# **Semester-Long Project Description**

**The aim of the project is to acclimate you to the process of conducting research in modern machine learning or data mining: 1) explore and analyze large-scale datasets 2) understand and replicate existing literature in hot research areas 3) think critically about the existing work and discover their pitfalls 4) innovate on top of the existing work incrementally.**

### **Grading**

**How well you answer the questions in the Topic Description and your deliverable satisfies the requirement below:**

1. **What you need to submit is a well-formatted Jupyter notebook (Colab link) optimized for readability) containing all answers to the questions and all training and testing records of your program. You are required to use Python (Pytorch, Pyro, and Tensorflow) in the project. Your code must be reproducible on** [**Google Colab**](https://colab.research.google.com/notebooks/intro.ipynb)**. Your code should be well documented (include comments on what your code is doing).**
2. **We provide reference code with you to help implement the papers, and you need to read and fully understand what the code is doing, run through the code (maybe debugging), and learn to revise and use the code smartly.**
3. **Your exposition should focus on summarization and highlighting the most salient/important aspects of the papers (aiming for introducing your work to an audience with little relevant background).**
4. **Answer the questions concisely and just hit the point. There will be penalties for verbose and irrelevant answers.**

**For bonus questions:**

1. **full grades: realize your ideas with your own code; produce better/competitive results**
2. **more than half of the grades: describe your ideas with some algorithms in detail; the ideas are reasonable and described concisely**
3. **less than half of the grades: write some general descriptions**

**For undergraduate students, you are only required to answer parts of questions, and your final grades will be scaled to 150 points. Specifically, you do not need to answer the following questions, but you will get bonus credits if you do answer them well:**

**Topic 1: Problem 7, Problem 8**

**Topic 2: Problem 5, Problem 6**

**Topic 3: Problem 7**

**Topic 4: Problem 5**

**Sample Jupiter notebooks:**

[**https://github.com/onefishy/am207\_fall19\_projects/blob/master/style\_transfer/Final\_Deliverabless\_AliceLi\_LinZhu.ipynb**](https://github.com/onefishy/am207_fall19_projects/blob/master/style_transfer/Final_Deliverabless_AliceLi_LinZhu.ipynb)

[**https://github.com/onefishy/am207\_fall19\_projects/blob/master/wormhole\_hmc/wongalexander\_140170\_9129284\_Project%20(2).ipynb**](https://github.com/onefishy/am207_fall19_projects/blob/master/wormhole_hmc/wongalexander_140170_9129284_Project%20(2).ipynb)

[**https://github.com/onefishy/am207\_fall19\_projects/blob/master/stochastic\_gradient\_hmc/sghmc\_1/Stochastic\_Gradient\_HMC.ipynb**](https://github.com/onefishy/am207_fall19_projects/blob/master/stochastic_gradient_hmc/sghmc_1/Stochastic_Gradient_HMC.ipynb)

### **Instructions**

**Team: Form a team of 3-5 people. Do not form a group between undergrad students and grad students.**

**Topic: Choose one topic from the four given choices described below. After your group members are finalized, you will bid for the topics you like, but there is no guarantee that you will get it. Specifically, you have 10 points as your budget, and you are asked to distribute the 10 points to the four topics (e.g., 5+3+2+0 for Topic 1, 2, 3, and 4, respectively). You will have higher chances to get the topic with larger points. If all groups bid all 10 points for one Topic, those who come earlier (depending on the time the TA receives your emails) will probably get the Topic, and the rest will be assigned a Topic randomly. The bid will start at 00:00 am on Sep. 22, and only one of your group members emails the TA (scao16@binghamton.edu) your group number and your scores for each topic (e.g., Group 2: 7+1+0+2). Make sure that the email’s subject follows "group number: topic scores", e.g., " Group 5: 0+1+7+2". Any emails before the starting time will be ignored.**

### **Topic Descriptions**

### **Topic 1: Topic Modeling**

**Introduction:**

1. [Prof. David Blei - Probabilistic Topic Models and User Behavior](https://www.youtube.com/watch?v=FkckgwMHP2s&t=139s)
2. [**Probabilistic Topic Modeling — Pyro Tutorials 1.7.0 documentation**](https://pyro.ai/examples/prodlda.html)
3. [MLSS 2019 David Blei: Variational Inference: Foundations and Innovations (Part 1)](https://www.youtube.com/watch?v=DaqNNLidswA&t=3809s)
4. [David Blei Variational Inference Foundations and Innovations Part 2](https://www.youtube.com/watch?v=Wd7R_YX4PcQ)

**Papers:**

**[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003).** [**Latent Dirichlet allocation.**](https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf)

**[2] Akash Srivastava, & Charles Sutton. (2017).** [**Autoencoding Variational Inference For Topic Models.**](https://arxiv.org/abs/1703.01488)

**[3]** [**TOP2VEC: DISTRIBUTED REPRESENTATIONS OF Topics**](https://arxiv.org/pdf/2008.09470.pdf)

**Problem 1: what is the problem the three papers aim to solve, and why is this problem important or interesting? (5 points)**

**Problem 2: 1) summarize the three methods, including high-level ideas as well as technical details: the relevant details that are important to focus on (e.g., if there’s a model, define it; if there is a theorem, state it and explain why it’s important, etc) 2) what are the major differences of the three methods? (15 points)**

**Problem 3: implement the ProdLDA [2] topic model and test it on the 20 newsgroups text dataset (report the 20 main topics you discovered from the data and visualize their word clouds) (30 points)**

**Reference code:** [**https://pyro.ai/examples/prodlda.html**](https://pyro.ai/examples/prodlda.html)

**Problem 4: implement Latent Dirichlet Allocation (LDA) [1] and test it on the 20 newsgroups text dataset (report the 20 main topics you discovered from the data and visualize their word clouds) (30 points)**

**Reference code:** [**https://pyro.ai/examples/lda.html#**](https://pyro.ai/examples/lda.html#)

**Problem 5: implement TOP2VEC [3] and test it on the 20 newsgroups text dataset (report the 20 main topics you discovered from the data and visualize their word clouds) (30 points)**

**Reference code:** [**https://github.com/ddangelov/Top2Vec**](https://github.com/ddangelov/Top2Vec)

**Problem 6: compare the experimental results of the three methods with the ground truth topics of the 20 newsgroups. Which method do you think is better? Explain why. (10 points)**

**Problem 7: fetch real-time tweets and discover real-time topics on them with the three methods, respectively. You need to report the top 5 topics for the last 5 seconds in real-time and the top 5 topics for each hour in the most recent 24 hours. (20 points)**

**Problem 8: fetch real-time news and discover real-time topics on them with the three methods, respectively. You need to report the top 5 topics for each day in the most recent week. (10 points)**

**Problem 9 (bonus): analyze the pitfalls of the existing topic modeling methods above and come up with one way to address the pitfalls. Can you realize your ideas? (60 points)**

### **Topic 2: Image Generation with Probabilistic Diffusion Models**

**Introduction:**

1. [DDPM - Diffusion Models Beat GANs on Image Synthesis (Machine Learning Research Paper Explained)](https://www.youtube.com/watch?v=W-O7AZNzbzQ&t=2184s)
2. [**Lil Log What are Diffusion Models?**](https://lilianweng.github.io/lil-log/2021/07/11/diffusion-models.html)

**Papers:**

**[1]**[**Denoising Diffusion Probabilistic Models**](https://hojonathanho.github.io/diffusion/)

**[2]** [**DENOISING DIFFUSION IMPLICIT MODELS**](https://openreview.net/pdf?id=St1giarCHLP)

**[3]**[**On Fast Sampling of Diffusion Probabilistic Models**](https://arxiv.org/pdf/2106.00132.pdf)

**Problem 1: what is the problem the two papers aim to solve, and why is this problem important or interesting? (5 points)**

**Problem 2: 1) summarize the three methods, including high-level ideas as well as technical details: the relevant details that are important to focus on (e.g., if there’s a model, define it; if there is a theorem, state it and explain why it’s important, etc) 2) what are the major differences of the three methods? (15 points)**

**Problem 3: implement DDPM [1] and test it on MNIST dataset. You need to generate samples and perform an interpolation experiment with your trained model. (30 points)**

**Reference code:** [**https://github.com/pesser/pytorch\_diffusion**](https://github.com/pesser/pytorch_diffusion)

**Problem 4: implement DDIM [2] and test it on MNIST dataset. You need to generate samples and perform an interpolation experiment with your trained model. (30 points)**

**Reference code:** [**https://github.com/ermongroup/ddim**](https://github.com/ermongroup/ddim)

**Problem 5: implement FastDPM [3] and test it on MNIST dataset. You need to generate samples and perform an interpolation experiment with your trained model. (30 points)**

**Reference code:** [**https://github.com/FengNiMa/FastDPM\_pytorch**](https://github.com/FengNiMa/FastDPM_pytorch)

**Problem 6: visualize the denoising process on CIFAR-10 and CelebA-HQ datasets for all three methods, respectively (e.g., Figure 6 of [1]). You are allowed to use the pre-trained models provided by the authors. (20 points)**

**Problem 7: interpolate source images on CIFAR-10/100 and CelebA-HQ datasets with all three methods, respectively (e.g., Figure 8 of [1]). You are allowed to use the pre-trained models provided by the authors. (20 points)**

**Problem 8 (bonus): is it possible to use diffusion models to generate text data? If you think that it is impossible or tough, explain why; If you think that it is possible, you will get bonus credits by realizing the idea. (80 points)**

### **Topic 3: Image Classification on Edges**

**Introduction:**

**Here are short videos to show you what is edge detection:**

1. [Dynamic Feature Fusion for Semantic Edge Detection (DFF)](https://www.youtube.com/watch?v=wSCKTepMfhY)
2. [Finding the Edges (Sobel Operator) - Computerphile](https://www.youtube.com/watch?v=uihBwtPIBxM)
3. [Canny Edge Detector - Computerphile](https://www.youtube.com/watch?v=sRFM5IEqR2w)

**However, the main task of this project is not to detect edges but to do image classification tasks with those edges you detected (with deep learning methods described in the paper [1] and [2]).**

**Papers:**

**[1]**[**Dynamic Feature Fusion for Semantic Edge Detection**](https://arxiv.org/pdf/1902.09104.pdf)

**[2]** [**Dense Extreme Inception Network: Towards a Robust CNN Model for Edge Detection**](https://arxiv.org/pdf/1909.01955.pdf)

**Problem 1: what is the problem the two papers aim to solve, and why is this problem important or interesting? (5 points)**

**Problem 2: 1) summarize the two methods, including high-level ideas as well as technical details: the relevant details that are important to focus on (e.g., if there’s a model, define it; if there is a theorem, state it and explain why it’s important, etc) 2) what are the major differences between the two methods? (15 points)**

**Problem 3: perform standard image classification with ResNet 18 on the CIFAR-10, CIFAR-100, SVHN, and Tiny ImageNet datasets and report the classification accuracy. (15 points)**

**Reference code:** [**https://github.com/kuangliu/pytorch-cifar**](https://github.com/kuangliu/pytorch-cifar)

**Problem 4: generate and visualize the edges you detected with [1] on the CIFAR-10, CIFAR-100, SVHN, and Tiny ImageNet datasets. You are allowed to use the pre-trained models provided by the authors. (15 points)**

**Reference code:** [**https://github.com/Lavender105/DFF**](https://github.com/Lavender105/DFF)

**Problem 5: 1) generate and visualize the edges you detected with [2] on the CIFAR-10, CIFAR-100, SVHN, and Tiny ImageNet datasets. You are allowed to use the pre-trained models provided by the authors. (20 points)**

**Reference code:** [**https://github.com/xavysp/DexiNed**](https://github.com/xavysp/DexiNed)

**2) combine the edges you detected and the original images to create “edge-enhanced images”. You may want to store the edges and the “edge-enhanced images” as datasets, to solve Problem 6 and Problem 7 later. (10 points)**

**Problem 6: perform image classification with ResNet 18 on the edges of the CIFAR-10, CIFAR-100, SVHN, and Tiny ImageNet datasets and report the classification accuracy. (30 points)**

**Problem 7: Perform image classification with ResNet 18 on the “edge-enhanced images” of the CIFAR-10, CIFAR-100, SVHN, and Tiny ImageNet datasets and report the classification accuracy. (30 points)**

**Problem 8: compare the results from Problem 3, Problem 6, and Problem 7. Do you see improvements by using edge information for image classification? Explain why. (10 points)**

**Problem 9 (bonus): analysis the pitfalls of existing edge detection methods above and come up with one way to address the pitfalls. Realize your ideas. (60 points)**

### **Topic 4: Natural-Image Reconstruction from fMRI**

**Introduction:**

1. [Deep image reconstruction from human brain activity (Paper Explained)](https://www.youtube.com/watch?v=YrO1v7-KcXs)
2. [Self-Supervision in Natural-Image Reconstruction from fMRI](https://www.youtube.com/watch?v=xArhRHDOU5Q&t=130s)

**Papers:**

**[1]** [**Deep image reconstruction from human brain activity**](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006633)

**[2]** [**From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI**](https://arxiv.org/pdf/1907.02431.pdf)

**Problem 1: what is the problem the two papers aim to solve, and why is this problem important or interesting? (5 points)**

**Problem 2: 1) summarize the two methods, including high-level ideas as well as technical details: the relevant details that are important to focus on (e.g., if there’s a model, define it; if there is a theorem, state it and explain why it’s important, etc) 2) what are the major differences of the two methods? (15 points)**

**Problem 3: reproduce experimental results in [1]. You need to show both the stimulus and the reconstructed images (Figure 2 and Figure 6 of [1]). (30 points)**

**Reference code:** [**https://github.com/KamitaniLab/End2EndDeepImageReconstruction**](https://github.com/KamitaniLab/End2EndDeepImageReconstruction)

**Problem 4: reproduce experimental results in [2]. You need to show both the ground truth and the reconstructed images (Figure 5 of [2]). (30 points)**

**Reference code:** [**http://www.wisdom.weizmann.ac.il/~vision/ssfmri2im/**](http://www.wisdom.weizmann.ac.il/~vision/ssfmri2im/)

**Problem 5: run both methods on “**[**Generic Object Decoding**](https://openneuro.org/datasets/ds001246/versions/1.0.1)**” dataset and test them on** [**Deep Image Reconstruction**](https://openneuro.org/datasets/ds001506/versions/1.3.1) **dataset to see if the models have strong generalization ability. (30 points)**

**Problem 6: which method is better based on your results of Problem 3, Problem 4, and Problem 5? What is the main reason why it performs better (theoretically and practically)? (20 points)**

**Problem 7: Based on your answer to Problem 2 and Problem 6, what are the major obstacles of the task of image reconstruction from fMRI? Can you think of ways to address them? (20 points)**

**Problem 8 (bonus): try to improve the quality of the reconstructed images by innovating on top of [2] incrementally. For example, you may use SOTA neural network architectures (e.g., Vision Transformer) or adding specific regularizations that you think will help. Or can you think of a way to combine the merits of [1] and [2]? (80 points)**

**Reproducing the existing literature is a fundamental ability no matter whether you are in academia or industry. It is also crucial to think critically about the experimental results and the method itself. Have fun!**