Career Development Seminar Overview: Big Data/Data Science

Sharada Tilve, PhD – Post-doctoral Fellow, NHLBI, FelCom Career Development Subcommittee
December 13, 2016

The Career Development seminar this month focused on careers in Big Data and Data Science. The panel included specialists working with government agencies, companies and non-profit institutions speaking about the demand of the field and their journey to obtain their current appointments.

Our first speaker was Dr. Subha Madhavan, Founding Director of the Innovation Center for Biomedical Informatics (ICBI) at the Georgetown University Medical Center and Associate Professor of Oncology. She is responsible for several biomedical informatics efforts including the Georgetown Database of Cancer (G-DOC), the NCI *in Silico* Research Center for Excellence, and the Lombardi Cancer Center's Biostatistics and Bioinformatics shared resource. She leads the Biomedical Informatics program of the Georgetown-Howard Universities CTSA. She was the PI on the Breast and Colon Cancer Family Registries data center that coordinates public health and epidemiology data across 12 sites in the US, Australia, and Canada. More recently, she has partnered with the FDA on the Center for Excellence in Regulatory Science program to develop evidence bases for pharmacogenomics and vaccine safety.

The main focus of her talk was to convey the necessary skills needed to build a career in Data Science. According to Dr. Madhavan, an ideal candidate would require minimal supervision and would possess the ability to learn new programming skills on their own time. In addition, the candidate should have superior organization skills and understanding of standard parameters. Oral and written communication skills are also a must. Finally, it is important that the candidate shows a willingness to continue learning. This can be demonstrated by taking certification courses from online sources such as Coursera or statistics.com.

Our second speaker was Dr. Karen Ketchum, a molecular genome biologist with experience in genomics, proteomics, and integrative technologies for managing high throughput sequencing and other big data ('omic) programs. She has also participated in cross functional disease-association studies, biomarker discovery for personalized medicine, business development and is familiar with FDA regulations for drug approval and device clearance through graduate course work in the Johns Hopkins University Bioscience Regulatory Affairs program. Dr. Ketchum holds a Ph.D. from McGill University and completed a postdoctoral fellowship in the Department of Genetics at Yale University.

Dr. Ketchum pointed out that an applicant for a data science job position should be able to truly realize the goal of the hiring company. They should have basic to intermediate understanding of software engineering and/or bioinformatics. She commented that accuracy in data analysis and efficiency of exchanging data are key in every job in this sector. Among soft skills, Dr. Ketchum stressed the importance of email style and content (particularly the ability to summarize work clearly and concisely), efficient teleconference presentations, and active participation in offering alternate conclusions.

Our third speaker was Dr. Wendy Wong. She has in-depth knowledge in statistics, computer science and biology. She was previously a computational biologist at the Wellcome Trust Sanger Institute where she worked on genetic interaction map for *C. elegans*. As a statistician, she also worked at Winton Capital Management in London on Transaction Cost Analysis. Before joining Inova Translational Medicine Institute (ITMI), she was a bioinformatics scientist at Illumina Ltd. in Cambridge, UK, working on algorithms on next-generation sequencing data. She is currently the interim director of bioinformatics at ITMI. Dr. Wong has a Ph.D. degree in Biometry (Biological Statistics) from Cornell University, where she developed statistical models to model the processes of molecular evolution using sequence data. She holds M.S. degrees in Biometry and Computer Science from Cornell University. She also has a B.S. degree in Genetics, Bacteriology, Mathematics, Computer Sciences and Statistics from University of Wisconsin - Madison.

Dr. Wong commented that it is very important to be proficient in programming skills such as R, Python, C++ and/or Java. Understanding of machine learning and good skills in statistics are key. Anyone new to data science should expect to spend their initial months on the job re-organizing and cleaning up data that's already in the company data base. They should be willing to deviate from their original job description if necessary to learn new skills which might address immediate needs of the firm. The ability to adapt to new programs and constantly learning new ones on the job is very important.

Our fourth and final panelist was Sean M. Gonzalez, Co-Founder of Data Community DC and President of General Influence, LLC, operating as a Data Science Consultant and Recruiter. Sean believes great people are all around us, and focuses on helping people find each other through data, visualization, and building relationships. Sean has 12 years' experience as a DoD contractor, focusing on algorithms for missile defense and networks, and integrating small business technologies through SBIR/STTR. Since leaving DoD, Sean has transitioned into data science consulting and supporting the data science community in Washington, DC.

He noted that the majority of people hired into the data science field attend professional meetups and network extensively. He suggested that those who are interested in this field from the DC area should join "Data Community DC" or other societies like Night Owl. He shed some light on consultancy as an option for people who want more freedom in terms of choosing their own projects. In terms of honing independent skills set, he urges candidates who are interested in data science to join local hackathons and practices on open-source big data sets.

Take-home Messages:

1. If you do not have a degree or much experience with Big Data and Data Science, online certification courses are a good place to start. Because programming trends are always changing, it is more important to demonstrate your willingness and ability to learn rather than to master one particular programming language.

2. If possible, apply for a fellowship with a company or academic setup to gain hands on experience of data management projects. This will give you an edge over other applicants when the time comes to apply for a job.
3. Make use of professional meetups and local hackathons for networking to get leads on job openings or short-term consulting opportunities.
4. A lot of work in this field requires the cleaning of existing data rather than creating something from scratch. Thus, employers are looking for accuracy, the ability to effectively summarize, upholding standard parameters and the ability to learn on the job.