NIST National Institute of Standards and Technology • U.S. Department of Commerce

Data Science Symposium Proceedings

March 4-5

2014

Abstracts for posters presented at the 2014 NIST Data Science Symposium on March 4th 2014

Version 1.1

TABLE OF CONTENTS

A CONCEPTUAL FRAMEWORK FOR HEALTH DATA HARMONIZATION

LEWIS E. BERMAN, & YAIR G. RAJWAN,

ICF INTERNATIONAL & VISUAL SCIENCE INFORMATICS

To improve the speed of health science discovery the National Institutes of Health (NIH) is focusing effort on trans-disciplinary science and mega-epidemiology. This focus notionally addresses the need to increase scientific breakthroughs through the collaborative efforts of large geographically disparate multi-national teams and study respondents. Multi-site data collection and local study protocol modifications may necessitate data harmonization to produce high quality analytic datasets. Harmonization is often a complex and tedious operation, but it is an important antecedent to data analysis as it increases the sample size and analytic utility of the data. Typical harmonization efforts are ad hoc which can lead to poor data quality or delays in data release.

To date, we are not aware of any efforts to formalize data harmonization using a pipeline process and techniques to easily visualize and assess the data quality prior to and after harmonization. Therefore, we propose a lifecycle model for data harmonization to formalize the process and improve the quality of harmonized data. The lifecycle model consists of several steps. The harmonization portion includes four stages used to improve the quality of datasets and data elements considered for data analysis. The first two stages are an external view of the studies being considered. These two stages consider study design and structural factors. The third and fourth stages are an internal view of the data content and filter studies or data elements. Techniques include qualitative and quantitative measures. The lifecycle model also includes a step to assess disclosure risks prior to production of a public use and internal non-public use harmonized dataset.

Lewis E. Berman Dr. Berman has over 25 years of experience in informatics, imaging, software and systems development and architecture, and epidemiologic health studies and operations. His professional experience spans large public health organizations, private industry, and academia. Dr. Berman has held the position of Deputy Director with the Centers for Disease Control and Prevention's National Health and Nutrition Examination Survey (NHANES), where he was responsible for international, state, and local community health survey consultations. He also led the methodological research into dried blood spot collection and home health examinations. Dr. Berman was responsible for the design of the survey information technology architecture to support NHANES and Community HANES initiatives. Prior to working at CDC, Dr. Berman held positions with the National Library of Medicine and the Naval Research Laboratory where he was responsible for research and software and systems development in radar and medical imaging. He has a bachelor's degree in Computer Science from the University of Maryland and a master's degree and a doctorate in Computer Science with a minor in Public Health from George Washington University. He also completed a fellowship in e-Government at the Council for Excellence in Government.

Yair G. Rajwan Dr. Yair G. Rajwan is the director of Analytics Visualization at Visual Science Informatics, LLC. Dr. Rajwan provides 'big data' analytics capabilities and information visualization techniques enabling better engagement effectiveness, better decision-making collaboration, and better insights perspective on outcomes and impact. He use a variety of proven processes and easy-to-use tools helping understand your content competitive advantage, craft memorable visual storytelling, run continues improvement of visual communication outreach campaigns, and evaluate your visual intervention on outcomes. Dr. Rajwan serves on the Consumer Technology Workgroup (CTWG) of the Health IT Standards Committee (HITSC) at the Office of the National Coordinator (ONC) for Health Information Technology, at the US Department of Health and Human Services (HHS). He received his MSc (1995) and DSc (2000) in computer science from the George Washington University School of Engineering and Applied Science in Washington, DC. Dr. Rajwan was a postdoctoral research Fellow of the National Library of Medicine (2011) in the Division of Health

Sciences Informatics at the Johns Hopkins University School of Medicine. He also received his MSc (2012) in Health Sciences Informatics from the Johns Hopkins University School of Medicine in Baltimore, MD.

REAL-TIME ANALYTICS FOR DATA SCIENCE

HIROTAKA OGAWA

NATIONAL INSTITUTE OF ADVANCE INDUSTRIAL SCIENCE AND TECHNOLOGY, JAPAN

The amount of datasets in the world is exploding, and today we face new challenges in analyzing extremely largescale datasets, so-called Big Data. In the era of Big Data, data analytics platforms should meet three requirements: handle bigger data, perform deeper analytics, and process in real-time. However, existing platforms including MapReduce and CEP, have only partly met these requirements. As a complementary solution, we are now developing Real-time Analytics Platform, which can process massive data streams of continuously generated Big Data. It can support deep and real-time analytics based on online machine learning, with distributed and scale-out architectures.

In the era of Big Data, we face new challenges in real Big Data applications, such as analyzing nation-wide M2M sensor networks for the automotive field, real-time healthcare monitoring for millions of prospective diabetes patients, and online advertising optimization for millions of customers. For such applications, data analytics platforms should meet three requirements at one time: handle bigger data, perform deeper analytics, and process in real-time. However, existing approaches are not always applicable or feasible for such applications. For example, Data analytics on datasets performed after storing all data into a single database restricts the size of the datasets. Analyzing the data on distributed data stores in a batch-processing manner, e.g., with MapReduce, causes higher latency and longer turn-around time. And, real-time processing on streams of data, e.g., with Complex Event Processing (CEP), is not suitable to complex analysis and calculation.

Hirotaka Ogawa received his Ph.D. in Computer Science from Tokyo Institute of Technology in 2009. He is currently a senior researcher of Information Technology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Japan. His research interests include online machine learning, big data and cloud platforms.

UTILIZATION OF A VISUAL ANALYTICAL APPROACH TO DETECT ANOMALIES IN LARGE NETWORK TRAFFIC DATA

LASSINE CHERIF, SOO-YEON JI, DONG HYUN JEONG

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY, UNIV. OF THE DISTRICT OF COLUMBIA AND DEPT. OF COMPUTER SCIENCE, BOWIE STATE UNIVERSITY

In this paper, we develop a method for detecting anomalies in large network traffic data through interactive visual analysis. While the amount of network traffic data is growing exponentially, our ability to analyze complex network traffic data is limited. To overcome this limitation, researchers have started to utilize interactive visual analytics

systems, which allow them to explore and examine large amounts of data quickly and efficiently. In this work, we performed two research challenges: (1) analyzing the large amounts of network traffic data efficiently, and (2) visually analyzing the data effectively. Due to the size of the network traffic data, traditional approaches of using desktop computing and statistics do not work successfully. Instead, a massive parallel processing approach (i.e. high-performance computing) is generally considered to be an ideal solution to be adopted for handling big data.

In this study, publicly available internet traffic data [1] is used. The data covers the collection comprised of 10 datasets containing about twenty to seventy thousands of objects (i.e. instances). Each dataset includes 248 attributes. With this data, we execute the two research challenges in a parallel yet integrated manner. To address the first challenge of analyzing large network traffic data, a solution for extracting hidden, but important features from the data is studied. This is a generalization process of extracting significant features. We extend this computationally expensive feature extraction to an integrated approach of leveraging the power of parallel processing and machine learning. To accomplish the second challenge, an interactive visual analytics approach is applied. Specifically, a visual analytics system – iPCA (interactive Principal Component Analysis [2]) – is used to perform interactive and intelligent visual analysis. In the system, multiple coordinated views are utilized, such that each view can store an independent representation of the data to support users' unique hypotheses and needs. With this system, the user will be able to analyze large scale data sources uninterruptedly in real-time. In addition, it improves our ability to extract insights from large scale network traffic data.

For future work, we will continue to identify a best feature extraction solution and to extend our visual analysis approach to support seamless visual analysis on the network traffic data through decent cloud computing technology.

ACKNOWLEGEMENT: This study is supported by the Department of Defense under Grant No. 62702-CS-REP

REFERENCES

[1]Moore, A., Zuev, D., Crogan, M., Discriminators for use in flow-based classification, Technical Report RR-05-13, Department of Computer Science, Queen Mary, University of London, August, 2005.
[2]Jeong, D.H., Ziemkiewicz, C., Fisher, B., Ribarsky, W., Chang, R., iPCA: An Interactive System for PCA-based Visual Analytics, Computer Graphics Forum (Eurovis 2009). 28(3), pp.767- 774, 2009.

Lassine Cherif is a student in the Department of Computer Science at the University of the District of Columbia. . His research areas include Data Visualization, Big Data, Anomaly Detection, Cloud Computing, and Cloud Computing Security. He is member of ACM and Cloud Security Alliance. Contact him at lassine.cherif@udc.edu.

Soo-Yeon Ji is an Assistant Professor in the Department of Computer Science at Bowie State University. Her research areas include Bioinformatics, Pattern Recognition, Machine Learning, Statistical Analysis, Artificial Intelligence, and Network Security. She received PhD from the Computer Engineering Department at Virginia Commonwealth University. She is a member of IEEE. Contact her at sji@ bowiestate.edu.

Dong Hyun Jeong is an Assistant Professor in the Department of Computer Science at the University of the District of Columbia. He is a deputy director of NSF-sponsored ARCTIC (Assurance Research Center for Trusted Information Computing). He is a member of ACM, IEEE, and IEEE computer society. His research areas include Information Visualization, Visual Analytics, and HCI (Human Computer Interaction). He received PhD from the Computer Science Department at the University of the North Carolina at Charlotte. Contact him at djeong@udc.edu.

RE-PRESENTING DATA: END-TO-END ARCHITECTURES FOR DATA SCIENCE

MALLIKARJUN SHANKAR

OAK RIDGE NATIONAL LABORATORY

Data science-based systems must be built upon foundational architectural principles that accommodate the end-to-end requirements of (i) large-scale data collection, storage, and processing, (ii) flexible analytics and interpretation of the collected data, and (iii) continual structural evolution of the data in the form of additions and modifications in the associated meta-data. One way to arrive at foundational architectural tenets for data science is to describe normative architectures for the data "stack". Such a stack must outline the generic categories of the data science activity, thus facilitating better design of systems and enabling assessment and optimization through measurement and refinement. We propose a set of components that comprise such a data stack and outline the stack's role in better constructing end-to-end analytic workflows, the internal interface points of which would suggest sites for measurement.

The surge of activity in the area of systems concerned with "big data" owes its origins to the proliferation of sensor data and its associated communications networks, storage systems improvements, and the improved ability to network and harness workstations. These drivers of data-growth and the advances in data processing ability have brought about the remarkable engineering innovations emerging from industry leaders like Google and Amazon, and from a variety of software start-ups. Although these modern systems enable us to process large data volumes cheaply, they have given rise to increased competition between what may appear to be new data processing paradigms and the established (and foundationally vetted) approaches of relational databases. In addition, the architectural place for emerging technologies that specialize in rudimentary structured data (such as key-value pairs or graphs) or unstructured data (e.g., plain text and blob-stored imagery) is still not well delineated in an end-to-end data systems framework.

We discuss the structural components in a data stack that correspond to separate and orthogonal technology areas of emphasis within data science. We argue that a general framework built upon such a data stack underlies a vast majority of applications. Each of these components may be associated with a separate design focus (e.g., one such component area is data representation and semantics and its concerns about capturing vocabularies and ontologies). By understanding how a system may be constructed with due attention to the data stack, system designers may be able to better instrument the system for measurement and improvement.

DATA INTENSIVE WORKFLOWS ON THE OPEN SCIENCE DATA CLOUD

ALLISON HEATH, MARIA PATTERSON, MATTHEW GREENWAY, RAYMOND POWELL, RENUKA ARYA, JONATHAN SPRING, RAFAEL SUAREZ, DAVID HANLEY, ROBERT GROSSMAN

UNIVERSITY OF CHICAGO

The Open Science Data Cloud or OSDC (www.opensciencedatacloud.org) is a petabyte-scale science cloud managed and operated by the not-for-profit Open Cloud Consortium (OCC) that has been in operation for four

years. It contains over a petabyte of scientific datasets across a variety of scientific disciplines, including the biological sciences, physical sciences and the social sciences.

The OSDC allows scientists to manage, analyze, integrate, share and archive their datasets, even if they are large. It provides a home for reference datasets and challenge problems across a variety of disciplines. Datasets, including reference datasets, can be downloaded from the OSDC by anyone. Small amounts of computing infrastructure are available without cost so that any researcher can compute over the data managed by the OSDC. Larger amounts of computing resources are also made available to research projects through a selection process so that interested projects can use the OSDC to manage and analyze their data. In addition, larger amounts of computing infrastructure are made available to researchers at cost.

For many projects, scientists utilize the OSDC as they would other cloud computing infrastructures, with the advantage of having easy access to a number of large data sets, high performance storage and high performance networks. For a select set of applications, we are developing a framework that automates running analyses over large datasets, called the OSDC Wheel. The name derives from the idea that it can run continuously. In this way, new algorithms and workflows can be improved and benchmarked and existing algorithms and workflows can be compared against a common reference dataset.

In this poster, we describe the OSDC, our experience running the OSDC Wheel, and our experience running large-scale workflow on datasets that in range in size from terabytes to hundreds of terabytes for different scientific communities.

UTILIZING BIG SOCIAL MEDIA DATA FOR HUMANITARIAN ASSISTANCE AND DISASTER RELIEF

FRED MORSTATTER, SHAMANTH KUMAR, HUAN LIU

ARIZONA STATE UNIVERSITY

Disaster response agencies have started to incorporate social media as a source of fast-breaking information to understand the needs of people affected by the many crises that occur around the world. Twitter, one example of social media, produces over 500 million status updates each day. The volume of this data is too much for first responders to consume through manual analysis. New systems and approaches are needed to help first responders obtain situational awareness during a disaster using social media data. We discuss two social media tools that we have developed to assist first responders with the challenge of understanding this deluge of big social media data.

TweetTracker is a tweet collection and aggregation system that addresses the problem of collecting big social media data for first responders in disaster scenarios. Using TweetTracker, first responders enter queries related to a disaster as the event unfolds. From here, TweetTracker addresses the ETL process of our big data workflow. Once these queries are entered into the system, TweetTracker extracts matching tweets from Twitter using Twitter's APIs. To answer queries in real-time, TweetTracker performs several transformations on each tweet, including: keyword extraction, user profile location translation, and several optimizations for indexing. Finally, TweetTracker loads the data into a NoSQL database that is then queried from TweetTracker and TweetXplorer for real-time analysis.

TweetXplorer is a visualization system that addresses the challenge of understanding the big data generated during crisis on social media. It helps first responders to understand the data via some visual analytic components. TweetXplorer focuses on emphasizing some key facets of disaster data to first responders, including: when relevant keywords are important, who are the most influential tweeters, and where are the geographic regions with the most requests for help.

These systems have helped first responders to find areas of need in recent crises including Typhoon Haiyan to generate an after-action report of the areas of need. They were also used during Hurricane Sandy to assess how social media could best be used to deliver aid. Our systems are the first of their kind to aid first responders in making sense of the fast, noisy, and big data generated on social media.

Fred Morstatter is a PhD student in computer science at Arizona State University in Tempe, Arizona. He is a research assistant in the Data Mining and Machine Learning (DMML) laboratory. Fred won the Dean's Fellowship for outstanding leadership and scholarship during his time at ASU. He is the Principal Architect for TweetXplorer, an advanced visual analytic system for Twitter data. He has also worked on TweetTracker. He has published in ICWSM, WWW, KDD, IEEE Intelligent Systems. He has also published a book: Twitter Data Analytics. Contact him at fred.morstatter@asu.edu. A full list of publications and updated information can be found at http://www.public.asu.edu/~fmorstat.

Shamanth Kumar is a Ph.D. Student in Computer Science at Arizona State University, and is interested in social media mining, online user behavior analysis, and information visualization. He obtained his B.E in Information Science and Engineering from Visveswaraiah Technological University, India. He works on social media based data analysis tools targeted towards information gathering and information analysis during Humanitarian Assistance/Disaster Relief events. He is the Chief Architect of TweetTracker (http://tweettracker.fulton.asu.edu/), which is a Twitter data aggregation and analysis system. He also works on TweetXplorer, which is an advanced visual analytics system for Twitter data. He has published research papers in several peer-reviewed conferences and workshops. He has also served as an external reviewer for various conferences and journals. He has served as a Program Committee member at SBP 2013, IJCAI 2013, and SBP 2014. A full list of his publications and updated information can be found at http://www.public.asu.edu/~skumar34/

Dr. Huan Liu is a professor of Computer Science and Engineering at Arizona State University. He obtained his Ph.D. in Computer Science at University of Southern California and B.Eng. in Computer Science and Electrical Engineering at Shanghai JiaoTong University. Before he joined ASU, he worked at Telecom Australia Research Labs and was on the faculty at National University of Singapore. He was recognized for excellence in teaching and research in Computer Science and Engineering at Arizona State University. His research interests are in data mining, machine learning, social computing, and artificial intelligence, investigating problems that arise in many real-world, data-intensive applications with high-dimensional data of disparate forms such as social media. His well-cited publications include books, book chapters, encyclopedia entries as well as conference and journal papers. He serves on journal editorial boards and numerous conference program committees, and is a founding organizer of the International Conference Series on Social Computing, Behavioral-Cultural Modeling, and Prediction (http://sbp.asu.edu/). He is an IEEE Fellow and an ACM Distinguished Scientist. Updated information can be found at http://www.public.asu.edu/~huanliu.

LARGE SCALE AVIATION DATA ANALYSIS

ADRIC ECKSTEIN, CHRIS KURCZ

MITRE

The MITRE Center for Advanced Aviation Systems Development (CAASD), a Federally Funded Research and Development Center (FFRDC), has been working on behalf of the Federal Aviation Administration (FAA) to perform analysis in safety, security and efficiency in the National Airspace (NAS) using data driven methods. These activities include identifying systematic safety risks, measuring the effectiveness of safety mitigation strategies, assessing benefits of existing and proposed route and/or airspace designs, and determining efficiency generation and propagation. We describe how CAASD has embraced data intensive computing technology to make progress in these areas and we will discuss challenges and strategies for evolving to changing data, data processing workflow, analytics and visualizations of complex systems.

Analysis that CAASD performs often requires a large historical archive of flight data and associated contextual information (e.g. end-to-end trajectory, air traffic control actions, weather, nearby aircraft, airspace restrictions, and aircraft capability). However, the control systems that govern the NAS are designed for real time operations and, thus, significant data capture and processing efforts are made to reconstruct a flight's history. To reconstruct this history, the analytic workflow begins with capture of surveillance data from Air Traffic Control (ATC) facilities. Due to the lack of a unique flight identifier, data from these facilities are associated and synthesized into an end to end flight by means of extensive use of map/reduce processing via hadoop. Once generated, many derivation and inference functions are applied to provide additional contextual information (e.g. high energy approaches, runway occupancy time, routes used, winds aloft, airport weather, fuel consumed, and runway used). We discuss the implementation of a directed acyclic processing graph that manages the workflow of map/reduce jobs which generates this data and describe its implications to provenance. These derivation functions require inputs from previously generated data and, therefore, propagate uncertainties and the underlying data must be consistent with changes to processing algorithms. In particular, we illustrate the complexities of propagating error through a fuel consumption model which requires inputs from derived data where potential issues include missing data, input measurement errors and fidelity, model error and its effect on cumulative fuel mass. In addition we illustrate the analyst workflow, describe the importance of data immutability for our purposes, and discuss challenges in data management, sharing, quality and availability.

KNOWLEDGE EXPANSION USING INFERENCE OVER LARGE-SCALE UNCERTAIN KNOWLEDGE BASES

DAISY ZHE WANG, YANG CHEN

UNIVERSITY OF FLORIDA, CISE

Google Knowledge Graph (KG) can improve search engines by understanding the concepts in documents and in queries to provide answers beyond keywords and strings. Current Google KG is a very sparse graph with large amounts of missing links. We propose to expand the knowledge graph by interpolating missing links using two

methods: First, design a probabilistic knowledge base that can incorporate uncertainty data sources in addition to the high-confidence data sources. Second, develop a scalable statistical inference engine that can probabilistically deduce missing links based on an existing KB and a set of first-order rules.

Project Goal Knowledge Graph (KG) is Google's attempt to improve search engines by understanding the concepts (e.g., entity and relations) in documents and in queries to provide answers beyond keywords and strings. As far as we know, the current Google KG contains 580 million objects and 18 billion facts about relations between them. While this is the largest knowledge graph constructed, it is also a very sparse graph: on average, only ~30 relations for one entity. We believe large amount of relations are missing because (1) only high-confidence data sources are used to constructed Google KG; (2) some of the relations are never recorded explicitly in any of the data sources.

The goal of our research is to expand a KG with uncertain facts such as those automatically extracted using openIE [openIE] and NELL[NELL] technologies, and to interpolate the missing links between entities using first-order soft and hard rules. To accommodate uncertainty in a KG, we design a probabilistic knowledge base (KB) that can represent the uncertainty and correlations between entities and relations as first-class citizens. To interpolate in the missing links, we develop a scalable inference engine that can deduce additional relations based on an existing KB and a set of first-order rules.

Datasets We use publically available KBs constructed from the Web from projects such as Reverb and NELL [ReVerb, NELL]. For link interpolation, we use Sherlock-Holmes dataset, which is a set of first-order rules extracted from the output of an openIE system [openIE]. It is worth noting that Reverb, Sherlock-Holmes and NELL all generate probabilities associated with each extracted entity, relation, or rule.

Outcome Three outcomes from this project are: (1) a probabilistic knowledge base that can store and represent an uncertain KB with facts, rules and relations; (2) a scalable inference engine that can perform grounding and inference over a huge probabilistic graphical model with hundreds of millions of nodes; (3) a set of new relations inferred, which can be evaluated using crowd sourcing services based our previous work [CASTLE2013].

References

[OpenIE] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the web. Commun. ACM, 51(12):68–74, 2008.

[ReVerb] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1535–1545. Association for Computational Linguistics, 2011.

[NELL] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an architecture for never-ending language learning. In Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010), volume 2, pages 3–3, 2010.

[CASTLE2013] Sean Goldberg, Daisy Zhe Wang, Tim Kraska. CASTLE: Crowd-Assisted System for Textual Labeling and Extraction. In Proceedings of HCOMP 2013.

Daisy Zhe Wang is an Assistant Professor in the CISE department at the University of Florida. She is also the director of the UF Data Science Research Lab. She obtained her Ph.D. degree from the EECS Department at the University of California, Berkeley in 2011 and her Bachelor's degree from the ECE Department at the University of Toronto in 2005. At Berkeley, she was a member of the Database Group and the AMP/RAD Lab. She is particularly interested in bridging scalable data management and processing systems with probabilistic models and statistical methods. She currently pursues research topics such as probabilistic databases, probabilistic knowledge bases, large-scale inference

engines, query-driven interactive machine learning, and crowd assisted machine learning. Her research is currently funded by DARPA DEFT, Greenplum/EMC, Google, Pivotal, Survey Monkey and Law School at UF.

Yang Chen is a third-year Ph.D. student at CISE, University of Florida. His research interest lies in large-scale machine learning, probabilistic graphical models and database systems. Yang is currently working on the ProbKB project led by Dr. Daisy Zhe Wang. Yang received his Bachelor's degree in Computer Science from University of Science and Technology of China (2007-2011).

MOLYTICS: MOBILE ANALYTICS TO DEAL WITH INTERNET OF THINGS SOURCED BIG DATA

ARKADY ZASLAVSKY, PREM JAYARAMAN, DIMITRIOS GEORGAKOPOULOS

CSIRO, AUSTRALIA

Today's smartphones equipped with a range of sensors and new emerging wearable mobile devices that complement smartphone sensing, ranging from unobtrusive sensors to gauge human emotional states to Google Glasses provide massive sources of valuable data for innovative and novel applications. These devices form the massively distributed Internet of Things infrastructure and will account for large segment of overall Internet traffic. This data from multitude of sensory devices can be explored, fused and processed to derive situational context information (e.g., via image processing and feature extraction) to present a coherent picture of the situation. Not all the data generated by the devices could be offloaded to back-end servers for processing (e.g. cloud computing) due to the cost involved in data transmission and computation. Further, the availability of abundant real-time sensory data leads to a typical data explosion problem and consequently to a big data problem.

Mobile participatory sensing offers the potential to deliver innovative and novel applications that could benefit a community of users by taking advantage of data obtained from the users, for example, continuously modelling the user's situation using sensor inputs from mobile phone and surrounding sensors providing the user with advice and recommendations. This can be applied in an application like monitoring workers in a large warehouse where individual worker can be monitored and a coordinated cloud system can process the data from a collection of workers. This data can help in optimizing and planning warehouse activities resulting in better facility management.

Currently, the participatory sensing frameworks focus only on collecting data from the mobile devices with not much attention given to on-board processing and analysis. In situations like the warehouse where the worker will have to perform his/her duties independently under uncertain circumstances, it is critical for a system to have local intelligence to arrive at efficient and effective decisions quickly. To accomplish this goal, the system needs to be equipped with intelligence, situation awareness and reasoning capabilities to support key decisions. Also, in most instances, it might not be accurate to assume that an individual node will have complete knowledge of all the other nodes. In such scenarios, a load balanced approach should work best where locally processed data is passed to the cloud for further processing, analysis, correlation and planning. At the same time, peer-based load balancing may also prove efficient.

Hence, this paper focuses on a context-aware mobile participatory sensing (opportunistic sensing) and analytics framework that follows a distributed participatory sensing approach with in-built load balancing capabilities to manage complex processing requirements and deliver results of processing to users/machines in the appropriate

manner. Therefore, data generated by the sensors on-board the smartphones are valuable and have the potential to drive many future applications in agriculture, manufacturing, intelligent transportation systems, logistics, health, environmental science, social networks, just to name a few. Hence it is critically important to address some of the key challenges around the area of mobile participatory sensing and on-board analytics.

We envision the development of a distributed system of mobile processing nodes namely mobile devices that can work in a coordinated fashion when required to jointly solve the user request/query and/or offer a service. Some of the key challenges that are being addressed include:

- Context and situation aware reasoning for mobile participatory sensing applications;
- Development of novel applications driven by data stream mining approaches that energy-efficiently monitors user activities e.g. continual monitoring of warehouse personnel for better resource optimization and allocation, monitoring of road potholes using participating driver community;
- Load balancing approaches enabling efficient processing on-board the mobile devices and at external resource, e.g. cloud servers;
- Use of powerful cloud computing services to fuse data from multiple sources including multiple user communities, social media etc to present a coherent picture of the situation and status of the system;
- Extending the scalable and highly successful data stream processing and data management engine SensorDB towards mobile nodes.

Dr **Arkady Zaslavsky** is a Senior Principal Research Scientst at CSIRO Comutational Informatics, leading research projects in Internet of Things and sensing middleware. He is also an Adjunct Professor at Australian National University. He has published more than 350 peer-reviewed papers in journals and conferences. He is a Senior Member of ACM, a member of IEEE Computer and Communication Societies.

Dr **Prem Prakash Jayaraman** is a Post-doctoral Fellow at CSIRO Computational Informatics and is involved in projects on Internet of Things, opportunistic sensing and mobile analytics on smartphones. Previously, he was a Research Fellow at the Centre for Distributed Systems and Software Engineering (DSSE), Monash University. Dr. Jayaraman received his PhD (2011) in Computer Science and Software Engineering from Monash University.

Dr **Dimitrios Georgakopoulos** is a Research Group Leader and is leading and managing projects in cyber-social computing area. His career spans mnany leading ICT companies, including Telcordia, GTE, MCC and others. He has extensively published in data management and service-oriented computing areas.

GETTING THE SCIENCE INTO DATA SCIENCE

NANCY GRADY

SAIC

Data Science is an empirical science, with a computational data system serving the same role as any experimental equipment: to measure something happening in the world in such a way as to learn something valuable from it. As such you need to follow a scientific methodology to ensure the end result you're providing is valid and has value. There has been a lot of discussion on making sure analytics are not influenced by poor data quality, but the same must be considered for the end-to-end process.

In table-top science you would put together your experiment to carefully measure a specific phenomenon. You needed to understand what measurement would prove your hypothesis, and make sure your equipment would provide a proper measurement. In the data arena this is equivalent to the statistician's computer design of experiments. This approach changed when data mining began to take hold. In most cases the data had already been created and was being re-purposed for a different analysis. The rigor of the traditional statistician gave way to an imprecise but deterministic modeling that still had to assume certain data distributions. Data Mining has now been expanded into? Data Science?, which has changed the landscape in two ways. First, sampling not as significant a concern as it is beginning to be understood that in some cases more data beats better algorithms and bad data will be insignificant in light of the quantity of data. Second, it is often not necessary to provide a deterministic answer; trending without knowing the precise magnitude of the trend is often sufficient. This is often expressed as the need for correlation but not causation.

The data mining community came together to introduce the CRISP-DM (The CRoss-Industry Standard Process for Data Mining) begun in 1996 by DaimlerChrysler, SPSS and NCR. While we again are hearing the concept that you just "let the data speak for itself", this is no more true now than it ever has been. While people (Data Scientists or teams) and (big data) technology have garnered most of the attention, the third leg in the stool that has been overlooked is process. We need to begin with process models like CRISP-DM (and there are others) to determine the best practices to provide an efficient focus for efforts, with a methodology that will give repeatable results.

This presentation will expand the CRISP-DM methodology to be more inclusive of new Data Science needs, and give real-world examples of some of the steps.

Data Science is a science. Rigor and understanding is needed at each step. Reviewing lessons learned the hard way from a data science project provides context for suggesting some general best practices, and ways to expand the data mining process to encompass data science.

Nancy Grady is an SAIC Technical Fellow with 25 years of Data Science experience, specializing in data and text mining, and computer modeling and simulation. She currently leads the Big Data R&D for SAIC providing rapid analytics across data silos. She is a former Wigner Fellow at Oak Ridge National Laboratory, is on program committees for KDD, IEEE big data, the ISO/IEC JTC 1 Study Group on Big Data; and leads the NIST Big Data subgroup for Definitions and Taxonomy.

LARGE-SCALE INFERENCE AND SCALABLE STATISTICAL METHODOLOGY FOR COMPLEX (BIG) DATA

ALI ARAB

DEPARTMENT OF MATHEMATICS AND STATISTICS, GEORGETOWN UNIVERSITY

The increasing interest in data-driven approaches to scientific problems requires data scientists, and in particular statisticians, to develop or adopt methodology that is appropriate for complex, high-dimensional, and potentially unstructured data. The existing statistical methodology for complex data (e.g., data with spatial/temporal dependence structure) is often not scalable for massive data sets. The scalability issues of these methods are rooted in both theoretical and computational issues. In this paper, I discuss large-scale statistical inference which is designed for high-dimensional data as well as statistical methodology that maintains computational scalability. In

particular, I discuss efficiently parameterized models (based on prior scientific knowledge or reduced dimension mathematical representations) for detecting anomalies in data with temporal, spatial, or spatio-temporal structure. I will illustrate the proposed methodology for applications using satellite imagery for environmental monitoring and identifying human rights violations. Finally, I will address potential extensions of the proposed statistical methods for Big Data motivated by distributed and scalable database frameworks (e.g.Hadoop).

NATIONAL DATA SCIENCE LABORATORY: AN EXPERIMENTAL BENCHMARKING INFRASTRUCTURE

CHAITAN BARU, HOWARD LANDER, ARCOT RAJASEKAR, JUSTIN ZHAN

New models of computation, highly distributed architectures and the emergence of data-intensive analyses provide a unique opportunity for advancing science. To achieve the full potential of data-driven science and research, adequate mechanisms for testing, comparing, and studying data oriented architectures and software systems are necessary. What is needed is a community infrastructure – a National Data Science Laboratory (NDSL) - for testing and benchmarking that would be leveraged in the continued development of large-scale data-oriented system architectures. In this abstract we define several salient features of such a laboratory; the complete specification will be created through a community driven process designed to maximize the usefulness of such a national facility.

The architecture of an NDSL should be distributed, flexible, adaptive, reservable and modular so that researchers can use it to build experimental end-to-end systems and conduct performance analysis and comparative studies. In our design, we identify seven main NDSL components (we discuss two below). These components form a layered architecture and provide modularity for plug-and-play integrated testing. We chose these component abstractions from our experience in dealing with large-scale distributed systems and with data-intensive computation projects and also to motivate the type of people who will be involved in the community design process.

Information Stores: The emergence of NoSQL databases has been driven by the insight that not all types of massive data require the rigor of the ACID properties that are a key advantage of relational database frameworks. A cursory glance at the Wikipedia entry on NoSQL (Wikipedia) lists over sixty system implementations – many with similar architectural models. A recent outcome of the NoSQL movement has been renewed interest in providing consistency and integrity without compromising availability and partitionability; this interest has given rise to the emergence of NewSQL systems. Studying the contrasts in Relational, NoSQL and NewSQL systems would be of interest to Big Data researchers and may set the stage for convergence of these systems into a common framework or may provide distinct but inter-operable frameworks that can take advantage of the promises of each system. A community facility such as NDSL would be a catalyst for performing comparisons across these systems and for characterizing their capabilities and applicability in domain-specific problem space.

Data Analytics Support: The model in scientific computing (and also other types of computing in business, finance, legal, medical, etc.) is changing from a computing-driven model to a data-driven model. Increasingly computations are using very large data sets, performing data fusion across diverse data, and utilizing an analyze/synthesize paradigm implemented by sharding jobs across a large number of processes working in isolation. Moreover, service oriented paradigms have also shifted computational models away from monolithic operations connected through MPI. Concepts such as scheduling, reservations, allocations, and accounting, which have been used in main-stream high performance computing are giving way to brokering through dynamic service level agreements

(sla), orchestration of distributed services through message bus, and allocated on-demand elastic computing clouds. These emerging models need to be studied for their effectiveness, and for defining classes of problems for which they are applicable. A central facility such as the NDSL will provide a platform that can be used for setting up a virtual networked environment that can be instrumented for tuning and optimization. Students and researchers will use the system to try novel ideas in data intensive computation without a major fiscal outlay in infrastructure building.

Other functionalities in the NDSL laboratory include: benchmarking for data mining and statistical analysis operations, effectiveness of data sharing capabilities, tenability of data communication protocols, availability of a diversity of reference data collections meeting the 5V characteristic of Big Data, tools and problems for Big Data Benchmarking.

Chaitan Baru is Distinguished Scientist and Associate Director Data Initiatives at the San Diego Supercomputer Center, University of California San Diego, where he also directs the Center for Large-scale Data Systems Research(CLDS). He initiated the Workshops on Big Data Benchmarking, WBDB (http://clds.sdsc.edu/bdbc/workshops) in May 2012 in San Jose, CA. The fifth WBDB will be held in August 2014 in Potsdam, Germany. Baru also co-Chairs the NIST Big Data Public Working Group, with Wo Chang, NIST and Bob Marcus, ET-Strategies (http://bigdatawg.nist.gov).

Howard Lander has worked extensively at the Renaissance Computing Institute with the application of Cyber Science and Engineering to data intensive scientific disciplines including physical oceanography and evolutionary biology. He is a Co-PI on the DataBridge project, an NSF funded research effort investigating the use of sociometric techniques to produce a Facebook like social network for data sets from the long tail of science. Howard currently co-chairs the RENCI Data Working Group, an institute wide cross-cutting working group involved in many aspects of Data Science. He has extensive experience with HPC technologies.

Arcot Rajasekar is a Professor in the School of Library and Information Sciences at the University of North Carolina at Chapel Hill, a Chief Scientist at the Renaissance Computing Institute (RENCI) and co-Director of Data Intensive Cyber Environments (DICE) Center at the University of North Carolina at Chapel Hill. He has been involved in research and development of data grid middleware systems for over a decade and is a lead originator behind the concepts in the Storage Resource Broker (SRB) and the integrated Rule Oriented Data Systems (iRODS), two premier data grid middleware developed by the Data Intensive Cyber Environments Group. A leading proponent of policy-oriented large-scale data management, Rajasekar has several research projects funded by the National Science Foundation, the National Archives, National Institute of Health and other federal agencies.

Justin Zhan is the Director of ILAB Interdisciplinary Research Institute and an Associate Professor of Computer Science at North Carolina A&T State University. His research interests include Big Data, Information Assurance, Social Computing, and Health Science. He is currently an editor-in-chief of International Journal of Privacy, Security and Integrity, International Journal of Social Computing and Cyber-Physical Systems.. He has served as a conference general chair, a program chair, a publicity chair, a workshop chair, or a program committee member for 160 international conferences and an editor-in-chief, an editor, an associate editor, a guest editor, an editorial advisory board member, or an editorial board member for 30 journals. He has published 150 articles in peerreviewed journals and conferences and delivered above 30 keynote speeches and invited talks. He has been involved in a number of projects as a PI or a Co-PI, funded by the National Science Foundation, Department of Defense, National Institute of Health, etc.

WHAT A DATA SCIENTIST DOES AND HOW THEY DO IT

BRAND L. NIEMANN

SEMANTIC COMMUNITY

I am going to tell and show What a Data Scientist Does and How They Do It with the following topics:

- The State of Federal Data Science
- Data Science Team Examples
- Data Science, Data Products, and Data Stories Venn Diagrams
- NAS Report on Frontiers in Massive Data Analysis
- Graph Databases and the Semantic Web
- Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG
- Federal Big Data Working Group Meetup

I am both a data scientist and a data journalist. I engage in a wide range of data science activities, organize and participate in Data Science Teams and Meetups, and use data science to produce data products and stories like this one for the NIST Data Science Symposium. I will also show highlights of four recent Blog Posts to Data Science DC and Semantic Community.

The Semantic Medline – YarcData Graph Appliance Application for the Federal Big Data Senior Steering topics includes:

- Graphs and Traditional Technologies and The YarcData Approach
- Semantic Medline at NIH-NLM and Bioinformatics Publication
- Semantic Medline Database Application and Work Flow
- Predication Structure for Text Extraction of Triples
- Visualization and Linking to Original Text
- New Use Cases and Making the Most of Big Data

Two YouTube Videos will be shown (Schizo-7 minutes, and Cancer -21 minutes).

Finally some highlights of the new Federal Big Data Working Group Meetups will be shown. All are welcome to participate in this new Meetup and their upcoming workshop in June.

The Fourth Meetup will be held tonight (March 4th) at The National Science Foundation, 4201 Wilson Boulevard, Arlington, Virginia 22230, from 6:30-9:30 p.m. NIH Welcome and Introduction to Biomedical Big Data Research: NIH Program Director by Dr. Peter Lyster. Brief Demo of NIH Semantic Medline/YarcData by Tom Rindflesch and Aaron Bossett. NSF Welcome and Introduction by NITRD Program Office Director, Dr. George Strawn. Presentation on BRAIN by Dr. Barend Mons. Discussion of A Data Fairport Workshop Summary. Open Discussion. Networking. RSVP: bniemann@cox.net

Brand Niemann, former Senior Enterprise Architect & Data Scientist with the US EPA, works as a data scientist, produces data science products, publishes data stories for Semantic Community, AOL Government, & Data Science & Data Visualization DC, and co-organizes the new Federal Big Data Working Group Meetup with Kate Goodier.

TRIDENT: VISIONING A SHARED INFRASTRUCTURE FOR DATA RESEARCH AT SCALE

CHAITAN BARU, MICHAEL CAREY, TYSON CONDIE, VAGELIS HRISTIDIS, DAVID LIFKA, RICH WOLSKI, SREERANGA RAJAN, ARNAB ROY

SAN DIEGO SUPERCOMPUTER CENTER, UC IRVINE, UC RIVERSIDE, CORNELL UNIVERSITY, UC SANTA BARBARA, CLOUD SECURITY ALLIANCE, BIG DATA WORKING GROUP

In this talk, we identify the need for a shared infrastructure for data research at scale, and provide a vision for addressing this need. Scalable data management is a computer science endeavor that is currently enjoying widespread interest and a sizable industry investment. There is a pressing need to establish objective and scientific approaches to big data and data science research by providing a common platform for software experimentation. Such a platform could enable objective benchmarking and comparative analysis of software and algorithmic performance thereby improving the current situation where most researchers work on different computing platforms using different algorithms, different data, and different environmental settings.

Our vision for an open distributed platform for data science research would consist of multiple sites hosting clusters of varying scale. For example, with appropriate funding, an initial configuration may consist of a large cluster with 100's of nodes sited at, say, the San Diego Supercomputer Center, and two smaller clusters with 10's of nodes sited at, say, UC Santa Barbara and Cornell University. Such a configuration would allow for scaling experiments at a single site as well as distributed experiments involving multiple sites at varying network distances. We assume that such sites would be connected via 10Gbps networks. Where possible, sites could utilize 100Gbps links (for example between SDSC and UCSB), to enable studies on the impact of differing link speeds.

The initial design points for such a platform—which we call Trident—are (i) use of commodity hardware—to mimic typical systems that operate at scale in industry, (ii) large node count—to enable significant scaling experiments; and, (iii) large IO capability—to model big data systems. Trident would provide a software-defined infrastructure using Eucalyptus, an open-source, industry-strength cloud management software compatible with Amazon Web Services (AWS). A variety of experiment modalities would allow the scaling of experiments from small to large node counts; multi-tenancy versus controlled allocation of jobs to nodes; "bare metal" experiments; and, "full scale" runs that utilize the entire Trident cloud infrastructure in exclusive mode. Additional open sites could be added to this platform—for example one of the sites could be located within a NIST laboratory—and other sites may host different types of hardware. Trident would enable research on a variety of data science topics related to measurement science, including data and elasticity; geo-replication and processing; IaaS vs PaaS for big data applications; security and privacy; and large-scale data-intensive application development.

Importantly, one of our objectives would be to assemble and make available reference datasets and workloads via a Trident Data Repository, in order to create a "live" benchmarking environment for sharing of datasets, workloads and schema definitions in order be able to evaluate systems on the same cluster configurations. The data science research community and supportive companies could provide a diverse portfolio of datasets, workloads, schemas, and large-system failure profiles. Examples of such datasets that could be assembled include anonymized Twitter data, astronomy data, synthetic medical records datasets, the California Digital Newspaper Collection, data from smart grids, emergency response and security-related data, and continuous system monitoring data from the

Trident platform itself. Thus, Trident could enable researchers to reproduce each other's experiments using different software but with the same test data and hardware.

Trident could also make available, "out of the box", the most common data platforms, e.g., those based on Hadoop, Hive, Pig, etc., for researchers to use in comparing their newly proposed systems against existing systems at scale. Currently, access to "industrial-strength" large system configurations is available primarily to staff at the "big Web companies" in industry. An initiative like Trident would go a long way towards leveling the playing field in big data and cloud computing research by making it possible for academic researchers nationwide to validate their work at much larger scales than ever before. We wish to underscore the importance and urgency in funding such a platform in order to fill an important gap in computer science and data science research.

Chaitan Baru is Distinguished Scientist and Associate Director Data Initiatives at the San Diego Supercomputer Center, University of California San Diego, where he also directs the Center for Large-scale Data Systems Research (CLDS). Baru's interests are in research and development in the areas of parallel database systems, scientific data management, data analytics, and the challenges of data-driven science and data-driven enterprises. He initiated the Workshops on Big Data Benchmarking, WBDB (see http://clds.sdsc.edu/bdbc/workshops) with the first workshop held in May 2012 in San Jose, CA. The fifth WBDB will be held in August 2014 in Potsdam, Germany. He has played a leadership role in a number of national-scale cyberinfrastructure R&D efforts across a wide range of science disciplines from earth sciences to ecology, biomedical informatics, and healthcare. Baru is also co-Chair of the NIST Big Data Public Working Group, along with with Wo Chang, NIST and Bob Marcus, ET-Strategies (http://bigdatawg.nist.gov).

EXPERIMENTAL DESIGN GUIDANCE FOR LARGE, COMPLEX SIMULATIONS

DONALD E. BROWN

UNIVERSITY OF VIRGINIA

Much of the current research in science and engineering requires experimentation using large, complex simulations. The complexity and size mean that researchers cannot afford any or many reruns in order to obtain scientifically meaningful results. To minimize reruns experimenters need a well-structured Design of Experiments (DOE). This presentation describes the Experimental Design Guidance Engine (EDGE), a system we built for programs in the Department of Defense to provide an automated process for many aspects of DOE with modern, high fidelity simulations. EDGE supports three high level functions of the experimenter working with complex simulations. First, EDGE provides a structured way of specifying an experiment's variables (controlled, exogenous, and response) through an XML schema. Second, it gives experimenters the ability to define arbitrarily complex and interdependent constraints on the variables through an industry-standard Constraint Programming (CP) API. Finally, EDGE uses a novel experimental design algorithm with an open-source CP solver and a k-medoid clustering algorithm for large data applications to generate a space-filling DOE where all experiments satisfy the constraints. Government planners have used EDGE for important questions of resource allocation.

STANDARDIZING DATA MANAGEMENT AND INFRASTRUCTURE VOCABULARY: THE RDA COMMUNITY EFFORT

GARY BERG-CROSS

RESEARCH DATA ALLIANCE

Launched in Gothenburg, Sweden March 18-20, 2013, the Research Data Alliance (RDA) is a new, international effort to meet the challenge of the current global research data landscape which is highly fragmented, by disciplines & domains. RDA's focus, reflects its slogan "Open Access Research Data without Barriers", to build socio-technical bridges across cross-disciplinary projects and groups to enable open sharing of data from a variety of sciences. The cross-disciplinary effort employs a metaphoric strategy of common data infrastructures "building blocks" as well as building specific "data bridges" to enable research data sharing across the data lifecycle and handle data complexity. Major RDA efforts are concentrated in Working and Interest Groups, one of which is the Data Foundation and Terminology (DFT). DFT is an 18 month effort to describe a basic, abstract (but clear) data organization model and vocabulary that systemizes the already large body of definition work on data management terms, especially as involved in RDA efforts. Such a vocabulary is needed because there is no agreed on vocabulary to discuss/agree on a common international e-Infrastructures.

The model, vocabulary and its derived reference data under development are practically centered around existing models based on an analysis of about 2 dozen community's data organizations & infrastructure of. Initial common abstractions have been developed and use case with products have been agreed to within the RDA community for use across communities and stakeholders to better synchronize data conceptualization. Another goal is to enable better understanding within and between communities, and to stimulate tool building, such as for data services, supportive of the basic model's use. Candidate vocabulary terms have been extracted from the analyzed data infrastructure and sharing models as well as other RDA Work Group efforts such as Data Type Registries, Metadata, Data Policy and PID Information Types. Extant standards such as ISO 12620 & 15926 have also been leveraged but the general finding is they fail to detail concepts needed for definitions across different practice areas.

Controlled vocabulary along with candidate definitions developed by the community are currently stored in an online prototype RDA Term Collection Tool using a semantic media wiki. Representative term areas include:

1. Persistent Identifiers (PIDs and their types)
2. Digital Object - Data Object
3. Collection - Data Set - Aggregation
4. Repository (Registries and related Policies)

Work is currently underway to vet the core set of terms and the vocabulary building process at the 3rd RDA Plenary set for Dublin in March 2014.

THE NATIONAL ECOLOGICAL OBSERVATORY NETWORK: OVERVIEW AND STRATEGIES FOR MANAGING THOUSANDS OF SIMULTANEOUS MEASUREMENTS ACROSS THE CONTINENT

J. R. TAYLOR, E. AYRES, H. LUO, S. METZGER, N. PINGINTHA-DURDEN, J. ROBERTI, M. SANCLEMENTS, D. SMITH, S. STREETT, AND R. ZULUETA

NATIONAL ECOLOGICAL OBSERVATORY NETWORK

The National Ecological Observatory Network (NEON) is responsible for making observations of terrestrial, aquatic, and organismal ecology in 20 different eco-climatic domains across the continent. NEON will provide localized data on key physical, climate, and chemical forcing, as well as their associated biotic responses, in an effort to inform climate change, land-use change, invasive species, and other impact studies. The sheer volume of data is expected to exceed hundreds of Terabytes per year and will present challenges for data management on an unprecedented scale. This talk will provide an overview of NEON as a whole, while specifically focusing on how to develop and implement a standardized ecological observatory that must accommodate such a large volume of data without sacrificing quality. Highlights will include preliminary first results hosted on the NEON data portal and look toward first article science results.

Jeff Taylor is the Director of the Fundamental Instrument Unit at the headquarters of The National Ecological Observatory Network in Boulder, CO. He is responsible for overseeing the construction, deployment, and data analysis of thousands of ground-based instruments that will make atmospheric, terrestrial, and soil measurements across the continent for the next 30 years. Prior to working at NEON, Jeff was in the Atmospheric Chemistry Division at the National Center for Atmospheric Research where his work focused on satellite-based observations of the upper troposphere-lower stratosphere. He has a PhD in physics from the University of Toronto.

BACKGROUND INTENSITY CORRECTION FOR TERABYTE-SIZED TIME-LAPSE IMAGES

JOE CHALFOUN, MIKE MAJURSKI, KIRAN BHADRIRAJU, STEVE LUND, PETER BAJCSY, MARY BRADY

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

Time-lapse epifluorescence microscopy using fluorescent protein reporters provides an opportunity for imaging and analyzing the dynamics of gene expression and morphological changes in live human pluripotent stem cell (PSC) cultures. The motivation of our work comes from the need to quantify the dynamics of OCT-4 gene expression in human stem cell colonies, OCT-4 being a critical gene in the regulation of pluripotency. To move the analysis in a meaningful statistical way, the imaging must be done on high spatial and temporal resolutions that generate terabyte-sized image sets spanning hundreds of Field of Views (FOV) through time. There are several technical challenges to overcome before quantitative biological information can be obtained from these big data sets. Images of live cells such as PSCs must be acquired with low power illumination, due to the light sensitivity of cells. The acquisition requirement of minimally perturbing cells leads to a low signal-to-noise ratio (SNR) of the fluorescent signal and its sensitivity to correcting for dark current, flat-field and background media sources of noise.

This paper addresses the problem of large-scale background image correction of terabyte-sized fluorescent images with the focus on background correction in order to minimize the remaining errors in the corrected background and maximize the SNR. Dark current, flat-field and background correction models are applied over a mosaic of hundreds of spatially overlapping fields of view (FOVs) taken over the course of several days, during which the background diminishes as cell colonies grow. Our approach to background correction is formulated as an optimization problem over two image partitioning schemas and four analytical correction models. The optimization objective function is evaluated in terms of (1) the minimum Root Mean Square (RMS) error remaining after image correction, (2) the maximum SNR reached after down-sampling, and (3) the minimum execution time.

Based on the analyses with measured dark current noise and flat-field images, the most optimal GFP background correction is obtained by using a data partition based on forming a set of sub mosaic images with a polynomial surface background model. We show that the background noise in terabyte-sized fluorescent image mosaics can be corrected computationally with the optimized triplet (data partition, model, SNR driven down-sampling) such that the total RMS value from background noise does not exceed the magnitude of the measured dark current noise. In this case, the dark current noise serves as a benchmark for the lowest noise level that an imaging system can achieve. In comparison to previous work, the past fluorescent image background correction methods have been designed for single FOV and have not been applied to terabyte-sized images with large mosaic FOVs, low SNR and diminishing access to background information over time as cell colonies span entirely multiple FOVs.

COMPLEX BIG DATA ON A BUDGET

JOSEPH SCHNEIBLE

TECHNICA CORPORATION

Data production is increasing exponentially and creating a corresponding demand for Big Data processing. Although some Big Data processing can be done efficiently at a reasonable cost, more complex processing can require expensive investments in hardware and software. For example, for tasks that are “embarrassingly parallel”, a Hadoop-based cluster provides an efficient solution. However, more complicated tasks with significant data inter-dependencies, such as graph analysis, often require large, shared-memory solutions. These solutions use specialized hardware with large amounts of RAM, high-speed interconnections, and many CPUs.

One alternative is the Parallel Sliding Window approach[1], which is an out-of-core graph processing technique that can process large data sets in an efficient manner on commodity hardware. The Parallel Sliding Window approach organizes data into partitions that can be loaded into memory separately. Within these partitions, data is organized such that portions needed when updating other partitions are in contiguous blocks, minimizing the required number of I/O operations. Communication between partitions occurs when other partitions read these blocks with their updated information.

Additionally, a well-designed out-of-core algorithm that orders independent operations to minimize I/O can operate as efficiently as an in-core algorithm. In this case, the computation can be limited by the speed of the processors. Due in large part to physical constraints, the speed of individual CPU cores has not increased significantly in recent years. To overcome this constraint, graphics-processing units (GPUs) have increasingly been used for general computation because they provide a large number of simple processors. By exploiting this

massive parallelism, it is possible to speed up many applications by an order of magnitude or more. However, applying this to big data applications is complicated by the irregularity of data commonly found in the real world.

For example, many real world networks, such as social, computer, and even biological networks, exhibit a scale-free behavior. This means that the vertices in the network have a power-law degree distribution in the number of edges in which they participate. Because the Parallel Sliding Window approach is vertex-centric, this uneven degree distribution makes it challenging to fully utilize GPUs. With many algorithms, the amount of computation as a function of the number of edges grows faster than the amount of data. In this case, higher degree vertices present an opportunity for greater data reuse and are efficient to process on the GPU, whereas lower degree vertices are better processed on CPUs.

We propose a solution that combines the I/O efficiency of the Parallel Sliding Window approach with the improved data processing of out-of-core hybrid CPU/GPU algorithms. Supporting this is a combination of computational models and micro-benchmarks used to determine an efficient work distribution between the CPU and GPU, depending on the algorithm and system. Using this approach, off-the-shelf desktop hardware can perform comparably to dedicated solutions in some scenarios, providing a high performance to cost ratio with a low barrier to entry.

References

[1] Kyrola, A., Blelloch, G., and Guestrin, C. GraphChi: Large-scale graph computation on just a PC. In OSDI (2012).

A BLENDED APPROACH TO BIG DATA ANALYTICS

RICHARD HEIMANN

DATA TACTICS CORPORATION

Topics expected to be broached:
Technical approaches to complex workflow components of Big Data systems, analytics, visualization, human-system interaction and major forms of analytics employed in data science.

Data Scientists are rewarded heavily for clever solutions to nontrivial problems. But, is this the optimal solution for users and decision makers alike?

No Free Lunch for Theorems tells us that no algorithm performs better than any other when their performance is averaged uniformly over all possible problems of a particular type. Meaning there is no such thing as a general purpose algorithm and analytics must be designed for a particular problem. This insight naturally leads us to rewarding data scientists for cleaver solutions. However, NFL also provides good evidence for analytical pluralism, which happens to be pretty central for a blended approach.

The evaluation of hybrid modes such as the blended approach includes objective and subjective pattern discovery. The objective element of deployment is important broadly speaking but really important as we include users. The objective element of data mining is the production of valid, accurate and nontrivial patterns as well as fixing pattern paradoxes. The subjective elements of data mining are allowing users to determine if patters are useful, novel and comprehensible.

The blended approach is both a mixture objective and subjective pattern discovery, facilitated by interactive analytics as well as overlapping solutions. An example would be the overlapping of two solutions to analyze Data D with Analytics A and Analytic B where A provides some insight to smooth pattern detection like a summary analytics (perhaps PCA or ICA) and B offered some insight to rough patterns in the data such as outlier detection. These two methods offer unique insights and may at times validate each other. Users would understand both structural patterns and structural breaks in D.

The lesson is that the elegance of analytics lies, at times in its inelegance. Analytics can validate and at points of divergence offer unique insights. Overlapping solutions can argue and agree with each other. We can feel conflicted with one analytic, but analytical pluralism is a representative democracy that functions by competition among parties who all believe they know a way to solve the problem. NFL for theorems shows us that this may be the best environment for analytics.

Overlapping solutions may be the thing that unlocks the power of analytics. Many data scientists still approach the problem by assuming theres a best way to solve a problem, but ignore alternate solutions and most egregiously ignore the user. The lesson is to abandon the question "What is the cleverest way to solve the problem" in favor of "Are there multiple, overlapping ways to solve this problem?"

Richard Heimann focuses on data science, big data and social scientific research. He has recently supported special programs at DARPA and MAP-HT and is now Lead Data Scientist at Data Tactics Corporation in McLean VA. In addition, Richard is currently an Instructor of Human Terrain Analysis at George Mason University and adjunct faculty at The University of Maryland, Baltimore County teaching Spatial Statistical Reasoning.
Richard also writes for the Big Data Republic on related topics and has an upcoming book titled Social Media Mining in R covering sentiment analysis and opinion mining.

METRICS FOR AND ASSESSMENTS OF BIG DATA EXPLOITATIONS SYSTEMS: A USER-CENTERED APPROACH

DAN TRAVIGLIA, JOSHUA C. POORE, DAVID REED, JANA L. SCHWARTZ

THE DRAPER LABORATORY

The volume and complexity of data continues to grow dramatically, and with them, the roles and responsibilities of data scientists and analysts. To address their needs and dynamically changing missions, the landscape for the development of tools to aid data scientists in their mission is moving to an open-component and/or self-authoring model. In this agile model, big data evaluation systems (BDES) can be constructed from a variety of analytic, visualization, and IT components to bring complex and varied data to address evolving missions. However, this raises new challenges for BDES system benchmarking, performance metrics, and assessment methodologies; they must be equally agile to be effective within new, open systems of rapidly assembled components. Among the most dynamic components within a BDES is the user; there are currently no institutionalized models for how best to assess them in conjunction with other components to ensure that a BDES (including the user) is performing efficiently within different mission contexts at the system level. In this paper presentation, we will introduce new insights and approaches for BDES assessment with a focus on metrics for user performance and experimental requirements for capturing variance at the BDES component level.

Previous approaches to understanding how users' learn and interact with complex software applications rely on retrospective or interruptive protocols (i.e., questionnaires, "talk-aloud" protocols), post-hoc interviewing, parallel physiological monitoring, high-frequency behavior clustering, or experimental design. Such approaches typically use contrasts (i.e., physiological arousal taken against "baselines"), data once-removed from the context of the human-computer interaction loop (i.e., questionnaires, interviews), or artificial constraints placed on the context itself to enable easier interpretation of the data (i.e., scripted workflows, as in tutorials). These strategies prohibit the generalization of findings across the wide range of user experiences and interfaces, and often average over subtle user variance that could be informative for assessment purposes.

Draper Laboratory has formalized a context-sensitive approach to understanding the user—fusing information about a user's activities (behavior) within the virtual context they are embedded, such as a BDES analytic or visualization interface. We achieve this approach by categorizing critical functions and features of the interface, hierarchically linking those categories to germane functions, and temporally decomposing the sequence of these features as events nested within a time-series. Discrete user behavior, metadata, and physiology collected in parallel can be similarly oriented along the time-series, enabling us to model and identify how users are responsive to elements of the interface, how workflow changes across usage, and how users' strategies mature as they gain experience. Our context-sensitive approach enables meaningful metrics for users' experience and workflow across both substantively different kinds of interfaces and usages of the same interfaces. Most importantly, our approach neither constrains organic user behavior, nor how they explore and learn different tools.

Draper has demonstrated this approach on a number of projects, both externally and internally funded. Leveraging simple, elegant mixed ANCOVA experimental designs, we have developed a systematic approach for isolating BDES component-level variance in human subject testing, so that interactions can be modeled that capture system-level performance in the context of different mission-focused challenge problems. We are currently utilizing this approach to model intelligence analysts' workflow with novel analytic tools, with a goal of using our approach as the basis of a recommendation system for selecting the right tool to the right user for use in a given analytic mission space.

Acknowledgments The project or effort depicted (XDATA) is sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL).

Dr. Joshua Poore is a Senior Member of the Draper Laboratory Technical Staff and experimental psychologist. He has applied this work to developing comprehensive metrics for user states (i.e., immersion, proficiency) within human computer interfaces, including virtual environments, simulations, gaming, and analytic tools;

Dr. Jana Schwartz is a Principal Member of the Technical Staff at Draper Laboratory, Group Leader for Human-Centered Engineering, and PI of Draper's XDATA team. She has interests in human/system collaborative operations, complexity theory, and the use of real-time psychophysiological response for closed-loop performance.

QUANTIFYING SOURCES OF UNCERTAINTY THROUGH TRACEABLE AND EMPIRICAL APPROACHES AT THE NATIONAL ECOLOGICAL OBSERVATORY NETWORK

JOSHUA A. ROBERTI, JANAE L. CSAVINA, STEFAN METZGER, SARAH STREETT, AND JEFFREY R. TAYLOR

NATIONAL ECOLOGICAL OBSERVATORY NETWORK, INSTITUTE OF ARCTIC AND ALPINE RESEARCH, UNIVERSITY OF COLORADO

The National Ecological Observatory Network (NEON) is a continental-scale research platform with a projected lifetime of 30 years. NEON's purpose is to provide high quality data products that will facilitate discovering and understanding the impacts of climate change, land-use change, and invasive species on ecology. To accomplish this, NEON will perform in-situ, sensor-based measurements of approximately 55,000 data streams, among others. With this comes great responsibility for ensuring and reporting data quality, and for providing reliable uncertainty estimates. Only when uncertainty is sufficiently quantified, can meaningful interpretations be made about mean quantities and their interrelations. These in turn are the main ingredients for constructing or constraining process-based models and the like. Given that NEON data will be publically available, our goal is to ensure that all sources of uncertainty are identified and if possible, quantified in a standardized and traceable manner. To meet this goal, laboratory calibrations and measurement uncertainty estimates follow ISO protocols. As a result, quantifiable as well as unquantifiable (i.e., those that can only be identified) uncertainties are provided in publically available documents. To complement this approach, empirical uncertainty estimates are formulated and realized as data are being collected. Using this approach we aim to gain a better understanding of sensor-specific uncertainties to quantify known, though previously unquantifiable uncertainties. Here, we provide an example of the process for NEON's air temperature measurements.

SO YOU'VE SEQUENCED A GENOME – HOW WELL DID YOU DO?

JUSTIN ZOOK

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST)

Next Generation Sequencing is being widely adopted for clinical applications. At present, there is no widely accepted set of metrics to understand the trustworthiness of variant calls from NGS. The Genome in a Bottle Consortium (www.genomeinabottle.org) is developing standards to address this need.

We have developed methods to understand accuracy of genotype calls at millions to billions of positions across a genome, which are publicly available through www.genomeinabottle.org and are being rapidly adopted by clinical, research, and bioinformatics groups performing Next Generation Sequencing. Previous work showing high discordance between sequencing methods and algorithms has highlighted the need for a highly confident set of genotypes across a whole genome that could be used as "truth" for understanding accuracy. Therefore, we have developed methods to make highly confident SNP, indel, and homozygous reference genotype calls for NA12878, the pilot genome for the Genome in a Bottle Consortium. To minimize bias towards any sequencing method, we integrate 10 whole genome and 3 exome datasets from 5 different sequencing platforms. We also integrate

variant calls from variant callers using only mapping, mapping with local de novo assembly, and global de novo assembly. The resulting genotype calls are more accurate than calls from any individual dataset, and allow performance assessment of more difficult variants than using microarrays. When assessed against fosmids, inheritance, and other validation datasets, we estimate that our calls contain less than one false positive or false negative per 10^8 bases in our confident regions, which make up ~85% of the genome. In addition to making our calls publicly available, we have integrated our highly confident variants into the GCAT website (www.bioplanet.com/gcat) so that anyone can interactively generate performance metrics for different combinations of sequencing datasets, mapping algorithms, and variant callers. In addition, as part of the Genome in a Bottle Consortium, we are now developing highly confident structural variant calls, and we plan to continue to improve the characterization of the pilot genome and additional genomes.

Justin Zook is currently working on developing reference materials, reference data, and reference methods for human genome sequencing with the Genome in a Bottle Consortium. Specifically, he is developing bioinformatics methods to compare and integrate whole genome DNA sequencing data from multiple platforms and sequencing runs, thereby generating “ground truth” for large sequencing datasets. As an NRC postdoctoral research associate at NIST from 2009-2010, Justin’s previous research interests included analytical chemistry, microfluidics, and nanotoxicity measurements.

A COORDINATED VIEW OF THE TEMPORAL EVOLUTION OF LARGE-SCALE INTERNET EVENTS

ALISTAIR KING, ALBERTO DAINOTTI, BRADLEY HUFFAKER, KC CLAFFY

UNIVERSITY OF CALIFORNIA, SAN DIEGO

We present a method to visualize large-scale Internet events, such as a large region losing connectivity, or a stealth probe of the entire IPv4 address space. We apply a well-known technique in information visualization "multiple coordinated views" -- to Internet-specific data. We animate these coordinated views to study the temporal evolution of an event along different dimensions, including geographic spread, topological (address space) coverage, and traffic impact. We explain the techniques we used to create the visualization, and using two recent case studies we describe how this capability to simultaneously view multiple dimensions of events enabled greater insight into their properties.

http://www.caida.org/publications/papers/2013/coordinated_view_internet_events/

CERTIFIED ANALYTICS PROFESSIONAL (CAP®) PROGRAM

LOUISE WEHRLE

INFORMS

Given the widespread use and continuing development and popularity of analytic methodologies, INFORMS leadership determined a need for a personnel qualification standard. Thus was born the Certified Analytics Professional (CAP®) program. Those earning the CAP designation will have verified education, experience, ethics, effectiveness and successful completion of an examination. The credential is based on an analysis of practice and thus every question on the exam and every requirement for eligibility is tied directly to what a professional analyst does. There are seven areas of assessment that mirror the analytics project development: Business problem framing, analytics problem framing, data, methodology selection, model building, deployment and lifecycle management.

How could this be applicable in industry, government and/or nonprofit organizations?

CAP is applicable to all of these and more: it provides a qualification standard for analytics practice. It establishes a benchmark for individuals, a standard for hiring, promotion, tenure and provides a career path for those engaged in analytics. This foundational credential is both vendor and software neutral, thus allowing the holder to work in any area of choice and perform analytics for data science, finance, healthcare, banking, energy, pharmaceuticals, insurance, retail and more.

How is the project innovative?

While there are many new academic and professional programs that teach and provide information on analytics, the CAP is the only assessment and qualification program that can be independently sought. The CAP program, based on practice, does not depend on the content of any academic curriculum or of any course content. CAP takes the practice of analytics, the expertise of analytics professionals and the regulations governing development of certification programs and melds them into one coherent, criterion referenced program that is the hallmark of analytics professionals.

SUPPORT FOR LEVERAGE POINTS IN MULTIVARIATE VISUALIZATION USER DATA

MARK A. LIVINGSTON, KRISTEN LIGGETT, PAUL HAVIG, JASON MOORE, JONATHAN W. DECKER, ZHUMING AI

NAVAL RESEARCH LABORATORY, AIR FORCE RESEARCH LABORATORY

Visual data representations aim to elicit a response in the user, generally to identify an interesting pattern or feature indicated by the visual representation as a property of the underlying data. Recently, leverage points were proposed as stages between sensation (perception) of data and cognition of information in data1. Our goal is to

examine quantitative studies of multivariate visualization (MVV) techniques to determine whether this evidence agrees with the theory behind leverage points.

The first leverage point is the focusing of exogenous attention by salient cues which alert users to changes or important aspects of the data. Examples include varying color and texture in multivariate visualizations. Data driven spots (DDS) use hue to identify variables and intensity to encode value; oriented slivers (OS) use orientation and intensity (right), while Attribute blocks (AB) use grid position and hue (respectively). These cues are all pre-attentive, but hue is not easily understood as a metric. This could explain why DDS and OS performed well in most of our studies. But in our most recent study7, DDS fared poorly. By image analysis, we find the target density correlated with user error (graphs, below). DDS confounded this pre-attentive cue by dropping the relative density of the target below 1.0 more often than OS or AB. Many users said they used density to respond. This confounding of users' exogenous attention apparently increased error. While further evidence would be needed to validate this interpretation, it does show how leverage points may give insight to user performance with MVV.

Another leverage point occurs when the visualization provides strong grouping cues. As seen in the images above, grouping cues come through a variety of pre-attentive mechanisms: proximity, color, and orientation respectively. We explore how this may have helped or hindered users' performance. We further address how the other leverage points (endogenous attention, mental models, structure, and training) were unlikely to have been exercised by these particular studies, giving direction for future work.

DISASTER RISK MANAGEMENT CALLS FOR BIG EARTH OBSERVATION DATA SCIENCE (BIGEODS)

PESARESI MARTINO, FERRI STEFANO, FLORCZYK ANETA J., KEMPER THOMAS, SYRRIS VASILEIOS, SOILLE PIERRE

GLOBAL SECURITY AND CRISIS MANAGEMENT UNIT, INSTITUTE FOR THE PROTECTION AND SECURITY OF THE CITIZEN OF THE EUROPEAN COMMISSION'S JOINT RESEARCH CENTER.

The poster summarizes the work done by the GLOB-HS project team in support to Disaster Risk Management policies of the European Union. The project aims to produce the first global fine-scale representation of the human settlements derived from automatic processing of high and very-high spatial resolution remotely-sensed imageries (GHSL - Global Human Settlement Layer). In the specific application scenario, human settlement information is used for exposure and damage assessment mapping inside decision support systems. A new EO data processing framework has been proposed in order to solve the complexity of information retrieval from bulk, heterogeneous and inconsistent high-spatial resolution image data set. [ref. M.Pesaresi et al. "A Global Human Settlement Layer From Optical HR/VHR RS Data: Concept and First Results", JSTAR2013 DOI:10.1109/JSTARS.2013.2271445] . Updates on the GHSL processing methodology applied to new data scenarios (European, Global) experimented during 2013 will be provided.

Why it is big data

Big data is commonly referred to as data characterised by Volume, Velocity, Variety, and Veracity (the so-called 4V principle). Indeed, because disasters can virtually happen at any place around the globe, disaster risk management calls for global availability of EO and ancillary data with suitable spatial resolution. In addition, the diversity of disasters requests data in various domains of the spectrum and other modalities such as elevation measurements or even non-image data such as census data (Variety). Furthermore, the constant natural and anthropogenic change of our planet calls for time series (Velocity). This has been epitomized recently by the commercial availability of 1m spatial resolution video streams captured from orbiting satellites. All previous characteristics naturally lead to sheer data Volume. Finally, because the information extracted from Earth Observation Data in the context of risk management may lead to decisions that involve saving life and assets as well as decide where to spend public money, it needs to be based on trusted data taking its uncertainty into account while the extracted information itself must be validated and associated with accuracy measurements (Veracity).

Why big data *science* is needed

Often, big data is addressed using data mining tools aiming at automatically revealing correlations and other patterns/clusters in the data. For the special case of big Earth observation data, data mining tools need to be used in conjunction with hypothesis driven science. Indeed, take for instance the simple and widely used vegetation index that successfully reveals the degree of vegetation occurring at a given location. It is unlikely that the ratio of specific spectral band differences could have been automatically revealed by knowledge discovery approaches. We believe that successful Big Earth Observation Data Science calls for the exploitation of both expert domain knowledge and data mining/machine learning techniques. In short, we need to make sure that big data does not preclude (big) judgement! This approach was considered in the GHSL project. Indeed, in order to solve the computational and semantic complexity of the GHSL task, hypothesis-driven reasoning in conjunction with newly-defined optimized image processing tools has been demonstrated.

Details on EO data used and data challenges

Geo-information deals with information placed in the spatial domain, consequently related to scale and spatial tolerance notions. A global description of the Earth land mass at a spatial resolution of 1m corresponds to 150 Tera-image-elements (Pixels). This poses the problem of generating/accessing appropriate ground truth for guiding the information extraction and/or validate the extracted information (issue of performance measurement with limited or no ground truth). The main data challenges can be listed as i) positional uncertainty of the pixel measurements; ii) high variability of sensors and associated quality, iii) high variability of illumination, atmospheric, and topographic conditions influencing the data collection, iv) high variability of built-up and settlements across the globe both in terms of materials and spatial patterns, v) only partial access to technical specs of available data and sensors.

Contribution to Datasets to Enable Rigorous Data Science Research

The generation of a Global Human Settlement Layers and their characterization contribute to the creation of reference datasets of general interest in the context of (but not limited to) disaster risk management. More generally, the European Commission through its Copernicus programme will actively participate to the creation of data science reference datasets by the free and open access delivery of the data produced by the forthcoming SENTINEL missions. This will foster advances in Big EO Data processing, analytics, and visualization. In this context, the Joint Research Centre of the European Commission is co-organizing with the European Space Agency and the European Union Satellite Centre, the 1st Conference on Big Data from Space that will take place at ESA-ESRIN, Frascati, Italy, from the 12th to the 14th of November 2014. The focus of the conference is on the whole data life

cycle, ranging from data acquisition by space borne and ground-based sensors to data management, analysis and exploitation in the domains of Earth Observation, Space Science, Space Engineering, Space Weather, etc.

LARGE DATASET GENERATION AND ANALYSIS OF OPTICAL MICROSCOPY IMAGES FOR QUANTIFYING DYNAMIC CHANGES IN PLURIPOTENT STEM CELL CULTURES

MICHAEL HALTER

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST)

Optical imaging of complex biological samples such as tissues and cells can produce very large multidimensional datasets (GB to TB in single experiments). The size of these datasets can prohibit effective manual and quantitative analysis with the software and hardware tools readily available to bioscience researchers. At the same time, large datasets are needed to assure (or test) for adequate sampling and for meeting high quality criteria of bio-manufacturing products. The sample size required for quantitative imaging experiments is partly determined by measurement uncertainty, but in this case, biological variability is often a much larger component of dispersion in the measurements. Biological variability is a common feature of biological systems, and understanding how biological variability arises in genetically identical populations is likely to provide fundamental insight into the mechanisms of biological control. The parameters measured from complex biological specimens frequently have a large dispersion and require that large numbers of images and replicates be examined.

We report the challenges of acquiring and processing of TB-sized images per single experiment. For the dataset that has already been acquired, for each of three replicate experiments ~280 cell colonies over an area of 225 square mm were imaged. To accomplish this, 352 fields of view were stitched together to produce a single image of all colonies that is approximately 1 GB (in size). The 225 square mm area was imaged every 45 min for 5 days to produce, for each replicate, a full dataset of 400 GB.

In terms of acquisition, a significant challenge in biology is the very large (nearly infinite) parameter space that can affect the biological system. Acquiring high quality datasets from biological samples can be challenging, but is essential. After all, large quantities of bad data are not useful. Designing experiments is challenging because of the potential for the experiment to perturb the biology, particularly when imaging live cells over time. Exposure to light can perturb living systems, and difficulty in maintaining temperature, humidity, nutrients, and other conditions can greatly impact the usefulness of the experimental data. Data from a recent study involving 5-day imaging experiments with living cells required the use of relatively low light levels and relatively short exposure times, requiring binning of pixels for analysis in order to produce a statistically reliable signal to noise ratio. Novel flat field and background correction procedures are required to minimize artifacts that result from image stitching and alignment. Other compromises in the experiment must be considered. That dataset for example was collected with a relatively low power objective lens, which precluded a spatial resolution that would allow analysis at the single cell level. Instead, these data are analyzed at the level of many cells that are residing as colonies, and a different experiment will be required to achieve cell by cell resolution.

The purpose of these experiments is to understand how to quantify the growth and function of stem cell colonies, as related to the expression of a fluorescent marker of cell state. Since such data have never been taken before,

the images themselves provide hints about the kinds of features and trends that may be meaningful to quantify. This requires understanding the hardware and software challenges that include acquisition, processing, viewing and exploring very large datasets.

INFORMATION THEORETIC EVALUATION OF DATA PROCESSING SYSTEMS

MICHAEL HURLEY

MIT LINCOLN LABORATORY

An important aspect of data science technologies that is often thought to be poorly understood is how to measure the performance of these technologies and correctly interpret the results. We believe that the foundation already exists in information theory. When these technologies are used to drive critical decision making, the data processing chains can be viewed as information channels. Information theory then provides a general and extensible set of performance metrics that can be used to evaluate the efficiency of the chains. Development is primarily needed to determine how to best apply this foundation to the specific requirements in data science. Our group has developed information theoretic techniques to evaluate the overall performance of systems like multi-target trackers and classifiers, as well as human-in-the-loop work processes. We have developed techniques to estimate information theoretic measures and variances from truth datasets, where the truth and output can be mapped to an NxM matrix of symbols. The variance measures provide the means to determine the statistical significance of test results and to decide if sufficient data were used for an evaluation. The information theoretic measures also provide unifying measures that can relate more traditional, noncommensurate performance measures (like time, processing power, communications bandwidths, storage capacity) by their relative impact on the information loss in a system. We will present an overview of the information techniques that we have developed, examples of the systems that we have evaluated, and what work is needed to extend the utility of this approach to the evaluation of other systems.

This work is sponsored by the Assistant Secretary of Defense for Research & Engineering under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

Michael Hurley is a staff member at MIT Lincoln Laboratory in the Intelligence and Decision Systems Group. He has worked on a multitude of projects including system architecture studies, operations research analysis, detection, tracking, and data fusion algorithms, and information theoretic techniques for assessing decision system performance. He holds a PhD in high energy particle physics from the University of Pennsylvania.

MASSIVELY SCALABLE DISTANCE-BASED DISTRIBUTED OUTLIER DETECTION ALGORITHMS

ONUR SAVAS, TUNG THANH NGUYEN, JULIA DENG

INTELLIGENT AUTOMATION, INC.

Distance-based outlier detection is used to find anomalies in many domains such as Earth sciences, astronomy, and space applications. However, the problem of distance-based outlier detection is difficult to solve efficiently in massive datasets because of potential quadratic time complexity. In this work, we address this problem by combining simple yet effective indexing and pruning techniques that further improves the state-of-the-art. The indexing scheme is based on sorting the data points in order of increasing distance from a fixed reference point and then accessing those points based on this sorted order. The indexing is efficiently implemented in parallel by using trie partitioning such as in Hadoop TeraSort. To speed up the basic outlier detection technique, we propose two distributed algorithms using a master/worker architecture suited for modern multi-core clusters. The first algorithm assumes a ring topology and can be employed even only the master node has access to the whole dataset. The second algorithm assumes a star topology and is empoyed when all nodes have access to the whole dataset. The first algorithm passes data blocks from each machine around the ring, incrementally updating the nearest neighbors of the points passed. By maintaining a cutoff threshold, it is able to prune a large number of points in a distributed fashion. The second algorithm operates in a similar fashion though the communications is minimized because the worker nodes can access the whole dataset. We have implemented both algorithms using multi-threaded MPI. Both algorithms have been shown to scale close to linear on datasets with millions of points in addition to speed-ups close to the number of worker nodes.

TO MEASURE OR NOT TO MEASURE TERABYTE-SIZED IMAGES?

PETER BAJCSY

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST)

This talk will elaborate on a basic question “To Measure or Not To Measure Terabyte-Sized Images?” posed by William Shakespeare if he were a bench scientist at NIST. This basic question is a dilemma for many traditional scientists that operate imaging instruments capable of acquiring very large quantities of images. However, manual analyses of terabyte-sized images and insufficient software and computational hardware resources prevent scientists from making new discoveries, increasing statistical confidence of data-driven conclusions, and improving reproducibility of reported results.

Motivated by NIST’s mission and the above dilemma, I will provide a high level overview of overarching questions that are related to big image data research in terms of (1) interactive visualization of terabyte-sized images to leverage human visual inspection, (2) access to image sub-sets and samples from any geographical location to support collaborative research, (3) seamless transitions of computations from desktop to cluster/cloud computing environments to speed up learning and discoveries, and (4) orchestration of on-demand computational services and client-server communication to provide ubiquitous hyperlinking of image data with extracted image

characteristics and derived knowledge. The presentation will include live demonstrations of prototype software systems developed for cell biologists and material scientists.

Peter Bajcsy received his Ph.D. in Electrical and Computer Engineering in 1997 from the University of Illinois at Urbana-Champaign (UIUC) and a M.S. in Electrical and Computer Engineering in 1994 from the University of Pennsylvania (UPENN). He worked for machine vision, government contracting, and research and educational institutions before joining the National Institute of Standards and Technology (NIST) in 2011. At NIST, he has been leading a project focusing on the application of computational science in biological metrology, and specifically stem cell characterization at very large scales. Peter's area of research is large-scale image-based analyses and syntheses using mathematical, statistical and computational models while leveraging computer science fields such as image processing, machine learning, computer vision, and pattern recognition. He has co-authored more than more than 24 journal papers and eight books or book chapters, and close to 100 conference papers.

A TAXONOMY FOR THE BIG DATA LANDSCAPE

PRAVEEN MURTHY, ARNAB ROY, SREE RAJAN

FUJITSU LABS OF AMERICA

In this work, we develop a taxonomy of big data domains and the various technologies and algorithms that apply to each of these domains. The primary purpose of the taxonomy is to benchmark the various technologies and algorithms, and also to orient newcomers to the area towards understanding and categorizing the big data landscape. Given the cornucopia of acronyms and terms, technologies, algorithms, and concerns, it is difficult for the uninitiated to determine where one should begin. The goal of our work is to be able to not only explain the terms and technologies but also to provide a big picture view of how and where they fit in, and what the current state-of-the-art is in practice.

Towards that end, we want to benchmark several aspects related to the big data landscape: a categorization of the various domains in which big data arise, the storage architectures that are required and their relative strengths and weaknesses, the compute infrastructure requirements, both for batch processing and real-time analytics, provenance for data origins, machine learning algorithms that are used for analytics, metrics for evaluating them, and security and privacy requirements of the data domains and the infrastructure.

While the most well-known big data processing infrastructure is the Hadoop ecosystem, the reality is that there are several compute infrastructures that are used in various domains. We will catalog these infrastructures, and map them to the various axes of the domain space from which big data is drawn and discuss their strengths and weaknesses.

The topic of machine learning has a rich history and corpus of approaches, philosophies, and algorithms. We will categorize these various approaches along different axes including algorithm types, and by the type of data that is processed. Into these broad categories, we will map the huge number of algorithms that are used in practice, and summarize the various types of metrics that are used to evaluate their performance. We also develop a hierarchy

or partial ordering of these algorithms by complexity and cost so that one can better evaluate and match the type of algorithm one should use for a specific use case and data domain.

We also discuss the types of security and privacy concerns that arise in different data domains, types of threats and attacks that one should be concerned with, and the various approaches that are currently used to achieving security and privacy in big data infrastructure and analytics.

> **Praveen Murthy** is a researcher in the software systems innovation group at the Fujitsu Labs of America. His research interests lie in developing program analysis techniques for analyzing security problems in web and big data infrastructure. He co-leads the attack surface reduction subgroup of the Big Data Working Group in the Cloud Security Alliance. Praveen holds a Ph.D. degree in EECS from UC Berkeley.

TPC-BIG DATA BENCHMARK INITIATIVE

RAGHUNATH NAMBIAR

TPC

Industry standard benchmarks have played, and continue to play a crucial role in the advancement of the computing industry. Demands for them have existed since buyers were first confronted with the choice between purchasing one system over another. Historically we have seen that industry standard benchmarks enabled healthy competition that results in product improvements and the evolution of brand technologies. Now, Big Data has become an integral part of mainstream IT ecosystem across all verticals. Industry and research community are challenged with effective means to measure the performance and price-performance hardware and software dealing with big data. Considering the importance the Transaction Processing Performance council (TPC.org) has formed a committee to develop set of industry standards to measure these aspects. This session presents the status a report from this committee.

INSURING THE QUALITY OF THE NATIONAL ECOLOGICAL OBSERVATORY NETWORK'S TOWER SENSOR DATA

S. STREETT, D. SMITH, J. TAYLOR

NATIONAL ECOLOGICAL OBSERVATORY NETWORK

The National Ecological Observatory Network's Fundamental Instrument Unit (NEON-FIU) is responsible for making automated terrestrial observations at 60 different sites located within the 20 eco-domains into which NEON has partitioned the United States. FIU will collect and process data on key physical and chemical climate properties for a time period of 30 years. The sheer volume of data that will be generated from the 10's of thousands of remotely deployed sensors far exceeds that of any other observatory network or agency (i.e., > 45 Tb/year).

NEON will produce high-level quality assured data products from these measurements that will be publicly accessible through NEON's data portal. Here we address the question of how to maintain the quality of our data products in near real-time. Results from plausibility testing will be presented for early temperature data collected at NEON's tower site in Sterling, CO.

STOP WRITING CUSTOM DATA PARSERS -- WRITE DFDL INSTEAD!

STEPHEN LAWRENCE

TRESYS TECHNOLOGY

This talk gives an introduction to the Data Format Description Language (DFDL), how it can be used to parse both textual and binary data in a standardized way, and how this leads to less time spent on custom data parser development and consequently, more time spent on data processing and analysis. The talk will then describe the current DFDL implementations, with focus on the open-source Daffodil project and its design. It will conclude with a brief walkthrough of real DFDL examples, including commercial and scientific formats, and a demonstration of the parsing capabilities of Daffodil.

The DFDL specification, which has completed a second round of public comments as part of the Open Grid Forum (OGF), is a modeling language for describing general text and binary data using a subset of XML Schema augmented with data format annotations. DFDL allows data to be read from its native format and presented as an instance of an information set or an XML document. DFDL also allows the reverse, through conversion of an information set back to its native format. By using the information set, this cleanly integrates with common XML utilities (e.g. XProc, XSLT, XQuery) for data processing and analysis regardless of the format of the native data.

Two implementations of DFDL exist, as is required by the OGF to become a standard. The first, created by IBM and already shipped in several IBM products (such as IBM Integration Bus v9), is written in both Java and C and includes graphical tools for modeling, running, and debugging DFDL schemas. The second implementation, Daffodil, is an open-source project written in Scala, with a design focused on speed and correctness. With the two implementations making great strides, and the DFDL specification nearing standardization, DFDL is becoming a promising tool that will ease data parsing, processing, and analysis.

> **Stephen Lawrence** has worked as a software engineer at Tresys Technology since 2007, while contributing to the open-source Daffodil project as a core maintainer for almost two years. He works alongside Michael Beckerle, the co-chair of the DFDL Working Group, to develop Daffodil and improve the DFDL specification. Outside of Daffodil, he focuses on computer security applications, including file inspection and sanitization, Security Enhanced Linux (SELinux), and cross domain solutions.

SEMANTIC GRAPH-SEARCH ON SCIENTIFIC CHEMICAL AND TEXT-BASED DATA

TALAPADY BHAT

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST)

Application Establish social media-based infrastructure, terminology and semantic data-graphs to annotate and present technology information using 'root' and rule-based methods used primarily by some Indo-European languages like Sanskrit and Latin.

Current approach Many reports, including a recent one on Material Genome Project finds that exclusive top-down solutions to facilitate data sharing and integration are not desirable for federated multi-disciplinary efforts. However, a bottom-up approach can be chaotic. For this reason, there is need for a balanced blend of the two approaches to support easy-to-use techniques to metadata creation, integration and sharing. This challenge is very similar to the challenge faced by language developer at the beginning. One of the successful effort used by many prominent languages is that of 'roots' and rules that form the framework for creating on-demand words for communication. In this approach a top-down method is used to establish a limited number of highly re-usable words called 'roots' by surveying the existing best practices in building terminology. These 'roots' are combined using few 'rules' to create terms on-demand by a bottom-up step.

> Y(uj) (join), O (creator, God, brain), Ga (motion, initiation) –leads to 'Yoga' in Sanskrit, English
> Geno (genos)-cide–race based killing – Latin, English
> Bio-technology –English, Latin
> Red-light, red-laser-light –English.

A press release by the American Institute of Physics on this approach is at
http://www.eurekalert.org/pub_releases/2013-07/aiop-ffm071813.php

Results Our efforts to develop automated and rule and root-based methods (Chem-BLAST -. http://xpdb.nist.gov/chemblast/pdb.pl) to identify and use best-practice, discriminating terms in generating semantic data-graphs for science started almost a decade back with a chemical structure database. This database has millions of structures obtained from the Protein Data Bank and the PubChem used world-wide.

Subsequently we extended our efforts to build root-based terms to text-based data of cell-images. In this work we use few simple rules to define and extend terms based on best-practice as decided by weaning through millions of popular use-cases chosen from over hundred biological ontologies.

Currently we are working on extending this method to publications of interest to Material Genome, Open-Gov and NIST-wide publication archive - NIKE. - http://xpdb.nist.gov/nike/term.pl

These efforts are a component of Research Data Alliance Working Group on Metadata https://www.rd-alliance.org/filedepot_download/694/160 & https://rd-alliance.org/poster-session-rda-2nd-plenary-meeting.html

ALGORITHM CHARACTERIZATION AND IMPLEMENTATION FOR LARGE VOLUME, HIGH RESOLUTION MULTICHANNEL ELECTROENCEPHALOGRAPHY DATA IN SEIZURE DETECTION

TINOOSH MOHSENIN

UNIVERSITY OF MARYLAND

Ubiquitous bio-sensing for personalized health monitoring is slowly becoming a reality with the increasing availability of small, diverse, robust, high fidelity sensors. This oncoming flood of data begs the question of how we will extract useful information from it. In this paper we explore the use of a variety of representations and machine learning algorithms applied to the task of seizure detection in large volume of high resolution, multi-channel EEG data. We explore classification accuracy, computational complexity and memory requirements with a view toward understanding which approaches are most suitable for such tasks as the number of people involved and the amount of data they produce grows to be quite large. In particular, we show that layered learning approaches such as Deep Belief Networks excel along these dimensions. We also present the implementation of these algorithms on different hardware approaches including Virtex-7 FPGA, GPUs and 65 nm-CMOS ASIC.

A SURVEY AND COMPARISON OF METHODS FOR TOPIC MODELING

THOMAS H. WOTEKI

ACENTIA

Topic Models are probabilistic models for uncovering the latent sematic structure of a collection of documents. Topic models are a tool for exploring and browsing large collections of natural language documents. This paper will survey and compare topic models based on Latent Dirichlet Analysis and related probabilistic models. Techniques for validating topic models and optimizing the semantic coherence of topic models will be discussed. References to software for constructing topic models will be provided.

Thomas Woteki, PhD in Statistics, is CTO of Acentia, a provider of technology and management solutions to clients in the Federal Government. Woteki has over 30 years of experience in a variety of Federal and industry positions. His prior positions include CIO of the American Red Cross, chief engineer for the design of the SEC's EDGAR system, chief statistician in the Energy Consumption Division of the Energy Information Administration, and Assistant Professorships in the departments of statistics at University of Texas, San Antonio and Princeton University.

PLATFORMS FOR BIG DATA ANALYTICS AND VISUAL ANALYTICS AT THE CSIRO AUSTRALIA

TOMASZ BEDNARZ AND JOHN TAYLOR

CSIRO AUSTRALIA

CSIRO Computational Simulation Sciences is building collaboratively a platform with the aim to unify and standardize way of using big data frameworks inside the organization. It aims to speed up building Virtual Laboratories and Visualization Platforms, connect data analytics, statistical modelling, imaging, visualization, machine learning into one big stack ready to go solutions. Hybrid system would connect HPC with Big Data frameworks, GPGPUs and Cloud Computing. The aim is to support various compute platforms, systems in interoperable way and build decision support solutions: mobile applications and web interfaces APIs. This will dramatically increase the productivity of data intensive applications development and accelerate scientific discoveries. By providing user-friendly access to computing resources and new workflow-based solutions, it will also enable the researchers to carry out many challenging data intensive tasks that are currently impossible or impractical due to the limitations of the existing interfaces and the local computer hardware.

This presentation will showcase various projects executed by CSIRO over last couple of years, their outcomes, challenges and impact. Some of the examples to be presented include: cloud-based image analysis and processing toolbox (http://cloudimaging.net.au) executed on the National eResearch Collaboration and Resources (NeCTAR, www.nectar.org.au) infrastructure, big data computational frameworks in clouds, computational imaging and visualisation, human-computer interactions, big data analytics, visual analytics, collaborative platforms. Also, CSIRO Data Acces Portal will be demonstrated - it provides access to data published by CSIRO across a range of disciplines to enhance collaboration and data exchange in science.

Dr **Tomasz Bednarz** currently works as a Computational Research Scientist and Projects Leader at the CSIRO's Computational Informatics division. Initially, he joined CSIRO in early 2009 to work as a 3-D Visualization Software Engineer at CSIRO's Queensland Centre for Advanced Technologies in Brisbane. Then, he moved to Sydney to work on computational and accelerate science using compute clusters and GPGPUs. In 2012-13, he led the project "Cloud-based Image Analysis and Processing Toolbox" (http://cloudimaging.net.au) powered by the NeCTAR infrastructure, and now he is back to Brisbane leading the project "Platforms for Big Data Analytics and Visual Analytics". He is actively pursuing activities in the field of Computational Simulation Sciences and Visualization - his broad range of expertise spanning from image analysis/processing, through numerical simulations and experiments with fluids, visualization, computer graphics, demoscene to human-computer interactions is evidenced by the quality and number of publications - more than 90 journal and conference publications. He is a member of: the Khronos Group, IEEE Computer Society, ACM and ACM SIGGRAPH. In the past, he's received couple of awards, including prestigious Julius Career Award in 2012. More information can be found here: au.linkedin.com/in/tomaszpbednarz/

Dr **Taylor** is currently CSIRO Director of eResearch & Computational and Simulation Sciences. Dr Taylor has written more than 200 articles and books on computational and simulation science, climate change, global biogeochemical cycles, air quality and environmental policy, from the local to the global scale, spanning science, impacts and environmental policy. Dr Taylor's research has been widely cited and attracted significant media attention. Dr Taylor has worked as a Computational Scientist and group leader both at the Mathematics and Computer Science Division, Argonne National Laboratory and at the Atmospheric Science Division at Lawrence Livermore National Laboratory. Dr Taylor was Senior Fellow in the Computation Institute at the University of Chicago. Dr Taylor has served on the Advisory Panel of the Scientific Computing Division of US National Center for Atmospheric Research (NCAR) and the US National Energy Research Scientific Computing Center NUGEX Advisory Committee. Dr Taylor currently serves on

the Board of the National eResearch Collaboration Tools and Resources (NeCTAR) a federal Government SuperScience initiative. Dr Taylor is a Fellow of the Clean Air Society of Australia and New Zealand.

AN IN-DEPTH LOOK AT NOSQL

WILL LAFOREST

MONGODB

The relational database has been incredibly successful since its inception 40 years ago but Edward Codd was not thinking of distributed computing, data variability, and petabyte scale datasets. The database market is undergoing a massive period of innovation driven by the pressures of big data. The new generation of databases are non-relational and collectively referred to as NoSQL. This talk will cover the key concepts and divergent approaches in the NoSQL space while focusing on the 3 major classes of key-value, big table, and document oriented databases. How should data be distributed? What does the data model look like? How does hadoop relate to these databases? These questions and more will be addressed.

Will LaForest heads up the Federal practice for MongoDB. Will focuses on evangelizing the benefits of MongoDB, NoSQL, and (OSS) open source software in solving Big Data challenges in the Federal government. He believes that software in the Big Data space must scale not only from a technical perspective but also from a cost perspective. He has spent 7 years in the NoSQL space focused on the Federal government, most recently as Principal Technologist at MarkLogic. His technical career spans diverse areas from data warehousing, to machine learning, to building statistical visualization software for SPSS but began with code slinging at DARPA. He holds degrees in Mathematics and Physics from the University of Virginia.

MONTE CARLO SIMULATION AND THE ENTERPRISE DATA WAREHOUSE

WILLIAM CARSON

TERADATA PROFESSIONAL SERVICES

Monte Carlo simulation (MCS) is a tool frequently used by decision-makers to characterize and evaluate risk in stochastic systems. Although computationally intense, MCS sees use in finance, engineering, and many other disciplines in which performing live experimentation to quantify uncertainty is infeasible or cost-prohibitive. For example, since financial regulators cannot induce the conditions required to determine whether a given institution's portfolios are robust to challenging market conditions, stress tests are performed with MCS. Similarly, systems engineers use MCS to help plan for fleet downtime given the performance characteristics of constituent components in individual aircraft. As systems become more complex and the risks associated with their implementations more substantial, MCS is being adopted by audiences who rely upon data-driven decision-making but may lack traditional high-performance computing resources.

The enterprise data warehouse (EDW) is a primary catalyst for data-driven decision-making. In a proper EDW, detailed historical and real-time data from disparate sources are integrated to form a complete view of the

enterprise. These data help characterize enterprise processes and drivers, which inform the MCS system model and its stochastic inputs, respectively. Since the relevance of MCS output for risk management is contingent upon the adequacy of the system model and its stochastic inputs, data from the EDW is an integral part of high-quality MCS. Furthermore, the hardware, software, and integration of modern EDW are highly sophisticated and lend themselves to quickly performing analytical tasks across system nodes. MCS is both compute- and data-intensive and its demand on these pathways logically increases with the quantity of input variables and the complexity of their interactions. Enterprise-scale MCS can readily leverage the parallelized capabilities of an EDW for high-performance results, using resources already present or without capital expenditure beyond that required to implement a modern EDW.

The benefits of performing rigorous MCS on a high-performance, production EDW are many-fold. First, both enterprise data and data generated by the MCS reside on the same system, which reduces data movement and, thereby, resource consumption. The analytical tools that fit MCS input distributions to enterprise data, create stochastic variates, and convert MCS data to actionable intelligence also reside on the same system, so the performance of these tools is optimized for the EDW architecture. Furthermore, development time of the MCS itself is reduced, as SQL is the underlying language used to build the MCS and is familiar to a broad audience. Finally, substantial performance gains are realized in MCS execution time on the EDW, which grant practitioners opportunities to increase sample size and iteratively tune input parameters for higher-confidence output. These benefits allow users to allocate more effort to performing high-level tasks such as developing the system model and specifying risk preferences, both of which contribute to better decision-making under uncertainty.

William Carson is a technology and analytics consultant who serves government, finance, healthcare, and life sciences clients. His research interests lie in simulation and modeling, and he has published work in *Medical Engineering and Physics*, *Prostatic Cancer and Prostatic Diseases*, and the IEEE's *Proceedings of the Systems and Information Engineering Design Symposium*. He has M.S. and B.S. degrees in systems engineering from the University of Virginia School of Engineering and Applied Science and is currently employed by Teradata Corporation.

THE CHALLENGE OF ACQUIRING ACCURATE, COMPLETE, NEAR-PATIENT CLINICAL DATA FOR DATA SCIENCE ANALYSIS

JULIAN M. GOLDMAN, MD

MASSACHUSETTS GENERAL HOSPITAL, HARVARD MEDICAL SCHOOL

Acquiring, managing, and analyzing "Big Data" is the subject of intense national discussion. As we embark on this new era, it is important to recognize that physiologic point-of-care or "near patient" clinical measurements create unique challenges to acquiring accurate and complete data sets. These challenges include:

Multiple sources of physiological variables are available, but are not mapped to a comprehensive adopted clinical and device taxonomy: For example, Heart Rate (HR) may be derived from the electrocardiogram (ECG), transduction of arterial blood pressure continuously (IABP), intermittent measurement by automated pneumatic non-invasive blood pressure monitor (NIBP), continuous photoplethysmographic analysis (e.g. via pulse oximeter) and other means. Physiological, body site, and instrumentation variations can generate different heart-rate values in each of these channels, especially with rapid time-varying signals. In addition, device settings such as signal

averaging time, filtration, and other settings may affect the measured value. Currently, clinical data sets usually do not include meta-data that describes measurement technology and site, and signal processing attributes.

Limited data granularity: Physiologic data, such as multi-channel ECG, continuous blood pressure, lung ventilation pressures and gas concentrations, and oxygen concentration measured by pulse oximetry, are typically part of a standard monitoring regimen for critically ill and intra-operative patients. The medical device or "sensor data" may be digitized in a range of 40-200 samples-per-second (SPS) and stored temporarily for clinical review and assessment. However, electronic health records (EHR) typically store data at discrete intervals of 1-15 minutes; continuous data is not retained. In addition to the significant data that occurs from continuous signals to discrete EHR storage, clinical systems do not permit control over the processing of the continuous data stream to obtain the "single" EHR value that will represent the clinical state.

Clinical care is performed with a combination of on-body sensor data, lab data, and expert observations and assessments. While device data is suitable for high-fidelity storage, the observational and contextual data is more difficult to measure or describe, record, and interpret. Expert observers investigate the veracity of data that is incongruous with the clinical state, and may discard suspected data. Consequently, it has been said "data appears more valid the farther it is from the patient's bedside".

These are only some of the challenges in creating complete and accurate near-patient point-of-care data sets that can be analyzed with emerging data science technologies to improve the quality and safety of clinical care and foster innovation in healthcare technology.

Julian M. Goldman, MD is Medical Director of Biomedical Engineering for Partners HealthCare, a practicing anesthesiologist at the Massachusetts General Hospital, Instructor at Harvard Medical School, and formerly principal anesthesiologist in the MGH “Operating Room of the Future”. Dr. Goldman is the Director of the Program on Medical Device Interoperability at MGH/CIMIT, which is a multi-institutional federally funded program that is advancing medical device interoperability to improve patient safety. He is deeply involved with diverse efforts to improve the safety and efficacy of systems of medical devices and HIT.
Dr. Goldman completed his anesthesiology residency and a fellowship in medical device informatics at the University of Colorado, and served as a Visiting Scholar in the FDA Medical Device Fellowship Program as well as an executive of a medical device company. He is Chair of ISO Technical Committee 121, Chair of ASTM Committee F29, Co-Chair of the FCC mHealth Task Force, and serves in leadership positions of AAMI and UL standardization committees. E-card: www.jgoldman.info

BY TITLE

BY AUTHORS