CS5284 : Graph Machine Learning

Administrative (Week 5)

Semester 1 2025/26

Xavier Bresson

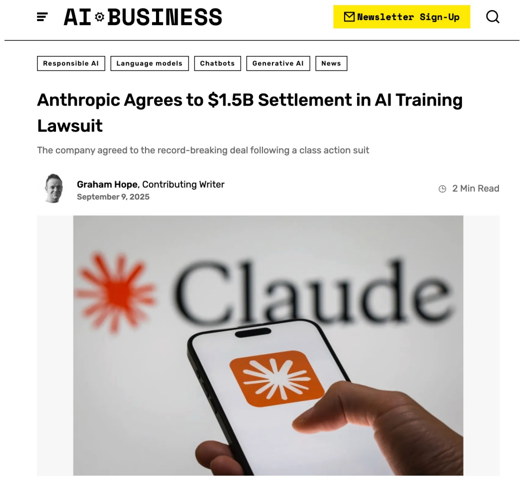
https://x.com/xbresson

Department of Computer Science
National University of Singapore (NUS)

QR Attendance

Please, scan the new below QR image for attendance.

Anthropic and several other companies s.a. OpenAI have been sued for copyright infringement.
LLMs require very large training datasets to pre-train neural networks with a large number of learnable parameters.
When you scrape the internet, there's a high chance of collecting copyrighted material.
What do you do then? Do you filter it?
Filtering is imperfect, and there's a temptation not to, since copyrighted data is usually high quality.
In any case, authors should be rewarded for their work, but it seems companies prefer to deal with lawsuits later rather than fix this issue proactively.
https://aibusiness.com/responsible-ai/anthropic-agrees-to-1-5b-settlement-in-ai-training-lawsuit



Anthropic Agrees to $1.5B Settlement in AI Training Lawsuit

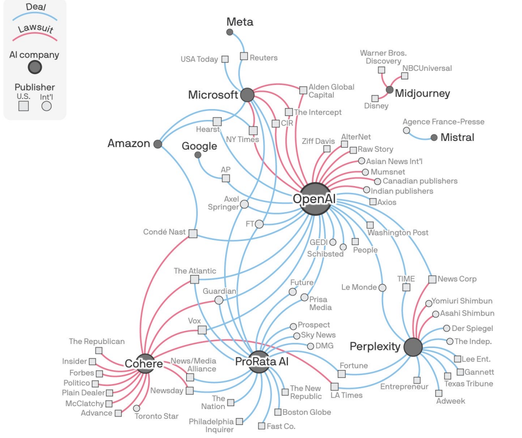
The company agreed to the record-breaking deal following a class action suit

Graham Hope, Contributing Writer
September 9, 2025

2 Min Read

Select publisher deals and lawsuits with AI companies

As of Sept. 4, 2025



Data: Axios research; Chart: Erin Davis/Axios Visuals

Admin

Change of Venue for Tutorial 2

- **Due to an IT issue, the temporary venue for this week's Tutorial 2 will be COM1-02-10.**

In-lecture questions

In-lecture question [Answer]

- **How do you compute an approximate solution of the Normalized Cut optimization problem? What is the operator B?**

$$\min_{F\in\{0,1\}^{n\times k}} \sum_{q=1}^{k} \frac{F_{\cdot,q}^T L F_{\cdot,q}}{F_{\cdot,q}^T D F_{\cdot,q}}, \quad \text{with } L = D - A \quad \text{and} \quad \sum_{q=1}^{k} F_{i,q} = 1 \ \forall i \in V$$

$$\min_{Y\in \text{ binary}^{n\times k}} \ \text{tr}(Y^T B Y) \ \text{ s.t. } \ Y^T Y = \text{I}_k, \ B = \text{I} - D^{-1/2} A D^{-1/2}$$

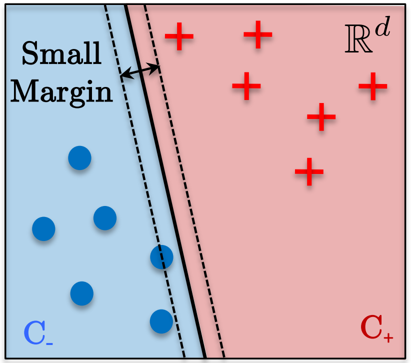
- **Answer : As before, relaxing the binary constraint Y from binary$^{\text{n x k}}$ to the nearest convex set R$^{\text{n x k}}$ makes the optimization continuous and tractable. This relaxation provides an approximate solution given by the spectral theorem, specifically, the k smallest eigenvectors of the normalized graph Laplacian B.**

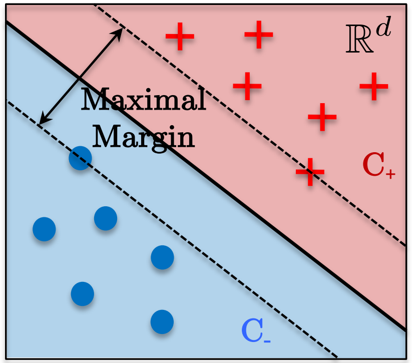
$$B = \Theta^{1/2} A \Theta^{1/2} \overset{\text{EVD}}{=} U \Lambda U^T \in \mathbb{R}^{n\times n}$$

$$\text{Solution} \quad Y^\star = U_{\cdot,1:k} \in \mathbb{R}^{n\times k} \quad (k \text{ smallest eigenvectors})$$

In-lecture question [Answer]

- How many class separators exist for linearly separable distributions? Why is it important to maximize the margin between the two classes? Please justify.
- In Slack #lectures
  - Identify the question and Reply in thread with a short response
- Answer : Multiple hyperplanes can separate the two classes in a linearly separable distribution. However, we aim to choose the hyperplane that generalizes best — this is the one that maximizes the margin between the classes. A larger margin typically leads to better generalization on unseen data, reducing the risk of overfitting.



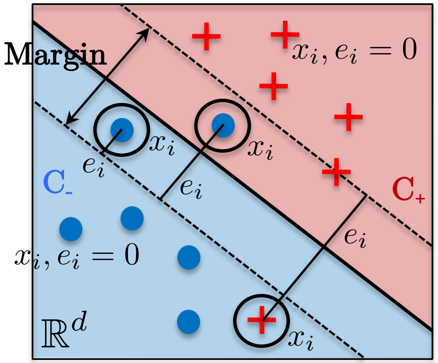


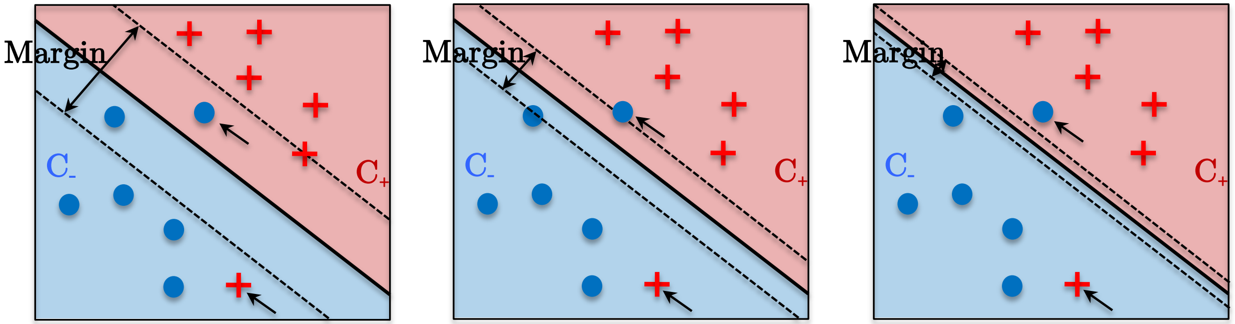
In-lecture question [Answer]

- What is the impact of varying the regularization parameter λ? Specifically, how do small versus large values of λ affect the model? Please justify.

- In Slack #lectures
  - Identify the question and Reply in thread with a short response

- Answer : For small λ values, more misclassification errors are allowed, the margin is larger. For large λ values, misclassification errors are penalized, leading to either no errors or very few, resulting in a smaller margin.





Small λ value

Intermediate λ value

Large λ value

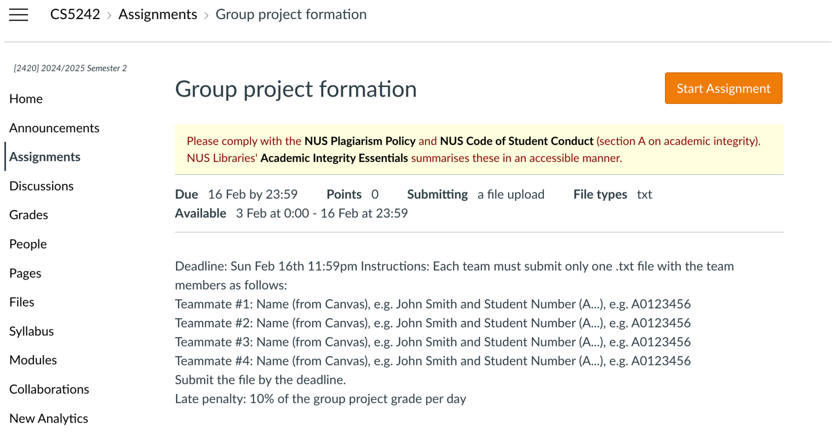
Group project

Group project formation

- Your team
  - You are free to select your teammates (each group may have 2-5 members), but you will have to make an agreement/contract to distribute and contribute equally to the tasks (for minimizing future conflict).
  - Contribute equally does not mean contributing like an expert in coding, maths, engineering, presentation, etc (some people are beginners) -- it means that the effort and attitude to make the project successful must be at the same level than others.
  - Each group, i.e. each teammate, will receive the same grade.
    - Choose your group wisely!
    - Not only people you know, i.e. your friends, but people willing to work on the project (good friend ≠ good teammate).
  - Note that you can select a teammate which is not in your tutorial group.

Group project formation

- Each team must submit a *unique* .txt file with the team members at Canvas>Assignments>Group project formation
- Submission deadline : Sun Sep 21st 2025 11:59pm (Week 6)
- Penalty : 10% of the group grade per late day



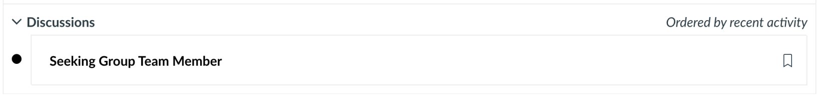
Group project formation

Please comply with the **NUS Plagiarism Policy** and **NUS Code of Student Conduct** (section A on academic integrity). NUS Libraries' **Academic Integrity Essentials** summarises these in an accessible manner.

Due 16 Feb by 23:59 **Points** 0 **Submitting** a file upload **File types** txt
Available 3 Feb at 0:00 - 16 Feb at 23:59

Deadline: Sun Feb 16th 11:59pm Instructions: Each team must submit only one .txt file with the team members as follows:
Teammate #1: Name (from Canvas), e.g. John Smith and Student Number (A...), e.g. A0123456
Teammate #2: Name (from Canvas), e.g. John Smith and Student Number (A...), e.g. A0123456
Teammate #3: Name (from Canvas), e.g. John Smith and Student Number (A...), e.g. A0123456
Teammate #4: Name (from Canvas), e.g. John Smith and Student Number (A...), e.g. A0123456
Submit the file by the deadline.
Late penalty: 10% of the group project grade per day

- Looking for teammates : Use e.g. Canvas discussion at Canvas>Discussions>Seeking Group Team Member



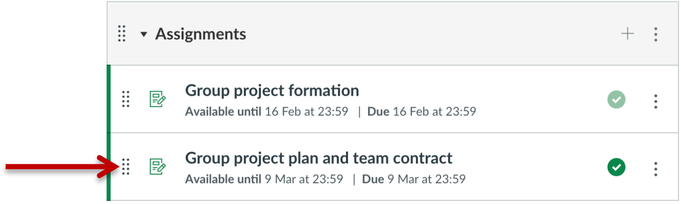
● Seeking Group Team Member

Project plan and team contract

- Project plan and team contract
  - Write a clear and concise one-page description of the project.
    - Strictly one-page limit. If $>$ one-page limit then project plan grade will be a 0.
    - Exception : References (if any) can be provided as extra pages.
    - Any style and format can be used, e.g. single/double columns, etc.
  - Possible template of project plan
    - Project motivation, description, proposed solution, project milestones.
  - Team contract/agreement
    - Add an additional page which describes the tasks assigned to each team member.
    - Each teammate must contribute equally to the project.
    - Each teammate must sign the contract.
    - If the signed contract is not submitted with the project plan, then the project grade will be a 0.

Project plan and team contract

- Project plan and team contract
  - Submit the project plan and team contract in Canvas:
    - Upload one .pdf file into Canvas > Assignment > Group project plan and team contract
    - Submit a *single* project plan and team contract per team.
    - File name must be “project_plan_contract_groupIDXX.pdf”, (for example project_plan_contract_groupID31.pdf) (see later slide for groupIDXX).
    - Deadline : Sun Oct 5th 2025 11:59pm (Week 7)
    - Penalty : 10% of the group grade per late day
    - Project plan and contract do not bring any point to the project grade
    - The TA allocated to your group will provide a feedback in Canvas by Fri Oct 10th (Week 8)



Assignments

Group project formation
Available until 16 Feb at 23:59 | Due 16 Feb at 23:59

Group project plan and team contract
Available until 9 Mar at 23:59 | Due 9 Mar at 23:59

Group ID

- After the deadline of group formation, your team will be assigned an ID number, i.e. IDXX.
  - For example, team ID27: John Smith and Joe Doe
  - Your team ID number will be available at Canvas > Home > W07 > list_ID_project_groups.pdf
    - Please, check and use your group ID for any future communication and submission.

TA allocated to Groups

- TA allocated to your group will be as follows :
  - Groups ID XX to XX(included) : Mr. Wang Jiaming, e0942816@u.nus.edu
  - Groups ID XX to XX(included) : Mr. Ryoji Kubo, e1583584@u.nus.edu
  - Groups ID XX to XX(included) : Mr. Liu Nian, e1154528@u.nus.edu
- Reminder : Group TA ≠ Tutorial TA
- If you have any question about the project, please ask the TA in charge of your group.

Project philosophy

- This project focuses on
  - The understanding of the fundamental concepts of graph machine learning techniques,
  - The practical skills required to develop a data analysis project.
- It is not about learning to use GitHub codes.
- It is not about winning a Kaggle competition.
- It is not about three lines of Keras' code to run machine learning techniques.
- It is not about running long experiments with the best possible GPUs.
  - Google Colab, Google Cloud, and your computer/laptop are enough.
- It is not about getting 90% of accuracy.
- It is about how to design from scratch, debug, understand and train learning algorithms.
- It is about to understand why it works and why it does not.

Project philosophy

- This project focuses on :
  - Theoretical knowledge received in this module.
  - Practical skills with data acquisition, exploration, exploitation, analysis.
  - Teamwork with management of tasks.
  - Concise and clear communication with written report and oral presentation.

Project goals

- Project goals
  - Download or prepare a dataset(s)
    - This dataset(s) can be novel or not.
  - Implement graph machine learning techniques on this dataset(s)
    - Use simple model(s) as baseline.
  - Propose improvement(s)
    - Motivation, description, equation, implementation, result, discussion.
  - Demonstrate initiatives
    - Develop own scrapper, dynamic visualization, discover new data insights, etc.
  - Deliveries
    - Python notebook for code demo
    - Project report (it can be merged with the notebook)
    - Video presentation and slides.

Pre-trained models

- The primary goal of the project is not to achieve 90% accuracy or better, but rather to assess the students' understanding of graph machine learning fundamentals.
- Students are expected to build something from first principles or "from scratch."
- However, it is allowed to use pre-trained models from DGL, PyG, HuggingFace, etc.
- But the use of pre-trained networks should be well justified within the context of the project.

Dataset(s)

- Dataset(s) can be collected from an existing repository
  - UCI : https://archive.ics.uci.edu/datasets
  - Kaggle : https://www.kaggle.com/datasets
  - Paperswithcode : https://paperswithcode.com/datasets
  - GitHub : https://github.com/topics/dataset
  - DGL : https://docs.dgl.ai/en/2.2.x/api/python/dgl.data.html
  - PyG : https://pytorch-geometric.readthedocs.io/en/2.5.2/modules/datasets.html

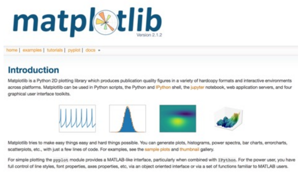
- Dataset(s) can be new, i.e.
  - Scrap using an API, e.g. Twitter API or Meta API
  - Collect data with your hand-crafted scrapper

Project development

- Step 1: Identify a data analysis problem that can be solved with graph machine learning.
  - You may use your own field of expertise or your personal interests.
  - The problem is neither too easy nor too difficult!
- Step 2: Dataset collection
  - Use existing dataset(s)
  - Develop new dataset(s)

Project development

- Step 3: Data exploration (analyze your data, get insights)
  - Use statistics
  - Use visualization libraries, for example
    - Matplotlib: https://matplotlib.org
    - Bokeh: https://bokeh.pydata.org
    - Graphlab: https://gephi.org

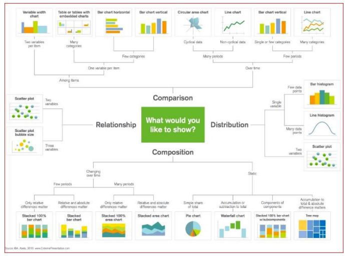


Introduction

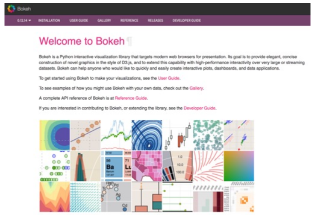
Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the jupyter notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the `pyplot` module provides a MATLAB-like interface, particularly when combined with `IPython`. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.



Source: ©A. Abela, 2010. www.ExtremePresentation.com



Welcome to Bokeh ¶

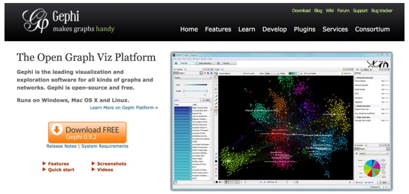
Bokeh is a Python interactive visualization library that targets modern web browsers for presentation. Its goal is to provide elegant, concise construction of novel graphics in the style of D3.js, and to extend this capability with high-performance interactivity over very large or streaming datasets. Bokeh can help anyone who would like to quickly and easily create interactive plots, dashboards, and data applications.

To get started using Bokeh to make your visualizations, see the User Guide.

To see examples of how you might use Bokeh with your own data, check out the Gallery.

A complete API reference of Bokeh is at Reference Guide.

If you are interested in contributing to Bokeh, or extending the library, see the Developer Guide.



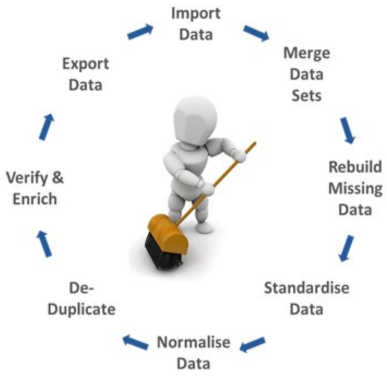
The Open Graph Viz Platform

Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.

Runs on Windows, Mac OS X and Linux.

Learn More on Gephi Platform »

- Step 4: Pre-processing
  - Data cleaning (missing features)
  - Data normalization (unbalanced scaling)
  - Important and consuming step to prepare data as clean as possible for analysis



Import Data

Merge Data Sets

Rebuild Missing Data

Standardise Data

Normalise Data

De-Duplicate

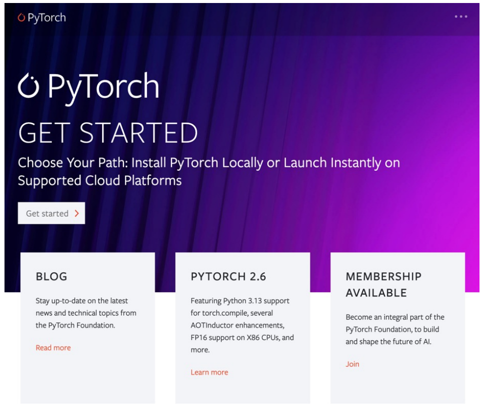
Verify & Enrich

Export Data

Project development

- Step 5: Data analysis with deep learning
  - Apply machine learning to solve your data problem :
    - Regression, classification, etc
  - Compare different models

- Step 6: Numerical results
  - Analysis, interpretation, conclusion



GET STARTED

Choose Your Path: Install PyTorch Locally or Launch Instantly on Supported Cloud Platforms

BLOG

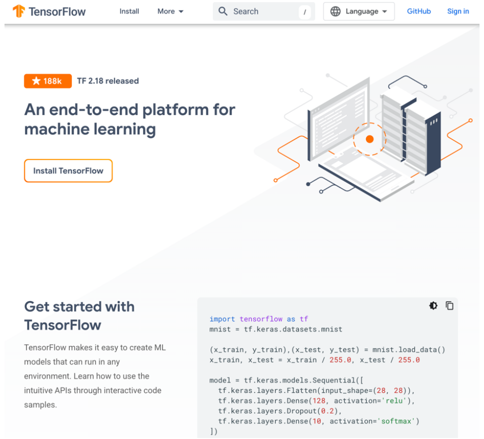
Stay up-to-date on the latest news and technical topics from the PyTorch Foundation.

PYTORCH 2.6

Featuring Python 3.13 support for torch.compile, several AOTInductor enhancements, FP16 support on X86 CPUs, and more.

MEMBERSHIP AVAILABLE

Become an integral part of the PyTorch Foundation, to build and shape the future of AI.



TF 2.18 released

An end-to-end platform for machine learning

Get started with TensorFlow

TensorFlow makes it easy to create ML models that can run in any environment. Learn how to use the intuitive APIs through interactive code samples.

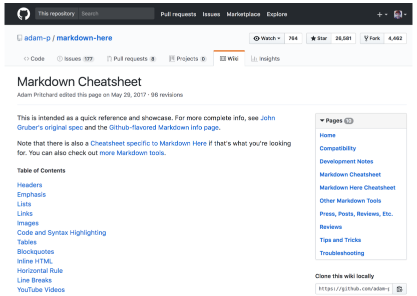
```
import tensorflow as tf
mnist = tf.keras.datasets.mnist

(x_train, y_train),(x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0

model = tf.keras.models.Sequential([
  tf.keras.layers.Flatten(input_shape=(28, 28)),
  tf.keras.layers.Dense(128, activation='relu'),
  tf.keras.layers.Dropout(0.2),
  tf.keras.layers.Dense(10, activation='softmax')
])
```

Project development

- Step 7: Report
  - Standard approach
    - Word/latex report
  - Modern approach
    - Use Python Notebook and Markdown: https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet
    - Future of scientific reports:
      - Code + description + analysis merged into a single document.
      - Code is reproducible, transferable to a new dataset, can be extended with new ideas.
  - You are free to select the mode of report.



Markdown Cheatsheet

Adam Pritchard edited this page on May 29, 2017 · 96 revisions

This is intended as a quick reference and showcase. For more complete info, see John Gruber's original spec and the Github-flavored Markdown info page.

Note that there is also a Cheatsheet specific to Markdown Here if that's what you're looking for. You can also check out more Markdown tools.

Table of Contents

Headers
Emphasis
Lists
Links
Images
Code and Syntax Highlighting
Tables
Blockquotes
Inline HTML
Horizontal Rule
Line Breaks
YouTube Videos

Project development

- Step 8: Video presentation
  - The project presentation must present concisely the project:
    - Project motivation and description, data acquisition, data exploration, pre-processing, proposed graph machine learning solutions, analysis of results, future development.
  - Each teammate must present her/his contribution to the project.
    - You will receive a project grade 0 if you do not present your contribution.
  - Use slides (one slide is 1-2min).
  - The length of the presentation is maximum 10min.
    - The time is strict, no more than 10min.
      - You will receive a project grade 0 if you video is beyond 11min.
    - Each member has 3-4min if your group size is 3, and 2min/member for a group size of 5.
  - Convince us you understood what you did !

Team communication

- Conflicts arise from few and lack of communication between teammates.
- It is strongly recommended the team (online) meet and discuss regularly the project status, the progress and the challenges faced.

Weekly monitoring & Zoom meeting

- Work weekly on the project.
  - Do not wait for the last weeks to start working on the project.
- We will monitor the project progress :
  - Each team must send a short update (s.a. one paragraph of 1-2 lines) of the week's progress to the TA allocated to your group.
  - Deadline : Every Friday by 6pm from week 8 to week 13.
  - Note that the update can be "no work done this week because of ---". It is fine, the update is not evaluated (it is for us to understand the overall project development).
  - You can use the update to ask for question(s) or a Zoom meeting with your TA.

Weekly monitoring & Zoom meeting

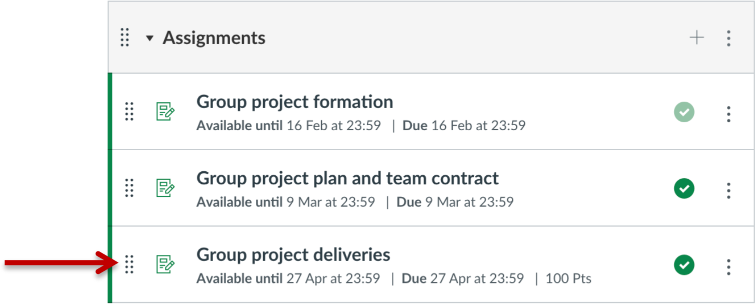
- It is required to have at least one Zoom meeting scheduled between you and the TA :
  - Ideally by Week 11, but multiple Zoom meetings (i.e. before/after Week 11) can be organized, as needed.
  - You are responsible for scheduling one Zoom meeting with the TA, e.g. in Week 11.
- Weekly updates and one zoom meeting count for 10 pts of the project grade.
- Note that the monitoring & Zoom meeting points are awarded independently of any content, making them easy to earn.

Marking scheme

- Project plan & team contract do not count.
- Weekly updates and one zoom meeting count for 10 pts.
- Steps 1-8 count for 65 pts.
- Anything that demonstrates initiatives will receive up to 25 pts additional points.

Project submission

- Submit notebook, report, presentation slides and video recording :
  - Canvas > Assignment > Group project deliveries
  - Create a .zip file with your notebook, report, presentation slides and video recording.
    - Use the format “project_groupID.zip” (for example project_group12.zip).
  - Note that the maximum upload file size is 500MB.
  - Contact your allocated TA if your submission is larger than 500MB.
  - Deadline : Sun Nov 23rd 2025 11:59pm (Week 14)
  - Penalty : 10% of the group grade per late day



Assignments

Group project formation
Available until 16 Feb at 23:59 | Due 16 Feb at 23:59

Group project plan and team contract
Available until 9 Mar at 23:59 | Due 9 Mar at 23:59

Group project deliveries
Available until 27 Apr at 23:59 | Due 27 Apr at 23:59 | 100 Pts

Deliveries and deadlines

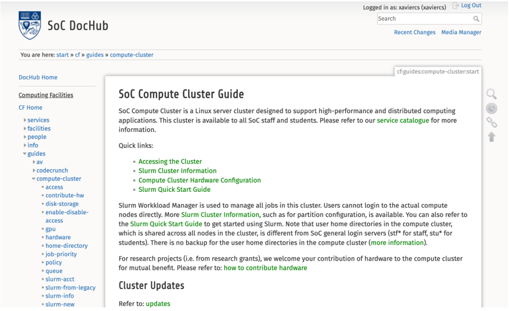
- Week 6 : Group formation, deadline: Sun Sep 21st 2025 11:59pm
- Week 7 : Project proposal and team contract, deadline: Sun Oct 5th 2025 11:59pm
- Week 14 :
  - A working/reproducible python notebook
  - Project report (it can be merged with the notebook)
  - Presentation slides and video recording (with e.g. Zoom)
  - Deadline : Sun Nov 23rd 2025 11:59pm
- Penalty : 10% of the group grade per late day

GPU

- The project should not require extensive experimentation with top-tier GPUs.
- If you need to run GPU, here are a few options :
  - SoC Compute Cluster (free but queue-based)
  - Google Colab (free version but limited to 12hr/24)
  - Google Cloud (first-time user receives USD 300)
  - Google Educational program (application pending)

SoC GPU

- To have access to SoC Compute Cluster, you need to do the following steps :
  - Create SoC account at : https://mysoc.nus.edu.sg/~newacct
  - Enable "SoC Compute Cluster" service for the SoC account at : https://mysoc.nus.edu.sg/~myacct/services.cgi
  - If accessing from outside SoC, you need to use SoC-VPN: https://dochub.comp.nus.edu.sg/cf/guides/network/vpn
  - SSH login to Slurm Login Node (xlogin) : ssh username@xlogin.comp.nus.edu.sg
  - Submit job to the Slurm Workload Manager : https://dochub.comp.nus.edu.sg/cf/guides/compute-cluster/slurm-quick
- More info about SoC Compute Cluster available at : https://dochub.comp.nus.edu.sg/cf/guides/compute-cluster
- List of GPU clusters available at : https://dochub.comp.nus.edu.sg/cf/guides/compute-cluster/hardware



You are here: start » cf » guides » compute-cluster

SoC Compute Cluster Guide

SoC Compute Cluster is a Linux server cluster designed to support high-performance and distributed computing applications. This cluster is available to all SoC staff and students. Please refer to our service catalogue for more information.

Quick links:

- Accessing the Cluster
- Slurm Cluster Information
- Compute Cluster Hardware Configuration
- Slurm Quick Start Guide

Slurm Workkload Manager is used to manage all jobs in this cluster. Users cannot login to the actual compute nodes directly. More Slurm Cluster Information, such as for partition configuration, is available. You can also refer to the Slurm Quick Start Guide to get started using Slurm. Note that user home directories in the compute cluster, which is shared across all nodes in the cluster, is different from SoC general login servers (stf* for staff, stu* for students). There is no backup for the user home directories in the compute cluster (more information).

For research projects (i.e. from research grants), we welcome your contribution of hardware to the compute cluster for mutual benefit. Please refer to: how to contribute hardware

Cluster Updates

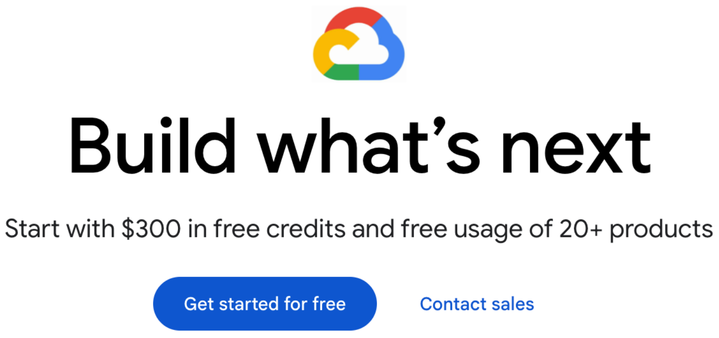
Refer to: updates

Google Colab GPU

- Google Colab
  - Free GPU, easy to use with Google Drive
  - Limited to 12hr/day
  - Colab Pro SGD 14.46/month, but compute it is too limited.
    - https://colab.research.google.com/signup
    - Tesla T4: 50 hours
    - Tesla V100: 20 hours
  - See slides : admin_week05_google_colab_gpu.pdf

Google Cloud GPU

- Google Cloud platform
  - https://cloud.google.com
  - Offers USD 300 / 150 hrs of free GPU (for first time user)
  - Instructions in setting up GPU available at :
    - admin_week05_google_cloud_gpu.pdf

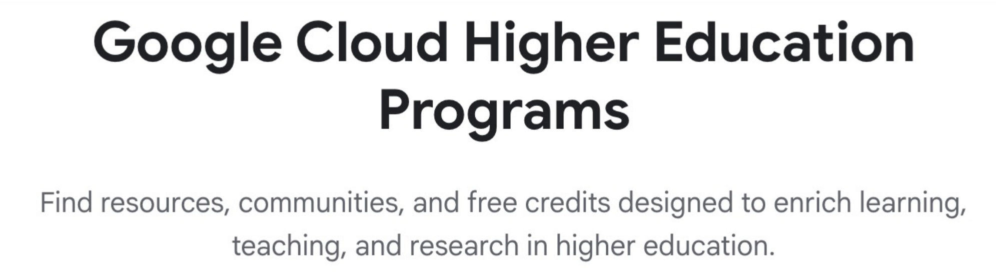


Build what’s next

Start with $300 in free credits and free usage of 20+ products

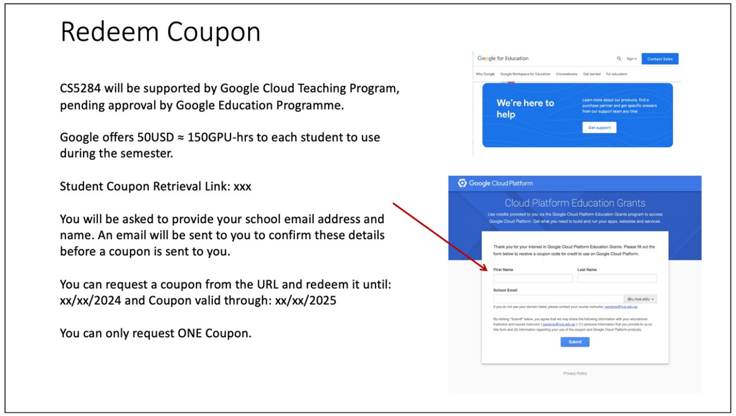
Google Education GPU

- Google Educational Program (application pending)
- USD50/student credit for 24/7 GPUs
- See slides : admin_week05_google_academic_gpu.pdf



Google Cloud Higher Education Programs

Find resources, communities, and free credits designed to enrich learning, teaching, and research in higher education.



Redeem Coupon

CS5284 will be supported by Google Cloud Teaching Program, pending approval by Google Education Programme.

Google offers 50USD ≈ 150GPU-hrs to each student to use during the semester.

Student Coupon Retrieval Link: xxx

You will be asked to provide your school email address and name. An email will be sent to you to confirm these details before a coupon is sent to you.

You can request a coupon from the URL and redeem it until: xx/xx/2024 and Coupon valid through: xx/xx/2025

You can only request ONE Coupon.

Learn more about our products, find a purchase partner and get specific answers from our support team any time

Use credits provided to you via the Google Cloud Platform Education Grants program to access Google Cloud Platform. Get what you need to build and run your apps, websites and services.

Thank you for your interest in Google Cloud Platform Education Grants. Please fill out the form below to receive a coupon code for credit to use on Google Cloud Platform.

If you do not see your domain listed, please contact your course instructor: xaviercs@nus.edu.sg

By clicking "Submit" below, you agree that we may share the following information with your educational institution and course instructor (xaviercs@nus.edu.sg): (1) personal information that you provide to us on this form and (2) information regarding your use of the coupon and Google Cloud Platform products.

Teaching assistants

- Teaching assistants are available to support the development of your project as best as possible.
- When the group ID are announced, then
  - Ask the TA in charge of your group for any questions.
  - Do not hesitate to communicate with them to clarify anything.
- Reminder : Group TA $\neq$ Tutorial TA

Midterm

Midterm

- **I will present the Midterm next week.**

Questions?