# CSC413 Final Project

The project in this course is an opportunity to develop a sequence-based deep learning application in an area of your own choosing. It also provides the chance to complete a deep learning project that is much closer to a real-world application.

While this project has some structure, you will be required to deal with the ambiguity and significant decision making that make up the life of a deep learning practitioner.

## Logistics

Projects must be done in groups of 3-4. Please form groups on Markus by March 14, 10pm. Exceptions to this rule can be made only in rare cases provided there is good reason to do so. Email the instructor if this applies to you. If you do not know anyone in class feel free to post a message on Piazza. We will also set aside some time during the tutorial for students who are looking for collaborators to find each other and discuss forming a group.

A 1-2 page project proposal is due March ~~18~~ 21, 10pm. You will also be asked to summarize the data set that you are using for this proposal.

Each team will submit a github repository page that describes the deep learning model built in the project. The repository should also contain the code that you wrote.

## Project Requirements

Your project must either take a sequence (of variable length) as an input, or produce a sequence as an output, or both. Your model should thus usually involve an RNN or a Transformer component. Here are some examples of possible projects:

- Using an RNN (or transformer) to classify sequences (e.g. whether a restaurant review is positive or negative)
- Using a generative RNN to produce sequences (e.g. South Park TV scripts)
- Using a Siamese network to determine whether two StackOverflow questions are duplicates
- Predict the next item in a sequence (e.g. Stock market)
- Predict the outcome of a patient based on some sequential factors

Before choosing a project, consider whether there is data available for you. Since the project deadline is about a month away, consider tailoring your project ideas to what data is available to you.

You are encouraged to use transfer learning and data augmentation ideas in your project.

You can use deep learning packages (e.g. pytorch, huggingface). However, you should be able to explain the steps involved in the forward pass computation of your model.

## Project Proposal and Data Collection Report (3%)

A 1-2 page project proposal is due March ~~18~~ 21, 10pm. Please use 12-point font and standard margins. You will also be asked to summarize the data set that you are using for this proposal.

The proposal should:

- Clearly describe the task that your model will perform. (2pt)
  - 2/2 for clearly describing the task using standard deep learning terminology
  - 1.5/2 for describing the task in a way that is understandable to the grader, but that uses non-standard terminology
  - 1/2 for describing the task generally (e.g. "sequence classification" without stating the exact classes)
  - 0/2 for a proposal that does not align with the project requirements
- Clearly describe the model that you intend to use (2pt)
  - 2/2 for clearly describing the model using standard deep learning terminology; the grader can picture exactly how the model could be used.
  - 1.5/2 for describing the task in a way that is understandable to the grader, but that uses non-standard terminology
  - 1/2 for describing the models generally (e.g. sequence-to-sequence model, without describing which ones)
  - 0/2 for a model that does not align with the project requirements

- Outline the data set that you intend to use, and provide some statistics about the amount/type of data that is available (4pt)
  - 1 point for convincing the grader that you are able to acquire the data that you need (with the appropriate license/permission for educational use)
  - 1 point for convincing the grader that the type and amount of data is sufficient (e.g. via summary statistics, examples data set)
  - 2 points for convincing the grader that you have explored the data, and considered information about your data relevant to your model (like in A1 Q1)
- Discuss any ethical implications of your model—how might the use (or misuse) of this model help or hurt people? (2pt)
  - 2/2 For a thoughtful discussion that considers the ethical implications across many groups of people (that different groups may be impacted differently).
  - 1/2 For a discussion that is generic, or considers the ethical implications for only one group of people.
- Describe how work will be divided amongst the team members. We recommend pair-coding for parts of the project, but consider the work that it might take to load/format your data, write a first model, "overfit" to a single data point, etc... (2pt)
  - 2/2 The description provides enough detail so that if a team member is replaced, they know exactly what their responsibilities will be.
  - 1/2 There is clearly an attempt to describe the division of tasks, but the communication is unclear and/or only the tasks listed above are assigned.
  - 0/2 Only vague assertions are made (e.g. "we will divide the work equally", "everyone will work on everything", or "we will determine who will work on what as the project progresses).
- Proper formatting (2pt)
  - 2/2 Proposal is 1-2 pages. The proposal is formatted so that readers can find specific information quickly (e.g. via the use of paragraphs and topic sentences)
  - 1/2 Proposal is slightly over the length limit. There was clearly an attempt to format the proposal, but information is still scattered in various places.
  - 0/2 Proposal runs extremely long. It is difficult to understand the structure of the proposal.

## Project Github Repository (25%)

The project github repository is due at the end of term April 8, 10pm. The repository should be either public, or privately viewable to the instructors and TAs. We recommend a public repository to showcase the work that you are able to do!

The repository should have a README file with the following component:

- Introduction that states the deep learning model that you are building
- Model:
  - A figure/diagram of the model architecture that demonstrates understanding of the steps involved in computing the forward pass
  - Count the number of parameters in the model, and a description of where the parameters come from
  - Examples of how the model performs on two actual examples from the test set: one successful and one unsuccessful
- Data:
  - Describe the source of your data
  - Provide summary statistics of your data to help interpret your results (similar to in the proposal)
  - Describe how you transformed the data (e.g. any data augmentation techniques)
  - If appropriate to your project, describe how the train/validation/test set was split. (Note that splitting the training/validation/test set is not always straightforward!)
- Training:
  - The training curve of your final model
  - A description how you tuned hyper-parameters
- Results:
  - Describe the quantitative measure that you are using to evaluate your result
  - Describe the quantitative and qualitative results
  - A justification that your implemented method performed reasonably, given the difficulty of the problem—or a hypothesis for why it doesn't (this is extremely important)

- Ethical Consideration:
  - Description of a use of the system that could give rise to ethical issues. Are there limitations of your model? Your training data?
- Authors
  - A description of how the work was split—i.e. who did what in this project.

The github repository will be graded based on:

- 70% - The quality of your README (i.e. you written report), including ~20% (TBD) for the justification that your implemented method performs reasonably—or a hypothesis for why it doesn't
- 20% - The quality of your code/documentation (i.e. whether the TA can generally understand what your code does, how it is organized, and where to find specific settings)
- 10% - The "difficulty" of your project. You can get the full credit by having a project that involves at least one of the following:
  - Data Augmentation
  - Transformer
  - Generative Model (e.g. that uses teacher-forcing)
  - Sequence-to-Sequence Architecture
  - Other advanced concepts—please contact an instructor