Press Release I September 28th, 2016

Northeast Big Data Innovation Hub Awarded $3.3 million to Create Solutions to Pressing Challenges in Health, Education and Data Sharing

The National Science Foundation (NSF) has announced $3.3 million in grants to researchers affiliated with the Northeast Big Data Innovation Hub. A common theme underlying the projects below is the need to address both technology and other challenges to fully exploit the value of big data, across priority areas ranging from education to health care.

The Northeast Big Data Innovation Hub is one of four consortia set up by NSF in 2015 to develop ambitious data-driven solutions to substantial societal challenges. Coordinated by Columbia University, the Northeast Hub builds partnerships across companies, government agencies and nonprofit organizations, including universities, across the region. The Northeast Hub awards, summarized below, grew out of months of discussion among group members, representing a community-driven effort to identify, articulate, and act on data innovation priorities for the Northeast region. Further information about the Hub's mission and activities may be found at http://nebigdatahub.org.

A Data Sharing Licensing Model and Platform
Principal Investigators: Jane Greenberg (Drexel), Tim Kraska (Brown,) Samuel Madden (MIT)
Co-Principal Investigators: Carsten Binnig (Brown), Daniel Weitzner (MIT)

Big discoveries await those who mine big data, but concerns about security, privacy, legal challenges, and other policies frequently limit how much data can be shared, and can even derail the data sharing promise. To remove these barriers, the above researchers propose a coordinated licensing model and data-sharing platform, ShareDB, which will automatically enforce licensing terms.

A "Northeast Data Sharing Group" made up of industry leaders and academic researchers will ensure that ShareDB is widely accessible in all fields, from health care to finance. Unlike Creative Commons and other data sharing platforms, ShareDB will allow data owners to share their data, with some

limitations, and keep sensitive information safe. Access controls, secure and validated metadata, and audit logs will provide additional mechanisms to automatically enforce and monitor the terms of licensing agreements.

Advancing Big Data Literacy
Principal Investigator: Stephen Uzzo (New York Hall of Science)
If big data is a language, what counts as fluency? The New York Hall of Science will take the lead in helping to define the skills and knowledge that make someone data-science literate. The results of this planning project will include a plan for creating, implementing and evaluating programs to teach data science at all academic levels. It will be developed with input from Northeast Hub members. Among the project's goals is to identify high-quality education resources as well as gaps that should be filled.

Overcoming Data Privacy and Security Challenges
Principal Investigators: Adam Smith (Penn State), Rebecca Wright (Rutgers)

Safeguarding individual privacy is essential to unlocking big data's potential without compromising the people behind the numbers. This planning project will include two workshops hosted by Rutgers' Center for Discrete Mathematics and Computer Science (DIMACS) will explore how individuals and data owners can gain more control. The first, "Overcoming Barriers to Data Sharing," will bring together experts to understand how modern privacy demands limit data sharing. The second workshop will explore privacy and security challenges facing Northeast Hub affiliates. The insights gained from these events will inform the design of future projects aimed at addressing the most important challenges identified.

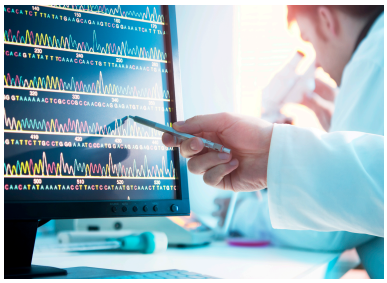
Combating Cyber Threats with Collaboration
Principal Investigator: John Yen (Penn State)
Co-Principal Investigators: Vijayalakshmi Atluri (Rutgers), George Cybenko (Dartmouth), Peng Liu (Penn State), Andrew Sears (Penn State)

Cyber attacks threaten companies, government agencies, and nonprofits alike, posing substantial risks to privacy and security and imposing heavy costs. As hackers grow more sophisticated, organizations are trying to beef up their defenses and develop ways for intrusion-detection analysts to secretly share information across organizations, even while an attack is underway. Led by John

Yen, the above researchers will organize a workshop and customer-driven planning activities to develop collaborative cyber defense solutions.

As a demonstration, the researchers will build a platform allowing intrusion-detection analysts at Penn State, Rutgers, and Dartmouth to post a Python script to another organization's cyber security data repository, facilitating coordinated, cross-organization defenses during an ongoing cyber attack. These represent the first steps in designing a platform that can substantially improve the ability of institutions across sectors to rapidly detect cyber attacks.

Mining Environmental Data to Find Links to Human Health
Principal Investigators: Gregory Cooper (U. Pittsburgh), Noemie Elhadad (Columbia), Vasant Honavar (Penn State), Chirag Patel (Harvard)


Vast repositories of information currently held in hospital, government, and industry databases could provide extraordinary insights into the influence of poverty, air pollution, climate, and other factors on human health. However, much of that data is locked in silos.

The above researchers will integrate and anonymize patient records with databases tracking weather, climate, and air pollution, as well as income, occupation, and other demographic trends. The researchers will use machine-learning techniques to hunt for causal links between risk factors and health outcomes. They will also lead a workshop and provide other training materials on how to use the integrated health data set for research and training in biomedical data science.

Improving Teaching and Learning with Big Data
Principal Investigators: Ivon Arroyo (Worcester Polytechnic Institute), Ryan Baker (U Penn), Beverly Woolf (U. Mass, Amherst)
Co-Principal Investigator: Neil Heffernan (Worcester Polytechnic Institute)

Education software has given researchers valuable new insights into how


students learn best, provided valuable feedback to teachers, and enabled personalized instruction tailored to the needs of individual students. Despite this progress, the dissemination of these powerful tools is limited by a lack of training in how to most effectively use these resources. The above investigators will introduce researchers and teachers to these data-driven techniques in a series of competitions, hackathons and workshops, sharing valuable insights and enabling participants to come away with the practical tools and techniques needed to improve classroom outcomes.

Energy Innovation and Economic Development
Principal Investigator: Abani Patra (U. Buffalo)

Long thought of as a Rust Belt town, Buffalo is reinventing itself for the digital age. The city is currently redeveloping former industrial areas to attract technology companies with massive data processing and storage needs. Patra, based at the University of Buffalo, will lead a series of workshops to explore ideas and partnerships addressing the data and information needs tied to providing energy for the city's brownfield redevelopment project. The workshops will also provide a template for the Northeast Hub's Energy Workgroup to develop projects that address important challenges in the energy ecosystem.