A Novel Automated Approach for Software Effort Estimation Based on Data Augmentation

Liyan Song[1,2], Leandro L. Minku[1], Xin Yao[1,2]

[1] University of Birmingham, UK
[2] Southern University of Science and Technology, China

Software Effort Estimation (SEE)

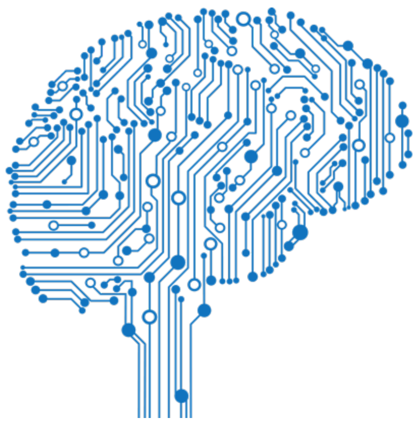
- Estimation of the effort required to develop a software project (e.g., in person-hours).
- Based on project features such as:
  - estimated size,
  - required reliability,
  - programming language,
  - development type,
  - etc.
- Both over and underestimations can be problematic.

SEE as a Machine Learning Problem

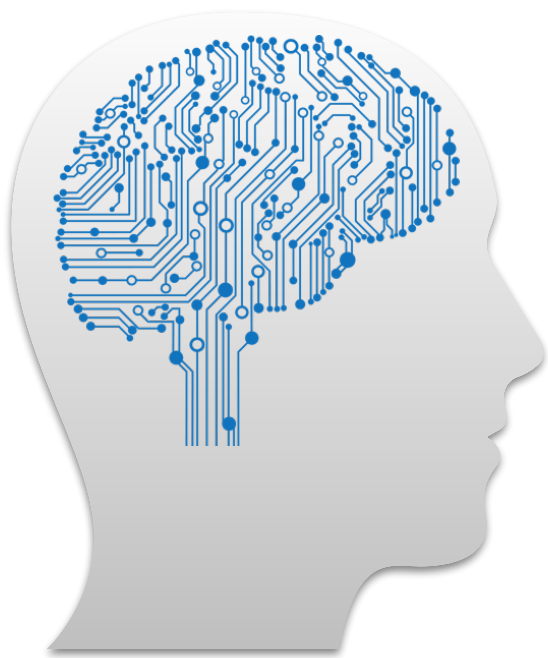
Previous projects are used as training data

Project id	x1 = size	x2 = reliability	x3 = language	…	y = effort ?
1	1000	medium	Java	…	850
2	1000	low	Matlab	…	500
3	900	large	C#	…	1000
…	…	…	…	…	…

→ Machine Learning Algorithm →



New project x →



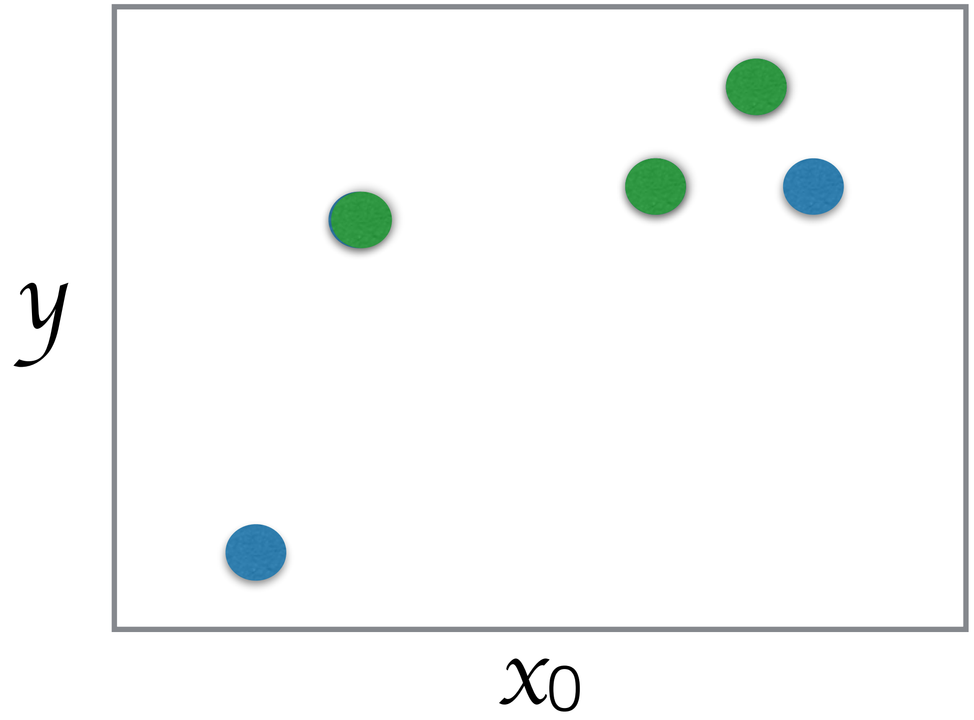
→ required effort y

A Key Challenge

- High cost of collecting effort required to develop projects.
- Scarcity of training data.
- Small training sets can lead to poor predictive performance.
- Most existing work investigates different machine learning algorithms to try to tackle this issue.

Data Augmentation

We generate additional synthetic projects based on existing ones.



Synthetic projects can enrich the representativeness of the area where they are generated, potentially leading to better SEE models.

How to Create Synthetic Projects?

1. Randomly select an existing training project.
2. Create a clone of this training project.
3. For each of the clone's input features.
    1. Displace this input feature with a certain probability.
4. Displace the clone's effort.

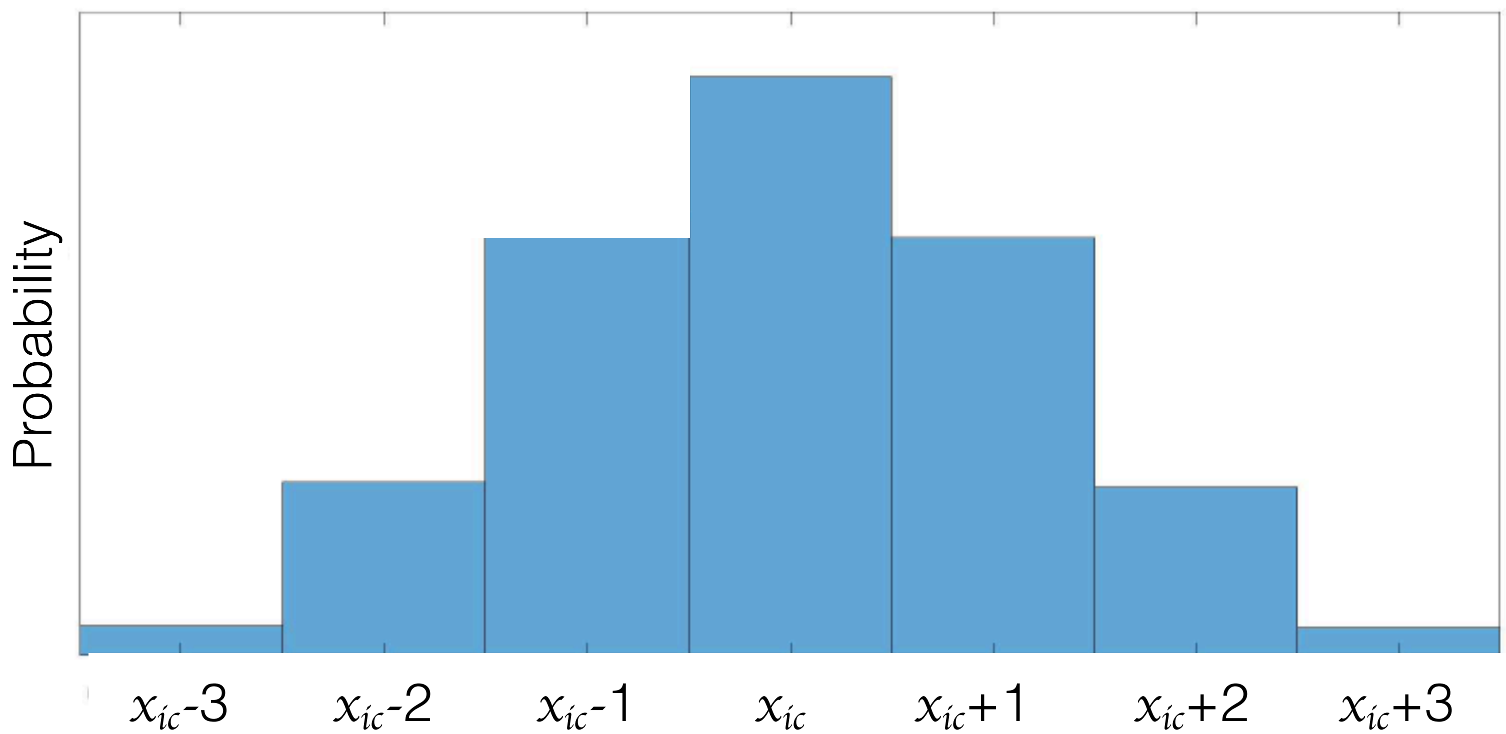
Displacing Categorical Input Features

With probability τ, uniformly sample a new value from:

$$\{v_1, v_2, \ldots, v_k\} \setminus \{x_{ic}\}$$

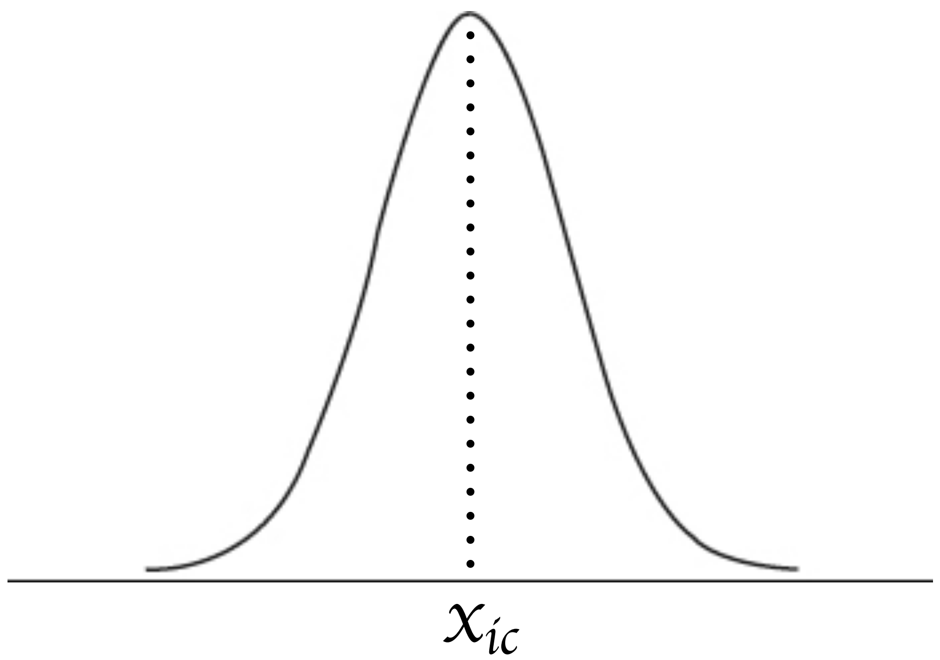
Displacing Ordinal Input Features

Sample a new value from $\mathcal{B}(n{=}2x_{ic}, p{=}1/2)$



Displacing Numeric Input Features

Sample a new value from $x_{ic} + \mathcal{N}(0,\sigma^2)$, where σ is a pre-defined parameter that should assume small values.



Displacing the Effort

Sample a new value from $y + \mathrm{sign}(e) \cdot |\mathcal{N}(0,\sigma^2)|$

e = sum of all Normal values used to displace the numeric size-related features.

Experiments

- Aims:
  - Evaluate the effect of synthetic data on predictive performance.
  - Understanding when and why the synthetic projects can help improving the baseline predictive performance.
- Machine learning algorithms:
  - LR, ATLM, k-NN, RVM, RT, SVR.
  - Proposed data augmentation.
  - SMOTE for SEE.
- MAE_{log} = Mean Absolute Error of the estimations in the log scale.

Datasets

Training set size

Size	Data set	#Fea	#Data	#Data/#Fea	Small	Medium	Large
Small	Maxwell	23	62	2.70	0.3	0.7	LOO
	Cocomo81	17	63	3.71			
	Nasa93	17	93	5.47			
	Albrecht	7	24	3.43			
	Kemerer	6	16	2.67			
Medium	Desharnais	8	77	9.63	0.1	0.3	0.7
	Org2	3	32	10.67			
	Org5	3	21	7.00			
	Org6	1	22	22.00			
	Org7	1	20	20.00			
Large	Kitchenham	3	145	48.33	0.04	0.08	0.7
	Org1	3	76	25.33			
	Org3	3	162	54.00			
	Org4	3	122	40.67			

ISBSG (International Software Benchmarking Standards Group)

SEACRAFT (Software Engineering Artifacts Can Really Assist Future Tasks)

RQ1

- Given a learning algorithm, can our data augmentation approach help improving prediction performance over its baseline? When? Could it be detrimental?
  - For most baselines and training set sizes, the proposed approach significantly improved MAE_{log}, according to Wilcoxon Rank Sum tests with Holm-Bonferroni corrections across data sets.
  - The proposed approach was never significantly worse across data sets.
  - Effect size (A12) of improvement depends on the baseline and training set size.
    - Small ($A12 \geq 0.56$), medium ($A12 \geq 0.64$) and large ($A12 \geq 0.71$)

RQ1 - LR and ATLM

MAE_{log} for Small Training Set Size

Data	syn.LR	bsl.LR	syn.ATLM	bsl.ATLM
Maxwell	**0.645±0.095**	**1.314±0.549**	**0.649±0.101**	**14.470±32.537**
Cocomo81	**0.654±0.135**	**8.596±13.643**	**0.668±0.140**	**17.543±39.428**
Nasa93	**0.534±0.082**	**0.942±0.877**	**0.540±0.081**	**0.927±0.836**
Kitchenham	**0.653±0.149**	**0.765±0.243**	**0.657±0.180**	**0.757±0.253**
Albrecht	**0.823±0.261**	**3.499±4.208**	**0.817±0.267**	**48.975±235.167**
Kemerer	1.058±0.573	1.712±2.152	**1.121±1.059**	**5.703±17.881**
Deshar	**0.695±0.193**	**2.163±4.230**	**0.699±0.188**	**3.235±4.758**
Org1	**1.324±1.361**	**2.133±2.337**	**1.004±0.400**	**668.348±3648.420**
Org2	1.092±1.634	1.343±2.222	0.785±0.325	0.975±0.846
Org3	0.684±0.146	0.744±0.229	**0.682±0.148**	**0.745±0.231**
Org4	**0.902±0.258**	**2.341±3.983**	**0.916±0.342**	**5.298±20.085**
Org5	**2.177±2.983**	**3.837±4.172**	**1.231±1.287**	**2.413±3.398**
Org6	**1.003±0.508**	**2.680±3.728**	**1.111±0.576**	**2.123±2.408**
Org7	**1.156±0.651**	**1.868±2.498**	**1.179±0.674**	**1.890±2.494**
aveRank	1.00	2.00	1.00	2.00
Wilcoxon	1	0.000122	1	0.000122

Improvements were frequently large when training sets were small or medium, especially for the small training sets.

RQ1 - RVM and RT

MAE$_{log}$ for Small Training Set Size

Data	syn.RVM	bsl.RVM	syn.RT	bsl.RT
Maxwell	**0.584±0.064**	**0.643±0.090**	**0.667±0.100**	**0.693±0.111**
Cocomo81	**0.684±0.115**	**0.779±0.143**	**1.100±0.176**	**1.172±0.135**
Nasa93	0.532±0.113	0.534±0.121	**0.728±0.078**	**0.796±0.083**
Kitchenham	**0.696±0.153**	**0.831±0.225**	**0.802±0.170**	**0.832±0.107**
Albrecht	**0.673±0.171**	**0.766±0.303**	**0.806±0.182**	**0.920±0.130**
Kemerer	0.665±0.151	0.615±0.170	0.799±0.155	0.818±0.150
Deshar	**0.583±0.095**	**0.626±0.160**	**0.639±0.094**	**0.692±0.068**
Org1	**0.922±0.228**	**0.988±0.234**	1.027±0.297	1.000±0.256
Org2	0.645±0.179	0.637±0.129	0.762±0.209	0.747±0.189
Org3	**0.753±0.192**	**0.855±0.188**	**0.835±0.142**	**0.971±0.101**
Org4	0.836±0.096	0.846±0.109	0.897±0.136	0.892±0.116
Org5	1.042±0.270	1.287±1.366	1.060±0.195	1.036±0.172
Org6	**0.999±0.320**	**1.089±0.260**	1.159±0.244	1.165±0.248
Org7	0.953±0.235	0.946±0.155	0.959±0.167	0.946±0.137
aveRank	1.21	1.79	1.36	1.64
Wilcoxon	1	0.006714	0	0.057983

Improvements were frequently medium or large when training sets were small or medium.

RQ1 - k-NN and SVR

MAE$_{log}$ for Small Training Set Size

Data	syn.k-NN	bsl.k-NN	syn.SVR	bsl.SVR
Maxwell	0.724±0.090	0.731±0.085	**0.570±0.087**	**0.598±0.075**
Cocomo81	1.266±0.142	1.297±0.142	**0.640±0.118**	**0.703±0.152**
Nasa93	0.990±0.111	0.984±0.107	0.519±0.079	0.544±0.148
Kitchenham	0.744±0.168	0.748±0.156	**0.621±0.119**	**0.676±0.138**
Albrecht	0.724±0.113	0.717±0.121	0.580±0.128	0.574±0.110
Kemerer	0.643±0.142	0.685±0.143	**0.526±0.147**	**0.575±0.157**
Deshar	0.622±0.088	0.618±0.084	0.526±0.051	0.526±0.067
Org1	**0.895±0.203**	**0.907±0.134**	**0.853±0.209**	**0.874±0.160**
Org2	0.659±0.208	0.671±0.185	0.633±0.182	0.645±0.201
Org3	**0.767±0.132**	**0.782±0.114**	**0.647±0.124**	**0.701±0.189**
Org4	0.860±0.156	0.863±0.135	**0.800±0.099**	**0.840±0.131**
Org5	0.971±0.232	1.009±0.239	**0.771±0.188**	**0.938±0.265**
Org6	0.959±0.255	0.961±0.273	0.860±0.236	0.888±0.262
Org7	0.917±0.168	0.909±0.150	0.923±0.220	0.892±0.148
aveRank	1.29	1.71	1.14	1.86
Wilcoxon	0	0.056274	1	0.005249

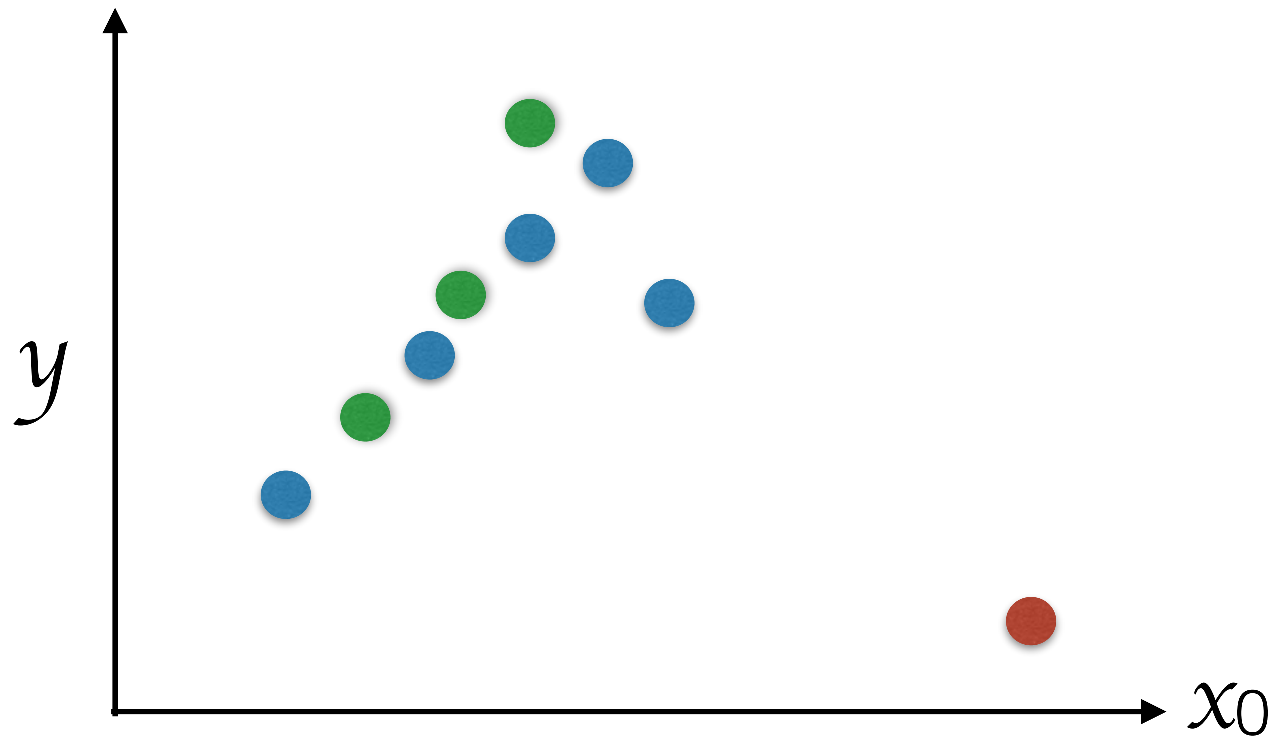
Improvements had small or insignificant effect size for all training set sizes, but there was no significant detrimental effect.

RQ2

- Why our synthetic projects are helpful? Why the magnitude of improvement varies depending on the baseline model?

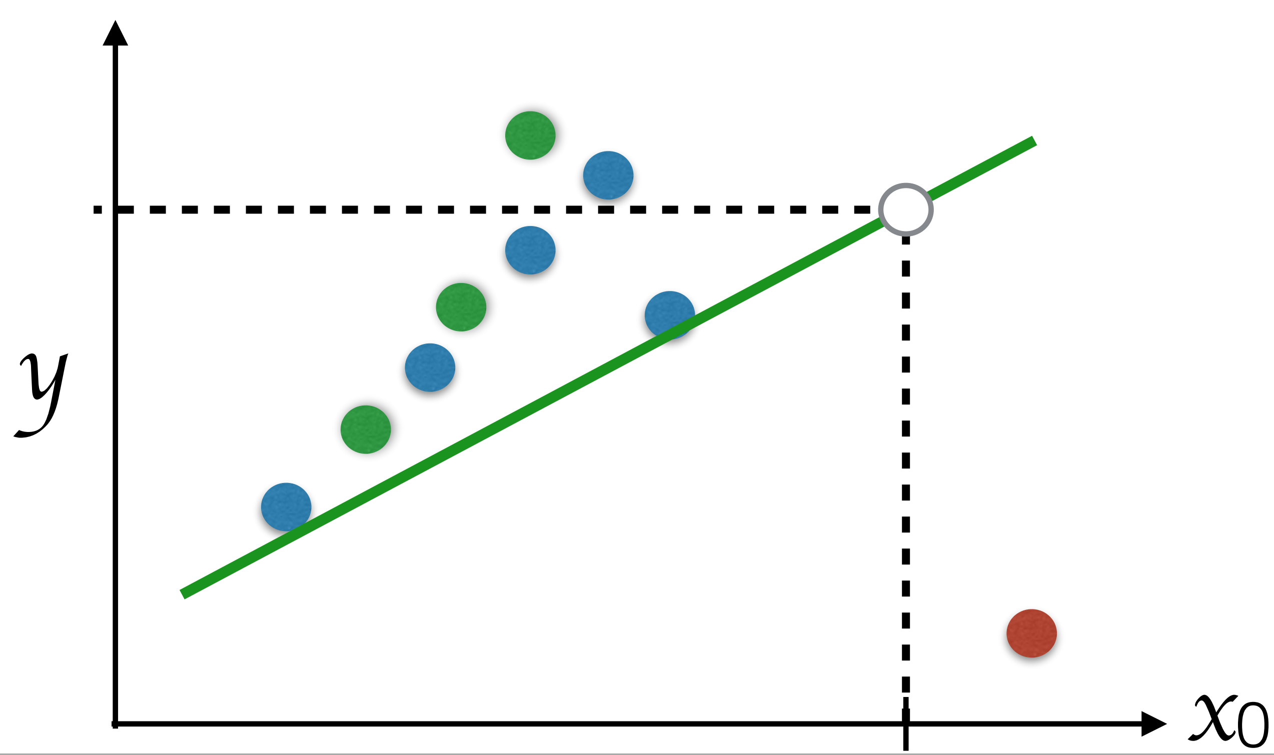
RQ2

- **Why our synthetic projects are helpful?** Why the magnitude of improvement varies depending on the baseline model?
    - Increasing the training set size helps to cope with lack of data and large noise.



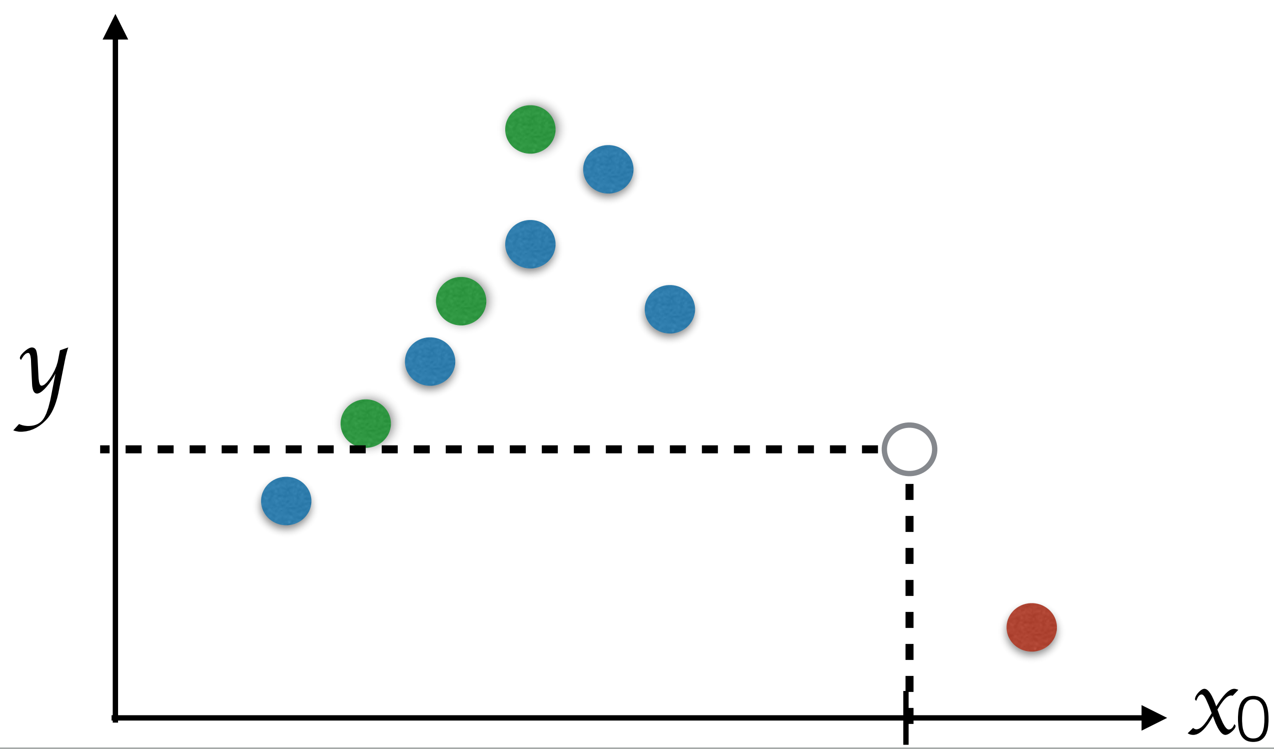
RQ2

- Why our synthetic projects are helpful? **Why the magnitude of improvement varies depending on the baseline model?**
  - LR/ATLM — global approaches.
    - Effect of synthetic data will impact predictions in the entire space.



RQ2

- Why our synthetic projects are helpful? **Why the magnitude of improvement varies depending on the baseline model?**
  - k-NN — local approach.
    - Synthetic data will only influence estimations if they are neighbours, reducing the effect of synthetic data.



RQ3

- How well does our data augmentation approach perform against the existing data augmentation approach from the SEE literature?

RQ3

MAE$_{log}$ for Small Training Set Size

Data	syn.LR	SMOTE .LR	syn.ATLM	SMOTE .ATLM
Maxwell	**0.645±0.095**	1.336±0.487	**0.649±0.101**	14.682±32.480
Cocomo81	**0.654±0.135**	8.596±13.643	**0.668±0.140**	12.827±28.833
Nasa93	**0.534±0.082**	0.958±0.913	**0.540±0.081**	0.949±0.863
Kitchenham	**0.653±0.149**	0.772±0.259	**0.657±0.180**	0.767±0.263
Albrecht	**0.823±0.261**	3.499±4.208	**0.817±0.267**	4.835±6.721
Kemerer	1.058±0.573	1.306±1.506	**1.121±1.059**	19.447±92.795
Deshar	**0.695±0.193**	2.189±4.224	**0.699±0.188**	2.787±3.025
Org1	**1.324±1.361**	2.133±2.337	**1.004±0.400**	668.348±3648.420
Org2	1.092±1.634	1.343±2.222	0.785±0.325	0.975±0.846
Org3	**0.684±0.146**	0.751±0.234	**0.682±0.148**	0.750±0.238
Org4	**0.902±0.258**	2.346±3.980	**0.916±0.342**	5.305±20.084
Org5	**2.177±2.983**	3.837±4.172	**1.231±1.287**	2.413±3.398
Org6	**1.003±0.508**	2.680±3.728	**1.111±0.576**	2.123±2.408
Org7	**1.156±0.651**	1.868±2.498	**1.179±0.674**	1.890±2.494
Wilcoxon	1	0	1	0
p-value	0.000091	0.250000	0.000091	1.000000

Proposed approach performs always similarly or better across data sets, with larger effect sizes for small or medium training sets when using LR, ATLM, RVM or RT.

Conclusions

- Proposed a novel data augmentation approach for SEE.
- RQ1: proposed approach leads to similar or better MAE_{log} than its baselines. Effect size of improvements is larger for small/medium training sets when using LR/ATLM and RT/RVM.
- RQ2: improvements are obtained due to larger datasets presenting better robustness to large noise. Their effect depends on intrinsic aspects of the base learner such as globality and locality.
- RQ3: proposed approach leads to similar or better MAE_{log} than an existing data augmentation approach for SEE. Effect size is larger especially for small/medium training sets when using LR/ATLM and RT/RVM.

The proposed approach can help to improve predictive performance when there is lack of training data.

Future Work

- Proposal of new strategies to displace the effort.
- Analysis with more performance metrics.
- Investigation of the proposed approach for other problems.