University of Puerto Rico Mayaguez Campus
Faculty of Arts and Sciences
Department of Mathematical Sciences
call box 9000, Mayagüez, P.R. 00681

THE DEPARTMENT OF MATHEMATICAL SCIENCES PROUDLY PRESENTS

COLLOQUIUM

SPRING 2015

Statistics, Data Mining, Big Data Analytics, And Data Science: more than a play of names

Dr. Edgar Acuña

UPRM

May 7, 2015

ABSTRACT

Since some time ago, some well renowned statisticians, such a J. W. Tukey (1915-2000), L. Breiman (1928-2005) and J. H. Friedman(1939 -), have been worried about the treatment of data analysis by statisticians. In 2001, UC Berkeley's Professor Breiman wrote a controversial paper-- "Statistical Modeling: The Two cultures"—which appeared in the journal of Statistical Science. There, Breiman proposed that statisticians must focus more on algorithmic models rather than on data models if their goal is for the statistics field to use data to solve problems. In the mid 90's, people from the database community had entered the world of data analysis, originating a new area of research named Data Mining. This field of research aimed to obtain knowledge discovery from large datasets utilizing tools from Statistics, Machine Learning, and new ones that were being produced by data miners. In the mid 2000's, research in Data Mining decreased when researchers realized that the use of high performance computing was necessary to analyze datasets which size were quickly increasing from gigabytes to petabytes. At the same time Google and Yahoo introduced new computational tools such as Hadoop and MapReduce to analyze very large and complex data.

Although the term "Big data" was introduced by the end of the 90's, it was not until 2011 when MacKinsey Global Institute's report was published and the Gartner Group defined the term, that "Big data" gained popularity. Researchers analyzing "Big data" were called "data scientists".

In 2009, Data Science, appeared as a new discipline, propelled by the world of social media (Twitter, Netflix, Facebook, LinkedIn, etc.) , where plenty of data must be analyzed. However, already, two statisticians—J. Wu (1997) and W. Cleveland (2001)— have proposed to rename the discipline of "Statistics" as "Data Science" .

In this talk, we compare the four disciplines previously mentioned and we will give some ideas as to how Statistics can compete with the emerging discipline of Data Science.

This talk will be delivered in Spanish.

Monzón Building, Room 201, 10:45 AM
Refreshments will be served
15 minutes before the colloquium, M213