Data-centric Trusted AI

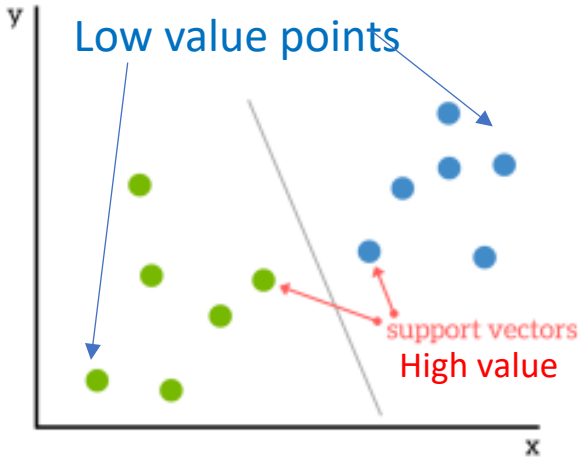
Soumi Das, Sourangshu Bhattacharya, Suparna Bhattacharya
IIT Kharagpur, HPE

http://cse.iitkgp.ac.in/~sourangshu/

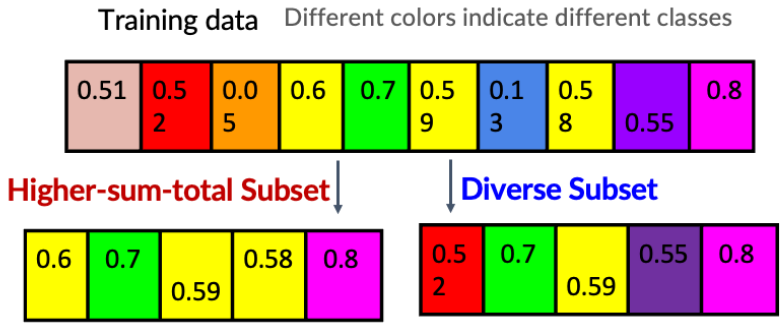
Email: sourangshu@cse.iitkgp.ac.in

Data Valuation and Subset Selection (DVSS)

Data Valuation – Estimate *contribution* of a training datapoint towards a task



Data Subset Selection – Select any *high-value subset* of a training dataset, which is of a fixed size.
- Coreset.



A value function is defined on a subset, need not be additive.

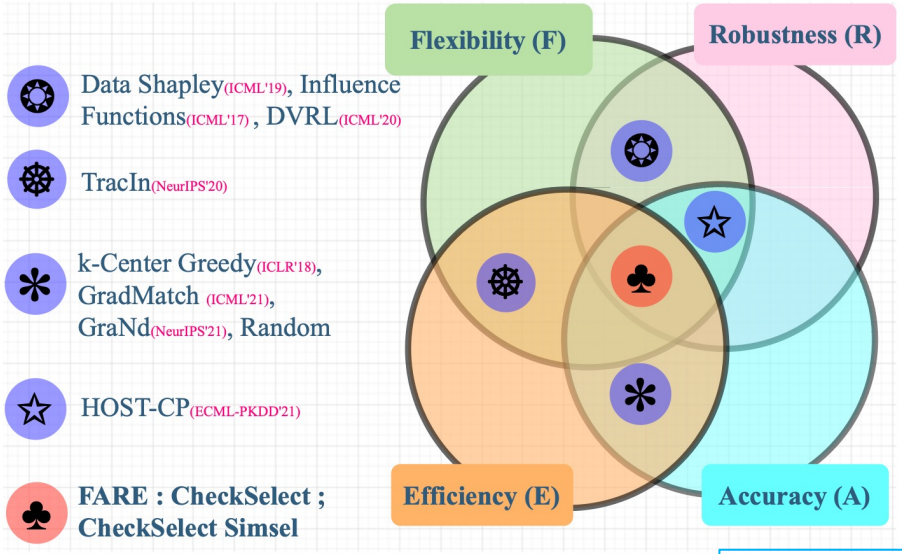
Recent Methods:
- Influence Functions (ICML 2017)
- Data Shapley (ICML2019)
- TracIn (Neurips 2020)

Recent Methods:
- K-center Greedy (ICLR 2018)
- GraNd (Neurips 2021)
- HOST-CP (ECML 2021)

Desirable Properties of DVSS Techniques

Flexibility: The technique should work with value functions capturing TAI objectives e.g. robustness, fairness, generalization, etc.

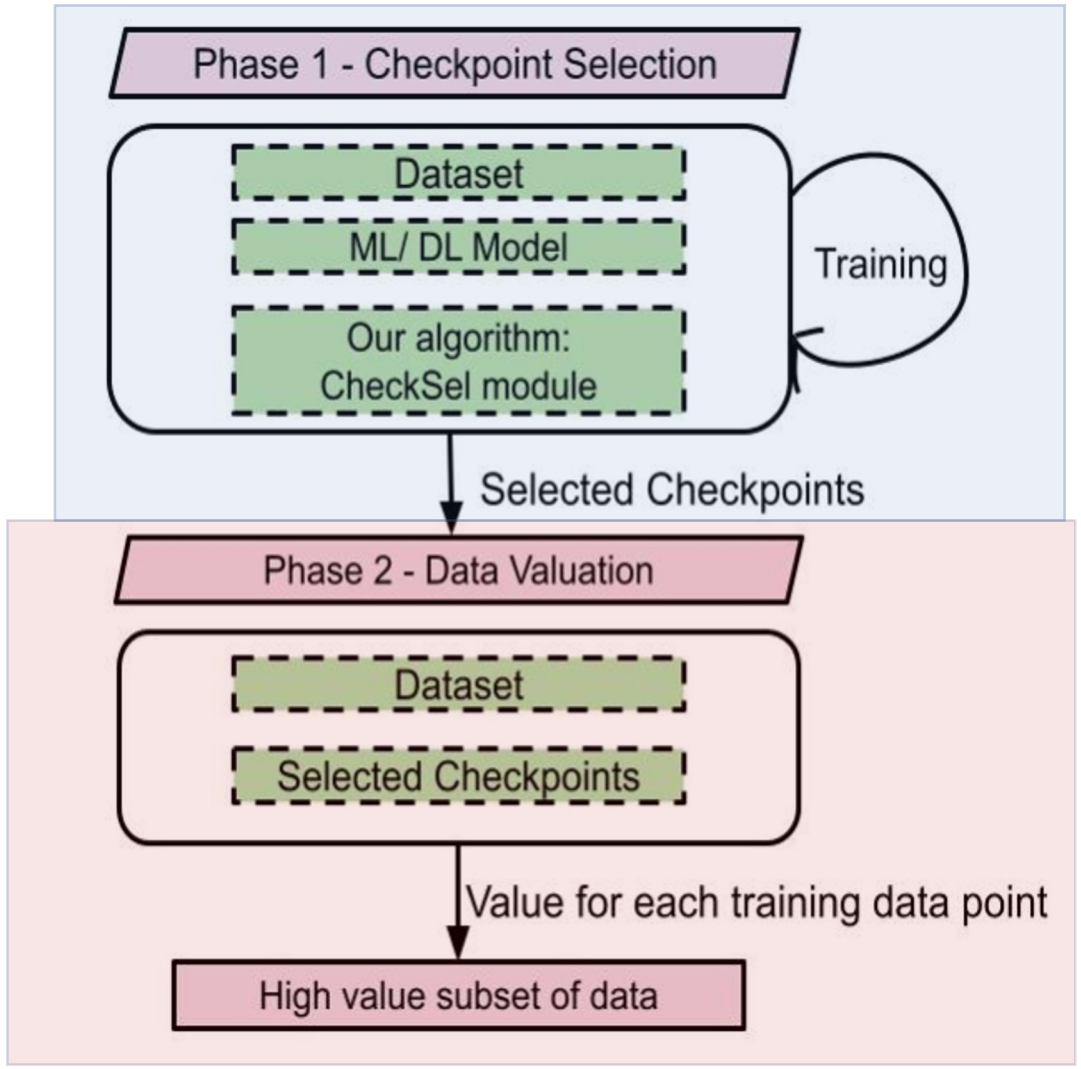
Robustness: The technique should work with a related dataset without re-training e.g. transfer from real world images to art images.



Efficiency: Time complexity of DVSS should be comparable to model training on entire data.

Accuracy: A model trained on the selected subset of a dataset should perform as well as the whole dataset.

Architecture and algorithms for DVSS



- Decrease in Validation loss through a training trajectory can be estimated as:

$$TracIn(d, d') = \sum_t \eta_t \nabla l(\theta_t, d).\nabla l(\theta_t, d')$$

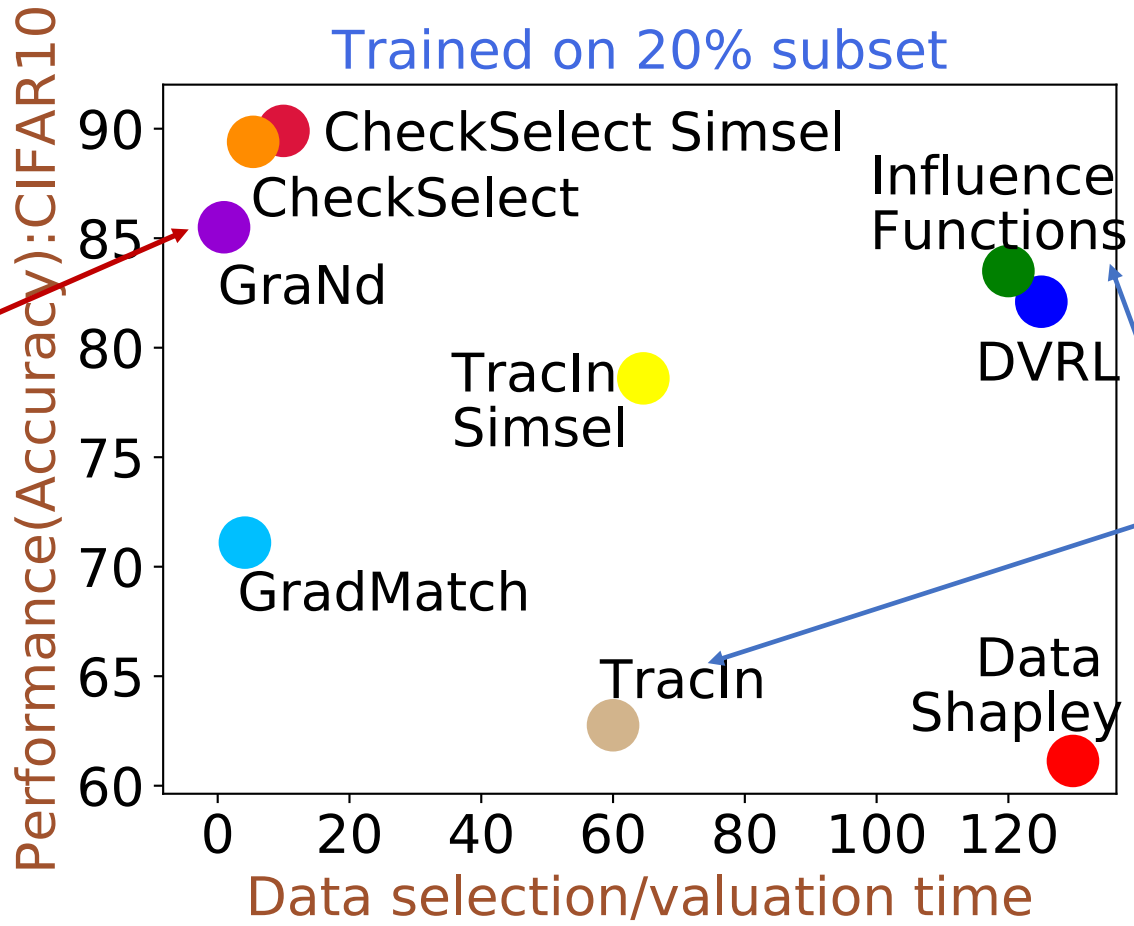
- This equation scores the influence of training datapoint *d* on loss of test datapoint *d'*
- All checkpoints are impossible to store.
- Hence, select influential checkpoints
- Checkpoint selection takes time similar to training

- Decouple the checkpoint selection and data valuation or subset selection module.
- Simsel algorithm can be used for selection of diverse
- Data Valuation takes time similar to inference.
- Subset selection time depends on validation and training set size.

https://github.com/SoumiDas/CheckSel

Empirical Results on CIFAR10

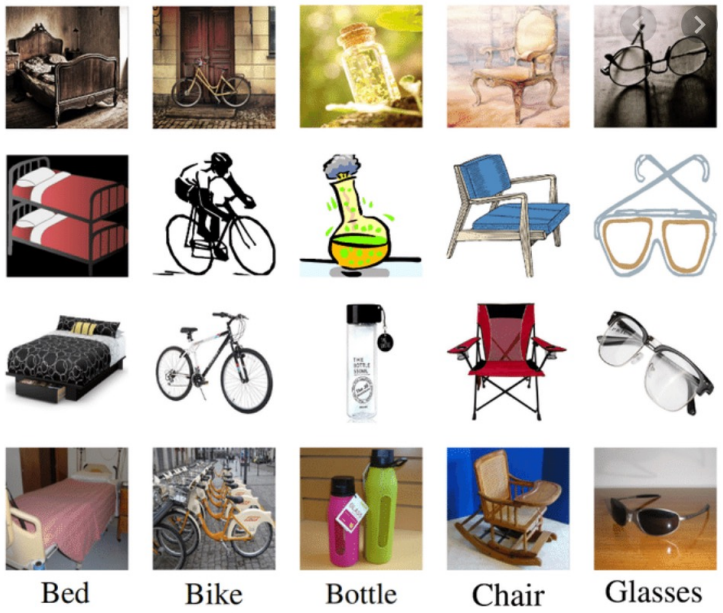
Better accuracy over the closest baseline which is not flexible.



An Order of magnitude faster than other flexible methods

Empirical Results on MS Office-Home dataset

Methods	Source ->Target											
	A->C	A->P	A->R	C->A	C->P	C->R	P->A	P->C	P->R	R->A	R->C	R->P
Random	45.03	57.74	56.5	48.9	57.74	56.5	48.9	45.03	56.5	48.9	45.03	57.74
k-center [5]	46.3	67.41	51.4	49.48	61.5	59.45	52.47	40.54	57.56	48.81	40.54	60.0
GraNd[1]	45.86	62.2	58.74	51.5	64.4	64.3	53.62	49.52	61.2	50.53	45.5	54.42
TracIn[4]	31.4	33.95	34.63	36.34	39.76	34.87	34.62	30.85	36.28	33.76	29.9	33.25
CheckSelect	**47.99**	**70.11**	**67.49**	**51.61**	**69.18**	**65.48**	**54.4**	**57.32**	**69.03**	**50.75**	**48.64**	**64.41**
Δ	**1.69**	**2.7**	**8.75**	**0.11**	**4.78**	**1.18**	**0.78**	**7.8**	**7.83**	**0.22**	**3.14**	**4.41**



7 – 8 % better accuracy for 20% subset selection over closest baseline.