"How not to do it": Anti-patterns for Data Science in Software Engineering

Tim Menzies
Computer Science
North Carolina State University, USA
tim.menzies@gmail.com

ABSTRACT

Many books and papers describe how to do data science. While those texts are useful, it can also be important to reflect on anti-patterns; i.e. common classes of errors seen when large communities of researchers and commercial software engineers use, and misuse data mining tools. This technical briefing will present those errors and show how to avoid them.

CCS Concepts

• Software Engineering • Decisions support systems →Data analytics

Keywords

Data Science; Software Analytics

1. INTRODUCTION

To say the least, a lot of industrial workers and researchers in SE use data mining. Data mining is widely used in the software industry. Graduates from data mining master degrees are in high demand. Any recent conference on SE contains dozens of papers on data science. In the SE journals of (TSE, the ASE journal, EMSE), data mining papers are the (first, first, and fourth) most cited papers.

While there are many Data Mining 101 tutorials for software engineering, there is a dearth of material devoted to advanced users of these tools. The main selling point of this tech briefing is the concept of "Anti-Pattern". As we mature as a field, it is becoming realized by the broader community that we need only share our successes- but also our failures. Hence, it is important to report "how not to do it".

2. STRUCTURE

This technical briefing will present those errors and show how to avoid them. The errors will be presented at several levels:

- **For researchers:** Prior to the study: how not to design a study, how to miss important result. Also, after the study: how to make unfounded conclusions, how to get a paper rejected from EMSE journal, ICSE, ASE. This talk will also talk about research directions in this field. This will include speculations on how to use these data mining anti-patterns as a novel research direction for data science.
- **For industrial practitioners:** Ways to design your data mining team in order to ensure the failure of your project.

3. WHO SHOULD ATTEND

This talk would be suitable for industrial practitioners who do not want to fail in their data mining projects.

This talk would also be suitable for academic researchers interested in either

- new research directions in software analytics;
- increasing the odds that their publications of current topics in software analytics might get accepted.

4. ABOUT THE PRESENTER

Tim Menzies(Ph.D.,UNSW,1995) is a full Professor in CS at North Carolina State University where he teaches software engineering and automated software engineering. His research relates to synergies between human and artificial intelligence, with particular application to data mining for software engineering. He is the author of over 230 referred publications. In his career, he has been a lead researcher on projects for NSF, NIJ, DoD, NASA, USDA, as well as joint research work with private companies.

Working closely with the software engineering community, Dr. Menzies has recently co-edited and released three summaries of the state of the art in data science for software engineering, published by Morgan Kaufmann. A third book ("Perspectives on Data Science for Software Engineering") is being developed with a particular focus on industrial practitioners.

Prof. Menzies is the co-founder of the PROMISE conference series devoted to reproducible experiments in software engineering (http://openscience.us/repo). He is an associate editor of IEEE Transactions on Software Engineering, Empirical Software Engineering, Information Systems and Technology, the Automated Software Engineering Journal and the Software Quality Journal. In 2015, he served as co-chair for the ICSE'15 NIER track. In 2016, he serves as co-general chair of ICMSE'16. In 2017 he will serve as PC chair of SSBSE'17. For more, see his vita (http://goo.gl/8eNhYM) or his list of publications https://goo.gl/qNQAIq) or his home page http://menzies.us.