NISS WORKSHOP

NISS

R & SPARK: TOOLS FOR DATA SCIENCE WORKFLOWS

DATE: May 30-31, 2018, 9 A.M to 5 P.M.

VENUE: Bureau of Labor Statistics Conference and Training Center
Conference Rooms 1-2. Postal Square Building, 1st Street, NE,
Washington, D.C. 20212-0001.

CLASS CAPACITY: 40

FEES:

- $760 for employees of NISS Affiliates, or for students
- $990 for non-NISS Affiliates

COURSE OUTLINE: R is a flexible, extensible statistical computing environment, but it is limited to single-core execution. Spark is a distributed computing environment which treats R as a first-class programming language. This course introduces data structures in R and their use in functional programming workflows relevant to data science.

The course covers the initial steps in the data science process:

- Extracting data from source systems
- Transforming data into tidy form
- loading data into distributed file systems, distributed data warehouses, and NoSQL databases, i.e., ETL.

This workflow is illustrated by using the SparkR and sparklyr package frontends to Spark from R.

SparkR and sparklyr are then used as interfaces for modeling big data using regression and classification supervised learning methods. Unsupervised learning methods, such as clustering and dimension reduction, are also covered. Additional methods, such as gradient boosting and deep learning, are illustrated using the h2o and rsparkling R packages. Finally, methods for analyzing streaming data are presented. The course finishes with an in-depth example. The infrastructure and content is containerized for easy download to your laptop using Docker.

PREREQUISITES FOR THIS COURSE: Differential calculus, basic matrix algebra, a statistics course covering regression, basic R.

OPERATING SYSTEMS: MacOS 10.11 (El Capitan) or higher or Windows 10 Professional. Students must bring their own laptops.

EVENT LOCATION: Bureau of Labor Statistics Conference and Training Center. Conference Rooms 1-2. Postal Square Building, 1st Street, NE, Washington, D.C. 20212-0001. Attendees are required to enter through the visitor entrance on First Street NE (between Massachusetts Avenue and G Street, NE) across from Union Station. Do not use the main entrance on 2 Massachusetts Avenue. Please note that food is not allowed in any of the classrooms. Only drinks with caps or lids are allowed into the classroom.

ID REQUIREMENT AND SCREENING: All visitors must present a valid photo ID at the visitor's entrance and pick up a visitor's badge. Visitors and packages will be processed through the x-ray and metal detector screening equipment. Equipment brought into BLS requires property passes. Equipment passes can be picked up from the receptionist. Please arrive 10 minutes early to allow enough time to go through security. Attendees must also check-in with JPSM onsite assistant each day of the course.

CONTACT US: Direct questions about this course to the Instructor E. James Harner at **eharner@mail.wvu.edu** or call him on his cell phone at **304-376-4170**.

INSTRUCTOR:

E. JAMES HARNER

E. James Harner is Professor Emeritus of Statistics at West Virginia University (WVU). He was the Chair of the Department of Statistics for 17 years and the Director of the Cancer Center Bioinformatics Core for 15 years at WVU. Currently, he is the Chairman of the Interface Foundation of North America which has partnered with the American Statistical Association to organize the annual Symposium on Data Science and Statistics (SDSS) beginning in May, 2018. The areas of his technical and research expertise include: bioinformatics, high-dimensional modeling, high-performance computing, streaming and big data modeling and statistical machine learning.

National Institute of Statistical Sciences

1150 Connecticut Avenue NW, 9th Floor, Washington, DC 20036;
Tel: (202) 862-4316;
Fax: (202) 828-4130