

BIOMEDICAL INFORMATICS & DATA SCIENCE

100 College Street, ninth floor, informatics@yale.edu
<https://medicine.yale.edu/biomedical-informatics-data-science>

Professors H. Altalib (*Neurology*), C.A. Brandt, K. Cheung, J. Deng (*Therapeutic Radiology*), G. El Fakhri, A.L. Hsiao, A. Justice (*Medicine*), S.H. Kleinstein (*Pathology*), P.L. Miller (*Emeritus*), L. Ohno-Machado (*Chair*), X. Papademetris, H. Xu, K. Xu (*Psychiatry*)

Associate Professors N. Bakshi (*Pediatrics*), J. Canterino (*Medicine*), S. Chen (*Genetics*), A. Chou (*Medicine*), M. Davis (*Radiology and Biomedical Imaging*), T. Durant (*Laboratory Medicine*), K. Harris Nwanyanwu (*Ophthalmology and Visual Science*), R. Hauser (*Laboratory Medicine*), S. Jarad (*Emergency Medicine*), T. Kuo, C. Liu (*Radiology and Biomedical Imaging*), Y. Liu (*Pharmacology*), D. Meeker, E. Melnick (*Emergency Medicine*), J. Miller (*Medicine*), R. Polimanti (*Psychiatry*), M. Rodriguez Martinez, W. Schulz (*Laboratory Medicine*), M. Sharifi (*Pediatrics*), R. Taylor

Assistant Professors S. Aneja (*Therapeutic Radiology*), K. Aneni (*Child Study Center*), I. Arhuidese (*Surgery*), G.I. Ash (*Medicine*), Q. Chen, H. Cheng, H. Cho, J.M. Cohen (*Dermatology*), B. Coleman (*Emergency Medicine*), E. Erson Omay (*Neurosurgery*), R. Hao (*Medicine*), M. Iscoe, R. Khera (*Medicine*), J. Kim, I. Leeds (*Surgery*), A. Lisi, X. Liu (*Radiology and Biomedical Imaging*), Y. Lu (*Medicine*), T. Marin (*Radiology and Biomedical Imaging*), J. Montalvo-Ortiz (*Psychiatry*), C. Oliveira (*Pediatrics*), E.M. Powers (*Pediatrics*), K. Wang (*Medicine*), L. Wang (*Medicine*), S. Zhang, H. Zhou (*Psychiatry*)

Instructors N. Hong, A. Loza, K. Raja

Senior Research Scientist N. Rajeevan

Research Scientists H. He, H. Rajeevan, R. Wang

Associate Research Scientists E.C. Coppola, F.A. Kidwai (*Medicine*), V. Kuttichi Keloth, B. Ondov

Assistant Clinical Professors C. Fernandes (*Child Study Center*), S. Ong (*Medicine*)

Lecturer R. Zhu

CB&B 5710a, Data Science Grant-Writing Practicum Lucila Ohno-Machado

This is a hands-on course where students review funded and non-funded grant proposals for different types of NIH awards, as well as the critiques provided by the reviewers. Proposals in informatics and data science are different than traditional basic sciences proposals and clinical research proposals, so this course is specific for those proposing data science and informatics innovation that can be applied in biology and/or medicine. Although there is an emphasis on K (mentored) and F awards, we also cover the basics of R (non-mentored) research awards. Instructors and classmates review proposals that students prepare as part of the course.

CB&B 5740b, Biomedical Natural Language Processing: Methods and Applications

Qingyu Chen

This course examines current natural language processing (NLP) methods and their applications in the biomedical domain. It provides a systematic introduction to basic knowledge on NLP and AI (e.g., linguistics, machine learning, and deep learning algorithms), advanced NLP tasks (e.g., information extraction, information retrieval, question answering), and corresponding approaches including the recent large language models (LLMs) and hands-on experience in developing biomedical NLP systems for different applications, ranging from biomedical literature mining to clinical decision support. Assessment in this course consists of technical exercises, exams, and projects, to demonstrate the applicability of skills learned during the course.

CB&B 5750b, Bioinformatics Applications in Biomedicine Jihoon Kim

This advanced graduate course prepares students to analyze real-world biomedical data at scale across omics data and electronic health record (EHR) data. Emphasizing reproducible workflows, statistical rigor, and clinical translation, students work directly with large-scale omics datasets and biobank resources, including the All of Us Research Program. Through coding-intensive assignments, students develop Linux-based pipelines, Slurm-driven HPC workflows, cloud-native analyses, and EHR-based cohort construction for translational investigations. Student-led journal presentations strengthen algorithmic understanding, reproducibility, and translational thinking. Designed to complement theory-driven coursework, the course focuses on practical execution—equipping students to build scalable pipelines, generate biologically and clinically meaningful insights, and communicate methods with professional technical clarity. Proficiency in Python and R for data analysis.

CB&B 5760b, Foundations of Real World Data Science: Electronic Health Records

Mark Iscoe and Daniella Meeker

The course covers scientific principles, best practices, and limitations of using observational data from administrative records, including hypothesis generation, feasibility assessments, and causal inference. Students learn pragmatic skills required to prepare analytic data from large, complex transactional databases. We cover methods for data quality characterization and profiling for study planning. Coursework includes application of methods for creation and validation of computable phenotypes, electronic clinical quality measures, and derived analytic variables. Skills include preparation of real-world data for visualization and reporting in business intelligence tools commonly used in population health and health administration. Students reproduce results from published literature using existing databases for predictive modeling, public health, and outcomes research. Completion of this course positions students for externships in healthcare analytics and health data science. Prerequisites: Course must be approved by the instructor BIS 638, Clinical Database Management Systems and Ontologies (or equivalent); proficiency in SQL and Python, or R; HIPAA and HSR training; YNHHS Research Basic Access; COS550 or equivalent; execution of DUAs for public and proprietary databases; demonstrated use of YU-YNHHS data science platform.

CB&B 5790b, Distributed Artificial Intelligence on Biomedical Data Tsung-Ting Kuo

This course covers the basic concepts of artificial intelligence (AI), distributed ledger, and distributed AI with applications on biomedical data. We plan to cover AI pipeline, blockchain and smart contracts, and blockchain-based federated learning methods. By the end of this course, students should be able to: (1) select and use AI, distributed

ledger, and distributed AI methods appropriate for a given clinical research content; (2) understand and apply AI, distributed ledger, and distributed AI approaches in the context of biomedical problems in research settings; (3) adapt, and validate AI, distributed ledger and distributed AI approaches to a specific biomedical problem; and (4) communicate audience-appropriate biomedical informatics content for AI, distributed ledger and distributed AI.

CB&B 5800a, Bioinformatics Algorithms in Genomics Haoyu Cheng

This course introduces key algorithms used in computational genomics, with a focus on both classical bioinformatics methods and emerging machine learning and deep learning approaches. Topics covered include sequence alignment, genome assembly and comparative genomics, variant identification and analysis, and gene expression and regulation, along with advanced techniques for specialized applications such as cancer genomics. Through hands-on exercises and projects, students gain practical experience in implementing algorithms and analyzing real-world genomic data. By the end of the course, students are prepared to conduct independent genomic analyses or develop novel bioinformatics algorithms to tackle emerging challenges in genomics.

CB&B 5801b, Computational Proteomics and Protein Modeling Maria Rodriguez Martinez

This course explores advanced computational proteomics and protein modeling approaches. Students gain hands-on experience analyzing bulk, single-cell, and imaging-based proteomics data. The course also covers cutting-edge machine learning and deep learning techniques, including transformer models, graph neural networks, and generative models, for protein-related tasks such as structure and function prediction and protein optimization. Special attention is given to applications in immunology. This course is ideal for students interested in computational biology, AI-driven structural biology, and computational immunology. Prerequisites: mathematics and statistics (linear algebra; calculus [for optimization]; probability/statistics), Python for data analysis (NumPy, Pandas, plotting, Jupyter; basic Git/conda), machine learning basics (regression/classification, model evaluation, intro to deep learning [PyTorch]). Recommended: basic biology (proteins [sequence/structure], high-level proteomics), advanced machine learning (transformers, GNNs, generative models, command-line tools).

CB&B 7140a, Research Paper in Computational Biology and Biomedical Informatics

Anthony Lisi and Hua Xu

This two-semester single credit pass/fail course must be completed as part of the terminal M.S. degree program in computational biology and biomedical informatics (CB&B). Students work with a faculty supervisor in designing their project and writing their research paper. The syllabus details the intended scope and process for writing the research paper. In the broadest terms, the research paper must be of publishable quality and defensible in a public scientific forum. The student's research supervisor is responsible for managing the intended product. The preferred format of the research paper for students is one that is in the style and length of a publishable, peer-reviewed paper, templated based on the journal submission. Prerequisite: second-year enrollment in program.

CB&B 7150b, Research Paper in Computational Biology and Biomedical Informatics
(Part 2) Anthony Lisi and Hua Xu

This two-semester single credit pass/fail course must be completed as part of the terminal M.S. degree program in computational biology and biomedical informatics (CB&B). Students work with a faculty supervisor in designing their project and writing their research paper. The syllabus details the intended scope and process for writing the research paper. In the broadest terms, the research paper must be of publishable quality and defensible in a public scientific forum. The student's research supervisor is responsible for managing the intended product. The preferred format of the research paper for students is one that is in the style and length of a publishable, peer-reviewed paper, templated based on the journal submission. Prerequisite: second-year enrollment in program.

CB&B 7400a, Introduction to Health Informatics Andrew Loza

The course provides an introduction to clinical and translational informatics. Topics include (1) overview of biomedical informatics, (2) design, function, and evaluation of clinical information systems, (3) clinical decision-making and practice guidelines, (4) clinical decision support systems, (5) informatics support of clinical research, (6) privacy and confidentiality of clinical data, (7) standards, and (8) topics in translational bioinformatics. Permission of the instructor required.

CB&B 7520b / CPSC 7520b / MB&B 7520b and MB&B 753b and MB&B 754b / MB&B 753b and MB&B 754b / MB&B 754b / MCDB 7520b / S&DS 5520b, Biomedical Data Science: Mining and Modeling Mark Gerstein and Matthew Simon

Biomedical data science 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, normalization of microarray data, mining of functional genomics data sets, and machine-learning approaches to data integration. Prerequisites: biochemistry and calculus, or permission of the instructor.

*** BENG 4106b, Medical Software Design** Xenophon Papademetris

Software design and implementation for medical applications, with emphasis on how new ideas can be developed within today's healthcare regulatory environment. This is a project-based class. The lectures provide essential material to help the students successfully complete their projects. In particular, the lectures cover material in the following three broad areas: (i) Medical software design based on a clinical need. (ii) Needs identification, verification, validation, and overview of the FDA regulatory process. (iii) Introductory material in experimental design, image analysis, and machine learning as needed by the projects. We also examine the new proposed FDA regulations on the use of machine learning in medical devices and related issues related to the use of these techniques in medical software in general. Prerequisite: Some programming background in at least one programming language. Instructor permission required.

BIS 633a, Population and Public Health Informatics Brian Coleman

This is not a programming course or a mathematics course. The course provides an in-depth survey of the data standards, data analysis tools, databases, and information management systems and applications used in clinical population research, disease surveillance, emergency response information systems, and the like. It examines informatics techniques used on population-level data to improve health and the application of information and computer science and technology to public health practice, research, policy, and decision support. This scientific area focuses on the capture, management, and use of electronic public health data. While these backgrounds are prominent in the field, the purpose of this course is to provide the history and context of the field.