Program Meeting for: Predoctoral Training in Biomedical Big Data Science (T32/T15) - Room G1 - November 13, 2015

Moderators: Erica Rosemond (lead) and Susan Lim
In Attendance: Bill Noble, Justin Starren, Umit Catalyurek, Michael Kosorok, Mark Forest, George Hripcsak, Matteo Pellegrini, Vijayaraj Nagarajan (NIH intramural), Jarek Meller (BD2K-LINCS), Ebony Hughes (NLM)

Agenda:

- Introductions
- Program Overviews: Training Directors are invited to give a short introduction about their program to the group regarding the student pool, curriculum and courses offered, and evaluation efforts of their programs. (5 minutes each, followed by discussion)
- Potential Topics for Open Discussion:
  - What are the commonalities among the training programs (curricular structure, scientific areas, etc)?
  - What common metrics are being used to evaluate the program, students, and (if applicable) new courses developed for the program?
  - Discussion of providing career guidance at this early stage?
  - Can we identify best practices in training and education in big data at this early stage?
  - What are the core competencies for a biomedical Big Data Predoctoral Training Program?
  - How can we involve the predoctoral trainees in the BD2K AHM next year?

Meeting Synopsis:

This was the first meeting of the BD2K T32 and T15 programs. In attendance were the PIs from the 4 T32 programs and 2 individual supplements to T15 programs. Individual PIs introduced their programs by describing the number and types of departments involved, the PhD programs the students are recruited from and the coursework the program is developing. The approaches to training predoctoral students in biomedical big data were diverse and it was noted that different students will require tailored training depending on their academic backgrounds. As such, core competencies for students have to be flexible and be applicable to the next stages of their career. In addition, the development of a semester long course in a certain topic was not seen as an efficient use of student's time. It was suggested that as opposed to over-burdening the students with coursework, which is in addition to their home PhD program's required coursework, that offering didactics in a modular format with a hands-on or applied approach would be an optimal use of their time. It was noted that though the training programs only support 2-6 students per year there is added value to the students not being directly supported by the training grant by having access to the courses, seminars, and journal clubs. One PI suggested that a successful outcome of the training grant was that general biomedical big data competencies across the academic institution were elevated. Action items from the discussion included a venue (wiki, etc) for sharing curriculum and resources for education and suggested that the R25 PI could be used as consultants for their content development and could provide webinars to the T32 PIs to disseminate their developed content. Another point of discussion was a need for guidance on the use of metrics for short-term as well as long-term success of these training programs.

Meeting Notes:

Introduction to the Individual Training Programs:

Bill Noble (representing) (and Thomas Daniel, Adrienne Fairhall and Daniela Witten - not present) - UNIVERSITY OF WASHINGTON

- program includes 7 departments and applied math; leverage existing coursework - IGERT from NSF - in Computer Science and stats; IGERT provides an advanced data science option with the degree;
- core curriculum in neuroscience and genomics
- set aside money for a course - 2 courses - from biostats - intro to data science and intro to machine learning; for students with a genomics or neuroscience background
- not PhD granting; working towards receiving an advanced data science option on the student's transcript

Mark Forest and Michael Kosorok - UNIV OF NORTH CAROLINA CHAPEL HILL

- certificate program; 20 dept participate; 5 modules rolling out in the Spring - cancer genomics to virology etc - biomedical collaborations - to build; multiplier effect; 28 apps, 20 additional applications; modules are 5 weeks; introduction to new cultures - need a biomedical question being asked; mathematical or computational tool; domain coverage - gap area for the student;
- expectations after 5 week rotation - reproduce some of the work in papers from collaborations; quant students go into the lab and collect the data; biological student will learn how to analyze; get to apply for funding by the Training grant - interviewing opportunity for student and mentor; end result is publication

Justin Starren (and Diego Klabjan - not present) - NORTHWESTERN UNIVERSITY AT CHICAGO

- first students in the summer; joint effort with analytics and informatics program; covering the waterfront - genomics to population science; focus on tool builders; develop data science courses - one for biologists; and one for population researchers; Dean will require all biomedical students to take one of these courses; statistics and data science will be required in the future
- tool builders versus tool users - users may be testing the waters; have options for both types of students; teaching each other;
- courses would be very different for tool builders versus tool users;
- PhD granting program

George Hripcsak - COLUMBIA UNIVERSITY HEALTH SCIENCES

- 4 slots - 24 years into current T15 - biomedical informatics; IGERT CS program separately; new institute; new PIs/faculty; additional courses; partnership between the data scientists and the domain scientist; PhD in data science; data science track; intro course - acculturation to statistics and programing; flip classroom
- teaching currency - challenges

Umit Catalyurek (and Philip Payne - not present) - OHIO STATE UNIVERSITY

- T15 - recruiting students now; add 3 courses to the program; intermediate courses - get them into data science and visualization and HPC; external partners - external rotations in company - 1 semester; may produce new research projects for

Matteo Pellegrini - UNIVERSITY OF CALIFORNIA LOS ANGELES

- recruited 6 students; matching funds from UCLA - 1 extra student; build from bioinformatics program - 2-3 year - 4 electives in big data or biomedical informatics; joint mentorship with clinical mentor; require internship - industry in the summer; big data challenge - large data set and compete;

Discussion Items:

Core Competencies:

- NSF - workshop - Training Students to Extract Value from Big Data - provide PDF to PIs for a reference
- biological versus quants - what are the core competencies for reach?
- what are the competencies that are expected for the next level?
- programming, inference - common sense and literacy
- Big Data U - will provide a list of online courses when the education discovery index is developed
- suggest to provide small "chunks" of courses - students do not want ¼ or semester courses

How can we incorporate the predoctoral students in next year's PI meeting?

- E-science institute may be an option for hosting a student meeting and have our trainees involved
- other options - Sloan foundation, NSF joint meetings

ACTION ITEMS (from discussion):

- a venue to disseminate materials developed from the BD2K short courses (R25s)
    - suggestion for the R25s to provide short webinars
    - Suggest a wiki
    - suggest provide funds for course directors of the R25s to travel to consult with the training program directors
- need to identify evaluation metrics - short-term and long-term for the training programs

- perhaps look to the IGERT programs for metrics
- identify a consult for the training PIs for evaluation metrics
- surveys?
- suggest to look to the CTSA education metrics for ideas