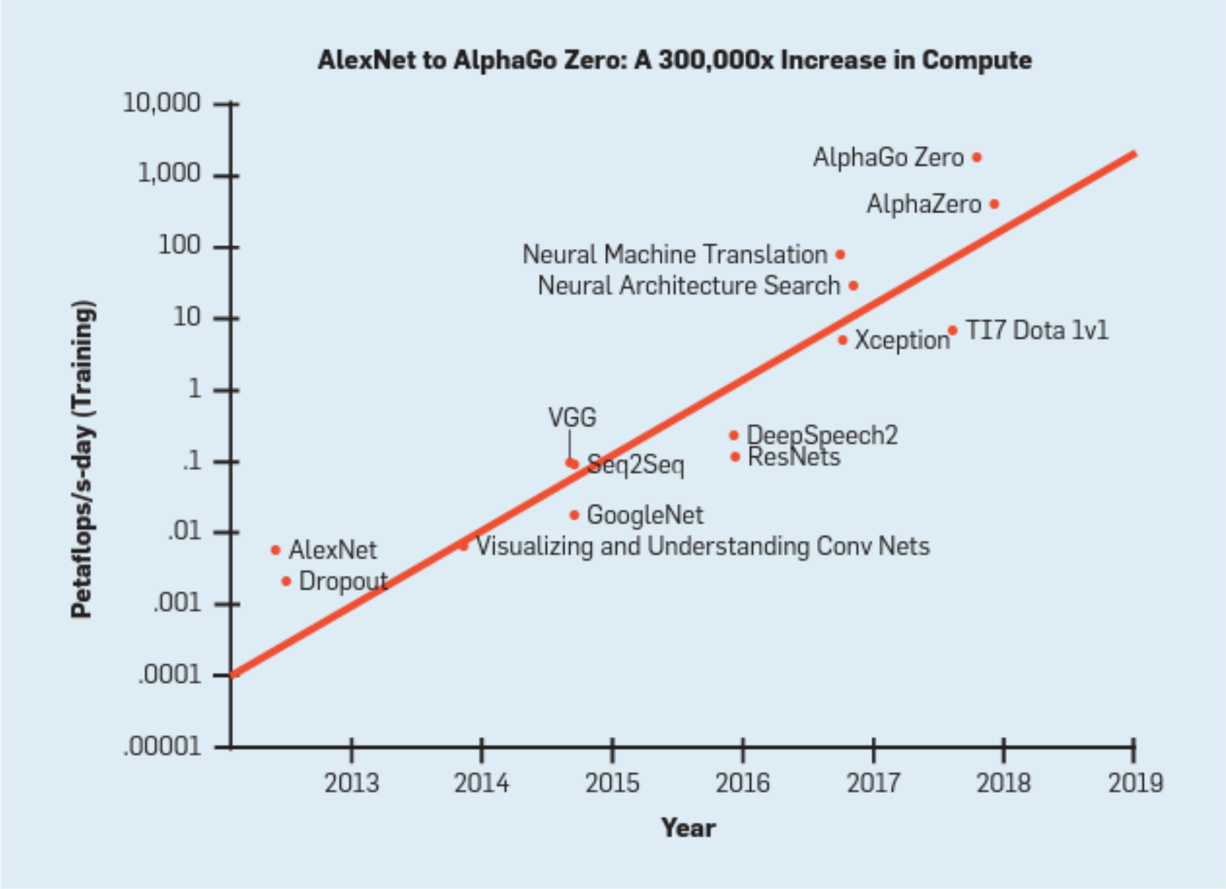
ORIENT: Submodular Mutual Information Measures for Data Subset Selection under Distribution Shift

Athresh Karanam, Krishnateja Killamsetty, Harsha Kokel, Rishabh K Iyer

Deep Learning is Computationally Expensive!

Figure 1. The amount of compute used to train deep learning models has increased 300,000x in six years. Figure taken from Amodel et al.[2]



Shwartz et. al. (2020)

SDA methods even more so!

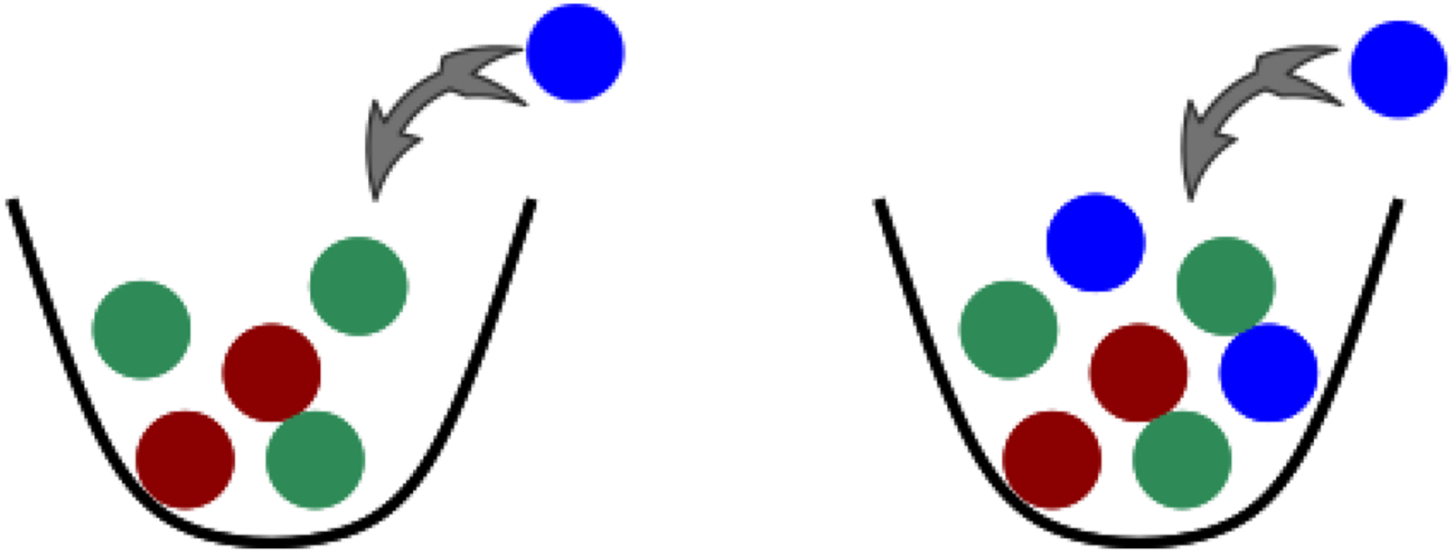
Office-Home dataset

Training ResNet50 on 3k train samples from Office-Home dataset using d-SNE loss takes >18 hours using GTX 1080Ti GPU

Xu et. al. (2019)

Submodular Functions

$$f(A \cup v) - f(A) \geq f(B \cup v) - f(B), \text{ if } A \subseteq B$$



$f = \#$ of distinct colors of balls in the urn.

Submodular Mutual Information

- **Entropy:** $H(X_A) = -\sum_{X_A} P(X_A) \log P(X_A)$**.** Entropy is submodular

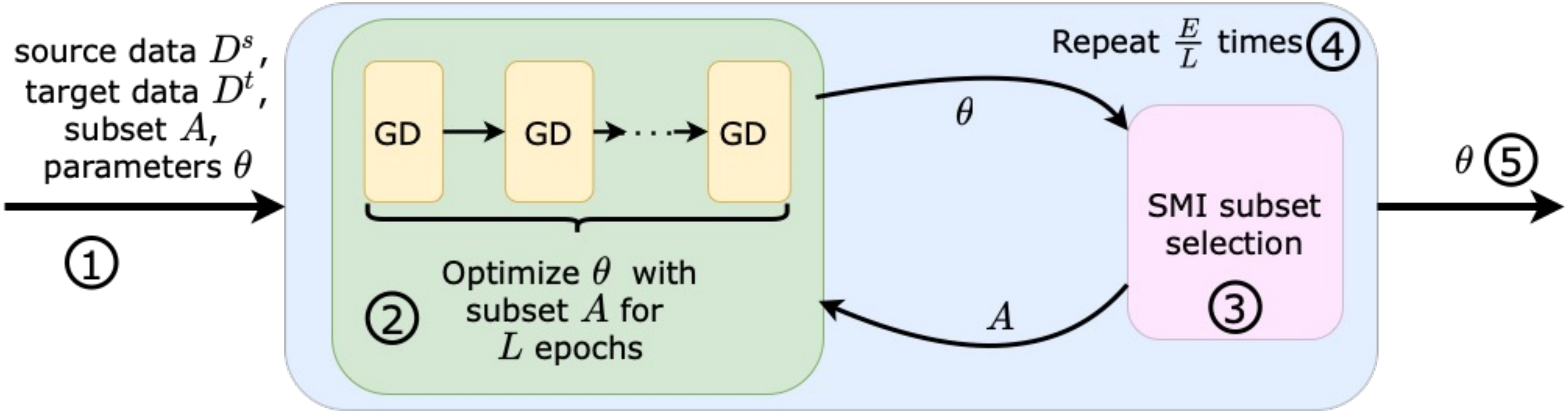
Submodular Mutual Information

- **Entropy:** $H(X_A) = -\sum_{X_A} P(X_A) \log P(X_A)$**.** Entropy is submodular
- **Mutual Information:** $I(X_A; X_B) = H(X_A) + H(X_B) - H(X_{A \cup B})$

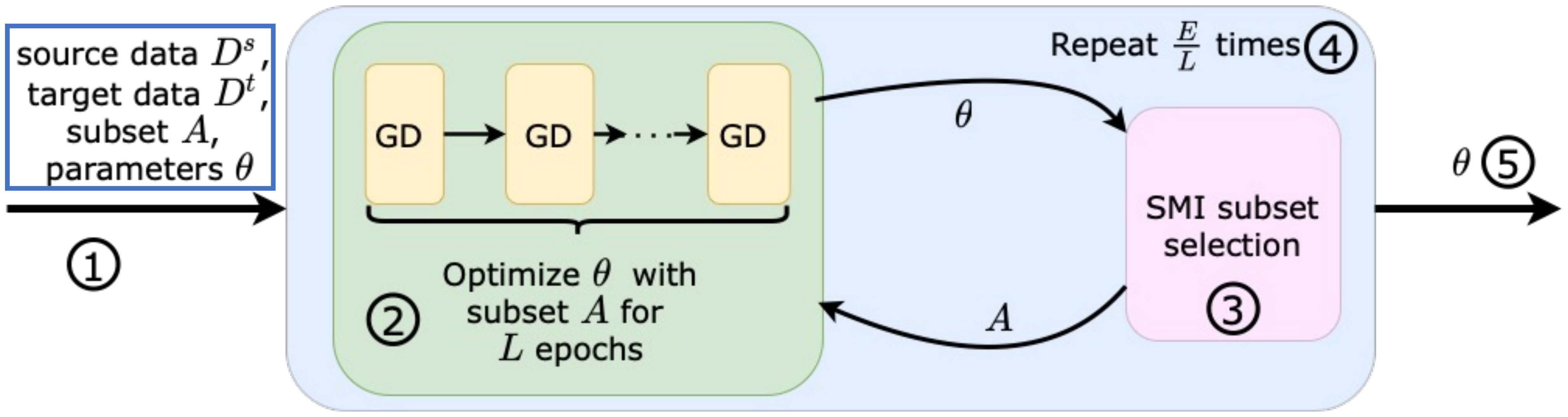
Submodular Mutual Information

- **Entropy:** $H(X_A) = -\sum_{X_A} P(X_A) \log P(X_A)$**.** Entropy is submodular
- **Mutual Information:** $I(X_A; X_B) = H(X_A) + H(X_B) - H(X_{A \cup B})$
- **Submodular Mutual Information:** $I_f(A; D^t) = f(A) + f(D^t) - f(A \cup D^t)$**.** Where the information of a set of points is $f(A)$ and f is submodular

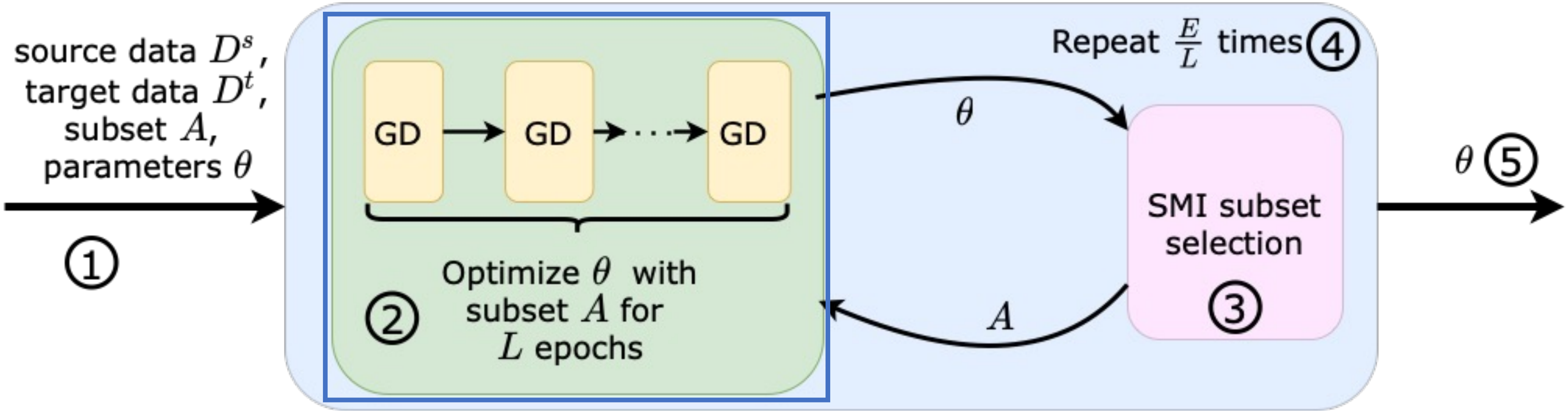
Orient Framework



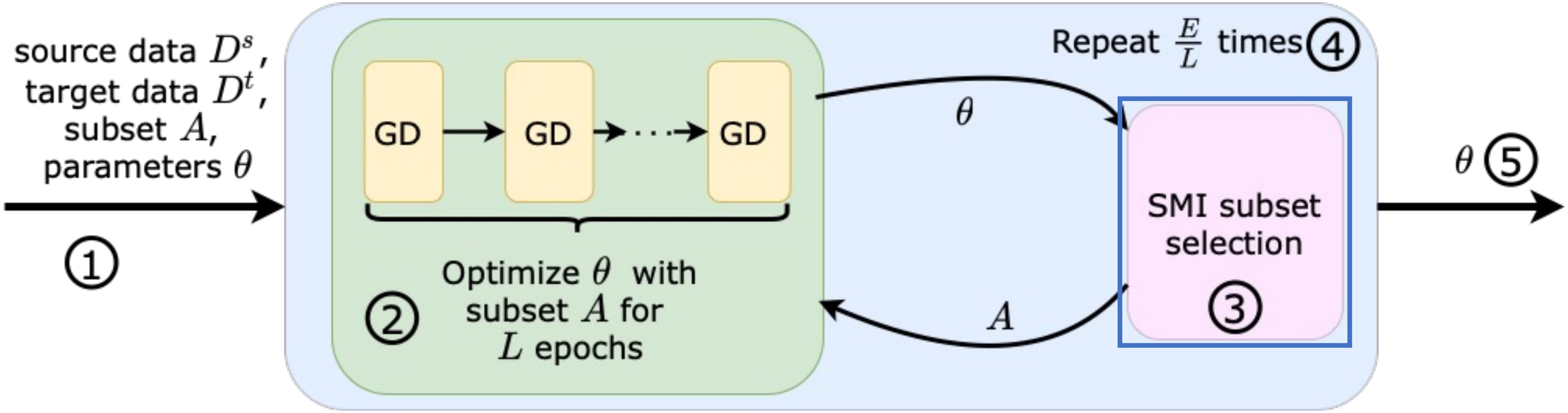
Orient Framework



Orient Framework



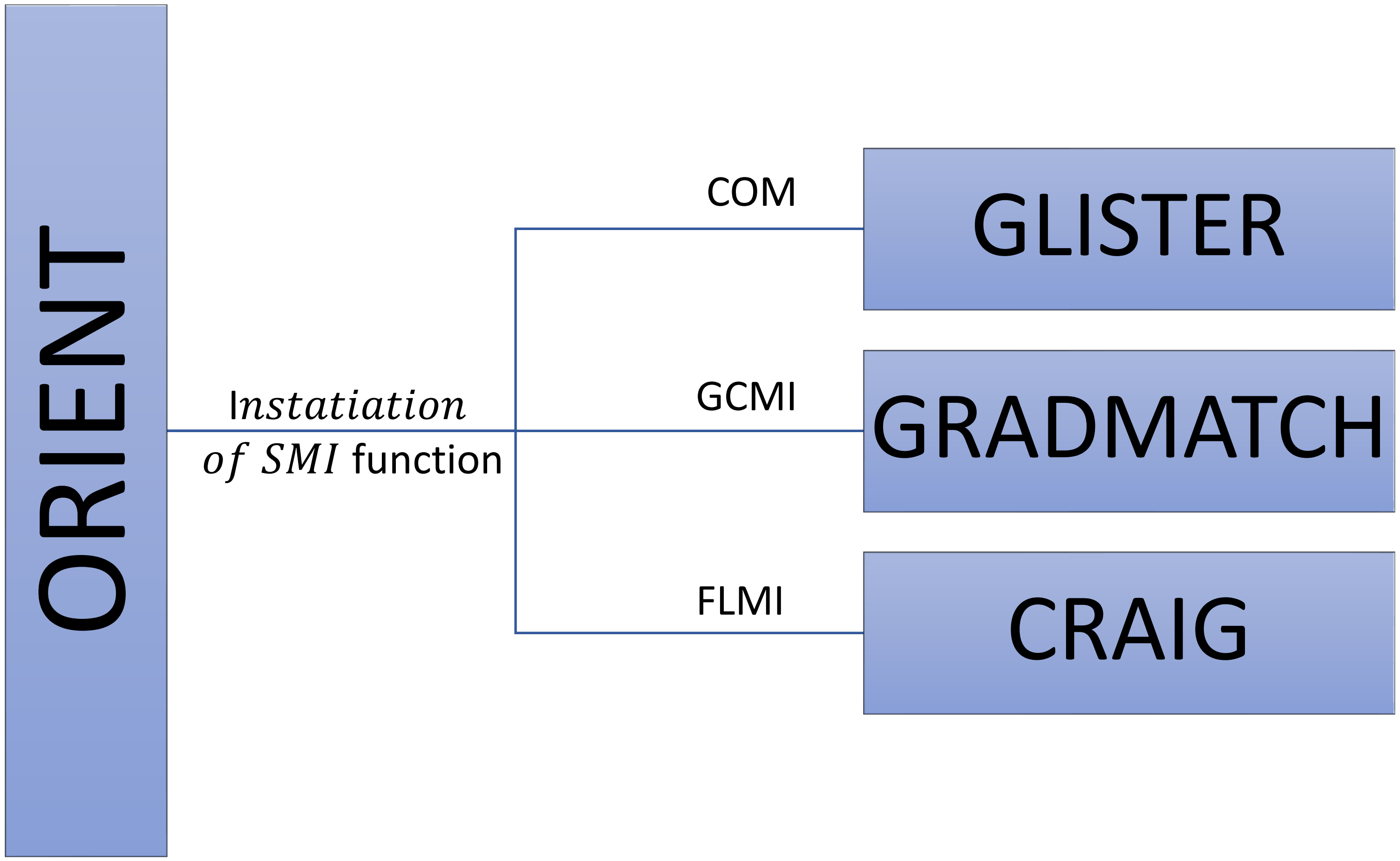
Orient Framework



$$\underset{A \subseteq D^S, |A| \leq b}{\operatorname{argmax}} \; I_f(A; D^t)$$

I_f maximizes the SMI between subset of source data A and target data D^t

Connections to other DSS methods



ORIENT subsumes existing DSS methods based on different instantiations of the SMI function

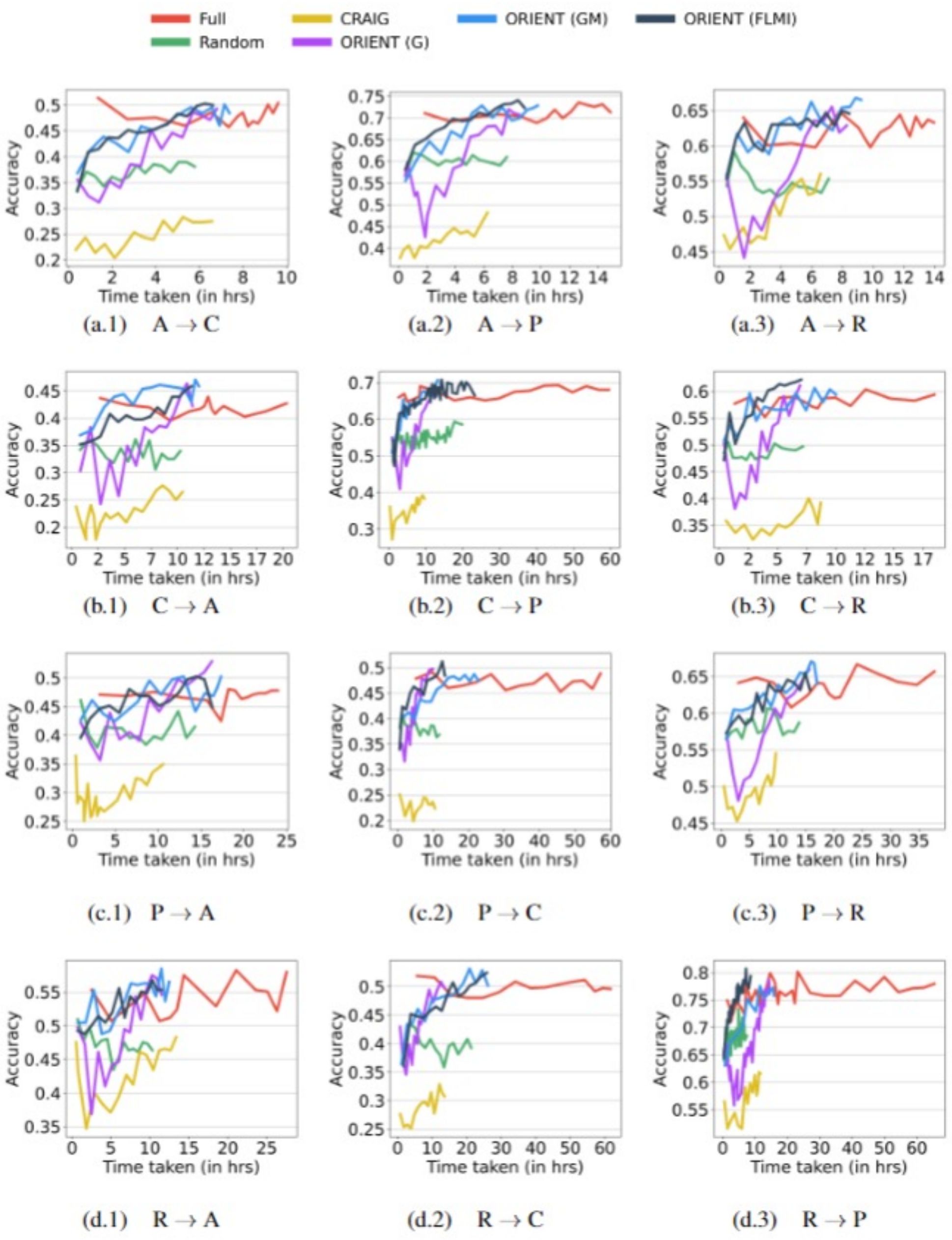
Experiments

Two domain adaptation datasets:

- Office-31 – 3 domains
- Office-Home – 4 domains

Integrating two domain adaptation techniques:

- CCSA
- d-SNE



Experiments

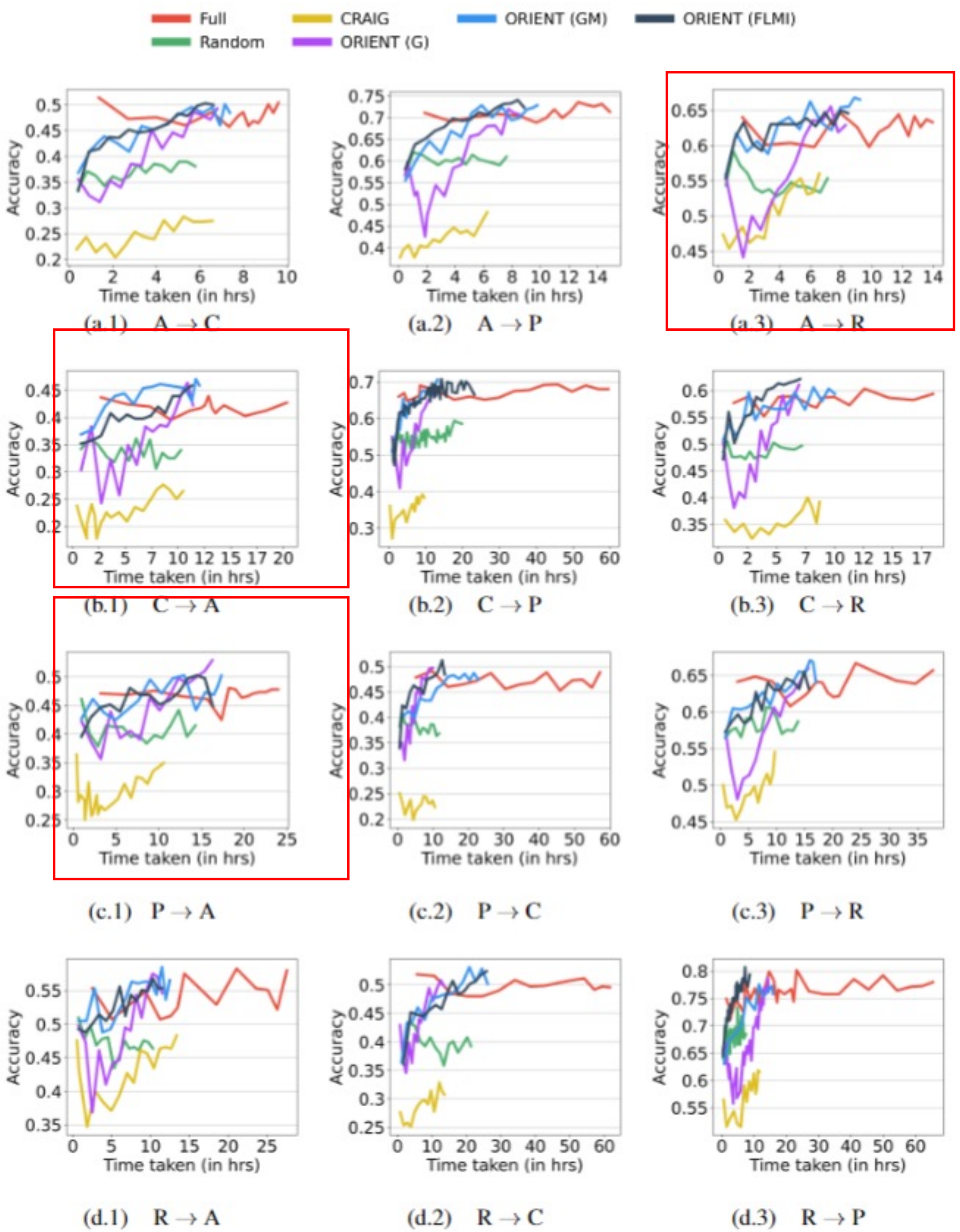
Two domain adaptation datasets:

- Office-31 – 3 domains
- Office-Home – 4 domains

Integrating two domain adaptation techniques:

- CCSA
- d-SNE

Better accuracy in some source-target combinations



Experiments

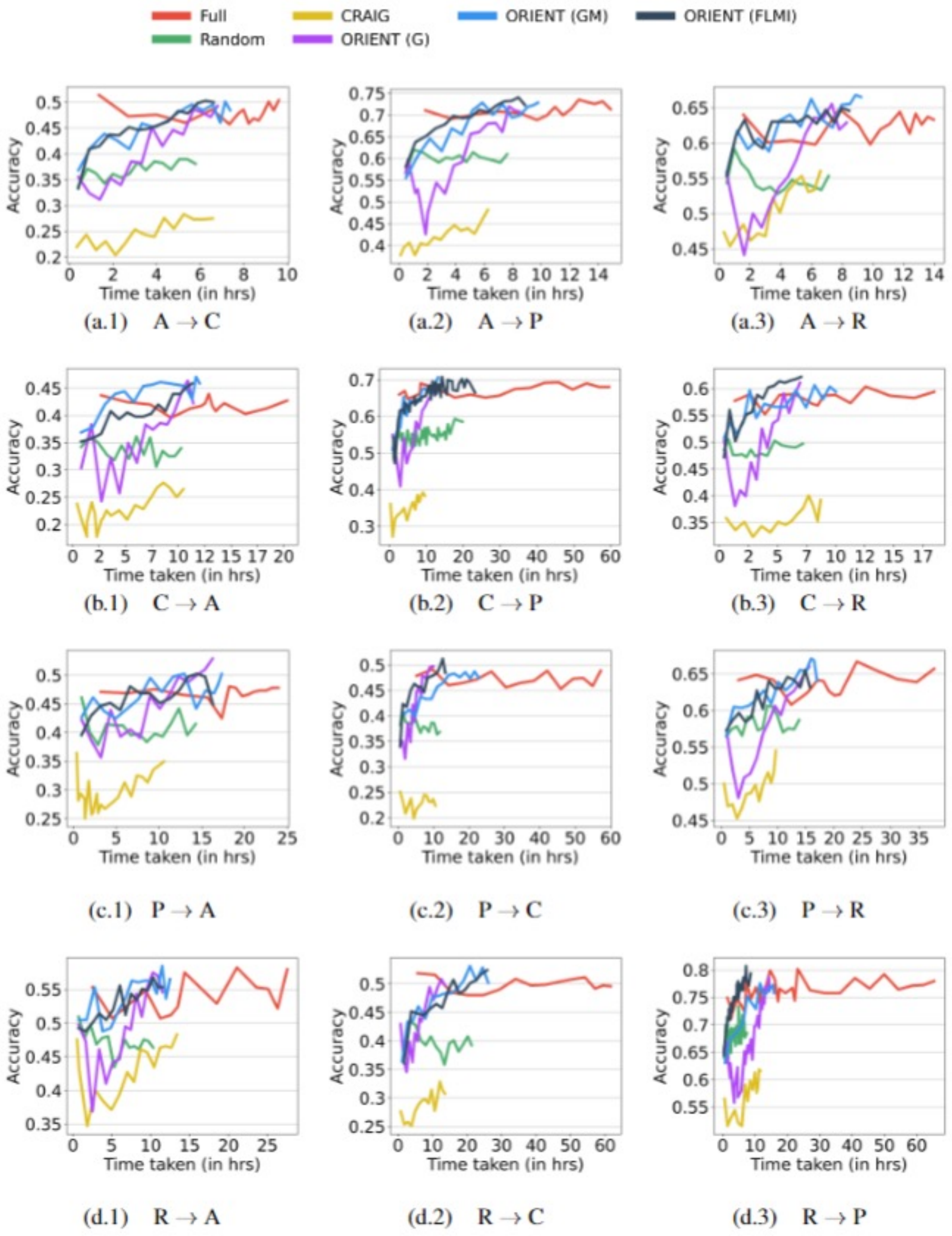
Two domain adaptation datasets:

- Office-31 – 3 domains
- Office-Home – 4 domains

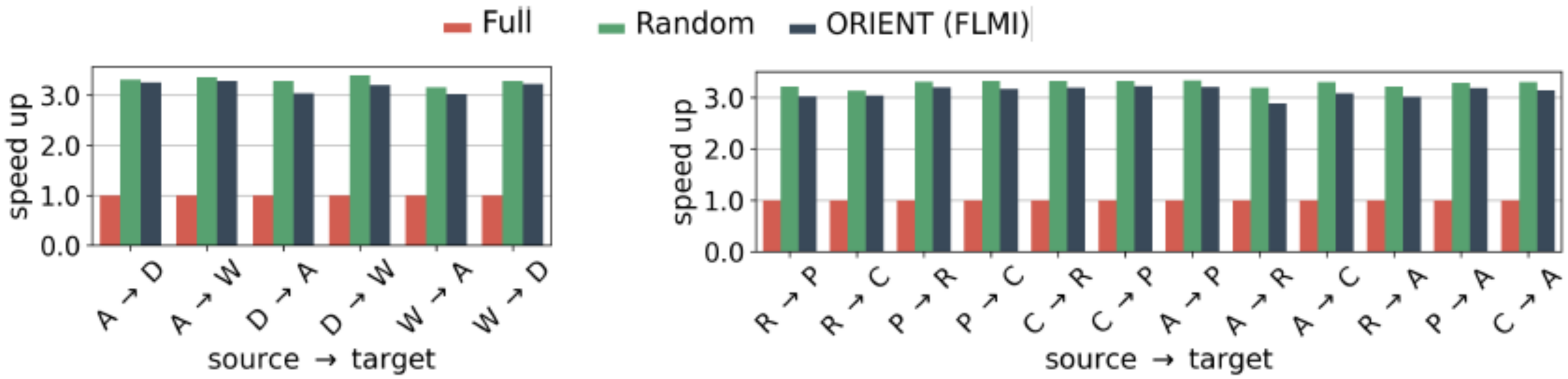
Integrating two domain adaptation techniques:

- CCSA
- d-SNE

2x-3x speed-ups over all source-target combinations



Experiments



(a) Office-31 dataset

(b) Office-Home dataset

Figure 4: Speed up achieved by combining d-SNE with ORIENT

ORIENT integrates with existing SDA methods to achieve ~3x speed-ups

Conclusion

- We propose ORIENT – data subset selection method based on SMI measures to improve computational efficiency of SDA techniques
- ORIENT subsumes existing data subset selection methods for different instantiations of the SMI function
- Experiments validate ORIENT's effectiveness in improving computational efficiency when integrated with state-of-the-art SDA techniques