UNIVERSITY OF TECHNOLOGY SYDNEY, AUSTRALIA.

DOCTORAL THESIS

Data Analytics and the Novice Programmer

Author: Alireza AHADI

Supervisor: Associate Professor Raymond LISTER (principal) and Dr Julia PRIOR (co-supervisor)

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in the Human Centred Technology Design School of Software

January 22, 2018
Declaration of Authorship

I, Alireza AHADI, declare that this thesis titled, “Data Analytics and the Novice Programmer” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.

- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.

- Where I have consulted the published work of others, this is always clearly attributed.

- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.

- I have acknowledged all main sources of help.

- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:
“When I was in college, my graduation thesis was called ‘Female Directors.’ I interviewed all of the important female directors from Mexico. There were four. That was it.”

Patricia Riggen
Abstract

Faculty of Engineering and Information Technology
School of Software
Doctor of Philosophy

Data Analytics and the Novice Programmer
by Alireza AHADI

The aptitude of students for learning how to program (henceforth Programming learn-ability) has always been of interest to the computer science education researcher. This issue of aptitude has been attacked by many researchers and as a result, different algorithms have been developed to quantify aptitude using different methods. Advances in online MOOC systems, automated grading systems, and programming environments with the capability of capturing data about how the novice programmer’s behavior has resulted in a new stream of studying novice programmer, with a focus on data at large scale. This dissertation applies contemporary machine learning based analysis methods on such “big” data to investigate novice programmers, with a focus on novices at the early stages of their first semester. Throughout the thesis, I will demonstrate how machine learning techniques can be used to detect novices in need of assistance in the early stages of the semester. Based on the results presented in this dissertation, a new algorithm to profile novices coding aptitude is proposed and its’ performance is investigated. My dissertation expands the range of exploration by considering the element of context. I argue that the differential patterns recognized among different population of novices is very sensitive to variations in data, context and language; hence validating the necessity of context-independent methods of analyzing the data.
CONFIRMATION OF ETHICS CLEARANCE

Human Negligible Low Risk Ethical clearance was granted for this PhD project by the University of Technology Sydney Human Research Ethics Committee under approval numbers ETH16-0340 (see Appendix D).
Acknowledgements

I would like to express my special appreciation and thanks to my advisor associate professor Raymond Lister, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on my research have been priceless.

A special thanks to my family. Words cannot express how grateful I am to my father, mother and my beloved sister for all of the sacrifices that you’ve made on my behalf.

Alireza Ahadi
KEYWORDS

Novice Programmer
Human Factors
Measurement
Computer Science Education
Data Mining
Online assessment
MOOC
Databases
SQL Queries
Learning Edge Momentum
Bimodal Grade Distribution
Machine Learning
Programming Source Code Snapshot
Pattern Recognition
Classification
Supervised Machine Learning
Contents

Declaration of Authorship iii

Abstract vii

Acknowledgements xi

1 Introduction 1

1.1 Motivation .................................. 1

1.2 Motivation .................................. 2

1.3 Research Questions ........................ 2

1.4 Research Design .......................... 2

1.5 Thesis Structure .......................... 3

Chapter 2: Background ................... 3

Chapter 3: Results Overview ............. 3

Chapter 4: Method ........................ 3

Chapter 5 to Chapter 12 .................... 3

Chapter 13: Discussion and Conclusion .... 3

1.6 Significance and Contribution ........ 3

1.7 Conclusion ............................. 4

2 Background 5

2.1 Introduction ............................. 5

2.2 Who is a novice programmer? .......... 5

2.3 Data analysis ........................... 7

2.3.1 Static success factors ................. 8

2.3.2 Success factors and contradictory reports .................. 10

2.3.3 Dynamically accumulating data ........ 11

2.4 Dynamic Programming Source Code Snapshot Data Collection Tools ........ 14

2.4.1 Blackbox ........................... 14

2.4.2 CloudCoder .......................... 14

2.4.3 CodeWorkout ........................ 15

2.4.4 JS-Parsons .......................... 15

2.4.5 PCRS [118] ........................... 16

2.4.6 Problets ........................... 16

2.4.7 TestMyCode .......................... 16
2.4.8 URI Online Judge ........................................ 17
2.4.9 UUhistle .................................................. 17
2.4.10 Web-CAT .................................................. 17
2.5 Publicly Available Programming Source Code Snapshot Datasets 18
  2.5.1 Blackbox .................................................. 18
  2.5.2 Code Hunt ............................................... 18
  2.5.3 Code.Org ................................................ 19
2.6 Approaches for Data Analysis ................................ 19
  Educational data mining and learning analytics ........ 19
  Machine learning and data mining approaches ........ 20
2.7 Machine Learning in CSed ................................ 21
2.8 Assessment ................................................... 23
2.9 Studies of the Novice Programming Process ............ 24
2.10 The analysis of the novice programmer errors ........ 24

3 Results Overview ............................................. 29
3.1 Results Overview ........................................... 29
3.2 The Thesis About The Thesis .............................. 29
  3.2.1 Research Questions 1 ................................. 30
   Paper 1. Chapter 5: Geek genes, prior knowledge, stumbling points and learning edge momentum: parts of the one elephant? ........ 30
   Paper 2. Chapter 6: Exploring machine learning methods to automatically identify students in need of assistance .................. 30
   Paper 3. Chapter 7: A Quantitative Study of the Relative Difficulty for Novices of Writing Seven Different Types of SQL Queries ........ 30
   Paper 4. Chapter 8: Students’ Semantic Mistakes in Writing Seven Different Types of SQL Queries 31
   Paper 5. Chapter 9: Students’ Syntactic Mistakes in Writing Seven Different Types of SQL Queries and its Application to Predicting Students’ Success ................................. 31
  3.2.2 Research Question 2 ................................. 31
   Paper 7. Chapter 11: On the Number of Attempts Students Made on Some Online Programming Exercises During Semester and their Subsequent Performance on Final Exam Questions 32
<table>
<thead>
<tr>
<th>Paper</th>
<th>Chapter</th>
<th>Title</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>12</td>
<td>A Contingency Table Derived Methodology for Analyzing Course Data</td>
<td>32</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Method</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>4.1</td>
<td>Introduction</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>4.2</td>
<td>Background</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>4.2.1</td>
<td>Programming source code snapshots, collected at the University of Helsinki</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>4.2.2</td>
<td>Database source code snapshots, collected at the University of Technology Sydney</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Geek genes, prior knowledge, stumbling points and learning edge momentum: parts of the one elephant?</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>5.1</td>
<td>Introduction</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>5.1.1</td>
<td>Statement of Contribution of Co-Authors</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>5.2</td>
<td>PDF of the Published Paper</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>5.3</td>
<td>Discussion</td>
<td>53</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Exploring machine learning methods to automatically identify students in need of assistance</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>6.1</td>
<td>Introduction</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>6.1.1</td>
<td>Statement of Contribution of Co-Authors</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>6.2</td>
<td>PDF of the Published Paper</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>6.3</td>
<td>Discussion</td>
<td>69</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>A Quantitative Study of the Relative Difficulty for Novices of Writing Seven Different Types of SQL Queries</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>7.1</td>
<td>Introduction</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>7.1.1</td>
<td>Statement of Contribution of Co-Authors</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>7.2</td>
<td>PDF of the Published Paper</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>7.3</td>
<td>Discussion</td>
<td>81</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Students’ Semantic Mistakes in Writing Seven Different Types of SQL Queries</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>8.1</td>
<td>Introduction</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>8.1.1</td>
<td>Statement of Contribution of Co-Authors</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>8.2</td>
<td>PDF of the Published Paper</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>8.3</td>
<td>Discussion</td>
<td>93</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Students’ Syntactic Mistakes in Writing Seven Different Types of SQL Queries and its Application to Predicting Students’ Success</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>9.1</td>
<td>Introduction</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>9.1.1</td>
<td>Statement of Contribution of Co-Authors</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>9.2</td>
<td>PDF of the Published Paper</td>
<td>98</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>9.3</td>
<td>Discussion</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Performance and Consistency in Learning to Program</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>10.1</td>
<td>Introduction</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>10.1.1</td>
<td>Statement of Contribution of Co-Authors</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>10.2</td>
<td>PDF of the Published Paper</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>10.3</td>
<td>Discussion</td>
<td>117</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>On the Number of Attempts Students Made on Some Online Programming Exercises During Semester and their Subsequent Performance on Final Exam Questions</td>
<td>119</td>
<td></td>
</tr>
<tr>
<td>11.1</td>
<td>Introduction</td>
<td>119</td>
<td></td>
</tr>
<tr>
<td>11.1.1</td>
<td>Statement of Contribution of Co-Authors</td>
<td>119</td>
<td></td>
</tr>
<tr>
<td>11.2</td>
<td>PDF of the Published Paper</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>11.3</td>
<td>Discussion</td>
<td>129</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>A Contingency Table Derived Method for Analyzing Course Data</td>
<td>131</td>
<td></td>
</tr>
<tr>
<td>12.1</td>
<td>Introduction</td>
<td>131</td>
<td></td>
</tr>
<tr>
<td>12.1.1</td>
<td>Statement of Contribution of Co-Authors</td>
<td>131</td>
<td></td>
</tr>
<tr>
<td>12.2</td>
<td>PDF of the Submitted Paper</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Discussion and Conclusion</td>
<td>155</td>
<td></td>
</tr>
<tr>
<td>13.1</td>
<td>Introduction</td>
<td>155</td>
<td></td>
</tr>
<tr>
<td>13.2</td>
<td>Overview of Research</td>
<td>155</td>
<td></td>
</tr>
<tr>
<td>13.3</td>
<td>Research Questions</td>
<td>155</td>
<td></td>
</tr>
<tr>
<td>13.4</td>
<td>Research Outcome</td>
<td>156</td>
<td></td>
</tr>
<tr>
<td>13.5</td>
<td>Research highlights</td>
<td>156</td>
<td></td>
</tr>
<tr>
<td>13.6</td>
<td>Research Significance</td>
<td>157</td>
<td></td>
</tr>
<tr>
<td>13.7</td>
<td>Research Findings</td>
<td>157</td>
<td></td>
</tr>
<tr>
<td>13.7.1</td>
<td>Data</td>
<td>157</td>
<td></td>
</tr>
<tr>
<td>13.7.2</td>
<td>Method</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td>13.7.3</td>
<td>Context</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td>13.8</td>
<td>Discussion</td>
<td>159</td>
<td></td>
</tr>
<tr>
<td>13.8.1</td>
<td>Data</td>
<td>159</td>
<td></td>
</tr>
<tr>
<td>13.8.2</td>
<td>Method</td>
<td>159</td>
<td></td>
</tr>
<tr>
<td>13.8.3</td>
<td>Context</td>
<td>160</td>
<td></td>
</tr>
<tr>
<td>13.9</td>
<td>Limitations</td>
<td>161</td>
<td></td>
</tr>
<tr>
<td>13.9.1</td>
<td>Data</td>
<td>161</td>
<td></td>
</tr>
<tr>
<td>13.9.2</td>
<td>Context</td>
<td>161</td>
<td></td>
</tr>
<tr>
<td>13.10</td>
<td>Recommendations</td>
<td>161</td>
<td></td>
</tr>
<tr>
<td>13.10.1</td>
<td>Data</td>
<td>162</td>
<td></td>
</tr>
<tr>
<td>13.10.1.1</td>
<td>Static data</td>
<td>162</td>
<td></td>
</tr>
<tr>
<td>13.10.1.2</td>
<td>Dynamic data</td>
<td>162</td>
<td></td>
</tr>
</tbody>
</table>
List of Figures

4.1 An abstract view of the logical services where data instrumentation and collection are typically performed. Note that some data collection tools encompass multiple logical services (Adapted from Ihantola et al., 2015). .................. 34
4.2 Data can be collected at different levels of granularity, which implies different collection frequencies and associated data set sizes (Adapted from Ihantola et al., 2015) .......... 35
4.3 Student’s home screen at AsseSQL. Students may attempt questions in any order, as many times as they wish. ........ 38
4.4 Screen for an individual question. The question and the simple output to clarify the question are presented for each question in the test. ............................. 39
4.5 Feedback screen for an individual question with the output of model answer as well as the student’s answer are presented in this page. .......................... 39
4.6 Back to the student’s home screen: students may return to this at anytime. ............................. 40
4.7 A sample of the ERD used in the test. .......................... 41
## List of Tables

3.1 Research Questions, the thesis about the thesis, and the corresponding chapters ........................................ 29

5.1 Authors’ Area of Contribution for The Paper Corresponding to Chapter 5 ................................................. 44

6.1 Authors’ Area of Contribution for The Paper Corresponding to Chapter 6 .................................................. 56

7.1 Authors’ Area of Contribution for The Paper Corresponding to Chapter 7 .................................................. 72

8.1 Authors’ Area of Contribution for The Paper Corresponding to Chapter 8 .................................................. 84

9.1 Authors’ Area of Contribution for The Paper Corresponding to Chapter 9 .................................................. 96

10.1 Authors’ Area of Contribution for The Paper Corresponding to Chapter 10 ............................................... 108

11.1 Authors’ Area of Contribution for The Paper Corresponding to Chapter 11 ............................................... 120

12.1 Authors’ Area of Contribution for The Paper Corresponding to Chapter 12 ............................................... 132

G.1 List of questions and their corresponding covered topic ................................................................. 220

G.2 Proposed Answers for the questions presented in G.1 .............................................................. 221
List of Abbreviations

ACC  Accuracy
CMS  Course Management System
CTP  Computational Thinking Patterns
DM  Data Mining
DT  Decision Tree
EDM  Educational Data Mining
EQ  Error Quotient
FN  False Negative
FP  False Positive
FDR  False Discovery Ratio
FNR  False Negative Ratio
FPR  False Positive Ratio
IDE  Integrated Development Environment
JAR  Java Archive File
LMS  Learning Management System
LN  Logical Necessity
LS  Logical Sufficiency
LSI  Kolb’s Learning Style Inventory
ML  Machine Learning
MSLQ  Motivated Strategies Learning Questionnaire
NN  Neural Network
NPSM  Normalized Programming State Model
PCA  Principal Component Analysis
RF  Random Forest
RED  Repeated Error Density
RSE  Rosenberg Self-Esteem
SEN  Sensitivity
SPC  Specificity
SQL  Structured Query Language
SVM  Support Vector Machine
TN  True Negative
TP  True Positive
WATWIN  WATson & GodWIN
Web-CAT  Web-based Center for Automated Testing
List of Publications by Candidate

Below are the publications first authored by the Candidate which contribute to this thesis by publication.


DOI: https://doi.org/10.1145/3123814 (See Chapter 12)
This thesis is presented in the format of scholarly papers published during the period of my candidature, according to UTS regulations set out in the website http://uts.edu.au/current-students/dab/uts-graduate-research-school. The papers included in this thesis by publication form a research narrative which is summarized in Chapter 5. Each paper then forms a separate chapter of this thesis, inserted in its published format.
To all novice programmers…