How to Make Best Use of Cross-Company Data for Web Effort Estimation?

Leandro L. Minku
University of Leicester, UK

Leandro Minku, Federica Sarro, Emilia Mendes and Filomena Ferrucci.
How to Make Best Use of Cross-Company Data for Web Effort Estimation?
Proceedings of the 9th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'15)
(best paper award)

Introduction

- Software effort estimation is the estimation of effort (e.g., person-hours) required to develop software projects.

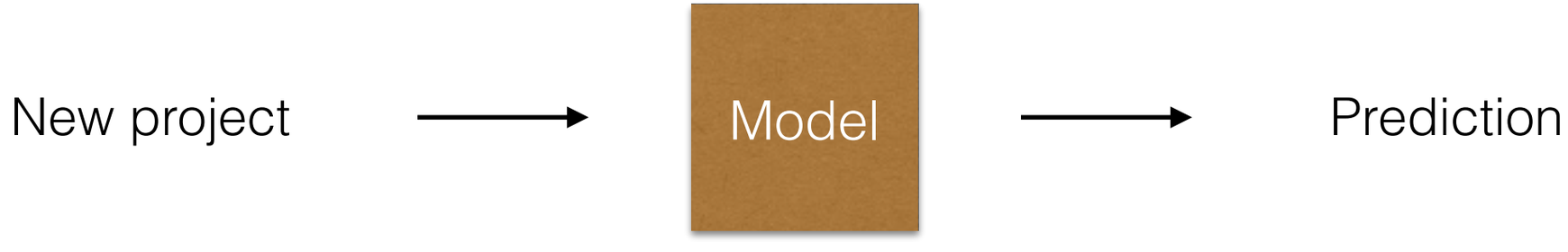
Introduction

- Web effort estimation is the estimation of effort (e.g., person-hours) required to develop web projects.
- Web effort estimation can be based on web project features, e.g., team expertise, number of web pages, number of images, etc.
- Over vs underestimations.

[17] E. Mendes. Practitioner's Knowledge Representation. Springer-Verlag, 2014, DOI: 10.1007/978-3-642-54157-5 2.

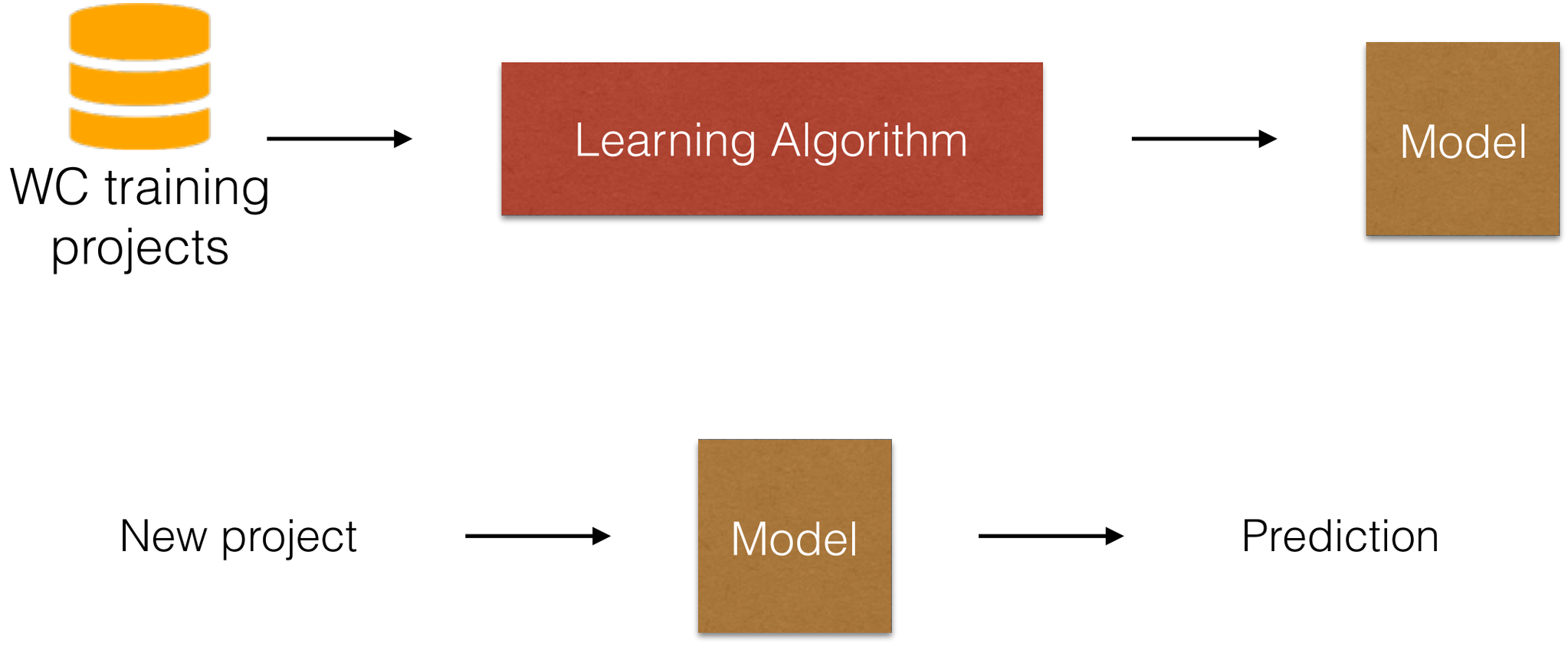
Machine Learning for Effort Estimation

Machine learning models can be used to perform effort estimations for a new project based on data describing past projects.



Within-Company (WC) Effort Estimation Models

Early studies suggested that general-purpose models (e.g., COCOMO) needed to be calibrated to specific companies.



Within-Company (WC) Effort Estimation Models

Problems of using only within-company (WC) data:

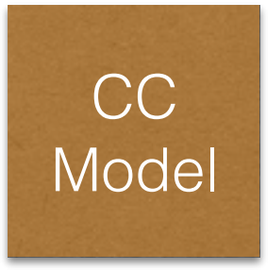
- Time to accumulate enough data may be prohibitive.
- By the time enough data are collected, they may be obsolete.
- Data need to be collected in a consistent manner.

[1] B. Boehm. Software Engineering Economics. Prentice-Hall, Englewood Cliffs, NJ, 1981.
[13] B. Kitchenham and N. Taylor. Software cost models. ICL Technical Journal, pages 73–102, 1984.
[16] P. Kok, B. Kitchenham, and J. Kirawkowski. The mermaid approach to software cost estimation. In ESPRIT, pages 296–314. 1990.

Cross-Company (CC) Effort Estimation Models

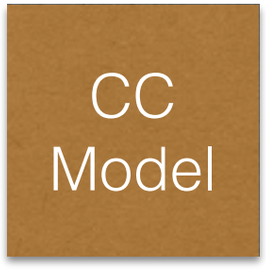
CC models are alternatives to WC models.
[CC term used loosely.]

CC training projects



E.g.:
ISBSG (www.isbsg.org)
PROMISE (http://openscience.us/repo/)

New WC project



Prediction

Cross-Company (CC) Effort Estimation Models

Problem: CC data may have different characteristics from WC data, leading to poorly performing models.

Making CC Data More Similar to WC Data

- Strategies to make CC data more similar to WC data (e.g., TEAK, NN filtering, Dycom) have been achieving more promising results.
  - Web projects:
    - TEAK provided competing performance (ties) against WC models in 6 out of 8 data sets.
    - NN-filtering provided competing (ties) performance in 7 out of 8 data sets.
  - Conventional projects:
    - Dycom provided competing (ties or wins) in 5 out of 5 data sets.

[15] E. Kocaguneli, T. Menzies, and E. Mendes. Transfer learning in effort estimation. Empirical Software Engineering, pages 1–31, 2014.
[33] B. Turhan and E. Mendes. A comparison of cross- versus single- company effort prediction models for web projects. In Euromicro Conference on Software Engineering and Advanced Applications, pages 285–292, 2014.
[28] L. L. Minku and X. Yao. How to make best use of cross-company data in software effort estimation? In ICSE, pages 446–456, 2014.

CC Web Effort Estimation

Our study is geared towards enabling Web development companies to make more efficient managerial decisions worthwhile, by investigating Dycom.

[17] E. Mendes. Practitioner's Knowledge Representation. Springer-Verlag, 2014, DOI: 10.1007/978-3-642-54157-5 2.

Research Questions

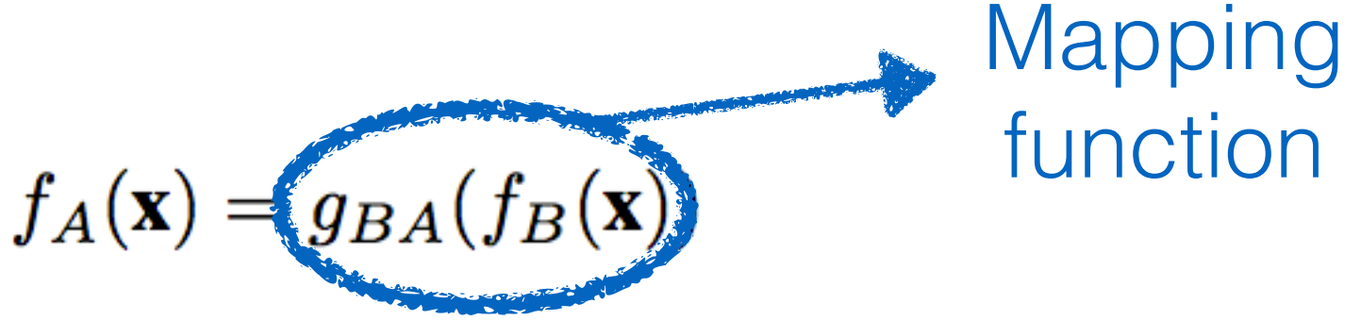
RQ1. How successful is a CC dataset at estimating effort for Web projects from a single company?

RQ2. How successful is the use of a CC dataset compared to a WC dataset for Web effort estimation?

RQ3. How does Dycom perform with respect to other techniques previously used for CC Web effort estimation?

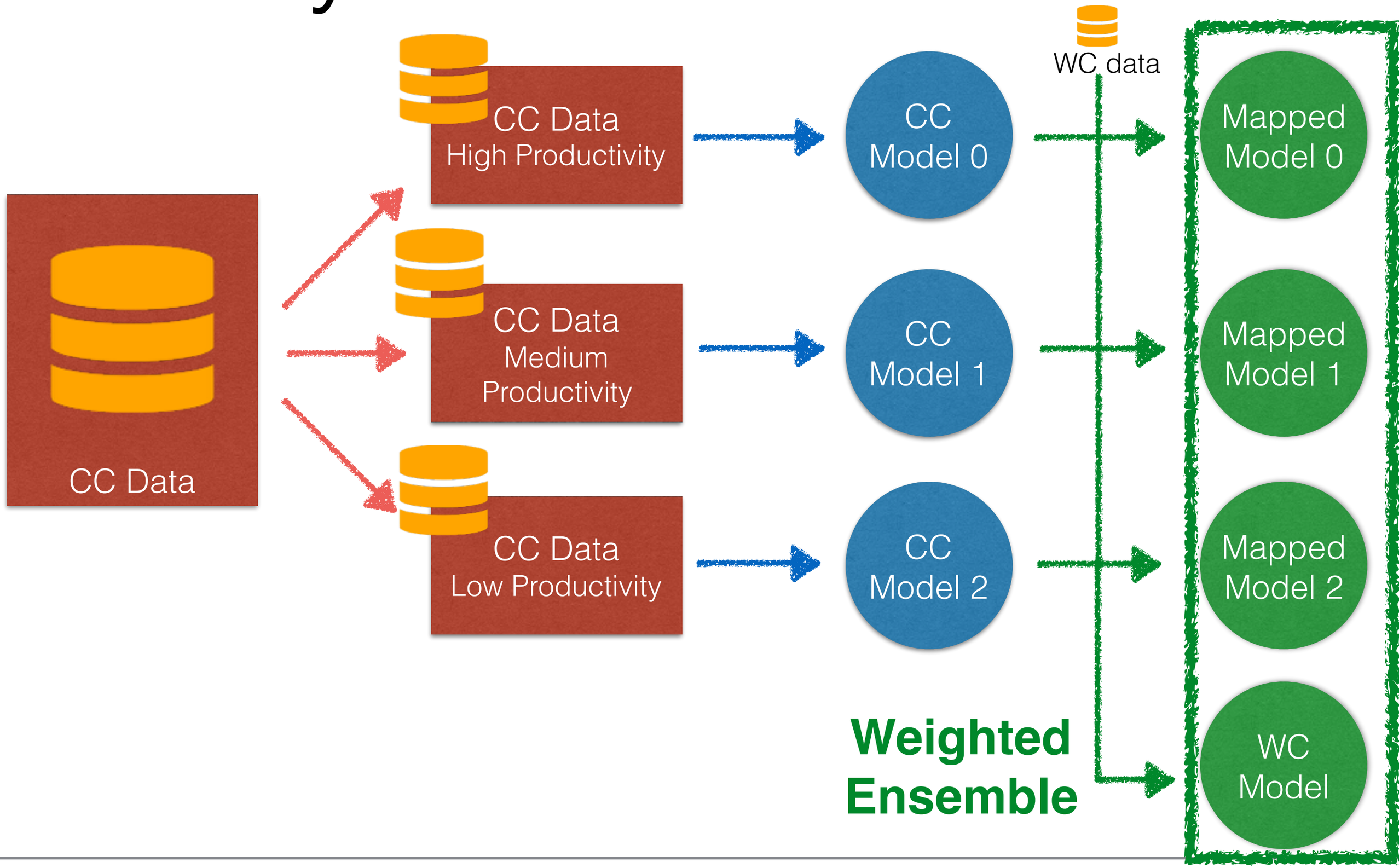
Dynamic **C**ross-C**o**mpany **M**apped Model Learning (Dycom)

There is a relationship between the effort of two companies A and B:



Effort estimation models can be built by learning (1) CC models and (2) mapping functions based on a limited number of WC data.

Dycom - Ensemble



Dycom - Learning a Mapping Function for a Cross-Company Model

$$f_{WC}(\boldsymbol{x}) = g_i(f_{CCi}(\boldsymbol{x})) = f_{CCi}(\boldsymbol{x}) \cdot b_i$$

$$b_i = \begin{cases} 1, & \text{if no WC training example has been received yet;} \\ \dfrac{y}{f_{CCi}(\boldsymbol{x})}, & \text{if } (\mathbf{x},y) \text{ is the first WC training example;} \\ lr \cdot \dfrac{y}{f_{CCi}(\boldsymbol{x})} + (1 - lr) \cdot b_i, & \text{otherwise.} \end{cases}$$

Data Sets

8 WC data sets from the Tukutuku database.

Variable	Description
nlang	Number of different development languages used.
DevTeam	Size of a project's development team.
TeamExp	Avg team experience with the development language(s) used.
TotWP	Total number of Web pages (new and reused).
NewWP	Total number of new Web pages.
TotImg	Total number of images (new and reused).
NewImg	Total number of new images created.
Fots	Number of features reused without any adaptation.
HFotsA	Number of reused high-effort features/functions adapted.
Hnew	Number of new high-effort features/functions.
totHigh	Total number of high-effort features/functions
FotsA	Number of reused low-effort features adapted.
New	Number of new low-effort features/functions
totNHigh	Total number of low-effort features/functions
TotEff	Actual total effort used to develop the Web application.

[23] E. Mendes, N. Mosley, and S. Counsell. Investigating web size metrics for early web cost estimation. JSS, 77(2):157–172, 2005.

Data Sets

8 WC data sets from the Tukutuku database.

WC Data	Avg Productivity	# of Projects	% Projects
C1	2.03	14	11.2
C2	4.61	20	16
C3	0.87	15	12
C4	2.49	6	4.8
C5	1.42	13	10.4
C6	0.67	8	6.4
C7	0.90	31	24.8
C8	1.20	18	14.4
Total	-	125	100

Experimental Analysis

RQ1. How successful is a CC dataset at estimating effort for Web projects from a single company?

- Comparison between Dycom and mean and median baselines.
  - For each WC data set, consider all other WC data sets as the CC data.
  - Amount of WC training data used by Dycom: 10% and 50% of original data set.
  - Base learner: regression trees.
  - Performance measures: MAE, MAEL, SA.
  - Wilcoxon Sign-Rank tests with Holm-Bonferroni corrections.
  - Thirty runs with different training and testing partitions.

RQ1 - Results

Test Set	Mean vs. Dycom	MAE	SA	MAEL
C1	Mean-P2	542.9989	-747.3199	3.0267
	Dycom-RT-P2	36.6201	42.8563	0.65105
	P-value	2.00E-06		2.00E-06
C2	Mean-P2	590.8896	-588.0089	4.6718
	Dycom-RT-P2	4.2004	95.1092	0.553
	P-value	2.00E-06		2.00E-06
C3	Mean-P2	2170.0519	13.5880	1.7619
	Dycom-RT-P2	734.1386	70.7664	0.23605
	P-value	4.10E-05		2.00E-06
C4	Mean-P2	392.4489	-71.6249	2.1131
	Dycom-RT-P2	110.96	51.4752	0.7837
	P-value	2.00E-06		2.00E-06
C5	Mean-P2	465.0405	-19.2921	1.6425
	Dycom-RT-P2	321.19815	17.6063	0.9441
	P-value	1.25E-01		4.00E-06
C6	Mean-P2	490.7463	-545.7188	2.0440
	Dycom-RT-P2	36.052	52.5632	0.41305
	P-value	2.00E-06		2.00E-06
C7	Mean-P2	802.1830	-13.2760	1.8029
	Dycom-RT-P2	23.234	96.7191	0.1244
	P-value	2.00E-06		2.00E-06
C8	Mean-P2	421.2622	-194.8180	1.5218
	Dycom-RT-P2	128.99745	9.7218	0.5614
	P-value	4.00E-06		2.00E-06

Dycom performed almost always better than mean.

RQ1 - Results

Test Set	Median vs. Dycom	MAE	SA	MAEL
C1	Median-P2	57.9686	9.5432	1.2305
	Dycom-RT-P2	36.6201	42.8563	0.65105
	P-value	5.72E-01		1.00E-05
C2	Median-P2	89.5760	-4.2988	2.8610
	Dycom-RT-P2	4.2004	95.1092	0.553
	P-value	2.00E-06		2.00E-06
C3	Median-P2	2523.5714	-0.4892	3.4836
	Dycom-RT-P2	734.1386	70.7664	0.23605
	P-value	2.80E-05		2.00E-06
C4	Median-P2	185.2500	18.9869	2.0480
	Dycom-RT-P2	110.96	51.4752	0.7837
	P-value	4.00E-06		2.00E-06
C5	Median-P2	325.4167	16.5242	1.2256
	Dycom-RT-P2	321.19815	17.6063	0.9441
	P-value	5.30E-01		4.90E-04
C6	Median-P2	29.6250	61.0197	0.4109
	Dycom-RT-P2	36.052	52.5632	0.41305
	P-value	3.82E-01		7.97E-01
C7	Median-P2	605.6667	14.4740	1.2335
	Dycom-RT-P2	23.234	96.7191	0.1244
	P-value	2.00E-06		2.00E-06
C8	Median-P2	94.6667	33.7481	0.7119
	Dycom-RT-P2	128.99745	9.7218	0.5614
	P-value	2.11E-03		3.85E-03

Dycom performed similar or better than median most of the time.

NN-filtering performed worse than median in five cases.

Experimental Analysis

RQ2. How successful is the use of a CC dataset compared to a WC dataset for Web effort estimation?

- Comparison between Dycom and WC model.
  - For each WC data set, consider all other WC data sets as the CC data.
  - Amount of WC training data used by Dycom: 10% and 50% of original data set.
  - WC model is trained with all WC data apart from one project used for testing, in a modified leave-one-out procedure.
  - Base learner: regression trees.
  - Performance measures: MAE, MAEL, SA.
  - Wilcoxon Sign-Rank tests with Holm-Bonferroni corrections.
  - Thirty runs with different training and testing partitions.

RQ2 - Results

	Approach	MAE	SA	MAEL
C1	WC-RT	22.8779	43.8107	0.7362
	Dycom-RT	36.6201	10.0591	0.6511
	P-value	4.99E-03		4.41E-01
C2	WC-RT	5.2373	26.1423	0.6399
	Dycom-RT	4.2004	40.7643	0.5530
	P-value	1.71E-03		2.77E-03
C3	WC-RT	627.7143	28.5761	0.2615
	Dycom-RT	734.1386	16.4667	0.2361
	P-value	8.59E-02		7.04E-01
C4	WC-RT	64.7500	73.4631	0.4000
	Dycom-RT	110.9600	54.5246	0.7837
	P-value	3.32E-04		1.70E-06
C5	WC-RT	374.0833	6.8672	0.9879
	Dycom-RT	321.1982	20.0337	0.9441
	P-value	7.34E-01		6.00E-01
C6	WC-RT	44.7500	-7.1856	0.5736
	Dycom-RT	36.0520	13.6479	0.4131
	P-value	5.71E-02		7.73E-03
C7	WC-RT	223.7953	68.5268	0.3517
	Dycom-RT	23.2340	96.7325	0.1244
	P-value	2.40E-06		4.20E-04
C8	WC-RT	76.6667	27.7487	0.4242
	Dycom-RT	128.9975	-21.5683	0.5614
	P-value	2.16E-05		4.53E-04

Dycom performed frequently similarly or better than WC model.

Other approaches that try to make CC data more similar to WC data did not perform better than WC model.

Experimental Analysis

RQ3. How does Dycom perform with respect to other techniques previously used for CC Web effort estimation?

- Comparison between Dycom and NN-filtering.
  - For each WC data set, consider all other WC data sets as the CC data.
  - Amount of WC training data used by Dycom: 10% and 50% of original data set.
  - Base learner: regression trees.
  - Performance measures: MAE, MAEL, SA.
  - Wilcoxon Sign-Rank tests with Holm-Bonferroni corrections.
  - Thirty runs with different training and testing partitions.

RQ3 - Results

	Approach	MAE	SA	MAEL
C1	NN-Filtering-RT	22.0922	45.7405	0.9009
	Dycom-RT	36.6201	10.0591	0.6511
	P-value	4.99E-03		4.41E-01
C2	NN-Filtering-RT	15.8203	-123.1032	1.0056
	Dycom-RT	4.2004	40.7643	0.5530
	P-value	1.71E-03		2.77E-03
C3	NN-Filtering-RT	670.8572	23.6671	0.2864
	Dycom-RT	734.1386	16.4667	0.2361
	P-value	8.59E-02		7.04E-01
C4	NN-Filtering-RT	125.8413	48.4257	0.7564
	Dycom-RT	110.9600	54.5246	0.7837
	P-value	3.32E-04		1.70E-06
C5	NN-Filtering-RT	400.0417	0.4046	1.1105
	Dycom-RT	321.1982	20.0337	0.9441
	P-value	7.34E-01		6.00E-01
C6	NN-Filtering-RT	35.8375	14.1617	0.5393
	Dycom-RT	36.0520	13.6479	0.4131
	P-value	5.71E-02		7.73E-03
C7	NN-Filtering-RT	226.3800	68.1633	0.4112
	Dycom-RT	23.2340	96.7325	0.1244
	P-value	2.40E-06		4.20E-04
C8	NN-Filtering-RT	73.0556	31.1518	0.4309
	Dycom-RT	128.9975	-21.5683	0.5614
	P-value	2.16E-05		4.53E-04

Dycom always performed similar or better than NN-filtering, except in one case.

Conclusions

RQ1. How successful is a CC dataset at estimating effort for Web projects from a single company?

- CC data can be successful in estimating effort for web projects from a single company when using Dycom -- it was almost always better than mean, median or random guess.

RQ2. How successful is the use of a CC dataset compared to a WC dataset for Web effort estimation?

- Dycom performed frequently similarly or better than a WC model while using only half of WC data.

RQ3. How does Dycom perform with respect to other techniques previously used for CC Web effort estimation?

- Dycom performed similarly or better than NN-filtering in all cases except for one.

Implications to Practice

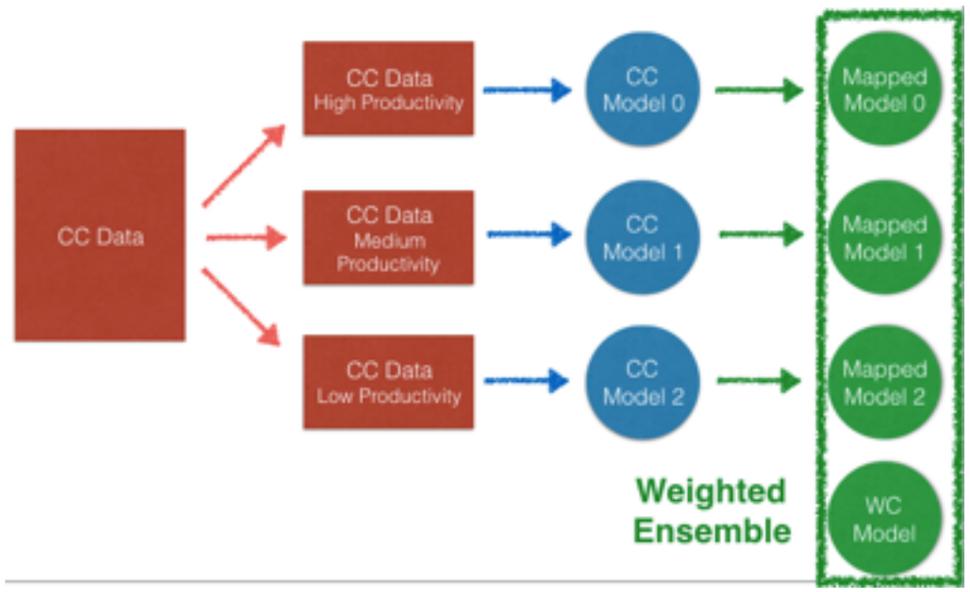
- Dycom can be a competitive choice for Web companies similar to the ones in this study and who have just a few WC projects.
  - A simple interface for use by companies should be implemented so that empirical studies on site can be performed.
- Dycom has the potential to provide a better understanding of the relationship between efforts of different companies.
  - This can in turn lead to insights into how to improve productivity.

Future Work

- Other base learners than regression trees should be investigated in future research.
- Experiments should be performed with additional data sets.
- Better strategies to split CC data should be investigated.
- More in depth understanding of why Dycom sometimes did not perform so well as a WC model.

Thank you!

Dycom



Dycom vs Mean

Test Set	Mean vs. Dycom	MAE	SA	MAEL
C1	Mean-P2	542.9989	-747.3199	3.0267
	Dycom-RT-P2	36.6201	42.8563	0.65105
	P-value	2.00E-06		2.00E-06
C2	Mean-P2	590.8896	-538.0089	4.6718
	Dycom-RT-P2	4.2004	95.1092	0.553
	P-value	2.00E-06		2.00E-06
C3	Mean-P2	2170.0519	13.5880	1.7619
	Dycom-RT-P2	734.1386	70.7664	0.23605
	P-value	4.10E-05		2.00E-06
C4	Mean-P2	392.4489	-71.6249	2.1131
	Dycom-RT-P2	110.96	51.4752	0.7837
	P-value	2.00E-06		2.00E-06
C5	Mean-P2	465.0405	-19.2921	1.6425
	Dycom-RT-P2	321.19815	17.6063	0.9441
	P-value	1.25E-01		4.00E-06
C6	Mean-P2	490.7463	-545.7188	2.0440
	Dycom-RT-P2	36.052	52.5632	0.41305
	P-value	2.00E-06		2.00E-06
C7	Mean-P2	802.1830	-13.2760	1.8029
	Dycom-RT-P2	23.234	96.7191	0.1244
	P-value	2.00E-06		2.00E-06
C8	Mean-P2	421.2622	-194.8180	1.5218
	Dycom-RT-P2	128.99745	9.7218	0.5614
	P-value	4.00E-06		2.00E-06

vs Median

Test Set	Median vs. Dycom	MAE	SA	MAEL
C1	Median-P2	57.9686	9.5432	1.2305
	Dycom-RT-P2	36.6201	42.8563	0.65105
	P-value	5.72E-01		1.00E-05
C2	Median-P2	89.5760	-4.2988	2.8610
	Dycom-RT-P2	4.2004	95.1092	0.553
	P-value	2.00E-06		2.00E-06
C3	Median-P2	2523.5714	-0.4892	3.4836
	Dycom-RT-P2	734.1386	70.7664	0.23605
	P-value	2.80E-05		2.00E-06
C4	Median-P2	185.2500	18.9869	2.0480
	Dycom-RT-P2	110.96	51.4752	0.7837
	P-value	4.00E-06		2.00E-06
C5	Median-P2	325.4167	16.5242	1.2256
	Dycom-RT-P2	321.19815	17.6063	0.9441
	P-value	5.30E-01		4.90E-04
C6	Median-P2	29.6250	61.0197	0.4109
	Dycom-RT-P2	36.052	52.5632	0.41305
	P-value	3.82E-01		7.97E-01
C7	Median-P2	605.6667	14.4740	1.2335
	Dycom-RT-P2	23.234	96.7191	0.1244
	P-value	2.00E-06		2.00E-06
C8	Median-P2	94.6667	33.7481	0.7119
	Dycom-RT-P2	128.99745	9.7218	0.5614
	P-value	2.11E-03		3.85E-03

vs WC model

	Approach	MAE	SA	MAEL
C1	WC-RT	22.8779	43.8107	0.7362
	Dycom-RT	36.6201	10.0591	0.6511
	P-value	4.99E-03		4.41E-01
C2	WC-RT	5.2373	26.1423	0.6399
	Dycom-RT	4.2004	40.7643	0.5530
	P-value	1.71E-03		2.77E-03
C3	WC-RT	627.7143	28.5761	0.2615
	Dycom-RT	734.1386	16.4667	0.2361
	P-value	8.59E-02		7.04E-01
C4	WC-RT	64.7500	73.4631	0.4000
	Dycom-RT	110.9600	54.5246	0.7837
	P-value	3.32E-04		1.70E-06
C5	WC-RT	374.0833	6.8672	0.9879
	Dycom-RT	321.1982	20.0337	0.9441
	P-value	7.34E-01		6.00E-01
C6	WC-RT	44.7500	-7.1856	0.5736
	Dycom-RT	36.0520	13.6479	0.4131
	P-value	5.71E-02		7.73E-03
C7	WC-RT	223.7953	68.5268	0.3517
	Dycom-RT	23.2340	96.7325	0.1244
	P-value	2.40E-06		4.20E-04
C8	WC-RT	76.6667	27.7487	0.4242
	Dycom-RT	128.9975	-21.5683	0.5614
	P-value	2.16E-05		4.53E-04

vs NN-Filtering

	Approach	MAE	SA	MAEL
C1	NN-Filtering-RT	22.0922	45.7405	0.9009
	Dycom-RT	36.6201	10.0591	0.6511
	P-value	4.99E-03		4.41E-01
C2	NN-Filtering-RT	15.8203	-123.1032	1.0056
	Dycom-RT	4.2004	40.7643	0.5530
	P-value	1.71E-03		2.77E-03
C3	NN-Filtering-RT	670.8572	23.6671	0.2864
	Dycom-RT	734.1386	16.4667	0.2361
	P-value	8.59E-02		7.04E-01
C4	NN-Filtering-RT	125.8413	48.4257	0.7564
	Dycom-RT	110.9600	54.5246	0.7837
	P-value	3.32E-04		1.70E-06
C5	NN-Filtering-RT	400.0417	0.4046	1.1105
	Dycom-RT	321.1982	20.0337	0.9441
	P-value	7.34E-01		6.00E-01
C6	NN-Filtering-RT	35.8375	14.1617	0.5393
	Dycom-RT	36.0520	13.6479	0.4131
	P-value	5.71E-02		7.73E-03
C7	NN-Filtering-RT	226.3800	68.1633	0.4112
	Dycom-RT	23.2340	96.7325	0.1244
	P-value	2.40E-06		4.20E-04
C8	NN-Filtering-RT	73.0556	31.1518	0.4309
	Dycom-RT	128.9975	-21.5683	0.5614
	P-value	2.16E-05		4.53E-04