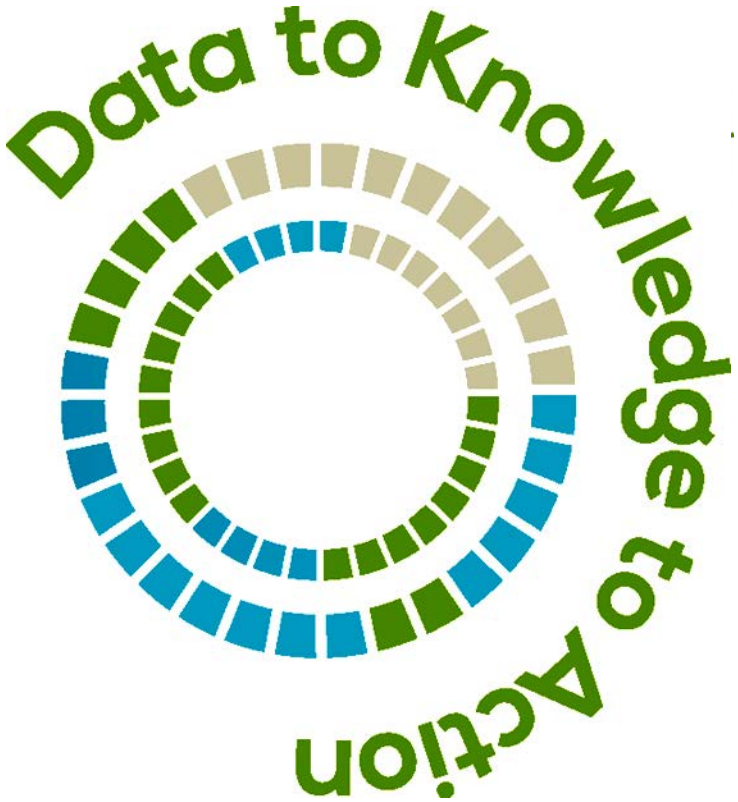


Building New Partnerships

Tom Kalil

Deputy Director for Technology and Innovation
Office of Science and Technology Policy

November 12, 2013

"The challenges we face today – from saving our planet to ending poverty – are simply too big for government to solve alone. We need all hands on deck."

Empowering the Patient, Curing Diseases, and Saving Lives

Data at the Forefront of Healthcare

Transforming cancer care and patient outcomes

American Society of Clinical Oncology

- Clinical trials generate most of the information used to improve patient outcomes, but make up only 3% of cancer care
- CancerLinQ™, **every patient's experience will help improve cancer care**
- ASCO has demonstrated a working prototype and phased **development of CancerLinQ is now underway**, with first tools expected by early 2015

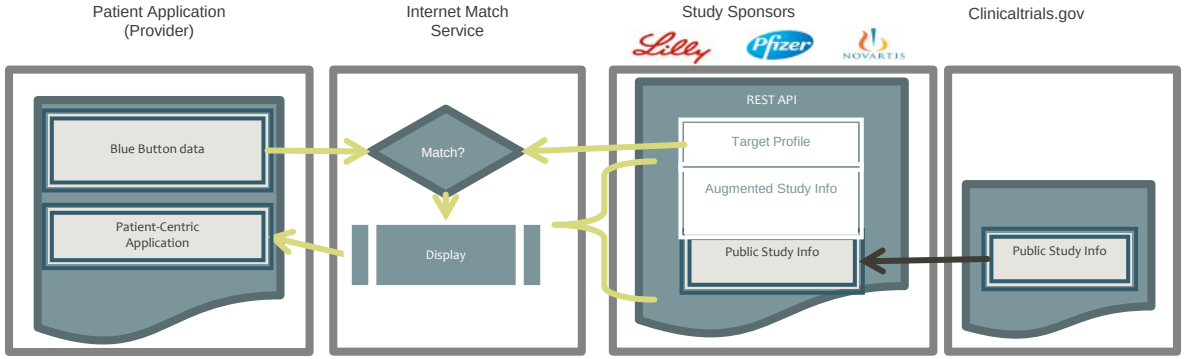
ASCO CANCER·LINQ™
Learning Intelligence Network for Quality

- This is a **$80 million project** over the next 5 years; currently **$27.8 M has been committed** by ACSO and multiple other organizational and individual supporters

Improving access to data about clinical trials

Novartis, Pfizer, and Eli Lilly and Company

- New platform will help connect patients and researchers by making **more and more useful data about clinical trials** available
- Provides machine readable **"target health profiles"** for trials
- One application of the platform allows patients to search for trials using their own **Blue Button data**
- **Platform is open** on both ends (to sponsors of clinical studies and to software companies) and is set **to launch in early 2014**

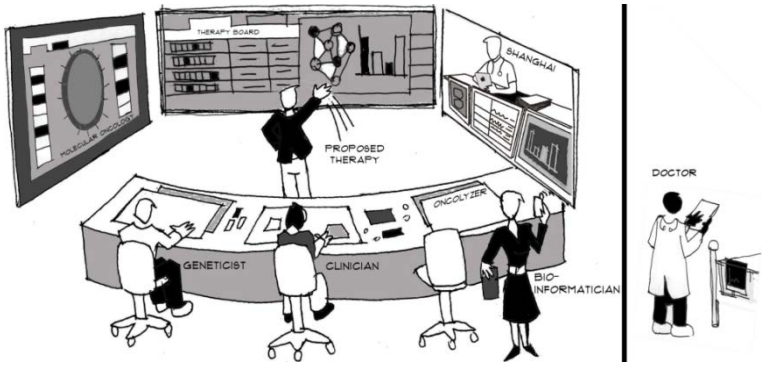


Accelerating real-time personalized medicine

SAP, Stanford School of Medicine, and the National Center for Tumor Diseases Heidelberg

SAP HANA Platform for healthcare

Real-time analytics of clinical and genomic data for breakthroughs in patient diagnostics, treatment and research



- Stanford seeing **17-600X** faster computations in genomics data analysis to ultimately benefit new treatments for autism, cardiovascular disease
- NCT analyzing clinical and genomics data in real-time to predict disease risk and match patients to clinical trials

SAP

STANFORD SCHOOL of MEDICINE

NCT NATIONAL CENTER FOR TUMOR DISEASES HEIDELBERG

New methods for early detection of heart disease

NIH, IBM, Sutter Health, and Geisinger Health System

- Supported by a $2 million joint research grant from the NIH
- Will both develop and apply new data analytics technologies to **help detect heart failure sooner than currently possible**
- Will create **practical and cost-effective** early detection methods for primary care practices using electronic health records

New methods for medical imaging

Rutgers University and the State University of New York, Stony Brook

- The NSF sponsored Center for Dynamic Data Analytics (CDDA) brings together **industry and academia** on multi-stakeholder research projects
- Partnership with BioClinica has developed **novel algorithms** to assess new therapies for **liver, spleen, and muscular diseases.**
- One disease studied, sarcopenia, **impacts 45% of the older U.S. population**

- The Scalable Indexing Project is developing techniques for storing and organizing **high-bandwidth streams of data** that can be queried in real time

Taking ownership of our personal health data

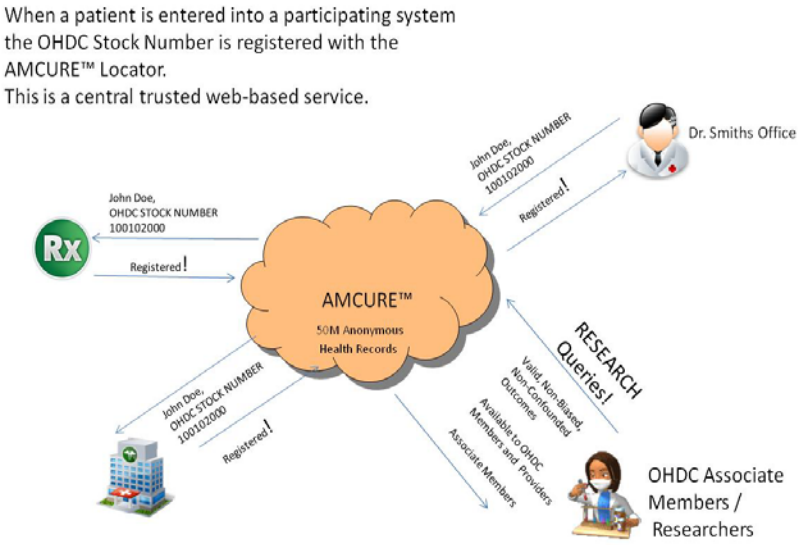
Our Health Data Cooperative (ODHC)

- A new cooperative developed to help patients **better manage and directly benefit** from the use of their health data while providing it **anonymously for clinical research uses**

- Enrollment is **open to all U.S. residents** with a goal of 10 million members enrolled by 2016

- North Carolina at Chapel Hill and RTI International will help build and design database access and educational infrastructure

- Partnership with Touchstone Energy Cooperative to enlist their 30M existing members



The next generation of genomic sequencing

DC-NET, CAAREN and George Washington University Medical Center

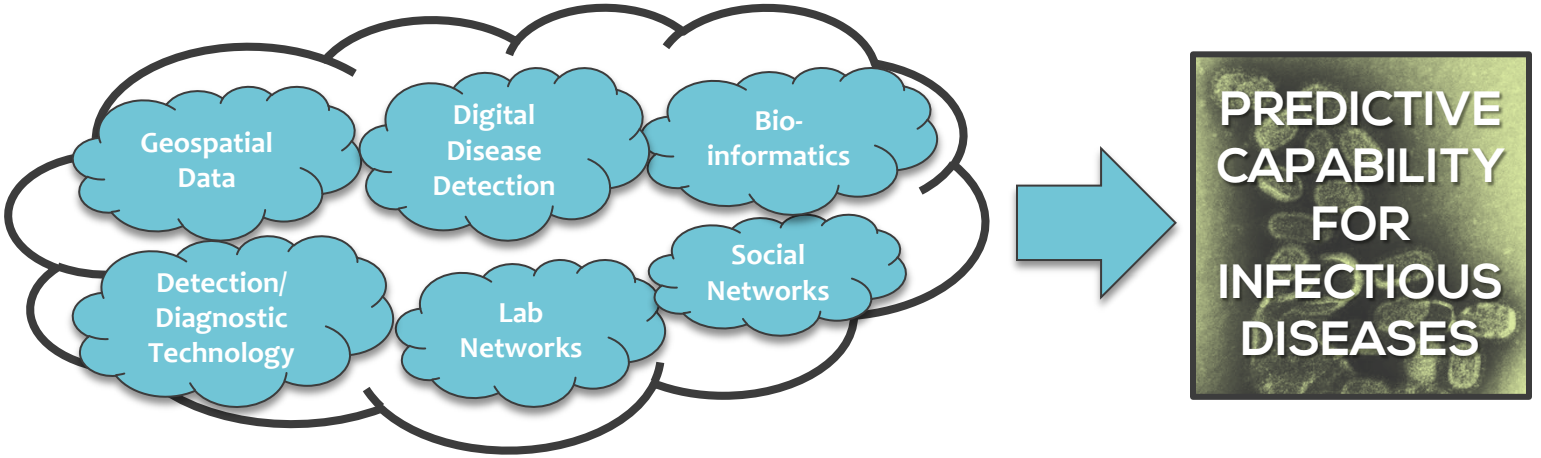
- DNA and RNA genomic sequences can be used **identify biomarkers and new targets for therapies** to treat diseases like cancer and heart disease.
- Next-generation sequencing machines produce **terabytes of data** and disease analysis require crunching data from **thousands of patients**
- GW has invested in **a massive new High Performance Computing cluster** for their genomic and other advanced research initiatives
- Partners are helping to create a **wide-area network** to communicate research data between the cluster, GW's multiple campuses, The St. Laurent Institute, and at the researchers' homes in the DC area.

A "weather service" to predict the next pandemic

Office of Science and Technology Policy

- **A new consortium of communities of interest** including government, NGOs, academic institutions, industry partners and others

- **A pilot project** will explore the drivers of biological events and determine how big data can be used to predict pandemic potential of novel infectious agents.



Disseminating open health data around the world

MedRed and BT Global

- MedRed BT Health Cloud (MBHC) is **a new international project** aimed at enhancing the integration and dissemination of open health data
- This multiyear, transatlantic effort makes available one of the **largest open health data repositories** anywhere in the world
- Includes several years of **de-identified population health** data encompassing all of England and Wales
- **U. S. public data sets**, such as the FDA adverse event reporting data, recently released Medicare data, from the Center for Medicaid Services, and data from other healthcare systems are currently being added

Growing the Economy

Data Driving Regional and National Development

Helping new businesses get off the ground

City of New York, Mayor's Office of Data Analytics

- The NYC Mayor's Office of Data Analytics (MODA) is **working across city government** to use city data to improve daily operations, help to prepare for and respond to disasters, and support economic growth

- Now working with **NYC New Business Acceleration Team** (NBAT) help new restaurants cut through red tape and open their doors to customers

- Using data on construction permits (Department of Buildings), restaurant inspections (Department of Health and Mental Hygiene), and NBAT counseling notes to see how free NBAT services can **cut a new business' time to open**

Engaging the financial sector on market stability

The Center for Technology in Government and the Institute for Financial Market Regulation, University at Albany, SUNY

- A workshop to establish a research agenda focusing on the **stability, openness and fairness of financial markets**
- To be held in DC on Nov 14, 2014 with government, academic, and industry attendees
- Will tackle the **information sharing and collaboration challenges** created through data complexity and volume, the dynamic and interconnected nature of markets, financial instruments, technologies, and institutions working within regulated markets and the regulatory process

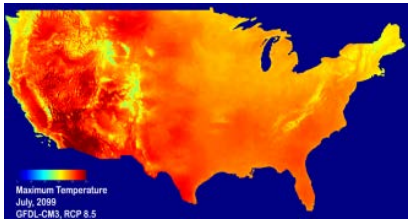
Supporting the Earth, Energy Use, and the Environment

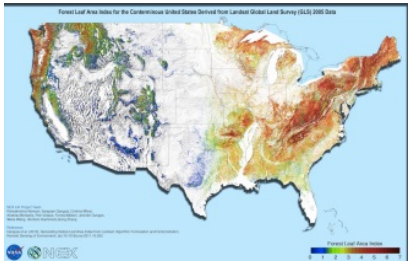
What Data can do Today to Help Sustain the World for Future Generations

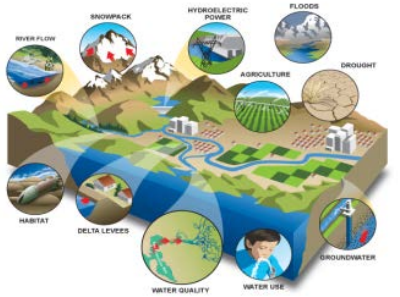
Bringing Earth data to the public

Amazon and NASA

- Amazon Web Services (AWS) and NASA are providing a significant amount of NASA's **Earth science data and models to the public** as an AWS Public Data Set.
- Gives everyone access to data and analytic techniques **previously only available to NASA researchers**
- Enables calculation of the next **National Climate Assessment** on the AWS cloud
- Hosting NASA NEX data in the cloud also enables **crowd-sourced citizen science** applications like those found on Zooniverse (zooniverse.com).







Protecting the world's forests in real time

Google and World Resources Institute

- Uses the Google Earth Engine to power Global Forest Watch 2.0, a **forest monitoring tool** that uses data from satellite imagery, monitoring systems, and mobile technology to provide **near real time information** on forests globally

- Will include **crowd-sourcing capability** so people on the ground can report deforestation when it takes place
- Project will be launched later this year

Creating a sustainable city

San Diego Supercomputing Center, Clean Tech San Diego, OSIsoft, and San Diego Gas & Electric

- Deploy a data infrastructure that connects physical systems (those managing **electricity, gas, water, waste, buildings, transportation and traffic**) in **San Diego**
- Aims to drive city-scale applications **that decrease electricity consumption and cost**, **discover and anticipate grid instabilities**, and educate the public
- Results will be published to **lead other communities on path to sustainabily**

Empowering the Nation and the World

The Community and the Public Use Data to Solve Society's Greatest Challenges

Top industry talent take on non-profits' data challenges

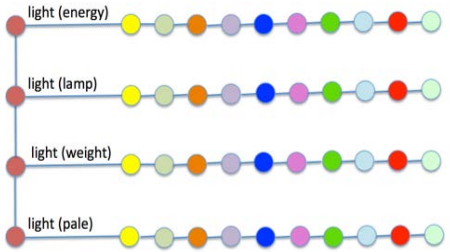
DataKind, Pivotal, Teradata, Medic Mobile, and the Mission Continues

- DataKind partners with Pivotal to bring **industry's top data analytics talent** to bear on society's greatest challenges currently being tackled by **non-profit organizations**; also exploring similar partnership opportunities with Teradata
- Partnership with The Mission Continues which to better understand the effects of their volunteer programs on **improving veterans' lives**
- Partnership with Medic Mobile to quantifiably measure the impact of their many health initiatives helping **under-served and disconnected communities** around the world

Translating every word in every language

The Kamusi Project, Long Now Foundation, Global WordNet, University of Ngozi, Swiss Federal Institutes of Technology

- The Kamusi Project created the Global Online Living Dictionary (GOLD), a system to **capture the full range of human linguistic data** from **all 7000 languages** spoken around the world and make it available for education and for technologies such **as speech recognition and automatic translation**



Multilingual big data: possible

- With EPFL in Switzerland, Kamusi is designing a unique **crowdsourcing system**
- Users who share their knowledge will earn **free access to the data**, a "play-to-pay" system that will include social networking and competitive gaming elements

Competing to find the best data solution

MIT and the City of Boston

- The Big Data Initiative at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) is **organizing a series of competitions** designed to spur innovative use of data to address major societal issues

- Defines **real-world challenges** in different areas such as transportation, health, finance, and education

- The first MIT Big Data Challenge, in collaboration with the **City of Boston**, to launch in November 2013, will make multiple data sets available to understand **public transportation patterns in the city**

Innovating civic engagement across the country

Splunk, NoticeandComment, the Sunlight Foundation , and Regulations.gov

- There are millions of federal documents and hundreds of millions of local notices posted as part of the **government decision making process**
- Splunk4Good will create a new, public interface that enables users to explore the federal regulatory data through **real-time graphs, dashboards, and visualizations**
- The Sunlight Foundation's Docket Wrench tool will be expanded using natural-language and machine-learning techniques to visualize **authorship, sentiment and content** of dockets
- NoticeandComment.com has developed a new publishing platform using the framework of Regulations.gov as their model for **all local governments to use**

Advancing Core Technologies

The Cutting Edge of Data Science and Engineering

Making leading edge technologies available for all

Berkeley's Algorithms, Machines, and People Laboratory (AMPLab)

- AMPLab is releasing their Spark software for lightning fast cluster computing on an **Apache Open Source License**

- 10 to 100x faster than traditional Map-Reduce, Spark is being used by **major industrial collaborations**, including Yahoo! and ClearStory Data. New startup **Databricks** will support the extensive new Spark ecosystem

- The AMPLab is supported by a combination of public and industrial sponsors, the National Science Foundation, DARPA, and **twenty-one industrial partners** including Founding Sponsors Amazon Web Services, Google, and SAP.

Supercomputing with the fastest real world data

SGI and Fedcentric Technologies

- The capacities of standard enterprise servers are increasingly unable to keep up with real world data streams
- Working with Fedcentric Technologies, SGI has created cost efficient commodity high density computing systems to manage **high velocity data**, in particular to address **high-speed detection, processing, and analysis for fraud prevention**
- This technology enabled the U.S. Postal Service to process the mail stream in real time and handle fraudulent postage, stolen meters, duplicate stamps, short pay postage prior to delivery
- In the next phase USPS will look to **dynamically route mail, to enable expedited shipping, delivery and service**

Collaborating to set benchmarking standards

The Big Data Top100 List

- **Objective benchmarks** for determining **application-level performance** and **price/performance** of Big Data systems are needed to foster competition and innovation

- The Big Data Top100 List is a new **open, community-based big data benchmarking initiative** coordinated by a board of directors from the San Diego Supercomputer Center, Pivotal, Cisco, Oracle, Intel, Brocade, Seagate, NetApp, Mellanox, Facebook, IBM, Google, and the University of Toronto

- As a part of this effort, the National Institute for Standards and Technology (NIST) has recently funded researchers at the San Diego Supercomputer Center to **study different strategies for data generation for big data**

Educating the Next Generation of Data Scientists

Our Nation's Thriving Data Workforce

Harnessing the potential of data for scientific discovery

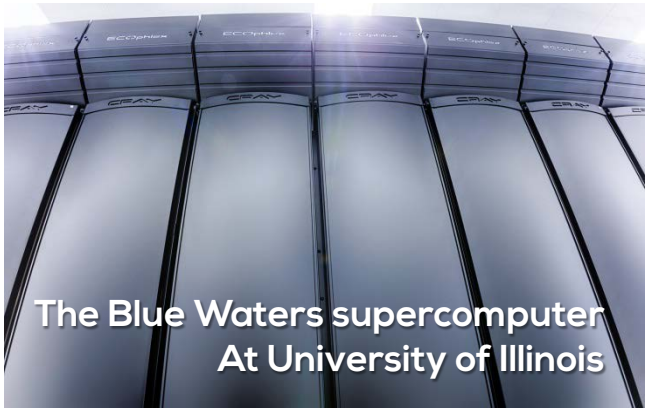
NYU, Berkeley, University of Washington, Gordon and Betty Moore Foundation, and Alfred P. Sloan Foundation

- While data science is already contributing to scientific discovery, substantial systemic challenges need to be overcome to **maximize its impact on academic research**
- **Collaboration is a five year, $37.8 million**, cross-institutional effort
- Hopes to **establish models** that will dramatically accelerate this data science revolution.
- Partnership is a **coordinated, distributed experiment** involving researchers at these leading universities

Developing Big Data as an engineering discipline

University of Illinois and the Grainger Foundation

- Built on a $100M gift from The Grainger Foundation as part of the **Grainger Engineering Breakthroughs Initiative**
- New multidisciplinary course in Big Data in Spring 2014, to be a **core class for all Engineering undergraduates** in 2016
- **13 new interdisciplinary senior faculty hired**



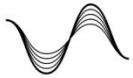
The Blue Waters supercomputer
At University of Illinois

- **Interdisciplinary Master of Engineering degree** planned for 2015
- Research teams formed around Sustainability, Energy, and the Environment; Health Sciences; Social Sciences

Engaging students on data science for social good

University of Chicago and The Schmidt Family Foundation

- The 2014 Summer Fellowship will connect undergraduate and grad students from around the country with **non-profits and government partners** on high impact problems
- Fellows tackle high impact data problems in health care, energy, education, disaster relief, transportation, and city services
- For institutions interested in **founding similar programs**, the fellowship will hold a workshop on November 15, 2013 in Chicago.



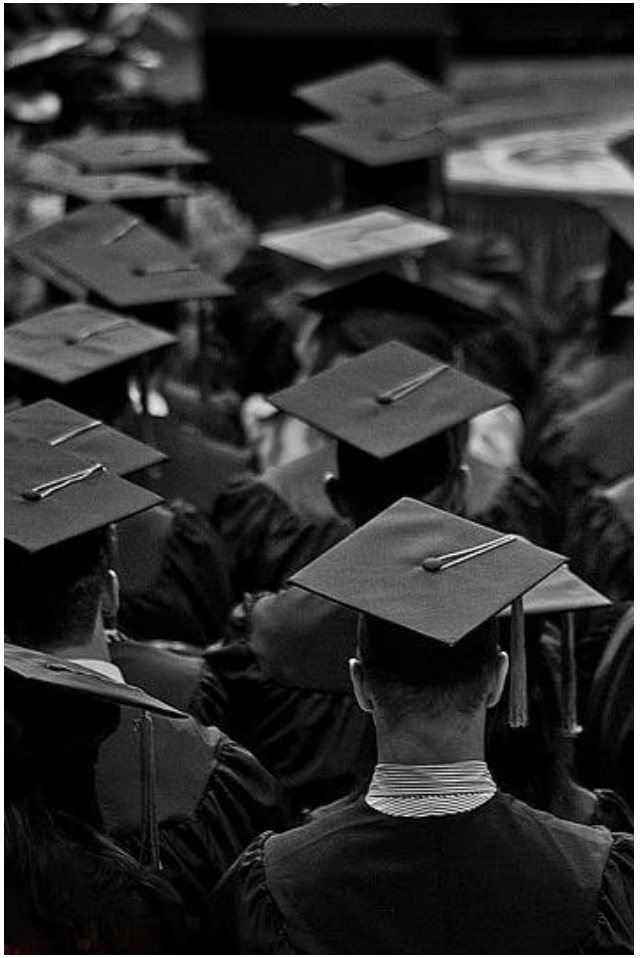
The Eric & Wendy Schmidt
Data Science for Social Good
Summer Fellowship

Positioning graduates for the most in-demand data jobs

IBM

- IBM will unveil a new **Analytics Talent Assessment**, a first of its kind online tool that university students can use to **gauge their readiness for public and private sector data-crunching jobs**, and, in turn, gain guidance on ways to further develop and position themselves for in-demand data analytics jobs

- The new assessment tool is the result of a **public and private sector collaboration** between IBM, academia and companies across multiple industries



Shaping the Future of a Data Driven Society

The Benefits and Implications of Our Data Use Today and Tomorrow

Understanding the social and cultural implication of data

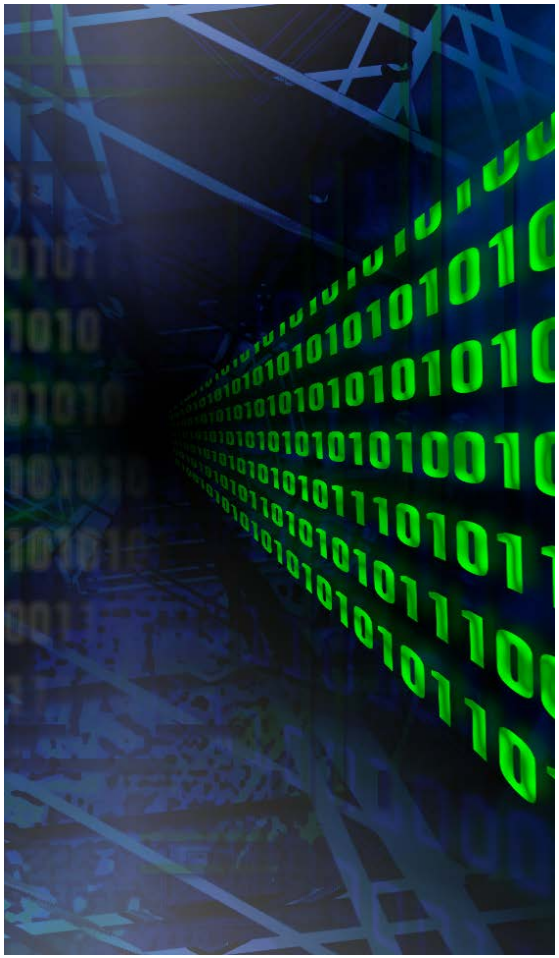
Council for Big Data, Ethics, and Society

- Council of scholars from diverse fields will focus on **social, ethical, legal, and policy issues** surrounding the Big Data phenomenon; co-Directed by danah boyd, Geoffrey C. Bowker, Kate Crawford, and Helen Nissenbaum
- **Will promote and conduct research**; develop new and powerful paradigms for leading edge ethical and political issues
- **Will work alongside technologists** guarding against known mistakes and inadequate preparation
- Will engage with **practitioners and law-makers**, participate in public deliberation, coordinate and host events.
- Will support **new research connections and bridge** academia, industry, government, and NGOs

Defining the role of technology in protecting privacy

MIT Big Data Initiative

- Bring together stakeholders from academia, industry, government and non-profits to focus on the future of Big Data and privacy
- The group will work toward collectively **articulating major privacy challenges and developing a roadmap** for future research needs
- The goal is to better understand and help **define the role of technology in protecting and managing privacy in the long term**, in particular when large and diverse data sets are collected and combined



Data stakeholders coming together across the nation

TechAmerica Foundation

- TechAmerica Foundation hosted a series of **Big Data Roadshows** this Fall in Silicon Valley, California; Austin, Texas, and Boston, Massachusetts

- Brought together industry representatives, senior federal and state government officials and academia; highlights the impact that Big Data is having on both the **public health and energy sectors** through insightful **analysis and case studies** from the private sector and government.

Where do we go from here?

- How can we **catalyze and celebrate** new and expanded partnerships to make the most of Big Data and data analytics?

- How can we **share information** on the progress that is being made?

- How can we **increase the number** of individuals and organizations (agencies, companies, foundations, non-profits, universities, states and cities, early adopters) that are involved?

Building New Partnerships