Training in Health and Biomedical Data Science at Columbia University

Noémie Elhadad, PhD
noemie.elhadad@columbia.edu
@noemieelhadad

Columbia DBMI Training Program

- 120 trainees and graduates (24 current PhD students)
- NLM T15
  - 2015: BD2K supplement on data science
  - 2017: NLM supplement on curriculum and faculty enrichment in data science

Data from Biology, Medicine, and Health

- Observational data from biology, medicine, and health are increasingly prevalent, in larger and larger amounts
  - Electronic health records, biomedical literature, self-reported and tracked health data, Internet and social media

- With the right approach, these data can
  - Help answer critical questions in a brand new way
  - Discover medical and public-health knowledge
  - Improve healthcare
  - Promote health of populations

Columbia DBMI Training Program

- Partnerships with healthcare institutions and international initiatives → Laboratory for innovation for our trainees
  - NewYork-Presbyterian Hospital
  - Observational Health Data Science and Informatics (OHDSI)
  - eMERGE

Data Science at Columbia University

- Columbia Data Science Institute
  - 7 research centers, including Health Analytics
  - 200+ faculty across 9 Schools (80 new faculty)
  - General training opportunities: Certificate, Masters in Data Science

- Fertile ground for research mentorship in data science + health
  - Experts in informatics, statistics, biostatistics, computer science, applied math, etc.

- But: unmet need to train students both in the **fundamentals** of data science and in the **health and biomedical ecosystem** that generated these data and will use the product of informatics research

Training objectives for health data science at Columbia

1. Train students in computational, data-driven methods that can solve biomedical and health problems

2. Promote understanding of the socio-technical processes that shape the way biomedical and health datasets are generated and used

3. Instill in students the methodological principles of “doing” data science as part of the biomedical and health ecosystems
   – e.g., be cognizant of and proactive about reproducibility needs in biomedical data science research

Research Mentorship Objectives

1. Train to work in multi-disciplinary, data-science teams
   – Interactions with researchers and fellow trainees from across departments and schools at Columbia
   – Co-mentorships between informatics and stats/CS faculty

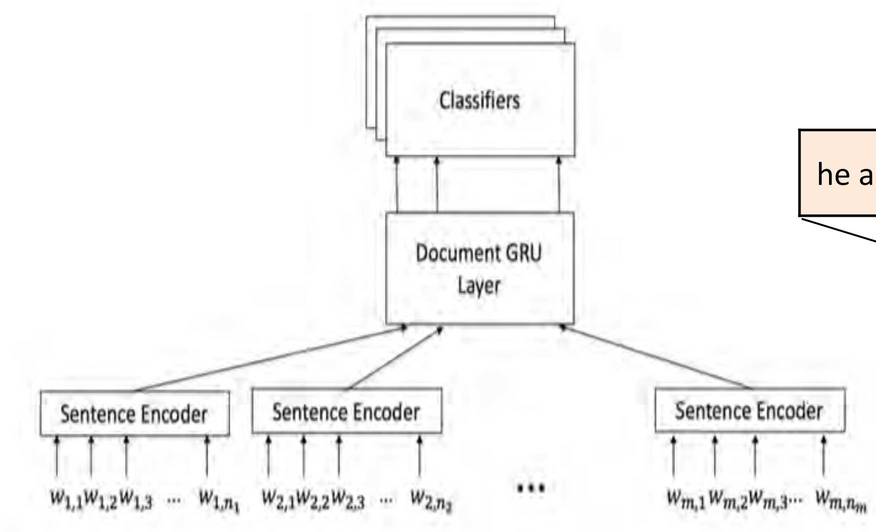
2. Support students to become the next generation of investigators in biomedical data sciences
   – Strong skill set in disseminating for audiences with varied backgrounds, all relevant to data and biomedical sciences.

Interpretable Deep Learning for Clinical Language Processing

Extreme, Multi-Label Classification:
Assign ICD code(s) to discharge summary
ICD9 codes: 9,000 potential labels

Contributions:

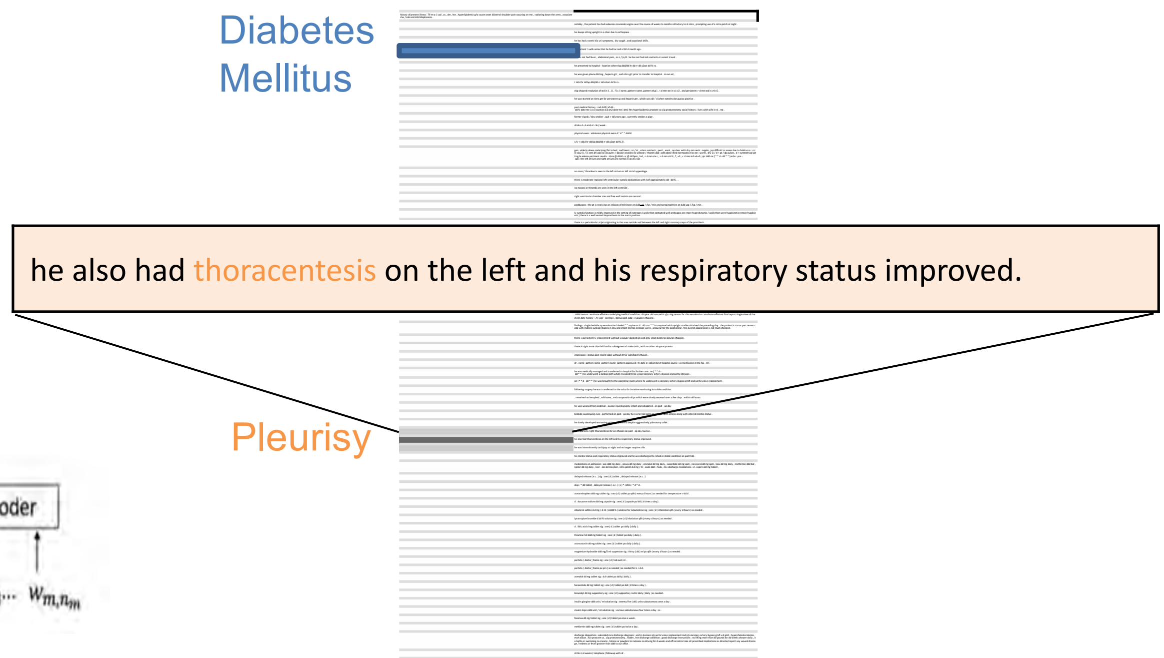
- Designed a hierarchical deep learning model (HA-GRU)
- Compared to two state of the art deep neural nets (CBOW and CNN)
- HA-GRU: Learn representation of words and sentences
- HA-GRU: Model can trace back significant sentences that explain model decisions



Results: (1) State of the art ICD coding algorithm (F-measure)

	ICD9 codes		Rolled-up ICD9 codes	
	MIMIC II	MIMIC III	MIMIC II	MIMIC III
SVM	**32.02%**	38.97%	47.61%	52.56%
CBOW	30.01%	31.18%	42.49%	42.16%
CNN	31.65%	**41.10%**	46.44%	52.14%
HA-GRU	-	-	**51.21%**	**53.02%**

(2) Visualizations for deep learning NLP model



Bayesian formulation of deep learning in healthcare

Proceedings of Machine Learning for Healthcare 2016 JMLR W&C Track Volume 56

Deep Survival Analysis

Rajesh Ranganath RAJESHR@CS.PRINCETON.EDU
Princeton University
Princeton, NJ 08540

Adler Perotte ADLER.PEROTTE@COLUMBIA.EDU
Columbia University
New York City, NY, 10032

Noémie Elhadad NOEMIE.ELHADAD@COLUMBIA.EDU
Columbia University
New York City, NY, 10032

David Blei DAVID.BLEI@COLUMBIA.EDU
Columbia University
New York City, NY, 10027

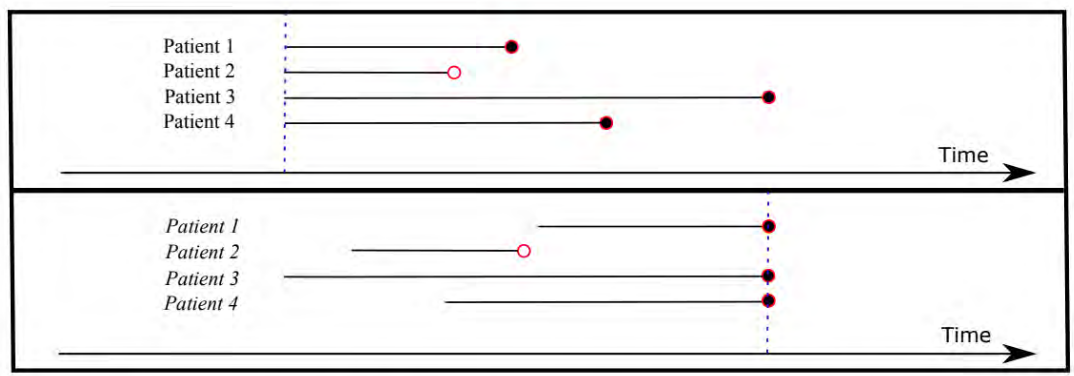


Figure 1: A comparison of traditional survival analysis (top frame) and failure aligned survival analysis (bottom frame). A filled circle represents an observed event, while an empty circle represents a censored one. In the case of standard survival analysis, patients in a cohort are aligned by a starting event. In failure aligned survival analysis, patients are aligned by a failure event.

JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY

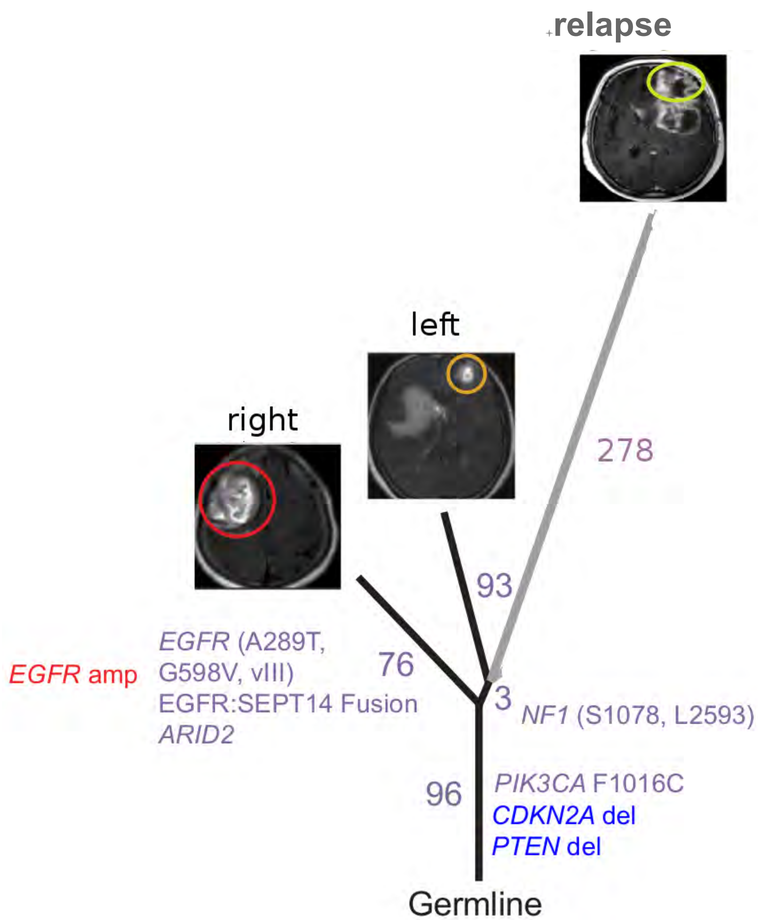
PUBLISHED BY ELSEVIER

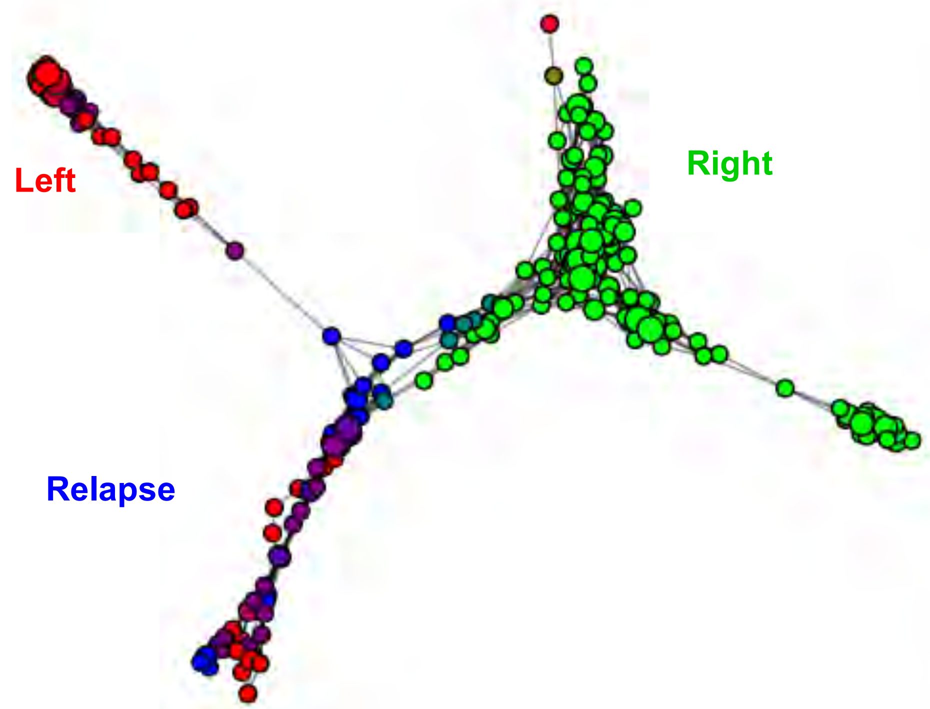
VOL. 68, NO. 16, 2016
ISSN 0735-1097/$36.00
http://dx.doi.org/10.1016/j.jacc.2016.07.761

Coupling Data Mining and Laboratory Experiments to Discover Drug Interactions Causing QT Prolongation

Tal Lorberbaum, MA,[a,b] Kevin J. Sampson, PhD,[c] Jeremy B. Chang, PhD,[b] Vivek Iyer, MD, MSE,[d]
Raymond L. Woosley, MD, PhD,[e] Robert S. Kass, PhD,[c] Nicholas P. Tatonetti, PhD[b]

Understanding the role of tumor heterogeneity in GBM under therapy:
Topological data analysis in single cells

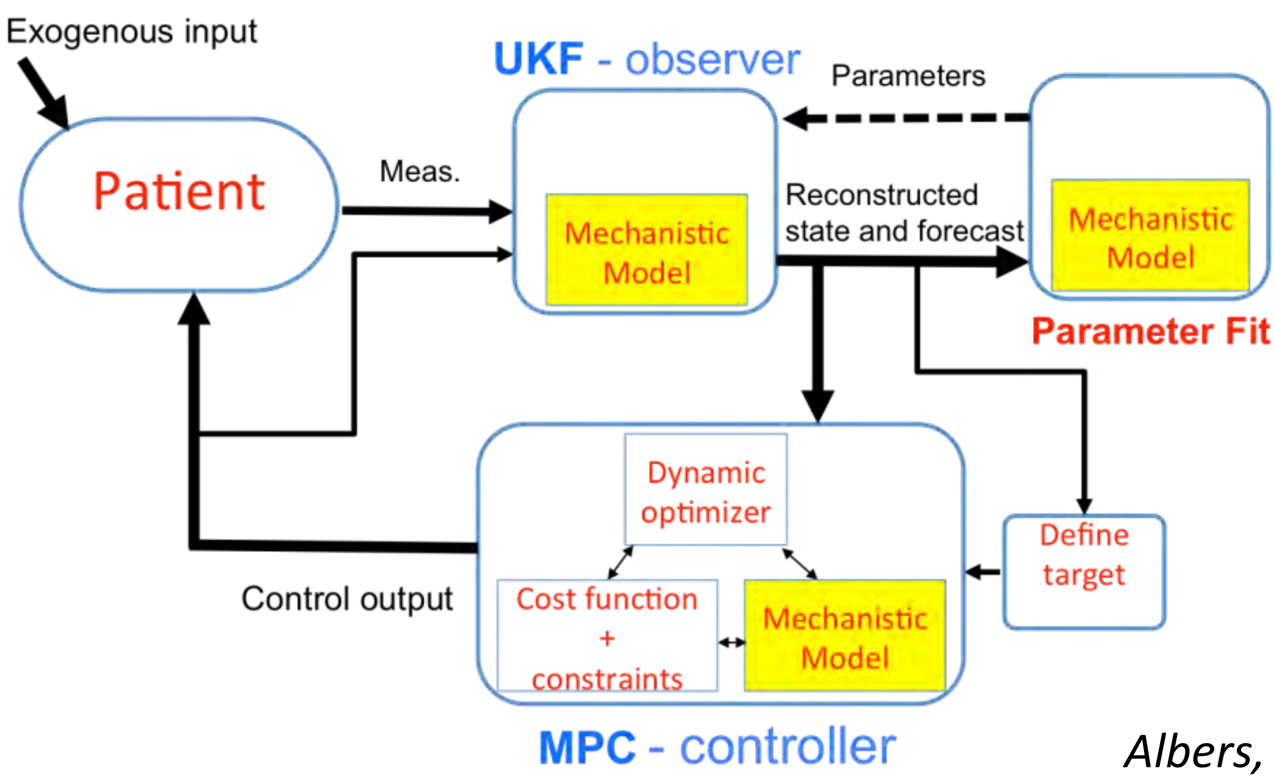


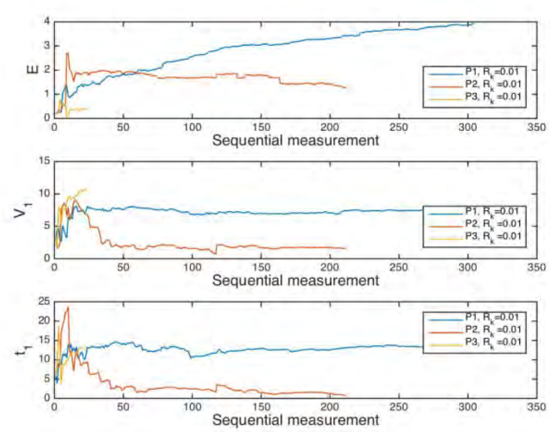


Nature Genetics (2017).

Data assimilation in diabetes

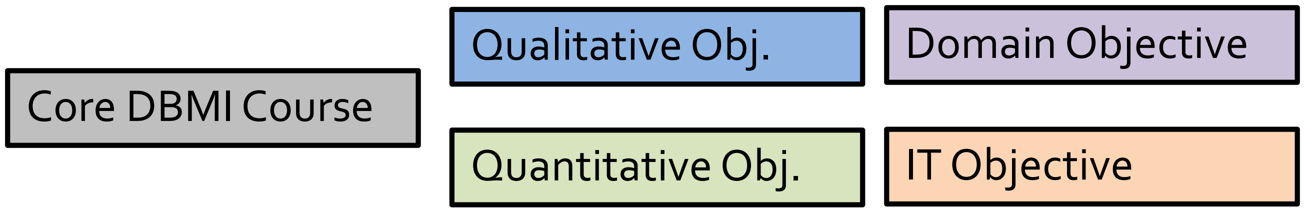
- **Joining mechanistic models & empirical data**
  - Glucose, insulin production, excretion, etc.
  - Estimate parameters from data
  - New: accommodate sparse, irregular, noisy data
  - Constrain the search space



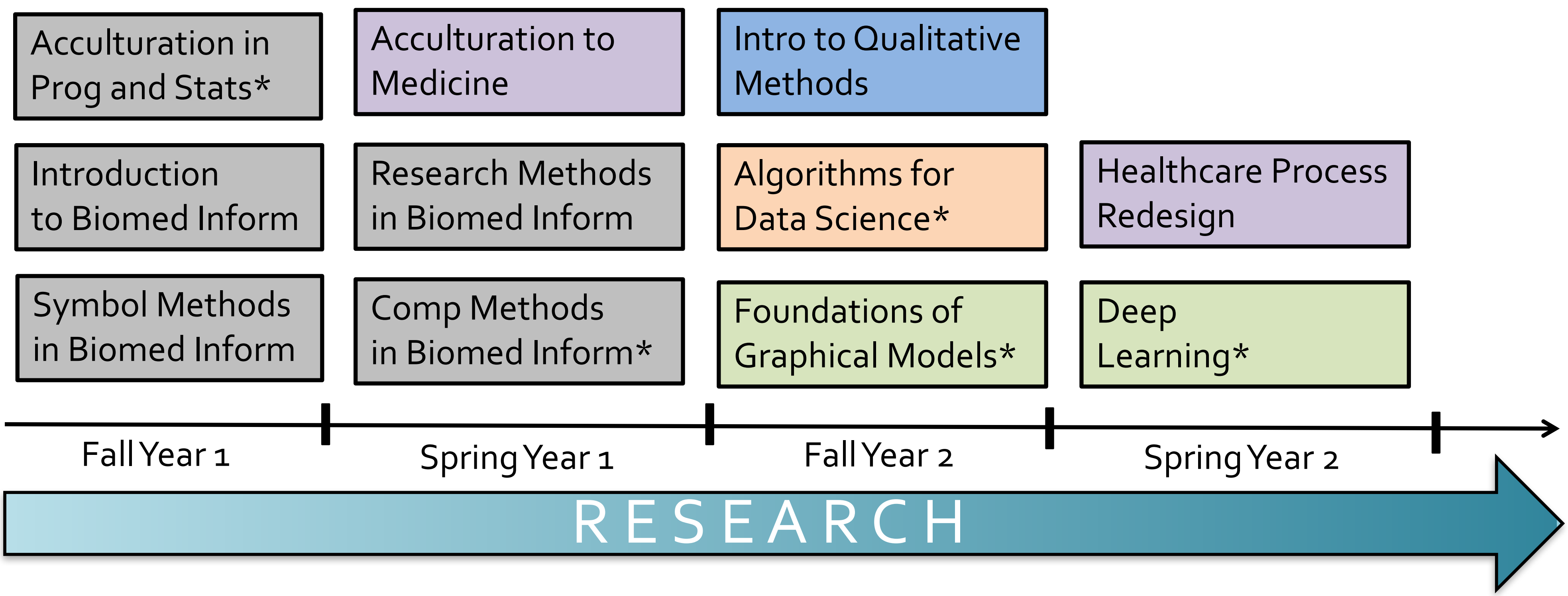


Albers, et al. PLOS Comp Bio (2017)

Curriculum



- Example course trajectory example for student in data science track with focus on EHR data and healthcare



Diversity of students and backgrounds: Acculturation to Programming and Statistics

- 1st-semester course (open to all DBMI students)
  - Introductory data science fundamentals
  - Computing (e.g., Linux environment, Python, Data Persistence)
  - Statistics (e.g., sampling, estimation, basics of prediction)
  - Reproducibility (e.g., Git, GitHub)

- Flipped classroom; focus on "doing"
  - Lectures/readings outside the classroom
  - Labs in the classroom with real-world, very large health datasets
  - Two instructors + 1 TA for 12 1st-year students
  - Rotating teams of 3 students for each lab

Evaluation

- Student Feedback
  - Formal course evaluation and direct interaction
- DBMI Training Committee Feedback
  - Review course evaluations, discuss feedback and the syllabi with the course instructors, and propose changes
  - Meet with elected student representatives regularly
- External Advisory Committee Feedback
  - Russ Altman, Ted Shortliffe, Kevin Johnson, Justin Starren
  - Senior researchers in data science: Dr. David Blei (CS and Statistics) and Dr. Shih-Fu Chang (Electrical Engineering, CS, Senior Vice Dean Eng)
- Student Enrollment
  - New data science courses and the overall track in data science
- Impact on data science research within and across DBMI
  - Number of research papers published by students enrolled in the courses
  - Number of projects and collaborations that started from a project in one of the proposed courses