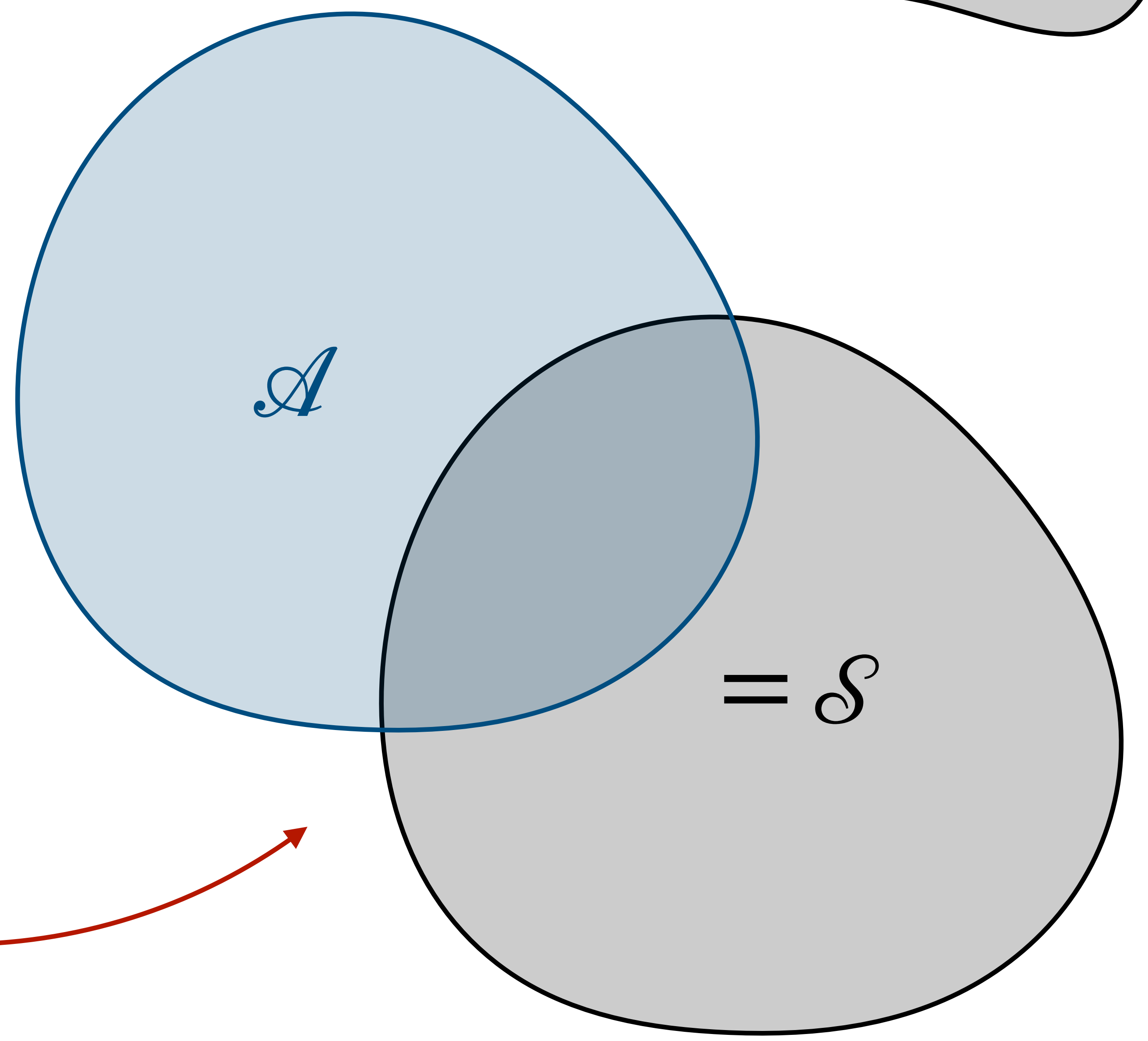
“only learn what is needed to solve a given task”

[MacKay, 1992]



(Inductive) Active Learning

“learn as much as you can”



studied in most prior works

Algorithms for transductive active learning

Probabilistic model of f:

- prior $p(f)$
- likelihood $p(D \mid f)$ of data D
- posterior $p(f \mid D)$

Algorithms: select data to minimize *posterior* uncertainty within $\mathscr{A}$
[MacKay, 1992]

Contributions

Algorithms: select the next sample to minimize *posterior* uncertainty within $\mathscr{A}$
[MacKay, 1992]

When f is a *Gaussian process* these algorithms are tractable:

- **Theory:** rates for the uniform convergence of uncertainty over $\mathscr{A}$
- **Applications:**
    - (1) active fine-tuning of neural networks
    - (2) safe Bayesian optimization

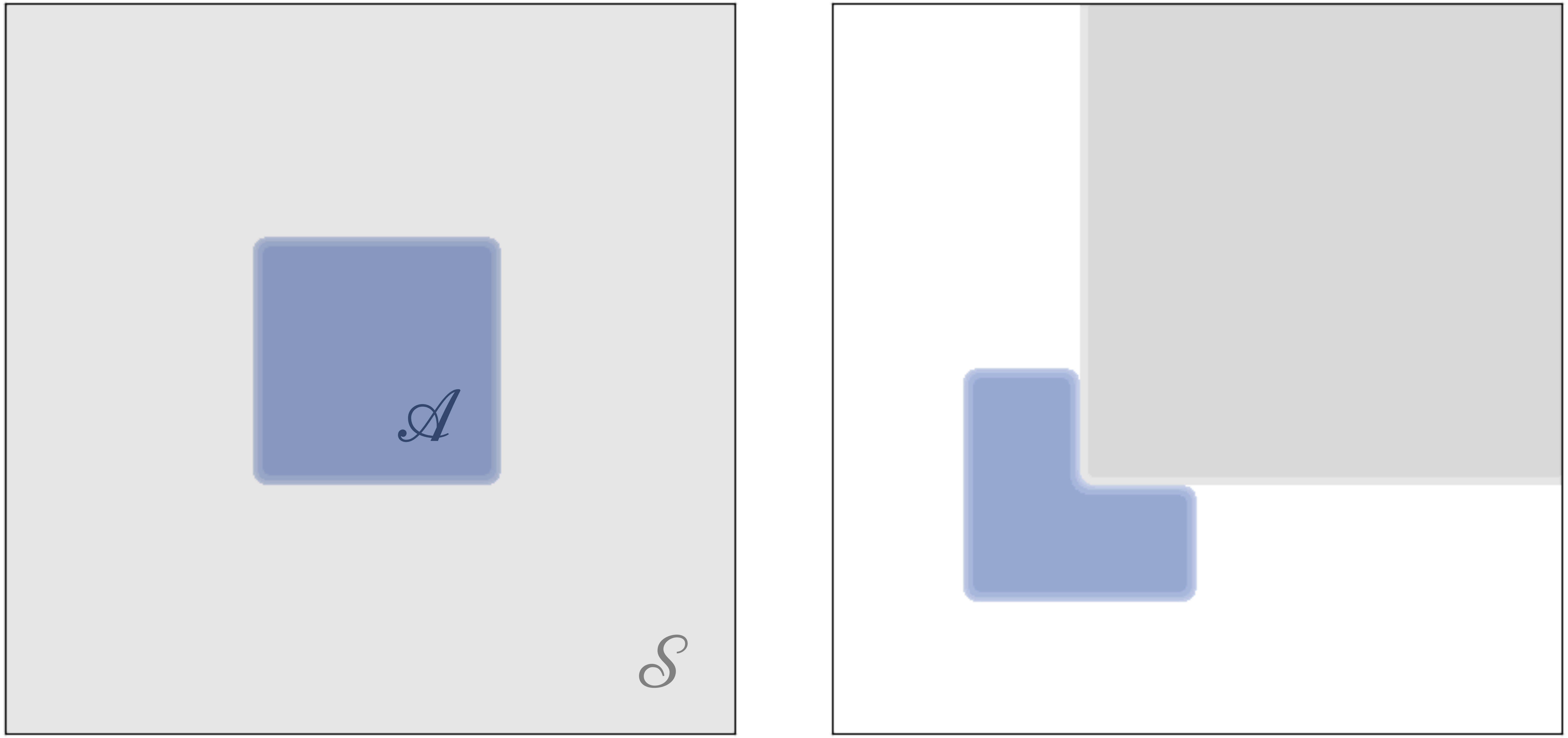
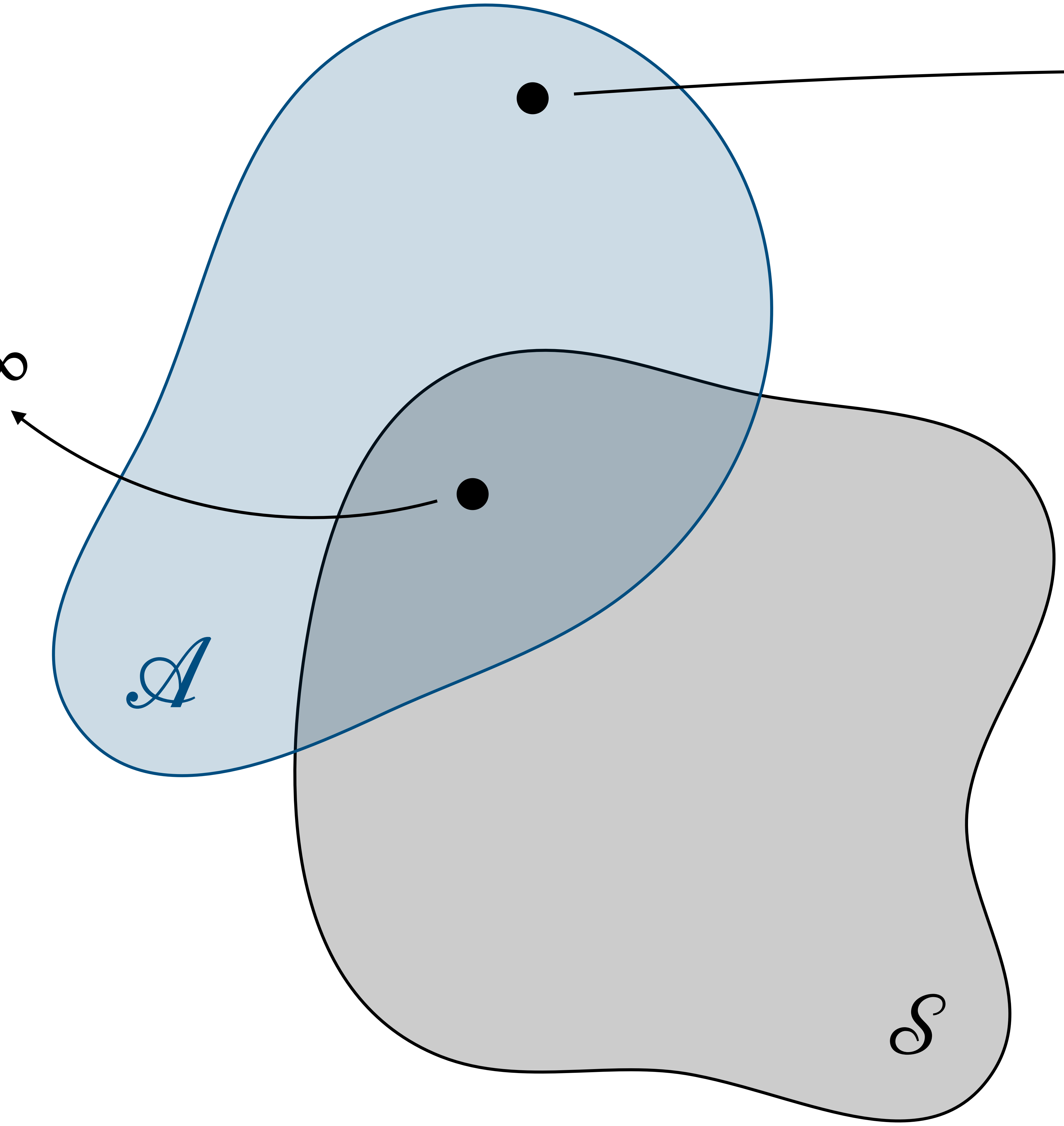


Illustration on a Gaussian process with RBF kernel

$\sigma_n^2(\boldsymbol{x})$ = posterior variance at $\boldsymbol{x} \in \mathscr{A}$

$\sigma_n^2(\boldsymbol{x}) \to 0$ as $n \to \infty$

e.g., by repeatedly sampling $\boldsymbol{x}$



what about the point $\boldsymbol{x}'$?

$\sigma_n^2(\boldsymbol{x}') \geq \eta_{\mathscr{S}}^2(\boldsymbol{x}') = \mathrm{Var}[f(\boldsymbol{x}') \mid \boldsymbol{f}(\mathscr{S})]$

is the **irreducible uncertainty**:

Theory

Uncertainty bound (informal)

For every $\boldsymbol{x}' \in \mathscr{A}$: $\quad \sigma_n^2(\boldsymbol{x}') - \eta_{\mathcal{S}}^2(\boldsymbol{x}') \leq C\gamma_{\mathcal{A},\mathcal{S}}(n)/\sqrt{n}$

irreducible

$\rightarrow 0$ for many kernels

where $\gamma_{\mathscr{A},\mathcal{S}}(n) = \max_{\substack{X \subseteq \mathcal{S} \\ |X| = n}} \mathrm{I}(\boldsymbol{f}(\mathscr{A}); \boldsymbol{y}(X))$

Agnostic error bound (informal)

If $f \in \mathscr{H}_k(\mathscr{X})$, then for every $\boldsymbol{x}' \in \mathscr{A}$ with probability at least $1 - \delta$:

$$|f(\boldsymbol{x}') - \mathbb{E}[f(\boldsymbol{x}') \mid D_n]|^2 \leq \beta_n^2(\delta)\left[\eta_{\mathcal{S}}^2(\boldsymbol{x}') + C\gamma_{\mathscr{A},\mathcal{S}}(n)/\sqrt{n}\right]$$

prediction

irreducible

reducible

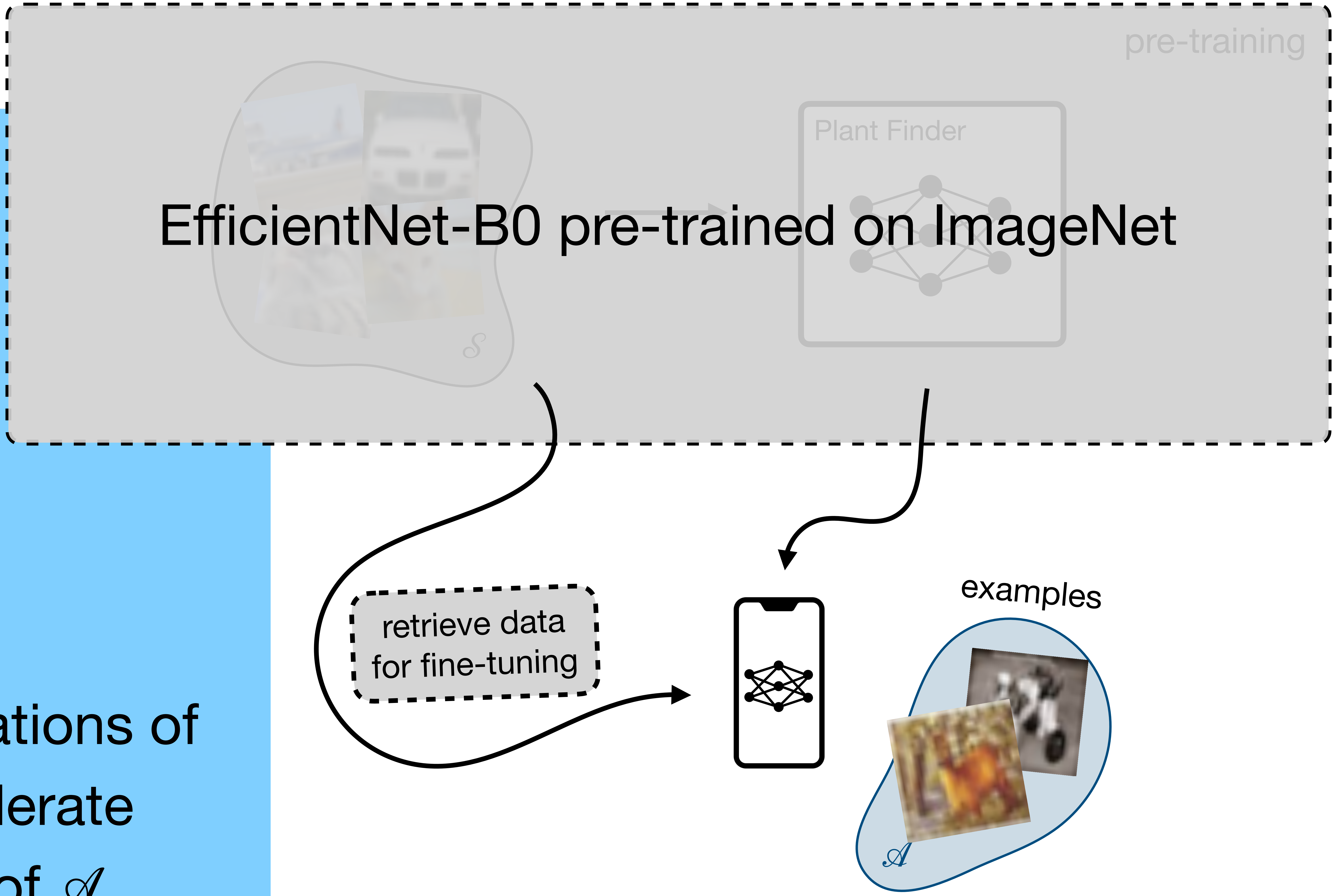
Applications

❶ Active fine-tuning

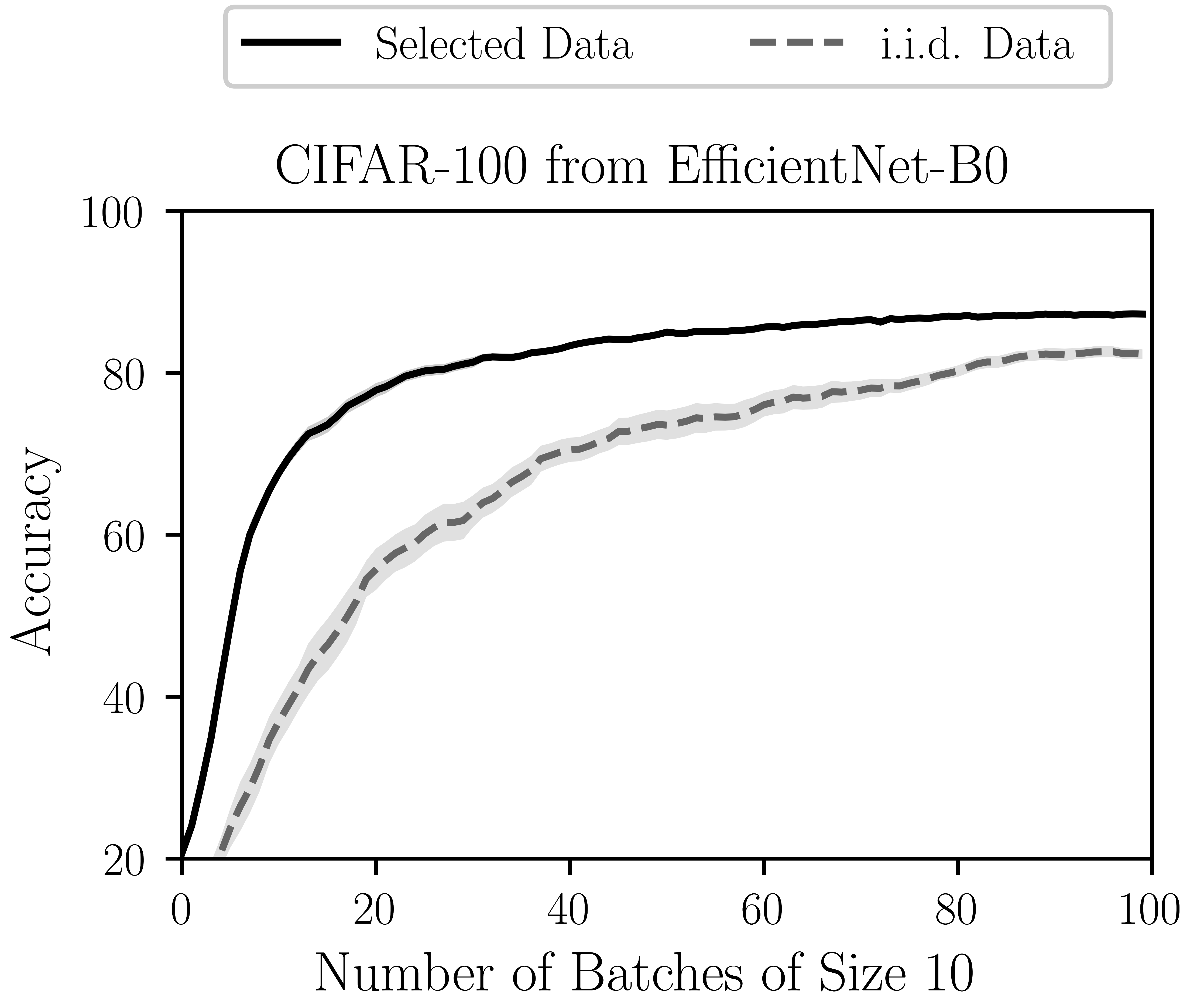
Given:

- pre-trained model
- $\mathscr{S}$ - training set
- $\mathscr{A}$ - test set

Goal: Leverage representations of pre-trained model to accelerate learning a good predictor of $\mathscr{A}$.



❶ Active fine-tuning of neural networks



Applications

❶ Active fine-tuning

Given:

- pre-trained model
- $\mathscr{S}$ - training set
- $\mathscr{A}$ - test set

Goal: Leverage representations of pre-trained model to accelerate learning a good predictor of $\mathscr{A}$.

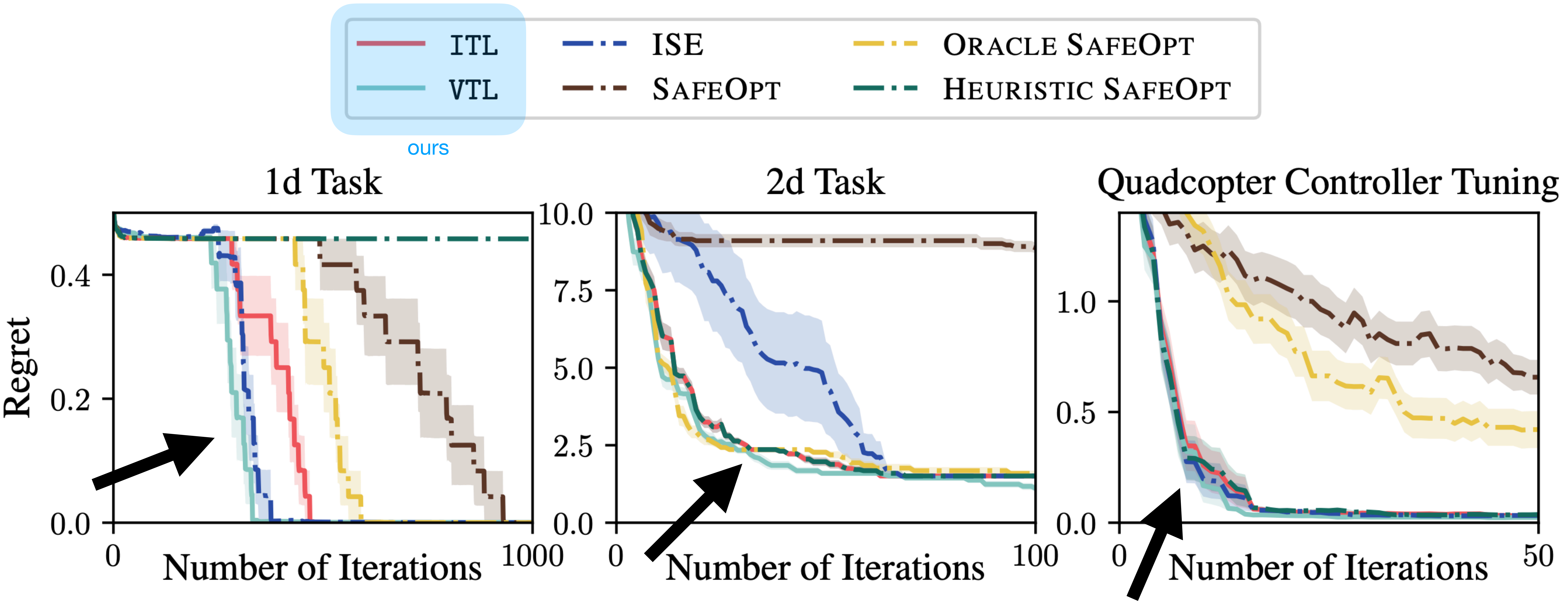
❷ Safe Bayesian optimization

Task: Optimize *unknown* function under *unknown* constraints that have to be satisfied at all times.

- $\mathscr{S}_n$ - pessimistic safe set
- $\mathscr{A}_n$ - set of potential safe optima

Theory: *Tighter* guarantees that *generalize* to continuous settings.

❷ Safe Bayesian optimization



Thanks for your attention!

jonas.huebotter@inf.ethz.ch