

TOWARDS REPRODUCIBLE ML RESEARCH IN NLP

Ana Lucic, Maurits Bleeker, Samarth Bhargav, Jessica Zosa Forde, Koustuv Sinha, Jesse Dodge, Sasha Luccioni, Robert Stojnic

ACL 2022

REPRODUCIBILITY AS A TEACHING TOOL

Maurits Bleeker, Sam Bhargav

OVERVIEW

How can we mitigate the challenges without reducing the benefits?

1. Introduction to Reproducibility
2. Reproducibility in NLP
3. Mechanisms for Reproducibility
4. Reproducibility as a Teaching Tool

OVERVIEW

1. Teaching through reproducibility
2. Examples of AI courses utilizing reproducibility as a teaching tool
 - a. Reproducedpapers.org (TU Delft)
 - b. FACT-AI (University of Amsterdam)
3. Guidelines for a successful reproducibility course
4. Lessons learned

TEACHING THROUGH REPRODUCIBILITY

EXAMPLES FROM OTHER ACADEMIC FIELDS

- Learning Networking by Reproducing Research Results (Yan et al. 2017)
 - Stanford CS course on reproducing work on networking systems
- Bringing Replication Into Classroom: Benefits For Education, Science, and Society (Ribotta, Blandine, et al 2022)
 - *"For more than a decade, research in psychology has been struggling to replicate many well-known and highly cited studies"*
- How to Use Replication Assignments for Teaching Integrity in Empirical Archaeology (Marwick, Ben, et al. 2020)
 - *"Here we argue for replications as a core type of class assignment in archaeology courses"*

MOTIVATION

Valuable experience for students:

- Practice implementing and extending existing research
- Recognize the importance (and difficulty) of reproducibility
- Helps students to develop critical thinking skills
 - This also helps with writing research papers
- Can be added to their portfolio, e.g., personal website, blog post, CV
 - Allows students to participate in the community

Contribute to existing research:

- New insights can direct future research
- Results can be published, e.g., in the *ReScience journal*

REPRODUCEPAPERS.ORG


TU Delft


REPRODUCEDPAPERS.ORG

"Is an open online repository for teaching and structuring machine learning reproducibility"

- Primary motivation: there exist several venues for reproducibility but there is a 'high barrier' to entry or a focus on 'short-term' (alternate years, etc)
- Propose: a low barrier, long term venue focused on reproducibility
- Reproduction aligns with several teaching goals:
 - Reading and critiquing literature
 - Implementing, executing and extending code
 - Comparing, analyzing and presenting results in a clear and concise manner

ONLINE REPOSITORY



[Reproductions](#) [Papers](#) [Help](#) [About](#)  [Sign in](#)

Reproductions

[Submit Reproduction](#)

Reproduced

New data

New algorithm variant

Reproduction of "SwinIR: Image Restoration Using Swin Transformer"

by Frans de Boer, Jonathan Borg, Adarsh Denga, Haoran Xia

We explain the technical details of the SwinIR paper in our own words, providing ample detail to understand the authors' contribution and algorithm. Furthermore we explore modifying the architecture used in the paper to allow it to run using reduced resources and thus use less energy.

[Detail](#)

Replicated

Reproduction of "Deep Learning with Differential Privacy"

by Deep Learning CS4240 Group66: Hengkai Zhang, Dong Shen, Yuxin Cheng

Benefits of machine learning techniques based on neuron networks are widely appreciated. While these methods require a large amount of data, sensitive information should be retained. Differential privacy is thus developed. This blog aims to present and describe our efforts to reproduce "Deep... [More](#)

[Detail](#)

ONLINE REPOSITORY

- Focus of the project: partial results, minor tweaks, etc.
- Well suited for use in teaching
- Badges (self-labeled):
 - **Replicated:** A full implementation from scratch without using any pre-existing code
 - **Reproduced:** Existing code was evaluated
 - **Hyperparams check:** New evaluation of hyperparameter sensitivity
 - **New data:** Evaluating new datasets to obtain similar results
 - **New algorithm variant:** Evaluating a different variant
 - **New code variant:** Rewrote/ported existing code to be more efficient /readable
 - **Ablation study:** Additional ablation studies

COURSE DETAILS

- Part of MSc CS - Deep Learning course, TU Delft
- Teaching team selects papers with two criteria:
 - Data availability
 - Computational demands
- Projects:
 - Teams should indicate which result to reproduce
 - Groups of 2-4 students, 8 week course
 - $\frac{1}{3}$ of the course time spent on reproduction
- Deliverables:
 - Blog about the repository (private/public)
 - PDF report

24 unique papers, 57 paper reproductions

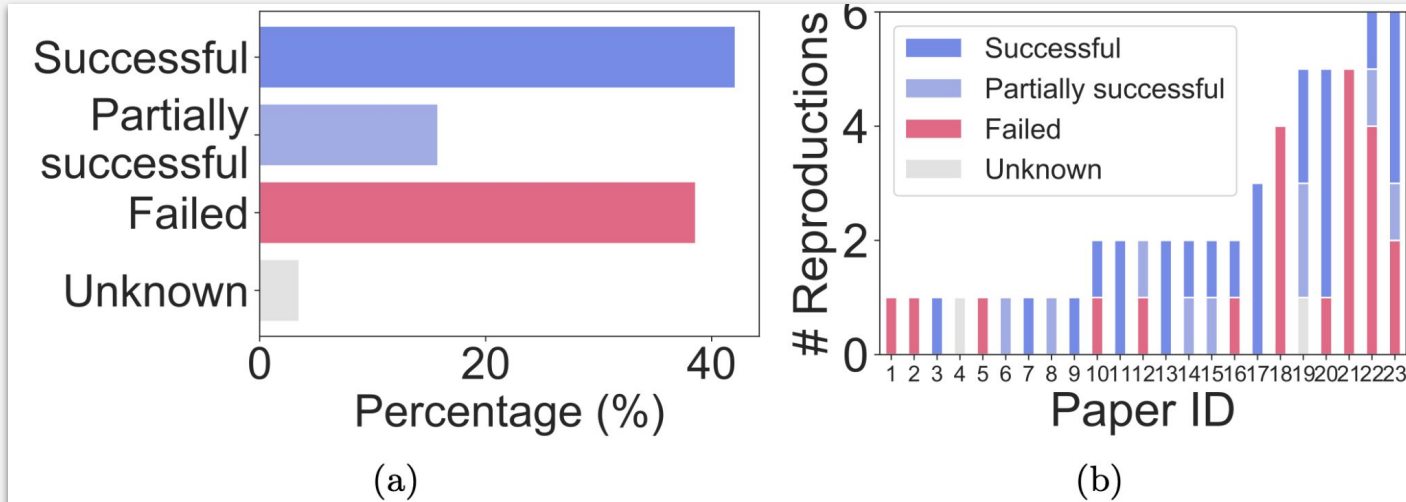
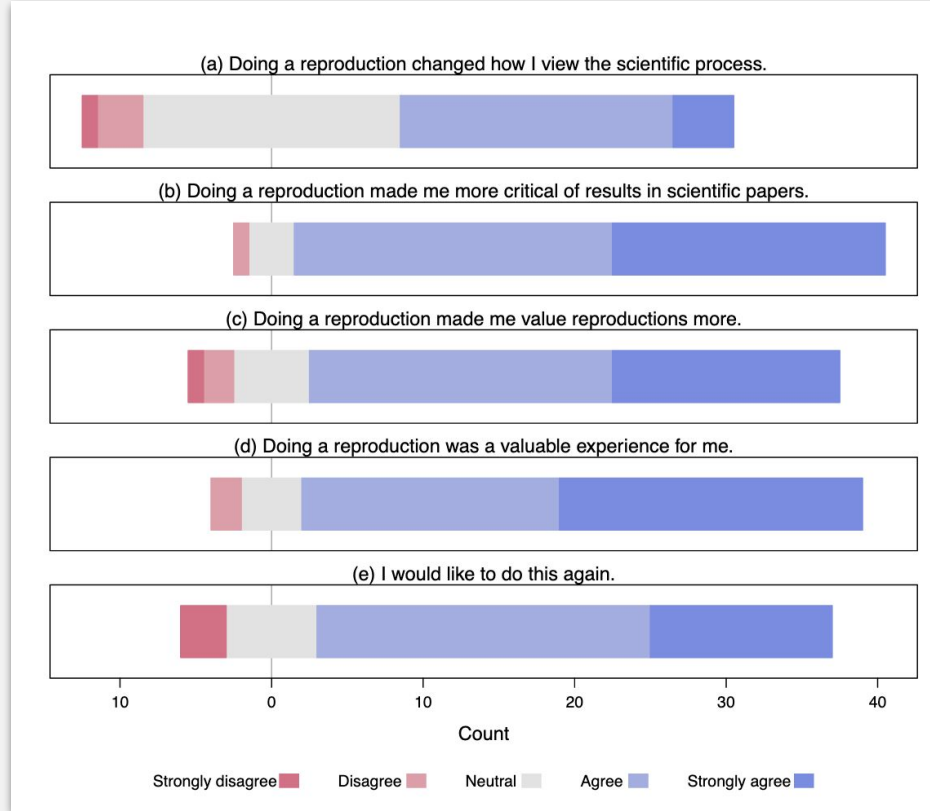
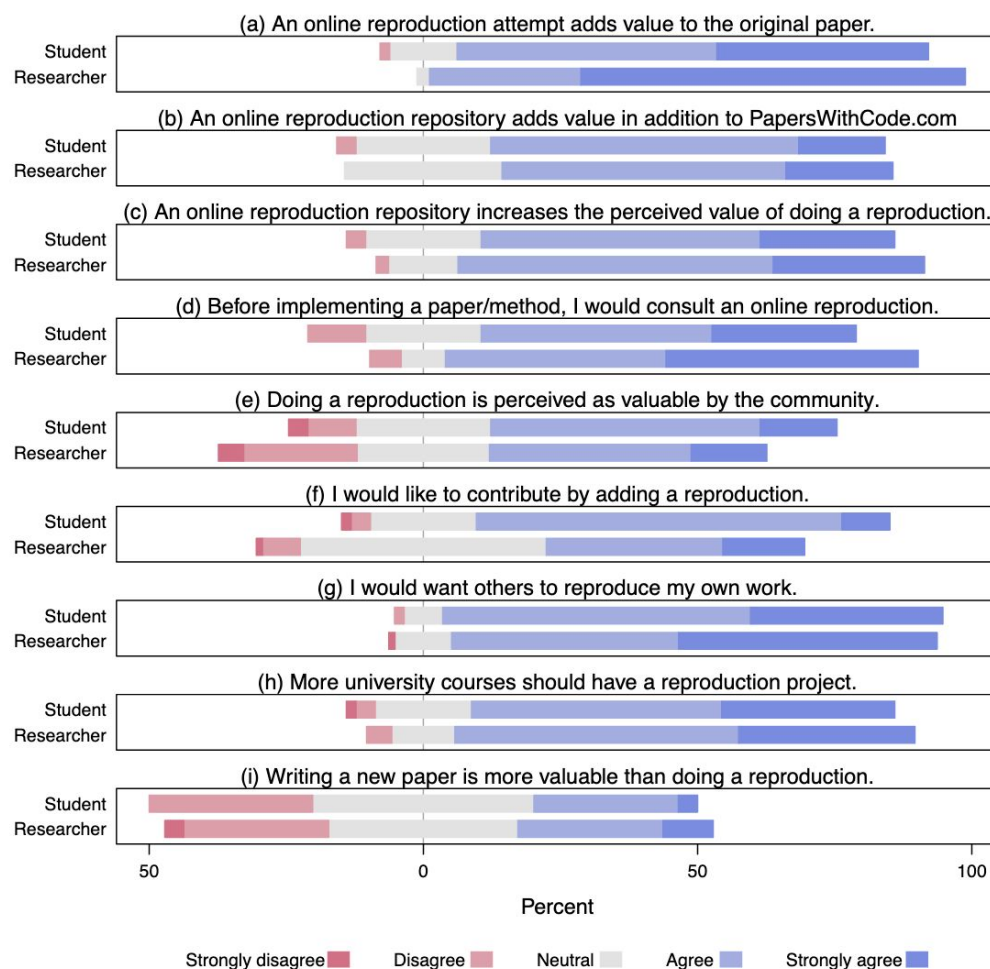


Fig. 2. Current [ReproducedPapers.org](https://reproducedpapers.org) statistics. (a) Reproduction success rates; (b) Number of reproductions per paper ID.



Student survey, N= 43



N = 144

43 course students + 14
other students

87 third-party AI
researchers

CONCLUSION

- Reproduction projects align closely with general course learning goals, and were received positively by most students
- These projects improve perceived value of reproductions, with an added incentive of publishing their work and adding to their portfolios
- *"We finally call on the community to add their reproductions to the website [ReproducedPapers.org](https://reproducedpapers.org)"*
- *"May the next generation of machine learners be reproducers"*

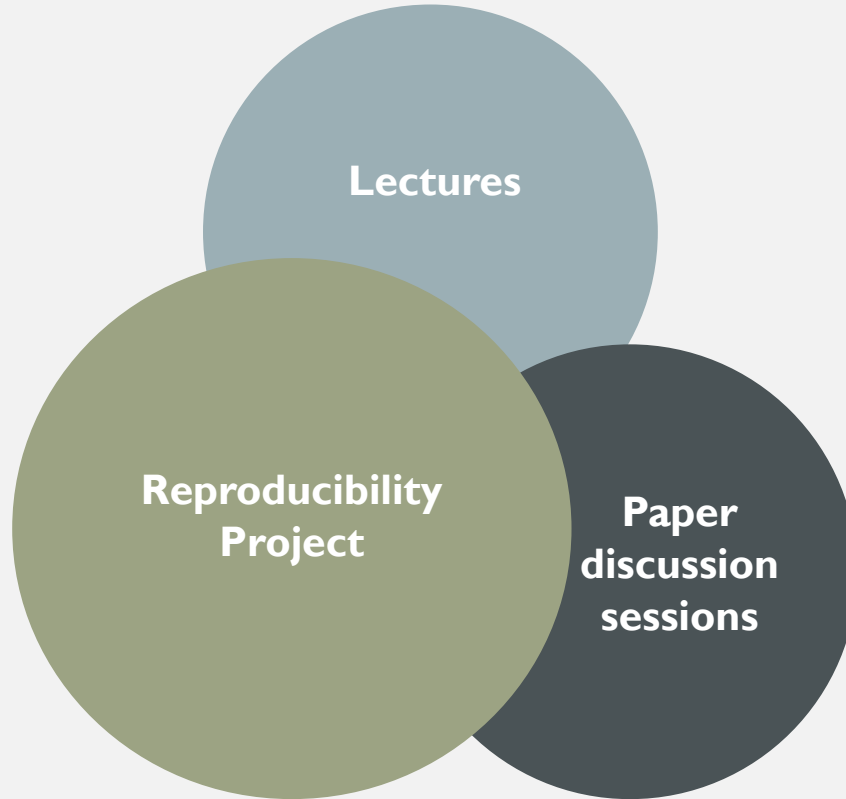
FAIRNESS, ACCOUNTABILITY, CONFIDENTIALITY, AND TRANSPARENCY IN AI COURSE

University of Amsterdam

COURSE MOTIVATION

- In 2019, we designed a new course on Fairness, Accountability, Confidentiality, and Transparency in AI (FACT-AI) at the University of Amsterdam (UvA)
 - Based on the requests of our students in the MSc AI: an increase in interest in ethical issues in AI
- The course aims to make students aware of two types of responsibility:
 - Towards society in terms of potential implications of their research
 - Similar to the NeurIPS Paper Checklist: discuss any potential negative societal impacts of your work
 - Towards the research community in terms of producing reproducible research

COURSE SETUP



LEARNING OBJECTIVES

- **LO #1:** Understanding FACT topics
- **LO #2:** Understanding algorithmic harm
- **LO #3:** Familiarity with FACT methods
- **LO #4:** Reproducing FACT solutions

LEARNING OBJECTIVES

Learning Objective #1: Understanding FACT topics

- Students can explain the major notions of fairness, accountability, confidentiality, and transparency that have been proposed in the literature, along with their strengths and weaknesses

Learning Mechanism:

- General lecture(s) per topic

LEARNING OBJECTIVES

Learning Objective #2: Understanding algorithmic harm

- Students can explain, motivate, and distinguish the main types of algorithmic harm, both in general and in terms of concrete examples where AI is being applied

Learning Mechanism:

- General lectures and guest lectures, where students can ask questions and are encouraged to participate in discussions
- This LO can be used for any AI course

LEARNING OBJECTIVES

Learning Objective #3: Familiarity with FACT methods

- Students are familiar with recent peer-reviewed algorithmic approaches in the FACT-AI literature

Learning Mechanism:

- Paper discussion sessions where students discuss a seminal FACT-AI paper in a small and interactive group, after reading the paper in advance

PAPER DISCUSSION

- Outline of how to dissect a paper ahead of time
 - Examples help!
- For the students, the goal of the paper discussion sessions is to:
 - Learn about prominent methods in the field
 - Reading a technical paper
 - Think critically about the claims made in the papers
 - Understanding a paper's strength and weaknesses
- All these (reading) skills are necessary for a good reproducibility study
 - If students can't understand the paper, how will they reimplement the algorithm?

PAPER DISCUSSION

- Students first read a seminal paper on their own trying to answer the following questions:
 - What are the main claims of the paper?
 - What are the research questions?
 - Does the experimental setup make sense, given the research questions?
 - What are the answers to the research questions? Are these supported by experimental evidence?
- Participate in small discussion sessions (ideally in person) with their peers to discuss their answers
 - Groups of 4 to 5 students

PAPER DISCUSSION

An instructor goes over the same paper, giving an overview of the papers' strengths and weaknesses

- In our case, each session was presented by a different instructor
- This to show:
 - There is no single way of examining a research paper
 - Different researchers will bring different perspectives to their assessment of papers
- We chose papers for their discussion sessions based on their impact on the FACT-AI field

LEARNING OBJECTIVES

Learning Objective #4: Reproducing FACT solutions

- Students can assess the degree to which recent algorithmic solutions are effective, especially with respect to the claims made in the original papers, while understanding their limitations and shortcomings

Learning Mechanism:

- Group project where students work in groups to reproduce FACT-AI papers from top AI conferences

GROUP PROJECT

- The group project is based on **reproducing existing algorithms** from top AI conferences and is the focal point of the course
- In our course, we focused on FACT-AI algorithms
- However, the setup for the course is not specific to FACT-AI and can be tailored to any topic
 - e.g., NLP, computer vision, information retrieval, general ML, etc.

GROUP PROJECT

- The group project is based on **reproducing existing algorithms** from top AI conferences and is the focal point of the course
- Students work in groups to reimplement existing algorithms from papers in top AI conferences (e.g., NeurIPS, ICML, ICLR, AAAI, etc).
 - FACT-AI course: groups of 3-4 students
- Students write up the results and submit reports
 - We encouraged them to submit their reports to the ML Reproducibility Challenge
- In our course, we focused on FACT-AI algorithms. However, the setup for the course is not specific to FACT-AI and can be tailored to any topic
 - e.g., NLP, computer vision, information retrieval, general ML, etc.

GROUP PROJECT

Benefits of participating in the ML Reproducibility Challenge:

- Motivates and incentivizes students
- Reports accepted by the ML Reproducibility Challenge are accepted for publication in the *ReScience* journal
- Exposes students to the paper submission cycle

GROUP PROJECT

Participating in the ML Reproducibility Challenge gives the students the opportunity to experience the whole research pipeline:

1. Reading a technical paper to understand its strength and weaknesses
2. Implementing (and perhaps also extending) the algorithms in the paper
3. Writing up the findings
4. Submitting to a venue with a deadline
5. Obtaining feedback from reviewers
6. Writing a rebuttal
7. Receiving the official acceptance/rejection notification

COURSES PARTICIPATE IN RC2021 FALL EDITION

Courses Participated in RC2021 Fall Edition

- DD2412 Deep Learning, Advanced. KTH (Royal Institute of Technology), Stockholm, Sweden
- CISC 867 Deep Learning, Queen's University, Ontario, Canada
- Special Topics in CSE: Advanced ML, Indian Institute of Technology, Gandhinagar, India
- FACT: Fairness, Accountability, Confidentiality and Transparency in AI, University of Amsterdam, Netherlands
- CSCI 662 -- Advanced Natural Language Processing, University of Southern California, USA
- Intelligent Systems and Interfaces, Indian Institute of Technology, Guwahati, India
- Intelligent Information Processing Topics, Tsinghua University, China
- Machine learning for data science 2, University of Ljubljana, Slovenia
- EECS 598-005: Randomized Numerical Linear Algebra in Machine Learning, University of Michigan, USA
- SYDE 671 - Advanced Image Processing, University of Waterloo, Canada
- BLG561E Deep Learning, Istanbul Technical University, Turkey
- CS 433 Machine Learning, EPFL, Switzerland

RESULTS OF THE ML REPRODUCIBILITY CHALLENGE

- See https://openreview.net/group?id=ML_Reproducibility_Challenge
- ML Reproducibility Challenge 2021
 - $\pm 40\%$ of the accepted papers were from the UvA FACT-AI course
- ML Reproducibility Challenge 2022
 - $\pm 50\%$ of the accepted papers were from the UvA FACT-AI course
 - Best paper award
 - 2 outstanding papers (out of 4)

FEEDBACK

First year MSc AI students

"I appreciate the critical view I have developed on papers as a result of this course. Normally I would easily accept the content of a paper, but I will be more critical from now on, as many papers are not reproducible."

"I really appreciated that this was the first course where students are judging state-of-the-art AI models. In other words, students were able to experience the scientific workfield of AI."

FEEDBACK

First year MSc AI students

"Replicating another study, seeing how (poorly) other research is performed was really eye-opening."

"I think it's really good that we get some practical insights into reproducing results from other papers, not all papers are as good as they seem to be."

QUESTIONS?

GUIDELINES FOR A SUCCESSFUL REPRODUCIBILITY COURSE

GUIDELINES FOR A SUCCESSFUL REPRODUCIBILITY COURSE

- INCLUDE A REPRODUCIBILITY LECTURE
- PAPER REQUIREMENTS
- GRADING
- TEACHING ASSISTANTS
- TIMING OF THE COURSE
- DURATION OF THE COURSE
- ADVANTAGES OF PARTICIPATING IN THE ML REPRODUCIBILITY CHALLENGE

INCLUDE A REPRODUCIBILITY LECTURE

Motivate reproducibility with a general lecture

- Position this lecture (ideally) at the beginning of the course
- Highlight papers examining reproducibility/replicability failures
 - For examples in NLP, see Part 2 of the tutorial
 - Include consequences of failure to reproduce (Part 2)
- Clearly outline scope of the project(s) and potential impact

PAPER REQUIREMENTS

- Choose 10-15 papers from the ML Reproducibility Challenge OpenReview portal that are suitable for your course
- Before the course starts, let the TAs check whether the selected papers are feasible for reproducibility study
 - Hire a team of experienced, graduate-level TAs
- Ideally assign each TA no more than 3-4 papers

PAPER REQUIREMENTS

- Select papers that are computationally feasible to reproduce
 - In our case, we were able to provide one GPU per team
 - Depends the available resources of the course and faculty
- At least one dataset should be publicly available and of a reasonable size
 - If the dataset is too big, it is an option to reproduce the work in a 'low-resource' data setting
- Select papers that are relevant to the topics covered in the course
- Emphasize the technical perspective of the sub-field
- It should be reasonable to reimplement the paper within the allotted time

GRADING

- Grading group projects on different papers in a fair manner is challenging
- Try to make the grading criteria as explicit as possible in order to make it clear for the students what is expected
- Organize a grade calibration session with the TAs after grading to align on expectations
- If participating in the ML Reproducibility Challenge, grade reports independently of the reviews

Grade	<= 5 (fail)	6 (sufficient)	7 (satisfactory)	8 (good)	9 (very good)	10 (excellent)
Project (40%)						
Project Design	Unsystematic and/or no validated use of research and design methodologies. Insufficient explanation. How are the results tested and/or verified?	Adequate use of research and design methodologies. Limited explanation.	Adequate use of research and design methodologies. Explained and justified.	Use of the right research and design methodologies. Well-explained and well justified.	Profound and critical use of research and design methodologies. Very clear and validated design.	Excellent demonstration of research and design methodologies.
Positioning of project	Project not positioned w.r.t. new literature, the FACT-field and reproducibility papers.	Project is somewhat positioned.	Project is sufficiently positioned in literature.	Project is correctly positioned in literature.	Project is well positioned within literature.	Project is integrated within literature, even from different fields/sources.
Creativity	The project does not make an original contribution. E.g. the picked paper is just said to be reproducible or not without any extra insights.	Project does not really make any original contribution. The results are reproducible, with limited effort or not reproducible with limited insights (why is this not working?).	Project team had at least one original contribution to reproduce the work and/or go beyond the original results of the paper.	Project team came up with several original ideas to reproduce the paper and/or go beyond the original results, design options and/or concepts not initiated or thought of by the supervisor.	Project team came up with many original ideas, design options and/or concepts to reproduce the work and/or go beyond the original results. Not initiated or thought of by the supervisor.	Project team surprised us all with some brilliant new ideas, design options and/or concepts, both in breadth and depth.

Code base (20%)						
Technical quality	Insufficient	Sufficient	Satisfactory	Good	Very Good	Excellent
Reproducibility of your results by the TA's.	Not reproducible. The project results should be reproducible by the TAs	N/A	With some effort the results are reproducible by the TAs.	N/A	Without any effort the results are reproducible by the TAs	N/A

Paper (30%)						
Content	Report shows no coherence of content. For example: What questions are you asking? What experiments do you run to answer them? What conclusions can you draw from these experiments?	Report shows sufficient coherence of content.	Report fulfils all requirements in terms of content.	Good report in terms of content.	Very good report in terms of content.	Excellent report in terms of content.
Form	Structure needs considerable improvement. General presentation of the content (text and figures) not very effective.	Structure needs some improvement. General presentation of the content (text and figures) is sufficient.	Structure is acceptable. General presentation of the content (text and figures) is satisfactory.	Clear structure. Good presentation of the content (text and figures).	Well-structured document. General presentation of the content (text and figures) is effective.	Very well-structured document. General presentation of the content (text and figures) is very effective.
Quality of writing	Poorly expressed. Document contains serious spelling and grammatical errors.	Reasonably expressed argumentation. Document contains some spelling and grammatical errors.	Sufficiently expressed argumentation. The document contains little spelling and grammatical errors.	Expressed and formulated well. Document has a nice flow. Document contains only minor spelling and grammatical errors.	Expressed and formulated very well. Document has a smooth flow with sufficient transitions. Document is without any spelling and grammatical errors.	Excellent expressed and formulated report. Document has a smooth flow with effective transitions. Spelling and grammatically error free.

Presentation (10%)						
Content	Presentation lacks detail and does not support conclusions. Irrelevant information presented.	Presentation lacks detail, and is just enough to support conclusions.	Presentation has sufficient detail to support conclusions.	Presentation has a good level of detail to support conclusions.	Presentation has the right level of detail to support the conclusions and to understand the recommendations.	Presentation has the perfect level of detail to support the conclusions and to understand the recommendations.
Form	Presentation is unstructured and not well organized. No (proper) use of visual aids.	Logical structure of presentation is poor. Improvements to the structure should be made. Use of visual aids can be improved.	Logical structure of presentation is reasonable but needs some improvement. Sufficient use of visual aids.	Presentation has good logical structure, the essentials are separated from the ancillary. Good use of visual aids.	Presentation has very good logical structure, the essentials are clearly separated from the ancillary. Good use of visual aids.	Presentation has excellent logical structure, the essentials are very well separated from the ancillary. Perfect use of visual aids.
Performance	Poorly expressed and formulated. Unclearly presented. Audience was ineffectively addressed.	Expression and formulation can be improved. Not always clearly presented.	Expressed and formulated adequately. Most of the time clearly presented. Audience was sufficiently addressed.	Well expressed and formulated. Clearly presented. Audience was well addressed.	Very well expressed, formulated and clearly presented.	Expressed, formulated and presented with great style, clarity and effectiveness. Audience was very well addressed and engaged.

TEACHING ASSISTANTS

- Have the TAs read the papers before the course starts to ensure they have a sufficient, in-depth understanding of their papers
 - Assign papers to TAs based on their interests
- To ease the load for the TAs, have several groups working on the same paper
- Ensure students have regular contact with their TA so no group gets stuck in the process
- Ask students halfway through the course to submit a draft report to their TAs in order to get feedback
 - We found this significantly increased the quality of the final reports

TIMING OF THE COURSE

Students need to have very strong programming skills

Table 1: The first year of the MSc AI program at the University of Amsterdam.

Course	Sem. 1	Sem. 2	EC
Computer Vision 1	■ □ □ □ □ □	□ □ □ □ □	6
Machine Learning 1	■ □ □ □ □ □	□ □ □ □ □	6
Natural Language Processing 1	□ ■ □ □ □ □	□ □ □ □ □	6
Deep Learning 1	□ ■ □ □ □ □	□ □ □ □ □	6
Fairness, Accountability, Confidentiality and Transparency in AI	□ □ ■ □ □ □	□ □ □ □ □	6
Information Retrieval 1	□ □ □ □ □ □	■ □ □ □ □	6
Knowledge Representation and Reasoning	□ □ □ □ □ □	■ □ □ □ □	6
Elective 1	□ □ □ □ □ □	□ ■ □ □ □	6
Elective 2	□ □ □ □ □ □	□ ■ □ □ □	6
Elective 3	□ □ □ □ □ □	□ □ ■ □ □	6

DURATION OF THE COURSE

- We strongly recommend to ensure that the students to have enough time to work on the project
- For our course, the students are working one month full-time on the project
 - We found this to be a beneficial setup since students didn't have to worry about any other courses during this time
- If it's not possible to work on the project full-time, then potentially adapt the weight of the course:
 - If students typically have 5 courses in one semester, consider making the reproducibility course worth 2 courses

ADVANTAGES OF PARTICIPATING IN THE ML REPRODUCIBILITY CHALLENGE

- Prioritize the ML Reproducibility Challenge by tying the reproducibility report directly to the grading
 - Students are graded on the same report that they submitted to the challenge therefore, participating is not an extra task
- Submitting to the challenge gives the students the opportunity to experience the whole research pipeline:
 - Submitting to a venue with a strict deadline
 - Obtaining feedback
 - Writing a rebuttal
 - Receiving the official notification

LESSONS LEARNED

SUMMARY OF THE LESSONS LEARNED

In our experiences, we found that the following were important components of a successful course:

- Including extension as part of reproducibility
- Having excellent teaching assistants
- Having students participate in the ML community
- Encouraging communication with the original authors

INCLUDING EXTENSIONS AS PART OF REPRODUCIBILITY

- We argue that the finding "*the original work is (not) reproducible*" is not insightful
- Require students to extend the paper if the source-code is already available
- Either extend the work to:
 - New domains, datasets or a low-resource regime (i.e., less data/compute)
 - New hyper-parameter settings or method different assumptions
 - Different model architecture
- Or explain why the work is not reproducible

INCLUDING EXTENSIONS AS PART OF REPRODUCIBILITY

There are two scenarios possible for the project:

- There already exists an open-source implementation of the selected paper. Students are allowed to use this:
 - The results the students obtain are different as described in the paper
 - The results are reproducible, meaning this method can now be used for further research
- There is no open-source implementation available, meaning the students need to reimplement everything themselves
 - Take this into account when grading

HAVING EXCELLENT TEACHING ASSISTANTS

- It is extremely important for the TAs to have **excellent programming experience** since this is the main aspect students need help with
- Have students meet with the TAs at least twice a week
- We had both second year MSc students and PhD students
 - PhD students are preferred, if possible
- Have the TAs help students with writing the rebuttal, since this is a new experience for them

HAVING EXCELLENT TEACHING ASSISTANTS

Since this is probably the first time the students are submitting a research paper, try to prevent the following common mistakes:

- Submitting single blind
- Referring to the course project in the introduction
- Motivation: "We had to do this for a course project"
- Submitting a non-anonymized code-base

HAVING STUDENTS PARTICIPATE IN THE ML COMMUNITY

- It is a motivating factor for students to create concrete output that is beneficial to the broader ML research community
- FACT-AI course 2019--2020
 - Creating a public repository with the best algorithm implementations
- FACT-AI course 2020--2021 and 2021--2022:
 - Participating in the ML Reproducibility Challenge

ENCOURAGING COMMUNICATION WITH THE ORIGINAL AUTHORS

- We strongly encourage students to contact the original authors
- It is beneficial for students to interact with scientists in the field
- It improves the papers' credibility, readability, and reproducibility
- Give the students some instructions how to do this:
 - Be aware that the authors are busy
 - Prevent that multiple teams are emailing at the same time
 - Have the TAs coordinate this

SUMMARY OF THE LESSONS LEARNED

In our experiences, we found that the following were important components of a successful course:

- Including extension as part of reproducibility
- Having excellent teaching assistants
- Having students participate in the ML community
- Encouraging communication with the original authors

QUESTIONS?

CONCLUSION

CONCLUSION

- We have shown two successful examples of graduate-level AI courses that focus on reproducibility with their course project
- We provided guidelines to successfully run a reproducibility project for any graduate-level AI course
- Implementing a course centred on a reproducibility project is fairly straightforward for the instructor and has many benefits for students
 - The course naturally "refreshes" itself every year when a new batch of papers is chosen

MOTIVATION

How can we mitigate the challenges of bigger, more complex models without reducing the benefits?

In this tutorial, we focus on the challenge of ensuring research results are reproducible

KEY TAKEAWAYS

SUMMARY OF TUTORIAL

In this tutorial, we've aimed to address the issue of **ensuring research results are reproducible**

- Part 1: We gave an introduction to reproducibility and presented some examples of (ir)reproducible results, both from within CS and from other disciplines
- Part 2: We went over some checklists in NLP as well as some examples of reproducibility research in NLP
- Part 3: We investigated existing mechanisms for reproducibility in ML/NLP such as Papers with Code and the ML Reproducibility Challenge
- Part 4: We discuss how to teach reproducibility to the next generation of AI researchers

BEST PRACTICES TO KEEP IN MIND

1. **Report** as much as much information as you can
 - Different types of papers have different requirements -- when creating a new dataset, consider the annotators! When running experiments, do a hyperparameter search!
2. **Share** dependency config files
3. **Release** code
 - If an experiment didn't work or provides evidence that doesn't support your main hypothesis (e.g., that your model is better than previous models), you should still report it!
4. **Run** multiple experiments (with different random seeds, or different data orders, etc.) and report error bars.
5. **Record** your carbon emissions
 - You can use tools like [CodeCarbon](#) or the [ML CO2 Calculator](#)
6. **Fill out** reproducibility checklists correctly, try to do any items that are appropriate (though we recognize the checklist isn't perfect)