CBB752 (Spr. '18) - Biomedical Data Science: Mining and Modeling

Course Description

Rapid developments in biotechnology and computing are changing the way that biomedical scientists interact with data. Traditionally, data were the end result of laborious experimentation, and their interpretation mostly involved careful thought and background knowledge. Today, data are increasingly generated much earlier in the scientific workflow and are much larger in scale. Also, before the data can be interpreted, extensive computational processing is often necessary. Thus, the data deluge now requires the mining and modeling of biomedical data at a large scale - ie biomedical data science.

This course aims to equip students with some of the concepts and skills relevant to biomedical data science, with an emphasis on bioinformatics, a sub-discipline of this broader field, through examples of mining and modeling of genomic and proteomic data. More specifically, bioinformatics encompasses the analysis of gene sequences, macromolecular structures, and functional genomics data on a large scale. It represents a major practical application for modern techniques in data mining and simulation. Specific topics to be covered include sequence alignment, large-scale processing, next-generation sequencing data, comparative genomics, phylogenetics, biological database design, geometric analysis of protein structure, molecular-dynamics simulation, biological networks, mining of functional genomics data sets, and machine learning approaches for data integration.

Overall Flow of the Class:
(Module = Group of Lectures)

- Introduction
- Module on "the Data" (Genomic, Proteomic & Structural Data), introducing the main data sources (their properties, where you access, &c)
- Module on Databases & Data Science Issues (Knowledge Representation incl. Sem. Web & Privacy, Provenance & Standards)
- Module on Mining (Alignment & Variant Calling, Supervised & Unsupervised Approaches, Networks)
- Module on Cell Modeling
- Module on Molecular Modeling

Lectures:

- MW 1:00 - 2:15 PM, Bass 305

Discussion Section:

- Bass 405 (subject to change)

Different headings for this class (4 variants)

- **CB&B752/CPSC752 - Grad. w/ programming**

  This graduate-level version of the course consists of lectures, in-class tests, discussion section, programming assignments, and a final programming project.

- **MB&B752/MCDB752 - Grad. w/o programming**

  This graduate-level version of the course consists of lectures, in-class tests,
  discussion section, written problem sets, and a final (semi-computational section and a literature survey) project. Unlike CBB752, there is no programming required.

- **MB&B 753b3/MB&B 754b4 - Modules**

  For graduate students the course can be broken up into two "modules" (each counting 0.5 credit towards MB&B course requirement):
  753 - Biomedical Data Science: Mining (1st half of term)
  754 - Biomedical Data Science: Modeling (2nd half of term)
  Each module consists of lectures, in-class tests, written problem sets, and a final, graduate level written project that is half the length of the full course's final project.

- **MB&B452/MCDB452/S&DS352 - Undergrad.**
  This undergraduate version of the course consists of lectures, in-class tests, discussion section, written problem sets, and a final (semi-computational section and a literature survey) project. The programming assignments from CB752 can be substituted for the written work by permission of instructor.
- **Auditing**

  This is allowed. We would strongly prefer if you would register for the class.

Prerequisites

The course is keyed towards CBB graduate students as well as advanced undergraduates and graduate students wishing to learn about types of large-scale quantitative analysis that whole-genome sequencing and forms of large-scale biological data will make possible. It would also be suitable for students from other fields such as computer science, statistics or physics wanting to learn about an important new biological application for computation.

Students should have:

1. A basic knowledge of biochemistry and molecular biology.
2. A knowledge of basic quantitative concepts, such as single variable calculus, basic probability & statistics, and basic programming skills.

These can be fulfilled by: MBB 200 and Mathematics 115 or permission of the instructor.

Class Requirements

Discussion Section / Readings

Papers will be assigned throughout the course. These papers will be presented and discussed in weekly 60-minute sections with the TFs. A brief summary (a half-page per article) should be submitted at the beginning of the discussion session.

In-class tests: Midterm & Quiz

- There will be a midterm covering the 1st half of the course.
- There will be a quiz covering 2nd half of the course comprising simple questions that you should be able to answer from the lectures plus the main readings.

For references, please refer the previous quizzes and answer keys from Fall 2012

Programming Assignments (Req'd for CBB and CS grad. students)

- There will be four homework assignments including assignment 0. We will try to promote the idea of reproducible research and using version control system, specifically GitHub, in facilitating the process of homework submission.

Non-programming Assignments

- There will be equivalent four homework assignments (including assignment 0), particularly for MB&B and MCDB students without a programming background. The programming part will be replaced with assignments involving the use of web-based tools or essay questions.

Pages from previous years

2018 is the 20th time Bioinformatics has been taught at Yale. Pages for the 19 previous iterations of the class are available. Look at how things evolve!

- 2017 Spring
- 2016 Spring
- 2015 Spring
- 2014 Spring
- 2012 Fall
- 2012 Spring
- 2011 Spring
- 2010 Spring
- 2009 and earlier (12 years of classes, staring in '98) (Note the pre-2010 course was Genomics & Bioinformatics; after 2010, the course contains all of the "Bioinformatics" of previous years and then more (!) with less "Genomics".)

Assignments	***Final Project***	***Materials***

CBB752 (Spr. '18) - Biomedical Data Science: Mining and Modeling

Announcements

First Meetings

The first lecture will be held on **Wed. Jan 17, 2018.** Somewhat confusingly, Friday, January 19th has been declared an "academic Monday," and so the course's second meeting will be on Friday, January 19th.

Snow Days (general policy)

We have built into the class schedule the potential for snow days. To avoid last minute uncertainty and confusion, we will not wait until Yale officially closes the university for snow (which only happens in the most extreme of blizzards). If the weather looks particularly problematic a few days before (e.g., on Sat. for a Mon. class), we will preemptively cancel via the class email list, which means it important for all to be on this list.

Comments

You do not have permission to add comments.

#	Day	Date		Topic	URL
Data Mining (First Half)					
1	W	17-Jan	MG	Introduction	
2	F	19-Jan	MDS	Data 1 - Genomics	
3	M	22-Jan	MDS	Data 2 - Genomics	
4	W	24-Jan	JR	Data 3 - Proteomics	
5	M	29-Jan	JR	Data 4 - Proteomics	
6	W	31-Jan	KC	Data 5 - Knowledge Representation & DBs	
7	M	5-Feb	MG-1	Data 6 - Introduction to Personal Genomes	
8	W	7-Feb		(Data Science Center Day - Skip)	
9	M	12-Feb	MG-2*	Mining 1 - Alignment (w/ cameo)	
10	W	14-Feb	MG-3	Mining 2 - Fast Alignment & Variant Calling	
11	M	19-Feb	MG*	QUIZ - 1	
12	W	21-Feb	ML	Mining 3 - Rare variants & EXaC	
13	M	26-Feb	MG-4	Mining 4 - More Variant Calling	
14	W	28-Feb	MG-5	Mining 5 - Unsupervised Mining	
15	M	5-Mar	MG-6	Mining 6 - Supervised Mining	
	W	7-Mar		(Open Day for Snow)	
Simulation (Second Half)					
16	M	26-Mar	MG-7*	Mining 7 - Network Prediction + Misc.	
17	W	28-Mar	MG-8	Mining 8 - Network Analysis	
18	M	2-Apr	CO	Protein Simulation I	
19	W	4-Apr	CO	Protein Simulation II	
20	M	9-Apr	CO	Protein Simulation III	
21	W	11-Apr	CO	Markov Models I	
22	M	16-Apr	CO	Markov Models II	
23	W	18-Apr	CO	Markov Models III / Protein Aggregation	
24	M	23-Apr	MG	QUIZ-2	
25	W	25-Apr	MG*	Presentations	

CBB752 (Spr. '18) - Biomedical Data Science: Mining and Modeling

Instructors

For general correspondence and questions, please contact us at:

cbb752 (at) gersteinlab.org

Instructor-in-Charge

Name	Office	Email
Mark Gerstein	Bass 432A	contact.gerstein.info

Guest Instructors

Name	Office	Email
Corey O'Hern	Mason Laboratory	corey.ohern (at) yale.edu
Jesse Rinehart	West Campus	jesse.rinehart (at) yale.edu
Matthew Simon	West Campus	matthew.simon (at) yale.edu
Kei Cheung	300 George St	kei.cheung (at) yale.edu
Monkol Lek	300 Cedar St	monkol.lek (at) yale.edu

Consultation is available UPON REQUEST or according to times stipulated by the individual instructors. Prof. Gerstein's office office hours will usually be right after some the classes.

Teaching Fellows (TA)

Name	Office	Email
William Meyerson	Bass 437	william.meyerson (at) yale.edu
Xiaotong Li	Bass 437	xiaotong.li (at) yale.edu

Comments

You do not have permission to add comments.

Grading Policy

Last year the following grade distribution was used for all students (CBB/CPSC/MBB/MCDB). We expect that this year the weighting scheme will be similar, to a first approximation.

Category	% of Total Grade
Midterm	15%
Quiz	15%
Discussion Section	10%
Homeworks	20%
Final Project	40%

Relevant Yale College Regulations

Students may have questions concerning end-of-term matters. Links to further information about these regulations can be found below:

- http://catalog.yale.edu/ycps/academic-regulations/reading-period-final-examination-period/
- http://catalog.yale.edu/ycps/academic-regulations/completion-of-course-work/
- Brief presentation on how to cite correctly : http://archive.gersteinlab.org/mark/out/log/2012/06.12/cbb752b12/cbb752_cite.ppt

Plagiarism

Below is a message from the Dean of Yale College about citing your references and sources of information and plagiarism:

" You need to cite all sources used for papers, including drafts of papers, and repeat the reference each time you use the source in your written work. You need to place quotation marks around any cited or cut-and-pasted materials, IN ADDITION TO footnoting or otherwise marking the source. If you do not quote directly – that is, if you paraphrase – you still need to mark your source each time you use borrowed material. Otherwise you have plagiarized. It is also advisable that you list all sources consulted for the draft or paper in the closing materials, such as a bibliography or roster of sources consulted.
You may not submit the same paper, or substantially the same paper, in more than one course. If topics for two courses coincide, you need written permission from both instructors before either combining work on two papers or revising an earlier paper for submission to a new course.
It is the policy of Yale College that all cases of academic dishonesty be reported to the chair of the Executive Committee.... "

" Academic integrity is a core institutional value at Yale. It means, among other things, truth in presentation, diligence and precision in citing works and ideas we have used, and acknowledging our collaborations with others. In view of our commitment to maintaining the highest standards of academic integrity, the Graduate School Code of Conduct specifically prohibits the following forms of behavior: cheating on examinations, problem sets and all other forms of assessment; falsification and/or fabrication of data; plagiarism, that is, the failure in a dissertation, essay or other written exercise to acknowledge ideas, research, or language taken from others; and multiple submission of the same work without obtaining explicit written permission from both instructors before the material is submitted. Students found guilty of violations of academic integrity are subject to one or more of the following penalties: written reprimand, probation, suspension (noted on a student's transcript) or dismissal (noted on a student's transcript). "

Also, it might be of interest to people, to look at this recent article regarding academic dishonesty.

Comments

You do not have permission to add comments.

CBB752 (Spr. '18) - Biomedical Data Science: Mining and Modeling

Section Readings

Weekly Discussion Sections & Readings

Timing TBD by class poll

Format

The standard discussion section involves student presentations on 1 or 2 papers. Some discussion sections will involve hands-on skill-building demos taught by the teaching fellows, such as the use of R, High Performance Computing, and GitHub.

The exact format will be determined based on the size of the class. However, tentatively, we require the following

- Students are expected to bring approx. a half page (2-3 paragraph) summaries of each paper to the section. (we will collect a hard copy during each session, but if you'd like to save some trees, we will accept electronic submission. Please submit **PDF** to **cbb752 (at) gersteinlab.org** BEFORE each section).
- Students will give approx. 20 **min presentations** about each paper.
- Students will be graded on a combination of the written summary, presentation, and participation in discussions.

Section Readings

Reading assignments for discussion sessions are listed below.

Session 1: Next-Gen Sequencing

- Goodwin S. et al. "Coming of age: ten years of next-generation sequencing technologies" Nature Reviews Genetics. 17 (2016) PDF
- Treangen T.J. and Salzburg, S.L. "Repetitive DNA and next-generation sequencing: computational challenges and solutions." Nature Reviews Genetics. 13:36-46 (2012) PDF

CBB752 (Spr. '18) - Biomedical Data Science: Mining and Modeling

Assignments

General

posted Dec 14, 2017, 1:55 PM by William Meyerson **[updated 11 minutes ago by Mark Gerstein]**

We expect to assign three to four homework assignments and a final project. They will be linked on this page.

To accommodate the diverse backgrounds and interests of the students, we offer students a choice of which style of assignments they wish to take. Students designate their track by selecting under which course number they wise to take the course. Quantitative track assignments involve more programming. Biology track assignments involve more writing. Some questions that are required on graduate track assignments will be counted as "extra credit challenge questions" for undergraduate students.

- Undergraduate Biology Track
- Undergraduate Quantitative Track
- Graduate Biology Track
- Graduate Quantitative Track

Listed below are tentative assignment topics and due dates, which will be finalized during the term.

- HW 0: Survey; due Wednesday, January 24, 2018
- HW 1: Genomic variant calling; due Wednesday, February 28, 2018
- HW 2: Biomedical data mining; due Wednesday, March 7, 2018
- HW 3: Protein simulations; due Wednesday, April 18, 2018
- Final Presentation; due April 25, 2018

The teaching fellows this year prefer for programming assignments to be written in R, which is the programming language they are most familiar with. The teaching fellows will also accept assignments in Python but will be less able to award partial credit and offer troubleshooting help in Python.
The course wants to encourage students to learn to use Yale's High Performance Computing Cluster since and GitHub, since these are important tools in biomedical data science.
Late homework assignments will be marked down 10% per day.
We expect that many students will find some of the assignments challenging due to the technical and conceptual demands of biomedical data science, but ample help will be offered to students who ask for it.
Students who are less familiar with R, HPC, and/or GitHub are especially encouraged to begin their assignments early so that they have time to get help from the teaching fellows.

CBB752 (Spr. '18) - Biomedical Data Science: Mining and Modeling

Final Project

There is a final project, TBA. Students will work collaboratively on biomedical data science projects and present on the last class. You can get a flavor for the Final Project by looking at the project from last year, linked below.

Previous Years

(Final Project from Spring 2017).

Showing **0** items

Name	Due Date	Description
Sort	Sort	Sort

Showing **0** items

Comments

You do not have permission to add comments.