NTNU
Norwegian University of
Science and Technology

Research method in AI

Reproducibility of results

Sigbjørn Kjensmo

Master of Science in Computer Science
Submission date: June 2017
Supervisor: Odd Erik Gundersen, IDI

Norwegian University of Science and Technology
Department of Computer Science

Abstract

Reproducibility of published computational research has seen increased interest the last twenty years. Regardless of academic field and the impact-factor of journals, studies of reproducibility of computational research have found low rates of reproducibility. Common issues relate to the availability of source code and data, even when original authors attempt to reproduce their own published research.

In this thesis, we investigate the state of reproducibility in artificial intelligence research. The objective is not to reproduce experiments, but to investigate and quantify the state of reproducibility in artificial intelligence research. Two hypotheses were investigated: 1) Documentation of AI research is not good enough to reproduce results, and 2) Documentation practices have improved in recent years. 400 research papers from two instalments of two top AI conference series, IJCAI and AAAI, have been surveyed to investigate the hypotheses. The results of our survey support the first hypothesis, but not the second. While common usage of public datasets is widespread, sharing of code is lagging behind. Facilitating sharing of source code, and data without disrupting the peer review process are necessary to improve the situation.

The contribution efforts of the research in this thesis are: (i) a survey design for evaluating documentation of published papers, (ii) an evaluation of two leading AI conference series, and (iii) suggested incentives to facilitate the reproducibility of AI research.

Sammendrag

Reproduserbarhet av publisert forskning har sett økende interesse og diskusjon de siste årene. Studier som undersøker reproduserbarhet har gjentatte ganger vist lav grad av reproduserbarhet innen flere akademiske felt, uavhengig av innflytelsesfaktoren til journalen. Ofte diskuterte problemer relaterer til hvor tilgjengelig kildekode og data benyttet er, selv for de opprinnelige forskerne.

I denne oppgaven ser vi på status for reproduserbarhet for publisert forskning innen kunstig intelligens. Målet er å undersøke og kvantifisere reproduserbarheten innen kunstig intelligens. To hypoteser ble gransket: 1) Publiseringer innen kunstig intelligens dokumenterer ikke nok til å tillate reprodusering, og 2) Dokumentasjonspraksis har bedret seg de siste årene. For å forske på hypotesene har en undersøkelse av 400 publiserte artikler fra to ledende konferanseserier innen kunstig intelligens, IJCAI og AAAI, blitt utført. To utgaver av hver konferanseserie har blitt undersøkt. Resultatene støtter kun den første hypotesen. Bruk av åpne datasett er utbredt, men deling av kildekode ser ut til å henge etter. For å forbedre reproduserbarheten til konferansene, vil intensiver til å dele kildekode, programmer, og data uten å påvirke peer-review prosessen være nødvendig.

Forskningen presentert i denne oppgaven har resultert i følgende bidrag: (i) design av reproduserbarhetsundersøkelse for publiserte artikler, (ii) en evaluering av to ledende konferanseserier, og (iii) forslag til intensiver som kan øke reproduserbarheten innen kunstig intelligens.

Odd Erik Gundersen, Anh Thy Tran

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

An introduction to the research topic, the hypotheses posited, the research methodology and what contributions the thesis provides is presented in this chapter.

1.1 Background, Motivation and Problem Outline

The motivation for this thesis is increased discussions of how a large amount of published research cannot be reproduced, even in the prestigious journals and by the original authors (Aarts et al. 2016, Begley & Ellis 2012, Begley & Ioannidis 2014, Prinz et al. 2011). Data from Scopus, as presented inGoodman et al. (2016), show that the discussion spans several scientific fields, although medical sciences are most involved. Within computational science, Peng (2011) states that *"the biggest barrier to reproducible research is the lack of a deeply ingrained culture that simply requires reproducibility for all scientific claims"*. The scientific method is built upon experiments being repeated so produced results can be confirmed and hypothesis verified, *"if other researchers can't repeat an experiment and get the same result as the original researchers, then they refute the hypothesis"* (Oates 2006, p. 285). If reproduction or validation of the methods used in a published experiment cannot be done, a key element of the scientific method is missing, and the trust in the publication deteriorates.

For preclinical medical research, large scale case studies have shown that a significant portion of published experiments are difficult to reproduce (Begley & Ellis 2012, Begley & Ioannidis 2014, Prinz et al. 2011). For computer science and artificial intelligence (AI), experiments reproducing a few papers (Hunold & Träff 2013, Fokkens et al. 2013), and surveys of researchers' experience with reproduction (Hunold 2015) report similar difficulties. Collberg & Proebsting (2016) finds that some researchers are not willing to share code and data, while many of those that do provide too little to reproduce. Some proposed solutions and platforms developed to facilitate reproducibility see little adoption with low ease-of-use or considerable time investments necessary to retroactively fit an experiment to them (Gent &

Kotthoff 2014). An increasing availability of such tools is still encouraging.

Previous reproduction experiments and the work on guidelines and best-practices for reproducible research (Sandve et al. 2013, Stodden & Miguez 2014) provide insight into what is necessary for reproduction. As does solutions developed to aide reproducibility, which mainly point towards open data and open source code. Our goal is to quantify the state of reproducibility at AI conferences, and find the areas of documentation that improvements in practices would most impact reproducibility.

1.2 Research Context

The research was conducted as my Master's thesis in the spring semester of 2017, 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 a specialization project during fall of 2016. Previous work by Gundersen (2015) presented at 3DOR2015[1] has influenced the research.

1.3 Hypothesis and Research Questions

Two hypotheses underlie the research in this thesis:

HYP1: *the documentation of experiments in publications at AI conferences is not good enough to consider the experiments reproducible.*

HYP2: *the documentation practices have improved in recent years.*

The following research questions were formulated to investigate these hypotheses:

RQ1: What is the state of reproducibility at AI conferences?

RQ2: What documentation is missing from AI papers to support reproducibility?

RQ3: Which practices have seen an improvement in recent years?

RQ4: What practices are encouraging reproducible research in AI?

RQ5: What incentives could be implemented to further encourage reproducible research?

RQ6: Does the author affiliation impact documentation practices?

All six research questions attempt to answer different aspects of the first hypothesis. RQ4 specifically examines the second hypothesis. Affiliation with industry have indicated a lower rate of reproducibility in studies of other research areas

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

(Collberg & Proebsting 2016), leading to the last research question, RQ6. To examine RQ4, what is encouraging reproducible research, the results of RQ3 are be necessary. As for RQ5, the results of RQ2 are essential to focus the suggested incentives.

1.4 Research Contributions

The following list is a short summary of the contributions contained in this thesis.

C1: *A survey design for evaluating documentation of experiments in AI research papers.* The survey design includes a description of the procedure taken to evaluate experiment documentation. Its documentation provides the means for other researchers to evaluate the results, as well as evaluating other conferences to compare documentation practices, or a future longitudinal study. The evaluation procedure is presented in section 3.2.

C2: *An evaluation of the state of reproducibility at two leading AI conference series.* This contribution is linked to RQ1 through RQ4, and RQ6. The evaluation gives an idea of what documentation practices are present at the investigated conferences, as well as shining a light on what documentation is missing to guide the approach to RQ5. Refer to chapter 4 for a presentation of the results, and to section 5.1 for a revisit of the research questions.

C3: *Suggested incentives to increase reproducibility of AI research.* The incentives are aimed at the problem areas discovered through the evaluation, and thereby depend on C2 to answer RQ5. Several proponents of reproducible research are careful with recommending hard requirements, to avoid increasing peer-review time and to not exclude experiments where it might be impractical, such as when legal issues hinder sharing of data. Thus, the incentives suggested attempt to provide best-practices, or reward reproducible research. See section 5.2 for further discussion.

1.5 Research Methodology

The following section provides an overview of the methodology used when conducting the research for this thesis. The literature review giving an overview of topics relevant to the thesis is covered first. Following is the survey design, data generation and analysis.

1.5.1 Literature Review

The literature review serves as an overview of relevant work and an introduction to research on reproduction. The discussion of reproducible research spans several fields, so some sources will be from fields quite different from AI. The study attempts to focus on sources relevant to AI, but includes important related studies from fields such as biomedical research and psychology. Not all the sources have

been published in reputable journals or at conferences, such as those from arXiv[2]. These have been included due to featuring often in related work from reputable publishers. Content from web pages have been avoided when possible, since the content may change over time.

The search for literature was based on queries with combinations of the following keywords: reproducible research, replication, repeatability, reproducibility, and computational. Queries began through the ACM Digital Library[3], Computer Science bibliography[4], and IEEE Xplore Digital Library[5]. The reference list included in publications from reputable journals and conferences were also included to explore relevant research.

Literature matching the queries made were examined to determine its relevance and quality. First, First, the abstract was examined to find the aim of the presented research and establish its relevancy for the thesis. If it is deemed relevant, and the study is cited by other researchers, its methodology was considered. If the presented methodology was found reasonable, for instance by the sample size used, the full publication was studied.

1.5.2 Data Generation

A survey of conference publications was used to answer the suggested hypotheses. To answer the first hypothesis on the state of reproducibility, a requirement is to cover a large amount of the conferences examined. Thus, a survey of the publications, rather than reproduction attempts, was deemed appropriate to cover a large enough sample. A probabilistic random sampling of the accepted papers was performed to generate a representable sample from a population of conferences covering similar topics and of similar recognition.

Chapter 3 covers the survey design in more depth, describing the requirements, data recorded, sampling and evaluation procedure.

1.5.3 Data Analysis

The data resulting from the survey was originally recorded in a Google Spreadsheet. To conduct the analysis, it was exported to a csv format, and examined through Python scripts included in the Appendix. Frequencies of all data recorded is reported in chapter 4, with additional patterns of analysis related to affiliation, publishing conference, and defined terms for reproducibility at the end of the chapter. Chapter 5 discuss the results of the survey and attempt to answer the investigated research questions.

[2] `https://arxiv.org/`
[3] `http://www.acm.org/dl/`
[4] `http://dblp.org/`
[5] `http://ieeexplore.ieee.org/Xplore/home.jsp`

1.6 Thesis Structure

The chapters in this thesis are divided into four parts. First, in chapter 1 and 2 the research context is presented with the introduction of the thesis, background information, and the state of the art in related topics. Second, chapter 3 and 4 present the results of the research with the survey design and its results. Third, evaluation, conclusion and future work is presented in chapter 5 and 6. Last, the appendices contain an abbreviated sample of the data, and scripts used throughout the research.

Part I: Research Context

Chapter 1: Introduction

Chapter 2: State of the Art

Part II: Research Results

Chapter 3: Research Method

Chapter 4: Results and Analysis

Part III: Evaluation and Conclusion

Chapter 5: Evaluation

Chapter 6: Conclusion and Future Work

Part IV: Appendices

Appendix A: Sample Selection

Appendix B: Survey Data

Appendix C: Analysis Code

Chapter 2

State of the Art

The terminology surrounding reproducibility and the practice of reproducibility in computer science is reviewed in this chapter. Research documenting empirical evidence of reproducibility in machine learning and computer science is reviewed, with some notable projects from other fields. Additionally, guidelines meant to facilitate computational reproducibility is presented. The review is necessary for the survey to be in line with related work on research reproducibility.

2.1 Reproducibility

Claerbout began experimenting with electronic documents as a means to transparently publish reproducible research in 1990. The effort included publishing a CD-ROM along papers containing software, data and documentation to reproduce the entire process from raw data to figures provided in his papers (Claerbout & Karrenbach 1992). Inspired by Claerbout, Buckheit & Donoho (1995) released WaveLab[1] in the early days of the internet. WaveLab was a package containing code and data for their papers, allowing anyone to reproduce their research and inspect, modify and reuse their code and data. They also condense Claerbout's ideas in the slogan *"An article about computational science in a scientific publication is* ***not*** *the scholarship itself, it is merely* ***advertising*** *of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures"* (Buckheit & Donoho 1995). Making it clear that the process leading up to the presentation is of more importance than the paper, and that reproducible research requires code, data, and documentation of this process to be made available with the paper.

Stodden (2011) distinguishes between replication and reproduction, stating that "replication using author-provided code and data, and independent reproduction work hand-in-hand". Replication is seen as rerunning the experiment with code and

[1] `https://statweb.stanford.edu/~wavelab/`

data provided by the author, while reproduction is a broader term *"implying both replication and the regeneration of findings with at least some independence from the [original] code and/or data"*. Replication is a means to resolve differences in reproductions. She argues that transparency in methodology arises from the notion of reproducibility. The openness of data or code is not the goal, but a requirement to achieve verifiable research through reproducibility. Drummond (2009) presents a dissenting opinion on the need for reproducibility, stating that replication, as the weakest form of reproducibility, can only achieve checks for fraud. Additionally, he raises the arguments that the majority of published code would not be used based on many papers having none, or few citations, and that the increased reviewer workload is not feasible. Drummond believes increased scepticism to results of experiments would reduce the impact of misconduct. However, there is no contradiction in his belief that replication is a weak form of reproduction and Stodden's view. For Stodden, replication is there to be able to analyse why differences have occurred in later reproductions. It is not there to corroborate the idea or experimental results, which both believe require independent reproductions in different environments. In the event that the original does not differ from reproductions, publishing of the original material to allow replication can still be of great value to facilitate further research through data and code reuse (Brown 2012, Stodden & Miguez 2014). Gent introduces *recomputation* as the ability to replicate a computational experiment, and encourages computational sciences to provide experiments in copies of virtual machines to make them easily recomputable for all time (Gent 2013).

To differentiate between the sources and possible solutions to reproducibility problems, Stodden (2013) introduces *empirical* and *computational* reproducibility, adding *statistical* reproducibility in a later publication (Stodden 2014). For computational reproducibility traditional scientific publications do not include enough information for computational methods to be reproduced, requiring supplementary material in the form of data, code, and implementation details. Lacking empirical and statistical reproducibility relates to the study power and bias (Ioannidis 2005), publication bias and ineffective peer review (Francis 2012), and misapplied methodology (Simmons et al. 2011). These require different remedies than computational reproducibility, where *"issues arise from an exogenous shift in the scientific research process itself - the broad use of computation"* (Stodden 2014).

Due to the inconsistencies in the use of the terms replicability and reproducibility, Goodman et al. (2016) proposes an extension of the most widely used term, reproduciblility, to be more precise. The proposed terms are *methods* reproducibility, *results* reproducibility and *inferential* reproducibility. They are defined as follows; methods reproducibility - *"the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results"*, results reproducibility *"the production of corroborating results in a new study, having used the same experimental methods"*, and lastly inferential reproducibility - *"the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study"* (Goodman et al. 2016). The approach is similar to Stodden (2013), and both results and

methods reproducibility are essential for computational reproducibility. Additionally, inferential reproducibility can be facilitated by computational reproducibility - especially in the case of a reanalysis. Goodman highlights that *"reporting of all relevant aspects of scientific design, conduct, measurements, data and analysis"* is necessary for all three types of reproducibility, in line with Stodden's view that availability of the computational environment is necessary for computational reproducibility. The goal of such transparency is, for Goodman, a way to ease evaluation of the weight of evidence from studies to facilitate future studies on actual knowledge gaps and cumulative knowledge, and reduce time spent exploring blind alleys from poorly reported research.

2.2 Reproducibility in Practice

Stodden (2010) presents a survey on what affects the decision to reveal code, data and ideas among registrants at the NIPS[2] conference up to and including 2008. Respondents tend to be steered by communitarian norms when sharing their work, such as one respondent stating that *"Much of my research would have not been possible if other people had not released their data or code. This is one of the main reasons for which I also want to contribute by releasing my own data and code"* (Stodden 2010). When not sharing material, however, private incentives dominate the decision. Examples of some prominent private incentives are the time necessary to clean up and document before a release and that the data or code may be used without citation. Stodden recorded positive improvements in the data and code sharing policies adopted by journals in Stodden et al. (2013).

While attempting to reproduce two NLP technologies, measuring WordNet similarity and Named Entity Recognition (NER), Fokkens et al. (2013) identified five main categories that influence reproduction. They note that omitting how *preprocessing* of data is done can break an experiment. The *experimental setup*, which steps to perform and how they are performed, had an impact on the NER experiment where the split of data for cross-validation lead to a major difference in results. Attention to *versioning* is important, as both experiments used datasets with different versions available and regular updates. This was also highlighted in Meng et al. (2015). *System output*, even output from intermediate steps, was found to be critical in the WordNet replication to identify why differences in output occurred. Finally, *system variations* may be inherent in the techniques used, such as how an algorithm reacts to a tie. Their reproductions provide new insights and better understanding of the techniques studied. They acknowledge that sharing of data and software was important for them to reproduce the work. Additionally, they observed that aspects with a huge impact on the results and conclusions of an experiment are often only mentioned in passing, or not at all if it is not the core of the research described (Fokkens et al. 2013).

Collberg & Proebsting (2016) studied the willingness of researchers to share data and code at eight ACM[3] conferences and five journals from 2011-13. Out of

[2]Neural Information Processing Systems: `https://nips.cc`

[3]Association for Computing Machinery: `http://www.acm.org/`

402 experimental papers they were able to reproduce 32.3% without communicating with the author, rising to 48.3% with communication. Papers by authors with industry affiliation showed a lower rate of reproducibility. Some notable build errors for papers they received distributions for but failed to build were: the distributed code is missing files, a third-party package could not be found, a pre-requisite tool could not be built, or a particular version of software, compiler, etc. was not available. Some of the reasons given for not sharing code for the papers they could not get code for (44%), were: could not find the final version corresponding to the paper, the code is being worked on for future release (too bad to release), the code was not intended to be released, the programmer was no longer available (student no longer part of program) or the code has restrictive licenses. An argument for why code should be published regardless is presented in LeVeque (2013), liking publishing of source code to the policy of publishing proof along with a mathematical theorem. As a low effort incentive to increase open source and open data, Collberg & Proebsting (2016) propose adding a mandatory sharing specification to the header of papers stating if code and data is accessible or not. The goal is to give the reader or reviewer an idea of how reproducible the research is up front, and create a contract the author has to fulfil. "The contract commits the author to making available certain resources ... and committing the reader/reviewer to take these resources into account when evaluating the contributions made by the paper" (Collberg & Proebsting 2016).

For preclinical medical research, Prinz et al. (2011) analysed data from 67 target validation projects in the pharmaceutical industry and found inconsistencies in published data and in-house data in almost two-thirds of the projects. The inconsistencies resulted in prolonged validation processes or in most cases termination of the projects as the evidence was insufficient to justify further investments. Further, they found no correlation between reproducibility and the impact factor of the journal the published data originated from. Among the observed reasons for a lack of reproducibility they mention incorrect or inappropriate statistical analysis, insufficient sample sizes, a bias towards publishing positive results and the control or reporting of experimental conditions, such as the description of materials and methods. There are signs that the competitive culture in science, where the number of high-impact-factor publications and grants matter more than the joy of discovery and collaborative contributions, encourages poor scientific practices (Casadevall & Fang 2011). Begley (2013) provides six questions to recognize whether the data from a preclinical paper is likely to stand up, focusing on the methodology presented. A later review of reproducibility problems in basic and preclinical biomedical research by Begley & Ioannidis (2014) highlights potential solutions to improve research reproducibility. Their recommendations are aimed at funding agencies, institutions, journals and investigators, and require each one to take part. For institutions they mention rewarding robust, rather than flashy studies and to reward researchers who comply with peer-generated guidelines and funding-agency requirements. These go hand in hand with a wish for journals and funding agencies to demand and reward good scientific methods. Generally, their recommendations aim at reducing the perceived need and incentives to publish early

by rewarding more robust publications and to label exploratory investigations as exploratory, highlighting that they need further validation.

2.3 Facilitating Reproducibility

Gent (2013) introduces a recomputation manifesto, using virtual machines to package an experiment with its environment. Additionally, Gent presents a site[4] with a free repository to store virtual machines aimed at recomputing experiments. It has been available since 2013. For the latest conference listed among its library, CP 2015[5], two out of more than 50 accepted papers published their experiments in virtual machines at the site. This initiative is not unique. ReproZip[6] packs an experiment with its necessary configuration to be easily reproducible on other machines, regardless of dependencies or operating system (Chirigati et al. 2013). General packaging tools like Docker (Boettiger 2015, Nagler et al. 2015) for software distribution or more environment specific tools such as iPython notebooks (Pérez & Granger 2007) (recently as a kernel for the more general Jupyter notebooks (Kluyver et al. 2016)), provide several means to publish reproducible experiments. Yilmaz (2012) introduce an e-Portfolio to publish code, data and scientific workflow in an ensemble of active documents. Though the means are available, Gent & Kotthoff (2014) observed that the main issue is their ease of use. The tools either have to be used from the very beginning of an experiment, or has a steep learning curve to retrofit the experiment. An attempt at making experiments using large-scale computing clusters easier to reproduce was presented at IEEE BigData 2016 (Monajemi et al. 2016). Similarly, Leitner et al. (2016) proposes the ACR Picking Benchmark[7], a robotics benchmark where all items necessary are easily available and common to allow wide spread use and easier reproduction of the benchmark experiment.

Hunold & Träff (2013) encourages the addition of a description of how to reproduce the findings in a publication. Highlighting that *"the description of how to reproduce findings should play an important part in every serious, experiment-based parallel computing research article"* (Hunold & Träff 2013). By describing how to reproduce findings in a publication, the reader is better equipped to verify the soundness of the experiment. Dolfi et al. (2014) provides a sample paper as a baseline model for other reproducible papers. This includes a guide for reproducing the experiment as an appendix, in line with the wishes of Hunold & Träff (2013). Additionally, the reader is given an idea of evaluation parameters to tweak to evaluate the robustness of the authors' evaluation.

A set of best practice principles for disseminating reproducible computational research is presented in Stodden & Miguez (2014). They are based around the idea that any publication of scientific work should include the material necessary to support its major claims, and enable other researchers to verify or reproduce

[4] `http://recomputation.org`
[5] `http://booleconferences.ucc.ie/cp2015`
[6] `https://vida-nyu.github.io/reprozip/`
[7] `http://juxi.net/dataset/acrv-picking-benchmark/`

the work. There are six principles, and they all relate to publication of materials supporting the dissemination. The first four consider code, methods, and data used during the experiment. They recommend open licensing of both data and code, and to make it easily accessible (for instance with a digital object identifier), tracking of workflow during the research process and that input values and randomization seeds that generated the results are included with the code. The fifth principle urges researchers to cite any and all 3rd party data and software used. Such citations might provide further incentive to publish material freely, seeing recognition for previous open sourced work. The last principle recommends following conditions associated with funded research, but to make the source of the requirements aware of improvements that could be made to be more in line with the previous principles (Stodden & Miguez 2014). Ten simple rules proposed by Sandve et al. (2013) mirror the principles presented by Stodden, with examples of how to implement the rules.

Chapter 3

Research Method

A literature survey was designed to investigate reproducible research, and its design is outlined in this chapter. Following the design is documentation of the evaluation procedure for each paper, and the chapter ends with a discussion of some known limitations of the survey.

3.1 Literature Survey Design

To investigate the reproducibility of published research, this survey employs manual evaluation of investigated papers, in contrast to previous related research attempting to run executables provided by authors. This decision is based on the ideal that any researcher should be able to attempt a reproduction, or evaluate the design of the research published in a paper. It does not mean that reproductions need to be successful, only that the information necessary to attempt one is made available for peers to critique. The decision to not communicate with authors is in part to save time the time necessary to gather supplementary materials, but also with the notion that the authors might not be available in the future. This notion is to some extent supported by Collberg & Proebsting (2016), where they found it quite common for the lead developer to be unavailable, due to students graduating, email addresses not working, or authors not having time to help.

A survey strictly based on published documentation has the benefit of allowing any reader to verify and critique evaluations made and the results presented in the future. Using material received from authors might differ based on when it is done and on the availability of the authors. Additionally, attempts to reproduce an experiment will depend on the investigators knowledge of the subject area, the time spent per experiment, and the computational resources available. A literature survey lowers the cost, and allows a larger sample population to analyse.

Disadvantages to a survey, however, includes a shift in focus to what can be counted and measured, as evidenced by the variables in section 3.1.1. Nuances and aspects not thought of when designing the survey may be overlooked, and the

depth of investigation into the research topic is limited.

The survey is an adaptation of the survey presented in Gundersen (2015), which evaluated 58 papers from the agent track at IJCAI 2013 as well as benchmarks from SHREC 2015.

3.1.1 Data requirements

The survey focuses on reproducibility in line with the terms methods reproducibility and results reproducibility defined by Goodman et al. (2016). Methods reproducibility requires enough information for another researcher to be able to, in theory, exactly perform the same procedures with the same data. As such, the experiment, methods, and data need to be made available. As for results reproducibility, another independent researcher should be able to corroborate the results following the same experimental procedures. What constitutes corroborating results is not well defined, and depends on the experiment. It is assumed that the data is not necessary, as the corroborated results should be in line with the results or analysis presented in the paper.

Directly topic related

With methods and results reproducibility in mind, the variables to record for each paper directly related to the topic cover the following categories: experiment documentation, methods documentation, and data documentation. The survey focuses on information available without contact with the authors, meaning resources need to be freely accessible and openly published.

Experiment documentation : How well documented is the experiment and the environment it was performed in.

- **Evaluation criteria** Are the criteria used to evaluate the method described?
- **Experiment set-up** Is the set-up for the experiment described? Are hyperparameters used during the experiment specified?
- **Hardware specification** Is the hardware used during the experiment specified?
- **Open experiment code** Is the code to run the experiment made available?
- **Software dependencies** Are software dependencies listed?

Method documentation : Documentation and availability of the method presented.

- **Pseudo-code** A textual description of the computational methods.
- **Open source code** The method source code is accessible.

Open data : Documentation and availability of the data used and generated during the experiment.

Open training data : Training data is available directly or through explicit mention of data split.

Open validation data : Validation data is available directly or through explicit mention of data split. Note that simply saying cross-validation was used is not enough, without specifying a type of cross-validation.

Open test data : Test data is available directly or through explicit mention of data split.

Open results data : Results data is published openly.

Indirectly topic related

To identify a paper and allow different analysis patterns, some indirectly topic related variables are recorded as well. These are sectioned into miscellaneous data, identifying data, and research transparency. Research transparency investigates explicit documentation of a natural-science based research method, to see if research methods in AI overlap with the traditional scientific method. The indirectly related data cover possible analysis patterns, such as: (I) reproducibility in relation to author affiliation, (II) reproducibility related to conference and instalment year, and (III) reproducibility related to novelty of research.

Identifying information : Includes recording of authors, the title and on-line link to paper.

Miscellaneous data : Data recorded to support analysis patterns and research characteristics.

Affiliation Are the authors affiliated with an academic institution, or industry?

Conference Notes the conference instalment the paper was published at.

Research type Separates theoretical and experimental papers.

Result outcome Does the paper present novel research?

Third-party citations Are third-party software and data cited correctly?

Research transparency : Explicit documentation of the research method in line with the scientific method.

Contribution Clear description of what the research contributes.

Research goal or objective Stated goals or objectives for the research.

Hypothesis Stated hypothesis to investigate.

Prediction Explicit mention of what the researchers predicted to see.

Problem description An explicit mention of what the investigated problem is.

Research method Description of the research method chosen.

Research question Explicit listing of the research question(s) of interest.

3.1.2 Data generation method

Data was generated by evaluating conference papers openly published in proceedings from two instalments of two different conference series. Physical copies can be ordered from the conferences, but all accepted papers are published on-line. This makes them easily available, and unobtrusive to obtain, allowing other researchers to scrutinize the research based on the original material.

The conference series investigated were the International Joint Conference on Artificial Intelligence (IJCAI) and the AAAI Conference on Artificial Intelligence (AAAI), specifically IJCAI-2013 and -2016, and AAAI-2014 and -2016. From these four instalments there are a combined population of 1910 accepted papers. IJCAI ran biennially until the first annual instalment in 2016, while AAAI was annual prior to 2014. Thus both conference series have had one instalment between the investigated years. Probabilistic random sampling was done for each conference, as documented in Appendix A[1]. A sample size of 100 from each conference was selected, restricting the necessary time to conduct the survey. The sample generation creates a list of 100 papers, with any sub-slice being a valid random sample.

For IJCAI-2013 the 58 papers from Gundersen (2015) were revisited, so only 42 of the 100 randomly sampled papers were included. This diminishes some of the representativeness of the IJCAI-2013 sample population. Due to the adapted survey, the papers were re-evaluated with the same procedure as the other papers.

The four conferences cover several disciplines within AI, and there may be differences within the disciplines. Between the conferences, however, it is assumed that the populations are not significantly different. This is based on the conferences covering the same disciplines at large, and employing blind peer review for acceptance with similar requirements in calls for papers. None of the conference series are vocal about open source or reproducible research. The AAAI conferences allow non-reviewed supplemental material, provided the documentation relevant for any claims is present in the paper itself, and the supplemental material adhere to the anonymity of the blind peer review. The IJCAI conferences do not allow supplemental material and discourage anonymized linking of supplemental material[2]. The confidence interval for each conference is reported in table 3.1, assuming that the populations are similar, a combined confidence interval is reduced to 4.36%.

[1] The sampling procedure is also available in a Jupyter notebook here: `https://github.com/sidgek/msoppgave`

[2] See FAQ for ICJAI-16, the same language was used for IJCAI-13: `https://docs.google.com/document/d/1hOOvHxLyxSen5mnjVZOfta2Yb8XK3Fr-5rsF9Rfc9Ag/`

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

The following enumerated list, shows the procedure followed for each paper. Evaluating a single paper was estimated to take 10 to 12 minutes. Theoretical papers are considerably quicker, due to most of the variables not being of interest.

1. Note down the title, authors, link, and conference instalment for the paper.
2. Look at the institutions the authors are affiliated with. If unsure, look the institutions up on-line.
   (a) If the institutions are research institutions or academic institutions, record affiliation as 0.
   (b) If industry institutions, record affiliation as 2.
   (c) If there are institutions from both industry and academia, record affiliation as 1.
3. Skim the abstract.
   (a) Is the presented research novel? (Record Result outcome as 1 for yes, 0 for no)
   (b) Does the abstract indicate an experiment? If not, scroll through the paper to look for an experiment. If no experiment is found record Research type as T for theoretical, if an experiment is found set E for experimental.
4. Search the paper file for explicit mentions of the following words: contribution, goal, objective, hypothesis, prediction, problem, research method, research question.
   (a) For each occurrence of a word, check the context it is mentioned in. If the context relates to the variables under *Research transparency*, record that variable as 1. Otherwise, record them as 0.
5. If the *Research type* was identified as theoretical, its evaluation is done and the remaining steps can be skipped.

6. Search through the paper.
   (a) If any pseudo-code is present, record pseudo-code as 1. Otherwise, 0.
   (b) If any references to supplementary materials is made, look it up and see if the method is included.
   (c) If any citations to datasets or source code are present, note third-party citations as 1. 0 otherwise.
7. The remaining variables are evaluated by skimming any sections identified as experimental or related to the experiment.
8. For mentions of datasets
   (a) If they are publicly available datasets or published by the authors, determine if any of the data is designated to training or validation. If neither set Open test data to 1, and the others to 0. If it is, set Open training data to 1.
   (b) If Open training data was set to 1, look for specification of validation and test split of the data in the paper and at the referenced location of the data (some published datasets come with designated splits). Set Open validation and Open test data to 1 respectively if found. Otherwise set to 0.
   (c) If any supplementary materials are referenced in the paper, look it up and see if the results data is available. Set to 1 if it is, 0 otherwise.
9. Look for mentions of CPU, RAM, GPU, AMD, Intel, GB.
   (a) Record Hardware specification as 1 if hardware is specified by model, 0 otherwise. Example of too little information: The experiment was run on a 4-core CPU. Accepted: The experiment was run on an Intel i5-4690K CPU at 3.5GHz.
10. Are any criteria to evaluate methods mentioned or discussed? Record 1 if yes, 0 if not.
11. Look for mentions of software dependencies.
   (a) If code is released, check for a requirements file or readme with requirements. If not see if there's any mention of software and its version in the paper. Set 1 if it is available, 0 if not. Note that including the version number is necessary.
12. Find any description of the experiment procedures themselves.
   (a) Are there procedures for running the experiment mentioned? Are hyperparameters used for methods given or discussed? This may be documented in the source code. If found set experiment set-up to 1, otherwise 0.

13. If any supplementary materials are referenced, see if they include code to run the experiment. Set open experiment code to 1 if they are, 0 otherwise.

Refer to Appendix B[3] for an abbreviated sample of how the survey data is structured. The data was recorded in a Google spreadsheet, and exported to a csv file.

3.3 Limitations of the Survey

Hardware specifications and software dependencies are difficult to evaluate. For software dependencies, it is difficult to specify software versions in a paper without sacrificing valuable paper space. If the code is not released, the value of it is low. For released code it is common best-practices to have a requirements file along with the code, it is highly recommended to do so. Hardware specification is difficult to set baselines for, as some experiments may be run through cloud services, or on multiple devices it can be difficult to be precise. Additionally, depending on the experiment and implementation, the hardware might not be particularly relevant. It does however give an indication of the resources necessary to perform the experiment. If speed and performance is an important factor in a paper, ranking of different methods is more valuable than exact speed, but it might still be valuable to discuss how the resources available to the methods impact the ranking.

Similarly, for experiment set-up, how detailed the description needs to be varies. A baseline for how detailed it is necessary to be for reproduction depends heavily on the reader. Ideally code to run the experiment is available which allows other researchers to examine the set-up. The paper should still include enough information on the steps done during the experiment for someone familiar with the field to recognize. During evaluations for the survey, experiment set-up became a check for discussion of hyper-parameters rather than experiment procedures. This shows an example of how evaluation bias can impact the survey, where a variable was modified to the investigators meaning rather than the intended.

The survey does not take licensing into account when evaluating the availability of data and code. Evaluations noted to make data and code available, may restrict the use without any indication in the evaluation data.

While it is intended, it is also important to repeat that the survey does not show successful attempts at reproduction. It merely investigates the availability of materials deemed necessary to attempt a reproduction. While such attempts would be valuable and interesting, the cost is substantial compared to the cost of the survey. Such attempts would likely require more manpower or a reduced sample size, but would also give a more precise indication of whether enough information is provided in a paper.

[3]The full dataset is also available at `https://github.com/sidgek/msoppgave`

Chapter 4

Results and Analysis

Results are presented in the following chapter. The chapter is separated into sections in line with the categories defined in section 3.1.1, in the following order: miscellaneous, research transparency, method documentation, experiment documentation, and open data. The chapter continues with an analysis of open source and open data in relation to author affiliation and conference instalment, ending with a look at methods and results reproducibility.

All figures presented were generated with the Jupyter notebook[1] in Appendix C[2].

4.1 Miscellaneous

The following variables are in the miscellaneous category: affiliation, conference, research type, result outcome, and third-party citation. The data for each variable except conference can be seen in figure 4.1. The conference distribution is seen in table 4.1. There is a clear dominance of academia affiliated papers, amounting to 82.8% (331) of the evaluated papers. Similarly, experimental papers dominate over theoretical, at 81.2% (325).

For Third-party citations, the intent was to record whether software and data used for an experiment is cited. For the most part, the papers noted with *Present* in figure 4.1d show correct citations to public datasets. A few papers did not include citations to papers even when the creators ask for a citation. However, they may not have asked for citations at the time of publication. The majority of papers noted as *Not present* may not have used citable third-party software. In general papers are good at citing datasets and methods they compare with, but there are few citations to software programs and libraries.

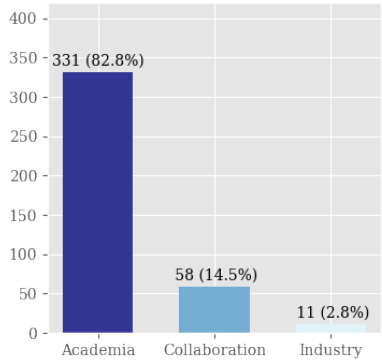
[1] Project Jupyter: `http://jupyter.org/`

[2] The full dataset and the Jupyter notebook is available at `https://github.com/sidgek/msoppgave`

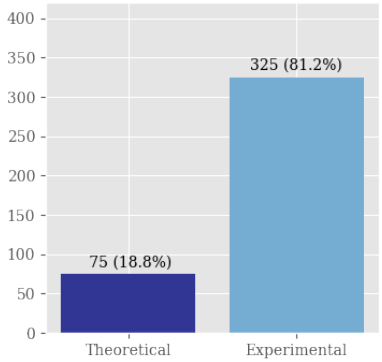
Result outcome (figure 4.1c) was erroneously recorded as a positive result rather than novel research. This would be any paper that presents confirmation of a hypothesis, or where the wording of their findings present a solution or improvement to something. Since very few papers include a hypothesis in the first place, the data for this variable will not be considered further and is merely presented for completeness.

Conference	**Papers**
AAAI 14	100
AAAI 16	100
IJCAI 13	100
IJCAI 16	100

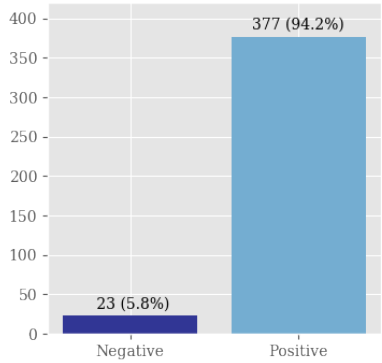
Table 4.1: Distribution of papers between conferences.



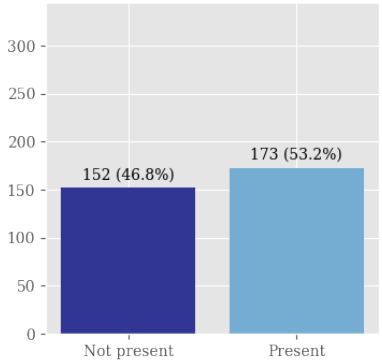
(a) Affiliation



(b) Research type



(c) Result outcome



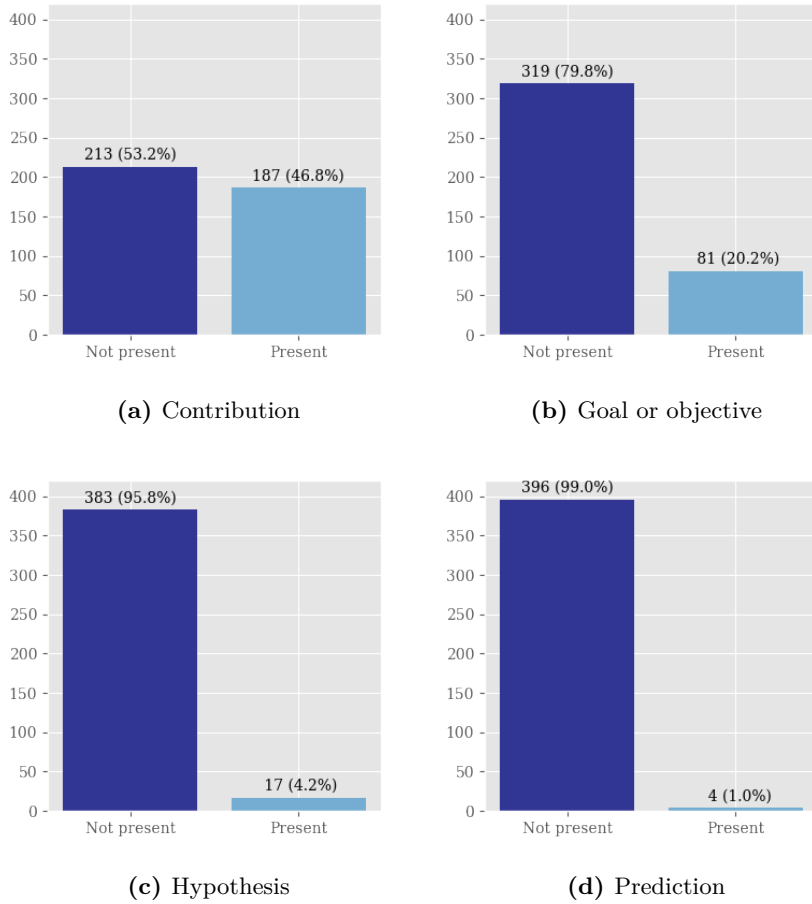
(d) Third-party citations

Figure 4.1: Frequencies for miscellaneous data. There are a total of 400 papers. However, for Third-party citation only the 325 experimental papers are relevant, accounting for the lower values of the y-axis.

4.2 Research Transparency

Research transparency variables relate to research methodology. This includes explicit mentions of: contribution, research goal or objective, hypothesis, prediction, problem description, research method, and research question. The distributions for each variable can be seen in figure 4.2 and 4.3. From the examined papers, contribution (46.8%), problem description (46.5%), and goal/objective (20.2%) are mentioned most. The remaining terms are seen in between 1 and 5 percent of the papers. However, the requirement for explicit mentions of the given terms may skew the data negatively. This is particularly true for contributions which was

seen in almost all papers, but not necessarily with the term contribution.



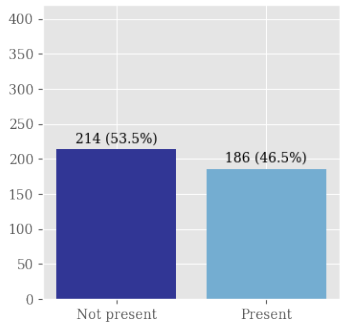
(a) Contribution

(b) Goal or objective

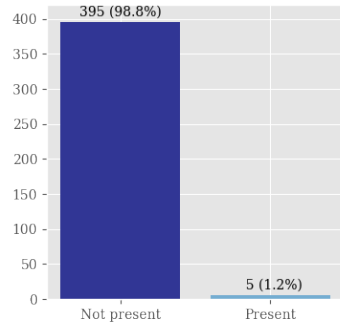
(c) Hypothesis

(d) Prediction

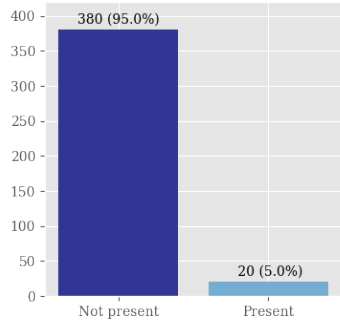
Figure 4.2: Data on research transparency. A term is *Present* if it is explicitly mentioned in a paper. These variables were recorded for all 400 papers.



(a) Problem description



(b) Research method

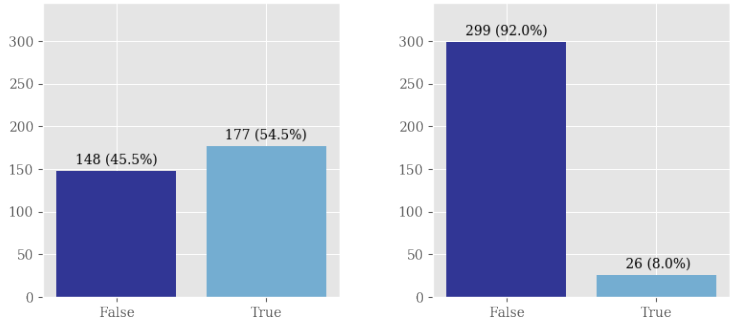


(c) Research question

Figure 4.3: Continuation of data on research transparency. A term is *Present* if it is explicitly mentioned in a paper. These variables were recorded for all 400 papers.

4.3 Method Documentation

Method documentation investigates the availability of the method under investigation through pseudo-code, and open source code. Only the 325 experimental papers are relevant, as seen by the left axis. The data is presented in figure 4.4. Pseudo-code is present in about half (54.5%) of the examined papers. The variable is not a good estimate for how many document their method, however, as there are other ways to present it. Open source code is only seen in 26 (8%) of the papers. A few papers referenced material that was no longer available during the evaluation, or that material would be published in the future.

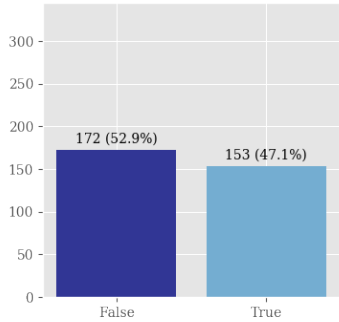


(a) Pseudo-code **(b)** Open source code

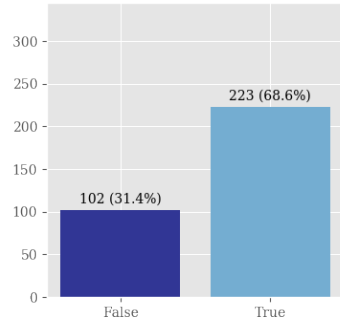
Figure 4.4: Method documentation data. These variables are applicable to the 325 experimental research papers.

4.4 Experiment Documentation

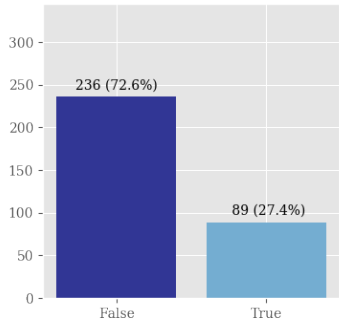
Experiment documentation relates to how well documented the experiment is and if the experiment is made available. The following variables are included: evaluation criteria, experiment set-up, hardware specification, open experiment code, and software dependencies. A summary of the data can be seen in figure 4.5. As in the method documentation category, only the experimental papers are relevant. Open experiment code (5.5%), hardware specification (27.4%), and software dependencies (16.0%) are the least documented. Sharing of experiment code is a little bit lower than sharing of source code (8% in figure 4.4b). Evaluation criteria seems low at 47.1%, but the evaluation was a little stricter than the procedure in section 3.2, requiring explicit mentions of the criteria and not just shown as results. Experiment set-up diverged from the original intent, becoming a measure of whether parameters used to instantiate the method and experiment are mentioned or discussed rather than a description of the experiment procedures.



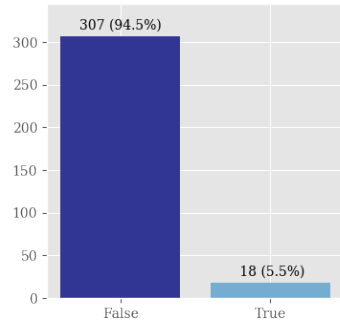
(a) Evaluation criteria



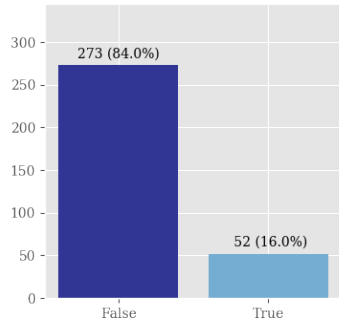
(b) Experiment set-up



(c) Hardware specification



(d) Open experiment code



(e) Software dependencies

Figure 4.5: Experiment documentation data. These variables are only applicable to the 325 experimental research papers.

4.5 Open Data

Open Data relates to the availability of data used during an experiment and documentation of dataset splits. The following variables are included: training data, validation data, test data, and results data. Figure 4.6 show the results for experimental papers. Most of the papers sharing open data do so by using public datasets, accounting for the higher proportion of training (32.0%) and test data (29.8%) compared to validation data (9.2%). The amount of papers with open validation data would be closer to open training data if more papers explicitly specified a type of cross-validation, instead of just mentioning cross-validation. Results data is rarely shared (3.7%), but occasionally bundled with the open source code.

(a) Open training data

(b) Open validation data

(c) Open test data

(d) Open results data

Figure 4.6: Results on the availability of open data. These variables are only relevant to the 325 experimental research papers. Of special note is the *N/A* column in (a) and (b), which show the amount of papers where either training or validation data was deemed to not have been used in a paper due to the nature of the methods presented. This skews the percentage distribution. Percentages without *N/A* amount to: (a) 43.8% *False* and 56.2% *True*, and (b) 83.5% *False* and 16.5% *True*.

4.6 Patterns for Analysis Revisited

The patterns for analysis from 3.1.1 were: (I) reproducibility in relation to author affiliation, (II) reproducibility related to conference and instalment year, and (III) reproducibility related to novelty of research. Since the result outcome variable for novelty of research was evaluated erroneously, data for this analysis is not presented. As open data and code are considered paramount for reproducible research, the variables analysed are open source code, experiment code, training, validation, test, and results data.

The differences in the variables open source code, experiment code, training,

validation, test and results data when author affiliation is accounted for can be seen in table 4.2. Papers affiliated with academia amount to 265, collaboration to 50 and industry to 10. For open source code, experiment code, validation, test, and results data there is little to suggest significant differences based on affiliation. For training data there seems to be more openness in academia affiliated papers, likely due to collaboration giving access private industry data.

Variable	Academia	Collaboration	Industry
Open source code	23 (8.7%)	2 (4.0%)	1 (10%)
Open experiment code	15 (5.7%)	2 (4.0%)	1 (10%)
Open training data	86 (61.0%)	16 (45.7%)	2 (22%)
Open validation data	25 (17.9%)	5 (14.7%)	0 (0%)
Open test data	80 (30.2%)	15 (30.0%)	2 (20%)
Open results data	11 (4.2%)	0 (0%)	1 (10%)

Table 4.2: Differences in adoption of open source and data based on affiliation. It is important to note that the amount of experimental papers affiliated with academia dominates at 265, compared to 50 and 10 for collaboration and industry respectively. For training data and validation data, some papers are N/A: respectively 124 and 125 for academia, 15 and 16 for collaboration, and 1 and 2 for industry.

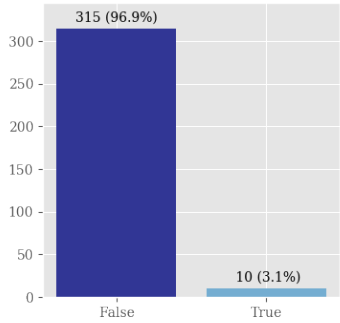
The split between conferences for experimental papers is as follows: 85 papers from AAAI-14, 85 papers from AAAI-16, 71 papers from IJCAI-13, and 84 papers from IJCAI-16. The high amount of papers not applicable to training and validation data from IJCAI 13 is likely due to the sample population involving 58 papers from the agent track, mentioned in section 3.1.2. With the confidence intervals calculated per conference per variable in mind, none of the variables can be said to show differences between conference series or series instalments (table 4.3). For IJCAI, there seems to be a slight increase in availability of test data from $18.3 \pm 8.5\%$ to $35.7 \pm 8.5\%$. However, this can be a result of the inconsistent sampling approach for IJCAI 2013.

Variable	AAAI 14	AAAI 16	IJCAI 13	IJCAI 16
Source code	7 (8.2±4.7%)	9 (10.6±5.5%)	2 (2.8±2.8%)	8 (9.5±5.2%)
Experiment code	4 (4.7±3.6%)	6 (7.1±4.6%)	0 (0%)	8 (9.5±5.2%)
Training data	25 (51.0±8.5%)	39 (60±8.7%)	9 (42.8±8.5%)	31 (51.7±8.9%)
Validation data	5 (10.4±5.2%)	9 (13.8±6.1%)	4 (20.0±6.8%)	12 (20.3±7.1%)
Test data	24 (28.2±7.6%)	30 (35.3±8.5%)	13 (18.3±6.6%)	30 (35.7±8.5%)
Results data	2 (2.4±2.6%)	2 (2.4±2.7%)	0 (0%)	8 (9.5±5.2%)

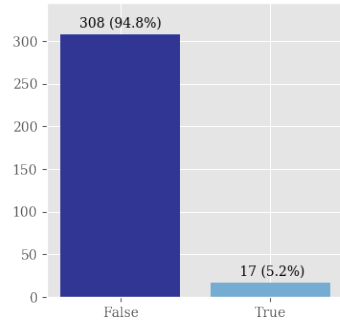
Table 4.3: Differences in adoption of open source and data based on conference instalment. The amount of experimental papers for each conference is as follows: 85 for AAAI 14, 85 for AAAI 16, 71 for IJCAI 13, and 84 for IJCAI 16. For training data and validation data, some papers were not applicable: respectively 36 and 37 for AAAI 14, 20 and 20 for AAAI 16, 50 and 51 for IJCAI 13, and 34 and 35 for IJCAI 16. The confidence intervals reported were calculated with `https://www.surveysystem.com/sscalc.htm`, with the population and sample size (100) specific for each conference and the percentages recorded from the survey, for a 95% confidence level. The population sizes are reported in table 3.1.

4.7 Reproducibility

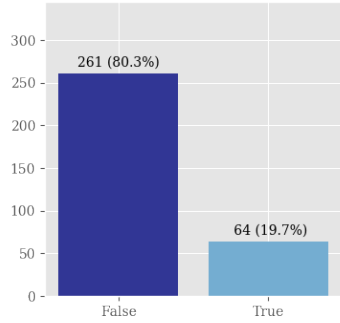
Open source and data are the most relevant to reproducibility. For methods reproducibility, a paper is considered good enough if experiment and source code, as well as all data except results data is available. For results reproducibility, the data requirements are removed. Figure 4.7 show how many of the 325 experimental papers cover all variables necessary for methods and results reproducibility, and how many papers cover the training, validation, and test data variables. As low as 3.1% (10 papers) make both code and data available to allow methods reproduction. Papers covering the variables for results reproduction amount to 5.2% (17 papers). Out of the 26 papers where the method source code is available, 17 of them include the experiment. Out of the 18 papers where the experiment code is available, 17 include the method code as well. Specifically for sharing of data, 27.4% (89 papers) provide training and test data if applicable. Requiring validation as well, reduces the number to 19.6% (64 papers).



(a) Methods reproducible papers



(b) Results reproducible papers



(c) Papers with open training, validation and test datasets. Relaxing the validation requirement results in 89 (27.4%).

Figure 4.7: The amount of experimental papers covering (a) methods and (b) results reproducibility through open source code and data. (c) Shows the amount of papers where training, validation and test sets are available if applicable, highlighting a drop from 19.7% with necessary data to 3.1% with data and code for methods reproducible papers. Papers where training or validation set is not applicable are counted as *True* if the remaining variables are 1.

Chapter 5

Evaluation

This chapter evaluates the conducted research by examining the research questions in light of the evidence found. A closer look at measures to improve the state of reproducibility is also presented.

5.1 Research Questions Revisited

In this section the research questions are revisited in turn. Each question will be examined in turn, following a repetition of the question being discussed.

RQ1: *What is the state of reproducibility at AI conferences?*

This question is central to evaluate the rest of the research questions and is the basis for the design of the survey conducted. An answer first needs to consider what is meant by reproducibility, and how to measure its state. The definition of reproducibility is covered by Goodman et al. (2016), and the survey was designed to cover methods and results reproducibility. While results reproducibility requires experiment and method source code, methods reproducibility additionally requires the data used. Section 4.7 highlights the relevant data from the survey, resulting in 3.1% methods reproducible and 5.2% results reproducible papers. The values are dismal compared to the 33.4% reported for the ACM conferences covered by Collberg & Proebsting (2016) with no contact with authors. It is first and foremost a result of few papers making code available. Note that results data was not included in the methods reproducible definition, which would reduce it further. This choice was based on a focus on similar or supporting results, rather than identical, since certain experiments might not be deterministic. Results data have been shown to help when differences occur in reproduction, to find the underlying causes (Fokkens et al. 2013).

RQ2: *What documentation is missing from AI papers to support reproducibility?*

Textual documentation practices are for the most part decent, but could see more explicit mentions of resources used in cases of hardware specification and software dependencies, providing readers an indication of the resources necessary to attempt a reproduction. Additionally, more precise information on dataset splits should be encouraged. However, there are three practices related to open code and data with particularly low values. In increasing order, they are: results data (3.7%), open experiment code (5.5%), and open source code (8.0%). Results data generally goes hand in hand with open source and experiment code, as it is common to see code repositories containing example data and results. Taking the additional work necessary to make one of the three available, can often lead to all three being made available with minimal extra effort. If the rate of open code was similar to open data, the state would be a little behind the 33.4% reported by Collberg & Proebsting (2016) for ACM conferences.

RQ3 *What practices have seen an improvement in recent years?*

There is no observable improvement from the earlier instalments to the later. The frequencies recorded for the later instalments are slightly higher, but not high enough to make confident claims considering the confidence intervals. For a closer look at the data, refer to table 4.3.

RQ4: *What practices are encouraging reproducible research in AI?*

This research question is difficult to answer through the survey data gathered, considering the lack of improvement observed in the previous research question. However, the use of publicly available datasets and benchmarks for comparisons of performance is still encouraging.

RQ5: *What incentives could be implemented to further encourage reproducible research?*

Incentives to encourage reproducible research should focus on making code available, as evidenced by the missing documentation mentioned in RQ2. As a bonus, some measures to improve code availability are also likely to facilitate sharing of data. Possible incentives and measures are discussed in section 5.2, and cover: making adapted good-enough practices; requiring a sharing contract for data, code, and support; requiring a guide on how to reproduce the experiment; providing a sharing platform in line with the double blind-review process; and introducing a voluntary reproduction review for accepted papers. The different possibilities range from little complexity and cost such as the sharing contract, to more complex, time consuming measures such as a reproduction review. As a minimum, IJCAI should stop discouraging authors to include links to supplemental material[1].

RQ6: *Does the author affiliation impact documentation practices?*

This is not conclusively answered by the survey conducted. With the bad practices overall for code sharing, there is little data to answer the question. Yet, the practices for open data indicate a slight decrease in availability for collaborative

[1] `http://ijcai-17.org/FAQ.html#q9`

and industry affiliated authors. An hypothesis for why academia affiliated papers have slightly more open data is that collaboration allows access to private datasets which are unlikely to be published openly, potentially for legal reasons to keep an industry advantage. The question would be better answered by the approach of Collberg & Proebsting (2016), where communication with authors is included in the survey. Attempting to receive data and code by contacting authors for reproduction may give more evidence for this hypothesis.

5.2 Encouraging Reproducible Research

Five recommended steps to encourage reproducible research at the investigated conference series are mentioned here. Each one is examined in turn, ranging from what is considered less complicated or time consuming, to more complex and time consuming measures. All of the approaches require work from authors and from the conference committee. The focus is on increased sharing of code. Some of the approaches will have the side-effect of increased sharing of data as well, or better textual documentation of experiments.

5.2.1 Sharing Contract

A sharing contract, or specification, is an explicit statement of the authors' intent to share open data, open code, and support peers who wish to examine the underlying experiment of the findings presented. It creates a contract which determines what readers can expect from the authors. It does not require all papers to publish data or code, but makes the reader more aware and potentially more sceptical when it is not shared. The contract, as introduced by Collberg & Proebsting (2016), is concise and contains the following information: location of resources and/or a contact email; external resources backing up the results, such as code, data, and media; and the level of technical support. For each resource, it specifies the accessibility, cost, and the form of the resource (e.g. source, binary, executable, as a service, sanitized). The support should state if it will be for bug fixes, installation, the cost, and a deadline for when it might no longer be available. An example contract for this thesis is included here.

Location: `https://github.com/sidgek/msoppgave`, `sigbjokj@stud.ntnu.no`

Code: access, free, source

Data: access, free

Support: bug fixes, free, 2018-01-01

5.2.2 How-to Reproduce Guide

As mentioned by Hunold & Träff (2013) and exemplified by Dolfi et al. (2014), including a detailed description of how to reproduce the findings of a paper would

be beneficial. Such a guide presupposes that the code and data is already available, but give a specific description of the steps taken to run an experiment and analyse the data. While Dolfi et al. (2014) append it as an appendix to the original paper, publishing it as supplemental material along the original paper can be a middle-ground for conferences with strict page limits. Even without code or data, publishing more detailed descriptions of experiments as supplemental material may still be a step in the right direction, allowing readers with a special interest to evaluate the design of the experiment more closely.

5.2.3 Publish Supplemental Materials

The AAAI conference series allows authors to upload supplemental material[2], specifying that it will not be reviewed. This is generally used for extended proofs, but could be used to encourage sharing of more resources. IJCAI would have to start such an implementation from scratch, as well as reverse their current practice of discouraging links to such material. The main issues stated are the need for papers to be self-contained and the blind peer-review. For the first issue, the current peer-review process should be enough to keep that requirement. For the second, encouraging addition after acceptance or not making the uploaded supplemental material available to reviewers before acceptance are two possible approaches. If the storage space necessary becomes a problem, exploring scientific repositories meant for data and code, and utilising them is also an option. Either way, it could limit the issue of links becoming dormant, which was the case for a couple of the papers from IJCAI 2013 previously recorded by Gundersen (2015) to have external resources available. Digital Object Identifiers for the supplemental materials would be a welcome bonus as well.

5.2.4 Good-enough Practices

A long term goal should be to create some good-enough practices as a guide to reproducible research within AI. As inspiration, Stodden & Miguez (2014), Wilson et al. (2016) and Sandve et al. (2013), are all good examples, and cover important parts of the research process. This approach would require educating researchers on the practices, which is likely to take a long time. Additionally, the practices should be adapted over time as new tools, and the adoption increases, potentially making the bar for good-enough practices higher over time.

5.2.5 Voluntary Reproduction Review

Including an optional reproduction review for accepted papers is likely the most time consuming, but also the most rewarding suggestion. ACM has had significant focus on reproducibility in recent years (Boisvert 2016), and SIGMOD has been a pioneer with their introduction of a reproduction review in 2008. Several other ACM conferences followed suit, and in 2015 the *ACM Transactions on Mathematical Software*(TOMS) journal began a similar initiative. Successful reviews lead to

[2] `http://www.aaai.org/Conferences/AAAI/2017/aaai17call.php`

increased recognition at the conferences, and prizes for excellence. For the TOMS Initiative and Policies for Replicated Computational Results[3], replication is done by an independent reviewer working together with the authors if the authors opt in after acceptance. The review process lead to a review report accompanying the paper, and the reviewer being acknowledged in the published paper as the author of the report. The additional exposure is an incentive for authors to commit to the review, while the acknowledgement of reviewers reward their work and effort. An implementation for AAAI or IJCAI can learn from the experience of these implementations, while also increase publication of reproduction attempts.

[3] http://toms.acm.org/replicated-computational-results.cfm

Chapter 6

Conclusion and Future Work

This chapter concludes the research, revisits the hypothesis and summarises the proposed measures to improve reproducibility. A list of proposed future work ends the chapter.

6.1 Conclusion

This thesis sought an overview of documentation practices for reproducible research at two AI conference series. A total of 400 papers were surveyed, split equally between two instalments of two conference series, recording the availability of code, data, and documentation of experiments. We hypothesized that the documentation of experiments is not good enough to consider papers reproducible, and that documentation practices have improved in recent years. The main drivers for reproducible research were determined to be open code and open data, with open code being required for results reproducibility and both being required for methods reproducibility.

The evidence supports the first hypothesis, but there is little evidence to support the second. None of the conference instalments have code or data policies, but the IJCAI instalments discouraged links to supplemental material and the AAAI instalments allowed uploading of supplemental material not in conflict with the blind peer-review. Surprisingly, there is no observable difference in availability of code or data between the conference series or when comparing the instalments within a conference series. About a third of the papers make data available, usually by using available open datasets, but some fail to document splitting of data. For code, only 5.5% make the experiment code available, while 8% make their methods source code available. Due to the low availability of source code, the reported amount of papers covering methods reproducibility and results reproducibility are 3.1% and 5.2% respectively.

Among the suggested ways to improve reproducibility, establishing some best-, or good-enough, practices for how to conduct reproducible AI research, and imple-

menting a voluntary reproduction review for future conference instalments should be a priority. Both allow the community to contribute, and the voluntary choice for a reproduction review makes it forgiveable during an introduction period.

6.2 Future Work

The following list contain topics for proposed future work:

Survey Specification: Improve the specification of the survey, to reduce the subjectivity of certain variables. This is especially relevant for experiment set-up, evaluation criteria, and software dependencies. Additionally, dataset availability should be evaluated as a whole, with documentation of dataset use as a separate variable instead of the division into three variables for training, validation and test data.

Expand Evidence: Evaluating future instalments of the conference series can measure the effect of new policies for documentation. While evaluation of other conference series or journals could provide evidence of what policies improve reproducibility, assuming policies differ.

Sharing Policies: Examine and evaluate policies adopted by other conference series and journals, both within AI and related computational research. The reproduction review processes for SIGMOD[1] and TOMS[2] are good starting points.

Reproduction Study: An in depth attempt at reproduction would provide a clear indication of what information is necessary for each paper and potentially the motives behind sharing or not. Two previous approaches are of note, the extensive reproduction projects at Open Science Framework Aarts et al. (2016), Errington et al. (2017), and the simpler attempt at running experiment code from Collberg & Proebsting (2016).

[1] SIGMOD Reproducibility: `http://db-reproducibility.seas.harvard.edu/`

[2] The TOMS Initiative and Policies for Replicated Computational Results: `http://toms.acm.org/replicated-computational-results.cfm`

Bibliography

Aarts, A. A., Anderson, C. J., Anderson, J., van Assen, M. A. L. M., Attridge, P. R., Attwood, A. S., Axt, J., Babel, M., Bahník, t., Baranski, E. & et al. (2016), 'Reproducibility project: Psychology'.
URL: *osf.io/ezcuj*

Begley, C. G. (2013), 'Reproducibility: Six red flags for suspect work', *Nature* **497**(7450), 433–434.
URL: *http://dx.doi.org/10.1038/497433a*

Begley, C. G. & Ellis, L. M. (2012), 'Drug development: Raise standards for preclinical cancer research', *Nature* **483**(7391), 531–533.
URL: *http://dx.doi.org/10.1038/483531a*

Begley, C. G. & Ioannidis, J. P. A. (2014), 'Reproducibility in science: Improving the standard for basic and preclinical research', *Circulation Research* **116**(1), 116–126.
URL: *http://dx.doi.org/10.1161/CIRCRESAHA.114.303819*

Boettiger, C. (2015), 'An introduction to docker for reproducible research', *ACM SIGOPS Operating Systems Review* **49**(1), 71–79.
URL: *https://arxiv.org/pdf/1410.0846.pdf*

Boisvert, R. F. (2016), 'Incentivizing reproducibility', *Commun. ACM* **59**(10), 5–5.
URL: *http://doi.acm.org/10.1145/2994031*

Brown, C. T. (2012), 'Our approach to replication in computational science', *Living in an Ivory Basement* .

Buckheit, J. B. & Donoho, D. L. (1995), Wavelab and reproducible research, *in* 'Wavelets and Statistics', Springer New York, New York, NY, pp. 55–81.

Casadevall, A. & Fang, F. C. (2011), 'Reforming science: Methodological and cultural reforms', *Infection and Immunity* **80**(3), 891–896.
URL: *http://dx.doi.org/10.1128/IAI.06183-11*

Chirigati, F., Shasha, D. & Freire, J. (2013), Reprozip: Using provenance to support computational reproducibility, *in* 'Presented as part of the 5th USENIX Workshop on the Theory and Practice of Provenance'.

Claerbout, J. F. & Karrenbach, M. (1992), Electronic documents give reproducible research a new meaning, *in* 'SEG Technical Program Expanded Abstracts 1992', Society of Exploration Geophysicists, pp. 601–604.

Collberg, C. & Proebsting, T. A. (2016), 'Repeatability in computer systems research', *Commun. ACM* **59**(3), 62–69.
URL: *http://doi.acm.org/10.1145/2812803*

Dolfi, M., Gukelberger, J., Hehn, A., Imriška, J., Pakrouski, K., Rønnow, T., Troyer, M., Zintchenko, I., Chirigati, F., Freire, J. et al. (2014), 'A model project for reproducible papers: critical temperature for the ising model on a square lattice', *CoRR, abs/1401.2000* .

Drummond, C. (2009), 'Replicability is not reproducibility: nor is it good science', *International Conference on Machine Learning* .
URL: *http://cogprints.org/7691/*

Errington, T. M., Tan, F. E., Lomax, J., Perfito, N., Iorns, E., Gunn, W., Nosek, B. A., Buck, S., Griner, E. M., Maherali, N. & et al. (2017), 'Reproducibility project: Cancer biology'.
URL: *osf.io/e81xl*

Fokkens, A., Erp, M. V., Postma, M., Pedersen, T., Vossen, P. & Freire, N. (2013), Offspring from reproduction problems: What replication failure teaches us, *in* 'Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics (ACL), pp. 1691–1701.

Francis, G. (2012), 'The psychology of replication and replication in psychology', *Perspectives on Psychological Science* **7**(6), 585–594.
URL: *http://dx.doi.org/10.1177/1745691612459520*

Gent, I. P. (2013), 'The recomputation manifesto', *CoRR, abs/1304.3674* .

Gent, I. P. & Kotthoff, L. (2014), Recomputation. org: Experiences of its first year and lessons learned, *in* 'Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference on', IEEE, pp. 968–973.

Goodman, S. N., Fanelli, D. & Ioannidis, J. P. A. (2016), 'What does research reproducibility mean?', *Science Translational Medicine* **8**(341), 341ps12–341ps12.
URL: *http://stm.sciencemag.org/content/8/341/341ps12*

Gundersen, O. E. (2015), Towards Scientific Benchmarks: On Increasing the Credibility of Benchmarks, *in* I. Pratikakis, M. Spagnuolo, T. Theoharis, L. V. Gool & R. Veltkamp, eds, 'Eurographics Workshop on 3D Object Retrieval', The Eurographics Association.

Hunold, S. (2015), 'A survey on reproducibility in parallel computing', *CoRR, abs/1511.04217* .

Hunold, S. & Träff, J. L. (2013), 'On the state and importance of reproducible experimental research in parallel computing', *CoRR, abs/1308.3648* .

Ioannidis, J. P. A. (2005), 'Why most published research findings are false', *PLoS Medicine* **2**(8), e124.
URL: *http://dx.doi.org/10.1371/journal.pmed.0020124*

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S. et al. (2016), Jupyter notebooks—a publishing format for reproducible computational workflows, *in* 'Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing', IOS Press, p. 87.

Leitner, J., Tow, A. W., Dean, J. E., Suenderhauf, N., Durham, J. W., Cooper, M., Eich, M., Lehnert, C., Mangels, R., McCool, C. et al. (2016), 'The acrv picking benchmark (apb): A robotic shelf picking benchmark to foster reproducible research', *CoRR, abs/1609.05258* .
URL: *https://arxiv.org/pdf/1609.05258.pdf*

LeVeque, R. J. (2013), 'Top ten reasons to not share your code (and why you should anyway)', *SIAM News, April* **1**.

Meng, H., Kommineni, R., Pham, Q., Gardner, R., Malik, T. & Thain, D. (2015), 'An invariant framework for conducting reproducible computational science', *Journal of Computational Science* **9**, 137–142.

Monajemi, H., Donoho, D. L. & Stodden, V. (2016), 'Making massive computational experiments painless', *IEEE BigData2016, Open Science in Big Data (OSBD 2016)* .

Nagler, R., Bruhwiler, D., Moeller, P. & Webb, S. (2015), 'Sustainability and reproducibility via containerized computing', *CoRR, abs/1509.08789* .

Oates, B. J. (2006), *Researching Information Systems and Computing*, SAGE Publications Ltd.

Peng, R. D. (2011), 'Reproducible research in computational science', *Science* **334**(6060), 1226–1227.

Pérez, F. & Granger, B. E. (2007), 'IPython: a system for interactive scientific computing', *Computing in Science and Engineering* **9**(3), 21–29.
URL: *http://ipython.org*

Prinz, F., Schlange, T. & Asadullah, K. (2011), 'Believe it or not: how much can we rely on published data on potential drug targets?', *Nature Reviews Drug Discovery* **10**(9), 712–712.
URL: *http://dx.doi.org/10.1038/nrd3439-c1*

Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. (2013), 'Ten simple rules for reproducible computational research', *PLoS Computational Biology* **9**(10), e1003285.
URL: *http://dx.doi.org/10.1371/journal.pcbi.1003285*

Simmons, J. P., Nelson, L. D. & Somonsohn, U. (2011), 'False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant', *Psychological Science* .

Stodden, V. C. (2010), 'The scientific method in practice: Reproducibility in the computational sciences', *Columbia Univeristy Academic Commons* .
URL: *http://hdl.handle.net/10022/AC:P:11417*

Stodden, V. C. (2011), 'Trust your science? open your data and code', *Amstat News* pp. 21–22.

Stodden, V. C. (2013), 'Resolving irreproducibility in empirical and computational research', *IMS Bulletin* **42**(8), 12–13.

Stodden, V. C. (2014), 'What scientific idea is ready for retirement? Reproducibility', *Edge.org* .
URL: *http://edge.org/response-detail/25340*

Stodden, V. C., Guo, P. & Ma, Z. (2013), 'Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals', *PLoS ONE* **8**(6), e67111.
URL: *http://dx.doi.org/10.1371/journal.pone.0067111*

Stodden, V. C. & Miguez, S. (2014), 'Best practices for computational science: Software infrastructure and environments for reproducible and extensible research', *Journal of Open Research Software* **2**(1), e21.
URL: *http://dx.doi.org/10.5334/jors.ay*

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L. & Teal, T. K. (2016), 'Good enough practices in scientific computing', *CoRR* **abs/1609.00037**.
URL: *http://arxiv.org/abs/1609.00037*

Yilmaz, L. (2012), 'Reproducibility in m&s research: issues, strategies and implications for model development environments', *Journal of Experimental & Theoretical Artificial Intelligence* **24**(4), 457–474.

Appendix

Appendix A: Sample Selection

The following pages are generated from Jupyter notebook and document the sample selection procedure. The notebook is also available in the supplementary materials or at `https://github.com/sidgek/msoppgave`, together with the generated samples and files with the population of accepted papers they were generated from.

paper_selection

June 11, 2017

0.1 Sampling of papers from conferences

This Jupyter notebook shows the procedure used to select a sub-sample of the accepted papers from each conference.

0.1.1 Accepted conference papers

Sampling of papers is based on the listing of accepted papers at the following locations:
AAAI-14 http://www.aaai.org/Library/AAAI/aaai14contents.php
AAAI-16 http://www.aaai.org/Library/AAAI/aaai16contents.php
IJCAI-13 http://ijcai-13.org/program/accepted_papers
IJCAI-16 http://ijcai-16.org/index.php/welcome/view/accepted_papers
These listings were used to generate the files available in the ../data/ folder. Each conference is represented by a textfile containing the papers accepted to the conference's main and special tracks. Each line in the textfiles represent a paper, including its title and the authors. Example:

```
Causality based Propagation History Ranking in Social Networks  Zheng Wang, Chaokun Wang, Jish
Intervention Strategies for Increasing Engagement in Volunteer-Based Crowdsourcing  Avi Segal,
```

Papers are available through AAAI Publications for all but IJCAI-16 (at the time of writing):
AAAI-14 http://www.aaai.org/ocs/index.php/AAAI/AAAI14/schedConf/presentations
AAAI-16 http://www.aaai.org/ocs/index.php/AAAI/AAAI16/schedConf/presentations
IJCAI-13 http://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/schedConf/presentations
For IJCAI-16, see the proceedings at: http://www.ijcai.org/Proceedings/2016
First, the accepted papers are loaded from files.

In [1]:
```
from glob import glob

    accepted_papers = {}
    track_files = glob('../data/accepted*'.format(dir))
    for file in track_files:
        conference = file.split('_')[-1]
        accepted_papers[conference] = []
        with open(file, 'r') as f:
            for line in f:
                accepted_papers[conference].append(line)
```

The resulting dictionary accepted_papers contains a list of the accepted papers for each conference.

```
In [2]: for conference, papers in sorted(accepted_papers.items()):
            print('{conference} includes {papers} accepted papers.'.format(
                conference=conference, papers=len(papers)))
```

```
aaai-14 includes 398 accepted papers.
aaai-16 includes 548 accepted papers.
ijcai-13 includes 413 accepted papers.
ijcai-16 includes 551 accepted papers.
```

0.1.2 Selection

A sample population of 100 papers is selected from each conference using Python's pseudo-random number module. As per the documentation on random.sample "*The resulting list is in selection order so that all sub-slices will also be valid random samples.*" The seed is set to the unix timestamp for Jan 10 14:46:40 2017 UTC: 1484059600.

```
In [3]: import random
        random.seed(1484059600)

        k = 100
        samples = {}

        # The order is set explicitly due to originally not sorting
        # accepted_papers.items().
        conferences = ['aaai-16', 'aaai-14', 'ijcai-13', 'ijcai-16']

        for conference in conferences:
            samples[conference] = random.sample(accepted_papers[conference], k)
```

Note that when originally generating the samples, the dictionary was iterated by the use of Python 3's dict.items() view. The order is not guaranteed, and I forgot to sort the iteration so repeated runs of the code would generate the same populations. Therefore, the order has to be set explicitly as above to generate the original populations.

The generated random samples are permanently stored to files in the ../data/ directory (Github: https://github.com/sidgek/msoppgave/tree/master/data/.

```
In [4]: for conference, papers in samples.items():
            outputfile = '../data/sampled_{conference}'.format(conference=conference)
            with open(outputfile, 'w') as f:
                for line in papers:
                    f.write(line)
```

Appendix B: Survey Data

Due to the amount of columns and rows in the dataset being impractical to add to the appendix, a sample of 10 abbreviated rows from the survey data is provided here to show the format. The entire evaluation dataset is provided as a .csv file in the supplementary materials and at `https://github.com/sidgek/msoppgave`.

index	title	resea..	result..	affil..	..	evalu..	comme..	confe..
1	A Gene..	E	1	0	..	1		IJCAI 16
2	Provin..	T	1	0	..	0		IJCAI 16
3	Effici..	E	1	0	..	1		IJCAI 16
4	Natura..	E	1	0	..	1		IJCAI 16
5	Learni..	E	1	0	..	1		IJCAI 16
6	Dynami..	E	1	0	..	1		IJCAI 16
7	A Unif..	E	1	0	..	1		IJCAI 16
8	Multi-..	E	1	0	..	1		IJCAI 16
9	Change..	E	1	2	..	1		IJCAI 16
10	Model..	E	1	1	..	1		IJCAI 16

Table 6.1: Abbreviated sample of survey data.

Appendix C: Analysis Code

The following pages are generated from Jupyter notebook and document the procedure to generate the figures found in chapter 4. The notebook is also available in the supplementary materials or at `https://github.com/sidgek/msoppgave`, together with the data used.

analysis

June 11, 2017

1 Evaluation analysis

We will be taking a look at the evaluations from the data folder ../data/ (notebook, github).

1.1 Setup

Before looking at the data, a list of imports and the version of libraries used is reported.

```
In [1]: # Built-in python libraries
        import platform
        from glob import glob
        from itertools import chain

        # 3rd-party libraries
        import matplotlib
        import matplotlib.pyplot as plt
        import numpy as np
        import pandas as pd
        import IPython
        from IPython.utils.coloransi import TermColors

        # Print versions.
        print('Python version: {}'.format(platform.python_version()))
        print('IPython version: {}'.format(IPython.__version__))
        print('matplotlib version: {}'.format(matplotlib.__version__))
        print('numpy version: {}'.format(np.__version__))
        print('pandas version: {}'.format(pd.__version__))


        # Initialize the backend for Jupyter
        %matplotlib notebook

        # Set style-sheet to grayscale.
        matplotlib.style.use('ggplot')
        colormap = plt.cm.get_cmap('RdYlBu_r')
        C = [colormap(x/5) for x in range(5)]
        # Set figure font to serif.
        plt.rcParams['font.family'] = 'serif'
```

```
        # Set how many columns to show in tables.
        pd.options.display.max_columns = 50
        pd.options.display.max_rows = 400
        # Set the format to print float values to 3 decimal points.
        pd.options.display.float_format = lambda x: '%.3f' % x

Python version: 3.6.1
IPython version: 5.3.0
matplotlib version: 2.0.2
numpy version: 1.12.1
pandas version: 0.20.1
```

2 The data

First we load the CSV file into a pandas DataFrame, print the amount of samples and take a look at the column headers of the dataset.

```
In [ ]: file = '../data/evaluations.csv'

        conversion_dict = {'research_type': lambda x: int(x == 'E')}

        evaluation_data = pd.read_csv(file, sep=',', header=0, index_col=0, converters=convers:

        print('Amount of samples: {}'.format(len(evaluation_data.index)))

        column_headers = evaluation_data.columns.values
        print('\nColumn headers: {}'.format(column_headers))
```

There are 400 samples with 27 columns in total for each sample. However, some columns are not necessary for further analysis: *title*, *authors*, *link*, *comments*. The *comments* column contains short messages such as *"Points to an extended paper"* or *"Links to appendix which links to code"* to give extra information in case an evaluation is unclear. The other three identify which paper was evaluated. These columns are therefore removed from the dataframe.

```
In [ ]: evaluation_data.drop(['title', 'link', 'authors', 'comments'], axis=1, inplace=True)
        column_headers = evaluation_data.columns.values
        print('\nColumn headers: {}'.format(column_headers))
```

The remaining 23 columns can be placed in more clarifying categories. All data is boolean with the value 0 or 1, unless otherwise specified below.

Miscellaneous Variables describing the research
research_type - Experimental (1) or theoretical (0).
result_outcome - Novel research or not.
affiliation - The affiliation of the authors; academia (0), collaboration (1), industry (2).
conference - The conference the paper was accepted to.
third_party_citation - Is third-party source code or data referenced?

Research Transparency How well documented is the research method?
problem_description - The problem the research seeks to solve.
goal/objective - The objective of the research.
research_method - Research method used.
research_question - Research question(s) asked.
hypothesis - Investigated hypothesis.
prediction - Predictions related to the hypothesis.
contribution - Contribution of the research.
Note: The variables under Research Transparency are 1 if explicitly mentioned in the paper, otherwise 0.

Experiment Documentation How well is the experiment documented?
open_experiment_code - Is the experiment code available?
hardware_specification - Hardware used.
software_dependencies - For method or experiment.
experiment_setup - Is the experiment setup described with parameters etc.?
evaluation_criteria - Specification of evaluation criteria.

Method Documentation How well is the method under investigation documented?
pseudocode - Method described in pseudocode.
open_source_code - Is the method code available?

Open Data How well is the data documented, and is it available?
train - Training set specification.
validation - Validation set specification.
test - Test set specification.
results - Raw results data.
Note: If no data is open sourced all will be 0. If data is open source but the sets are not specified train or test will be set to 1 depending on whether the research requires training or not. If the research does not require training, train and validation does not have a value set.

A look at the first two samples of the dataset show the difference between experimental and theoretical papers.

```
In [ ]: evaluation_data.head(2)
```

The first sample is an experimental paper (**research_type=1**) and has values set for all the columns. The second paper, however, is a theoretical paper (**research_type=0**) and only has values set for the *Miscellaneous*, and *Research Transparency* categories, excluding the *third_part_citation* column. Note that the datafile has Experimental noted as E and theoretical noted as T.

Cells with missing values are represented as `NaN` in pandas and can be seen for all the columns exclusive to experimental papers in the second sample above. For experimental papers where training is not relevant, both the *train* and *validation* columns will show as `NaN`. To add `NaN` to visualisations below, we fill them out with the value -1.

Additionally, we split the experimental papers into a separate dataframe for plotting later.

```
In [4]: evaluation_data = evaluation_data.fillna(-1)
        experimental_data = evaluation_data[evaluation_data.research_type == 1]
```

2.1 Miscellaneous

We start with the miscellaneous category, defining the plot function which will be used for all categories. The only variable not plotted is the `conference` variable, which has its frequencies

printed out instead.

Variables describing the research
research_type - Experimental (1) or theoretical (0).
result_outcome - Novel research or not.
affiliation - The affiliation of the authors; academia (0), collaboration (1), industry (2).
conference - The conference the paper was accepted to.
third_party_citation - Is third-party source code or data referenced?

```
In [5]: def plot_full_series(series, title, labels, width=0.4):
            bins=len(labels)
            Y, X = np.histogram(series, bins=bins)
            total_Y = sum(Y)
            fig = plt.figure(figsize=(4,4))
            ax = plt.subplot(111)
            plt.bar(X[:-1], Y, color=C, width=width, axes=ax)
            ax.set_ylim(0, total_Y + 20)
            ax.set_xticks(X[:-1])
            ax.set_xticklabels(labels)
            # ax.set_title(title) Removed in favor of captions in report.

            # Add amount labels to bars
            for y, x in zip(Y, X[:-1]):
                label = '{:3.0f} ({:.1%})'.format(y, y / total_Y)
                ax.text(x, y + 5, label, ha='center', va='bottom')
            plt.show()
            fig.savefig('../doc/report/fig/{}'.format(title.replace(' ', '_')))
```

```
In [ ]: print(evaluation_data.groupby('conference').size(), end='\n\n')

        plot_full_series(evaluation_data.affiliation, 'Affiliation', ['Academia', 'Collaborati
        plot_full_series(evaluation_data.research_type, 'Research Type', ['Theoretical', 'Expe
        plot_full_series(evaluation_data.result_outcome, 'Result Outcome', ['Negative', 'Posit
        plot_full_series(experimental_data.third_party_citation, 'Third-party Citation', ['Not
```

2.2 Research Transparency

How well documented is the research method?

problem_description - The problem the research seeks to solve.
goal/objective - The objective of the research.
research_method - Research method used.
research_question - Research question(s) asked.
hypothesis - Investigated hypothesis.
prediction - Predictions related to the hypothesis.
contribution - Contribution of the research.
Note: The variables under Research Transparency are 1 if explicitly mentioned in the paper, otherwise 0.

```
In [ ]: plot_full_series(evaluation_data.contribution, 'Contribution', ['Not present', 'Present
        plot_full_series(evaluation_data['goal/objective'], 'Goal or Objective', ['Not present
```

```
plot_full_series(evaluation_data.hypothesis, 'Hypothesis', ['Not present', 'Present'])
plot_full_series(evaluation_data.prediction, 'Prediction', ['Not present', 'Present'])
plot_full_series(evaluation_data.problem_description, 'Problem Description', ['Not pres
plot_full_series(evaluation_data.research_method, 'Research Method', ['Not present', 'I
plot_full_series(evaluation_data.research_question, 'Research Question', ['Not present
```

2.3 Experiment Documentation

How well is the experiment documented?
evaluation_criteria - Specification of evaluation criteria.
experiment_setup - Is the experiment setup described with parameters etc.?
hardware_specification - Hardware used.
open_experiment_code - Is the experiment code available?
software_dependencies - For method or experiment.

```
In [ ]: plot_full_series(experimental_data.evaluation_criteria, 'Evaluation Criteria', ['False
        plot_full_series(experimental_data.experiment_setup, 'Experiment Setup', ['False', 'Tru
        plot_full_series(experimental_data.hardware_specification, 'Hardware Specification', [
        plot_full_series(experimental_data.open_experiment_code, 'Open Experiment Code', ['Fals
        plot_full_series(experimental_data.software_dependencies, 'Software Dependencies', ['Fa
```

2.4 Method Documentation

How well is the method under investigation documented?
pseudocode - Method described in pseudocode.
open_source_code - Is the method code available?

```
In [ ]: plot_full_series(experimental_data.pseudocode, 'Pseudocode', ['False', 'True'])
        plot_full_series(experimental_data.open_source_code, 'Open Source Code', ['False', 'Tru
```

2.5 Open Data

How well is the data documented, and is it available?
train - Training set specification.
validation - Validation set specification.
test - Test set specification.
results - Raw results data.

```
In [ ]: plot_full_series(experimental_data.train, 'Training Data', ['N/A', 'False', 'True'])
        plot_full_series(experimental_data.validation, 'Validation Data', ['N/A', 'False', 'Tru
        plot_full_series(experimental_data.test, 'Test Data', ['False', 'True'])
        plot_full_series(experimental_data.results, 'Results Data', ['False', 'True'])
        all_sets = experimental_data[['train', 'validation', 'test']].all(axis=1)
        plot_full_series(all_sets, 'Data_Sets', ['False', 'True'])
```

2.6 Analasys patterns

The analysis patterns will be examined for variables related to open data and source code.

2.6.1 Author affiliation

```
In [ ]: labels_of_interest = ['open_source_code', 'open_experiment_code',
                             'train', 'validation', 'test', 'results']

        for label in labels_of_interest:
            print(experimental_data.groupby('affiliation')[label].value_counts())
```

2.6.2 Conference differences

```
In [ ]: for label in labels_of_interest:
            print(experimental_data.groupby('conference')[label].value_counts())
```

2.6.3 Novelty of research

This analysis pattern has been discarded due to problems with the evaluation of Result outcome.

2.7 Reproducibility

```
In [ ]: methods_reproducible = experimental_data[labels_of_interest[0:-1]].all(axis=1)
        results_reproducible = experimental_data[labels_of_interest[0:2]].all(axis=1)

        plot_full_series(methods_reproducible, 'Methods Reproducible', ['False', 'True'])
        plot_full_series(results_reproducible, 'Results Reproducible', ['False', 'True'])
```