Contents

Abstract v
Resumen vii
Resum ix
Acknowledgements xi
Contents xiii
List of Figures xvii
List of Tables xix

I Introduction 1

1 Introduction 3
  1.1 Motivation ...................................... 3
  1.2 Research Questions and Objectives .................. 6
  1.3 Research Methodology ................................ 7
  1.4 Thesis Outline .................................. 9

II Background 15

2 Inductive Programming 17
  2.1 Introduction .................................... 17
  2.2 Learning programs ................................ 21
  2.3 Inductive Bias .................................... 23
  2.4 General Purpose Inductive Programming Systems ..... 25

3 Data Science Automation 31
  3.1 Introduction .................................... 31
  3.2 Data Science Trajectories ......................... 33
  3.3 Data Wrangling ................................... 35
  3.4 Data Wrangling Automation ........................ 38
## Contents

### III BK-ADAPT: Automating Data Format Standardisation  
43

4  **Domain-specific Induction**  
4.1  Introduction  
4.2  Problem Definition  
4.3  Experiments  
4.4  Conclusions  

5  **Dynamic Background Knowledge**  
5.1  Introduction  
5.2  Upgraded Approach  
5.3  Method  
5.4  Experiments  
5.5  Conclusions  

### IV AUTOMAT[R]IX: Automating Matrix Transformations  
73

6  **Learning Simple Matrix Pipelines**  
6.1  Introduction  
6.2  Problem Definition  
6.3  Method  
6.4  Experiments  
6.5  Conclusions  

### V Conclusions  
95

7  **Conclusions and Future Work**  
7.1  Conclusions  
7.2  Future Work  

**Bibliography**  
103

**Appendices**  
117

A  **Data Collection**  
A.1  Survey to collect personal data  

B  **Data Wrangling: Competences and Skills**  
B.1  List of Data Wrangling Competences  
B.2  List of Data Wrangling Hard Skills  
B.3  List of Data Wrangling Soft Skills  

C  **Logging Data Scientists**  
125
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>BK-ADAPT: Supplementary Material</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D.1 Background Knowledge: List of Functions</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>D.2 Data: List of Meta-features</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>D.3 Experiments: Extended Results</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>D.4 Tool: System Overview</td>
<td>149</td>
</tr>
<tr>
<td>E</td>
<td>AUTOMAT[R]IX: Supplementary Material</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E.1 Background Knowledge: List of Functions</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>E.2 Prior Probabilities: List of Top Functions in GitHub</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>E.3 Text Hints: Frequent Terms from Function’s Documentation</td>
<td>156</td>
</tr>
<tr>
<td></td>
<td>E.4 Data: List of Examples</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>E.5 Pipeline of events for a simple example</td>
<td>161</td>
</tr>
</tbody>
</table>