**TITLE**: Scale-Out Data Science with R and Python

**Tomas Singliar, Mario Inchiosa, John-Mark Agosta, Hang Zhang**

Microsoft Corporation

One Microsoft Way   
 Redmond, Washington, 98052, USA   
 {tomas.singliar, marinch, joagosta, hangzh}@microsoft.com

**CORRESPONDING AUTHOR:** Tomas Singliar – Tomas.Singliar@microsoft.com

**HANDS-ON TUTORIAL DURATION:** 6 hours (2x 3 hour sessions)

**TARGET AUDIENCE:** Intermediate level in knowledge and practice of machine learning, R and Python

**ABSTRACT**   
 Python and R dominate the domain of data science software. However, when it comes to scalable analysis, or deployment of trained models into production, barriers still exist. Many data scientists are hindered by a limited suite of available functions to handle large datasets efficiently, and knowledge about the appropriate computing environments to scale R and Python scripts from desktop analysis to elastic and distributed cloud services. Another productivity limitation is the tedium of the experimentation loop in which the right preprocessing, model, and hyperparameters are found.

In this tutorial, we will demonstrate how to create scalable machine learning pipelines in R and Python with emphasis on scaling on Spark clusters. We will model the data science journey by first prototyping locally and then show how to move the data science process to the Cloud, to exploit larger compute resources and data colocation that various Spark implementations offer. In particular, the attendees will see how to build, persist, and consume machine learning models using distributed machine learning functions in Python and R. Armed with a distributed compute platform, we will show how Microsoft’s AutoML library can automate the search for the best model.

We will provide hands-on exercises drawing on recent examples from time series forecasting, Active Learning, and Reinforcement Learning. Code samples will be available in a public GitHub repository. Spark and AzureML Compute clusters will be the target distributed platforms; participants will do exercises on Data Science Virtual Machines using RStudio and Jupyter notebooks.

**TUTORIAL OUTLINE**

1. **Introduction**:

Scaling up your data science process - issues and solutions

What limits the scalability of your code in face of large data? What techniques can be used to overcome those limits? What libraries can I use in Python? In R?

What limits your modeling productivity? How do I navigate the space of modeling choices - preprocessing sequences, models, hyperparameters?

2. **Hands-on exercises and demonstrations**:

End to end scalable data process

Data exploration, wrangling, visualization, modeling and deployment on single node Data Science Virtual Machines and Spark clusters

Scalable analysis on single nodes: Analysis with data on disk, in-database, and in Spark

Distributed model search and parameter optimization in python with AutoML.

Deployment of ML models as web-services APIs with Azure ML python SDK, with parallel scoring on an elastic cluster.

**PREREQUISITES**

**Participants should come to the sessions with access to an Azure subscription. You can use Azure’s free tier.**

**(**https://azure.microsoft.com/en-us/offers/ms-azr-0044p/**)**