BMEG 3105 Fall 2022

Data analytics for personalized genomics and precision medicine

Course introduction

Lecturer: Yu LI (李煜) from CSE

Liyu95.com, liyu@cse.cuhk.edu.hk

Monday, 5 September 2022

Expected outcomes:

- Make good use of the resulting huge datasets to understand basic biology and medical conditions
- Apply AI and machine learning in clinical applications at the system level
- Learn the fundamental concept of data analytics
- Know the various data in genomics and medicine
- Apply the data analytics techniques to process the data and resolve problems in biology

Pre-course survey results:

Interested topics: 1. Biomedical imaging 2. Neural networks 3. Dynamic programming 4. Clustering 5. Data visualization	**Driver of taking this course:** 1. For data analytics/machine learning techniques 2. Exploring new fields 3. For biological/genomics/health applications 4. For knowledge and research experience
Help may be needed for: 1. Programming 2. Data analytics concepts 3. Practical application 4. Finding additional materials and resources	**What would be provided:** 1. Basic Python programming tutorials 2. Data analytics fundamentals 3. Application of techniques 4. Necessary biology concepts 5. Additional resources and materials, at the end of each lecture. All online, reading lists may be provided

Course logistics:

Lectures: Wed 9:30am-11:15am (11:05am), SC-L4 Fri 9:30am-10:15am, LSB-LT3
Tutorial: Fri 10:30-11:15am, LSB-LT3

- ✓ Tutorial sessions on Python programming (Exact schedule determined by the TA)
- ✓ Slides available the day before the lecture (On course website)
- ✓ Video recordings uploaded onto YouTube & Bilibili (With many other interesting lecture videos)

Prof. Yu LI's channels: https://space.bilibili.com/605551118
https://www.youtube.com/channel/UCcKnGmzsZ_PSPUSm0RZdVNA

Teaching team:

Yu Li: liyu@cse.cuhk.edu.hk Office hours: 3pm-5pm, Friday Location: SHB-106	Qinze Yu: qzyu22@cse.cuhk.edu.hk Office hours: 2pm-4pm, Monday Location: SHB-116
Yixuan Wang: yxwang@cse.cuhk.edu.hk Office hours: 2pm-4pm, Tuesday Location: SHB-116	Liang Hong: lhong22@cse.cuhk.edu.hk Office hours: 1pm-3pm, Monday Location: SHB-116

Grading:

- **Midterm (20%)**: **Oct 21 (Fri) 8:30am-11:15am.** In-class, **one bonus question (2%).**
- **Final (20%)**: TBA
- **In-class quizzes (10%)**: **Oct 19 (Wed) & Nov 30 (Wed).** The questions will be simple. Mainly for checking the participation.

> ❖ **All exams & quizzes:**
> - ✓ **Open-book, paper-based materials allowed**
> - × Phone/computer/other communication tools
> - × Discussion

- **Individual project (20%):**
  - ❖ **Potential projects:**
    - ✧ From reads to gene expression matrix processing pipeline.
    - ✧ Gene expression matrix processing pipeline.
    - ✧ Single-cell RNA-seq processing pipeline.
    - ✧ Bio-image classification.
    - ✧ Cancer gene identification.
    - ✧ Gene enrichment analysis.

    **Or give the teaching team your project and seek help*

  - ❖ **Project milestone report (5%): Due: Nov 9.** One page.
    1. *Title, author*
    2. *What problem do you want to do? Why is the problem interesting? (1%)*
    3. *What data are you going to process? The source, the size, the sample of the data (1%)*
    4. *What's the output of your method? (1%)*
    5. *How are you going to do it? Describe the method step by step, from input to output(1%)*
    6. *What are the expected results? How are you going to evaluate the results? (1%)*
    7. *What have you done?*

  - ❖ **Project Final report (7%): Due: Dec 2.** No length requirement.
    *Submit together with **codes** (**5%**, whether it is correct or not).

    1. *Title, author*
    2. *What problem do you want to do? Why is the problem interesting and important? (0.5%)*
    3. *What data have you processed? The source, the size, the sample of the data (0.5%)*
    4. *What have you done to resolve the problem? Describe the method step by step, from input to output (2%)*
    5. *What are the results? (1.5%)*
    6. *Result evaluation (1.5%)*
    7. *Any idea for further improvement? (1%)*

  - ❖ **Project Presentation (3%): Nov 25 & Dec2**, together with the tutorial course. 5 mins each.
    1. Logic (1%)
       - ✧ *What is the problem?*
       - ✧ *Why is it important?*
       - ✧ *How do you resolve it? Overview of your idea*
       - ✧ *Overview of the results*
    2. Clarity (1%)
       - ✧ *Whether the audience can understand and follow the presentation*
    3. Slides preparation (1%)
       - ➢ *Clear illustration*
       - ➢ *No typos, no grammar errors*

- **Homework (20%):**
  - ❑ **Programming Assignment 0:** Programming environment setup
    Posted: Sep 9 Due: Sep 14 **(No need to submit anything)**
  - ❑ **Assignment 1 (5%):** About the basic concept of data analytics-1
    Posted: Sep 16 **Due: Sep 28**
  - ❑ **Assignment 2 (5%):** About the basic concept of data analytics-2
    Posted: Oct 5 **Due: Oct 19**
  - ❑ **Programming Assignment 1 (5%):** About application of DA to biology **(Non-grading)**
    Posted: Oct 26 **Due: Nov 16**
    The assignment will cover the entire second module. So you can do part of it after each lecture
  - ❑ **Assignment 3 (5%):** DA in Personalized Genomics and Precision Medicine
    Posted: Nov 16 **Due: Nov 25**

- **Scribing (10%)**: Summarize one of the lectures. **Submit it within one week after the course.** Each student should do **at least one** lecture. Each lecture should have at least one scribing. **You can sign up for at most two, for additional 1%. Your note and scribing will be posted online, for others reference. You can choose to remove your name or not.**
  - ✧ Scribing sign up link:
    https://docs.google.com/spreadsheets/d/1lA_eLCQawKLG2IWNPJ3gDa9bKp9adOZcQMdb_ciLgEg/edit?usp=sharing
  - ✧ Deadline for signing up for scribing: **11:59 pm on 12th Sep**

- **Bonus (up to 6%)**:
  - ✧ One bonus question in Midterm: 2%.
  - ✧ One additional scribing: 1%.
  - ✧ **Pre-course survey** + **post-lecture survey**: 0.5% each, maximum 3%.

  Post-lecture survey submit deadline:** **11:59pm on the day before the next lecture

  ***Survey results**: Note down if you did the survey, for confirmation & scoring*
  https://docs.google.com/spreadsheets/d/1D6fW_VH6zv6gjejbQ6u5PGJIWbgWpFe-sDG-TuiFnsI/edit?usp=sharing

  **Course might be adjusted according to the surveys. Feedback and suggestions are encouraged.*

- **Late days:**
  - ✓ For A1, A2, A3, PA1, project mid-term report
  - × **For final project report & scribing report**
  - ✧ 6 late days total, **2 max for any assignment**
  - ✧ Grades will be deducted by 25% for each additional late day

Brief overview of DATA in personalized genomics and precision medicine: What we can use to model a human being and thus provide better healthcare

- ✓ Data analytics is useful:
  - ✧ Aggregate data
  - ✧ Generate hypothesis
  - ✧ Support conclusion
- ✓ Computers have become cheaper and more powerful
- ✓ Sequencing cost decreasing
- ✓ Single cell data accumulating

Methods for measuring a person:

1. Gene and mutations
2. Gene expression (Transcriptome)
3. Proteome
4. Metabolome
5. Molecular network & Cellular network
6. Microbiome (Oral and gut)
7. Organ (Biomedical imaging)
8. Hospital test (Blood test and so on)
9. Electrocardiography(ECG)
10. Demographic information (Age, gender, location and so on)
11. Drug history and disease history
12. Personal statement and doctor diagnosis
13. Living habit (Exercise)
14. Diet
15. Family history
16. Communications and social media data
17. Environment (Pollution)
18. Travel history (Global pandemic)

Useful links:

- Course website: https://yu979.github.io/BMEG3105-Spring-2022/
- Piazza: http://www.piazza.com/cuhk.edu.hk/fall2022/bmeg3105

  **Ask questions through Piazza (anonymous allowed)/email, or visit in-person during office hours*
  **For personal matters, please send private posts*

- Zoom link: https://cuhk.zoom.us/j/96301008009?pwd=V0dkcEgzdHMzcmljdEIxaHV1TFNkQT09

  **Please inform teaching team if absent because of pandemic*

- Python & Colab: https://colab.research.google.com/notebooks/intro.ipynb
- Scribing preference sign up: **Due 11:59 pm on 12th Sep** https://docs.google.com/spreadsheets/d/1lA_eLCQawKLG2IWNPJ3gDa9bKp9adOZcQMdb_ciLgEg/edit?usp=sharing
- Post-course survey (Week1, Intro): https://docs.google.com/forms/d/e/1FAIpQLSdRGoChZyAmWQbLIIZ5s7VvAyQkG_FZ5I9AccxBcRCj7b9HdQ/viewform
- Survey results: Note down if you did the survey, for confirmation & scoring https://docs.google.com/spreadsheets/d/1D6fW_VH6zv6gjejbQ6u5PGJIWbgWpFe-sDG-TuiFnsI/edit?usp=sharing