Enabling Seamless Data Sharing in Industry and Academia

SPEAKERS / SEPTEMBER 29-30, 2016 / PHILADELPHIA, PA

War Stories Part I (Session chair: Florence Hudson)

Robert Cheetham, Founder and President, Azavea

Bio: Robert Cheetham is the founder and CEO of Azavea, a B Corporation that applies geospatial technology for civic, social, and environmental impact. Azavea creates advanced geospatial analysis software for infrastructure, climate change, urban ecosystems, elections, public safety and human services. The firm's public service mission and research focus are at the heart of its work including: open source projects, R&D partnerships with universities, the Summer of Maps program, the OpenDataPhilly project, and donating a portion of its profits to charitable organizations. Prior to founding Azavea, Robert served as a software developer and GIS analyst for the University of Pennsylvania, the City of Philadelphia and the Philadelphia Police Department and as a civil servant in Japan. He has an MLA in Landscape Architecture and Regional Planning from the University of Pennsylvania and a BA in Japanese Studies from the University of Michigan.

Abstract: *History of Open Data in Philadelphia: Successes, Failures and Lessons Learned*
Robert will provide a brief history of municipal open data in Philadelphia, culminating in the development of the OpenDataPhilly web site and events since it's release. The talk will also describe areas in which OpenDataPhilly has both succeeded and failed as well outline thoughts on areas in which Philadelphia's experience represents an alternative model to that of other municipal governments and the potential for future academic collaboration.

Jason Bobe, Director of the Sharing Lab, Icahn Institute and Dept. of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai

Short Bio: Jason Bobe, Associate Professor, Director, Sharing Lab, Icahn Institute for Genomics and Multiscale Biology, Mount Sinai School of Medicine

Long Bio: http://www.jasonbobe.net/about.html

Title: "What if Your Biology Holds the Key that Protects Others from Disease? Changing the Discourse around Sharing Health Data"

Abstract: The protection of personal health and medical data has been recognized as an important goal for decades. The societal value of sharing data is immense, but to-date paid much less attention. Designing a biomedical research enterprise that provide individuals access to their own

data and improved options for sharing is paramount for addressing critical social concerns like better health, new therapies & strategies for disease prevention.

Joe Chaya, Director, Project Management for Global Data Warehousing and Technology Innovation, IMS Health

Bio: Currently leading a Global Project Management Office for the IMS Health Data Warehousing organization. Key areas of focus are on the deployment of Big Data technologies and technology integration associated with merger and acquisition activities. Served as Chair of the Industry Advisory Board for the National Science Foundation's Center for Visual and Decision Informatics. Responsible for outreach with academic institutions to seek synergies between IMS Health and academic research to further technology innovation. Career spans nearly 3 decades in information technology focused on making healthcare perform better in the acute care hospital and pharmaceutical industries.

Abstract: Data sharing challenges and IMS Health

- IMS Health – Who we are and what we do
- Types of data sharing sensitivities
- Case studies for real world data sharing challenges

John Lee, Senior Associate, Osage Partners

Bio: John is a Senior Associate at Osage University Partners focusing on infrastructure software, enterprise software, and hardware. At Osage, John has been actively involved in Algorithmia, Body Labs, Clarifai, Embodied, Jibo, Kymeta Corp, Paracosm, Seismos, and Streetlight Data.

Prior to Osage, John spent time at Lux Capital and was a manager in the investment arm of Mahindra & Mahindra, one of Asia's largest conglomerates, where he worked on diversified investments and in-house startups in technology. Previously, John co-founded two startups and served as the Director of Cover Africa, a non-profit organization working on a malaria prevention research project in rural Ghana.

In his spare time, John is an outdoor enthusiast and avid rock climber, surfer, and cyclist. John holds a BS from Cornell University.

IP, Policy, Legal issues (Session chair: Tim Kraska)

Anita Eisenstadt, Research Staff Member, Science and Technology Policy Institute

Bio: Anita Eisenstadt is a Research Staff Member at the Science and Technology Policy Institute (STPI) in Washington DC where she brings legal, international, and science policy expertise to analysis of complex issues for the White House Office of Science and Technology Policy and other Federal science agencies. Prior to joining STPI, Ms. Eisenstadt served as a Senior Foreign Affairs

Officer at the U.S. State Department where she promoted international science and technology policy cooperation and served as an expert on biotechnology research, intellectual property rights and the public domain. She led the U.S. delegation to the OECD Committee on Science and Technology Policy and its Working Party on Biotechnology. From 1990 to 2004, Ms. Eisenstadt served as an Assistant General Counsel at the National Science Foundation where her portfolio included Federal data policies, international law, research compliance, export controls, environmental law, employment law, grants and contracts, and legislation. Ms. Eisenstadt has also served as a Senior Marketing Attorney for Volkswagen of America. Ms. Eisenstadt holds a JD from Wayne State University Law School and a BA in Anthropology and Asian Studies from the University of Michigan.

Abstract: Anita Eisenstadt from the Science and Technology Policy Institute will provide an overview of legal mechanisms that can be used to promote sharing and use of research data. Legal interoperability is the ability to combine data from two or more sources without conflicts among restrictions on use of the data and without the need to seek authorization from the data provider on a case-by-case basis. Ms. Eisenstadt will describe how copyright and other forms of intellectual property may protect research data. She will then describe mechanisms such as agreements, policies, waiver of copyright, or use of common licenses that may be used to promote data sharing and maximize legal interoperability. Finally, she will provide some examples of international governmental and private initiatives that may be useful models for future data sharing efforts.

Paul Ragusa, Intellectual Property Law Partner, Baker Botts

Bio: Paul Ragusa is a partner in the Intellectual Property section of Baker Botts LLP in New York. His practice encompasses high technology and Hatch-Waxman patent litigation, patent portfolio management, counseling and licensing. His diverse technical experience spans audiovisual compression and transmission, semiconductor device fabrication, nanotechnology, software, techniques for delivering pharmaceutical products and medical devices. He was part of the team that drafted the patent licensing agreements directed to the MPEG-2 digital video compression standard, and continues to focus on Standards Setting Organizations (SSOs) and related licensing, including 802.11, ATSC, AVC, Blu-ray, DASH, HEVC, LTE, MVC and UHD-BD. He serves on the Advisory Board at *IP Litigator*, Chairs the American Intellectual Property Law Association's Antitrust Law Committee, and serves on the International Institute for Conflict Prevention and Resolution's (CPR) Intellectual Property Disputes Committee, using those positions to address the practical implications of important changes to the law. Mr. Ragusa holds a JD from the George Washington University Law School, and a BS in Physics from Binghamton University.

Abstract: Paul Ragusa will address licensing and IP issues that arise in the context of data sharing in industry and academia. He will provide an overview of standard data licensing agreements, including common approaches to such agreements and typical sections. He will then focus on commonly encountered pitfalls in negotiating data licensing agreements, including defining the data to be licensed, providing restrictions on use, addressing service and technical support issues, balancing accountability and remedies for breach, and treating intellectual property resulting from results or improvements which use the licensed data. The use of institutional agreements and policies to protect data will also be addressed.

Infrastructure Challenges and Opportunities (Session chair: Jane Greenberg)

Greg Madden, Senior Advisor for Research Computing and Cyberinfrastructure, Penn State

Bio: Greg Madden is Senior Advisor for Research Computing and Cyberinfrastructure for Penn State University, where his role is to identify and eliminate every point of friction in the research computing ecosystem across the university. Serving as an advocate for the research faculty, Mr. Madden's primary areas of focus at Penn State include high performance computing, advanced research networking, data centers, software and data acquisition and licensing, and research data governance. Mr. Madden has 30 years of experience in higher education, having worked at seven different colleges and universities in various positions ranging across the spectrum of IT, including programming, customer service, systems administration, network administration, and management. He holds an MA in Geography from the University of North Carolina at Chapel Hill and a BS in Computer Science and Philosophy from the University of Alabama.

Title: Towards Seamless Data Sharing

Abstract: Universities face a wide variety of interwoven problems on the road to a future of seamless data sharing. On the technology side, we take as given that data transfer and storage concerns can be solved with enough capital and ingenuity. More challenging are issues of cost, inter-institutional trust, service governance, and institutional prioritization; some of these will be more amenable to easy solutions, some less. Finally, there are a whole suite of issues surrounding research data governance, beginning with basic terminology, data classifications, and data ownership, and extending through more complex issues such as how institutional approaches to data sharing are encoded into policies, and how those policies relate to compliance and the role of the research administrators.

Sam Madden, Professor and Director of CSAIL, MIT

War Stories II (Session chair: Florence Hudson)

John Brzozowski (Fellow and Chief IPv6 Architect, Comcast) and Yana Kane-Esrig (Senior Principal Engineer, Comcast)

Bios: A proven technology leader John has been the driving force behind Comcast's IPv6 transition and deployment since the program began in 2005. His accomplishments and success leading Comcast's IPv6 deployment are recognized around the globe. John's tireless efforts continue as he advances the adoption of IPv6 across all of Comcast's infrastructure and services. He leverages his expertise and experience to drive the adoption and implementation of IPv6 in the real world and at scale, ensuring that innovative solutions are in place to support traditional and next generation services. His pioneering work has had significant impacts not only within Comcast but also across

the entire cable industry and beyond. He has and continues to be instrumental in encouraging the adoption of IPv6 across the broadband, content, and consumer electronics ecosystems.

Yana Kane-Esrig currently works for Comcast as a Senior Principal Engineer. She focuses on the analysis of Quality of Service and Quality of Experience data in the context of the High Speed Data Services. Her objective is to obtain and communicate the insights that help Comcast to reliably and cost-effectively deliver on its promises to its customers and to offer new products to the market. Yana received her Ph. D. degree in Statistics from Cornell University and her B. S. E. in Electrical Engineering and Computer Science from Princeton University.

Title: Data analysis projects: gotchas, remedies, and wish list

Abstract: We will discuss the common challenges in internal and external data analysis projects in the corporate world.

In the internal projects, problems arise due to unfamiliarity of participants with their roles in data analysis. For a project to be successful, the internal customer must be engaged in specifying the business objectives. The organization that provides the data must have the data documented and provide access to the Subject Matter Experts with the domain knowledge. The data engineers and the data analyst need to ensure correct communications among themselves about the specifications of the data set that is being extracted and integrated from multiple sources. The data analyst needs to ensure integration of these different roles into the project team.

The academic community can help by providing the training to the people who will be called to play the all these different roles in the data analysis projects, so they understand the responsibilities of their own role and how it fits into such projects. In addition, tools are needed to better enforce documentation of data, including the meaning of the data, how it is collected, discontinuities, abnormal values.

Jane Greenberg, Alice B. Kroeger Professor, Director of the Metadata Research Center, College of Computing & Informatics, Drexel University

Bio: Jane Greenberg is the Alice B. Kroeger Professor and Director of the Metadata Research Center at the College of Computing & Informatics, Drexel University. She serves on the advisory boards for the Dublin Core Metadata Initiative (DCMI) and the U.S. Research Data Alliance (RDA). She is 2016 ELATE Fellow, and was a 2014 Data Fellow at the National Consortium for Data Science, and held a 2012 Chair of Excellence at the University of Carlos III-Madrid, Spain. She is the principal investigator of Metadata Capital Initiative (MetaDataCAPT'L) and the HIVE (Helping Interdisciplinary Vocabulary Engineering) linked data initiative, and was co-Principal Investigator for the Dryad data repository from 2008-2015, overseeing metadata R&D and data curation.

Title: Dryad: An Industry-Researcher-Non-Profit Data-Sharing Partnership

Abstract: Dryad (http://datadryad.org/) is a "a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable." Dryad's success is grounded in stakeholder engagement during the conceptualization and planning stage. This

presentation will provide a brief overview of Dryad's start-up phase and day-to-day operation. Attention will be given to Dryad's JDAP-joint data archiving policy, governance, non-profit status, and licensing, and how these aspects support Dryad's sustainability and growth.

Industrial Requirements (Session chair: Jane Greenberg)

Kareem Aggour, Senior Computer Engineer, GE

Bio: Kareem Aggour leads a team of researchers within GE Global Research's Knowledge Discovery Lab, when he develops novel solutions to address GE's Big Data challenges. Kareem graduated with honors from the University of Maryland, College Park with B.S. degrees in Electrical Engineering and Computer Science. Upon graduation Kareem joined GE, and in 2000 Kareem transferred to GE Global Research in Niskayuna, NY, where he has worked ever since. While working, Kareem attended Rensselaer Polytechnic Institute (RPI) where he completed an M.S. in Computer Engineering. Kareem is currently pursuing a PhD in Computer Science at RPI, also part-time. Kareem is a member of the IEEE. He has 28 publications and holds 12 patents.

Abstract
We highlight two recent examples of external data sharing from the Knowledge Discovery Lab at GE's Global Research Center. The first use case entailed sharing an Industrial Big Data dataset with a commercial software vendor to evaluate their software on our data. The second use case entailed sharing a set of biomedical images with a university to explore a partnership around developing image preprocessing routines. Finally, we conclude with general thoughts and considerations regarding our experiences with data sharing.

Tony Orsini, Director of Decision Sciences, Experian

Bio: Tony has 30 years of management and operational experience crossing the Financial Services and Marketing Services business lines. His responsibilities range from managing a staff of highly dedicated statisticians and technicians to participating both sales and strategy meetings. This range allows the group to stay connected at all levels of the organization and thus leads to more accurate and timely solutions for our customers.

Tony joined Experian in June of 1996 and has been involved in many of the Marketing Service's areas. Initially he ran the computer services department responsible for address hygiene and merge/purge. He also managed the Experian's Catalog Coop, Business to Business Coop and Targeting Acquisition modeling. Currently, Tony is responsible for analytical development and operations for Targeting and Digital Advertising.

Prior to Experian, Tony was a manager at SEI a large financial services company. There, he led a team of programmers specializing in 401K management software and Pension payment systems. Clients included Fidelity Investments, Chase, CitiCorp, John Hancock, Marriott Corp, Transamerica.

Abstract: We will discuss a recent partnership between a large, metropolitan Fire Department and Experian to improve community outreach efforts by local fire departments to decrease fires and associated loss of life. Experian worked with FEMA NFIRS data and referential data to analyze,

model and standardize datasets for use. In addition, Experian created an iPad application to distribute this data and collect additional survey data for further analysis / reporting. We will conclude with our thoughts on what makes for successful partnerships between the private sector and academia / public sector.

City Perspective (Session Chair: René Bastón)

Amen Ra Mashariki, Chief Analytics Officer, City of New York

Bio: Amen Ra Mashariki is the Chief Analytics Officer for New York City and the Director of the Mayor's Office of Data Analytics (MODA).
Dr. Mashariki began his career as a software engineer at Motorola , then worked as a Computer Scientist and Research faculty at the Johns Hopkins University Applied Physics Laboratory. In 2012 he was appointed by President Barack Obama to serve as a White House Fellow. After his year as a White House Fellow Dr. Mashariki was appointed as the first Chief Technology Officer for the Office of Personnel Management in the federal government.
Dr. Mashariki holds a Doctorate from Morgan State, M.S. in Computer Science from Howard University, and a B.S. in Computer Science from Lincoln University