

# **2016 New York Scientific Data Summit (NYSDS 2016)**

**New York, New York, USA  
14-17 August 2016**



**IEEE Catalog Number: CFP16NYS-POD  
ISBN: 978-1-4673-9052-1**

**Copyright © 2016 by the Institute of Electrical and Electronics Engineers, Inc  
All Rights Reserved**

*Copyright and Reprint Permissions:* Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

***\*\*\*This publication is a representation of what appears in the IEEE Digital Libraries. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP16NYS-POD
ISBN (Print-On-Demand):	978-1-4673-9052-1
ISBN (Online):	978-1-4673-9051-4

**Additional Copies of This Publication Are Available From:**

Curran Associates, Inc  
57 Morehouse Lane  
Red Hook, NY 12571 USA  
Phone: (845) 758-0400  
Fax: (845) 758-2633  
E-mail: [curran@proceedings.com](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

CURRAN ASSOCIATES INC.  
**proceedings**  
.com

# 2016 New York Scientific Data Summit

## TABLE OF CONTENTS

### SESSION PAPERS

---

#### SESSION 1 STREAMING DATA ANALYSIS

##### **Stream Processing For Near Real-Time Scientific Data Analysis 1**

*Jong Youl Choi, Tahsin Kurc, Jeremy Logan, Matthew Wolf, Eric Suchyta, James Kress, David Pugmire, Norbert Podhorszki, Eun-Kyu Byun, Mark Ainsworth, Manish Parashar, and Scott A. Klasky*

##### **A Scalable Cyberinfrastructure for Interactive Visualization of Terascale Microscopy Data 9**

*A. Venkat, C. Christensen, A. Gyulassy, B. Summa, F. Federer, A. Angelucci, and V. Pascucci*

##### **Enabling Re-executable Workflows with Near-realtime Visualization, Provenance Capture and Advanced Querying for Mass Spectrometry Data 16**

*Mathew Thomas, Julia Laskin, Bibi Raju, Eric Stephan, Todd Elsethagen, Nhuy Van, and Son Nguyen*

##### **Auto-Tuning Intermediate Representations for In Situ Visualization 26**

*Steffen Frey and Thomas Ertl*

##### **A Framework to Visualize Temporal Behavioral Relationships in Streaming Multivariate Data 36**

*Shenghui Cheng, Klaus Mueller, and Wei Xu*

#### SESSION 2 LONG TERM DATA STORAGE, CURATION AND SHARING

##### **A General Purpose Tool-set for Representing Data Relationships: Converting Data into Knowledge 46**

*Joshua Stillerman, Thomas Fredian, Martin Greenwald, and John Wright*

##### **A Multiclass Classification Method Based on Deep Learning for Named Entity Recognition in Electronic Medical Records 52**

*Xishuang Dong, Lijun Qian, Yi Guan, Lei Huang, Qiubin Yu, and Jinfeng Yang*

##### **Data Storage and Sharing for the Long Tail of Science 62**

*Boyu Zhang, Line C. Pouchard, Preston M. Smith, Amandine Gasc, and Bryan C. Pijanowski*

### **SESSION 3 EXPERIMENTAL DATA ANALYSIS**

#### **Automatic Histopathology Image Analysis with CNNs 71**

*Le Hou, Kunal Singh, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, Roberta J. Seidman, and Joel H. Saltz*

#### **Deep Learning for Analysing Synchrotron Data Streams 77**

*Boyu Wang, Ziqiao Guan, Shun Yao, Hong Qin, Minh Hoai Nguyen, Kevin Yager, and Dantong Yu*

#### **Computer Vision Experiments in Crowdsourced Astronomy 82**

*Rasmi Elasmir*

#### **Software Tools for X-ray Photon Correlation and X-ray Speckle Visibility Spectroscopy 84**

*Sameera K. Abeykoon, Yugang Zhang, Eric D. Dill, Thomas A. Caswell, Daniel B. Allan, Arman Akilic, Lutz Wiegart, Stuart Wilkins, Annie Heroux, Kerstin K. van Dam, Mark Sutton, and Andrei Fluerasu*

### **SESSION 4 INDUSTRY SOLUTIONS AND CHALLENGES FOR BIG DATA**

#### **Streaming Data Analysis on the Wire 94**

*Dimitrios Katramatos, Meng Yue, Shinjae Yoo, Kerstin Kleese van Dam, Jin Xu, and Jiayao Zhang*

#### **Curating Identifiable Data for Sharing: the Databrary Project 101**

*Rick O. Gilmore, Karen E. Adolph, and David S. Millman*

### **SESSION 5 CONVERGENCE OF DATA AND HPC**

#### **Bringing the HPC Reconstruction Algorithms to Big Data Platforms 107**

*Nikolay Malitsky*

#### **Data Provenance Hybridization Supporting Extreme-Scale Scientific Workflow Applications 115**

*Todd Elsethagen, Eric Stephan, Bibi Raju, Malachi Schram, Matt MacDuff, Darrern Kerbyson, Kerstin Kleese van Dam, Alok Singh, and Ilkay Altintas*

## **POSTER SESSION PAPERS**

---

### **Analyzing Large Data Sets from XGC1 Magnetic Fusion Simulations using Apache Spark 125**

*R. Michael Churchill*

### **Parking Lot Delineation and Object Detection using a Localized Convolutional Neural Network 128**

*Daniel Cisek, Jediah Dale, Susan Pepper, Manoj Mahajan, and Shinjae Yoo*

### **Analysis of VSX1, HGF and LOX in the Pakistani Familial and Sporadic Keratoconus Patients 132**

*Pir Muhammad Siddique, Rida Khursheed Malik, Arsalan Anwar Chohan, Wajid Ali Khan, Sorath Siddiqui, Abdul Mannan Khan Minhas, Shaheena Bashir, Maleeha Maria, Sobia Shafique, Nadia Khalida Waheed, Maleeha Azam, and Raheel Qamar*

### **Visualization of the Higgs Interaction in the Standard Model and Beyond the Standard Model Physics 133**

*Han Soul Lee and Michael McGuigan*

### **Visualization and Simulation of the Material Properties of Carbon Nanotori 138**

*Yeeren Low and Michael McGuigan*

### **Analysis of Nanoparticle Growth in Environmental Transmission Electron Microscopy 142**

*Remi Megret, Shinjae Yoo, Dmitri Zakharov, and Eric Stach*

### **PUMA-V: An Interactive Visual Tool for Code Optimization and Parallelization Based on the Polyhedral Model 146**

*Harper Langston, E. Papenhausen, K. Mueller, B. Meister, and R. Lethin*

### **Model-Driven Visual Analytics for Big Data 150**

*Shenghui Cheng, Bing Wang, Wen Zhong, Cong Xie, Salman Mahmood, Jun Wang, and Klaus Mueller*

### **Computing Infrastructure, Software Optimization, and Real Time Analysis for High Data-Rate MX 152**

*Herbert J. Bernstein, Babak Andi, Kaden Badalian, Lonny E. Berman, Dileep K. Bhogadi, Shirish Chodankar, Jonathan DiFabio, Martin R. Fuchs, Jean Jakoncic, Edwin O. Lazo, Sean McSweeney, Lisa Miller, Stuart Myers, Dieter K. Schneider, Bruno Seiva Martins, Wuxian Shi, John Skinner, Hugo Slepicka, Alexei S. Soares, Vivian Stojanoff, Robert M. Sweet, and Ryan Tappero*