OCTOBER 23-25, 2016

GroupHealth
RESEARCH INSTITUTE

2nd Seattle Symposium on
Health Care Data Analytics

Short Course #2

Sunday, October 23

1 – 5 p.m.

Practical approaches to health care analytics in the presence of big data: Case studies in Python and Apache Spark

INSTRUCTORS:	Debashis Ghosh, PhD, and Evan Carey, MS
FORMAT:	One 4-hour session, including practicum. Laptop is strongly recommended.
TARGET AUDIENCE:	Epidemiologists, data scientists, informaticians, data analysts, and statisticians

About the course

The phrase “big data” has become widespread, but what does this mean for the practicing healthcare analyst? How does the presence of big data impact the actual workflow of a practicing analyst in health care? Is “machine learning” superior in the presence of big data? In this workshop, attendees will be exposed to multiple tools useful in the analysis of big data, including Python, SQL, Hadoop, and Spark. The workshop will predominately consist of practical examples with code, with the expectation that students follow along and run code throughout the workshop.

Through instructor-led examples, we will discuss and demonstrate the efficiency of various analytic frameworks for a binary classification problem. We will begin with examples of managing data in SQL and alternatively in a NoSQL environment. Various examples of dimensionality reduction for data relevant to healthcare in the pre-modeling environment will be covered.

We will further consider more complex dimensionality reduction techniques requiring an analytic platform beyond a simple database management system. In order to explore different approaches to a classification problem, penalized regression (LASSO), Random Forests, and support vector machines (SVM) will be presented. We will contrast traditional serial optimization approaches (such as Newton Raphson) with parallel optimization approaches (such as stochastic gradient descent). We will follow-up these models with a discussion on the impact of censoring.

Students will be provided with code to run all models ahead of the workshop, thus no experience in these languages is required. All software used will be open source; students will be expected to set up their computing environment prior to the workshop, further details and guidance will be sent to attendees.

GroupHealth.
RESEARCH INSTITUTE

DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY of WASHINGTON

REAGAN - UDALL
FOUNDATION
FOR THE
Food and Drug Administration

Specific learning objectives

- Understand options for managing large healthcare data sources in the pre-modeling environment.
- Understand the difference between traditional RDBMS and NoSQL alternatives such as Hadoop and Spark.
- Understand and define the differences between the model, loss function, regularizer, and optimization.
- Understand serial versus parallel model optimization techniques and the implications for practical approaches to analysis in the presence of big data.
- Understand the impact of increasing dimensionality on different analytic approaches.
- Gain a basic understanding of fitting models in Python
- Gain a basic understanding of fitting models in Apache Spark (using the Python API)

About the instructors

Debashis Ghosh, PhD, is Professor and Chair of the Department of Biostatistics and Informatics at the University of Colorado Denver. In addition, he serves as Associate Director for Bioinformatics for the Center for Personalized Medicine at the University of Colorado Anschutz Medical Campus. Prior to arriving at Colorado, Ghosh was Professor of Statistics and Public Health Sciences at Penn State University. He was previously Assistant and Associate Professor at the University of Michigan. Dr. Ghosh's has published more than 170 peer-reviewed articles, commentaries and book chapters in statistical and scientific literature. He is a Fellow of the American Statistical Association and was recently honored with the 2013 Mortimer Spiegelman Award for outstanding early career statistical contributions to public health.

Evan Carey, MS, received his masters in applied biostatistics from the Colorado School of Public Health (CSPH). He is currently a PhD candidate in Epidemiology at the CSPH. Mr. Carey is a statistician/data scientist with the Veteran Healthcare Administration and the CSPH. Mr. Carey has designed and implemented large administrative cohort studies in the VA system for the past 5 years, working with datasets exceeding billions of rows. Mr. Carey has designed and given seminars in R, Python, and Hadoop programming for data science for multiple fortune 500 companies and federal agencies.