Department of COMPUTER SCIENCE
THE UNIVERSITY OF TEXAS AT DALLAS

invites you to attend

# Data Science Conference

Saturday, October 6th 2018 @ JSOM building, UT Dallas

| | |
|---|---|
| **8 - 8:30am** | Registration & Breakfast |
| **8:30 – 12 noon** | Data Science hands-on workshops |
| **12 - 12:30pm** | Registration & Lunch & Networking |
| **12:30 – 6:20pm** | Technical Talks by several prominent speakers from industry and academia |
| **6:20 - 7pm** | Networking with speakers & industry folks |

## Flyer: utd.edu/t/4312

## Registration: bit.ly/cs-conf-utd

| Registration fee | Technical Talks ONLY | Whole day (Workshop & Tech Talks) |
|---|---|---|
| ***Guests*** | $60 | $100 |
| ***UTD folks*** | $10 (***until October 4***), $20 (***Oct 5 onwards***) | |

## Questions? jeyv@utdallas.edu

# Agenda for Data Science Conference

## Saturday, October 6, 2018 @ UT Dallas

### Hands-on Workshops

| Time | Description |
|---|---|
| 8- 8:30am | Registration & Breakfast (Foyer area) |
| 8:30 – 12pm (coffee break @ 10:15am) | Workshops (in parallel):<br>1. Data Science with Python – Prof. Kamran Khan, CS faculty – JSOM 1.118<br>2. Data Science with TensorFlow – Dr. Anurag Nagar, CS faculty – JSOM 1.107<br>3. Interactive Analytics and Story Telling: From Python-Pandas – Matplotlib to Tableau, Viswanath Puttagunta, Divergence Academy – JSOM 1.117 |

### Technical Talks – JSOM 1.118

| Time | Description |
|---|---|
| 12pm – 12:30 | Registration & Lunch & Networking |
| 12:30 – 12:40 | Welcome by Dr. Gopal Gupta, CS Department Head |
| 12:40 – 1:20 | *Keynote address:*<br>**Exponential Technologies and the Acceleration of Everything**<br>Dave Copps, Visionary technologist, entrepreneur, CEO |
| 1:20 – 2:00 | **Fraud Detection in Customer Loyalty Programs,**<br>Subhashini Tripuraneni, 7-Eleven |
| 2:00 – 2:40 | **An Inductive Logic Programming Approach to Explainable Machine Learning,** Farhad Shakerin, UT Dallas |
| 2:40 – 3:20 | **The Automation Mindset: Tips/Tools for Streamlining Data Science Workflows,**<br>Sydeaka Watson, AT&T |
| 3:20 - 3:40 | Tea Break & Networking |
| 3:40 – 4:20 | **Computational Intelligence Tool for Patient Behavioral Health Status Tracking,**<br>Dr. Michael Morgan, Morgan Analytics Research Institute |
| 4:20 – 5:00 | **The Power of Suggestion: An Introduction to Recommender Engines,**<br>Mimi Park, Slalom Consulting |
| 5:00 – 5:40 | **IoT Analytics using Azure Databricks and Pyspark,**<br>Prasad Chandravihar, Lennox International |
| 5:40 – 6:20 | **Big Data Stream Analytics for Cyber Security,**<br>Dr. Latifur Khan, CS faculty, UT Dallas |
| 6:20 – 7pm | Snacks & Networking |

# Description of Workshops & Presenter Information

**Workshop #1:** Data Science with Python

**Presenter:** Professor Kamran Khan, CS faculty, UT Dallas
cs.utdallas.edu/people/faculty/kamran-khan

Prerequisite: Basic familiarity with Python

Following topics will be covered with hands-on examples: Various Python libraries like scikit learn, Nltk, pandas,pydata, numpy etc. which are prominently used for data analytics with Python. Data Munging and Cleaning in Python.[Pandas]

**Workshop #2:** Data Science with TensorFlow

**Presenter:** Dr. Anurag Nagar, CS faculty, UT Dallas

Prerequisite: Basic familiarity with Python

This workshop will introduce participants to TensorFlow (TF), which is an open source, high performance library developed by Google. We will start with basics of TensorFlow and introduce you to stateful dataflow graphs, which are the primary computation engine for TF. Then, we will discuss Estimators, especially pre-made ones that facilitate computation of complex Machine Learning tasks, then move on to neural networks, and deep learning, areas where TFs are most commonly used. We will cover some applications of deep learning, such as Convolution Neural Networks using Keras framework. Throughout the workshop, we will use hands-on lab sessions using Google's Colaboratory environment.

**Workshop #3:** Interactive Analytics and Story Telling: From Python-Pandas-Matplotlib to Tableau

**Presenter:** Viswanath Puttagunta, CTO and Principal Data Scientist at Divergence.AI,
linkedin.com/in/viswanathputtagunta

Histograms, Box Plots and Scatter Plots form the foundation for understanding the distributions underlying the data. Once you understand the distributions, you can come to certain conclusions and communicate insights to the business stake holders, enabling them to make better decisions.

While Python-Pandas-Matplotlib are great while doing exploratory work and sharing data and ideas among fellow programmers and Data Scientists, I found they are not very effective while communicating insights with senior leadership and managers.

When you are in the same room as an executive, it is critical to be able to answer as many questions as possible in the same 15-20 session without scrolling up and down and dynamically programming. This is where Tableau comes in handy.

In this talk, we'll go through basic Exploratory Data Analysis with Python-Pandas-Matplotlib and then with Tableau.

# Description of Tech-Talks & Presenter Information

**Keynote Address:**
Exponential Technologies and the Acceleration of Everything

**Presenter:** Dave Copps, Visionary technologist, entrepreneur & CEO
linkedin.com/in/davecopps

We are living in a world of overlapping revolutions in areas like Robotics, Nanotechnology and Genetics. At the base of each of these revolutions is AI, an Apex technology that is the root cause of this acceleration and abundance.

Dave is known locally and internationally as a technologist and visionary on the role that advanced technologies will play in transforming markets and the world. For the past 15 years, as a technologist, entrepreneur and CEO, he founded, launched and sold three companies all focused on machine learning and artificial intelligence. In 2017 Dave was recognized as Emerging Company CEO of the Year in Texas.

Most recently Dave served as CEO of Brainspace Corporation (acquired by Cyxtera), a company he founded, that revolutionized the use of machine learning in digital investigations. Brainspace was recognized as North Texas' Emerging Company of the Year in 2016 for its work in Augmented Intelligence. The technology radically improved the effectiveness of analysts and agents through their interaction with machine learning and AI. The Brainspace platform has quickly become the industry standard in intelligent analytics and has been adopted by organizations around the world, including all major international consulting firms, many agencies in the Federal Government, the DoD and other international intelligence agencies.

Prior to Brainspace, Dave founded Engenium Corporation (purchased later by Marsh McLennan) where he led the adoption of the company's semantic search platform. Dave's companies have placed machine learning and AI in thousands of organizations and agencies around the globe.

Dave received his BA from the University of North Texas in Industrial Anthropology / Corporate Culture. He is an invited member of the Aspen Institute's Roundtable on AI, a frequent speaker at MIT's EmTech conferences and other events centered around machine learning, AI and the future of technology. He is an active mentor and investor in startup companies and lectures at Universities and technology incubators all over the world including Tech Wildcatters and the Dallas Entrepreneurs Center (The DEC) in his home town of Dallas Texas.

When Dave is not being a geek, he enjoys collecting custom guitars, brewing craft beer, ocean sailing and family time at his house on the island of Bequia in the Caribbean.

**Fraud Detection in Customer Loyalty Programs**

**Presenter:** Subhashini Tripuraneni, Head of AI & DS, 7-Eleven
linkedin.com/in/subhashini-tripuraneni

Fraud in customer loyalty programs is starkly different from credit card fraud. The data is highly imbalanced in that the percentage of overall transactions that are fraudulent is very low. Additionally, the fraud patterns are constantly changing – fraudsters quickly derive rules of existing fraud prevention system and devise inventive workarounds. In this presentation, we will focus on an artificial neural network, such as autoencoders, to conduct anomaly detection. The rapidly evolving fraud patterns can be efficiently identified through the above unsupervised learning technique, saving businesses money.

**An Inductive Logic Programming Approach to Explainable Machine Learning**

**Presenter:** Farhad Shakerin, PhD candidate, UT Dallas, linkedin.com/in/farhadshakerin

Despite widespread adoption, machine learning models such as XGBoost and Deep Learning remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction. Recently, new techniques have been proposed to explain behavior of black box classifiers locally. LIME and SHAP are examples of such techniques. These techniques provide local explanation of a model's behavior for individual data samples. In this talk we first introduce some of these techniques, then we present our algorithm based on Inductive Logic Programming (ILP) to capture global behavior of machine learning models. Compared to state-of-the art rule extraction system, our approach results in rules that explain the logic underlying a model succinctly and accurately.

**A Robust, Interactive and Dynamical Computational Intelligence Tool for Patient Behavioral Health Status Tracking**

**Presenter:** Dr. Michael Morgan, Managing Director, Morgan Analytics Research Institute,
linkedin.com/in/michael-morgan-phd-b31881a

The Morgan Analytics Research Institute, a Texas-based nonprofit corporation, conducts research and development of new applications in computational neuroscience and deep machine learning. These tools target U.S. national healthcare objectives by providing innovative inroads to patient health status monitoring and management. Primary target markets include military veterans and individuals within disadvantaged communities.

Our most recent project is a computational intelligence tool for mental and behavioral healthcare patients for tracking their recovery status. This is an area full of challenges that are of interest to researchers, teachers, data scientists, data engineers and other professionals seeking personal involvement in developing data science, machine learning and networked computing tools for the healthcare sector.

Challenges include:

1. Utilizing the "slave" nodes in a virtual server network to conduct more thorough, efficient and accurate deep analytics prior to downloading results to the "master" node server.
2. Improving, through the computational intelligence platform, maintenance of contact and communication with patients, ultimately to help avoid psychotic breakdowns and relapses.
3. Detecting and proactively reducing the propensity for patients to lie about how they are feeling, their internal impulses and behaviorally-linked reward seeking that has been proscribed by their treatment program.
4. Development of application-specific latent state space methodologies to model and predict observed patient data at the master node with very large data sets.
5. Using better stochastic methods of sampling and model-fitting to improve on latency issues and enhancing predictive capabilities. Again, this is a major problem with big data.

**The Automation Mindset: Tips/Tools for Streamlining Data Science Workflows**

**Presenter:** Sydeaka Watson, Senior Data Scientist, AT&T Chief Data Office, Owner & Lead Data Scientist, Korelasi Data Insights, L.L.C. linkedin.com/in/sydeakawatson

A data scientist's workflow generally consists of a series of small tasks and mundane decisions that often consume a lot of time and energy. To begin a predictive modeling workflow, for example, one might log into a database, download the data to a local file system, and then move the data into a different system where analytic tools are installed. After loading the dataset and applying data transformations, the data scientist creates descriptive summaries and builds a predictive model. Those results might be stored and manually organized into an annotated report that summarizes the analysis. Finally, the report is emailed to team members and stakeholders and possibly published to the web. In practice, the data scientist might have to iteratively re-execute the entire workflow multiple times in order to incorporate stakeholder feedback or include other analyses that are added as the project's scope expands.

The promise of automation is a future in which data scientists will yield responsibility of these burdensome tasks to programmable robots that will do our work. I am happy to report that the future has arrived! A number of open source and proprietary solutions are currently available to create, organize, and execute automated data science workflows. It is also quite possible to unleash the power of workflows with a handful of easy-to-use R packages and Linux/Bash commands.

A data-driven company that leverages this framework will see insights produced, published, and distributed more rapidly within their organization. Data scientists that embrace the automation mindset are more productive, yet have more time and energy available to inject creativity and innovation into their work.

In this presentation, I will highlight R functions and system commands that could be used to engineer efficient data science workflows. A sample end-to-end solution similar to the complex workflow described above will be used to demonstrate these principles.

**The Power of Suggestion: An Introduction to Recommender Engines**

**Presenter:** Mimi Park, Data Scientist, Slalom Consulting, linkedin.com/in/mimi-park-1557b219

For companies such as Amazon, Netflix, and Spotify, recommender engines drive significant conversion and revenue. Recommender engines provide a scalable way of personalizing content for users in scenarios with many products and services.

Building a successful recommender engine is complex. However, incorporating the right mix of business strategy, data, and analytical knowledge will allow you to build an engine that helps you accomplish your identified goals.

Benefits:

- Understand industry use cases and the resulting impacts
- Learn how to recommend products to the right user at the right time and place
- Be aware of best practices for building an engine that works

Agenda: What is a recommender engine? Examples and use cases. Business value and benefits. Key methodologies. Best practices.

**How Azure Databricks and Pyspark helped to make IoT Analytics a reality**

**Presenter:** Prasad Chandravihar, Lead Data Scientist, Lennox International, linkedin.com/in/prasadmc

At Lennox International, we have thousands of IoT connected devices streaming data into the Azure platform with a minute level-polling interval. The challenge was to use these data sets, combine with external data sources such as weather, and predict equipment failure with high levels of accuracy along with their influencing patterns and parameters. Previously the team was using a combination of on-premise and desktop tools to run algorithms on a sample set of devices. The result was low accuracy levels (around 65%) on a process that took more than 6 hours. The team had to work through several data orchestration challenges and identify a machine learning platform which enabled them to collaborate between our engineering SME's, Data Engineers and Data Scientists. The team decided to use Azure Databricks to build the data engineering pipelines, appropriate machine learning models and extract predictions using PySpark. To enhance the sophistication of the learning, the team worked on a variety of Spark ML models such as Gradient Boosted

Trees and Random Forest. The team also implemented stacking, ensemble methods using H2O driverless AI and sparkling water on Azure Databricks clusters, which can scale up to 1000 cores.

**Big Data Stream Analytics for Cyber Security**

**Presenter:** Dr. Latifur Khan, Professor, UT Dallas, utdallas.edu/~lkhan

Data streams are continuous flows of data. Examples of data streams include network traffic, sensor data, call center records and so on. Data streams demonstrate several unique properties that together conform to the characteristics of big data (i.e., volume, velocity, variety and veracity) and add challenges to data stream mining. In this talk we will present an organized picture on how to handle various data mining techniques in data streams. In addition, we will present a number of stream classification applications such as adaptive website fingerprinting, textual stream analytics (political actor identification over textual stream), attack trace classification and secure data analytics using trusted execution environment (TEE).

This research was funded in part by NSF, NASA, Air Force Office of Scientific Research (AFOSR), NSA, IBM Research, HPE and Raytheon.

**LEARN MORE ABOUT THE UT DALLAS COMPUTER SCIENCE DEPARTMENT AT OUR WEBSITE**

CS.UTDALLAS.EDU