NSF Cyber Carpentry: Data Life-Cycle Training

Overview

Massive data collections (i.e., “Big Data”) have ushered in a paradigm shift in the way scientific research is conducted and new knowledge is discovered. The traditional observe-hypothesize-test model of small-scale scientific endeavor is increasingly augmented and in some cases supplanted with collaborative scientific research applying complex patterns of data integration and analysis involving multi-disciplinary teams from distributed organizations brought together to solve a common problem. Emerging cyber-infrastructure solutions necessitate addressing the needs of domain scientists from multiple angles, including data access, metadata management, large-scale analytics and workflows, data and application discovery and sharing, and data preservation. The Cyber Carpentry workshop aims to make it easier for trainees to learn all aspects of data-intensive computing environment and more importantly to work together with other researchers with complementary expertise - pairing domain scientists with computer and information scientists.

This two-week workshop will provide an overview of best data management practices, data science tools and concrete steps and methods for performing end-to-end data intensive computing and data life-cycle management and for promoting reproducible science and data reuse. The workshop will be held at the University of North Carolina at Chapel Hill from July 15th to 26th 2019. Travel and accommodation will be provided and students will receive a certificate of completion from the UNC School of Information and Library Science on successful completion of the workshop.

Workshop topics will include concepts and practices in:

- Data Life-Cycle Management and Policy Automation for increasing sustainability
- Data and Metadata Curation for effective data preservation
- Metadata, Ontology and Provenance for increasing interoperability
- Concepts in Federation for effective collaboration and sharing
- Abstraction, Virtualization and Containerization for reproducible science
- Effective Collaboration Techniques
- Information Analytics and Scientific Workflows
- Computation in cloud- and cluster resources

This workshop is funded by the National Science Foundation under the Cyber Training program to prepare, nurture and grow the national scientific workforce for creating, utilizing, and supporting advanced cyberinfrastructure (CI) that enables cutting-edge science and engineering and contributes to the Nation's overall economic competitiveness and security.

Instructors

The instructors for this course are the investigators of the successful Datanet Federation Consortium (DFC) project which ran between 2013 and 2017. The DFC project implements a collaboration framework that promotes sharing in and across multiple scientific and engineering disciplines. The chief outcome of DFC is an integrated policy-oriented federation platform which intertwines human interactions, policy sets, and evolving cyberinfrastructure leading to a system that has proven to be extensible, sustainable and applicable across multiple disciplines. DFC meshes three core concepts of (a) virtualization, (b) policy-driven automations, and (c) federation to provide the extensibility, sustainability and technology independence needed for long-term scientific collaboration. The instructors for this course come from designers, developers and users of the DFC system. More information about the team can be found https://cybercarpentry.web.unc.edu/instructors.

Eligibility

Space for this workshop is limited. Post-doctoral fellows and doctoral students in science, humanities and engineering disciplines are encouraged to apply. Participants will be selected on the basis of their current research or work activities, previous experience with open science practices, data management techniques and analysis methods, and their current or former opportunities to access training in these areas.

Both doctoral students and post-doctoral researchers are eligible to apply. We seek applicants with the following types of expertise:

- Domain scientists who are using large data collections in their research
- Computational scientists who are designing tools and techniques to facilitate data-intensive work
- Information scientists who developing and implementing data management standards and practices

Participants will be selected on the basis of their academic qualifications, current research activities and prior experience with open science practices, data management techniques, and analysis methods.

Priority will be given to applicants who have not had opportunities to access training in data management practices and data science tools. Women, under-represented groups and persons with disabilities are especially encouraged to submit applications.

Travel and Accommodation Support

Course participants will receive support to cover the cost of an economy round trip airfare within the contiguous United States, and will be provided accommodation in Chapel Hill for the duration of the course.

How to Apply

Applicants should complete the online application form at the website below. The application form requests basic demographics in addition to information about research background and data science training and skills. A recommendation letter from the advisor or post-doc mentor will also be needed. Please also submit a 2-page Curriculum Vita in PDF using the NSF guidelines.

Your application will only be considered complete if we have received a completed application form, a recommendation letter from the advisor, and a 2-page CV. Details of how to send these materials are on the course web site: https://cybercarpentry.web.unc.edu/

Deadline

For applications to receive full consideration they must be received by **5 pm Pacific Time on Monday, APRIL 15th, 2019.**

Information

More information can be found at the training web site: https://cybercarpentry.web.