# Introduction to Deep Learning & Finance

# Course Project 2

**Instructors: Xiaolin Hu and Wu Zhu**
Department of Computer Science and Technology, Tsinghua University
Department of Finance, SEM, Tsinghua University

## Aims

This project offers you an opportunity to apply what you have learned in the course to real-world challenges.

## Options

Project 2 is an open project, in which you can pick up any topic you are interested in. The topic could be a new research in finance field, or you can find a real-world finance problem to solve. In case you can't find any topic you are particularly interested in, we provide a few options in the following list.

Suggestion list:

1. AI conference paper reproduction

   We encourage you to reproduce latest works from top AI conferences, here we offered several papers as a suggestion:

   1. MASTER: Market-Guided Stock Transformer for Stock Price Forecasting
      Link: https://ojs.aaai.org/index.php/AAAI/article/download/27767/27575

   2. MDGNN: Multi-Relational Dynamic Graph Neural Network for Comprehensive and Dynamic Stock Investment Prediction
      Link: https://arxiv.org/pdf/2402.06633

   3. CI-STHPAN: Pre-trained Attention Network for Stock Selection with Channel-Independent Spatio-Temporal Hypergraph
      Link: https://ojs.aaai.org/index.php/AAAI/article/download/28770/29478

   Note that these papers use different data, and if you want to do these tasks, contact TA for uploading these data to the server. These works require better coding skills compared to following options, so please be aware of it when you choose your project topic.

2. A paper titled "Empirical Asset Pricing via Machine Learning" by Gu et al. published in *The Review of Financial Studies*, May 2020. Available at: https://academic.oup.com/rfs/article/33/5/2223/5758276

   A comprehensive study of ML methods over US stock market. You can try to reproduce the neural network used in the paper, or further experiment on models you find suit to the case.

3. A paper titled "Autoencoder Asset Pricing Models" by Gu et al. published in Journal of Econometrics, May 2021. Available at: Autoencoder asset pricing models - ScienceDirect

   Explaining stocks return with a latent risk-factor model based on autoencoder.

The above two works use the same dataset, which you can find at:
https://cloud.tsinghua.edu.cn/d/4844d94f42eb4c8dbbcf/.

If you choose to work on paper 2 or 3, we require you to try with methods beyond those used in the papers, compare your method's performance to original papers' and analyze it.

4. A paper titled "Predicting Returns with Text Data" by Ke et al., which is still a working paper. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3389884

   Predict stock return through related news. This involved with language models for sentiment analysis. Unfortunately, we cannot provide data of this paper, but we encourage you to explore on text data and perform similar task. You may focus on news related to a small set of assets and see how it works instead of predicting the whole market.

   Tsinghua library includes some data sources that provide finance-related text data:

   https://ecollection.lib.tsinghua.edu.cn/databasenav/entrance/detail?mmsid=991021855125703966
   https://ecollection.lib.tsinghua.edu.cn/databasenav/entrance/detail?mmsid=991021855125503966
   https://ecollection.lib.tsinghua.edu.cn/databasenav/entrance/detail?mmsid=991021855125903966

   You can further explore the whole economy data sources from Tsinghua library to see if there is any data you are particularly interested in:
   https://ecollection.lib.tsinghua.edu.cn/databasenav/entrance/dataBaseSubSearchList?subject=%E7%BB%8F%E6%B5%8E%E5%AD%A6

If you choose to reproduce the results of paper 2 and paper 3, you only need to reproduce the main results. But only reproducing the results is not enough to get a very high grade unless you make further analysis or do some extensions (e.g., apply the method to the finance financial data of other countries that haven't been studied in the paper). An in-depth analysis would be encouraged with a high score, such as heterogeneity analysis to explore the potential channel driving your main results.

You are **welcome** to discuss your proposal with the instructors and TA in emails or face to face at any time.

## Important dates

5. Form groups of 2 or 3 and submit a proposal before: April. 26 (Sunday), 2024
6. Poster presentation on: June 16 (Sunday), 2024
   Submit PDF poster online before: June 17 (Monday), 2024
7. Submit the final report before: June 21 (Friday), 2024

## Project Proposal

The project proposal should be one paragraph (200-400 words). Your project proposal should describe:

- What is the problem that you will be investigating? Why is it interesting?
- What data will you use? If you are collecting new data, how will you do it?
- What method or algorithm are you proposing? If there are existing implementations, will you use them and how? How do you plan to improve or modify such implementations? You don't have to have an exact answer at this point, but you should have a general sense of how you will approach the problem you are working on.
- How will you evaluate your results? Qualitatively, what kind of results do you expect (e.g. plots or figures)? Quantitatively, what kind of analysis will you use to evaluate and/or compare your results (e.g. what performance metrics or statistical tests)?

**Submission:** Please submit your proposal as a PDF. **Every member in your group should submit** for the sake of grading.

**Grading (5%).** Based on the completeness of the above points.

## Poster presentation

Make a poster and report to your peer students. A template will be provided.
**Submission**: Please submit a PDF file version of your poster after the presentation day.

**Grading (5%, By peer students):** Based on the novelty, clarity, and intermediate results.

## Final report

Your final write-up should be between **5 - 8** pages using the provided template:
https://www.overleaf.com/latex/templates/neurips-2023/vstgtvjwgdng
**Submission:**

1. A pdf file of your final report
2. zip file with Supplementary Materials

*Report.* The following is a suggested structure for the report:

- Title, Author(s)
- Abstract: It should not be more than 200 words
- Introduction: this section introduces your problem, and the overall plan for approaching your problem
- Background/Related Work: This section discusses relevant literature for your project
- Approach: This section details the framework of your project. Be specific, which means you might want to include equations, figures, plots, etc
- Experiment: This section begins with what kind of experiments you're doing, what kind of dataset(s) you're using, and what is the way you measure or evaluate your results. It then shows in details the results of your experiments. By details, we mean both quantitative evaluations (show numbers, figures, tables, etc) as well as qualitative results (show images, example results, etc).
- Conclusion: What have you learned? Suggest future ideas.
- References: This is absolutely necessary.

*Supplementary Material* is not counted toward your 5-8 page limit.

- Source code. (required)

- Any text or lengthy math derivatives. (optional)
- Cool videos, interactive visualizations, demos, etc.(optional)

Examples of things to *NOT* put in your supplementary material:

- Datasets files
- Model / weights checkpoint files
- A computer virus

**Grading (30%):**

write-up: 10%

- clarity, structure, language, references: 4%
- background literature survey, good understanding of the problem: 3%
- good insights and discussions of methodology, analysis, results, etc.: 3%

technical: 10%

- correctness: 4%
- depth: 3%
- innovation: 3%

evaluation and results: 10%

- sound evaluation metric: 3%
- thoroughness in analysis and experimentation: 3%
- results and performance: 4%

## Example project reports

Your final report should structure like a computer science conference paper. You may find some example project reports of the CS courses in Stanford University here:
http://cs231n.stanford.edu/reports.html
http://cs229.stanford.edu/projects.html

## Collaboration with people not enrolled in this course

Your report PDF should list all authors who have contributed to your work; enough to warrant a co-authorship position. This includes people not enrolled in this course such as faculty/advisors if they sponsored your work with funding or data, significant mentors (e.g., PhD students or postdocs who coded with you, collected data with you, or helped draft your model on a whiteboard). All authors should be listed directly underneath the title on your PDF. Include a footnote on the first page indicating which authors are not enrolled in this course. All co-authors should have their institutional/organizational affiliation specified below the title.

If you have non-course contributors, you will be asked to describe the following: Specify the involvement of non-course contributors (discussion, writing code, writing paper, etc). For an example, please see the author contributions for AlphaGo (Nature, 2016).

## Other rules

- You can write the submission files in either Chinese or English.
- You may consult any papers, books, online references, or publicly available implementations for ideas and code that you may want to incorporate into your strategy or algorithm, so long as you clearly cite your sources in your code and your write-up. However, under no circumstances may you look at another

group's code or incorporate their code into your project.

- The same project should *NOT* be submitted to multiple courses. If discovered, it will be regarded as cheating, and your final scores will be divided by N, where N is the number of courses that you submitted to.
- Specify whether the project has been submitted to a peer-reviewed conference or journal. Include the full name and acronym of the conference (if applicable). This only applies if you have already submitted your paper/manuscript and it is under review as of the report deadline.

**Acknowledgement:** The design of this project follows that of the Stanford course CS231n: http://cs231n.stanford.edu/project.html.