

# Proposal Information

|  |  |
| --- | --- |
| **Track** | One of:  Paper – for consideration for the [MS Journal of Applied Research](https://www.yammer.com/microsoft.com/#/threads/inGroup?type=in_group&feedId=3950783&trk_event=search_ac&trk_model=group&trk_len=8&trk_pos=0) (MSJAR)  Poster/Demo – a booth at the poster/demo reception on December 7  Talk – a 25-minute technical presentation  Tutorial – a 2-hour hands-on tutorial  \* Remember: You can submit the same work in multiple tracks, but you need to create a separate submission for each so that they can be judged against other submissions in each track. |
| **Session Title (required)** | Using Team Data Science Process (TDSP): realizing the benefits of standardization and DevOps in data science |
| **Speakers/Authors (required)** | **Debraj GuhaThakurta**  Debraj GuhaThakurta is a senior data scientist lead in AI & Research, Cloud Data Platform, Algorithms and Data Science, where he focuses on developing the Team Data Science Process and the use of different Microsoft data platforms and toolkits (Spark, SQL-server, ADL, Hadoop, DL toolkits) for creating scalable and operationalized analytical processes. He has a PhD in chemistry and biophysics and many years of experience in data science and machine learning applications, particularly in biomedical and forecasting domains. Debraj has published more than 25 peer-reviewed papers, book chapters, and patents.  **Buck Woody**  Buck Woody (<http://buckwoody.com/>) works on the Microsoft Machine Learning and Data Science Team, using data and technology to solve business and science problems. With over thirty years of professional and practical experience in computer technology, he is also a popular speaker at many conferences around the world; the author of over 650 articles and seven books on databases and Machine Learning technologies, teaches Database courses and sits on the Data Science Board at the University of Washington, and specializes in data analysis techniques.  **Wei Guo**  Wei Guo is a Data Scientist in Microsoft Cloud AI group. He is working on various data science solutions using big data technologies and advanced analytics. He has extensive experience in credit risk, fraud detection, insurance pricing, and clinical trials. Wei Guo has PhD degree in Statistics.  **Xibin Gao**  Xibin Gao is a Data Scientist at Microsoft Cloud AI Platform (AI and Research) group. He works on building end-to-end predictive analytics solutions and develop toolkits to improve data science productivity. He has experience in the area of text analysis, web ranking, and recommendation systems. He has a PhD degree in Computer Science. |
| **Session Objectives (required)** | In addition to data platforms and data science tools, it is important to have an efficient process in order to improve the efficiency of the development and deployment of data science solutions. Many enterprise data science teams today are facing challenges related to standardization, collaboration, and incorporation of a mature process into their advanced analytics solution development and deployment. In this session, we will address the process-related challenges and how to address them using Microsoft Team Data Science Process (TDSP).  This is going to be a hands-on tutorial with the following objectives for attendees:   1. Learn about the process-related challenges our enterprise customer data science teams are facing today, and how Microsoft’s Team Data Science Process (TDSP) can help address those. 2. Learn how one can incorporate standardization, collaborative development, and DevOps practices in data science projects using TDSP. 3. Learn how to create data science projects using the standardized TDSP structure, artifacts, and documentation templates. 4. How to form collaborative data science teams, and plan and execute data science projects under an Agile development framework in Visual Studio. 5. How to perform collaborative code development and review using a version control system such as Git. 6. How to address DevOps practices, e.g. unit testing, continuous integration in data science projects. 7. How to assess data platform and security options for data science projects 8. Finally, learn how to instantiate and use TDSP in Azure ML workbench (Vienna). |
| **Session Audience**  **(required)** | This session is designed for data scientists and data engineers, level 200 and above. The audience should have experience in developing and deploying data science projects. |
| **Description (required)** | Many enterprise data science teams today are facing challenges related to standardization, collaboration, and incorporation of a mature process into their advanced analytics solution development and deployment. We have developed and released Team Data Science Process (TDSP) to address these challenges. Although work is ongoing to improve the process and its features, TDSP is currently assisting multiple data science teams (internal and external) to standardize their data science projects, and adopt collaborative development and DevOps practices.  This hands-on tutorial will show attendees how to use TDSP to incorporate standardization, collaboration and DevOps in your data science projects. We plan to have two segments in this session:   1. **Segment 1 (30 mins)**: Introductory presentation on current process challenges in data science, how TDSP is addressing those, key components of TDSP, and how TDSP integrates DevOps as well as Azure data platforms and tools. 2. **Segment 2 (1 hour and 30 mins)**: Hands-on session in which attendees will learn how to: 3. Instantiate TDSP structure, document and artifact templates for standardization of data science projects 4. Use Agile work planning and execution for data science projects 5. Use version control and code review practices when doing collaborative code development in data science projects 6. Incorporate DevOps practices (e.g. unit testing, continuous integration) in your data science projects 7. Assess data platform and security requirements 8. Use TDSP in Azure ML work bench, Vienna   We strongly feel many data science teams will benefit by adopting the components of TDSP. Several internal and externals teams are already using TDSP in all their data science projects (e.g. Algorithms and Data Science organization in Cortana Cloud Data Platforms, Microsoft Consulting Services, and partners). This tutorial will enable other data science teams in the Microsoft community to get familiarized with TDSP and consider its adoption. In the near future, we are planning to disseminate TDSP through other tutorials in external conferences; the ML-ADS tutorial will help us further assess and improve our content.  References for content:   1. TDSP: <http://aka.ms/tdsp> 2. TDSP in Vienna: <https://github.com/amlsamples/tdsp> |
| **Outline (optional)** | **Please see above** |
| **Paper (required for MSJAR submissions)** | **NA** |
| **Previous Presentations and Papers (optional)** | TDSP has been presented earlier at ML-ADS and other conferences:   1. TDSP talk at ML-ADS Spring 2017 2. TDSP talk at Ignite 2016 (link to presentation [video and deck](https://techcommunity.microsoft.com/t5/Microsoft-Ignite-Content-2016/BR021-Data-Science-Doesn-t-Just-Happen-It-Takes-a-Process-Learn/td-p/17096)) 3. TDSP talk at 5th Global Big Data Conference, Santa Clara, Aug 2017 ([link to deck](https://www.slideshare.net/DebrajGuhaThakurta/team-data-science-process-presentation-tdsp-aug-29-2017)) 4. TDSP Gartner Analyst Briefing (July 2017) [Slide available on request] |

# Guidelines to Create a Strong Proposal – Your **Description** above should describe how your session will meet these requirements

## Conference Focus

## A great proposal helps attendees:

## Stay cutting-edge on the latest machine learning, AI, data science, and big data tools and techniques; or

## Understand best practices in machine learning, AI, and data science; or

## Learn about Microsoft’s machine learning, AI, big data, and visualization tools

And shows how they can be applied to solve real-world problems for Microsoft or its customer, improve customer experiences, or help make better business decisions

## Analytical Depth

A great proposal is appropriate for a technical conference when it:

## Contains core material that is strong enough in analytical / technical terms to be presented -- for example, if a machine learning model was applied, the proposal shows details around the model applied, train-test set and splits, features, metrics defined, the alternative algorithms that were experimented with, and so on. For a more data science focused presentation, the talk should focus on data sources used, data cleansing that has been applied, and statistical analysis that has been performed, as well as actionable insights that have been derived.

## Includes sufficient detail such that a skilled practitioner could reproduce the results given the same input data and where applicable apply the presented method to other input data.

## Makes reference to internal and/or external related work where applicable and explains the differences and/or improvements

## Includes appropriate visuals or tables that help the audience to grasp the core ideas more easily.

## Business Impact A great proposal describes work focusing on a good target.

## A good target is so important to the business, so full of opportunity, that it engages top management commitment and creates momentum.

## It focuses on generating insight rather than merely information.

## It is both ambitious and approachable—ambitious in that it has business impact, and approachable in that it has access to the resources and capabilities to succeed.

## A good target is focusing the analytical investment on Microsoft’s distinctive technologies, which serve our customers in a way differentiated from our competitors, and which create the formula for the company’s success.