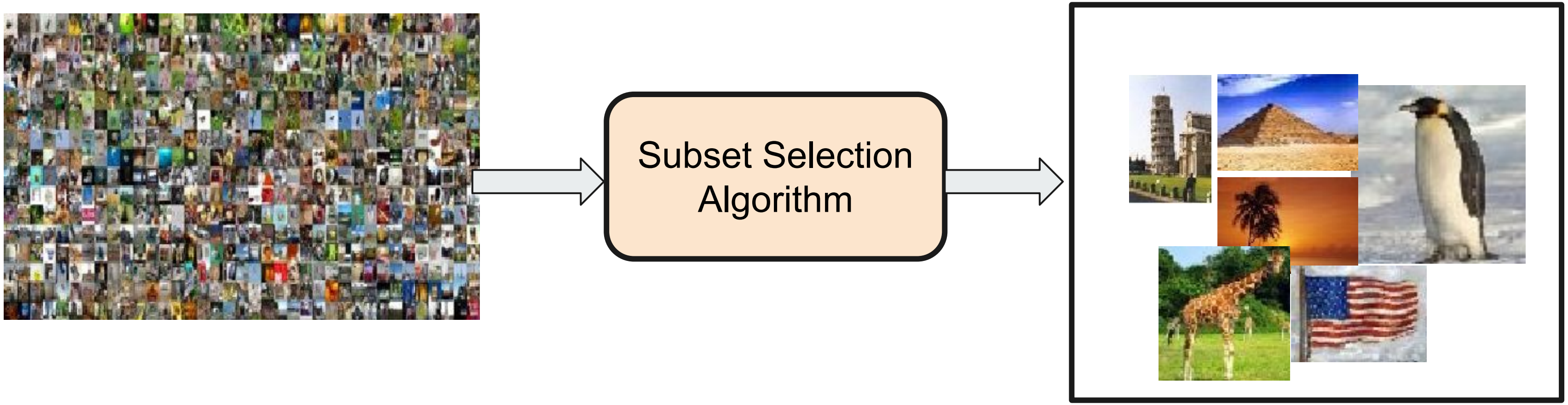
ECML PKDD 2021

Finding High-Value Training Data Subset through Differentiable Convex Programming

Soumi Das [†], Arshdeep Singh [†], Saptarshi Chatterjee [†], Suparna Bhattacharya*, Sourangshu Bhattacharya [†]

[†] Indian Institute of Technology, Kharagpur
* Hewlett Packard Labs, Hewlett Packard Enterprise

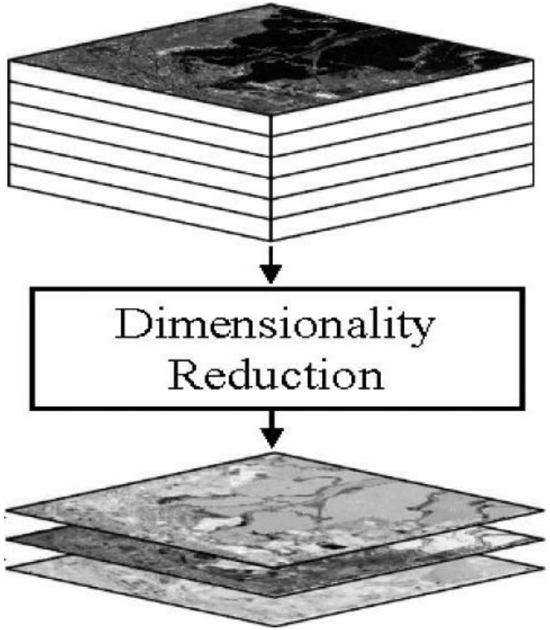
Subset Selection

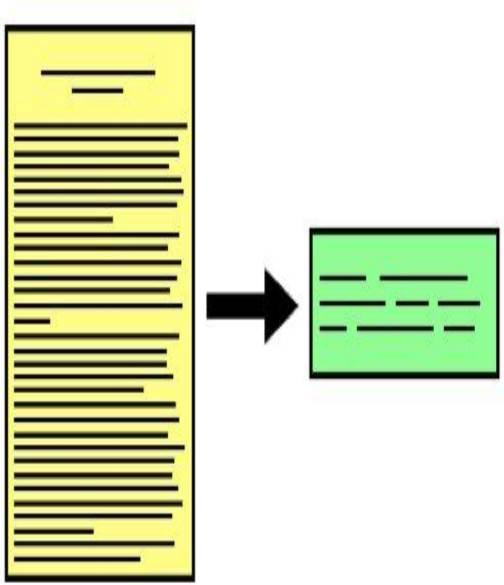


Research Question: How to select the most informative items?

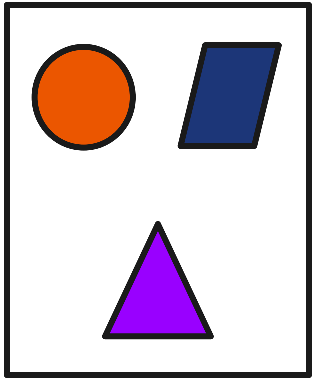
NP-Hard Problem

Context of Informativeness





Machine Learning Setup

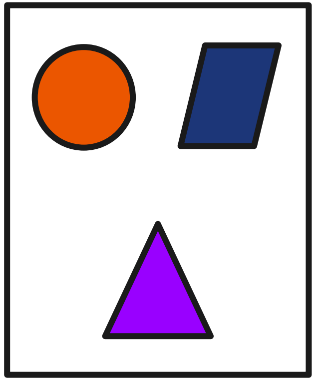


Training Data

Machine Learning Model

Test Data performance

Machine Learning Setup



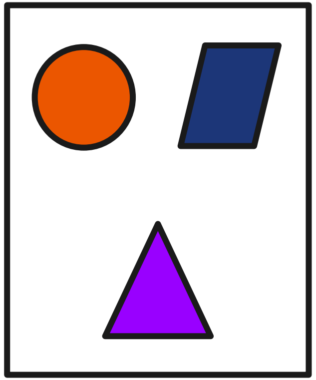
Training Data

Machine Learning Model

Test Data performance

How much does each data point contribute towards the test set performance?

Machine Learning Setup



Training Data

Machine Learning Model

Test Data performance

Applications: Explainability ; Debugging domain mismatch ; Fixing mislabelled examples

Related Works

1. Influence Functions (IF)
2. Data Shapley (DS)
3. TracIn using Checkpointing (TracIn-CP)
4. Data Valuation using Reinforcement Learning (DVRL)

1. Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions." *ICML*. PMLR, 2017.
2. Ghorbani, Amirata, and James Zou. "Data shapley: Equitable valuation of data for machine learning." *ICML*. PMLR, 2019.
3. Pruthi, Garima, et al. "Estimating training data influence by tracing gradient descent." *NeurIPS* (2020).
4. Yoon, Jinsung, Sercan Arik, and Tomas Pfister. "Data valuation using reinforcement learning." *ICML*. PMLR, 2020.

Proposed Method:

HOST-CP = **H**igh value **O**nline **S**ubset selection of **T**raining samples through differentiable **C**onvex **P**rogramming

High-Level Idea

1. Learnable subset selection formulation.
2. Jointly learn the selection parameters along with the model parameters for optimizing the value function.

Problem Statement

$$\max_{s \in \mathcal{S}} v(s) \quad \text{sub. to } |s| \leq \gamma n$$

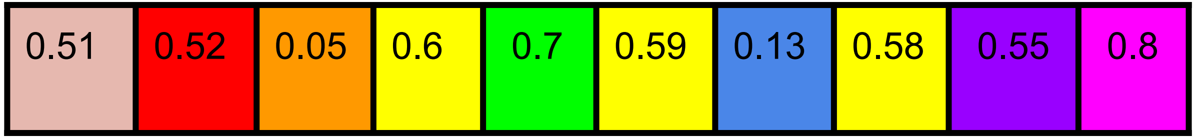
Value function

Fraction of incoming instances

$$v(s) = -\mathcal{L}(f(\theta^*(s)), \mathcal{D}^t)$$

Other value functions exist like test set accuracy, or expected return in the context of Reinforcement Learning

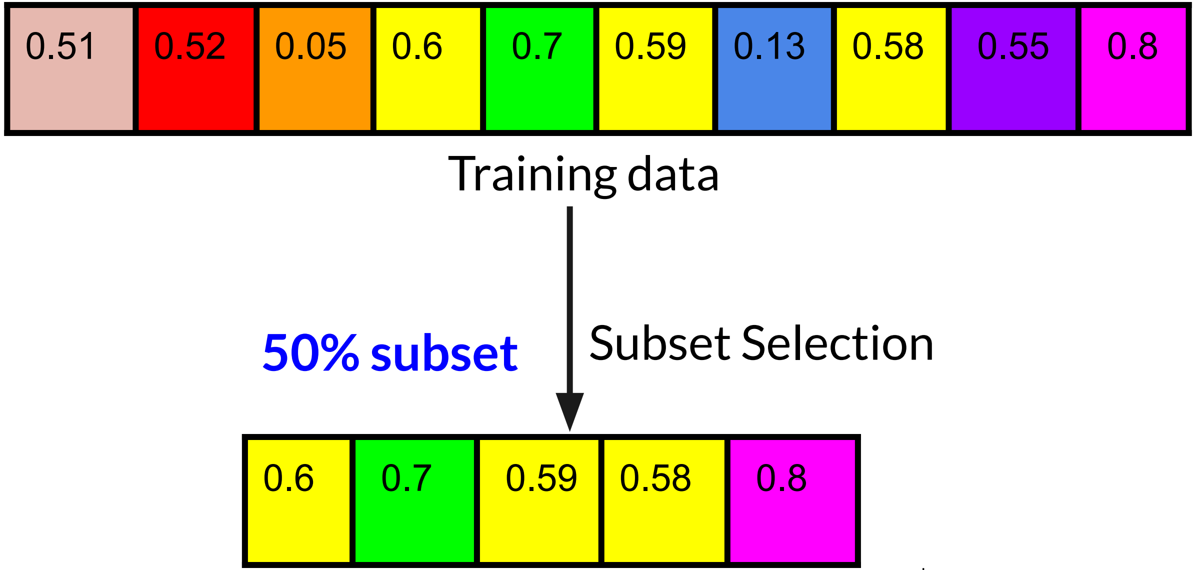
Related works: Data Valuation



Training data

* Different colors indicate different classes

Related works: Data Valuation → Subset Selection



Similarity between instances is not considered - can be a driving factor in subset selection

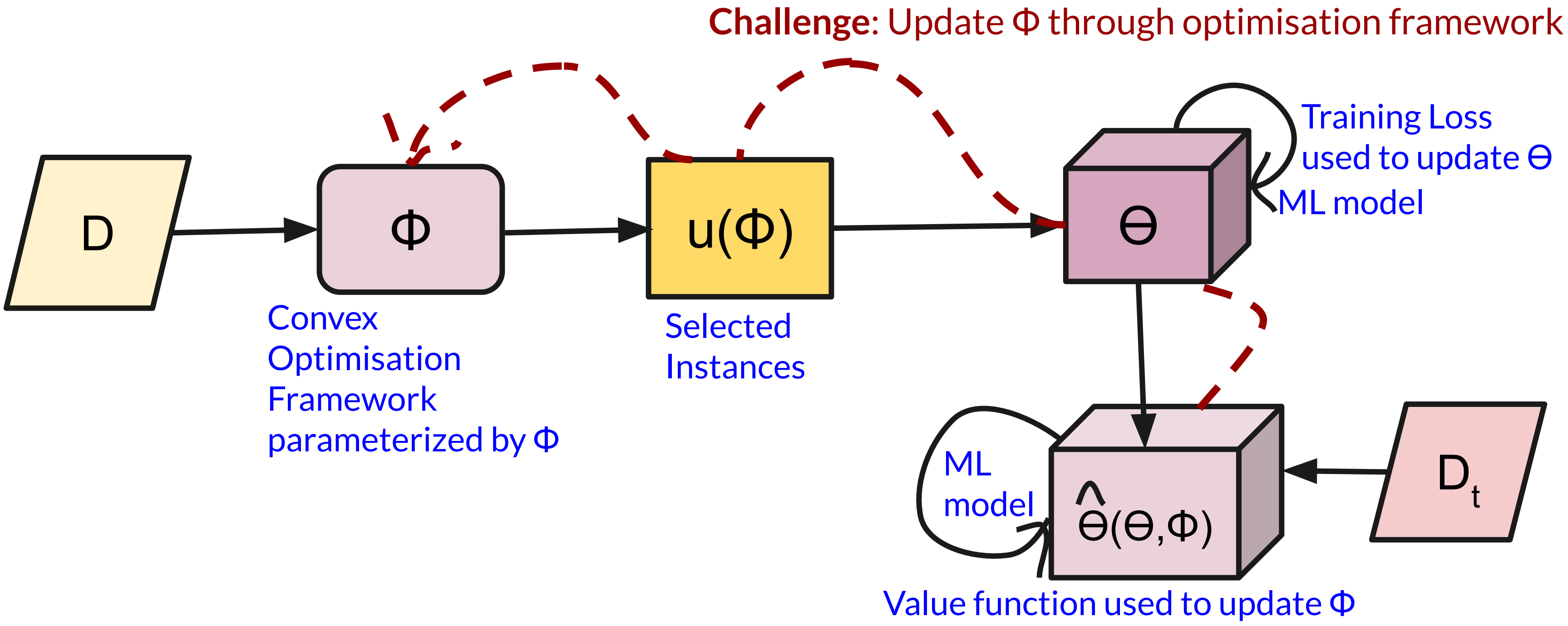
* Different colors indicate different classes

Examples from real data

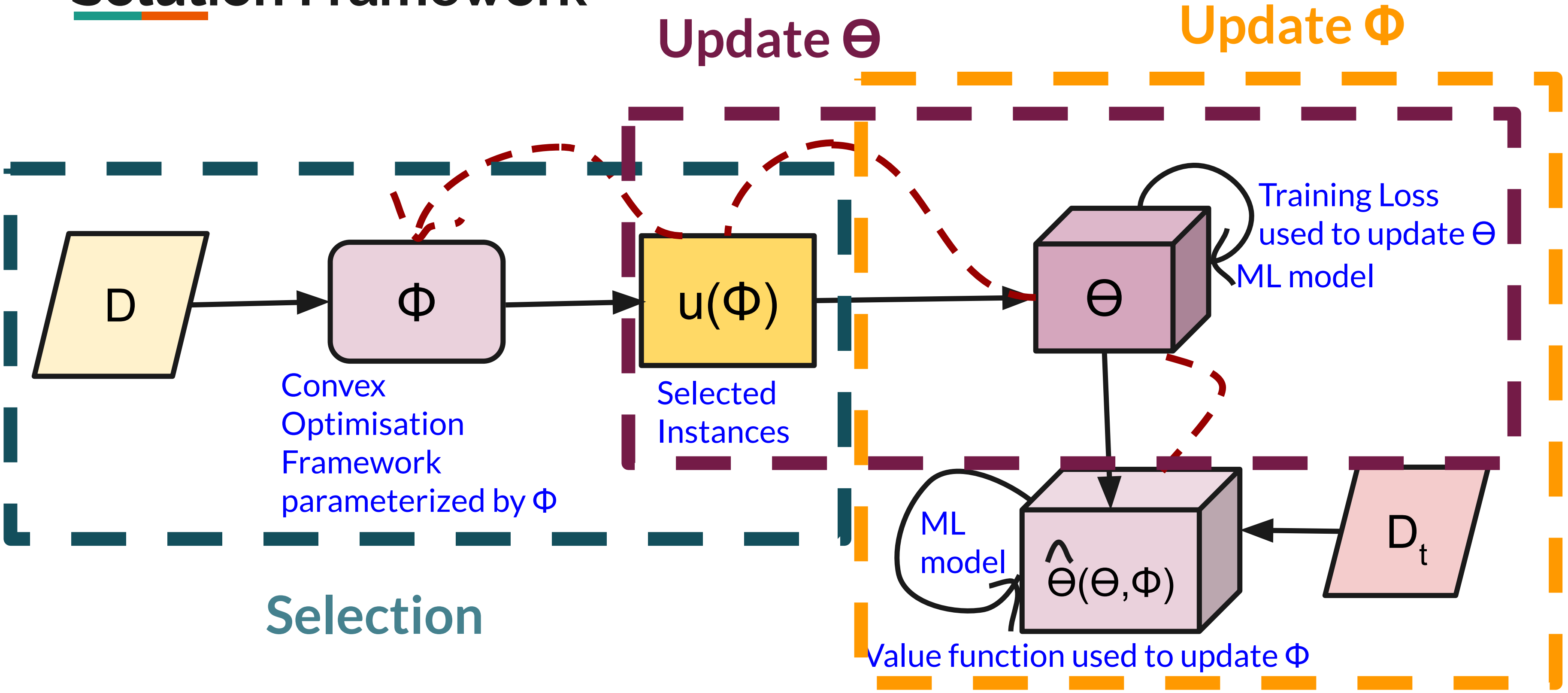
50,000 images of 10 classes

Top few images were from only the classes - horse and bird

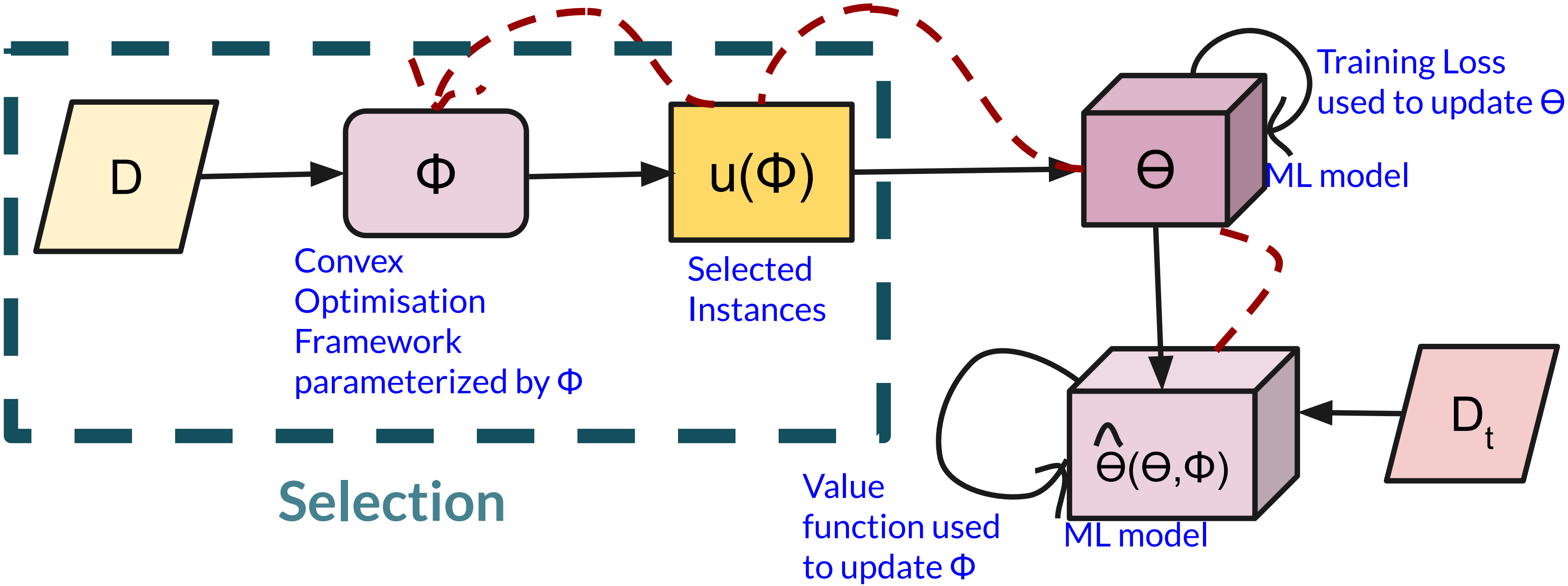
Solution Framework



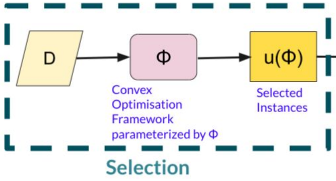
Solution Framework



Selection



D
Φ
Convex Optimisation Framework parameterized by Φ
u(Φ)
Selected Instances
Θ
Training Loss used to update Θ
ML model
Selection
Value function used to update Φ
$\hat{\Theta}$(Θ,Φ)
ML model
D_t

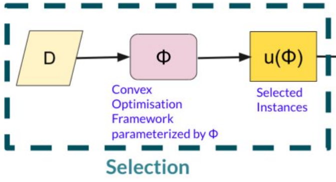


Selection Objective Function

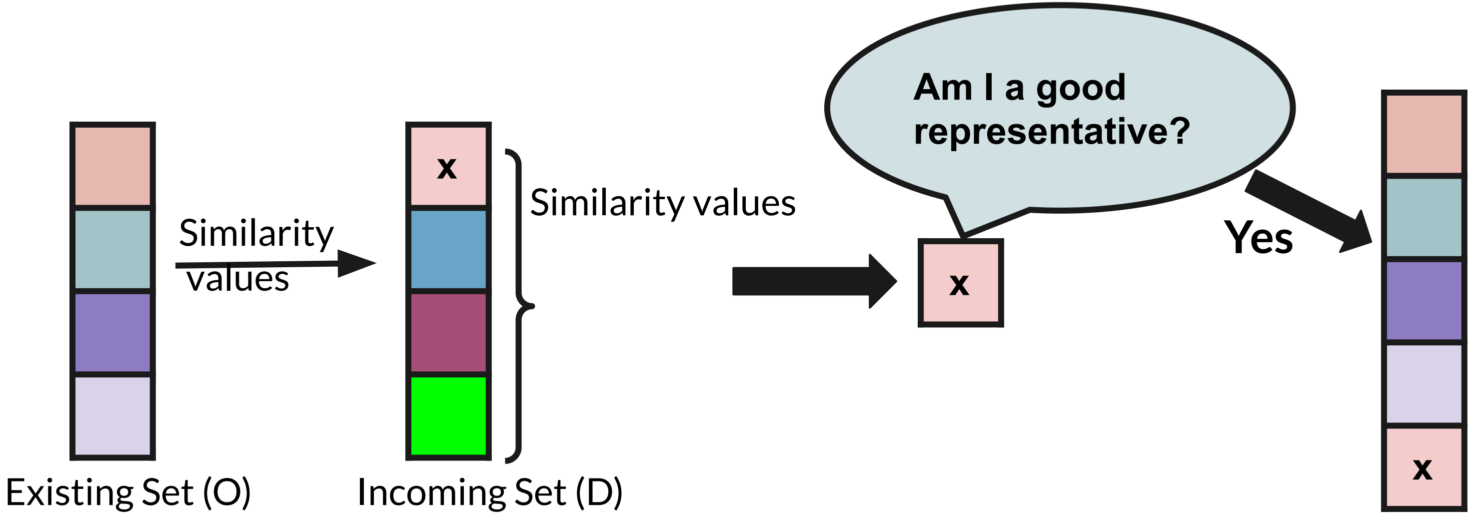
$$s^* = \min_{s \in S} \sum_{(x,y)\in D} \min_{(x',y')\in s} d(x, x')$$

This can be mapped to Facility Location problem and relaxed to convex linear programming problem

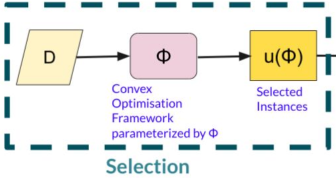
Instance Selection



We exploit the fact that the selection of a training point can depend on other training points also.



Convex Optimisation Selection



Affine function

$$\boxed{D(h(x_i,\phi),h(x_j,\phi))}$$

Distance / Dissimilarity

Incoming set

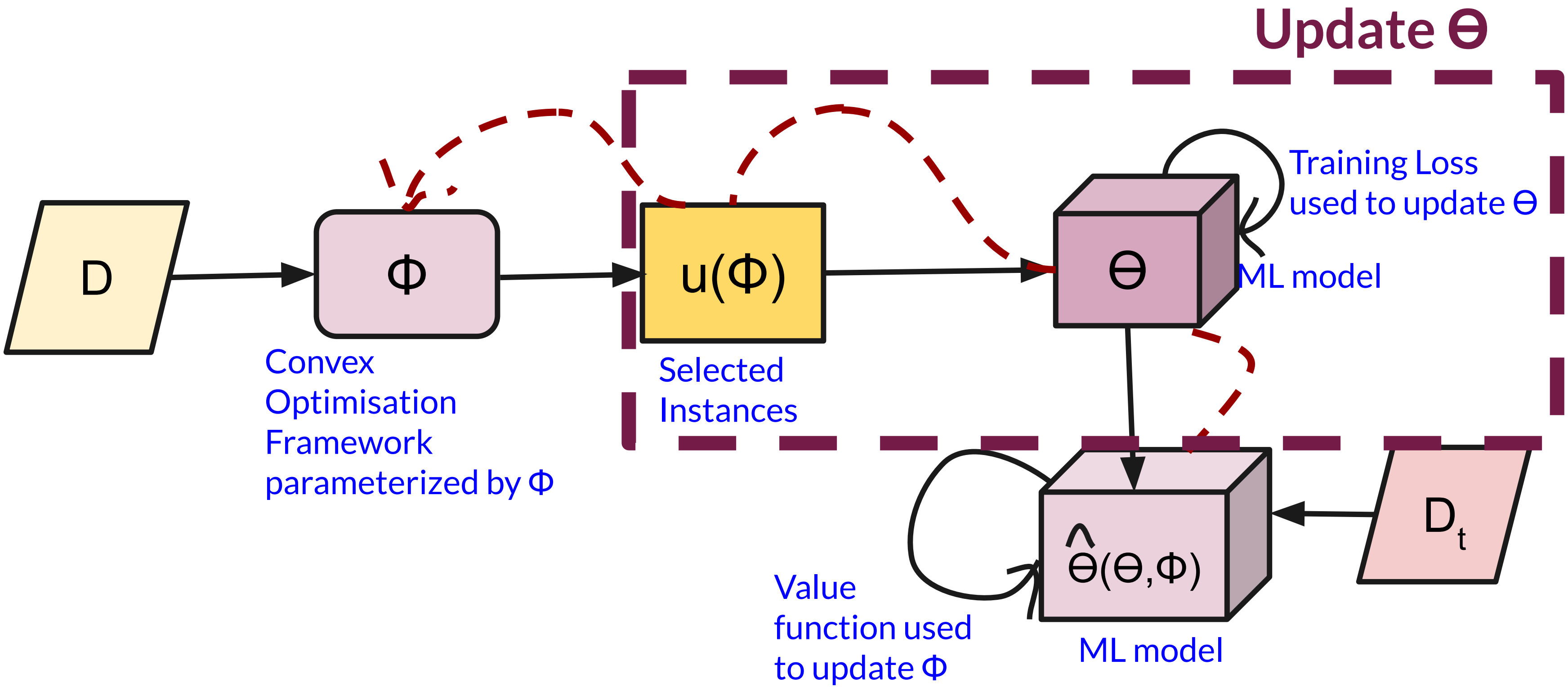
Existing set

$$\min_{z_{ij}^o, z_{ij}^n \in [0,1]} \sum_{x_i \in D_{i(t)}, x_j \in \mathcal{O}(t)} z_{ij}^o d(x_i, x_j) + \sum_{x_i \in D_{i(t)}, x_j \in D_{i(t)}} z_{ij}^n d(x_i, x_j)$$

$$\text{sub. to} \sum_{x_j \in \mathcal{O}(t)} z_{ij}^o + \sum_{x_j \in D_{i(t)}} z_{ij}^n = 1, \ \forall i = \{1, \ldots, n\}$$

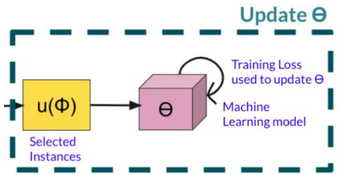
$$\sum_{x_j \in D_{i(t)}} u_j \leq \gamma |D_{i(t)}|$$

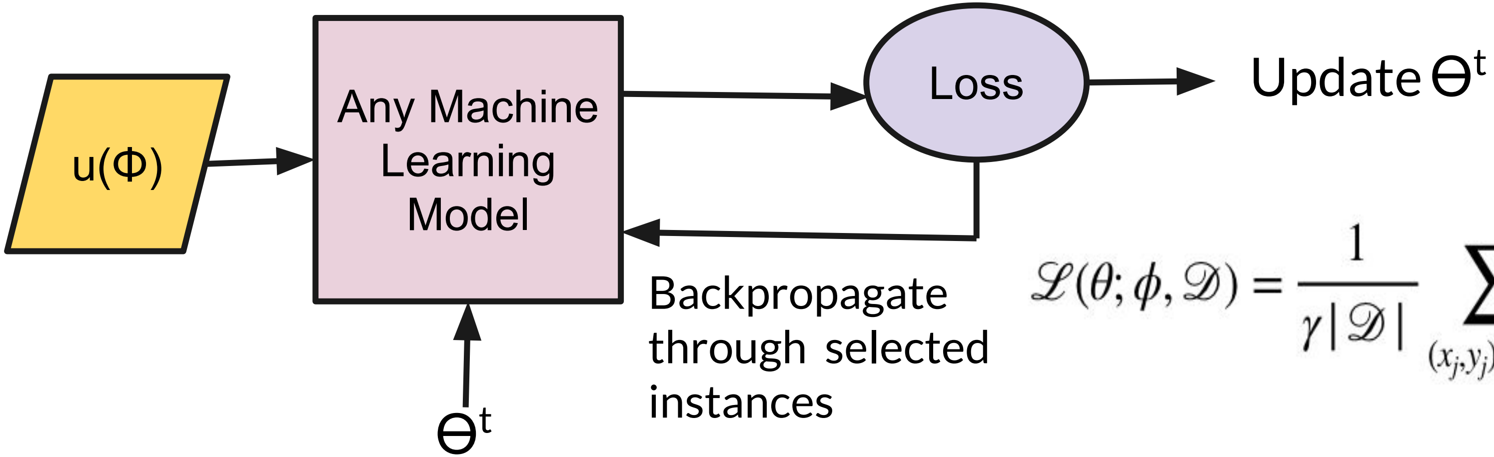
Update ϴ



Update ϴ
D
Φ
Convex
Optimisation
Framework
parameterized by Φ
u(Φ)
Selected
Instances
ϴ
Training Loss
used to update ϴ
ML model
θ̂(ϴ,Φ)
Value
function used
to update Φ
ML model
D
t

Update model parameter Θ



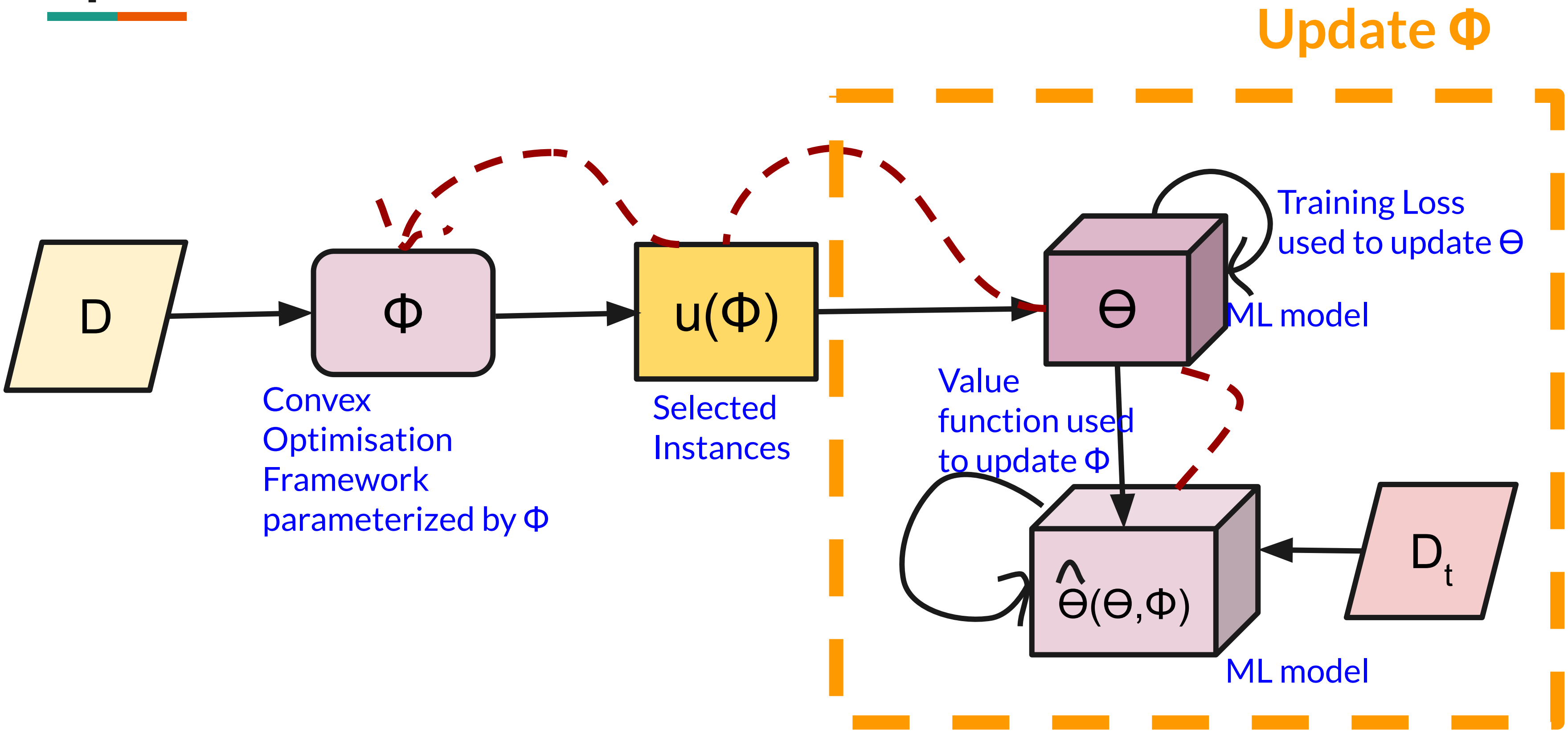


Representatives

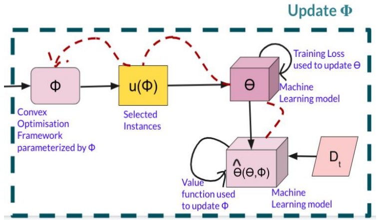
$$\mathscr{L}(\theta; \phi, \mathscr{D}) = \frac{1}{\gamma|\mathscr{D}|} \sum_{(x_j, y_j) \in \mathscr{D}} u_j(\phi) L(y_j, f(x_j; \theta))$$

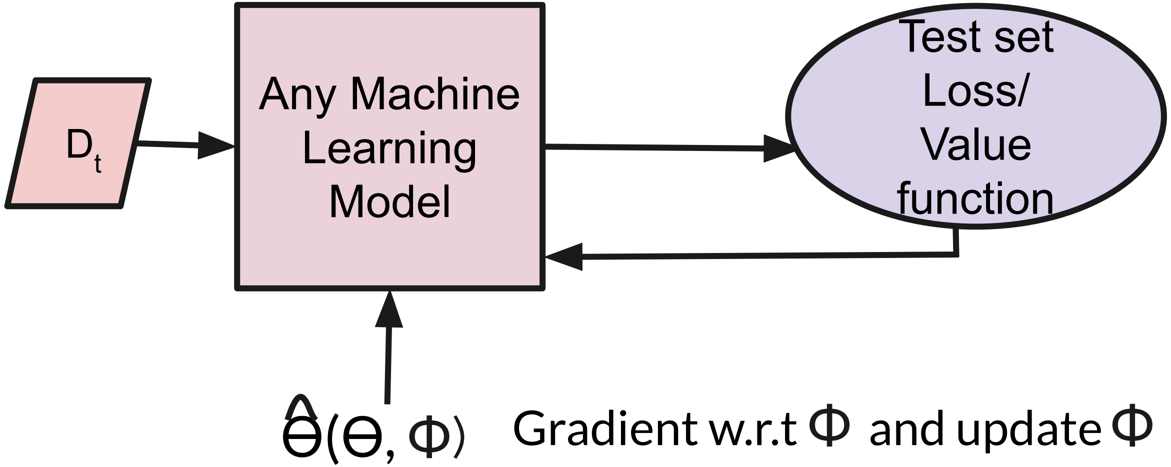
$$\theta^{t+1} = \theta^t - \alpha \frac{1}{k} \nabla_\theta \mathscr{L}(\theta; \phi^t, D_{i(t)})$$

Update Φ


Update Φ
D
Φ
u(Φ)
Θ
Convex Optimisation Framework parameterized by Φ
Selected Instances
Training Loss used to update Θ
ML model
Value function used to update Φ
$\hat{\Theta}(\Theta,\Phi)$
D_t
ML model

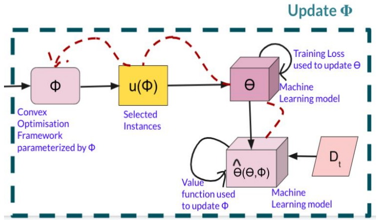
Update selection parameter Φ

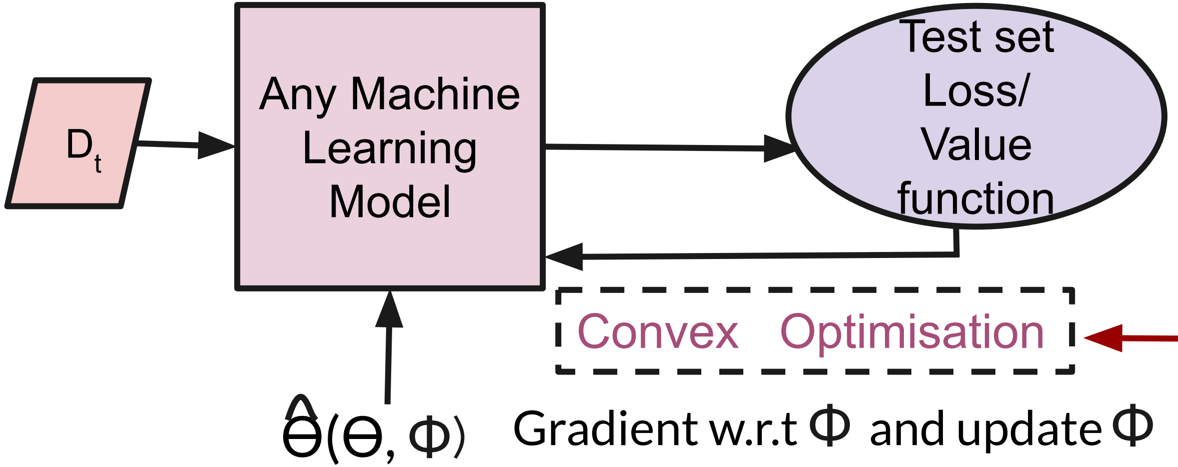




$$\phi^{t+1} = \phi^{t} - \beta \frac{1}{k} \nabla_{\phi} \mathscr{V}(\phi; \theta^{t}, D_{i(t)}, \mathscr{D}^{t})$$

Update selection parameter Φ

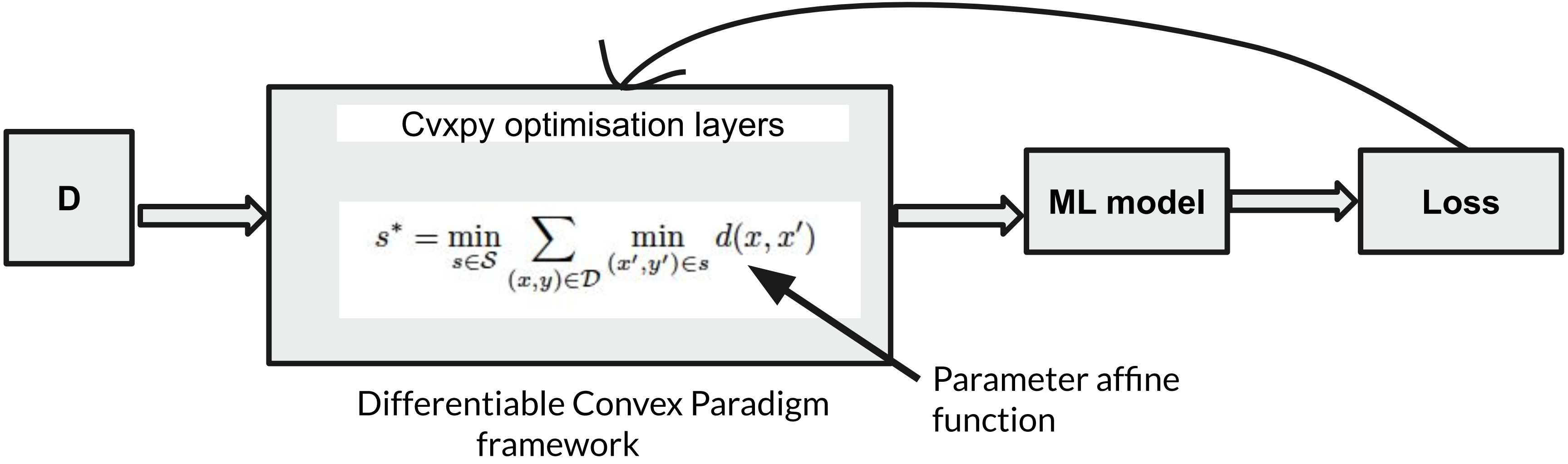




Differentiable[1]
Convex Paradigm

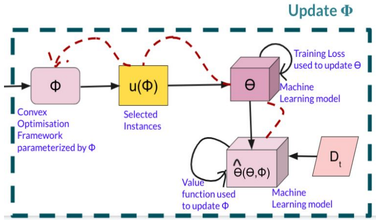
$$\phi^{t+1} = \phi^t - \beta \frac{1}{k} \nabla_\phi \mathscr{V}(\phi; \theta^t, D_{i(t)}, \mathscr{D}^t)$$

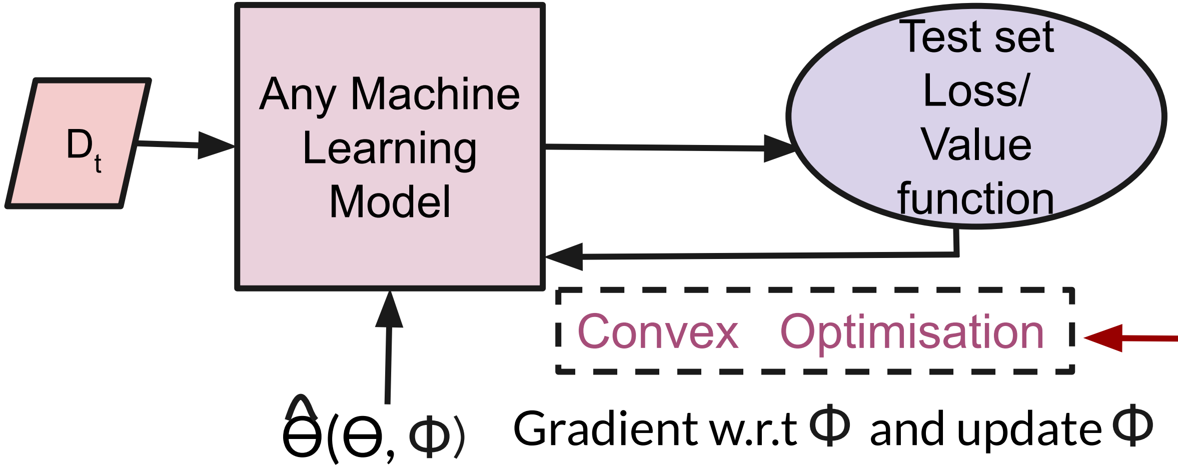
Differentiable Convex Paradigm (DCP)



[1] Agrawal, A., Amos, B., Barratt, S., Boyd, S., Diamond, S., Kolter, J.Z.: Differentiable convex optimization layers. In: NeurIPS (2019)

Update selection parameter Φ





$$\phi^{t+1} = \phi^t - \beta \frac{1}{k} \nabla_\phi \mathscr{V}(\phi; \theta^t, D_{i(t)}, \mathscr{D}^t)$$

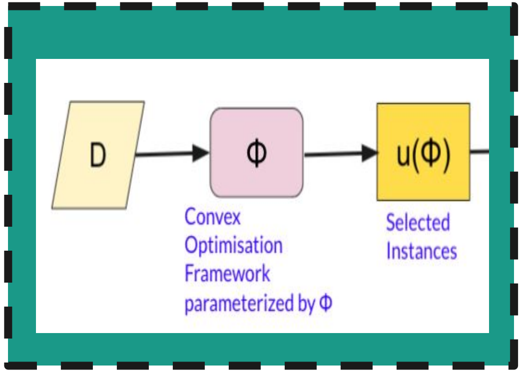
Differentiable[1]
Convex Paradigm

$$\mathscr{V}(\phi;\theta',\mathscr{D},\mathscr{D}^t) = v(\hat{\theta},\mathscr{D}^t), \text{ where } \hat{\theta} = \theta' - \alpha \nabla_\theta \mathscr{L}(\theta;\phi,\mathscr{D})$$

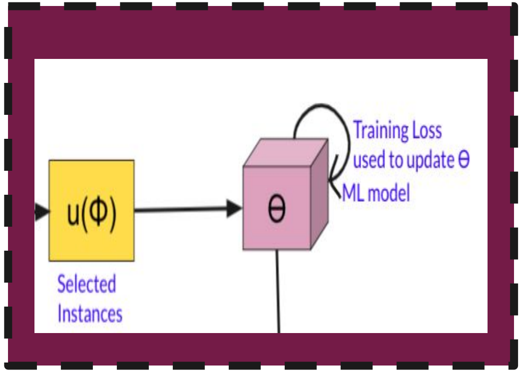
$$\mathscr{L}(\theta;\phi,\mathscr{D}) = \frac{1}{\gamma|\mathscr{D}|} \sum_{(x_j,y_j)\in\mathscr{D}} u_j(\phi) L(y_j, f(x_j;\theta))$$

Joining the threads

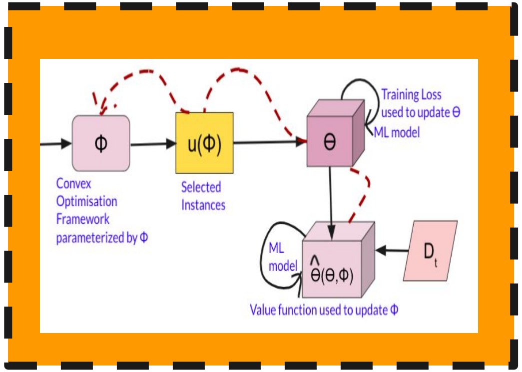
Selection



Update ϴ



Update Φ



Empirical Evaluation

Datasets

1. Image domain - CIFAR10
2. Biomedical domain - Protein
3. Text domain - 20 Newsgroups
4. Synthetic dataset

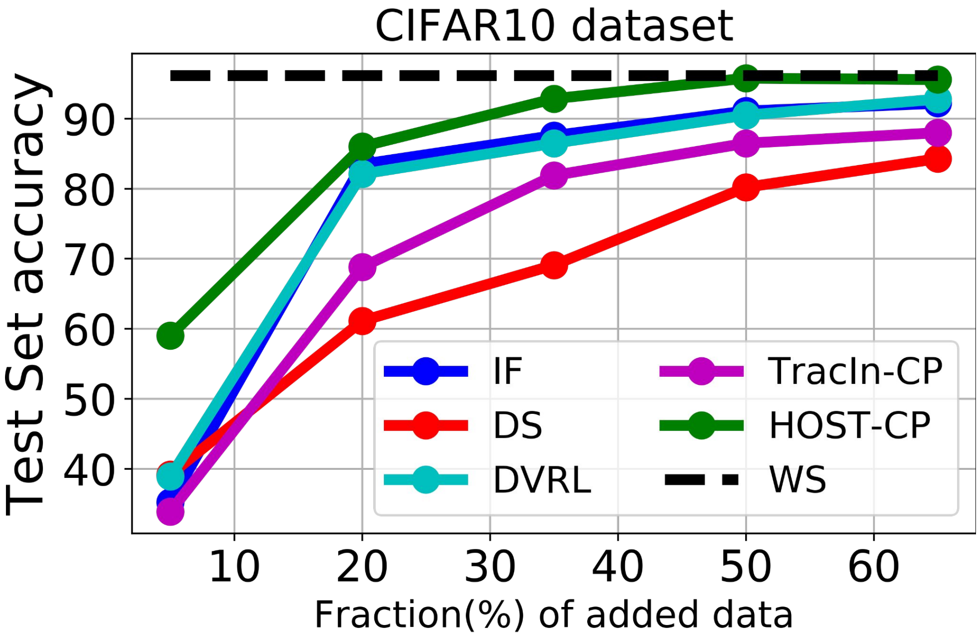
This is the first line of this text example.

This is the second line of the same text.

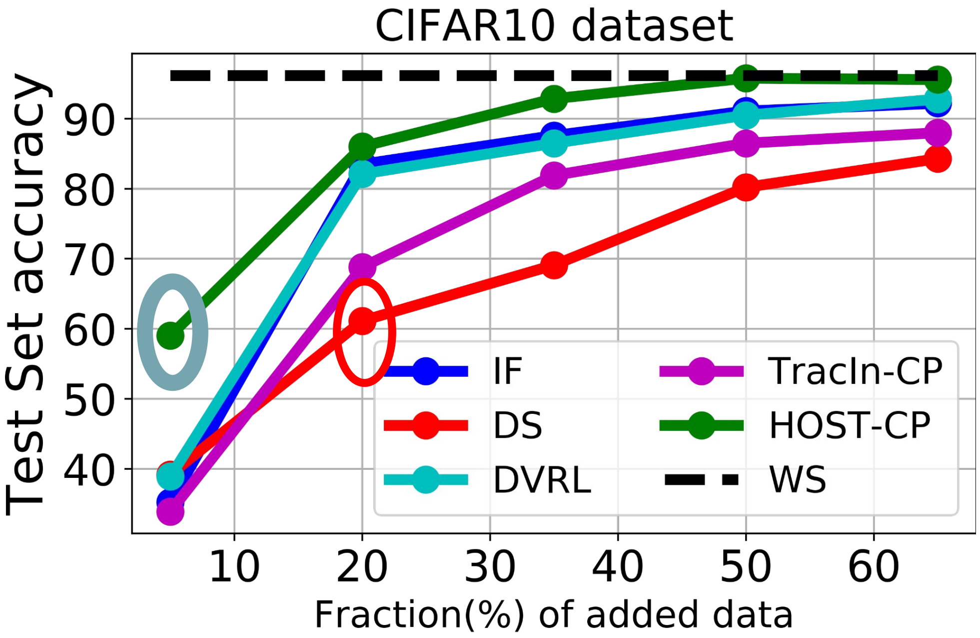
Applications

1. Explainability : Which set of examples best explains the test set predictions?
2. Diagnosing mislabelled examples : How early can the mislabelled examples be detected?

Experimental Results : Addition



Experimental Results : Addition



Data Shapley(DS) reaches the same accuracy level at a fraction of 20% that HOST-CP attained at a fraction of 5%

Denotes presence of redundant elements in selection using baselines.

Intuitive examples

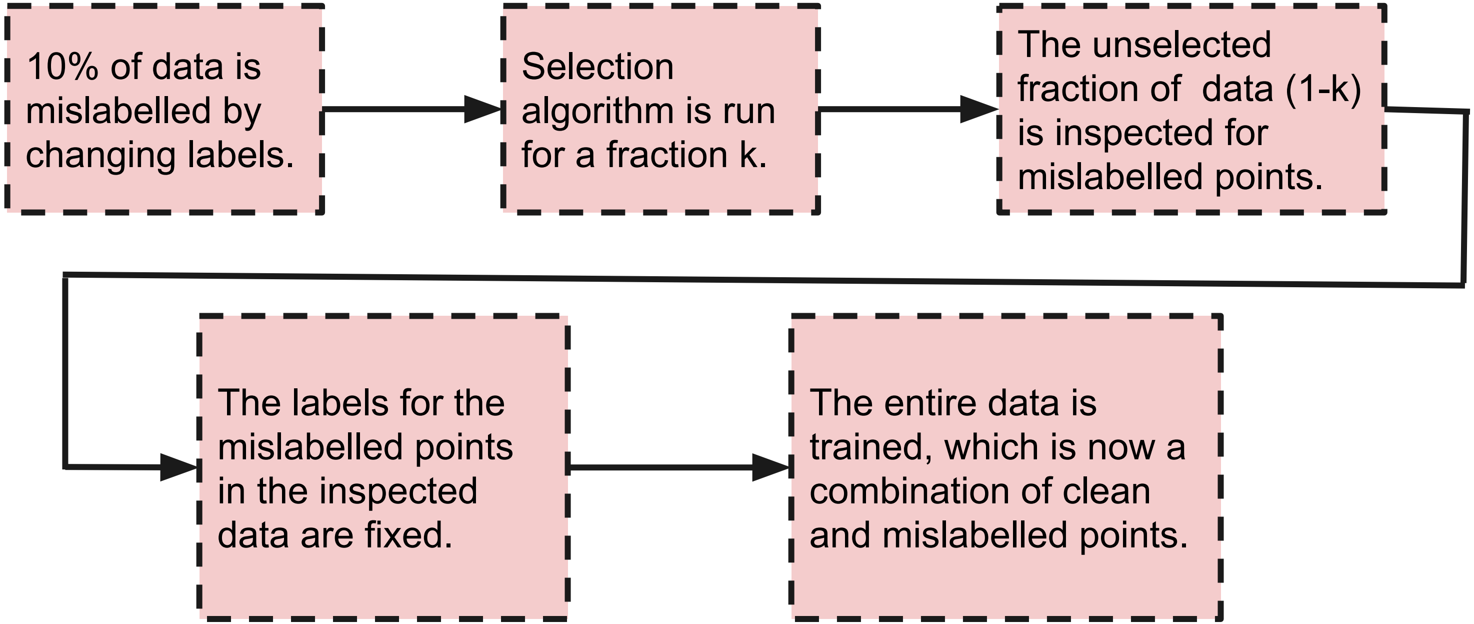
Top few images were from only the classes - horse and bird

Baseline method

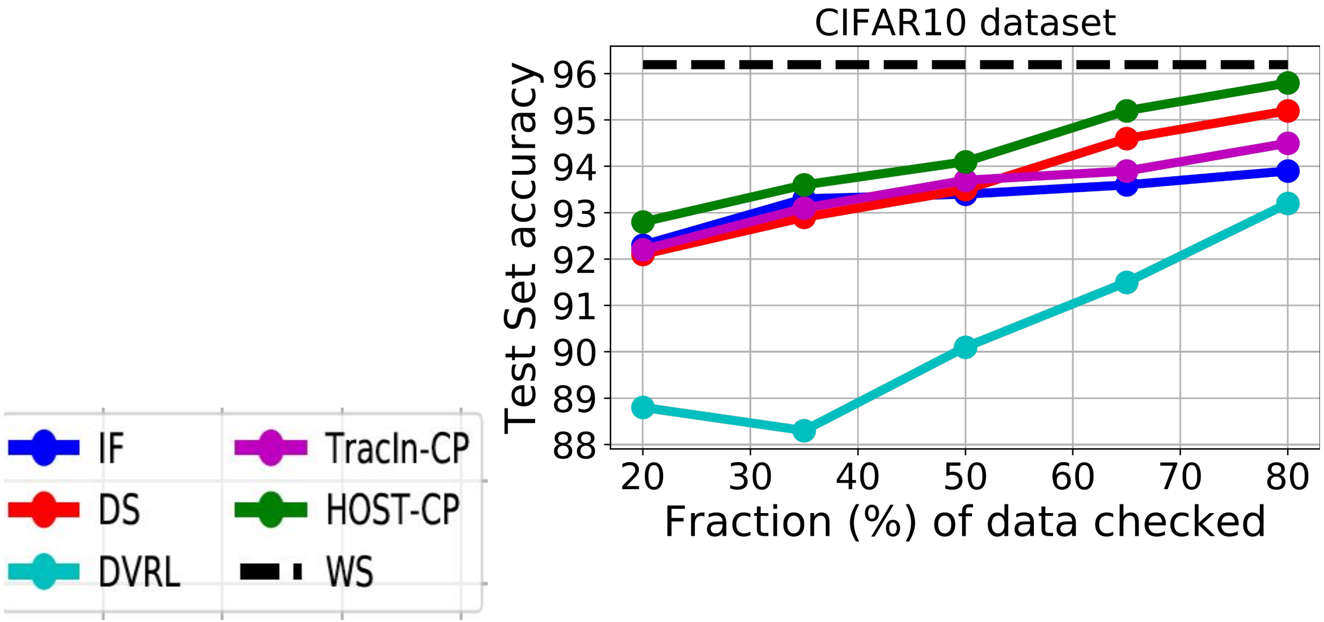
Top few images were from different classes - deer, bird, truck, dog, etc

HOST-CP

Experimental Results : Mislabelling



Experimental Results : Mislabelling



Variation of accuracies with fraction of inspected data from CIFAR10 dataset

Conclusion

1. Proposed a technique for finding high-values subsets essential for better test set predictions.
2. Designed a learning convex framework for subset selection.
3. Compared the method against S.O.T.A baselines to show that the proposed method performs comparatively better thus giving considerably better subsets.

Paper: https://arxiv.org/abs/2104.13794 ***Github***: https://github.com/SoumiDas/HOST-CP

THANK YOU!!