BMEG 3105
Data Analytics for Personalized Genomics and Precision Medicine
Lecture 1 Course Introduction

Lecturer: Professor Yu LI *Scribe: JIA Zihan*

Course Outline of Lecture 1:

- Review the pre-course survey results
- Course Logistics
- Brief overview of DATA in personalized genomics and precision medicine

I. Pre-course Survey Results Description

1.1 Classmates' Information

There are totally 46 students attend BMEG3105 this year, among which 28 students submitted the questionnaire. The whole class is mainly composed of BME students (75%), and other major include CS, Biology, AIST, CDAS, Cell and Molecular Biology as well as QFRM. Among these people most of them are year 3 undergraduate students (60.7%), some also come from year 4 (25%) and year 2 (10.7%).

1.2 Students' Backgrounds and Needs

According to the statistics, these students have a relatively poor foundation in machine learning (average 1.679/5) and algorithms (1.929/5). They perform better in subjects such as the probability & statistics (3.0/5) and biology (3.036/5). As for their aim to register this course, most of them are interested in biological/genomics/health applications, as well as data analytics/machine learning techniques. Briefly speaking, the need for mathematics, concepts in data analytics and programming is relatively greater than other basic knowledges like biology.

1.3 Teaching Team's Solution

To satisfy these needs and find out the solution, the teaching team will provide basic Python programming tutorials to students. The key concepts in data analytics will be introduced, together with some additional recourses and materials at the end of each lecture. Mathematics will not be heavily emphasized as it's not easy to set up in a short period of time. Moreover, some biology may be introduced as needed.

II. Course Content Description and Logistics

2.1 Meeting Dates & Teaching Staff

All the Lecture slides will be available the day before the lecture day, students can easily find them on the **course website** https://lim-zq.github.io/BMEG3105-Fall-2025/.

Meanwhile, from this year no video recording will be provided due to university's regulations, so the only way to listen to the class is to go to the classroom offline. The detailed information is listed below:

Lectures	Wednesday 9:30am-11:15am (11:05am), Science Center L4
	Friday 9:30am-10:15am, MMW703
Tutorial	Friday 10:30-11:15am, MMW703

Please be noted that the TA will have several sessions to help students on Python programming. Here is the information of the instructor and TAs:

- **Instructor: Professor Yu LI**

E-mail: liyu@cse.cuhk.edu.hk
Office hour: 3pm-5pm, Friday (SHB-106) or by request

TAs

- **Ziqian LIN**

E-mail: linziqian@link.cuhk.edu.hk
Office hour: 3pm-5pm, Tuesday (SHB-904)

- **Xinyuan LIU**

E-mail: 1155246738@link.cuhk.edu.hk
Office hour: 3pm-5pm, Thursday (SHB-904)

2.2 Software and Communications

- **Blackboard**

It is the main software to manage the course, and the grades of the assignments and tests will be released via blackboard. The scribing documents should also be submitted via blackboard.

- **Piazza**

The website is https://piazza.com/cuhk.edu.hk/fall2025/bmeg3105. Students can ask questions through Piazza (even anonymously). Students can also make the private post to TAs and instructor if they have personal matter.

- **Colab**

A free, cloud-based platform that allows users to write and execute Python code in notebooks. The TA will prepare recitation classes to introduce it, mainly for the non-grading homework and student's project.
Website: https://colab.research.google.com/notebooks/intro.ipynb

2.3 Grading Items

Announcement: All the exams and quizzes will be open book, but no computers or phones are allowed to take during the exams. Besides, the deadlines all due at 11:59 pm on the specific days.

- **Homework (20%)**

Three grading homework A1, A2, A3 (5%+5%+5%) and one non-grading programming assignment PA0 (5%). Be serious to all the assignments because they can be helpful when preparing midterm and final. Here are the starting time and deadlines of each assignment:

Assignment Number	Posted Dates	Due
Programming Assignment 0 (PA0): Programming Environment Setup [No need to Submit Anything]	September 5th	September 17th
Assignment 1 (A1): Basic Concept of Data Analytics -1	September 12th	September 24th
Assignment 2 (A2): Basic Concept of Data Analytics -2	October 3rd	October 15th
Programming Assignment 1 (PA1): Application of DA to the Biology [Cover the Entire 2nd Module]	October 24th	November 14th
Assignment 3 (A3): DA in Personalized Genomics and Precision Medicine	November 12th	November 21st

- **Scribing (10%)**

Summarize one of the lectures and submit it **within one week** after the lecture. Each student should do at least one lecture. The notes and scribing will be posted online for other classmates' reference. Students can choose to remove their names or not. Remember to fill the sign table before September 10th via https://docs.google.com/spreadsheets/d/18Rx5EfxcMS9lgl1dKtQfpKciQfF0B4OM23vLPom3leE/edit?usp=sharing. Additional one week can be extended for the scribers who are responsible for the first two lectures. Everyone can sign for at most two, for additional 1% bonus.

- **In Class Quiz (10%)**

Used for checking attendance. There are totally two sets of quizzes and the exact dates are Oct 15th and Nov 26th. The questions will be simple, mainly based on lecture slides.

- **Mid-term (20%)**

An in-class examination on Oct 17th. Students are allowed to bring any paper-based materials during the exam; however, discussion is not allowed. Include a 2% bonus question.

- **Final (20%)**

A centralized exam arranged by RES. Open book and open notes, but electronic devices are not allowed to use during the exam.

- **Project (20%)**

Description: Not allowed to team-up and students should finish it individually. Students can choose their own project topics and discus with the teaching team, or just choose their topics from the following list:

- 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
- …

Percentage Distribution:

(1) Project milestone report (totally 5%, 1 page. Due: Nov 7th):

- Title, author
- Problem statement and the interesting points (1%)
- The source, the size, the sample of the data (1%)
- The output of the chosen method (1%)
- Describe the method step by step, from input to output (1%)
- The expected results and the ways to evaluate these results (1%)
- Students' contribution on the project

(2) Project final report (totally 7%, no length requirement. Due: Dec 2nd):

- Title, author
- Problem statement and the interesting points (0.5%)
- The source, the size, the sample of the data (0.5%)
- Student's contribution to resolve the problem; describe the method step by step, from input to output (2%)
- The results of the project (1.5%)
- Result evaluation (1.5%)
- Ideas of further improvement (1%)

(3) Code (5%): submitted together with project final report.

(4) Project presentation (3%, 7 min for each student on Nov 21st or Nov 28th)
Evaluation Items:

- ✧ Logic (1%): problem statement, importance, overview of the idea and results.
- ✧ Clarity (1%): whether the audience can understand and follow the presentation.
- ✧ Slides Preparation (1%): clear illustration, no grammar error.

Bonus: (6% in total)

- ➢ One bonus question in Midterm 2%
- ➢ One additional scribing: 1%
- ➢ Pre-course survey + post-lecture survey: 0.5% for each, and the maximum is 3%. It is encouraged to complete all of them so that the instructor can receive feedback and adjust the course accordingly. Remember to fill in the form https://docs.google.com/spreadsheets/d/1W-a2wLq5agvn12-nkaVTkqRVU_6zTEgiDQj4cjq7m04/edit?usp=sharing

Noted that there exist **late dates policies** in three hand-written assignments, project mid-term report and the programming assignment. There are 6 late days total, and 2 max for any assignment. They *cannot* be used on final project report and scribing report.

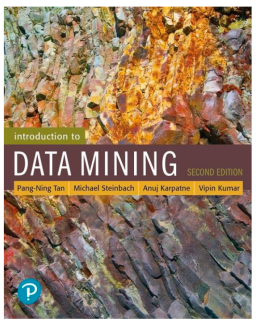
2.4 Usage of AI Tools & Academic Dishonesty

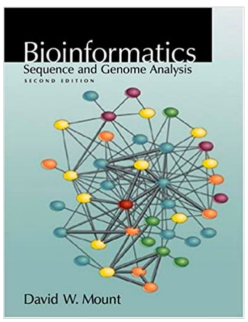
Approach 3 - Use only with explicit acknowledgement: AI tools is allowed in the project to polish the report. However, students are required to submit both their own version and the one polished using AI tools, and make it clear how they used AI tools and which part in the report.

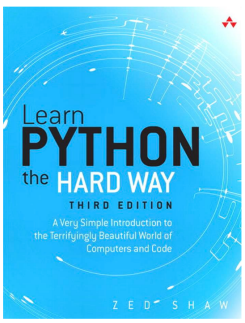
Also, be honest in exams and do not cheat. Consequences can be serious.

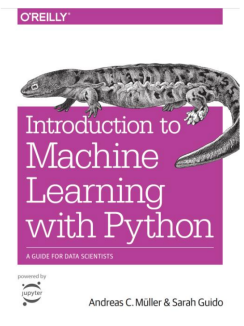
2.5 Reference Books

This is a new field and there is no book available online designed specifically for this course. Slides and the note during class is important in whole studying process. There do exist some reference books but mainly for module 1, as illustrated below:









III. Brief Overview of DATA in Personalized Genomics and Precision Medicine

3.1 The Need for Data Analytics

- Massive of data is being collected and warehoused.
- Computers have become cheaper and more powerful.
- Data analytics are useful.

3.2 Why Data Analytics in Personalized Genomics and Precision Medicine?

Because there are tons of sequencing and health data available and waiting to be analyzed.

3.3 Methods to Use Data to Measure a Person

- Gene and mutations
- Gene expression (Transcriptome)
- Proteome
- Metabolome
- Molecular network & Cellular network
- Microbiome (Oral and gut)
- Organ (Biomedical imaging)
- Hospital test (Blood test and so on)
- Electrocardiography (ECG)
- Demographic information (Age, gender, location and so on)
- Drug history and disease history
- Personal statement and doctor diagnosis
- Living habit (Exercise)
- Diet
- Family history
- Communications and social media data
- Environment (Pollution)
- Travel history (Global pandemic)
- Other applications: track 109 people to predict their health risk; the UK Biobank resource with deep phenotyping and genomic data…

3.4 Summary of the Course

- 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.