# Biomedical Data Science: Mining and Modeling

Home Syllabus Course Info Discussion Sections Assignments Quiz Archive

Final Project

## Course Description

Rapid developments in bio- and information- technology and 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 in biomedicine now requires mining and modeling on 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). This module also includes discussion of databases and knowledge representation issues.
- Module on Mining (Alignment & variant calling necessary for personal genomics; Basic multi-omics calculations; Supervised & unsupervised mining approaches towards multi-omic data; Networks)
- Module on Cell Modeling
- Module on Molecular Modeling

### Lectures

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

### Discussion Section

- F 9:30-10:30 AM or F 1:00-2:00 PM, Bass 405

## Different headings for this class (4 variants)

- CBB 752 / CPSC 752 - Grad. with programming
  - This graduate-level version of the course consists of lectures, in-class tests, discussion section, programming assignments, and a final programming project.
- MB&B 752 / MCDB 752 - Grad. without 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&B 452 / MCDB 452 / S&DS 352 - 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:

- A basic knowledge of biochemistry and molecular biology.
- 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: Quiz

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

For references, please refer the previous Quiz Archive

## 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

- 2020 Spring is the 23rd time Bioinformatics has been taught at Yale. Pages for the 22 previous iterations of the class are available. Look at how things evolve!
- 2019 Spring
- 2018 Spring
- 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".)

## Class data dump

- Syllabus and class info dump in single PDF file: PDF v1 or PDF v2
- Class poster: PDF

**Biomedical Data Science: Mining and Modeling is maintained by Jiahao Gao.**
This page was generated by GitHub Pages.

# CBB 752 - Biomedical Data Science: Mining and Modeling

Home Syllabus Course Info Discussion Sections Assignments Quiz Archive Final Project

## Syllabus

CBB752-spg-20-schedule-cbb752b20 : 2020 Public

| # | Day | Date | | Topic | Public Comme |
|---|---|---|---|---|---|
| | M | 1/13 | -- | *YALE* Spring term classes begin, 8.20 a.m. | |
| | F | 1/17 | -- | *YALE* Friday classes do not meet; Monday classes meet instead | |
| | M | 1/20 | -- | Martin Luther King Jr. Day; classes do not meet. | |
| | | | | Data Mining (1st Half) | |
| 1 | M | 1/13 | MG | Introduction | **Pack 1** |
| 2 | W | 1/15 | MDS | DATA 1 - Genomics | |
| 3 | F | 1/17 | MDS | DATA 2 - Genomics | |
| 4 | W | 1/22 | JR | DATA 3 - Proteomics I | |
| 5 | M | 1/27 | JR | DATA 4 - Proteomics II | |
| 6 | W | 1/29 | KC | DATA 5 - Knowledge Representation & Databases | |
| 7 | M | 2/3 | MG* | DATA 6 + MINING 1 - Personal Genomes Intro. (with an individual's perspective) | |
| 8 | W | 2/5 | MG | MINING 2 - Seq. Comparison + Multi-seq Alignment | |
| 9 | M | 2/10 | MG | MINING 2 - Fast Alignment & Variant Calling (incl. a focused section on SVs) | |
| 10 | W | 2/12 | MG | MINING 3 - Basic Multi-omics & Unsupervised Mining #1 (SVD & Spectral Methods) | |
| 11 | M | 2/17 | MG+TF | Quiz on 1st Half | |
| 12 | W | 2/19 | MG | MINING 4 - Unsupervised Mining #2 (Extensions & Applications) | |
| 13 | M | 2/24 | MG | MINING 5 - Supervised Mining & Network Analysis | |
| 14 | W | 2/26 | MG* | MINING 6- Deep Learning I | |
| 15 | M | 3/2 | MG* | MINING 7 - Deep Learning II | |
| 16 | W | 3/4 | MG* | held out day for snow | |
| | F | 3/6 | -- | *YALE* Spring recess begins, 5.30 p.m. | |
| | | | | Simulation (2nd Half) | |
| | M | 3/23 | -- | *YALE* Classes resume, 8.20 a.m. | |
| 17 | M | 3/23 | SK | Cell/Immune Modeling I | |
| 18 | W | 3/25 | SK | Cell/Immune Modeling II | |
| 19 | M | 3/30 | SK | Cell/Immune Modeling III | |
| 20 | W | 4/1 | CO | Protein Simulation I | |
| 21 | M | 4/6 | CO | Protein Simulation II | |
| 22 | W | 4/8 | CO | Protein Simulation III | |
| 23 | M | 4/13 | CO | Markov Models I | |
| 24 | W | 4/15 | CO | Markov Models II | |
| 25 | M | 4/20 | MG+TF | Quiz on 2nd Half | |
| 26 | W | 4/22 | MG* | Final Presentations | |
| | F | 4/24 | -- | *YALE* Classes end, 5.30 p.m.; Reading period begins | |
| | F | 5/1 | -- | *YALE* Final examinations begin, 9 a.m. | |
| | W | 5/6 | -- | *YALE* Final examinations end, 5.30 p.m. | |
| | F | 5/8 | -- | *TENT* Term grades due for seniors, 5:30 p.m. | |

2020 Public

>
<

# Biomedical Data Science: Mining and Modeling

Home Syllabus Course Info Discussion Sections Assignments Quiz Archive

Final Project

## Contact Info

### For general correspondence and questions, please contact us at:

cbb752 (at) gersteinlab.org

### Instructor-in-Charge

| Name | Abbr | Office | Email |
|---|---|---|---|
| Mark Gerstein | MG | Bass 432A | contact.gerstein.info |

### Guest Instructors

| Name | Abbr | Office | Email |
|---|---|---|---|
| Corey O'Hern | CO | Mason Laboratory | corey.ohern (at) yale.edu |
| Steven Kleinstein | SK | 300 George St | steven.kleinstein (at) yale.edu |
| Jesse Rinehart | JR | West Campus | jesse.rinehart (at) yale.edu |
| Matthew Simon | MDS | West Campus | matthew.simon (at) yale.edu |
| Kei Cheung | KC | 300 George St | kei.cheung (at) yale.edu |
| Carl Zimmer | CZ | 266 Whitney Ave | carl.zimmer (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 (TF)

| Name | Abbr | Office | Email |
|---|---|---|---|
| Hussein Mohsen | HM | Bass 437 | hussein.mohsen (at) yale.edu |
| Jiahao Gao | JG | Bass 437 | jiahao.gao (at) yale.edu |

# General Course Policy

## First Meeting

The first lecture will be held on Mon. Jan 13, 2020. Somewhat confusingly, Friday, January 17th has been declared an "academic Monday" due to Martin Luther King Jr. Day (01/20/2020), and so the course's **third** meeting will be on Friday, January 17th.

## 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.

## Grading Policy

We expect that this year the weighting scheme will be to a first approximation:

| Category | % of Total Grade |
|---|---|
| Midterm Quiz | 15% |
| Final Quiz | 15% |
| Discussion Section | 20% |
| Homeworks | 20% |
| Final Project | 30% |

## 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.

---

**Biomedical Data Science: Mining and Modeling is maintained by Jiahao Gao.**

# Biomedical Data Science: Mining and Modeling

Home Syllabus Course Info Discussion Sections Assignments Quiz Archive

Final Project

## Weekly Discussion Sections & Readings

### Time and Location

| Session | Time | Location | Note |
|---|---|---|---|
| Section 1 | 9:30-10:30 AM | BASS 405 | |
| Section 2 | 1:00-2:00 PM | BASS 405 | Room Change: BASS 434 for 2/21 |

### 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, we generally require the following:

- Each week, students should **read the assigned papers and write at a minimum of 200 words (half a page, single-spaced, per paper) summaries of each paper (two articles = approx. 1 page)**. We would like to encourage electronic submission, via Canvas. For those who have trouble accessing canvas, we will also accept submission over email to cbb752 (at) gersteinlab.org **BEFORE** the start of each section.
- Each student will give one presentation about a selected paper (approx. 20 min) in one of the sessions.
- Students will be graded on a combination of the written summary, presentation, and participation in discussions.
- If you are presenting, you are exempt from writing a summary.
- Please notify TFs in advance if you cannot come to the discussion session. Student **can miss up to one discussion section without a penalty**.

## Section Readings

Reading assignments for discussion sessions are listed below:

### (Optional) Suggested Reading for Week 1 (There will be a class instead of discussion session on 1/17/2020)

- How to (seriously) read a scientific paper, on your own. [Link]

## Session 1, 1/24, BASS 405 for both Sessions (First Discussion Sessions)

### Topic

- Next-Gen Sequencing and database

### Reading Assignment

- Goodwin S. et al. “Coming of age: ten years of next-generation sequencing technologies” Nature Reviews Genetics. 17 (2016) [PDF]
- Wheeler DA et al. “The complete genome of an individual by massively parallel DNA sequencing,” Nature. 452:872-876 (2008) [PDF]

## Session 2, 1/31, BASS 405 for both Sessions

### Topic

- Proteomics

### Reading Assignment

- A draft map of the human proteome. Nature 509,575–581 (29 May 2014) [PDF]
- Mass-spectrometry-based draft of the human proteome. Nature 509, 582–587 (29 May 2014) [PDF]

## Session 3, 2/7, BASS 405 for both Sessions

### Topic

- Debate I

### Reading Assignment

- Gencode vs Salzberg et al. debate
  - (Main paper) Salzberg et al. CHESS paper using GTEx [PDF]
  - (Main paper) GENCODE’s rebuttal [PDF]
  - (Optional) New human gene tally reignites debate [News Article]
- (Optional) Why most published research finding are false [PDF]

## Session 4, 2/14, BASS 405 for both Sessions

### Topic

- Help session on Quiz 1 - TFs prepare materials on SW alignments and Q&A session
- Sequence and Alignments

### Reading Assignment

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. Journal of Molecular Biology, 215(3):403-10. PMID: 2231712. [PDF]
- T.F. Smith and M.S. Waterman. (1981) Identification of common molecular subsequences. Journal of Molecular Biology,147(1): 195-7. PMID: 7265238. [PDF]

## Session 5, 2/21, BASS 405 for Session 1, BASS 434 for Session 2 (Note Room Change)

### Topic

- Debate II - Phylogenetics

### Reading Assignment

- Jarvis ED et al. (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. Science, 346(6215), 1320-1331. [PDF]
- Mitchel KJ, Cooper A, Philips MJ (2015) Comment on “Whole-genome analyses resolve early branches in the tree of life of modern birds.” Science, 349(6255) 1460 [PDF]

## Session 6, 2/28, BASS 405 for both Sessions

### Topic

- Deep learning for genomics

### Reading Assignment

- A primer on deep learning in genomics [PDF]
- Deep learning for biology [PDF]

## Session 7, 3/6, BASS 405 for both Sessions

### Topic

- Protein structure and biophysics

### Reading Assignment

- Zhou, AQ, O'Hern, CS, Regan, L (2011). Revisiting the Ramachandran plot from a new angle. Protein Sci., 20, 7:1166-71 [PDF]
- Dill KA, Ozkan SB, Shell MS, Weikl TR. (2008) The Protein Folding Problem. Annu Rev Biophys,9, 37:289-316. PMID: 2443096. [PDF]
- Bowman GR, Beauchamp KA, Boxer G, Pande VS. "Progress and challenges in the automated construction of Markov state models for full protein systems," J. Chem. Phys. 131 (2009) 124101 [PDF]

## (spring recess)

## Session 8, 3/27, BASS 405 for both Sessions

### Topic

- Immune system modelling and dynamics

### Reading Assignment

- Perelson AS. Modelling viral and immune system dynamics. Nat Rev Immunol. 2002 Jan;2(1):28-36. [PDF]
- Modeling the Spread of Ebola [PDF]

## Session 9, 4/3, BASS 405 for both Sessions

### Topic

Debate III - Cancer incidence

### Reading Assignment

- Debate reignites over the contributions of 'bad luck' mutations to cancer [Link]
- The simple math that explains why you may (or may not) get cancer [Link]

## Session 10, 4/10, BASS 405 for both Sessions

### Topic

- Help session on quiz 2 / final project

### Reading Assignment

- (no reading assignment this week)

---

**Biomedical Data Science: Mining and Modeling is maintained by Jiahao Gao.**
This page was generated by GitHub Pages.

# Biomedical Data Science: Mining and Modeling

Home Syllabus Course Info Discussion Sections Assignments Quiz Archive

Final Project

## Quiz Archive

- Spring 2019: Main Page, Quiz 1 key, Quiz 2 key, HW1, HW2
- Spring 2018: Main Page, Quiz 1, Quiz 1 Key, Quiz 2, Quiz 2 Key
- Spring 2017: Main Page, HW 1, HW 2, Midterm, Final
- Spring 2016: Main Page, Midterm, Midterm Key, Final Quiz, Final Quiz Key
- Spring 2015: Main Page
- Spring 2014: Main Page
- Fall 2012: Main Page, Quiz 1-4
- Spring 2012: Main Page, Quiz 1-4
- Spring 2011: Main Page, Quiz 1
- Spring 2010: Main Page, Quiz 1, Quiz 2, Quiz 3+4
- Spring 2009: Main Page, Quiz 1, Quiz 2
- Spring 2008: Main Page
- Fall 2006: Main Page, Quiz 1, Quiz 2
- Fall 2005: Main Page, Quiz 1, Quiz 2
- Spring 2005: Main Page, Quiz 1, Quiz 2
- Fall 2003: Main Page, Quiz 1, Quiz 1 Key, Quiz 2, Quiz 2 Chart, Quiz 2 Key
- Fall 2002: Main Page, Quiz 1, Quiz 1 Key, Quiz 2, Quiz 2 Key
- Fall 2001: Main Page, Quiz 1, Quiz 1 Key, Quiz 2
- Fall 2000: Main Page
- Fall 1999: Main Page, Quiz 2 with Key
- Spring 1999: Main Page
- Spring 1998: Main Page

**Biomedical Data Science: Mining and Modeling is maintained by Jiahao Gao.**
This page was generated by GitHub Pages.