CS229 Final Project Information

One of CS229's main goals is to prepare you to apply machine learning algorithms to real-world tasks, or to leave you well-qualified to start machine learning or AI research. The final project is intended to start you in these directions.

For group-specific questions regarding projects, please create a private post on Ed. Please first have a look through the frequently asked questions.

Note: *Only one group member is supposed to submit the assignment, and tag the rest of the group members* (do not all submit separately, or on the flip side forget to tag your teammates if you are the group's designated submitter). If you do not do this, then you can submit a regrade request and we will fix it, but *we will also deduct 1 point*.

Previous Projects

2022 (Spring)	2021 (Spring)	2020 (Spring)	2019 (Autumn)	2019 (Spring)	2018	2017
2016	2016 (Spring)	2015	2014	2013	2012	2011
2010	2009	2008	2007	2006	2005	2004

Project Topics

Your first task is to pick a project topic. If you're looking for project ideas, please come to office hours, and we'd be happy to brainstorm and suggest some project ideas.

Most students do one of three kinds of projects:

1. Application project. This is by far the most common: Pick an application that interests you, and explore how best to apply learning algorithms to solve it.

2. Algorithmic project. Pick a problem or family of problems, and develop a new learning algorithm, or a novel variant of an existing algorithm, to solve it.
3. Theoretical project. Prove some interesting/non-trivial properties of a new or an existing learning algorithm. (This is often quite difficult, and so very few, if any, projects will be purely theoretical.)

Some projects will also combine elements of applications, algorithms and theory. Many fantastic class projects come from students picking either an application area that they're interested in, or picking some subfield of machine learning that they want to explore more. So, pick something that you can get excited and passionate about! Be brave rather than timid, and do feel free to propose ambitious things that you're excited about. (Just be sure to ask us for help if you're uncertain how to best get started.) Alternatively, if you're already working on a research or industry project that machine learning might apply to, then you may already have a great project idea.

A very good CS229 project will be a publishable or nearly-publishable piece of work. Each year, some number of students continue working on their projects after completing CS229, submitting their work to conferences or journals. Thus, for inspiration, you might also look at some recent machine learning research papers. Two of the main machine learning conferences are ICML and NeurIPS. You can find papers from the recent ICML https://icml.cc/Conferences/2020/Schedule and NeurIPS conference https://neurips.cc/Conferences/2020/Schedule. Finally, looking at class projects from previous years is a good way to get ideas.

Once you have identified a topic of interest, it can be useful to look up existing research on relevant topics by searching related keywords on an academic search engine such as: http://scholar.google.com. Another important aspect of designing your project is to identify one or several datasets suitable for your topic of interest. If that data needs considerable pre-processing to suit your task, or if you intend to collect the needed data yourself, keep in mind that this is only one part of the expected project work, but can often take considerable time. We still expect a solid methodology and discussion of results, so pace your project accordingly.

Notes on a few specific types of projects:

- Deep learning projects: Since CS229 discusses many other concepts besides deep learning, we ask that if you decide to work on a deep learning project, please make sure that you use other material you learned in the class as well. For example, you might set up logistic regression and SVM baselines, or do some data analysis using the unsupervised methods covered in class. We prioritize

methodological and experimental rigor and standardize our grading, so note that pursuing (or not pursuing) a deep learning project will not itself impact your grade. Finally, training deep learning models can be very time consuming, so make sure you have the necessary computing resources.

- Preprocessed datasets: While we don't want you to have to spend much time collecting raw data, the process of inspecting and visualizing the data, trying out different types of preprocessing, and doing error analysis is often an important part of machine learning. Hence if you choose to use preprepared datasets (e.g. from Kaggle, the UCI machine learning repository, etc.) we encourage you to do some data exploration and analysis to get familiar with the problem.
- Replicating results: Replicating the results in a paper can be a good way to learn. However, we ask that instead of just replicating a paper, also try using the technique on another application, or do some analysis of how each component of the model contributes to final performance. In other words, your project does not need to be completely novel, but should not just duplicate previous work done by others.

Project Parts: Proposal, Milestone, Poster, and Final Report

This section contains the detailed instructions for the different parts of your project.

Submission: We'll be using Gradescope for submission of all four parts of the final project. We'll announce when submissions are open for each part. You should submit on Gradescope as a group: that is, for each part, please make one submission for your entire project group and tag your team members.

Evaluation

We will not be disclosing the breakdown of the 40% that the final project is worth amongst the different parts, but the poster and final report will combine to be the majority of the grade. Projects will be evaluated based on:

- The technical quality of the work. (I.e., Does the technical material make sense? Are the things tried reasonable? Are the proposed algorithms or applications clever and interesting? Do the authors convey novel insight about the problem and/or algorithms? Does the project have sufficient scope for the given team size?)

- Originality. (Did the authors add their own data processing, methods, or analysis? Does the final project avoid being a mirror image of existing papers/projects with no net new work?)
- Communication. (Are the authors able to clearly and effectively explain the work that they did, including context, methods, and results? Do the paper and poster balance clarity with rigor?)

In order to highlight these components, it is important you present a solid discussion regarding the learnings from the development of your method, and summarizing how your work compares to existing approaches.

Project Proposals

In the project proposal, you'll pick a project idea to work on early and receive feedback from the TAs.

In the proposal, below your project title, include the project category. The category can be one of:

- Athletics & Sensing Devices
- Audio & Music
- Computer Vision
- Finance & Commerce
- General Machine Learning
- Life Sciences
- Natural Language
- Physical Sciences
- Theory & Reinforcement Learning

Project Mentors Based on the topic you choose in your proposal, we'll suggest a project mentor given the areas of expertise of the TAs. This is just a recommendation; feel free to speak with other TAs as well.

Format	Your proposal should be a PDF document, giving the title of the project, the project category, the full names of all of your team members, the SUNet ID of your team members, and a 300-500 word description of what you plan to do. Your project proposal should include the following information: ● Motivation: What problem are you tackling? Is this an application or a theoretical result? ● Method: What machine learning techniques are you planning to apply or improve upon? ● Intended experiments: What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm? Presenting pointers to one relevant dataset and one example of prior research on the topic are a valuable (optional) addition.
Grading	The project proposal is mainly intended to make sure you decide on a project topic and get feedback from TAs early. As long as your proposal follows the instructions above and the project seems to have been thought out with a reasonable plan, you should do well on the proposal.

Milestone

The milestone will help you make sure you're on track, and should describe what you've accomplished so far, and very briefly say what else you plan to do. You should write it as if it's an “early draft" of what will turn into your final project. You can write it as if you're writing the first few pages of your final project report, so that you can re-use most of the milestone text in your final report. Please write the milestone (and final report) keeping in mind that the intended audience are the instructors and the TAs. Thus, for example, you should not spend two pages explaining what logistic regression is. Your milestone should include the full names of all your team members and state the full title of your project. Note: We will expect your final writeup to be on the same topic as your milestone.

Contributions	Please include a section that describes what each team member worked on and contributed to the project. This is to make sure team members are carrying a fair share of the work for projects. If you have any concerns working with one of your project teammates, please create a private Ed post.
Grading	The milestone is mostly intended to get feedback from TAs to make sure you're making reasonable progress. As long as your milestone follows the instructions above and you seem to have tested any assumptions which might prevent your team from completing the project, you should do well on the milestone.
Format	Your milestone should be at most 3 pages, excluding references. Similar to the proposal, it should include • Motivation: What problem are you tackling, and what's the setting you're considering? • Method: What machine learning techniques have you tried and why? • Preliminary experiments: Describe the experiments that you've run, the outcomes, and any error analysis that you've done. You should have tried at least one baseline. • Next steps: Given your preliminary results, what are the next steps that you're considering?

Poster

Format	Here are some poster guidelines (please note that 36x24in means 36in wide by 24in tall, i.e. it's better if your poster is formatted landscape). You can also look at posters from previous years. Note: Despite some examples given in the guidelines, posters with nice, illustrative figures are preferred over posters with lots of text.

Final Writeup

We know that most students work very hard on the final projects, and so we are extremely careful to give each writeup ample attention, and read and try very hard to understand everything you describe in it.

After the class, we will also post all the final writeups online so that you can read about each other's work. If you do not want your write-up to be posted online, then please create a private Ed post at least a week in advance of the final submission deadline.

Format Final project writeups can be at most 5 pages long (including appendices and figures). We will allow for extra pages containing only references. If you did this work in collaboration with someone else, or if someone else (such as another professor) had advised you on this work, your write-up must fully acknowledge their contributions. For shared projects, we also require that you submit the final report from the class you're sharing the project with.

Here's more detailed guidelines with a rough outline of what we expect to see in the final report: final-report-guidelines.pdf.

Contributions Please include a section that briefly describes what each team member worked on and contributed to the project. If you have any concerns working with one of your project teammates, please create a private Ed post. We may reach out and factor in contributions and evaluations when assigning project grades.

Code Please include a link to a Github repository or zip file with the code for your final project. You do not have to include the data or additional libraries (so if you submit a zip file, it should not exceed 5MB).

Grading	The final report will be judged based off of the clarity of the report, the relevance of the project to topics taught in CS229, the novelty of the problem, and the technical quality and significance of the work.

After CS229

After CS229, if you want to submit your work to a machine learning conference, the ICML deadline will probably be in early February next year (http://icml.cc), and the NeurIPS deadline is usually in late May (http://neurips.cc/). Of course, depending on the topic of your project, other non-machine learning conferences may also be more appropriate.

Project FAQs

1. **What are the deliverables as part of the term project?**

   The project has four deliverables:

   a. Proposal
   b. Milestone
   c. Poster
   d. Final report

   Please see the Logistics doc on the home page for due dates and deadlines.

2. **Should the final project use only methods taught in the class?**

   No, we don't restrict you to only use methods/topics/problems taught in class. That said, you can always consult a TA if you are unsure about any method or problem statement.

3. **Is it okay to use a dataset that is not public ?**

We don't mind you using a dataset that is not public, as long as you have the required permissions to use it. We don't require you to share the dataset either as long as you can accurately describe it in the Final Report.

4. **Can I use datasets on Kaggle?**

   Using datasets on Kaggle is allowed. However, the project needs to focus on model performance and achieve a high leaderboard score to receive high grades. This is because a significant amount of work is needed to formulate the problem, obtain data and preprocess data, whereas Kaggle challenges provide you well-defined problems and organized datasets at the start.

5. **Is it okay to combine the CS229 term project with that of another class ?**

   In general it is possible to combine your project for CS229 and another class, but with the following caveats:

   a. You should make sure that you follow all the guidelines and requirements for the CS229 project (in addition to the requirements of the other class). So, if you'd like to combine your CS229 project with a class X but class X's policies don't allow for it, you cannot do it.
   b. You cannot turn in an identical project for both classes, but you can share common infrastructure/code base/datasets across the two classes.
   c. In your milestone and final report, clearly indicate which part of the project is done for CS229 and which part is done for a class other than CS229. For shared projects, we also require that you submit the final report from the class you're sharing the project with.

6. **Is it okay for CS229 project to be part of my on-going research?**

   The CS229 project has to be new, and you cannot use existing work or research for the CS229 project. However, the project can depend on existing research that you have done, as long as the CS229 project is new work that you are doing for the class and sufficiently self-contained, but in this case, you must clearly state in your milestone and final report what part of the project was done before CS229 and for CS229. You can however, use the CS229 project for publications of your main research.

7. **Do all team members need to be enrolled in CS229?**

No, but please explicitly state the work which was done by team members enrolled in CS229 in your milestone and final report. This extends to projects that were done in collaboration with research groups as well. We generally don't encourage you to collaborate with non-Stanford people for the course project due to potential IP implications (Stanford owns the IP for all technology that's developed as a result of course projects). If you choose to collaborate with a company, please understand you will need to follow the IP policy here.

8. **What are acceptable team sizes and how does grading differ as a function of the team size ?**

   We recommend teams of 3 students, while team sizes of 1 or 2 are also acceptable. The team size will be taken under consideration when evaluating the scope of the project in breadth and depth, meaning that a three-person team is expected to accomplish more than a one-person team would.
   The reason we encourage students to form teams of 3 is that, in our experience, this size usually fits best the expectations for the CS229 projects. In particular, we expect the team to submit a completed project (even for team of 1 or 2), so keep in mind that all projects require you to spend a decent minimum effort towards gathering data, and setting up the infrastructure to reach some form of result. In a three-person team this can be shared much better, allowing the team to focus a lot more on the interesting stuff, e.g. results and discussion.
   In exceptional cases, we can allow a team of 4 people. If you plan to work on a project in a team of 4, please come talk to one of the TAs beforehand so we can ensure that the project has a large enough scope. You must submit a written proposal for a 4-person project to the head TA, which has to be approved.

9. **Do I have to be on campus to submit the final report?**

   No, the final report will be submitted via Gradescope.

10. **What fraction of the final grade is the project?**

    The term project is 40% of the final grade.

11. **What is the late day policy for group project?**

    Students can use late days for both the proposal and milestone (though note that these late days apply to each member of the group). Also note that while late days are allowed, we do not recommend using them since the assignments are more time consuming than writing up the proposal and milestone (and worth

more points). Note that there are no late days for the final poster and paper to provide enough leeway for grading.

12. **Does the team have to be all SCPD students or all on-campus students?**

    A team can have both on-campus and SCPD students.

13. **Can we use some Machine Learning libraries such as scikit-learn or are we expected to implement them from scratch?**

    You can use any library for the project.

14. **Is it ok to use a public repository for version control?**

    We recommend that students keep their repo private while working on the project. After the class ends, we understand that many students may want to have their work be public, so that they can point to it for interviews, outside advisors, etc, which is acceptable.

15. **What if two teams end up working on the same project?**

    It is okay if two teams end up working on the same project as long as they don't coordinate to do so, in order to not be biased in the way they tackle the problem. Alternatively the teams can coordinate to make sure they work on different problems.

16. **Will we be provided any cloud compute resource credit?**

    We are looking into getting cloud credit for the projects. We will announce on Ed once this is finalized. Also check out Google Colab for free GPU resources.

17. **Are we required to use Python for the project?**

    Any programming language is allowed for the project.

17. **Must I attend the poster session?**

    Attendance at the poster session is expected of every student unless a prior arrangement has been made with a legitimate reason (i.e. internships starting or the entire group is composed of SCPD students). Please post on Ed if you'd like to request an exception. The reason for this is that we'd like the poster session to

be interactive rather than just a grading session; ensuring students can feel comfortable leaving their poster to view other posters is critical for this!