

# Final Project Guidelines

Your class project is an opportunity for you to explore an interesting machine learning problem of your choice in the context of a real-world data set. Below, you will find some datasets, and you are encouraged to propose your own idea.

- Pick a real-world problem, describe your dataset, define your machine learning problem, apply a machine learning technique, and present the results. You are encouraged (not required) to explore a large-scale dataset, i.e., over 100K data instances.
- Projects can be done individually, or in teams of two students. For a group, group members are responsible for dividing up the work equally and making sure that each member contributes.
- You are allowed to use any ML libraries and Generative AI such as ChatGPT. Please clearly disclose what libraries, open source code, or Generative AI are used and how they are used in the report.
- If you are having trouble writing a proposal or executing the project, please feel free to consult with the instructor ([yfang@scu.edu](mailto:yfang@scu.edu)).

Your project will be worth 20% of your final class grade, and will have three deliverables:

- Proposal, Nov. 1, Week 6, Friday, by 1pm to Camino: 1 page (20%)
- Presentation, Dec. 6, Week 10, Friday (30%)
- Final Report, Dec. 13, Week 11 (exam week), Friday, by noon (50%)

Your project will be evaluated based on several factors including:

- The extensiveness of the study and experiments. A project that involves well-designed experiments with competitive baselines and thorough analysis of the experimental results are scored highly.
- The writing and the clarity of the final report.
- The novelty of the approach would be a plus.

## Project Proposal (submitted on Camino)

Include the following information:

- Project title
- Data set
- Project idea
- The approach you will use
- Software you will use
- References.
- Teammate (if any).
- Timeline.

## Project Report

The following is a suggested structure for the report:

- **Title, Author(s)**
- **Abstract:** It should not be more than 300 words.
- **Introduction:** This section introduces the problem, and your overall approach to the problem.
- **Background/Related Work:** This section discusses relevant literature for your project.
- **Approach:** This section details your approach to the problem. For example, this is the section where you would describe the architecture of your neural network. You should be specific - you may want to include equations, figures, plots, etc.
- **Experiments:** In this section, you describe:
  - The dataset(s) you used
  - How you ran your experiments (e.g. model configurations, learning rate, training time, etc.)
  - The evaluation metric(s) you used
  - Your results. It is very important to show both quantitative evaluation (show numbers, figures, tables etc. relating to your evaluation metric(s)) and qualitative evaluation (show example results, etc.).
- **Conclusion:** What have you learned? Suggest future ideas.
- **References:** Include references to all literature that informed your project work. This is absolutely necessary

## Submission

Your proposal is expected to be 1 page (in any template) in PDF submitted to Camino.

Your final report is expected to be between 4 - 8 pages using the [ACL template](#). Here are two things you need to turn in on Camino:

- Final project report in PDF.
- Source code as a zip file.  
(optional) Cool videos, interactive visualizations, demos, etc.

## Suggestions of Project Ideas:

## Useful resources of datasets:

Kaggle: <https://www.kaggle.com/datasets>

UCI Repository: <https://archive.ics.uci.edu/datasets>

## Text

---

- **Autonomous Tagging of StackOverflow Questions**
  - Make a multi-label classification system that automatically assigns tags for questions posted on a forum such as StackOverflow or Quora.
  - Dataset: [StackLite](#) or [10% sample](#)
- **Keyword/Concept identification**
  - Identify keywords from millions of questions
  - Dataset: [StackOverflow question samples by Facebook](#)
- **Topic identification**
  - Multi-label classification of printed media articles to topics
  - Dataset: [Greek Media monitoring multi-label classification](#)

## Natural Language Understanding

- **Automated essay grading**
  - The purpose of this project is to implement and train machine learning algorithms to automatically assess and grade essay responses.
  - Dataset: [Essays with human graded scores](#)
- **Sentence to Sentence semantic similarity**
  - Can you identify question pairs that have the same intent or meaning?
  - Dataset: [Quora question pairs](#) with similar questions marked
- **Fight online abuse**
  - Can you confidently and accurately tell whether a particular comment is abusive?
  - Dataset: [Toxic comments on Kaggle](#)

- **Open Domain question answering**

- Can you build a bot which answers questions according to the student's age or her curriculum?
- [Facebook's FAIR](#) is built in a similar way for Wikipedia.
- Dataset: [NCERT books](#) for K-12/school students in India, [NarrativeQA by Google DeepMind](#) and [SQuAD by Stanford](#)

- **Social Chat/Conversational Bots**

- Can you build a bot which talks to you just like people talk on social networking sites?
- Reference: [Chat-bot architecture](#)
- Dataset: [Reddit Dataset](#)

- **Automatic text summarization**

- Can you create a summary with the major points of the original document?
- Abstractive (write your own summary) and Extractive (select pieces of text from original) are two popular approaches
- Dataset: [CNN and DailyMail News Pieces](#) by Google DeepMind

- **Copy-cat Bot**

- Generate plausible new text which looks like some other text
- Obama Speeches? For instance, you can create a bot which writes some [new speeches in Obama's style](#)
- Trump Bot? Or a Twitter bot which mimics [@realDonaldTrump](#)
- Narendra Modi bot saying "*doston*"? Start by scrapping off his *Hindi* speeches from his [personal website](#)
- Example Dataset: [English Transcript of Modi speeches](#)

Check [mlm/blog](#) for some hints.

- **Sentiment Analysis**

- Do Twitter Sentiment Analysis on tweets sorted by geography and timestamp.
- Dataset: [Tweets sentiment tagged by humans](#)

- **De-anonymization**

- Can you classify the text of an e-mail message to decide who sent it?
- Dataset: [150,000 Enron emails](#)

## Forecasting

---

- **Univariate Time Series Forecasting**
  - How much will it rain this year?
  - Dataset: [45 years of rainfall data](#)
- **Multi-variate Time Series Forecasting**
  - How polluted will your town's air be? Pollution Level Forecasting
  - Dataset: [Air Quality dataset](#)
- **Demand/load forecasting**
  - Find a short term forecast on electricity consumption of a single home
  - Dataset: [Electricity consumption of a household](#)
- **Predict Blood Donation**
  - We're interested in predicting if a blood donor will donate within a given time window.
  - More on the problem statement at [Driven Data](#).
  - Dataset: [UCI ML Datasets Repo](#)

## Recommendation systems

---

- **Movie Recommender**
  - Can you predict the rating a user will give on a movie?
  - Do this using the movies that user has rated in the past, as well as the ratings similar users have given similar movies.
  - Dataset: [Netflix Prize](#) and [MovieLens Datasets](#)
- **Search + Recommendation System**
  - Predict which Xbox game a visitor will be most interested in based on their search query
  - Dataset: [BestBuy](#)
- **Can you predict Influencers in the Social Network?**
  - How can you predict social influencers?
  - Dataset: [PeerIndex](#)

## Vision

---

- **Image classification**

- Object recognition or image classification task is how Deep Learning shot up to it's present-day resurgence
- Datasets:
  - [CIFAR-10](#)
  - [ImageNet](#)
  - [MNIST Handwritten Digit Classification Challenge](#) is the classic entry point
  - [Character recognition \(digits\)](#) is the good old Optical Character Recognition problem
  - Bird Species Identification from an Image using the [Caltech-UCSD Birds dataset](#) dataset
- Diagnosing and Segmenting Brain Tumours and Phenotypes using MRI Scans
  - Dataset: MICCAI Machine Learning Challenge aka [MLC 2014](#)
- Identify endangered right whales in aerial photographs
  - Dataset: [MOAA Right Whale](#)
- Can computer vision spot distracted drivers?
  - Dataset: [State Farm Distracted Driver Detection](#) on Kaggle

- **Bone X-Ray dompetition**

- Can you identify if a hand is broken from a X-ray radiographs automatically with better than human performance?
- Stanford's Bone XRay Deep Learning Competition with [MURA Dataset](#)

- **Image Captioning**

- Can you caption/explain the photo a way human would?
- Dataset: [MS COCO](#)

- **Image Segmentation/Object Detection**

- Can you extract an object of interest from an image?
- Dataset: [MS COCO](#), [Carvana Image Masking Challenge](#) on Kaggle

- **Large-Scale Video Understanding**

- Can you produce the best video tag predictions?
- Dataset: [YouTube 8M](#)

- **Video Summarization**

- Can you select the semantically relevant/important parts from the video?
  - Example: [Fast-Forward Video Based on Semantic Extraction](#)
  - Dataset: Unaware of any standard dataset or agreed upon metrics? I think [YouTube 8M](#) might be good starting point.
- **Style Transfer**
  - Can you recompose images in the style of other images?
  - Dataset: [fzliu on GitHub](#) shared target and source images with results
- **Chest XRay**
  - Can you detect if someone is sick from their chest XRay? Or guess their radiology report?
  - Dataset: [MIMIC-CXR at Physionet](#)
- **Face Recognition**
  - Can you identify whose photo is this? Similar to Facebook's photo tagging or Apple's FaceId
  - Dataset: [face-rec.org](#), or [facedetection.com](#)
- **Clinical Diagnostics: Image Identification, classification & segmentation**
  - Can you help build an open source software for lung cancer detection to help radiologists?
  - Link: [Concept to clinic](#) challenge on DrivenData
- **Satellite Imagery Processing for Socioeconomic Analysis**
  - Can you estimate the standard of living or energy consumption of a place from night time satellite imagery?
  - Reference for Project details: [Stanford Poverty Estimation Project](#)
- **Satellite Imagery Processing for Automated Tagging**
  - Can you automatically tag satellite images with human features such as buildings, roads, waterways and so on?
  - Help free the manual effort in tagging satellite imagery: [Kaggle Dataset by DSTL, UK](#)

## Reinforcement Learning

---

- **Deep Q Learning**
  - Can you make AI play games and automate stuff by learning in an environment.

- Environments (dataset of Reinforcement Learning) [OpenAI GYM](#)
- T-REX Chrome Dino BOT [Git Repo](#)

## Music

---

- **Music Genre recognition using neural networks**
  - Can you identify the musical genre using their spectrograms or other sound information?
  - Datasets: [FMA](#) or [GTZAN on Keras](#)
  - Get started with [Librosa](#) for feature extraction

## Generative AI

---

### References:

[On the opportunities and risks of foundation models.](#)

Rishi Bommasani, et al., *arXiv preprint arXiv:2108.07258* (2021).

[ChatGPT is not all you need. A State of the Art Review of large Generative AI models](#)

Roberto Gozalo-Brizuela and Eduardo C. Garrido-Merchán. *arXiv preprint arXiv* (2023)

[Natural Language Processing with Transformers, Revised Edition.](#)

Lewis Tunstall, Leandro von Werra, Thomas Wolf (2022).

[Fine-tuning language models from human preferences.](#)

Daniel M. Ziegler, et al., *arXiv preprint* (2019).

### Online Resources:

[AI Tools](#) – A curated list by the University of Michigan of popular AI tools for a variety of applications, including: content, images, video, programming, productivity, and research.

[AwesomeLLM](#) – A curated list of LLMs from a technical and AI-focused perspective.

[Generative AI Essentials Course](#) – GenAI course by the University of Michigan Center for Academic Innovation.

[HuggingFace LLM Leaderboard](#) – Generative AI, and especially LLMs, are evolving rapidly. A curated list of top LLMs on a set of benchmarks can be found here.



[Papers with Code](#) – A collection of Generative AI (and non-Generative AI) papers with examples of code, datasets and more. A useful resource for seeing Generative AI applications in research.

[Prompt Engineering for ChatGPT](#) – How to apply prompt engineering effectively with ChatGPT.

[Stanford Helm](#) – “Holistic Evaluation of Language Models (HELM) is a living benchmark that aims to improve the transparency of language models.”

[thepi.pe](#) – “The thepi.pe platform provides a user-friendly interface for scraping and extracting data from various sources” – this page compares LLM evaluations.

### Other Dataset Suggestions

- [Data.gov](#): U.S. Government's open data
- KDD Cup: <http://www.kdd.org/kdd-cup>, annual competition in data mining, like Kaggle
- [Google public datasets](#).
- [NYC Taxi data](#) for 2013 (FOI'ed by Chris Wong). 2013 Trip Data (11.0GB). 2013 Fare Data (7.7GB). [Visualization for a days trip](#).
- [Yahoo WebScope](#)
- [Numerous APIs from Google](#) (e.g., Maps, Freebase, YouTube, etc.)
- [Trulia](#), [Zillow](#): real estate listing sites
- Numerous graph datasets (large and small): [SNAP](#), [Konect](#)
- Movies data: [Rotten Tomatoes](#), [IMDB](#)
- [List of lists of datasets for recommendations](#).
- [Million song dataset by Echo Nest](#).  
It contains not only the basic information of songs (artist, genre, year, length etc), but also some musical features (like tempo, pitch, key, brightness).
- [The Free 'Big Data' Sources Everyone Should Know](#)
- [Quandl - a dataset search engine for time-series data](#).
- [Amazon AWS Public Data Sets](#) (Thanks Jonathan!)
- [Classification datasets](#)
- [Various geophysical datasets](#) for the oceans (magnetism, gravity, seismology, etc).
- [Social trends](#)
- [Beer data](#)
- [Academic torrents \(terabytes\)](#)