Abstract

Write your abstract here...

Sammendrag

Skriv sammendrag her...

Odd Erik Gundersen, Anh Thy Tran

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter introduces the research performed and its results.

1.1 Background and Motivation

1.2 Background, Motivation and Problem Outline

1.3 Research Context

The research was conducted as my Master's thesis at the department of Computer and Information Science at the Norwegian University of Science and Technology. The research task was formulated by Odd Erik Gundersen, my supervisor, and is a continuation of previous work by Gundersen (2015) presented at 3DOR2015[1].

1.4 Hypothesis, Objectives and Research Questions

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

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

RQ1 What is the state of reproducibility at AI conferences?

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

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

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

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

1.5 Research Approach

1.6 Research Contributions

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

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

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

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

1.7 Thesis Structure

Chapter 2

Reproducible Research

2.1 Terminology

2.2 Reproducible research

Reproducible research was coined as a term in Claerbout & Karrenbach (1992) as the ability to recreate published figures and results from their data, parameters and programs. Claerbout and his colleagues began publishing CD-ROMs containing the text of their books as interactive documents where figures could be rebuilt from its original data and code. With the growth of the Internet and the world-wide-web, Buckheit & Donoho (1995) began distributing a software package called WaveLab[1] as freeware. WaveLab contained the software environment necessary to reproduce figures and results from their papers, developed in part due to Jon Claerbout's recommendations for really reproducible computational research. Buckheit & Donoho (1995) emphasized the idea that code and data for the process behind a presentation of results is required with a slogan condensing Claerbout's ideas:

> *"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 set of instructions which generated the figures."*

Peng et al. (2006) propose reproducible research as a minimum standard in epidemiologic research, stating that full replication is not always feasible. They define replication as *"multiple independent investigators using independent data, analytical methods, laboratories, and instruments."* Criteria for research to be reproducible is defined as making analytical data, code underlying any principal results and the software environment available, and accompanied by adequate documentation to enable repeating the original and similar analyses. This is an important distinction,

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

as reproducible research by this definition allows *"1) verifying published findings, 2) conducting alternative analyses of the same data, 3) eliminating uninformed criticisms that do not stand up to existing data, and 4) expediting the interchange of ideas among investigators."* (Peng et al. 2006) It does not, however, address issues prior to data analysis such as study design and data generation. As discussed in Vandewalle et al. (2007), open access benchmark problems and data sets fit well in with the ideas and concepts of reproducible research. Notably, benchmarks for competitions do no allow the designers to choose the test set in a biased way, potentially addressing some of the issues not covered by reproducible research.

Stodden V. C. 2010 Reproducible research: Addressing the need for data and code sharing in computational science yale law school roundtable on data and code sharing. Drummond 2012 Reproducible Research: a Dissenting Opinion

(Dror G. Feitelson. 2015. From Repeatability to Reproducibility and Corroboration. SIGOPS Oper. Syst. Rev. 49, 1 (January 2015), 3-11. DOI: http://dx.doi.org/10.1145/2723872.2723875)

2.3 Empirical Studies into Reproducibility

Collberg et. al. (2014) Repeatability and Benefaction in Computer Systems Research http://reproducibility.cs.arizona.edu/

J. Kovacevic, "How to Encourage and Publish Reproducible Research," 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu, HI, 2007, pp. IV-1273-IV-1276. https://doi.org/10.1109/ICASSP.2007.367309 - "Reproducible research (RR) refers to the idea that in âĂIJcomputationalâĂİ sciences, the ultimate product is not a published paper but rather the entire environment used to produce the results in the paper (data, software, etc.)."

P. Vandewalle, J. Kovacevic and M. Vetterli, "Reproducible research in signal processing," in IEEE Signal Processing Magazine, vol. 26, no. 3, pp. 37-47, May 2009. https://doi.org/10.1109/MSP.2009.932122 Strict requirements, and degrees of reproducibility.

Collberg & Proebsting (2016) unable to even get code from authors for 44% out of 402 papers from ACM conferences and journals.

Victoria C. Stodden, 2010, The Scientific Method in Practice: Reproducibility in the Computational Sciences, Columbia University Academic Commons, http://hdl.handle.net/100

2.4 Observations

Reproducible research provides readers the ability to verify findings presented in a publication rather than blindly trusting the results shown. Additionally, the more detailed documentation of the research allows thorough comparisons to corroborating research and attempts to independently implement experiments and methods. Thereby facilitating further research through more robust research and further corroboration of findings. Our approach to reproducible research focus on openness

and transparency in methods and experiments, to allow review and discussion of research.

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. Considering Collberg & Proebsting (2016) were unable to receive code from authors of 44% of the 402 papers they examined, and the percentage of papers they could run source code for went from 32.3% to 48.3% after contacting the authors, keeping to already published documentation is assumed to give a good estimate while saving the time necessary to gather supplementary materials through communication with authors.

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

The conferences 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. A sample size of 100 from each conference was selected, restricting the necessary time to conduct the survey. Confidence intervals are reported in table 3.1. Probabilistic random sampling was done for each conference, as documented in Appendix A[1]. IJCAI ran biannually until the first annual instalment in 2016, while AAAI was annual prior to 2014. Thus both conferences have had one instalment between the investigated years.

For IJCAI-2013 the 58 papers from Gundersen (2015) were revisited, so only 42 of the 100 sampled papers were included. The sample generation creates a list of 100 papers, however, any sub-slice will also be valid random samples. 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.

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`)

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 conferences are vocal about open source or reproducible research, though the AAAI conferences allow non-reviewed supplemental material provided the documentation relevant for any claims is present in the paper itself. The confidence interval for each conference is reported in table 3.1, assuming that the populations are similar, a combined

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

confidence interval is reduced to 4.36%.

3.2 Evaluation Procedure

The following enumerated list, shows the procedure followed for each paper. Evaluating a single paper generally takes about 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[2] 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.

[2]The full dataset is also available in the supplemental material and at `https://github.com/sidgek/msoppgave`

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.

Evaluating whether experiment set-up is discussed can be difficult. 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 to attempt a reproduction. While such attempts would be valuable and interesting, the cost is substantial compared to the cost of this 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.

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. Following these sections, researcher errors made during the evaluations is discussed. 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.

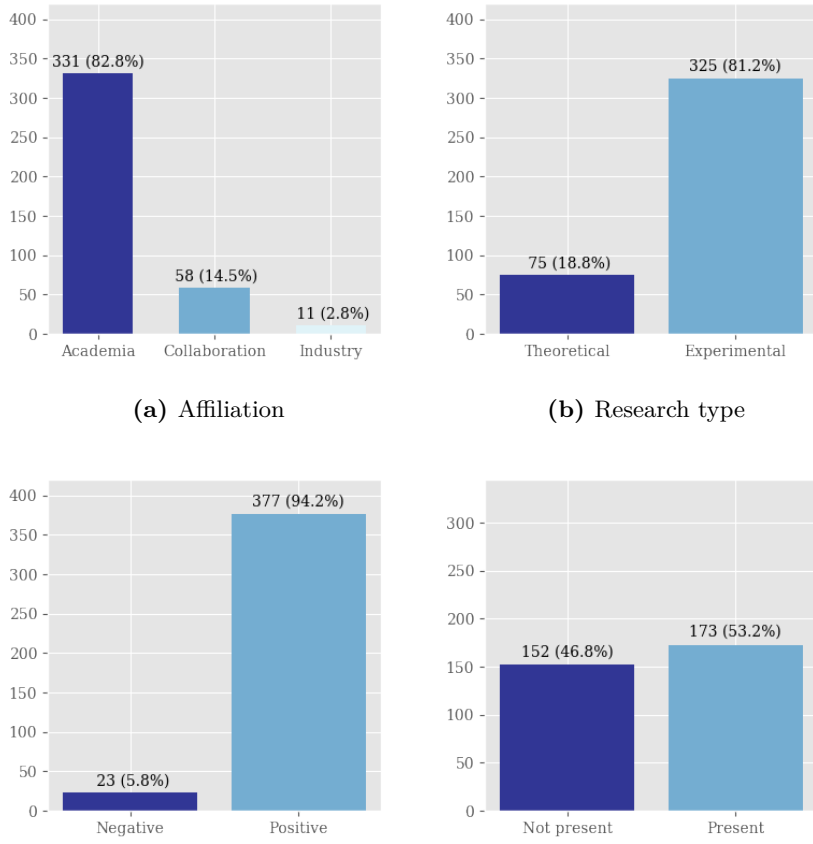
4.1 Miscellaneous

Variables in the miscellaneous category describe the research and include the following variables: 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).

Note that the Third-party citation data in figure 4.1d and Result outcome data in figure 4.1c should not be used for analysis due to researcher error as discussed in section 4.6, but is 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

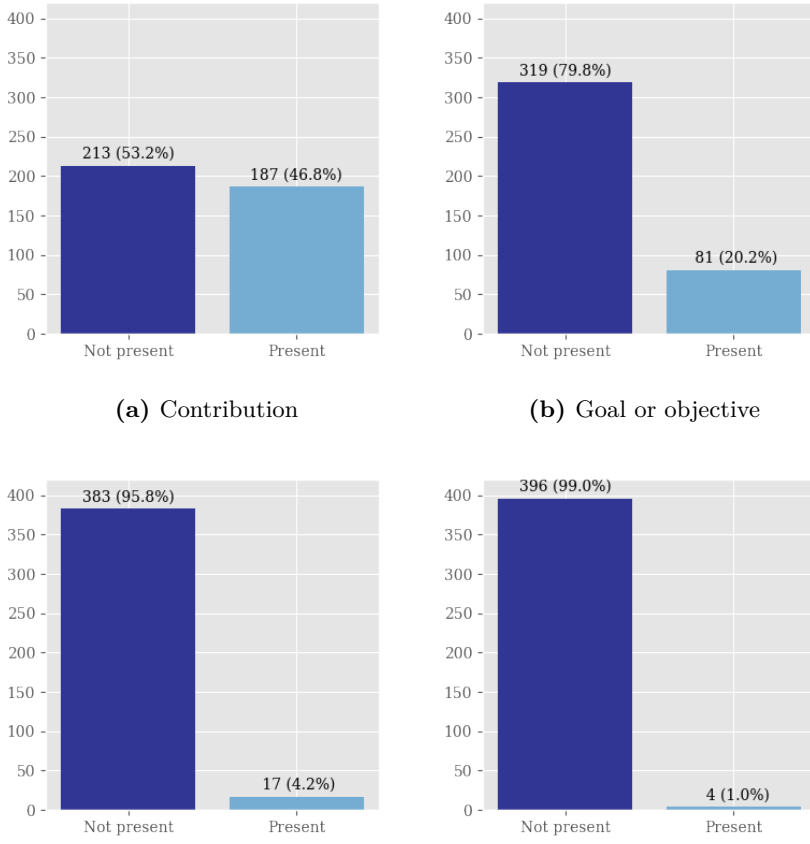
(c) Result outcome

(d) Third-party citations

Figure 4.1: Summary of miscellaneous data for all 400 papers. Note that for Third-party citation only the 325 experimental papers are relevant, which accounts for the lower values on the left axis.

4.2 Research Transparency

Research transparency variables describe how well the research method is documented. 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. The variables show little explicit mention the research background, with contribution (46.8%), problem description (46.5%), and goal/objective (20.2%) mentioned most. The remaining variables are seen in between 1 and 5 percent of the papers. This indicates that few papers present their work in line with the scientific method as used by natural sciences, but may be influenced by the terms chosen and the strict requirement of explicit mentions.



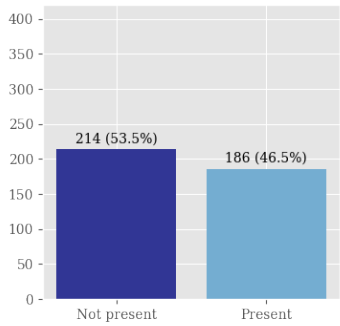
(a) Contribution

(b) Goal or objective

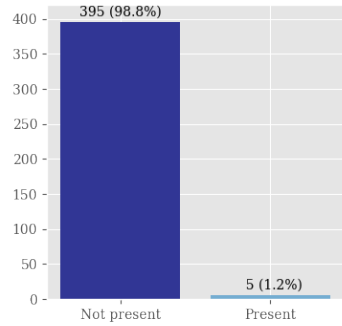
(c) Hypothesis

(d) Prediction

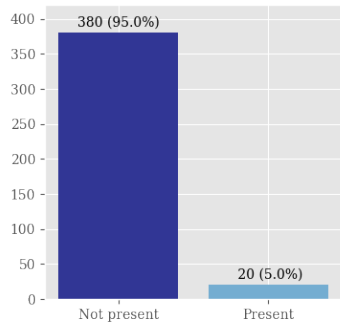
Figure 4.2: Summary of data on research transparency. A term is *Present* if it is explicitly mentioned in a paper. These variables are applicable to all 400 papers.



(a) Problem description



(b) Research method

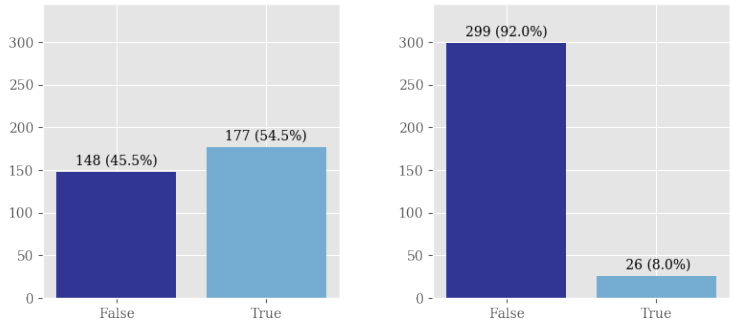


(c) Research question

Figure 4.3: Continued summary of data on research transparency. A term is *Present* if it is explicitly mentioned in a paper. These variables are applicable to all 400 papers.

4.3 Method Documentation

The method documentation category investigates the availability of the method under investigation through the pseudo-code, and open source code variables. Only the 325 experimental papers are relevant for these variables, as seen by the lower values on the left axis compared to the transparency data. The data is summarised 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. The papers without pseudo-code often include mathematical expressions to describe methods. Open source code is only seen in 26 (8%) of the papers. A few papers reference material that was not found during the evaluation or that material will be published in the future.



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

Figure 4.4: Summary of data for the method documentation category. These variables are only applicable to the 325 experimental research papers.

4.4 Experiment Documentation

Experiment documentation variables relate to how well the experiment is documented and if it 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 variables on experiments. Sharing of experiment code is a little bit lower than sharing of source code (8%). 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 is also covered in section 4.6, and covers papers that

mention and discuss hyper-parameters rather than describing the necessities of how to conduct the experiment.

4.5 Open Data

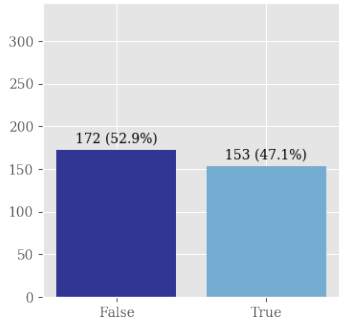
Open Data relates to the availability of data used during an experiment and documentation of dataset splits. The following variables are included: training data, validation data, test data, and results data. Figure 4.6 summarizes the results for experimental papers. Most of the papers sharing open data do so indirectly 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 specifying the type of cross-validation used was not required. Results data is rarely shared (3.7%), but occasionally bundled with the open source code.

4.6 Researcher Error

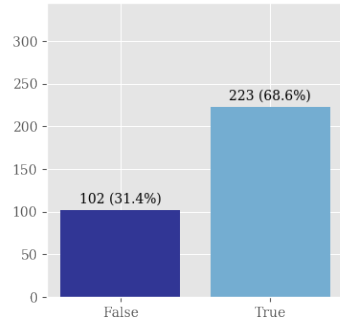
The data for Third-party citations and Result outcome have been excluded from further analysis due to inaccurate data and researcher error, respectively. Experiment set-up divulged from the original intent, becoming a measure of whether parameters used to instantiate the method and experiment is mentioned or discussed rather than a description of the experiment.

For Third-party citations, the intent was to record citations of software and data used for an experiment. This intent was to investigate if public datasets and published source code is cited when used by other researchers. For the most part, the papers noted with *Present* in figure 4.1d show correct citations to public datasets. However, it is difficult to say if a paper with a *Not present* used third-party software or data at all, or failed to cite it properly.

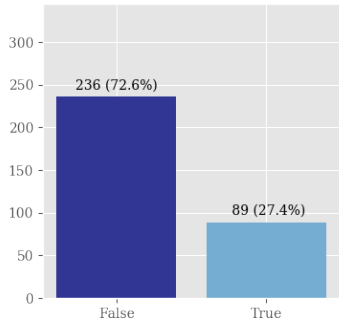
Result outcome (figure 4.1c) was erroneously recorded as a positive result, instead of a notion of the novelty of the research. This would be any paper that presents confirmation of a hypothesis, or where the wording of their findings present a solution or improvement to something. Since very few papers include a hypothesis in the first place, the data for this variable is excluded from any further analysis.



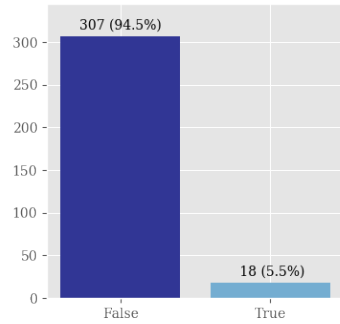
(a) Evaluation criteria



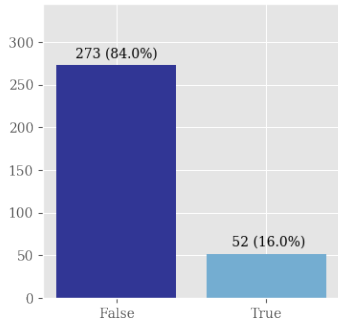
(b) Experiment set-up



(c) Hardware specification



(d) Open experiment code



(e) Software dependencies

Figure 4.5: Summary of data from the experiment documentation category. These variables are only applicable to the 325 experimental research papers.

(a) Open training data

(b) Open validation data

(c) Open test data

(d) Open results data

Figure 4.6: Summary of the open data category. These variables are only applicable to the 325 experimental research papers. Of special note is the *N/A* column in (a) and (b), which shows 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*.

4.7 Patterns of 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. The variables investigated for these relations are all variables related to open source code or open 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. There is little to suggest significant impact on sharing from these data, keeping in mind that papers affiliated with academia amount to 265, collaboration to 50 and industry to 10. The industry sample size is too small to compare with the others, while the differences in collaboration and academia is small. The largest difference can be seen in open training data, academia at 61.0% and collaboration at 45.7%, potentially due to collaborations with industry giving academic researchers access to industry data not shared publicly. This has not been investigated further, however.

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.

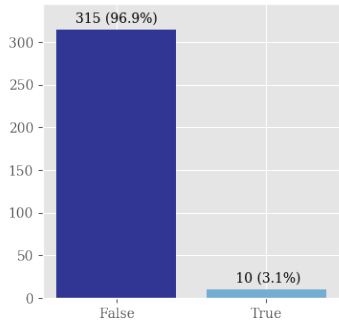
The split between conferences for experimental papers is as follows: 85 papers from AAAI 14, 85 papers from AAAI-16, 71 papers from IJCAI 13, and 84 papers from IJCAI 16. The high amount of papers not applicable to training and validation set from IJCAI 13 is likely due to the sample population involving 58 papers from the agent track, mentioned in section 3.1.2. Considering the confidence intervals for the 100 paper sample sizes for each conference are just below 9%, none of the variables in table 4.3 have detectable differences between instalments, or differences between the 2013/14 instalments compared to 2016.

Variable	AAAI 14	AAAI 16	IJCAI 13	IJCAI 16
Open source code	7 (8.2%)	9 (10.6%)	2 (2.8%)	8 (9.5%)
Open experiment code	4 (4.7%)	6 (7.1%)	0 (0%)	8 (9.5%)
Open training data	25 (51.0%)	39 (60%)	9 (42.8%)	31 (51.7%)
Open validation data	5 (10.4%)	9 (13.8%)	4 (20.0%)	12 (20.3%)
Open test data	24 (28.2%)	30 (35.3%)	13 (18.3%)	30 (35.7%)
Open results data	2 (2.4%)	2 (2.4%)	0 (0%)	8 (9.5%)

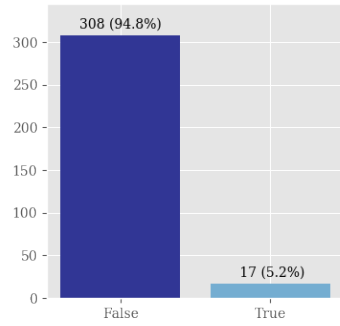
Table 4.3: Differences in adoption of open source and data based on conference instalment. The amount of experimental papers for each conference is as follows: 85 for AAAI 14, 85 for AAAI 16, 71 for IJCAI 13, and 84 for IJCAI 16. For training data and validation data, some papers are N/A: respectively 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.

4.8 Reproducibility

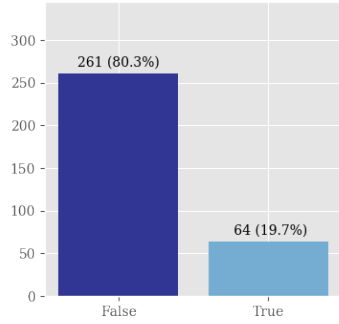
The importance of open source and data to methods and results reproducibility is shown in figure 4.7. For methods reproducibility, a paper is considered good enough if experiment and source code, as well as all data except results data is available. For results reproducibility, the data requirements are removed. 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.



(a) Methods reproducible papers



(b) Results reproducible papers



(c) Papers with open training, validation and test data sets.

Figure 4.7: The amount of experimental papers covering (a) methods and (b) results reproducibility as defined in Goodman et al. (2016). (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

Discussion

Of note: among the IJCAI 2013 papers previously examined by Gundersen (2015), two of the papers were found to have references to supplementary material which were no longer available. Suggesting that the availability and ease of use for scientific repositories for source code or data is important.

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

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

[Citation needed] attempt to run experiments from computer conferences with contact with authors.

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

Chapter 6

Conclusion

Bibliography

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

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*

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.

Peng, R. D., Dominici, F. & Zeger, S. L. (2006), 'Reproducible epidemiologic research', *American Journal of Epidemiology* **163**(9), 783.

Vandewalle, J., Suykens, J., Moor, B. D. & Lendasse, A. (2007), State-of-the-art and evolution in public data sets and competitions for system identification, time series prediction and pattern recognition, *in* '2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07', Vol. 4, pp. IV–1269–IV–1272.

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'])
```