CS5284 : Graph Machine Learning

Administrative (Week 6)

Semester 1 2025/26

Xavier Bresson

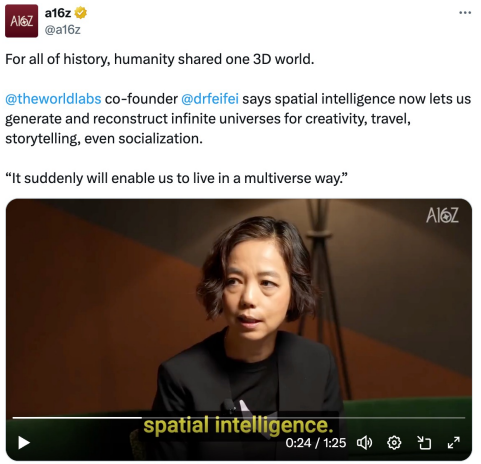
https://x.com/xbresson

Department of Computer Science
National University of Singapore (NUS)

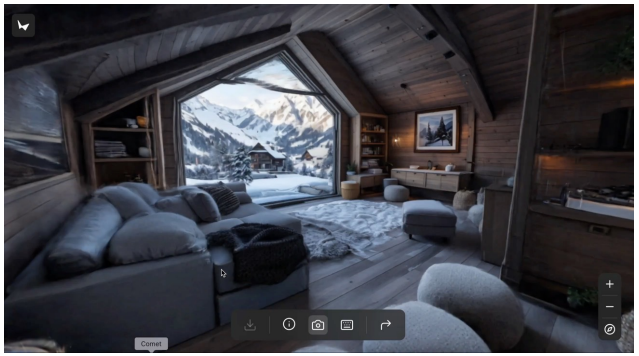
QR Attendance

Please, scan the new below QR image for attendance.

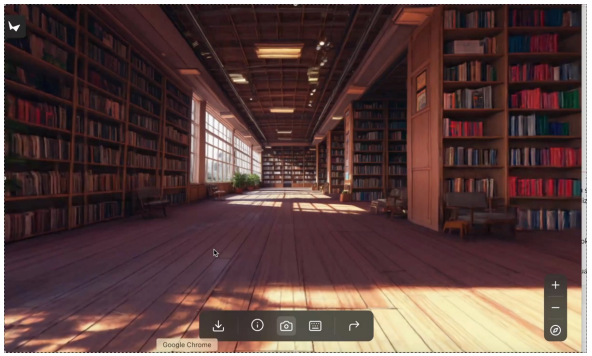
After language intelligence, spatial intelligence is the next AI frontier!
How to make AI agents understand the world? For self-driving cars, robots, etc
Challenges are object persistent, understanding physics, and visual reasoning with common sense.
https://www.worldlabs.ai/about
Researchers have developed Large Visual Models (LVMs), but does modeling the 3D world as a sequence of tokens the solution?



https://www.youtube.com/watch?v=_PioN-CpOP0



Prompt from a single image



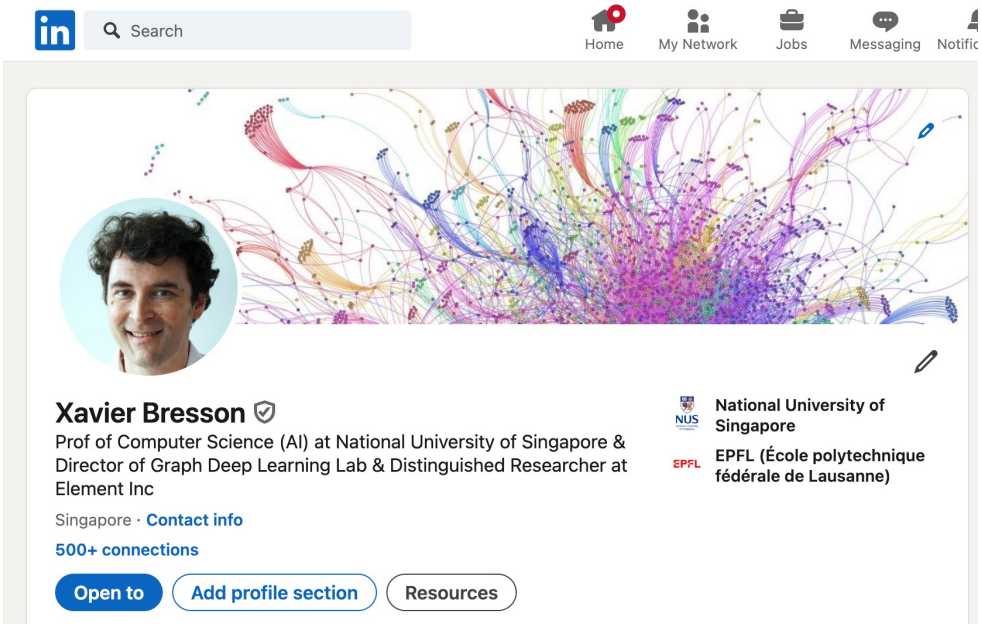
Prompt from a single image

Stay connected

Feel free to send me a LinkedIn invitation with the note "CS5284 GML"

so I can recognize that you are part of the course ☺

https://www.linkedin.com/in/xavier-bresson-738585b

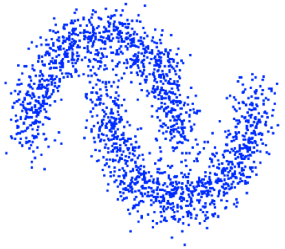


In-lecture questions

In-lecture question [Answer]

- **Propose an alternative graph regularization loss to the Dirichlet loss defined as:**

$$\int_{\mathcal{M}} |\nabla f|^2 \approx f^T L f = \sum_{ij \in V} A_{ij} |f(x_i) - f(x_j)|^2$$

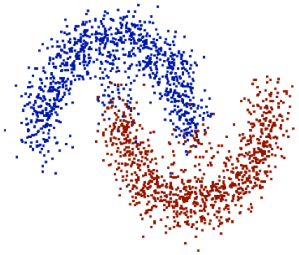


- **In Slack #lectures**
  - **Identify the question and Reply in thread with a short response**

- **Answer :**

$$L_{\text{graph}}(f, p) = \int_{\mathcal{M}} |\nabla f|_p^p \quad \text{p-Dirichlet energy}$$

$$\approx \sum_{ij \in V} A_{ij} |f(x_i) - f(x_j)|^p \quad \text{discrete p-Dirichlet energy}$$

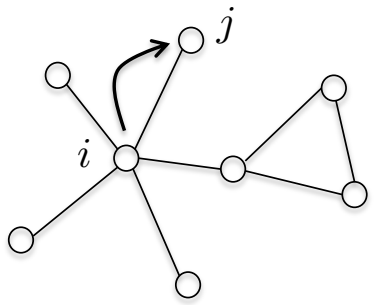


Specifically, $L_{\text{graph}}(f, 1) = \sum_{ij \in V} A_{ij} |f(x_i) - f(x_j)|$ Total Variation energy

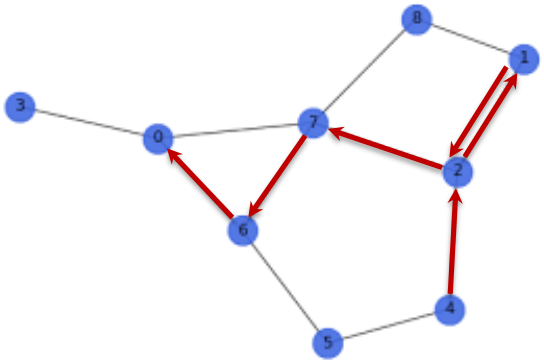
Proven to provide better solutions than Dirichlet.

In-lecture question [Answer]

- How can you interpret the stochastic matrix A? What does A_{ij} represent, and how can you design a random walk algorithm from A? (A random walk is a path that consists of a succession of random steps on some mathematical space.)
- In Slack #lectures
  - Identify the question and Reply in thread with a short response
- Answer : The matrix element A_{ij} represents the probability of transitioning from vertex i to vertex j in the graph domain. A random walk algorithm can be constructed by iteratively making random moves from node i to node j, with the transitions determined by sampling from the probability distribution given by A_{ij}. For example, this can be implemented using Bernoulli sampling to decide the next step based on the transition probabilities.



$$A_{ij} = \Pr(i \to j)$$



Student questions

Student question

From the graph SVM example in Lecture 4 Lab 4, I tried different values of k in k-NN graph construction, and tried different distance metrics, and many of them do not work. Is Graph SVM very sensitive to hyperparameters in practice?

- This is right, and true for most machine learning models – They require manual setting of hyper-parameters, which is often a trial-and-error process.
- A solution is to use cross-validation and grid search of a set of values for the hyper-parameters, but this is computationally expensive and hence not used with modern deep learning techniques.
- For graph SVM, the heuristic is to construct a graph with k-NN model that provides a good clustering of the data distribution, and then use it with kernel SVM.

Student question

In Lecture 4 Lab 4 solution, I found out that H is the diagonal matrix whose corresponding entries are 1 if the data is labelled and 0 if the data is not labelled.

L is the diagonal matrix whose corresponding entries can be either (i) or (ii), defined as the following:

(i) 1 or -1, corresponding to the two labelled classes

(ii) 0 if the data is not labelled.

Since both H and L are diagonal matrices, multiplying them would be equivalent to element-wise multiplication along the diagonals of H and L.

Therefore, I observe that H = L^2.

Therefore, LH = HL = L^3. Since the entries in L are either -1, 0 or 1, L^3=L.

Hence, $Q = LHK(I + \gamma \mathcal{L}K)^{-1}HL = LK(I + \gamma \mathcal{L}K)^{-1}L$, i.e. H is redundant. Likewise, for the solution ξ^*, which contains an HL term in the product that can be simplified to L.

This led me to think why H is added in the lecture slides.

I believe that H is there as a weighting factor for labelled data, with higher values in the diagonal meaning higher importance, and in the unweighted case, as explained above, H is redundant.

Is my understanding correct?

$$\min_{f \in \mathcal{H}_K} \; f^T K f + \lambda \sum_{i=1}^{n} L_{\text{Hin}}(f_i, \ell_i) + \gamma f^T \mathcal{L} f$$

with

$$\text{Representer theorem} : \; f(x) = \text{sign}\Big(\sum_{i=1}^{n} \xi_i^\star K(x, x_i) + b \Big) \in \pm 1$$

$$\text{Optimization problem} : \; \alpha^\star = \arg \min_{0 \le \alpha \le \lambda} \; \frac{1}{2} \alpha^T Q \alpha - \alpha^T 1_n \;\; \text{s.t.} \;\; \alpha^T \ell = 0$$

$$\text{with } Q = LHK(\mathrm{I} + \gamma \mathcal{L}K)^{-1}HL \in \mathbb{R}^{n \times n}$$

$$\text{Solution} : \; \xi^\star = (\mathrm{I} + \gamma \mathcal{L}K)^{-1} HL \alpha^\star$$

- **Yes, this is correct and additionally, it controls the data that are unlabeled with value 0.**

$$L = \text{diag}(\ell_1, \ell_2, ..., \ell_n) \;\; \text{e.g.} \;\; L = \text{diag}(-1, +1, -1, -1, -2, -2, +5, +5, ...)$$
$$H = \text{diag}(h_1, h_2, ..., h_n) \;\; \text{e.g.} \;\; H = \text{diag}(1, 1, 0.5, 0.1, 0, 0, ...)$$

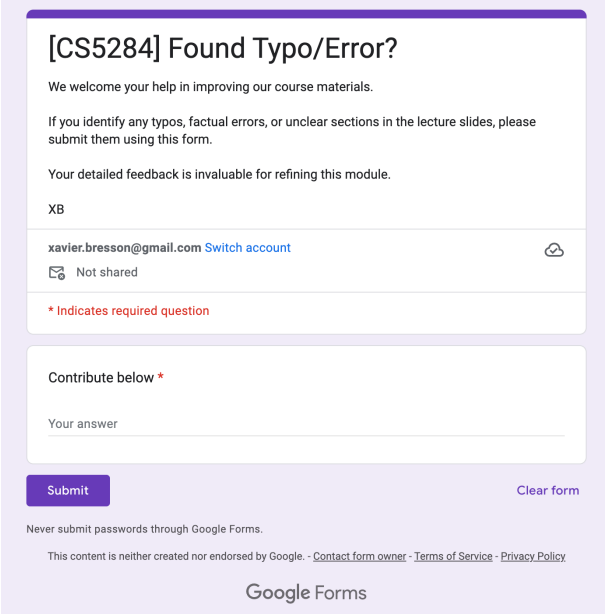
Admin

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.**

Contribute to the module

- Found a typo/error? Help us improve the course!
- https://forms.gle/HeCNfPEdqbHn28m3A
- Thank you for your service.

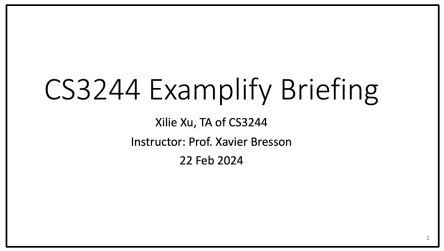


Midterm

Midterm

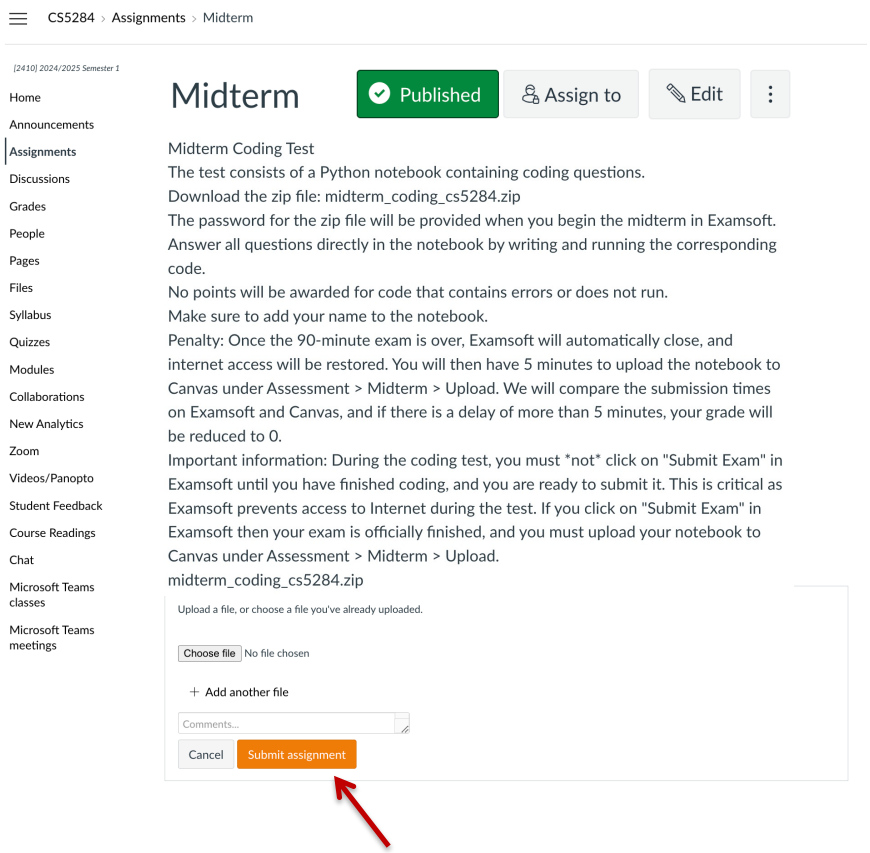
- Date & Time: Wed Oct 1st (Week 7), 6:45pm-8:15pm
- Duration: 90 minutes
- Location: LT19 (Seating map will be available at the entrance)
- 50% Multiple Response Questions (MRQ) via Examplify
- 50% Coding Questions in local Python environment
- Exam Conditions
  - Open book, NO Internet access
  - You may use anything available on your computer e.g. calculator, PDFs, except LLMs :)
  - You cannot use physical calculator, phone, etc
- Topics Covered: Everything up to "Lecture 5: Recommendation on Graphs" (including all material s.a. lectures, tutorials, etc.)
- Important
  - Ensure your laptop is fully charged.
  - It is highly recommended that you test Examplify on your device before the exam to avoid any conflicts. Some students faced issues last year with Examplify compatibility.
- If you have any questions about the exam, please contact the TA assigned to your tutorial group.

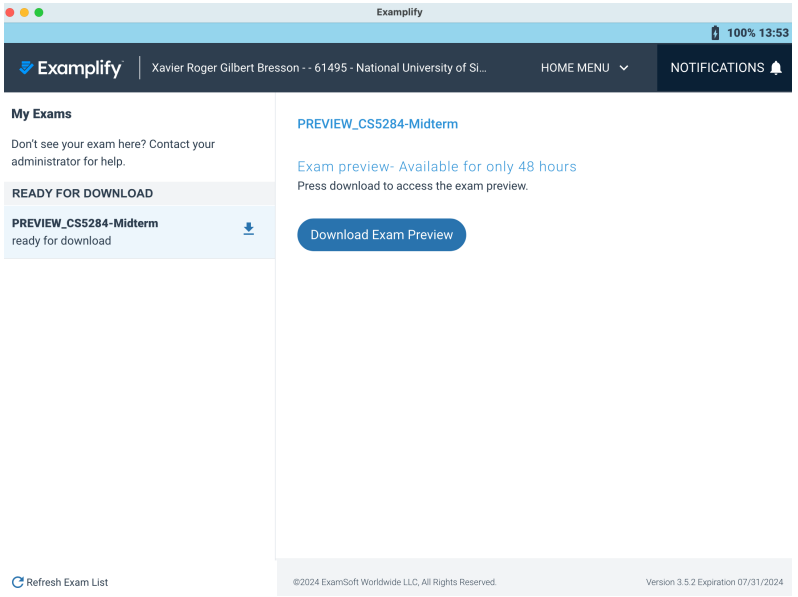
Midterm



- Examplify briefing is available in Canvas>Home>Week 6>
  admin_week06_examplify_briefing.pdf
- It is highly recommended to read these slides if you have never used Examplify before.
- If you have used Examplify previously, it is still strongly recommended to test it again on your laptop before the exam. Several students experienced conflicts last year between their system and the latest version of Examplify.
- Responsibility
  - You are fully responsible for ensuring that Examplify works on your device.
  - Test Examplify on your laptop at least 3 days before the exam, and if you encounter any issues, please contact the TA in charge of your tutorial group.
- Important
  - You will receive a grade of 0 if you are unable to use Examplify on the day of the exam.
- Prohibited devices
  - Tablets such as iPads are not allowed for the exam due to compatibility issues with Examsoft.

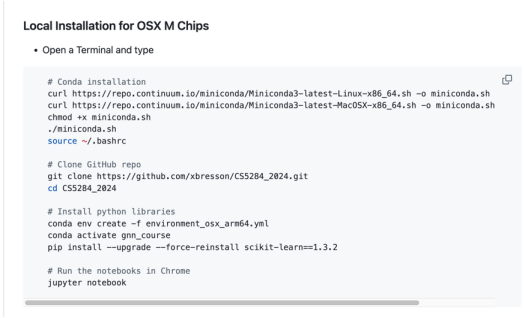
Midterm


CS5284 › Assignments › Midterm
[2410] 2024/2025 Semester 1
Home
Announcements
Assignments
Discussions
Grades
People
Pages
Files
Syllabus
Quizzes
Modules
Collaborations
New Analytics
Zoom
Videos/Panopto
Student Feedback
Course Readings
Chat
Microsoft Teams classes
Microsoft Teams meetings
Midterm
Published
Assign to
Edit
Midterm Coding Test
The test consists of a Python notebook containing coding questions.
Download the zip file: midterm_coding_cs5284.zip
The password for the zip file will be provided when you begin the midterm in Examsoft.
Answer all questions directly in the notebook by writing and running the corresponding code.
No points will be awarded for code that contains errors or does not run.
Make sure to add your name to the notebook.
Penalty: Once the 90-minute exam is over, Examsoft will automatically close, and internet access will be restored. You will then have 5 minutes to upload the notebook to Canvas under Assessment > Midterm > Upload. We will compare the submission times on Examsoft and Canvas, and if there is a delay of more than 5 minutes, your grade will be reduced to 0.
Important information: During the coding test, you must *not* click on "Submit Exam" in Examsoft until you have finished coding, and you are ready to submit it. This is critical as Examsoft prevents access to Internet during the test. If you click on "Submit Exam" in Examsoft then your exam is officially finished, and you must upload your notebook to Canvas under Assessment > Midterm > Upload.
midterm_coding_cs5284.zip
Upload a file, or choose a file you've already uploaded.
Choose file
No file chosen
Add another file
Comments...
Cancel
Submit assignment


Examplify
100% 13:53
Examplify
Xavier Roger Gilbert Bresson - - 61495 - National University of Si...
HOME MENU
NOTIFICATIONS
My Exams
Don't see your exam here? Contact your administrator for help.
READY FOR DOWNLOAD
PREVIEW_CS5284-Midterm
ready for download
PREVIEW_CS5284-Midterm
Exam preview- Available for only 48 hours
Press download to access the exam preview.
Download Exam Preview
Refresh Exam List
©2024 ExamSoft Worldwide LLC, All Rights Reserved.
Version 3.5.2 Expiration 07/31/2024

Midterm

- Detailed Python local installation steps are available at Canvas>Home>Week 3> admin_week03_local_installation.pdf
- Environment setup
  - The test notebook is designed for the course's Python environment: https://github.com/xbresson/CS5284_2025
  - Ensure that you are using the same package versions to avoid any conflicts during the test.



- Responsibility
  - It is your responsibility to ensure that you can successfully run the Python notebook on your local machine.
- Important
  - Do not attempt to set up Python on the day of the test!
- If issue
  - Contact the TA in charge of your tutorial group asap if you encounter any problems.
- Consequence
  - You will receive a grade of 0 if you are unable to run the exam notebook.

Midterm

- Samples of MRQ and Coding questions are available at Canvas>Home>Week06>
  - MRQ_sample.zip
  - coding_sample.zip

CS5284 Graph Machine Learning

MRQ Sample

Q: Which of the following statements are true? Select all responses that apply (i.e. you may select 0 response, 1 response, 2 responses, etc).

a) In general, the curse of dimensionality, i.e. high-dimensional features (e.g. text data with thousands of word features), degrades the performance of machine learning techniques.
b) Kernel techniques mitigate the curse of dimensionality by better adapting their predictions to the data distribution.
c) In spectral kernel K-means, we select the eigenvectors associated with the k smallest eigenvalues of $\theta^{-1/2}K\theta^{-1/2}$ (where K is the kernel matrix and θ is a weighting matrix), while for spectral Normalized Cuts we choose the eigenvectors associated with the k largest eigenvalues of the graph Laplacian $L = I - D^{-1/2}AD^{-1/2}$ (where A is the graph adjacency matrix and D is the degree matrix).
d) For large graphs, it is critical to have a very large degree value for each node to capture robust graph information, e.g. for n in the order of a million nodes, the mean degree should be at least a thousand.

Answer :

CS5284 Graph Machine Learning

Coding sample

Instructions

Name: Please, add your name here : e.g. JOHN SMITH

Answers: Please write your answers directly in this notebook by completing the code sections marked with

`# YOUR CODE STARTS HERE`

`# YOUR CODE` (it can span one or multiple lines)

`# YOUR CODE ENDS HERE`.

Remark: If certain conditions of the questions (for eg. hyperparameter values) are not stated, you are free to choose anything you want.

Exercise : Graph Clustering with World Happiness Dataset

Import libraries and utility functions

```
In [ ]: %reset -f
import datetime
print('Timestamp:',datetime.datetime.now().strftime("%y-%m-%d---%H-%M-%S"))
import numpy as np
import scipy.sparse
import pandas as pd
import sklearn.metrics.pairwise
import networkx as nx
import sys; sys.path.insert(0, 'lib/')
from lib.utils import compute_ncut
from lib.utils import interactive_vis_graph
from lib.utils import print_neighboring_countries
from lib.utils import get_same_label_countries
```

Load and visualize the dataset

```
In [ ]: wh = pd.read_csv("datasets/2019.csv") # load the dataset
# Features GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices, Generosity, Perceptions of corruption
data = wh.iloc[:,3:].to_numpy()
print('Number of data points is', data.shape[0], 'and the number of features is', data.shape[1])
print('\nFirst three data points:\n', str(data[:3]))
print('\nData statistics')
wh.describe()
```

Question 1 : Data normalization

Normalize the dataset `data` so that it has a mean of zero and a standard deviation of one along each data dimension.

After normalization, print both the mean and standard deviation of the normalized dataset.

Hint: You can use the functions `numpy.mean()` and `numpy.std()` for this task.

Midterm

- Step-by-step process on exam day (Wed Oct 1st) :
- Use Examsoft to download the midterm exam (available Oct 1st, noon). Refer to "admin_week06_examplify_briefing.pdf" for detailed instructions.
- Download the Python notebook from Canvas>Assessments>Midterm>midterm_coding_cs5242.zip (available Oct 1st, noon).
- The password for the zip file will be provided in Examsoft > Q1.
- Respond to all questions directly in the notebook by writing and running the corresponding code.
- Ensure your code runs without errors. No points will be awarded for buggy or incomplete code.
- Don't forget to add your name to the notebook.
- Complete all MRQs in Examsoft. Do NOT submit until you finish the notebook questions (to remain offline), or your coding exam will officially finish.
- Submission & penalty: After the 90-minute exam concludes, Examsoft will automatically close, and your internet connection will be restored. You will have 5 minutes to upload your completed notebook (e.g. midterm_coding_cs5284_john_smith.ipynb) to Canvas under Assessment>Midterm>Upload. We will compare the submission times between Examsoft and Canvas. Any delay of more than 5 minutes will result in your grade being reduced to 0.

Questions?


SO, IS THIS GOING TO BE ON THE TEST??
HOW TO ANSWER:
IF I SAY YES
IF I SAY NO
THAT'S TOO EASY.
THEY WON'T LEARN IT.
BUT THEN IF IT'S NOT ON THE TEST
IF I SAY MAYBE OR ANSWER VAGUELY
THEY'LL COMPLAIN
IT WASN'T GOING TO BE, BUT THANKS FOR THE SUGGESTION...
NOOOO!!
heh heh.
JORGE CHAM © 2016
WWW.PHDCOMICS.COM