Abstract

Write your abstract here...

Sammendrag

Skriv sammendrag her...

Preface

Write your preface here...

Acknowledgements (optional)

Table of Contents

List of Tables

List of Figures

Abbreviations

Symbol = definition

Chapter 1

Introduction

1.1 Background, Motivation and Problem Outline

1.2 Research Context

The research was conducted as my Master's thesis at the department of Computer and Information Science at the Norwegian University of Science and Technology. The research task was formulated by Odd Erik Gundersen, my supervisor, and is a continuation of previous work by **?** presented at 3DOR2015[1].

1.3 Hypothesis, Objectives and Research Questions

Underlying this thesis is the hypothesis that; *the documentation provided in experimental publications at AI conferences is not good enough to consider the experiments reproducible.*

Objective 1 *Evaluate the reproducibility of accepted papers to AI conferences.*

RQ1 What is the state of reproducibility at AI conferences?

Objective 2 *Recommend practices that could be adopted to aid the reproducibility of conference papers.*

RQ2 What is generally missing from AI papers to support reproducibility?

RQ3 What can ease the documentation of missing information from conference papers?

[1] `http://vc.ee.duth.gr/3DOR2015/`

1.4 Research Approach

1.5 Research Contributions

”We examine the common practices and challenges we see in recent OSN research, from which we propose a set of recommendations for the benefit of OSN researchers in all disciplines.”

C1: A survey of experimental research papers from AI conferences.

C2: An indication of the state of reproducibility at AI conferences.

C3: An approach to measure the reproducibility of AI conference papers.

1.6 Thesis Structure

Chapter 2

Literature Review

2.1 Reproducibility Terminology

Fehr 2016 Best Practices for Replicability, Reproducibility and Reusability of Computer-Based Experiments Exemplified by Model Reduction Software: https://arxiv.org/abs/1607.01191

Peng (2006): https://doi.org/10.1093/aje/kwj093 ”Scientific evidence is strengthened when important findings are replicated by multiple independent investigators using independent data, analytical methods, laboratories, and instruments.” ”An attainable minimum standard is reproducibility, where independent investigators subject the original data to their own analyses and interpretations. Reproducibility calls for data sets and software to be made available for 1) verifying published findings, 2) conducting alternative analyses of the same data, 3) eliminating uninformed criticisms that do not stand up to existing data, and 4) expediting the interchange of ideas among investigators.”

Claebout (1992): http://dx.doi.org/10.1190/1.1822162

- Thompson and Burnett (2012): https://nationalethicscenter.org/content/article/175 - Stodden - Goodman - Drummond - Recomputation manifesto? (Gent 2013)

2.2 Criteria for reproducible research

Fehr 2016 Best Practices for Replicability, Reproducibility and Reusability of Computer-Based Experiments Exemplified by Model Reduction Software: https://arxiv.org/abs/1607.01191

2.2.1 Focus on recreating figures

Peng (2006): https://doi.org/10.1093/aje/kwj093 ”Reproducibility is a minimum step that can be taken in any study. In the context of epidemiology, a study is reproducible when it satisfies the criteria in table 1, adapted from the paper by Schwab et al. (7) and others. We illustrate reproducibility requirements separately for each of the following research components: data, methods, documentation, and distribution.” ”Table 1: Research component

Requirement Data Analytical data set is available. Methods Computer code underlying figures, tables, and other principal results is made available in a human-readable form. In addition, the software environment necessary to execute that code is available. Documentation Adequate documentation of the computer code, software environment, and analytical data set is available to enable others to repeat the analyses and to conduct other similar ones. Distribution Standard methods of distribution are used for others to access the software, data, and documentation."

Schwab (2000) Making scientific computations reproducible (ReDoc): http://dx.doi.org/10.1109/5992.88
"Conditionally reproducible result files that require proprietary data, licensed software, or more than 10 minutes for recomputation. The author nevertheless supplies a complete set of source files; this ensures that readers can reproduce the results if they possess the necessary resources"

2.2.2 Computational Experiments

Laine (2007): DOI: 10.7326/0003-4819-146-6-200703200-00154 "Reproducibility involves methods to ensure that independent scientists can reproduce published results by using the same procedures and data as the original investigators. It also requires that the primary investigators share their data and methodological details. These include, at a minimum, the original protocol, the dataset used for the analysis, and the computer code used to produce the results."

Hutton (2015): Towards reproducibility in online social network research https://doi.org/10.1109/TETC.2
""

2.3 Best Practices for Reproducibility

- Stodden Miguez 2014: http://doi.org/10.5334/jors.ay - Stodden, Donaho++ 2009: https://doi.org/10.1109/I
- Sandve et al (Ten simple rules) `http://dx.doi.org/10.1371/journal.pcbi.1003285` Davison (2012): http://rrcns.readthedocs.io/en/cns2012/index.html

- (LeVeque? Top ten reasons to not share code) - Yilmaz 2012 - Collberg Proebsting 2016 - Claerbout 1992 - Casadevall 2011 - Buckheit 1995 - Brown 2012

2.4 Reproducibility in Practice

The Recomputation Manifesto (Gent 2013) Six theses; 1. Computational experiments should be recomputable for all time. 2. Recomputation of recomputable experiments should be very easy. 3. It should be easier to make experiments recomputable than not to. 4. Tools and repositories can help recomputation become standard. 5. The only way to ensure recomputability is to provide virtual machines. 6. Runtime performance is a secondary issue.

- Hunold (Trff) - Kluyver (Jupyer notebooks) - Leitner (acrv picking benchmark) - See spec. project

BioCatalogue: a catalogue of services for biological evaluation and analysis, displays availability of endpoint to the service as well. (https://nationalethicscenter.org/content/article/175)

2.5 Observations

Chapter 3

Research Method

3.1 Litterature Survey Design

Advantages of doing a survey - Can be replicated on similar documents or on original documents provided the method is shared and documents are accessible - Can produce a lot of data at a low cost, in a relatively short time compared to attempting full replications of experiments - Allows a larger sample population due to the shorter time necessary to evaluate a paper

Disadvantages - The depth is restricted, does not provide detail on the research topic - Focuses on what can be counted and measured, other aspects may be overlooked

3.1.1 Data requirements

Want to investigate the reproducibility of experiments published in papers at AI conferences.

/* Explain variables and relate to best practices!!! */

Directly topic related : Is source code or data open for the experiment and method? Is the method documented? Is the experiment documented? etc.

Indirectly topic related Research transparency (hypothesis, predictions...) Author affiliation (uni/industry/both) Novel research? Conference view on supplementary material Theoretical / Experimental research

Possible analysis patterns

1. reproducibility related to author affiliation
2. reproducibility related to conference view on supplementary material?
3. reproducibility related to publishing year (improvement over time?)

3.1.2 Data generation method

Documents, conference papers. (Ch. 16) - Existing conference papers, published openly in the proceedings of the conferences. Physical copies can be ordered, but all accepted papers are available on-line.

Advantages: - easy to obtain, accessible and are obtained unobtrusively - allows later longitudinal studies - other researchers can check and scrutinize the research based on original material

Disadvantages: -

Sampling frame: accepted papers at IJCAI-13, -16 and AAAI-14 and -16 (can be seen in repo files for sample generation) Sampling technique: probabilistic random sampling of each conference separately. ”Probability sampling, as its name suggests, means that the sample has been chosen because the researcher believes that there is a high probability that the sample of respondents (or events) chosen are representative of the overall population being studied. That is, they form a representative cross-section of the overall population.” Oates p.96 Discuss representativeness of sample method.

Sample size: 100 for each conference, restricts the necessary time to conduct the survey while still providing informative accuracy ranges when considering previous research (cite?)

Conference	Population Size	Sample Size	Confidence Interval
AAAI 2014	398	100	8.49
AAAI 2016	548	100	8.87
IJCAI 2013	413	100	8.54
IJCAI 2016	551	100	8.87
Combined	1910	400	4.36

Table 3.1: Confidence intervals of survey sample populations given a 50/50 yes/no split with confidence level of 95%. (`https://www.surveysystem.com/sscalc.htm`)

3.2 Evaluation Procedure

- Step by step ’instructions’ - Sampling documentation - Evaluation documentation - Example evaluations (variable X: ”Exhibit A” covers, ”Exhibit B” is not enough)

3.3 Limitations of the Survey

- Evaluation bias (modification of variables) - Sample inconsistency for IJCAI-13 (50 papers) - Not an actual attempt at reproducing experiments, researcher’s view that discussion of a variable is missing?

Chapter 4

Results and Analysis

- All variables contain Nominal (or categorical) data, which means the frequency is what is of interest.
    - Bar charts common for visualizing frequency - Pie charts for showing proportions
    - Frequency of each variable
    - Relationships between variables

Chapter 5

Discussion

Anonymous publication of source code and data along with papers (blind review)

Hunold and Trff 2013: Add a description of how to reproduce the findings in a publication

ACM TOMS: Independent replication review http://toms.acm.org/replicated-computational-results.cfm

Chapter 6

Conclusion

Appendix

Write your appendix here...

Appendix A

Population selection code

Appendix B

Survey data (and population samples)

Appendix C

Analysis code