Introduction

You are expected to work in a group to survey and select suitable data sets (examples are given below). You are also expected to apply the data analysis skills which have been covered in this course. You are advised to make use of data computing solutions (you have learned in the course) for understanding and visualizing data with parallel computing elements. Please try to think about the data creatively and present your findings in your report and presentation in a concise manner.

- Group Size : 3 to 4 (any group with more than 4 students will be dismissed without any prior notice)
- Size of data set : Should be sufficiently reasonable to observe interesting/significant findings

Deliverables and Grading

- **Deliverables** : One presentation (30%)
  One report (70%)
- **Contents:**
  The content skeleton of your work should follow or be modified from the suggested *Sample Project* below, plus any interesting discussions related to our learning in this course.
- **Grading:**
  In your group report, please state the contribution (%) and brief details for each member clearly. e.g.

LEE Siu Ming (smchan123)	: 40% (Data set preparation, analysis, reporting)
WONG Chi Keung (ckwong456)	: 25% (Background research, analysis)
ZHANG Jiecong (jzhang78)	: 35% (Data set preparation, reporting, coordination)

  Each group will give a 10 minutes presentation. Another 5 minutes are reserved for questions from the audience and the instructor.

  During the presentation, each member should present a speech with good contents and illustrations. Please also note the quality of the language (i.e. spoken English and grammar) and presentation styles (body language, confidence, organization and flow of the talk etc.)

  Report format: At least 10 pages long. Please use reasonable font size and margins.

  Group score or individual score?
  Each student will receive a score which basically reflects the quality of the work of the whole group, with adjustments according to *individual performance*. Such *individual performance* is assessed by the contribution stated in the report, the presentation quality, and Q/A performance.
  Note: Be prepared that the instructor may name a particular group member to answer some questions.
- **Submissions:**
  - After your presentation, you should submit the following files to Canvas before the deadline:
    Final report, presentation slides, data sets, reference papers (if any), and other materials (if any)

Sample Project (not limited to this one)

"Parallel Data Computing Solutions to Hong Kong Real Estate Data"

1. Collect the Hong Kong real estate data from several sources (e.g. https://data.gov.hk/en/).
   - Document the source of the data clearly in the report.
2. Preprocess and Visualize the data with histograms, scatterplots, and other diagrams you have learned;
   - Preprocess the data so that you can visualize it.
   - Implement data visualizations so that we know better about the data.
3. Analyze the data and discuss your own findings
   - Perform advanced analysis on the data (e.g. data clustering and association rule mining)
   - Explain the findings, and try to make conjectures about the findings you obtained.
4. Discuss how parallel computing is applied to accelerate the data computing process
   - Describe what kind of parallel computing strategy you have implemented (e.g. parallel for loop)
   - Explain why such a parallel computing strategy has been adopted (e.g. memory hierarchy)
5. Conclusion and Future work.
   - State your conclusions and the related pros / cons.
   - If you have enough time, what you can do? What problems are there to be investigated further?

Milestones and Schedules

- Before Week 4: Join a group on Canvas. (Leftovers will be assigned randomly)
- Week 4: Grouping confirmed -- no more change of grouping.
- Week 6: You are advised to confirm a project topic with your groupmates for work load distribution.
- Week 12-13: Project presentation during lecture time. The presentation schedule will be randomly assigned.
- Week 13: Submit all final deliverables on Canvas.

Project Presentation Attendance

Project presentation attendance will not be recorded but you are highly encouraged to join and learn from the group projects of your classmates and to support their hard works during their project presentations.

Possible Data Sources

(You are encouraged to find your own datasets you are interested in; below are just examples that you can choose.)

Hong Kong Government Data: https://data.gov.hk/en/
US Government Data: https://www.data.gov/
Singapore Government Data: https://data.gov.sg/
UC Irvine Machine Learning Repository: http://archive.ics.uci.edu/ml/
Panama Papers Graph Data (i.e. Network): https://github.com/amaboura/panama-papers-dataset-2016
Stanford Large Network Dataset Collection: https://snap.stanford.edu/data/
Offshore Leaks Database (i.e. Text Data): https://offshoreleaks.icij.org/

Miscellaneous:
https://toolbox.google.com/datasetsearch
http://www.kdnuggets.com/2011/02/free-public-datasets.html
https://r-dir.com/reference/datasets.html
https://www.springboard.com/blog/free-public-data-sets-data-science-project/
http://www.datasciencecentral.com/page/search?q=data+sets

Example Project Topics

(Your own project topics/ideas are encouraged and preferred.)

Analyze factors relating the gaming performance in League of Legends
Exploration of Factors Relating to Movie Box Office Performance
Historical Buildings in Hong Kong
FIFA players' statistics and Professional Football Clubs' Seasonal Performance
A visual exploration of aircraft crashes since 1908
NBA in Data: An analytical report on Los Angeles Lakers
Hong Kong Housing Trend
Gastronomy and Ingredients Matching Across the World
Exploring of factors relating to League of Legend world championship performance
The frequency of earthquakes
Homeless, Hong Kong
The Relationship among Gender, Education and Employment in Hong Kong
Renewable energy in the European Union
Flight Networking and On-time Performance Analysis
Secondary School in Hong Kong
World University Rankings and Statistics
Exploring currency exchange rate
Mass Shooting in America
An evaluation of workplace environment in Hong Kong
Shootings in NBA
Exploration of typhoon in Hong Kong in 21st century
IMDB Movie Analysis
Data mining in conditions and predictions of G20 countries by continent
The Analysis of Mandatory Provident Fund (MPF) Schemes
Understanding people's reactions to new movies cia Twitter and film review websites
Mobile Application (ios and android system) Ranking and the relevant factors on America market
Unemployment rate and major indices of US, Germany and Japan
Analysis on the 2016 Legislative Council Election
Analysis of Factors Affecting Global Temperature Rise

--- END OF DOCUMENT ---