Weak-to-Strong Generalization Through the Data-Centric Lens

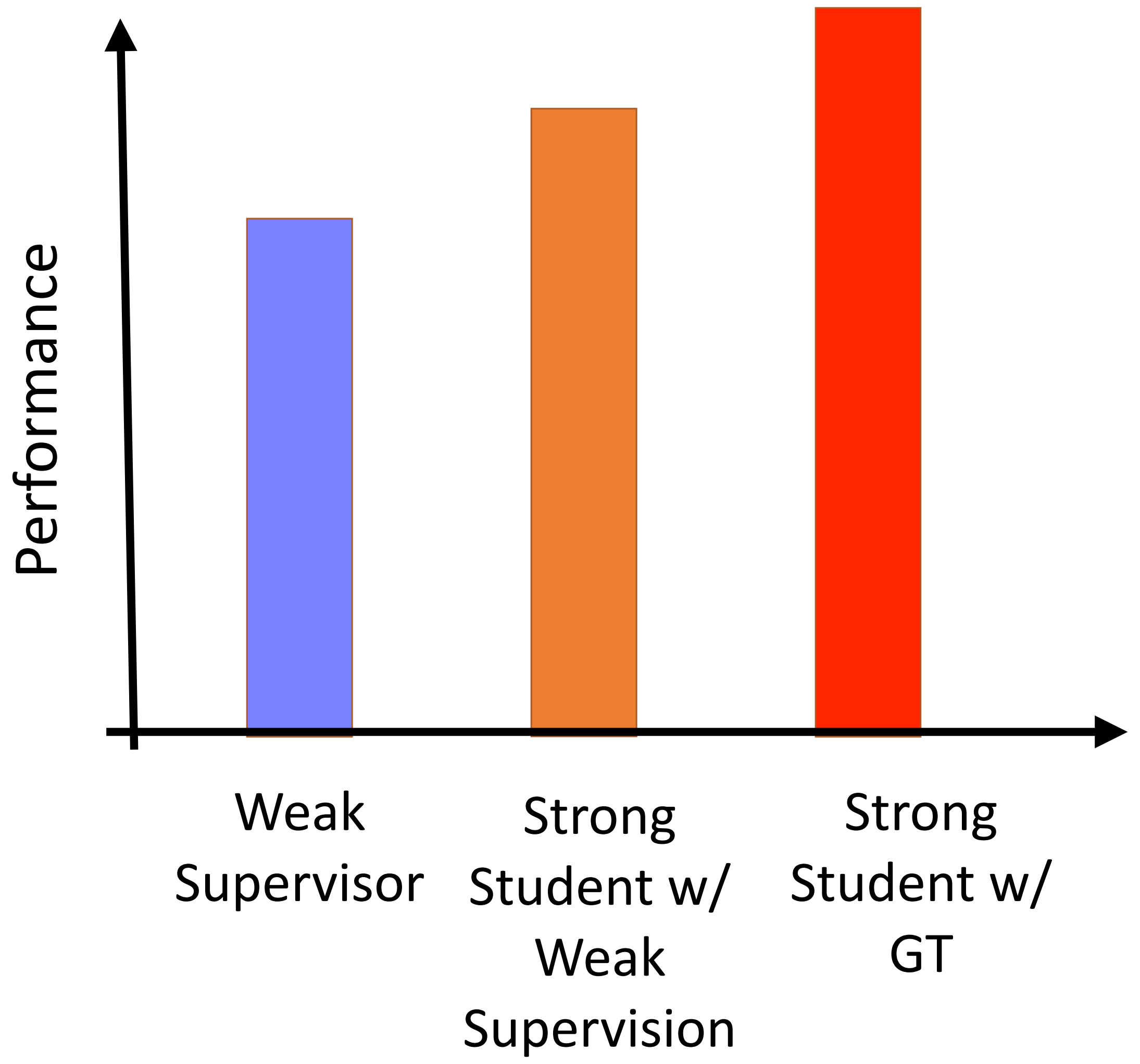
Changho Shin, John Cooper, Frederic Sala

University of Wisconsin–Madison

Paper Link

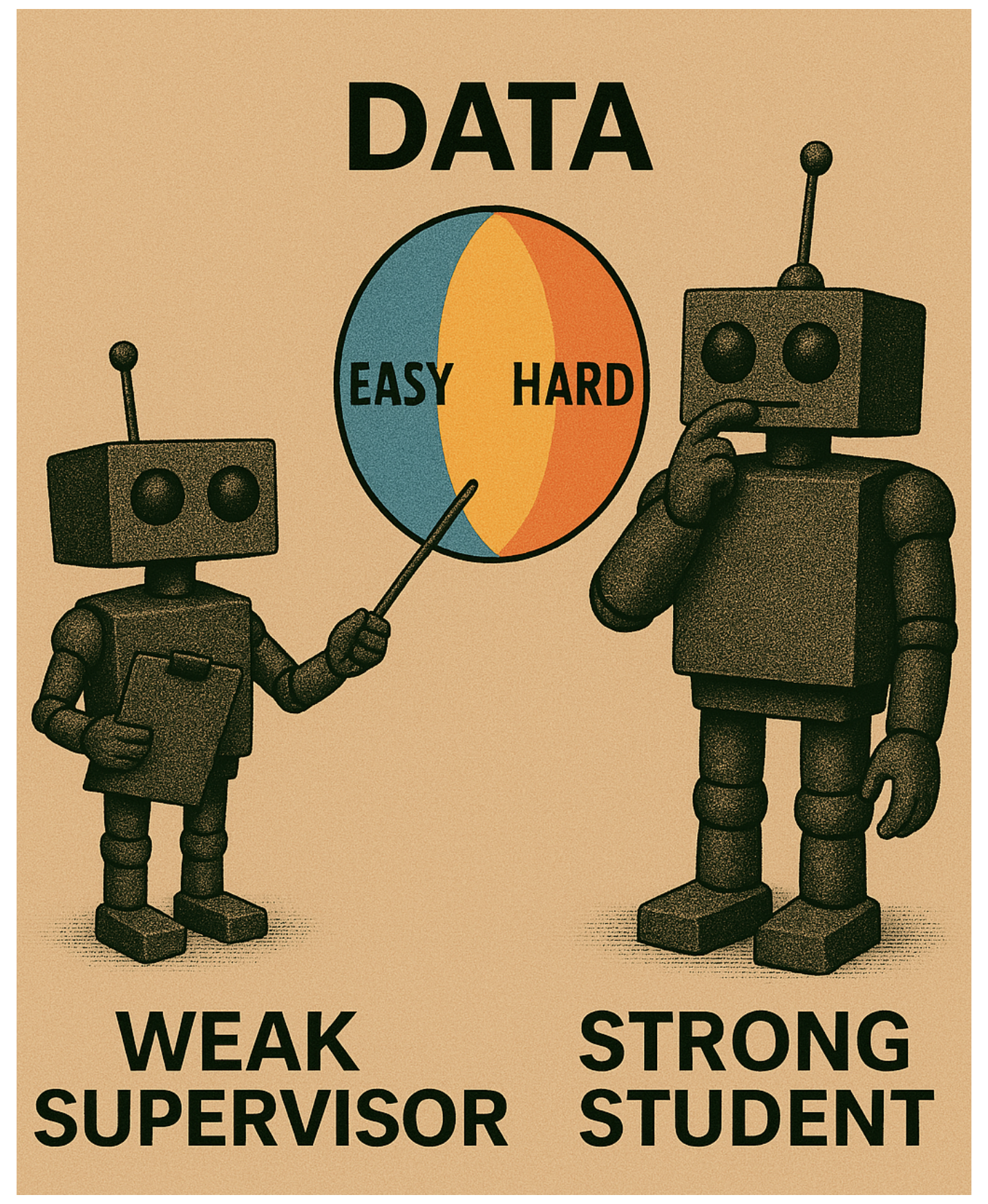
Background

Weak-to-Strong Generalization

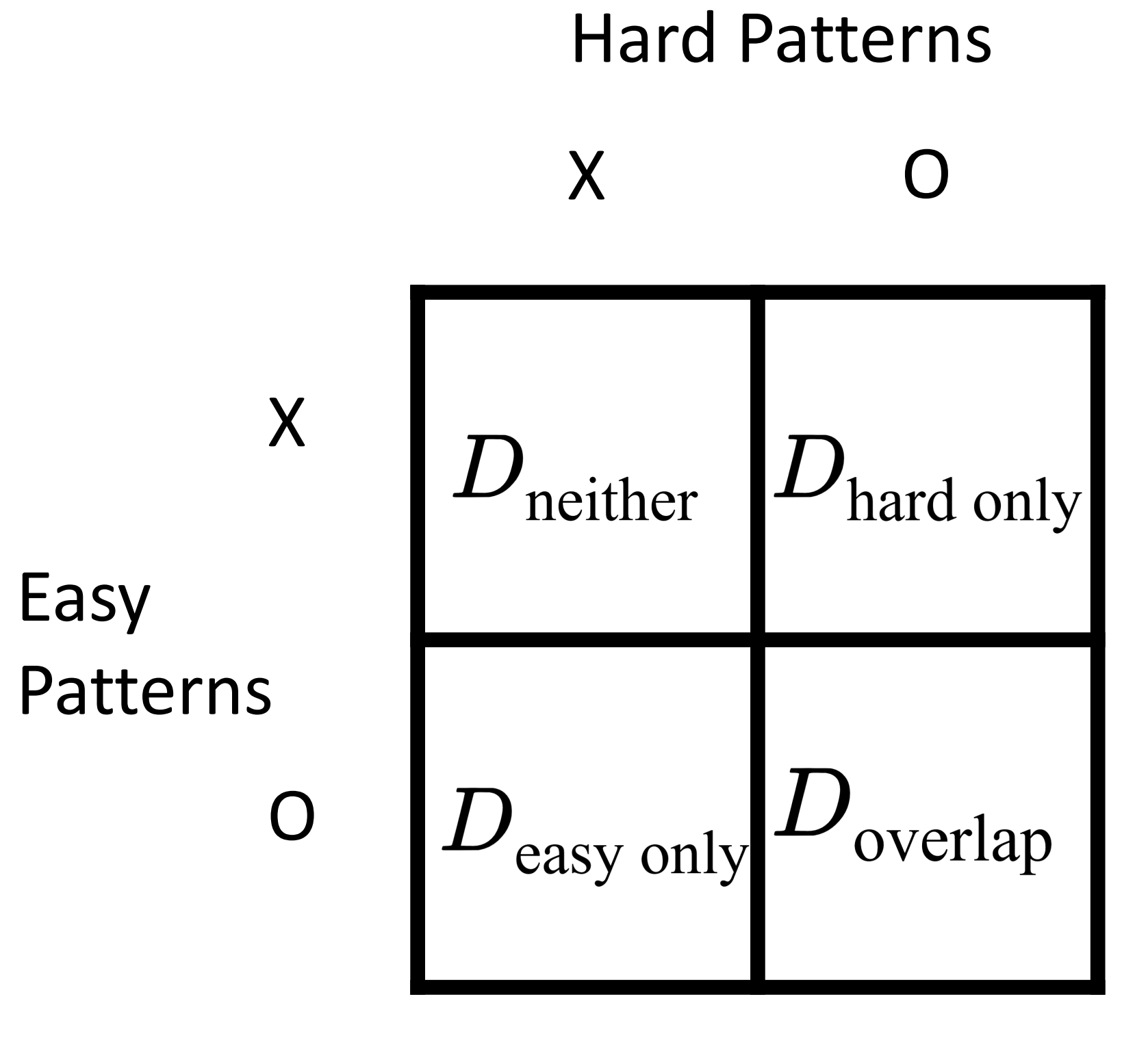


❓ How can a strong student outperform a weak supervisor?

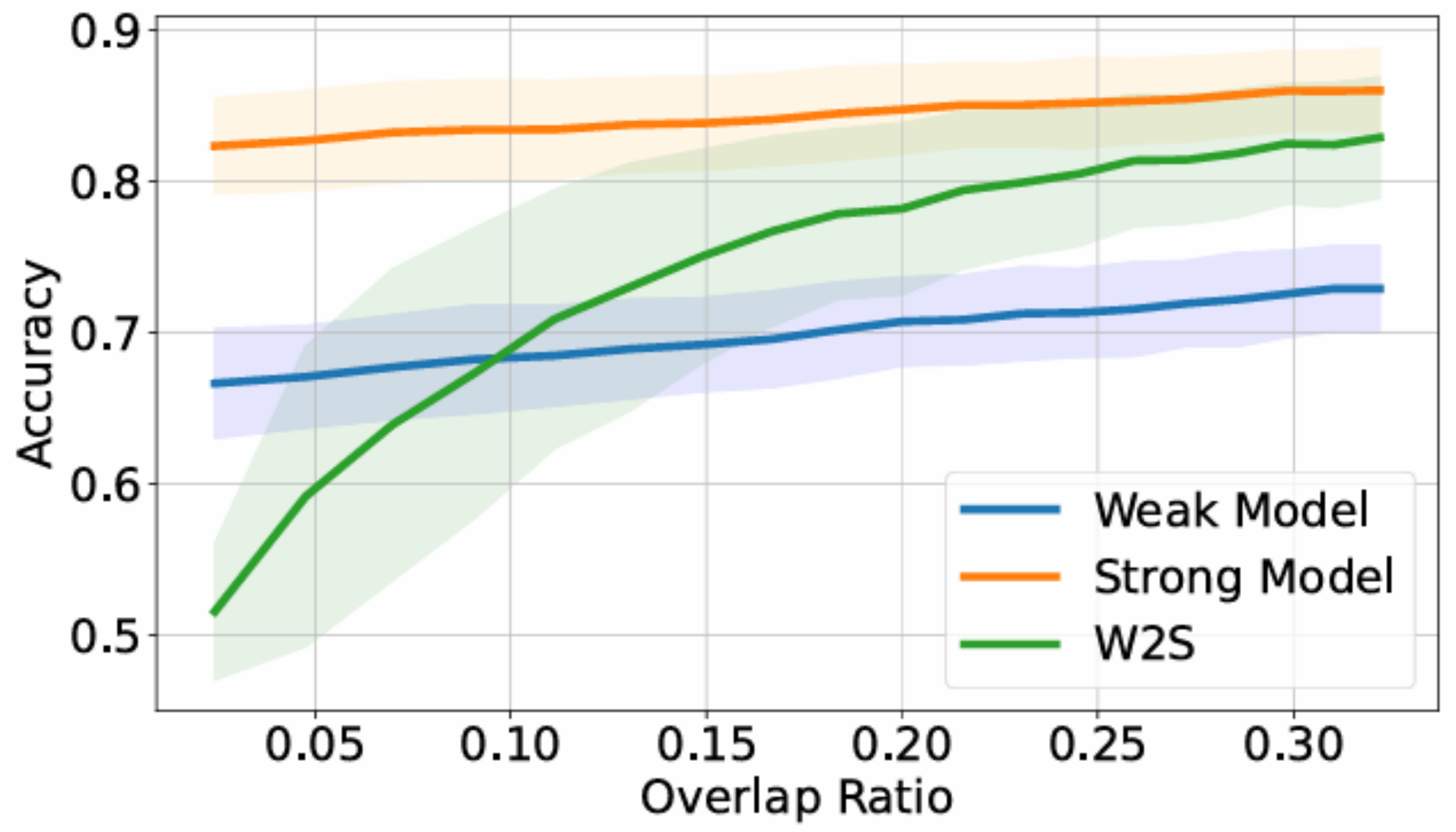
💡 We propose a "data mechanism" that drives weak-to-strong generalization.



Overlap Mechanism



Latent Data Partition

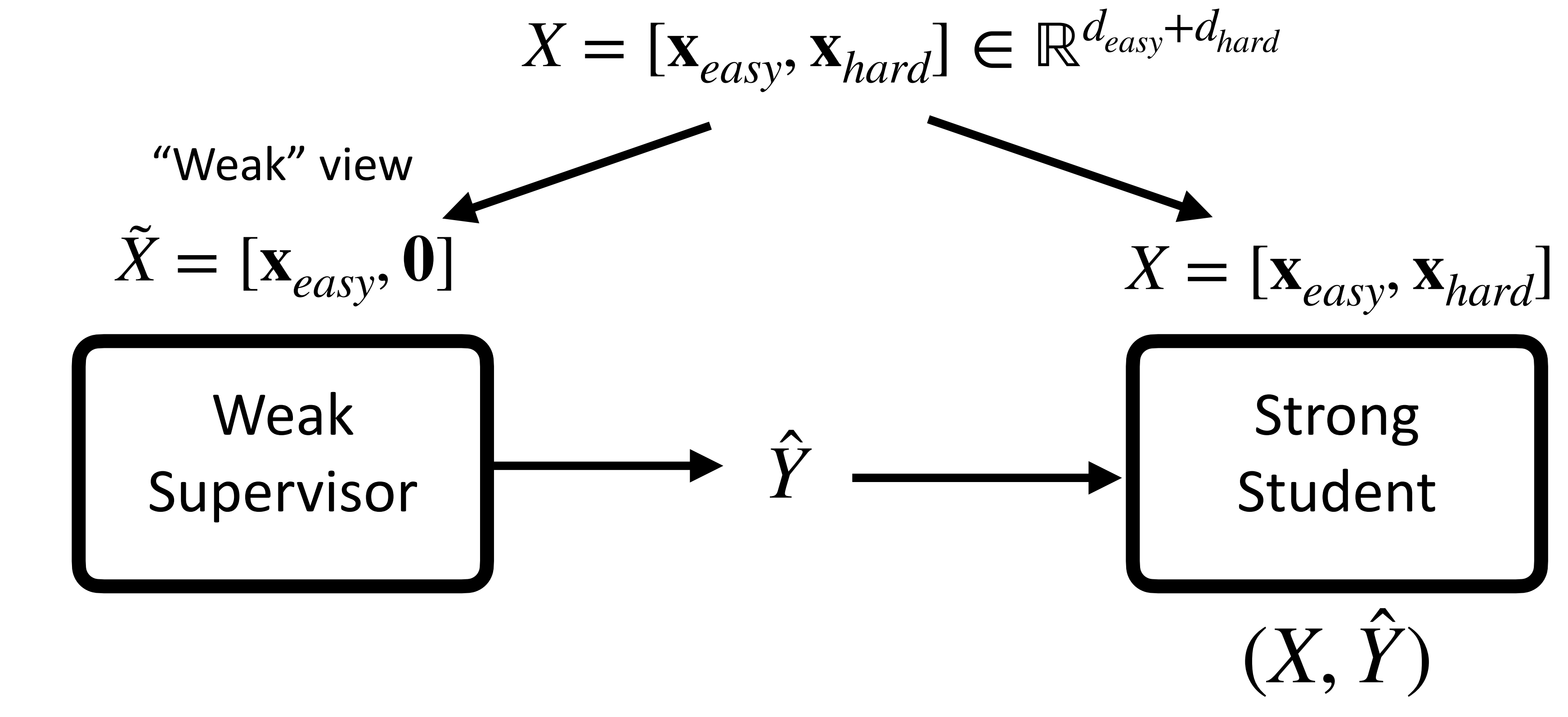


Overlapping points enable weak-to-strong generalization

1) **Weak supervisor** can **provide accurate pseudo-labels** based on **easy patterns**

2) **Strong student** can learn **hard patterns** using accurate pseudo-labels

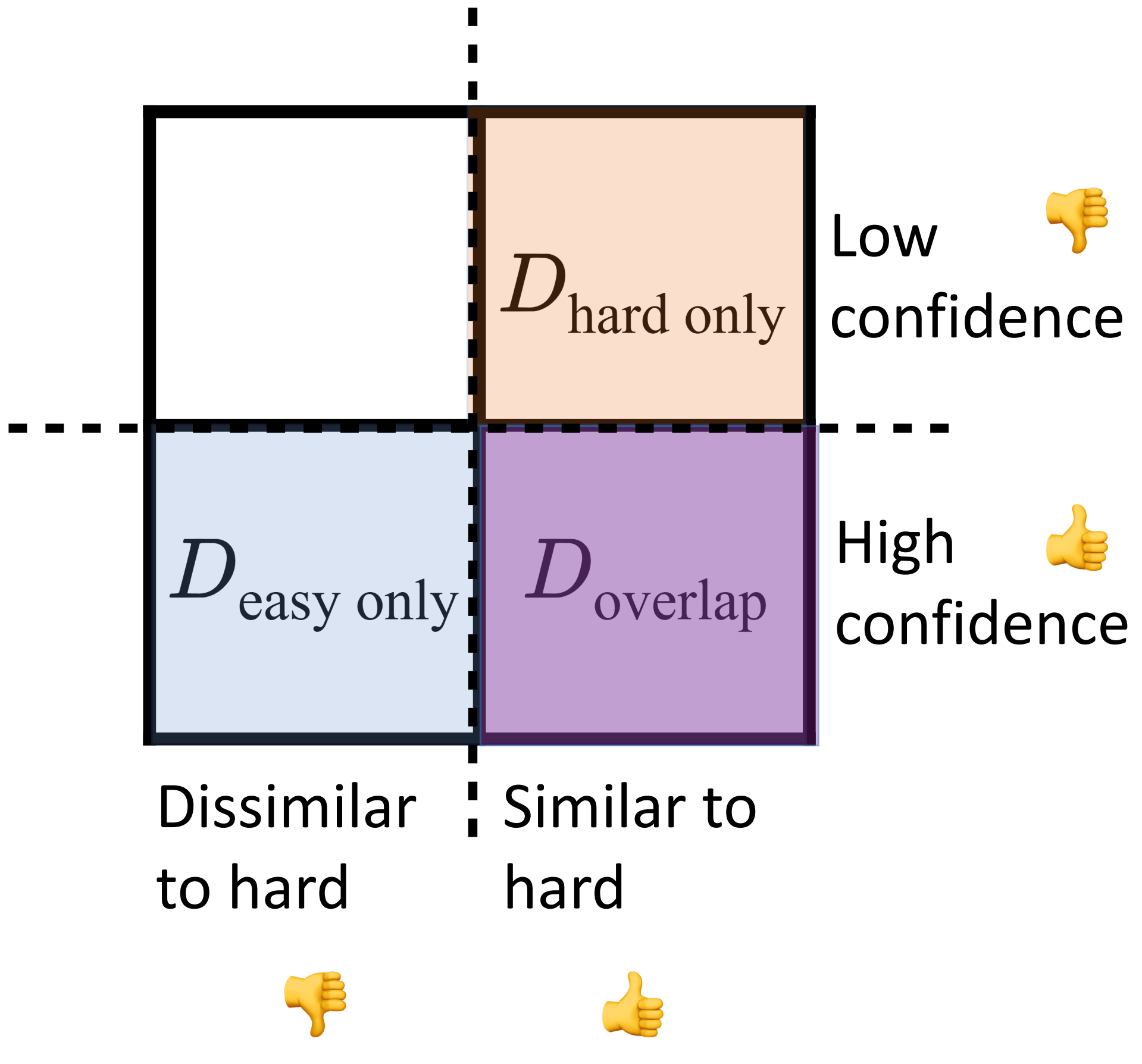
Data Model


$X = [\mathbf{x}_{easy}, \mathbf{x}_{hard}] \in \mathbb{R}^{d_{easy}+d_{hard}}$
"Weak" view
$\tilde{X} = [\mathbf{x}_{easy}, \mathbf{0}]$
$X = [\mathbf{x}_{easy}, \mathbf{x}_{hard}]$
Weak Supervisor
$\hat{Y}$
Strong Student
$(X, \hat{Y})$

$X_{easy\ only} = [x_{easy}, \mathbf{0}]$ $X_{hard\ only} = [\mathbf{0}, x_{hard}]$ $X_{overlap} = [x_{easy}, x_{hard}]$

Overlap Detection

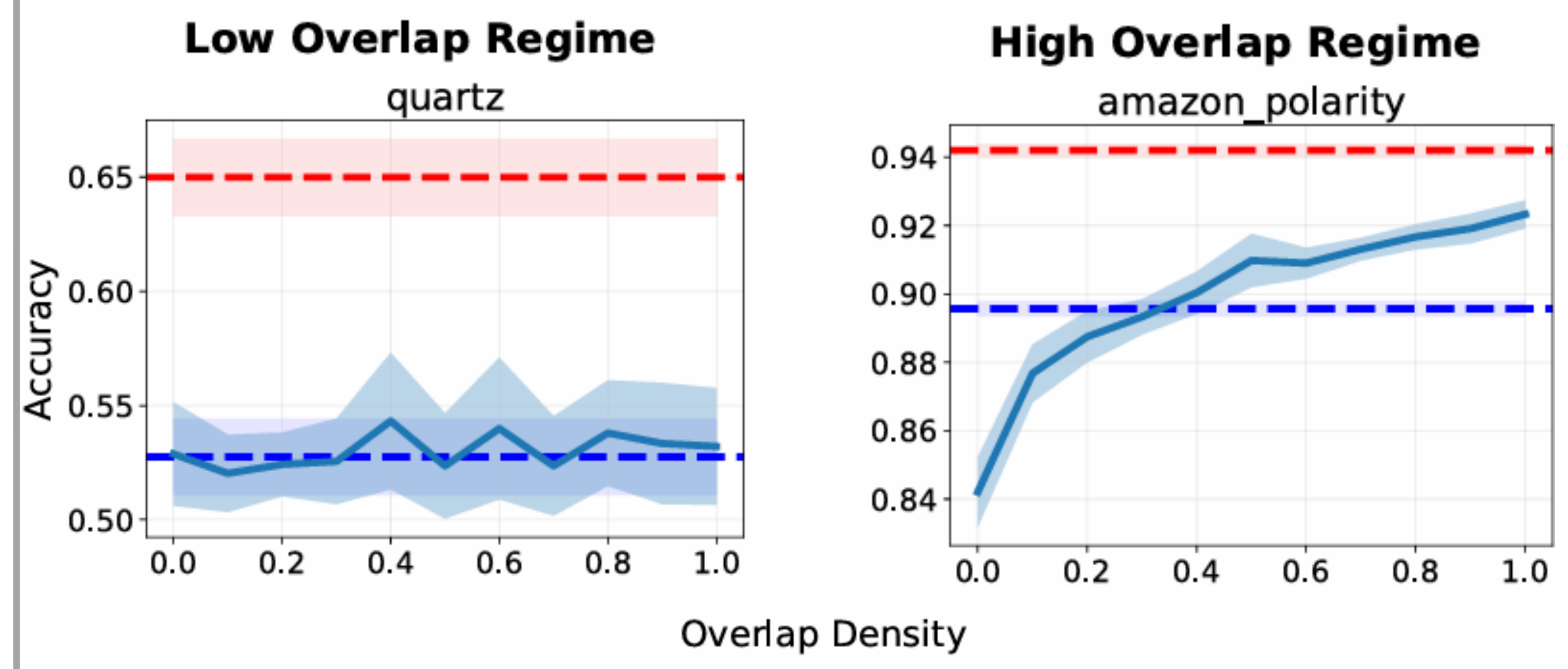
How can we find overlapping points in real-world data?



➤**Step 1**. Use **confidence scores** to identify examples likely to contain easy patterns (i.e., **high-confidence predictions from the weak supervisor**).

➤**Step 2.** Among the easy-pattern data, apply similarity-based filtering to **isolate examples that also resemble hard patterns**.

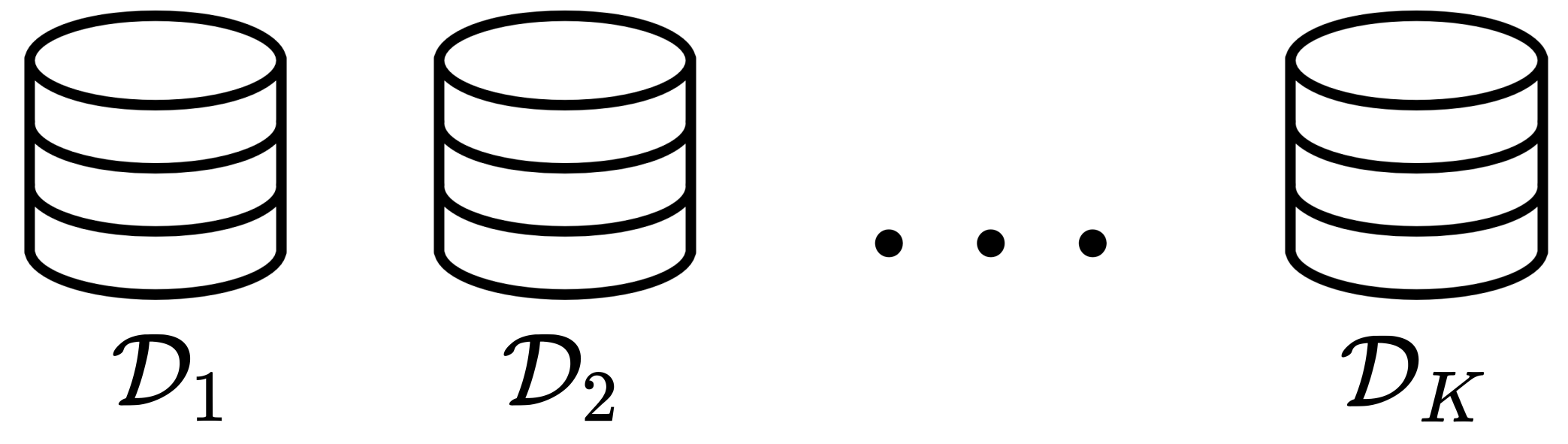
The extent of overlap is predictive of W2S generalization!



$n_{overlap} = 347$ $n_{overlap} = 3850$

Data Source Selection

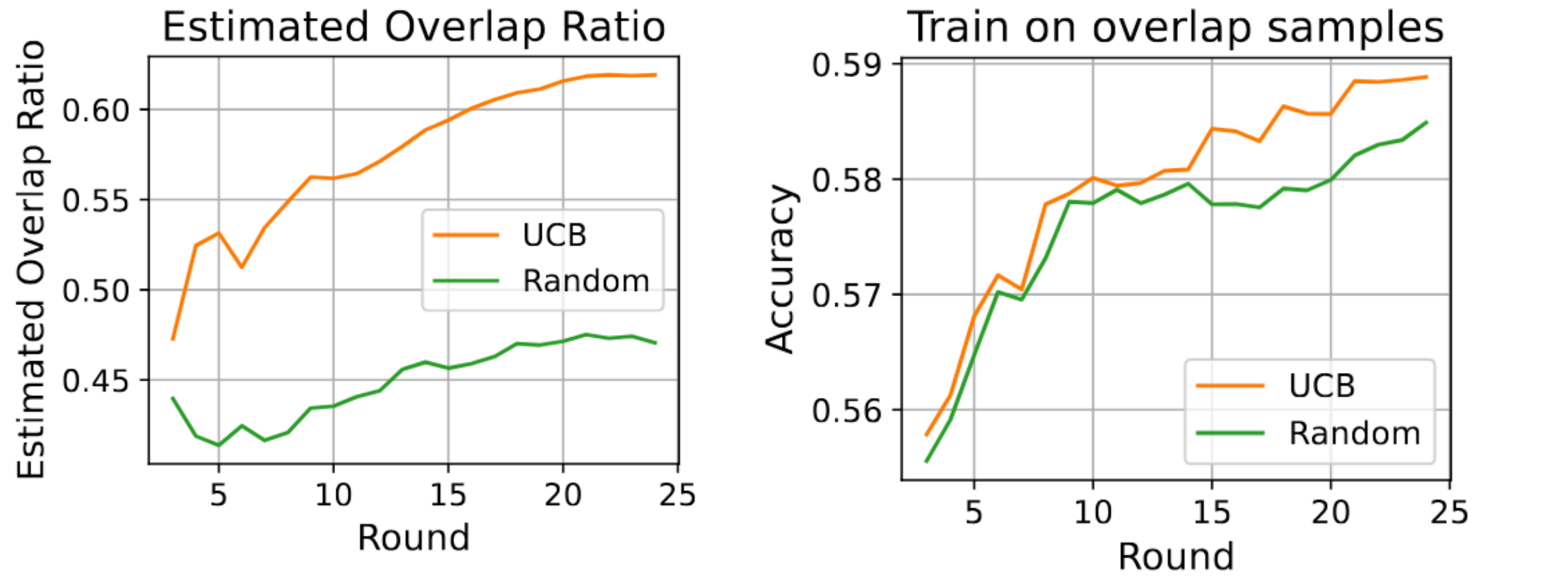
Setup: Sample data from multiple sources under a limited budget.



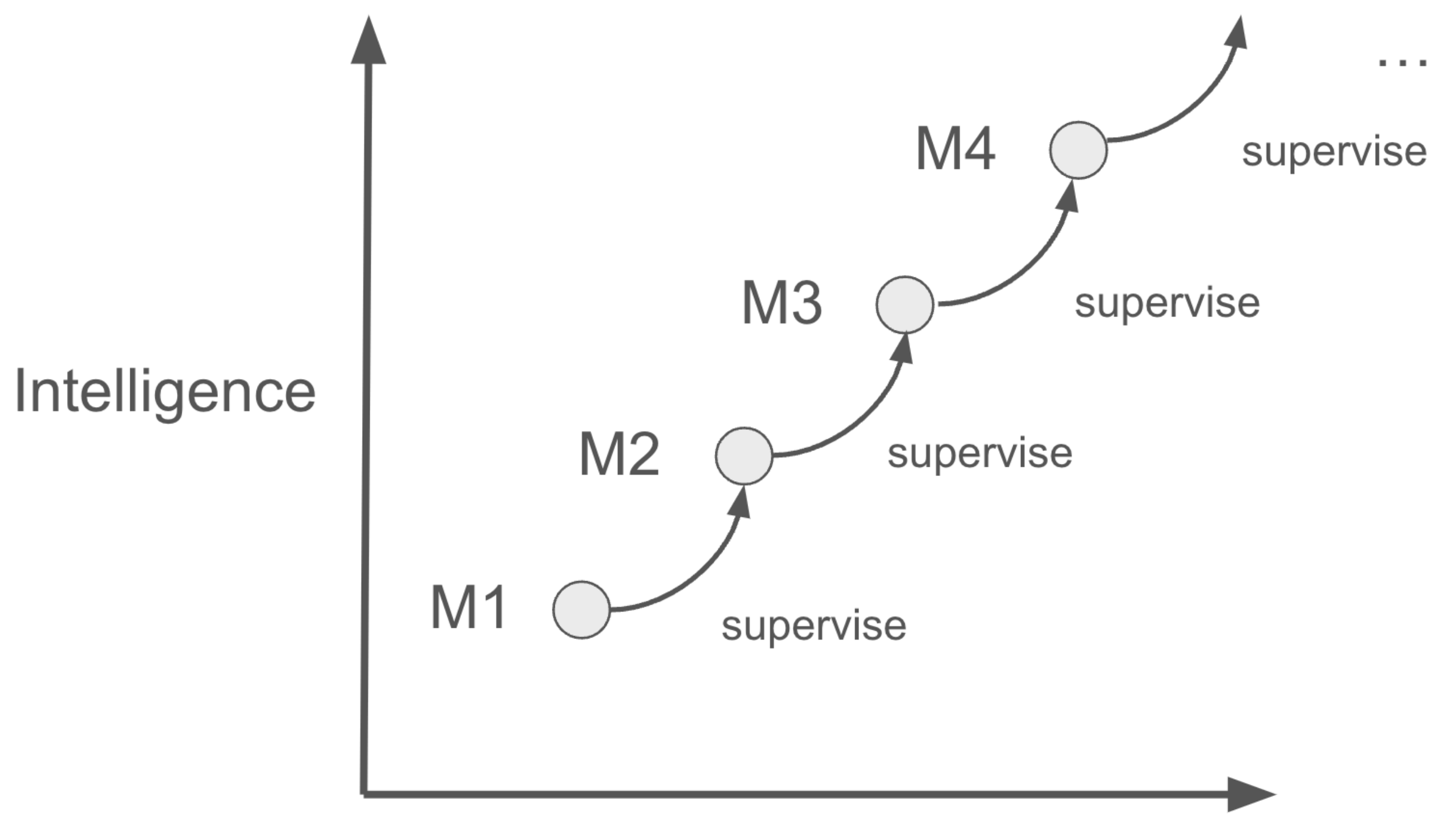
Goal: Maximize the overlap ratio.

Method: Use an Upper Confidence Bound (UCB) strategy to prioritize sources likely to yield high overlap.

Result: UCB increases overlap —> improves weak-to-strong generalization



Moving Forward...



- Generative tasks (e.g., reasoning)
- Synthetic data generation for W2S