**Proposal Type**: Full Day Short Course  
**Theme**: Data Science and Big Data  
**Title**: Big Data, Data Science and Deep Learning for Statistician

**Description**:

With recent big data, data science and deep learning revolution, enterprises ranging from FORTUNE 100 to startups across the world are hungry for data scientists and deep learning scientists to bring actionable insight from the vast amount of data collected.

In this course, you’ll develop a clear understanding of the big data platform, motivation and use case for deep learning through hands-on exercises to prepare statisticians to be successful data scientists and deep learning scientist. We will also cover the “art” part of data science and machine learning to guide participants to learn typical agile data science project flow, general pitfalls in data science and machine learning, and soft skills to effectively communicate with business stakeholders.

**Outline and Objectives**:

The big data platform, data science, and deep learning overviews are for the audience with statistics background. The data science workflow, pitfall and soft skills are highlight through real-world data science and machine learning problems. The Databricks community edition cloud platform will be used throughout the training course to cover the hands-on sessions: (1) big data platform using Spark through R sparklyr package; (2) introduction to Convolutional Neural Network (CNN); (3) deep learning examples using Tensorflow through R kreas package; (4) soft skills.

The primary audiences for this course are (1) statistician in traditional industry sectors such as manufacturing, pharmaceutical and banking; (2) statistician in government agencies; (3) statistical researchers in universities; and (4) Ph.D. and M.S. graduate students in statistics departments. The prerequisite knowledge is MS level education in statistics and entry level of R knowledge. No software installation is needed to students’ laptop and the cloud platform is easily accessed through browsers such as Chrome or Firefox with the internet connection.

**Instructors:**

Both instructors have Ph.D. in Statistics from Iowa State University and have worked in data science and machine learning areas for a number of years. Dr. Li is a Sr. Data Scientist at Amazon and Dr. Lin is a Data Scientist at Netlify. With deep statistics background and a few years’ industrial experiences in data science, they have trained and mentored numerous junior data scientist. They have taught a similar continue education course without the deep learning part at the 2017 Joint Statistics Meetings at Baltimore, and they will teach similar courses at Joint Research Conference on Statistics in Quality, Industry, and Technology (Santa Fe, NM June 2018), ICSA Applied Statistics Symposium (New Brunswick, NJ June 2018) and Fall Technical Conference (West Palm Beach, FL October 2018).

Dr. Li organized and will present at the Introductory Overview Lecture “Leading Data Science: Talent, Strategy, and Impact” at the 2018 Joint Statistical Meetings. Dr. Li is also an Instructor of Amazon’s internal Machine Learning University and was one of the key founding members of Walmart’s Analytics Rotational Program which bridges the skill gaps between new hires and productive data scientists.

Before Netlify, Dr. Lin was a leader at DuPont on applying advanced data science to enhance Marketing and Sales Effectiveness from 2013-2018 where she provided data science leadership for a broad range of predictive analytics and market research analysis. She is the co-founder of Central Iowa R User Group and blogger of [scientistcafe.com](http://scientistcafe.com).

**Relevance to Conference Goals:**

This short course fit the conference goals very well. It focuses on Big Data and Data Science applications in real-world problems including the new development of deep learning. With the focus on the cloud platform, students can learn the current trend of software and big data infrastructure used by tech companies such that they can expand their programming scope to cover more applications. The short course also includes the needed soft skills discussion to prepare students with better understanding of the data science project flow, pitfalls and communication skills. This course keeps statistician’s background in mind to bridge the gaps between a traditional statistician and a successful data scientist. After taking the course, students will be confident to positively impact their organization by transforming their current traditional statistics team into a data science team or to explore data scientist opportunities for their future career development.  
  
  
**Software Packages**:

We will use R as the programing language with sparklyr, dplyr, ggplot, and keras. The big data platform of Spark and deep learning framework TensorFlow are installed and configured by the above-mentioned packages and the Databricks cloud platform. No installation or configuration of software in students’ laptop is needed. The only requirement is a laptop with a browser such as Chrome or Firefox and internet connection which is one of the great advantages of cloud platform (zero setup and zero maintenance effort).

 