New NSF-Supported Data Science Training Program

Opportunities for PhD students

UNIVERSITY of WASHINGTON

NRT Program goals and details

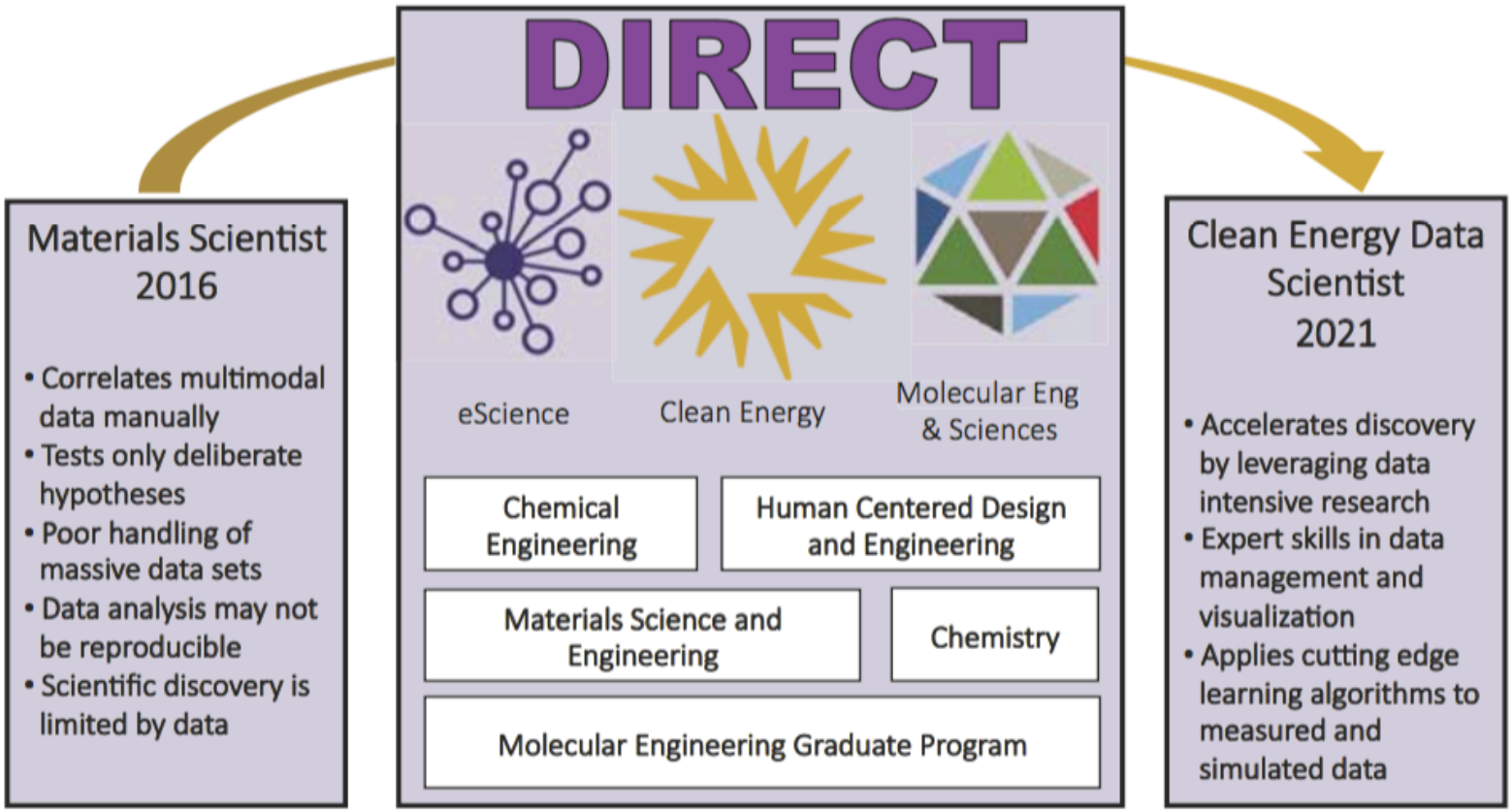
- NSF National Research Traineeship (NRT) is designed to encourage the development and implementation of bold, new, and potentially transformative models for STEM graduate education training. The NRT program seeks proposals that ensure that graduate students in research-based master's and doctoral degree programs develop the skills, knowledge, and competencies needed to pursue a range of STEM careers

NRT vs IGERT

- **IGERT: NSF's previous graduate training program, last awards made 9/2013. A** traineeship **focused on supporting** individuals **in a thematic, interdisciplinary research setting/program.**

- **NRT: New type of trainineeship started in 2014. Focus is on creating** programs **that serve PhD and MS students and developing needed job skills for 21st century. Much less emphasis on supporting individual traineeships. Much more emphasis in growing a sustainable campus effort.**

NRT DIRECT: Data Intensive Research Enabling Clean Technologies



Who can participate?

- **PhD students (recommended entry point is 2nd year and higher) with interests in materials for clean energy and data science (*no prior data science experience is required!*)**
- **Students must apply (October 15 deadline) and be accepted into the program**
- **Cohort 1 size will be capped at 30 students (mix of PhD and MS)**

Program requirements

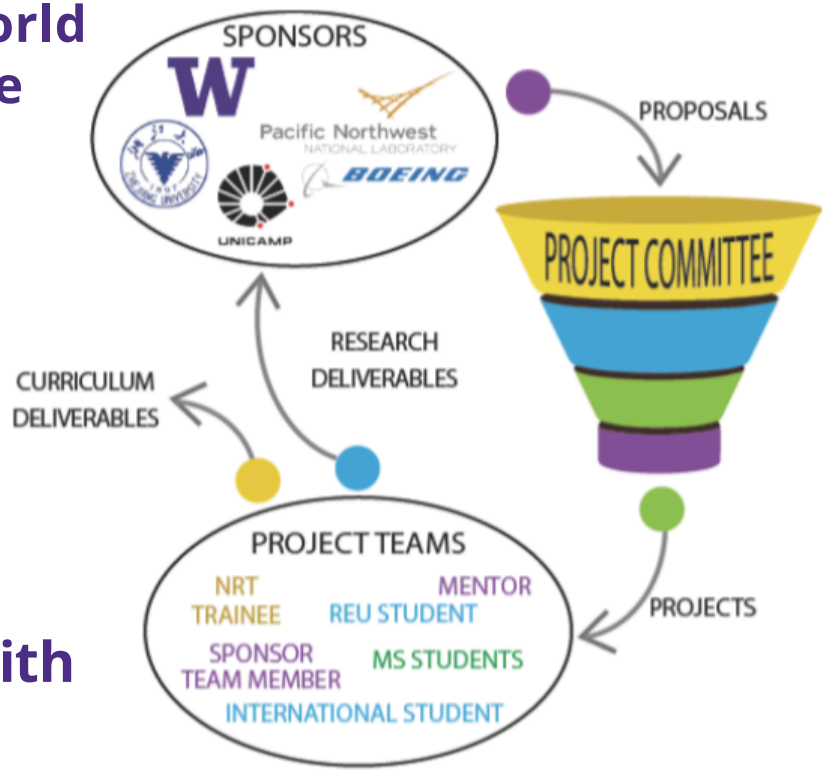
- No prior data science training is needed!
- (1) Students must apply and be accepted into the program: http://www.cei.washington.edu/opportunities/direct/application/
  - Applications will likely be selective
  - Write strong essays about your interest and relevance to your career!
- (2) Complete NRT/DIRECT coursework (WIN17) and attend CEI seminar while you are in the DIRECT program
- (3) Complete a team project related to data science and materials for clean energy

Program benefits and requirements: coursework

- Students who are accepted into the NRT/DIRECT program will have the following benefits:
  - Access to special NRT/DIRECT courses during WINTER 2017:
    - Python Software Carpentry workshop (late AU2016)
    - "Software Engineering for Data Scientists" – customized course taught by Prof. David Beck
    - "Molecular/Materials Data Science" – data science survey class taught by Prof. Jim Pfaendtner

Program benefits and requirements: project based learning experience

- Team based research experience
  - Work on a team based project to apply data science skills to real world research/technical problems in the area of materials for clean energy
- Projects will fall into the cycle of creating new materials for clean energy
- Opportunity to design your own project and lead a team, or join a project team
- NRT faculty and staff will assist with this process
- The goal is for the projects to lead to a co-authored chapter of your thesis!

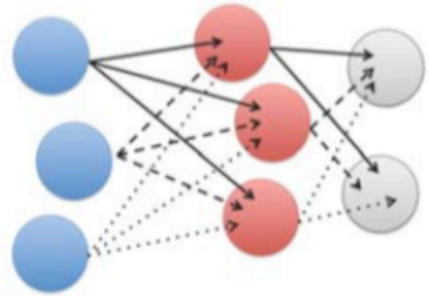


Additional Program Benefits

- **Students who are accepted into the NRT/DIRECT program will have the following benefits:**
  - **Leadership / professional training: access to special seminars for developing project management skills, presentation skills, capacity for leadership**
  - **Completing the NRT/DIRECT requirement is a credentialed item you can put it on resume**
  - **We are planning for an additional "Data Science" credential (e.g.,** "ChemE PhD degree + Data Science Option"**).**
    - **But not 100% sure it will be ready for Cohort 1 students who graduate in 2016.**
    - **I'm trying! ☺**

Project examples: materials design and synthesis

A representative PBL project in the **Design** area would automate generation of neural network models from massive campaigns of MD simulations of liquid solutions – the project sponsor would provide the simulations or the tools to run the simulations and the team would work on the data management and analysis portions. The models would provide new descriptors of electrolyte additives and example data sets for the students in the DIRECT coursework to work on.



The area of data intensive **Synthesis** suggests many new PBL projects. For example, students could develop detailed quantitative structure-property relationships combining predictions from simulations and heuristics from known synthesis routes to design and optimize and validate with new experiments a combinatorial synthesis pathway for novel battery electrolytes.

Project examples: materials characterization

As just one simple example among many possible focus areas for PBL projects in the area of **Characterization**, consider the student who wants to understand a complex compositional relationships arising from growth processes in a ternary semiconductor for optoelectronics applications, Fig. 6(A-D).[31] She obtains hundreds of GB of data on her samples that include hyperspectral energy dispersive x-ray spectroscopy (EDS) images with full x-ray spectra at each pixel. *On the same set of samples* she also has local surface potential, dark conductivity, and photoconductivity data, all spatially correlated. To the eye, the images show what appear to be strong correlations between local elemental ratios and physical properties such as surface potential, but the correlations are not perfect. There is no direct theoretical model for experiment to test, so researchers intuitively search for correlations, but which are causal? What new experiments will optimally reduce uncertainty in any correlations? This data set is also typical in that our publications to date, have contained only a small subset of the available data, and this set has been screened for *anticipated* possible functional relationships – perhaps missing any unexpected discoveries that may lie in the remaining unreduced, uncategorized, and unpublished data.

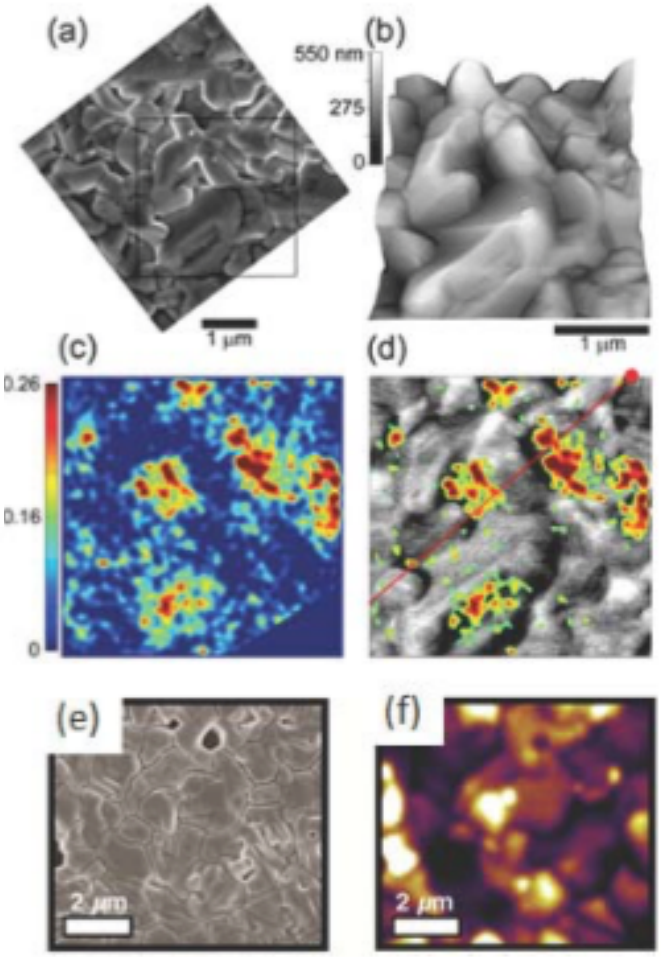


Figure 6: Examples of multimodal, multidimensional image data for energy materials. (A) SEM of copper zinc tin sulfur/selenium film, (B) AFM topography of same region, (C) subset of EDS spectra showing S/Se ratio of same area, (D) overlap of (C) with SKPM surface potential image. (E) SEM image of hybrid $CH_3NH_3PbI_3$ perovskite. (F) Fluorescence lifetime image of same region in (E) each data point is a multidimensional histogram of photon count rates and arrival times from which local lifetimes can be fit. (A)-(D) reproduced from ACS Nano[31], (E)-(F) from Science[27].

Funding opportunities for PhD students in DIRECT Cohort 1

- Students who are engaged with the program and complete the required program elements (seminar, courses and project) will be able to propose a 2Q funded top-off (from CEI/NRT funds) to support a data science themed project related to your thesis work (WI/SP18 or SP/SU18)
- We plan for applications for this top-off, and funding opportunities for future Cohorts to coincide w/CEI grad fellows funding cycle