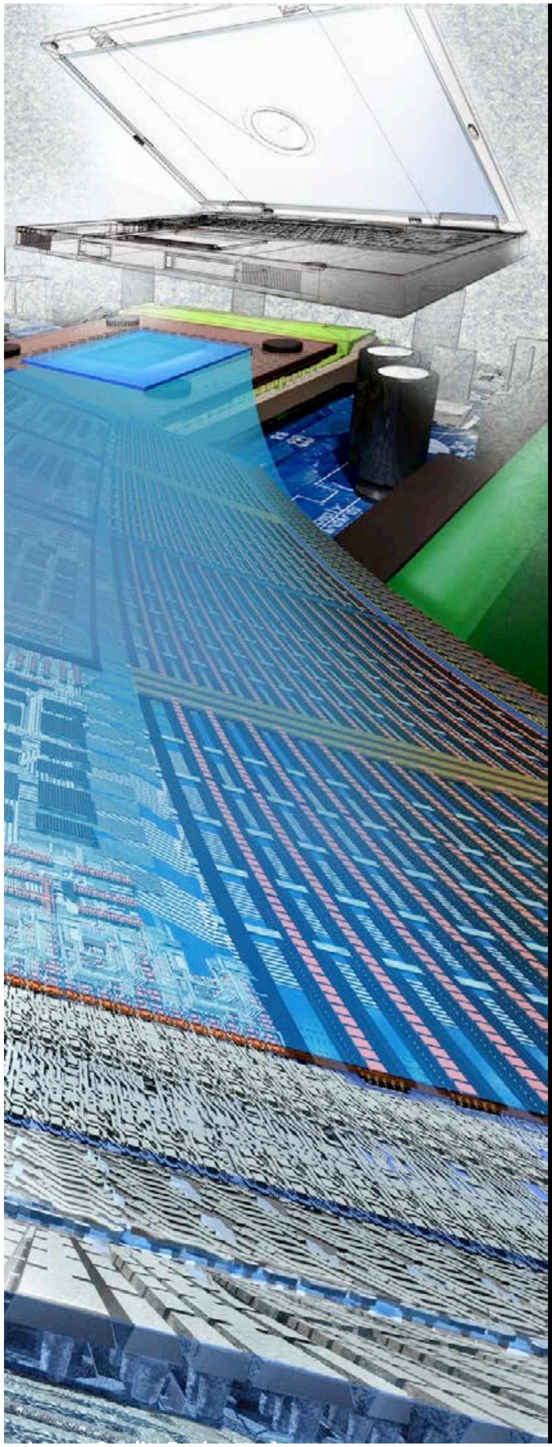


Accelerating Innovation in Big Data: From Data to Knowledge to Action

Farnam Jahanian
National Science Foundation

Ontology Summit Symposium
April 28, 2014

The Promise of Big Data

Advances in information technologies are transforming the fabric of our society, and data represents a transformative new currency for science, engineering, education and commerce.

Image Credit: CCC and SIGACT CATCS

Seizing the Big Data Revolution

- **Data Tsunami: Explosive Growth in Size, Complexity, and Data Rates**
  - Enabled by experimental methods, observational studies, scientific instruments, simulations, email, videos, images, click streams, Internet transactions … and sensors everywhere!
- **The Age of Data: From Data to Knowledge to Action**
  - Widespread use of data to create actionable information leads to timely and more informed decisions and actions.

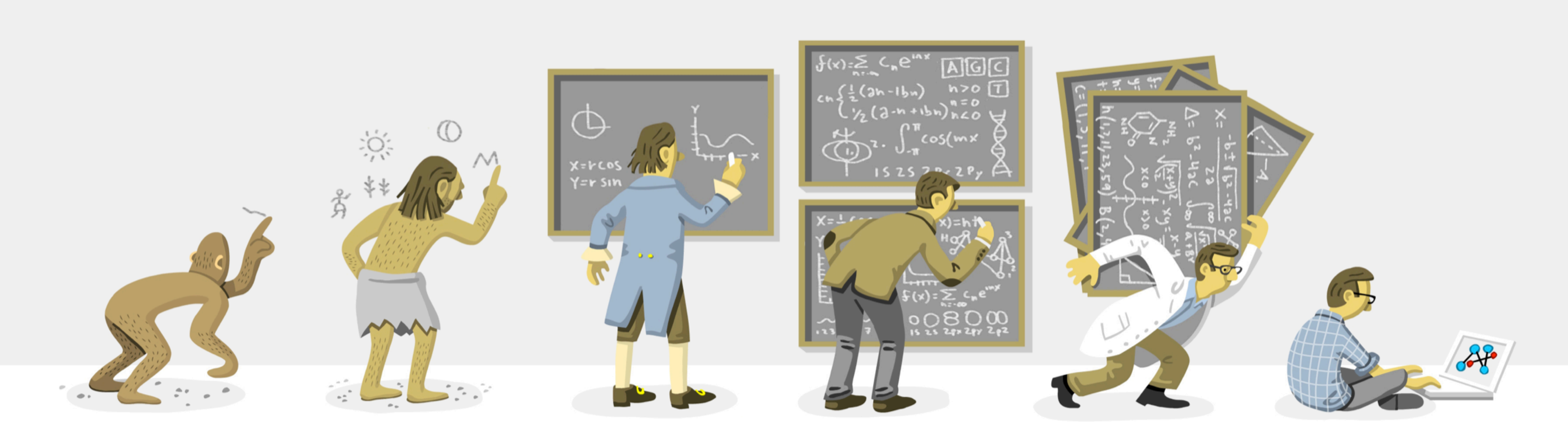


Image Credit: *Chi Birmingham*

Big Data … “the next frontier for innovation, competition and productivity”

[2011 McKinsey Report]

Why is Big Data Important?

- Transformative implications for **commerce and economy**
- Critical to accelerating the **pace of discovery** in almost every science and engineering discipline
- Potential for addressing some of **society's most pressing challenges**

Data-driven Discovery and Innovation Addresses Societal Challenges

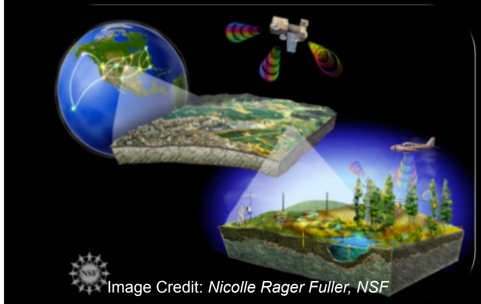


Image Credit: *Nicolle Rager Fuller, NSF*

Environment and Sustainability

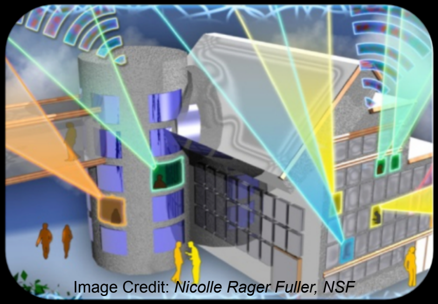


Image Credit: *Nicolle Rager Fuller, NSF*

Energy Production and Delivery

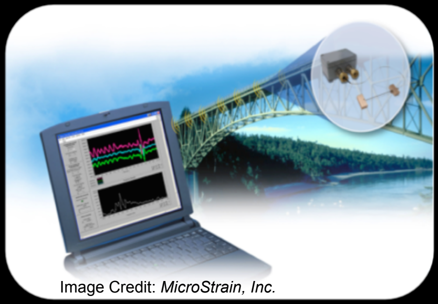


Image Credit: *MicroStrain, Inc.*

Manufacturing and Smart Systems

Image Credits: *Texas A&M University*

Emergency Response and Disaster Resiliency

Image Credit: *ThinkStock*

Secure Cyberspace

Health and Wellbeing

Image Credit: *Cisco, Inc.*

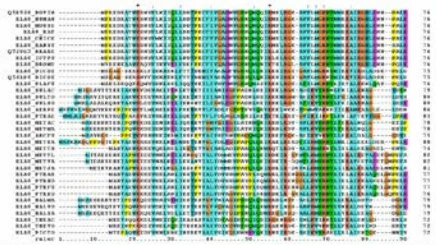
Transportation

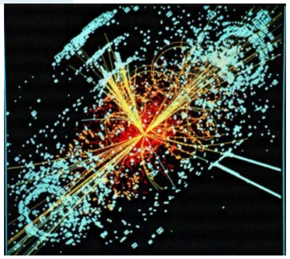
Image Credit: *Georgia Computes! Georgia Tech*

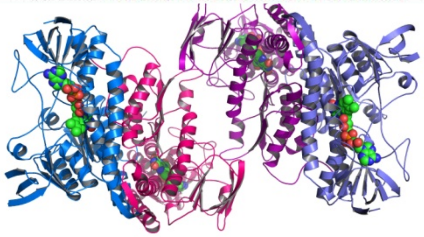
Education and Workforce Development

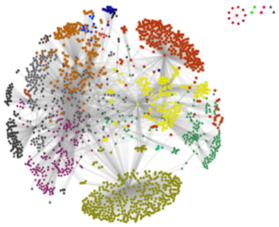
Era of Data and Information

Scientific Data









Digital Media

VIDEO

EMAIL

BLOGS

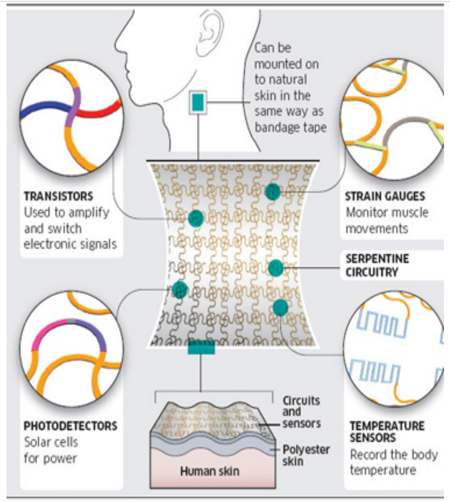
MOBILE

VOIP

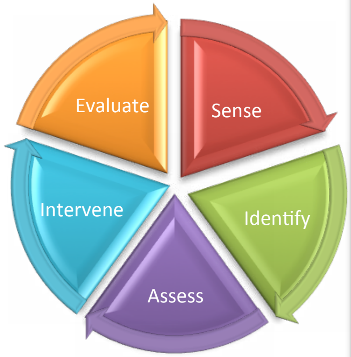
IM

Human Sensors

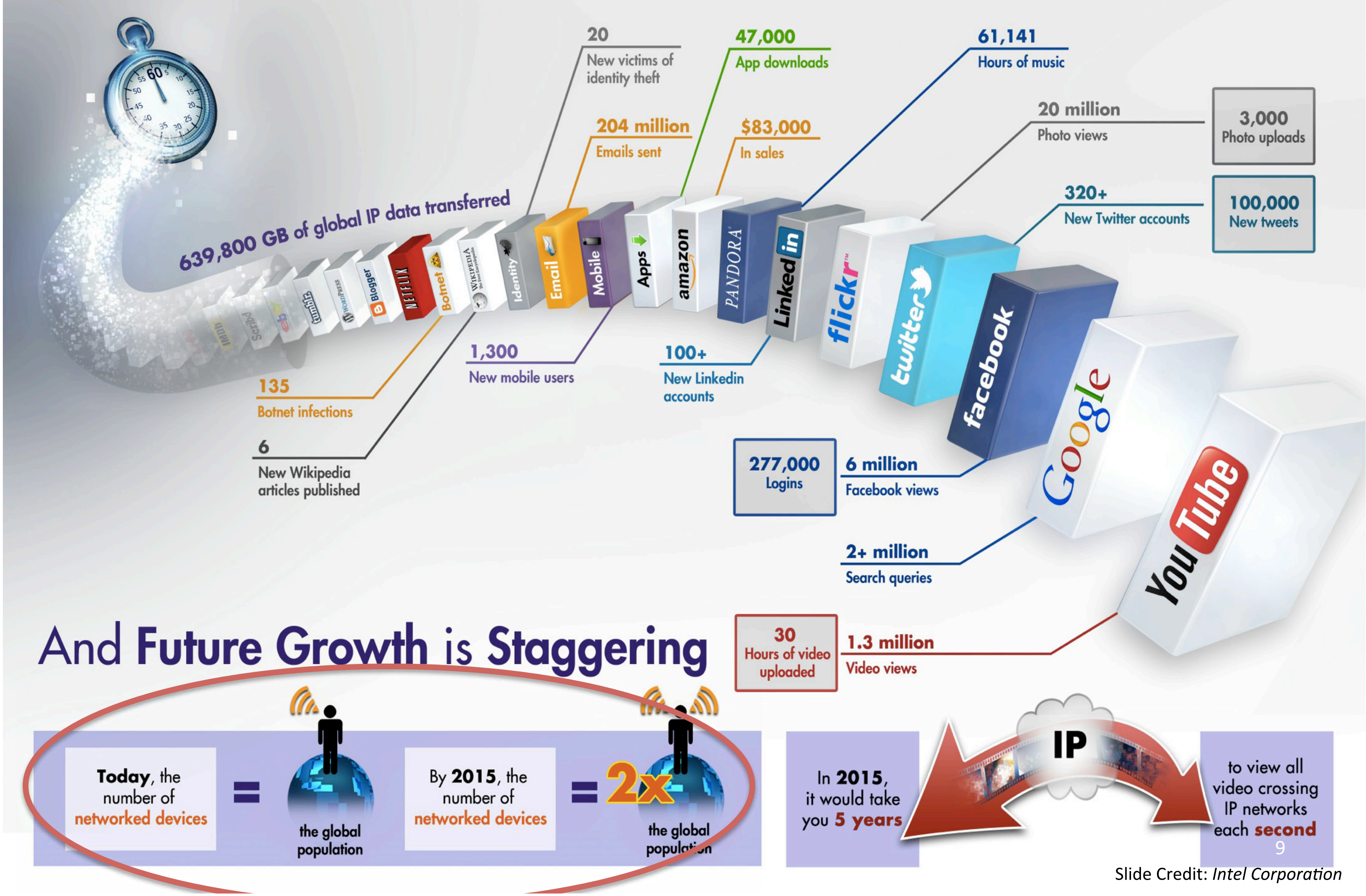
Health Care


Can be mounted on to natural skin in the same way as bandage tape
TRANSISTORS Used to amplify and switch electronic signals
STRAIN GAUGES Monitor muscle movements
SERPENTINE CIRCUITRY
PHOTODETECTORS Solar cells for power
Circuits and sensors
Polyester skin
Human skin
TEMPERATURE SENSORS Record the body temperature

Sources: Sciencemag.org, Department of Electrical and Computer Engineering, University of Wisconsin



What Happens in an **Internet Minute?**



639,800 GB of global IP data transferred

20
New victims of identity theft

47,000
App downloads

61,141
Hours of music

204 million
Emails sent

$83,000
In sales

20 million
Photo views

3,000
Photo uploads

320+
New Twitter accounts

100,000
New tweets

135
Botnet infections

6
New Wikipedia articles published

1,300
New mobile users

100+
New Linkedin accounts

277,000
Logins

6 million
Facebook views

2+ million
Search queries

30
Hours of video uploaded

1.3 million
Video views

And **Future Growth** is **Staggering**

Today, the number of networked devices = the global population

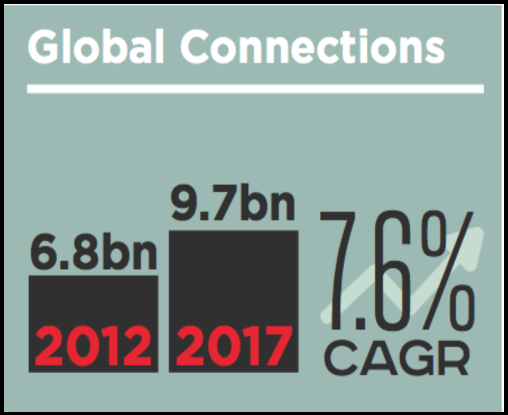
By **2015**, the number of networked devices = 2x the global population

In **2015**, it would take you **5 years** to view all video crossing IP networks each **second**

Slide Credit: *Intel Corporation*

Mobile Devices & Cellular Networks are Pervasive

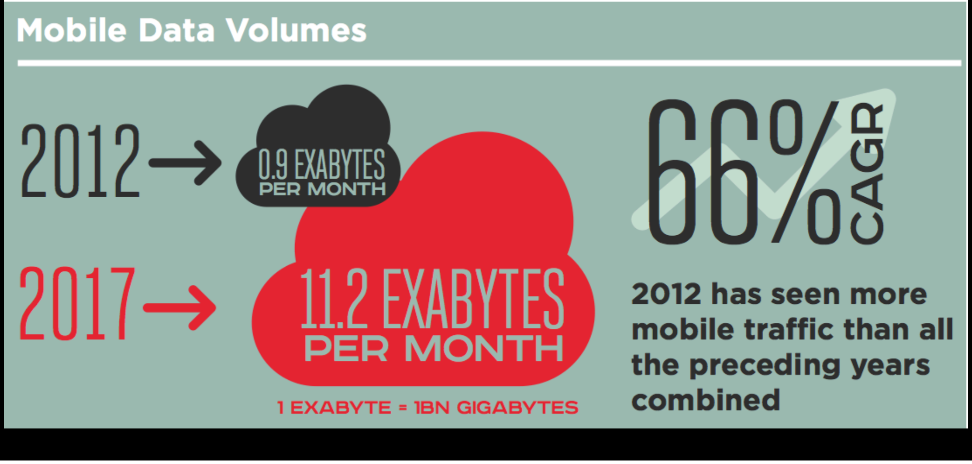
The number of mobile-connected devices will soon exceed the number of people on earth.



vs

World Population
7,145,407,06

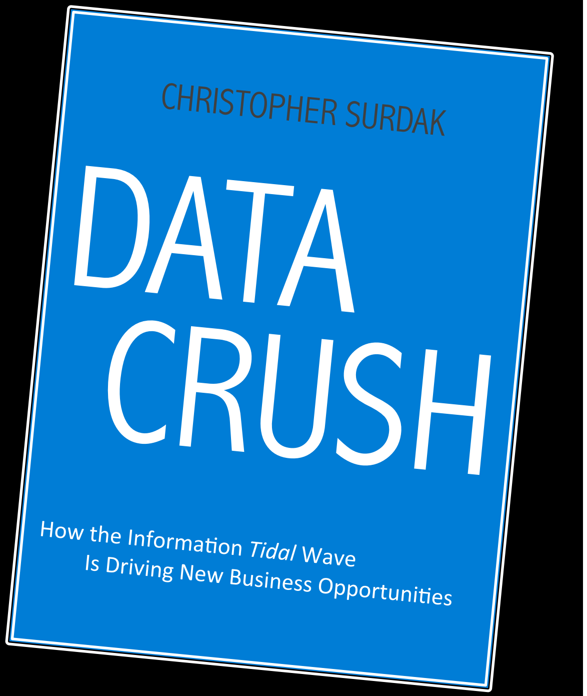
Mobile data traffic will grow at a compound annual growth rate (CAGR) of 66 percent from 2012 to 2017, reaching 11.2 exabytes per month by 2017.



2012 has seen more mobile traffic than all the preceding years combined

Mobility and Data Growth

- **Pervasiveness and Connectedness:** Mobile devices have eclipsed traditional computers as points of entry for the Internet – "always-on always connected"
- **Mobile Data Services and Applications:** Smartphones able to access data in a wide variety of forms from text messaging and web browsing to apps for everything and steaming services
- **Contextual Computing:** Apps that take advantage of a user's presence in time and space, combined w/ other relevant data, to deliver targeted and tailored content and services
- FUTURE: SMART EVERTHING and CONNECTED



CHRISTOPHER SURDAK

DATA CRUSH

How the Information *Tidal* Wave
Is Driving New Business Opportunities

Social Media and Data Growth

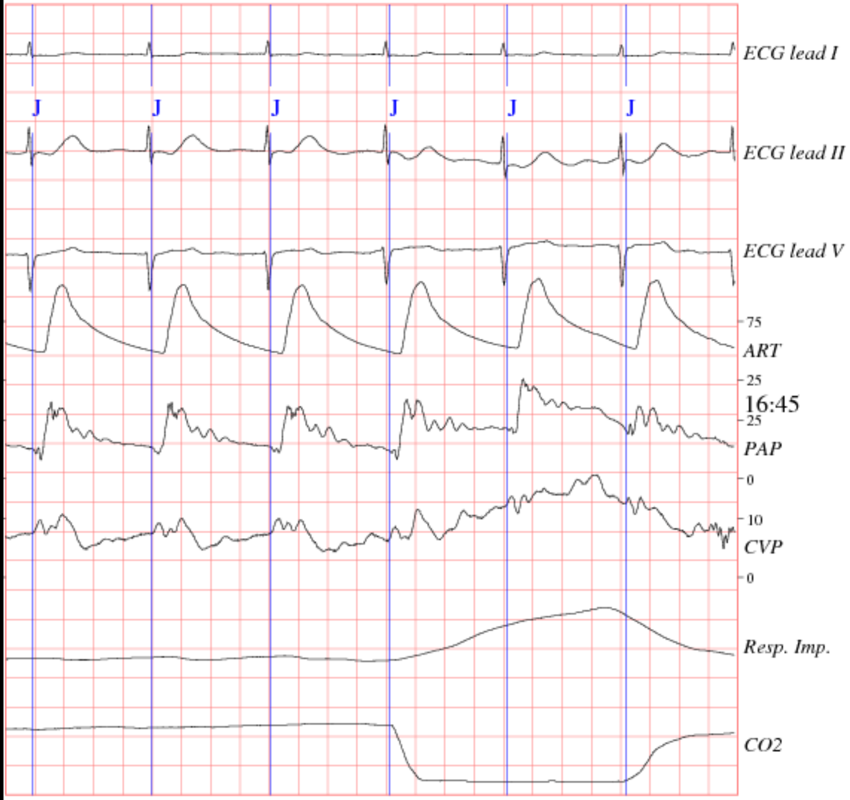
- How could companies, such as Facebook and Twitter, that provide free service, be worth billions of dollars?

Monetizing User Data

- Selling data to third parties: Targeted advertising; detailed profiling of customers' behaviors, opinions, and preferences; predictive analysis
- OLD vs. NEW marketing: This level of intimacy is potentially intrusive but highly valuable
- OLD CRM systems vs. NEW customer engagement: Consumers' expectations are changing, expecting much deeper engagement with companies and personalized service
- New business models enabled by online transactions and predictive data analysis: *e.g.*, eBay, Amazon, Groupon, Bonobos, Target

Era of "Big Data" in Healthcare

- Large volumes of data currently collected
  - EHRs and PHRs
    - Multi-scale and multi-source
  - During hospitalizations
    - For safety and diagnosis
  - On an out-patient basis
    - Typically event monitors
  - Via ubiquitous mobile sensors
    - Behavior, physiology, environment
  - As part of clinical studies
    - To evaluate safety and efficacy
  - From growing body of scientific
    - knowledge in biomedical research literature



- Gigabits/patient/day
  - High sampling rates
  - Multiple signals
- Accumulating data is getting easier, but using data is hard

Wireless Health

GPS

EEG

Pulmonary Function

SpO_2

Posture

ECG

Gait

Blood Pressure

Balance

Step Height

Step Size

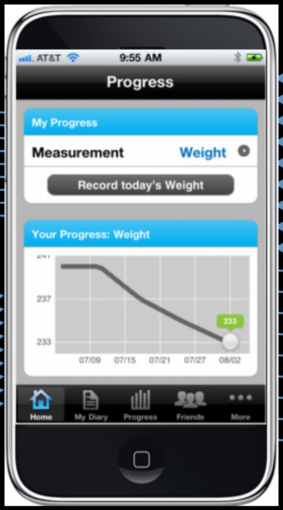
Training

Coaching

Chronic Care

Social Networks

Health Information



AT&T 9:55 AM

Progress

My Progress

Measurement Weight

Record today's Weight

Your Progress: Weight

Performance

Prediction

Early Detection

Decision Support
Population Statistics
Epidemiology
Evidence

Inference

Datamining

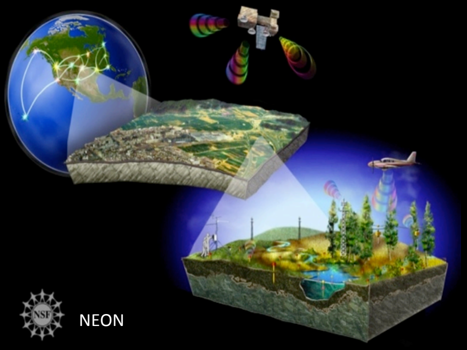
Paradigm Shift: A patient-centered framework for health and wellness

Conceptualizing Big Data for Science

- Science gathers data at an ever-increasing **rate** across all **scales** and **complexities** of natural phenomena
- Sloan Digital Sky Survey in 2000, collected more data in its 1st few weeks than had been amassed in the entire history of astronomy
  - Within a decade, over 140 terabytes of information collected
- The Large Synoptic Survey Telescope due in Chile in 2016 will amass double that quantity of data **every week**

Sloan Digital Sky Survey telescope. Credit: Fermilab Photo

LSST



NEON

From Hypothesis-driven to Data-driven Discovery

The Economist, The data deluge and how to handle it: A 14-page special report (Feb 25, 2010).

The Fourth Paradigm: Data-Intensive Scientific Discovery (2009, Microsoft Corporation).

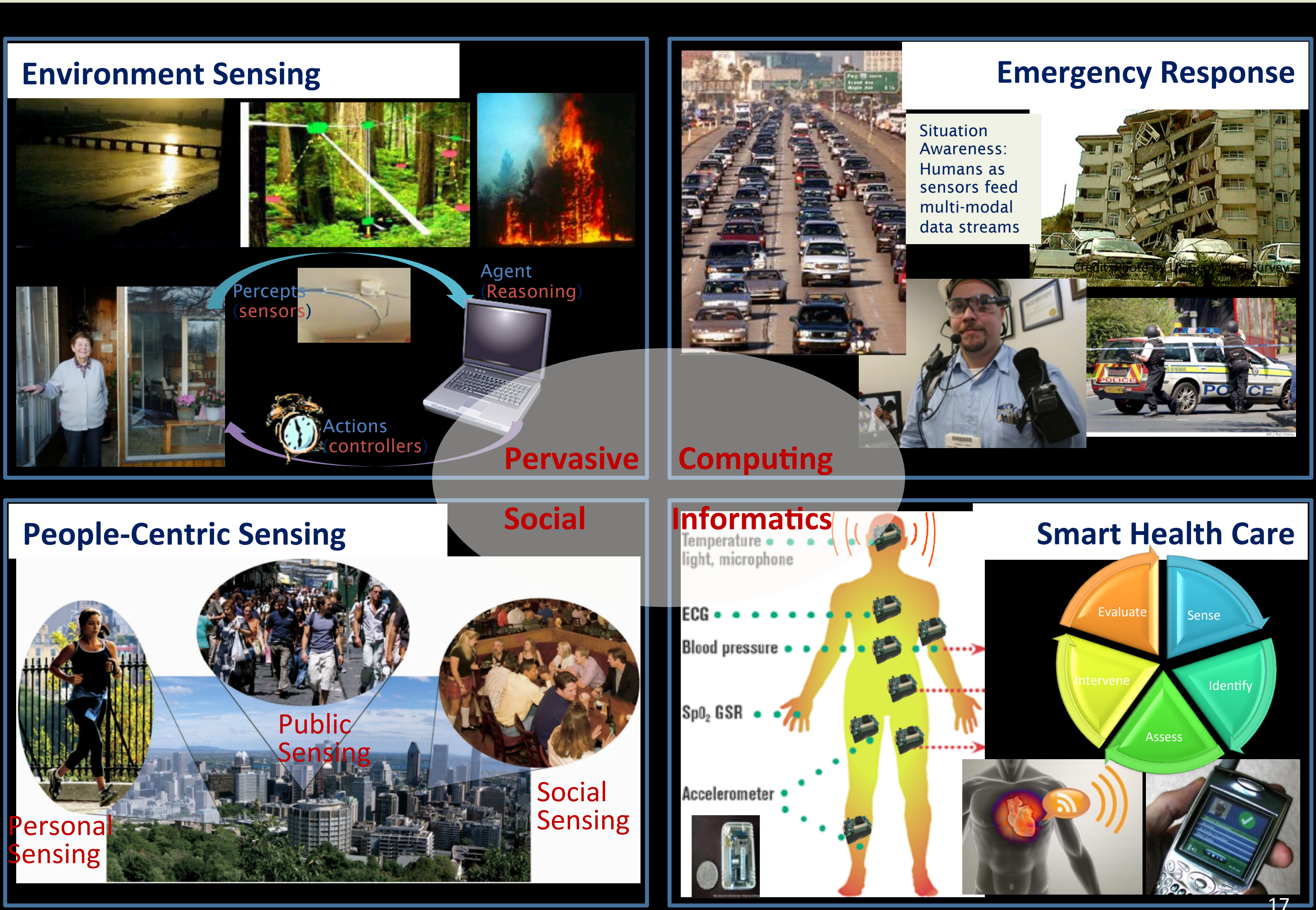
http://www.sciencemag.org/site/special/data/

http://www.economist.com/node/15579717

http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Data are motivating a profound transformation in the culture and conduct of scientific research.

Smart Sensing, Reasoning and Decision



Environment Sensing

Emergency Response

Situation Awareness: Humans as sensors feed multi-modal data streams

Credit: Photo by US Geological Survey

People-Centric Sensing

Smart Health Care

Source: Sajal Das, Keith Marzullo

> **“Our ability to generate data far exceeds our ability to digest it!”**

Too Big to Ignore: The Business Case for Big Data
By Phil Simon

What is possible?

From **data** to **knowledge** to **discovery** by

- Enabling **extraction of knowledge** from very large, heterogeneous data sets
- Providing novel approaches to **driving discovery** and **decision-making**
- Yielding increasingly more **accurate predictions**, potentially saving lives
- Providing deeper understanding of **causal relationships** based on advanced data analysis

What’s Different? Why Now?

Why Now?

Confluence of Social, Technical and Policy Interests

- Decades of advances in technology
- Plummeting costs of computation, communication and storage
- Consumer technologies, broadband connections and social media
- Data is no longer regarded as static:
  - now a raw material or a corporate asset, used to created economic value, and foundation for new business models
- Platform economy and innovation ecosystem
- Increasing transparency of democratic governance (open gov)
  - Public access to high value datasets (data.gov)
- Democratization of data and tools

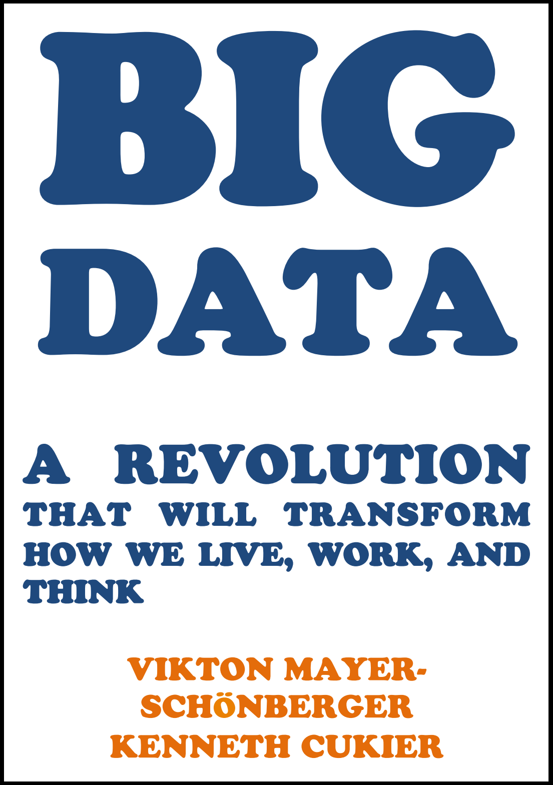
Moore's Law
Kryder's Law
Pervasive Sensors
Data Mining
Machine Learning
NL Understanding
Info Retrieval
Computer Vision
Video Analytics
Data Visualization
Crowd Sourcing
Social Networks
…

Characteristics of Big Data

- Structured vs. unstructured
- Metadata: data about data
- Data inside the organization vs. outside generated by always-on consumers, ubiquitous sensors, and social media
- Incomplete, fragmented and long-tailed
- Multiple format, heterogeneous
- …

Three Shifts in the Way We Analyze Information

- **More:** ability to collect, manage and analyze far more data rather than be artificially limited by sampling
- **Messy:** loss of accuracy and exactness at micro-level, but gain insight at the macro-level
- **Good enough:** discover patterns and correlations rather than causality



BIG DATA

A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK

VIKTON MAYER-SCHÖNBERGER
KENNETH CUKIER

http://big-data-book.com

Federal Big Data R&D Initiative

A National Imperative

REPORT TO THE PRESIDENT AND CONGRESS

DESIGNING A DIGITAL FUTURE: FEDERALLY FUNDED RESEARCH AND DEVELOPMENT IN NETWORKING AND INFORMATION TECHNOLOGY

Executive Office of the President

President's Council of Advisors on Science and Technology

DECEMBER 2010

- PCAST calls on the Federal government to increase R&D investments for collecting, storing, preserving, managing, analyzing, and sharing the increasing quantities of data.

- Furthermore, PCAST observed that the potential to gain new insights … to move from data to knowledge to action has tremendous potential to transform all areas of national priority.

Source: PCAST (December 2010), "Report to the President and Congress: Designing a Digital Future…"– A periodic congressionally-mandated review of the Federal Networking and Information Technology Research and Development (NITRD) Program. Available at: http://www.whitehouse.gov/administration/eop/ostp/pcast

Federal Big Data R&D Launch

March 29, 2012

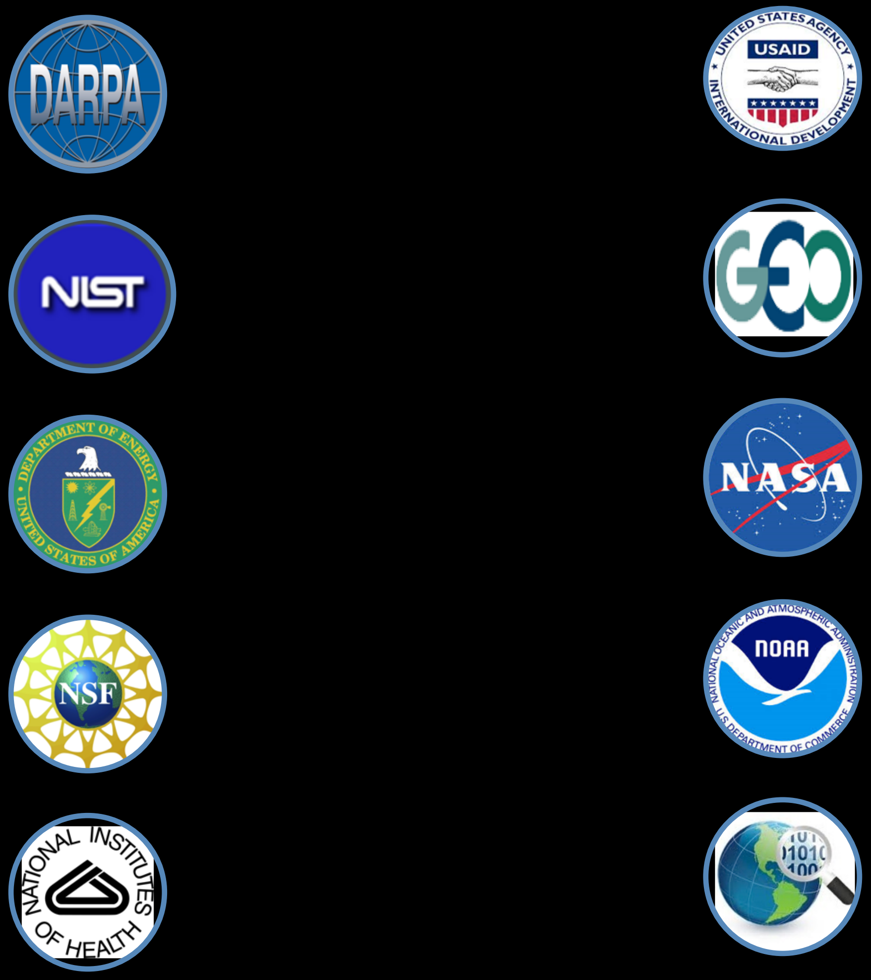
- **Led by cross-agency "Big Data" Senior Steering Group –** chartered in spring 2011 by the White House OSTP:
  - Co-chaired by NSF and NIH
  - Significant research community input
  - Members from 18 agencies
  - Charged with developing a framework and a plan

- **Major Announcements**: NSF, NIH, USGS, DoD, DARPA, DOE

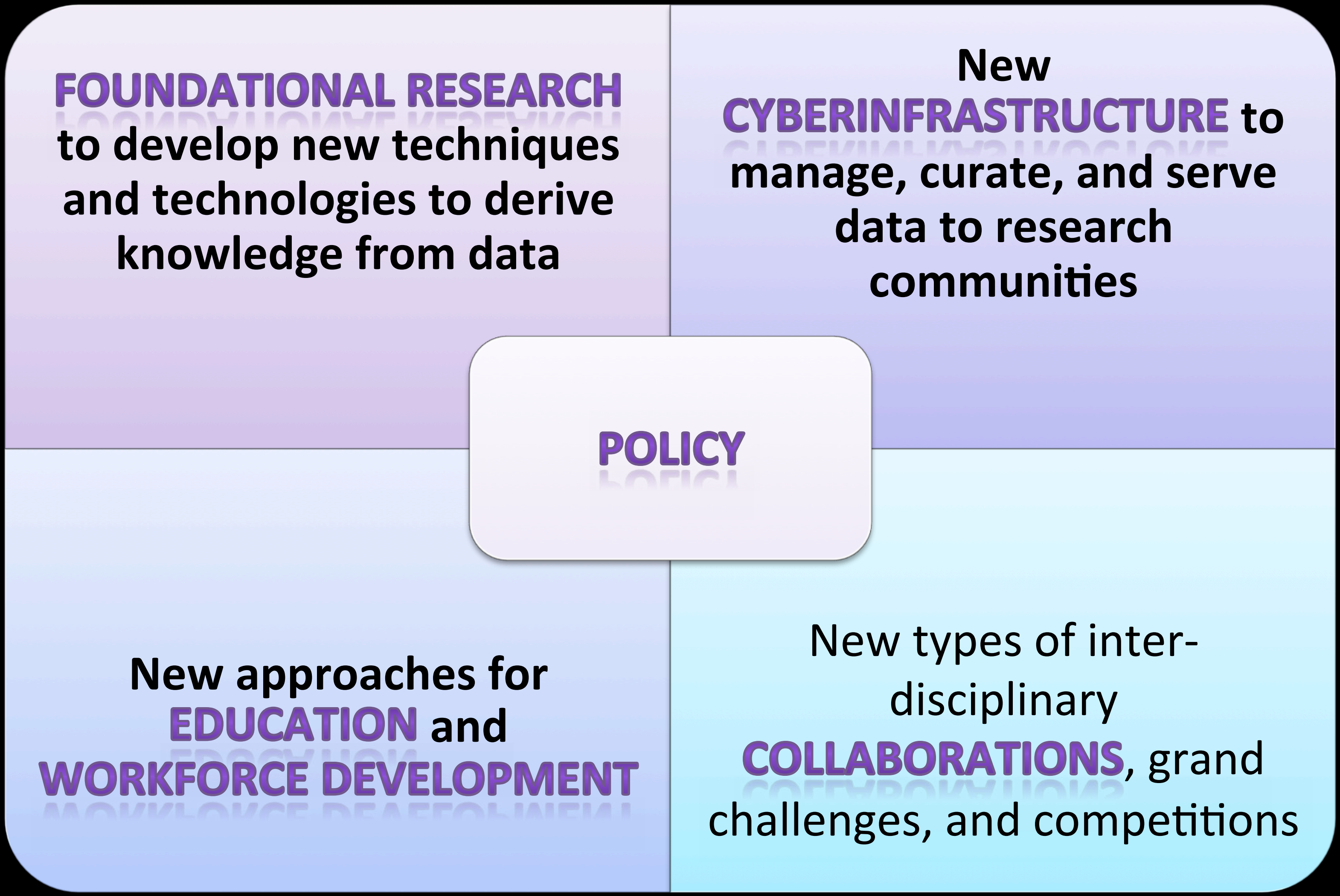
- **Cornerstone Announcement**: *Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIG DATA) Solicitation*
  - All NSF Directorates and 8 NIH Institutes
  - Research Thrusts: Collection, Storage, and Management; Data Analytics; Research in Data Sharing and Collaboration

More information available at: http://nsf.gov/news/news_summ.jsp?org=CISE&cntn_id=123607&preview=false

Agency Program Highlights


DARPA
UNITED STATES AGENCY
USAID
INTERNATIONAL DEVELOPMENT
NIST
GEO
DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA
NASA
NSF
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NOAA
U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTES OF HEALTH
1010

NSF Framework for Investments



FOUNDATIONAL RESEARCH to develop new techniques and technologies to derive knowledge from data

New **CYBERINFRASTRUCTURE** to manage, curate, and serve data to research communities

POLICY

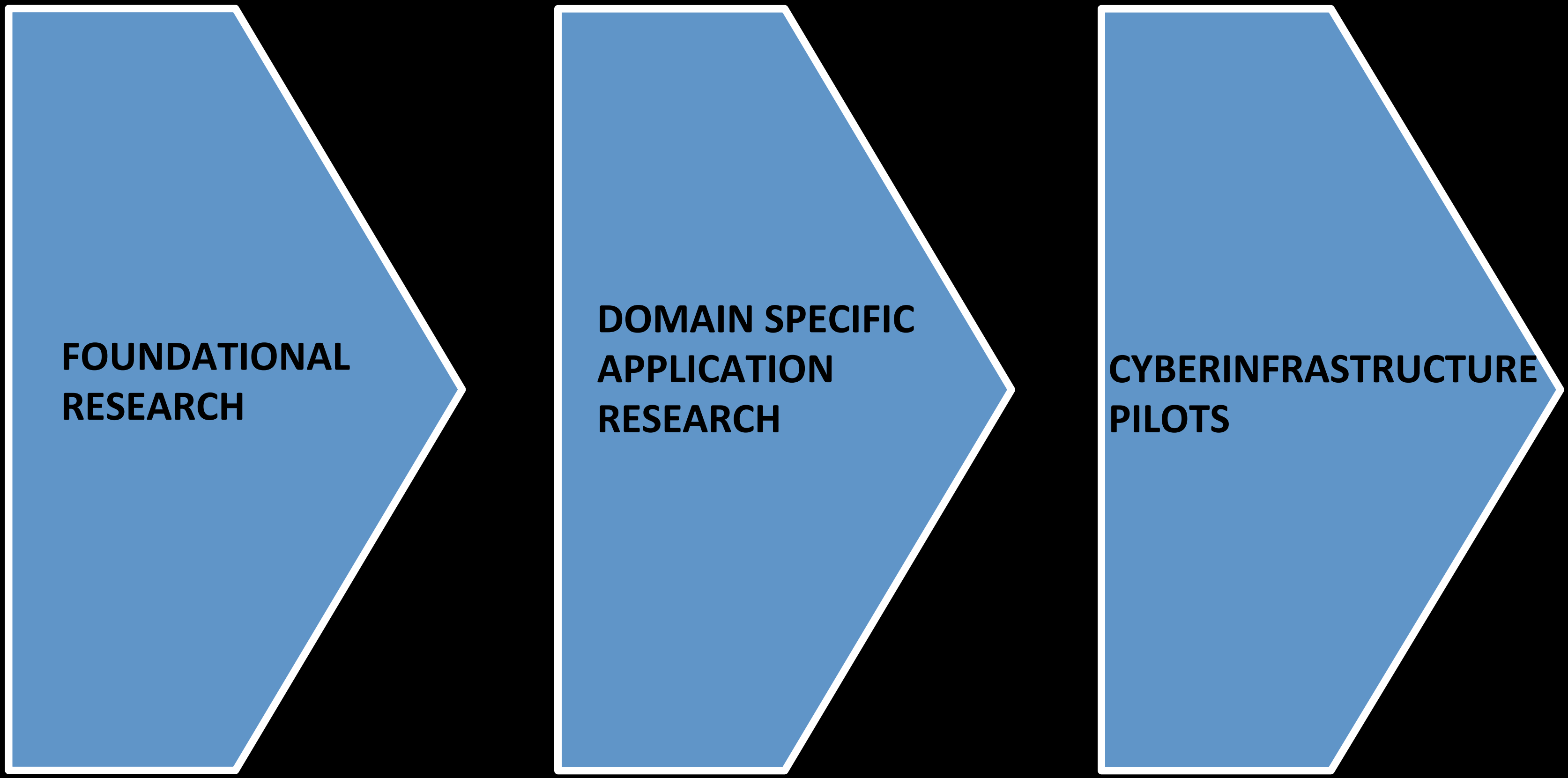
New approaches for **EDUCATION** and **WORKFORCE DEVELOPMENT**

New types of inter-disciplinary **COLLABORATIONS**, grand challenges, and competitions

Critical Techniques and Technologies for Advancing Big Data Science and Engineering (NSF 14-543)

- Two categories for submission
  - Foundational: Encourages fundamental, novel techniques, theories, methodologies and technologies of broad applicability
  - Innovative Applications: Encourages novel techniques, theories, methodologies, and technologies of interest to at least one specific application
- Due Date: June 9, 2014
- Size: up to $500K per year for up to 4 year

Building a Big Data R&D Pipeline



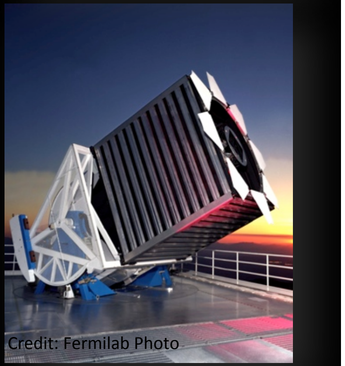
FOUNDATIONAL RESEARCH

DOMAIN SPECIFIC APPLICATION RESEARCH

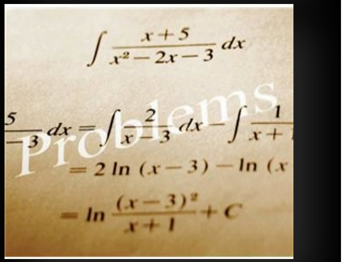
CYBERINFRASTRUCTURE PILOTS

Complex Policy Setting

- Researchers want data.
- Public policy requires access to data.
- Public policy also requires protection of privacy, intellectual property, and sensitive information.
- Business model challenges for publishers and societies.
- White House Memo on Feb. 22 directs United States federal agencies to develop a plan to support "increased public access" of results from federally funded research.

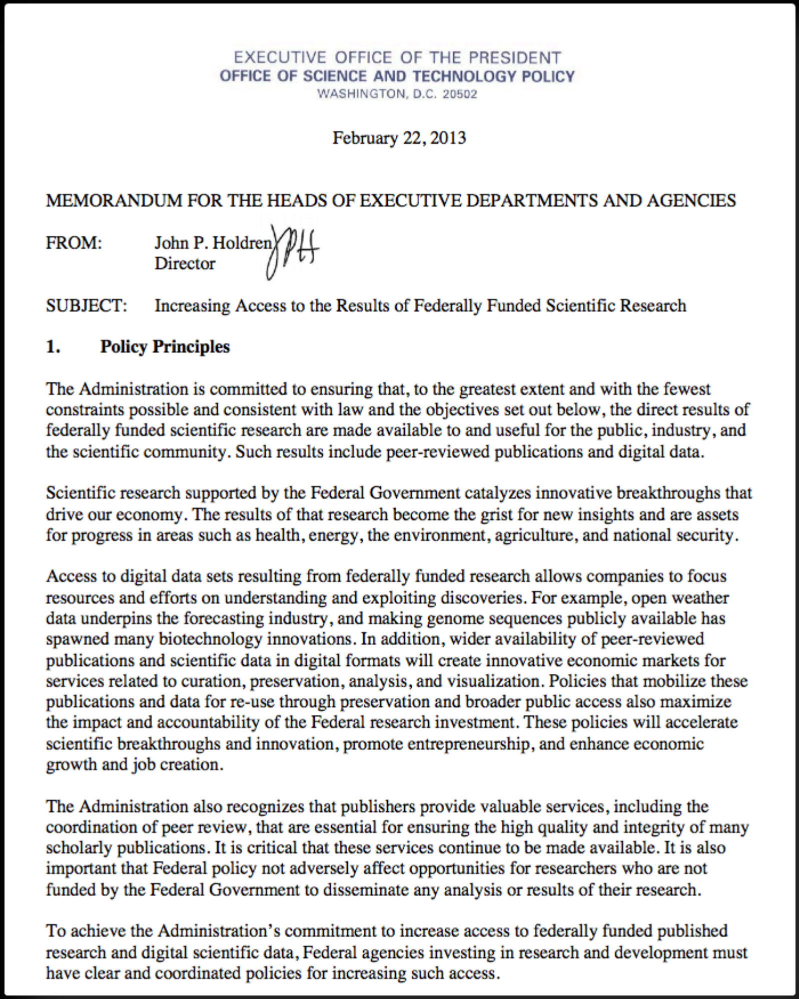


Credit: Fermilab Photo



Public Access

- White House Memo on Feb. 22 directs United States federal agencies to develop a plan to support "increased public access" of results from federally funded research.
- *Peer-reviewed publications … should be stored for "long-term preservation and publicly accessible to search, retrieve, and analyze in ways that maximize the impact and accountability of the Federal research investment."*
- *Digitally formatted scientific data resulting from unclassified research … "should be stored and publicly accessible to search, retrieve, and analyze."*



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather data underpins the forecasting industry, and making genome sequences publicly available has spawned many biotechnology innovations. In addition, wider availability of peer-reviewed publications and scientific data in digital formats will create innovative economic markets for services related to curation, preservation, analysis, and visualization. Policies that mobilize these publications and data for re-use through preservation and broader public access also maximize the impact and accountability of the Federal research investment. These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation.

The Administration also recognizes that publishers provide valuable services, including the coordination of peer review, that are essential for ensuring the high quality and integrity of many scholarly publications. It is critical that these services continue to be made available. It is also important that Federal policy not adversely affect opportunities for researchers who are not funded by the Federal Government to disseminate any analysis or results of their research.

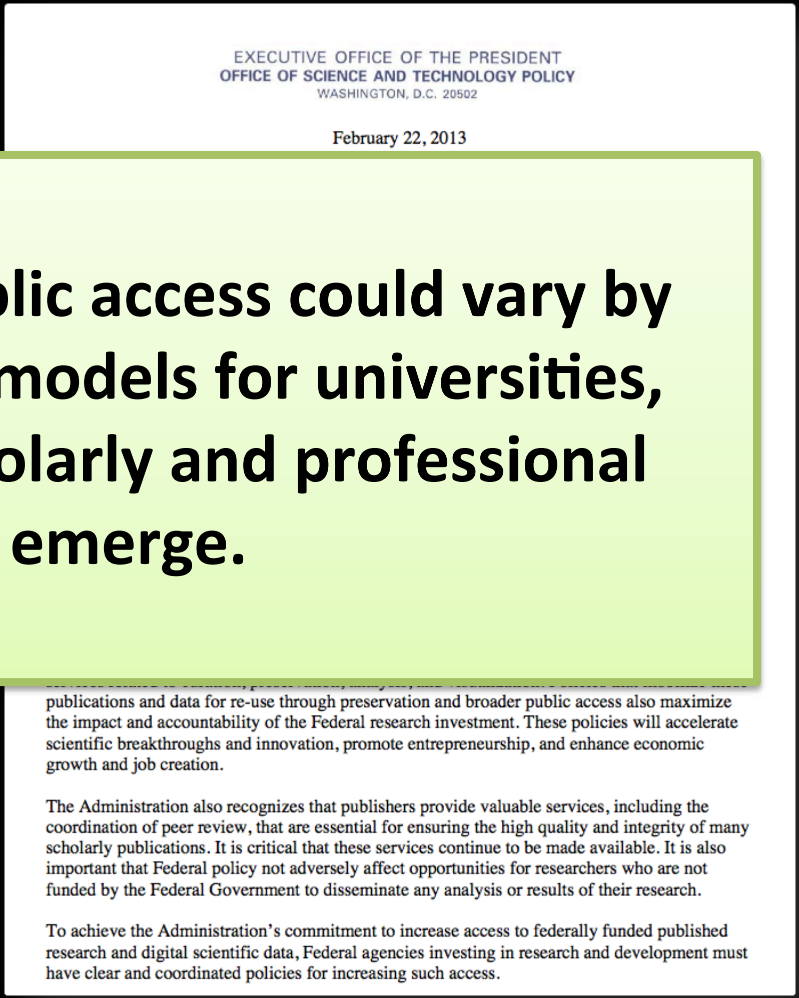
To achieve the Administration's commitment to increase access to federally funded published research and digital scientific data, Federal agencies investing in research and development must have clear and coordinated policies for increasing such access.

Public Access

- White House Memo on Feb. 22 directs United States federal agencies to develop a plan to

Implementation plans for public access could vary by discipline, and new business models for universities, libraries, publishers, and scholarly and professional societies could emerge.

- *Digitally formatted scientific data resulting from unclassified research … "should be stored and publicly accessible to search, retrieve, and analyze."*



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

publications and data for re-use through preservation and broader public access also maximize the impact and accountability of the Federal research investment. These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation.

The Administration also recognizes that publishers provide valuable services, including the coordination of peer review, that are essential for ensuring the high quality and integrity of many scholarly publications. It is critical that these services continue to be made available. It is also important that Federal policy not adversely affect opportunities for researchers who are not funded by the Federal Government to disseminate any analysis or results of their research.

To achieve the Administration's commitment to increase access to federally funded published research and digital scientific data, Federal agencies investing in research and development must have clear and coordinated policies for increasing such access.

Data to Knowledge to Action:

White House event encouraging public-private partnerships across the country

November 12, 2013

Materials Genome Initiative

Goal: Decrease the time-to-market by 50%

To help businesses discover, develop, and deploy new materials twice as fast, we're launching what we call the Materials Genome Initiative. The invention of silicon circuits and lithium ion batteries made computers and iPods and iPads possible, but it took years to get those technologies from the drawing board to the market place. We can do it faster.

-President Obama, Carnegie Mellon University, June 2011

Goals:

- Develop a Materials Innovation Infrastructure
- Achieve National goals in energy, security, and human welfare with advanced materials
- Equipt the next generation materials workforce

Themes

- Incentivizing open paradigms of sharing & access of tools
- Facilitating the development of innovation ecosystems & access to all stakeholders
- Driving innovative techniques across computation, informatics & experimentation
- Catalyzing shift in culture across the entire materials continuum & scaling the movement

Cognitive Science and Neuroscience

Goal: Understanding the human brain

- White House BRAIN Initiative launched in April 2013 (NSF, NIH, DARPA).
- Addresses critical challenge of research integration across multiple scales ranging from molecular to behavioral levels.
- Builds on NSF's unique ability to catalyze multi-disciplinary research and ongoing NSF investments.
- New neuroscience discoveries will
  - enable us to foster brain health;
  - engineer solutions that enhance, replace or compensate for lost function;
  - improve the effectiveness of formal and informal educational approaches;
  - promote learning across the lifespan; and
  - build brain-inspired smarter technologies for improved quality of life.

BRAIN INITIATIVE

BRAIN RESEARCH THROUGH ADVANCING INNOVATIVE NEUROTECHNOLOGIES

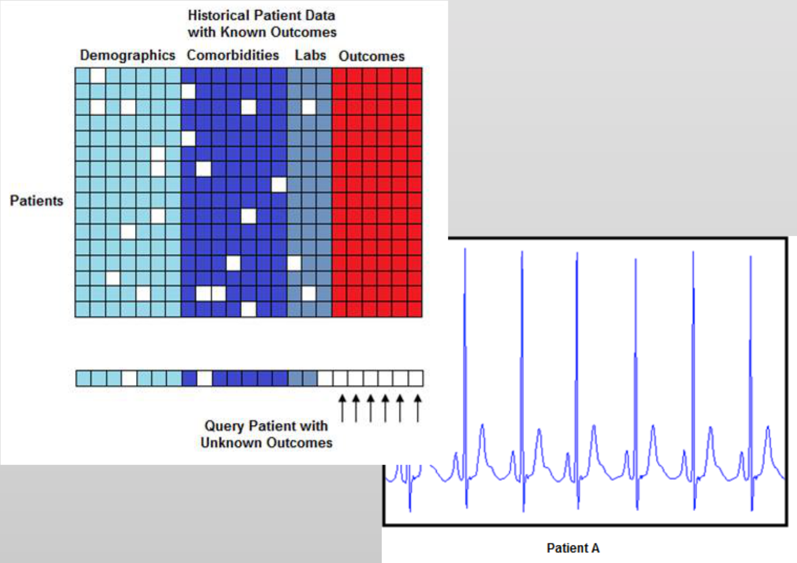
Some success stories…

Enabling the Patient, Curing Diseases, and Saving Lives

Data at the Forefront of Diagnosis

Predicting Risk of Cardiovascular Death

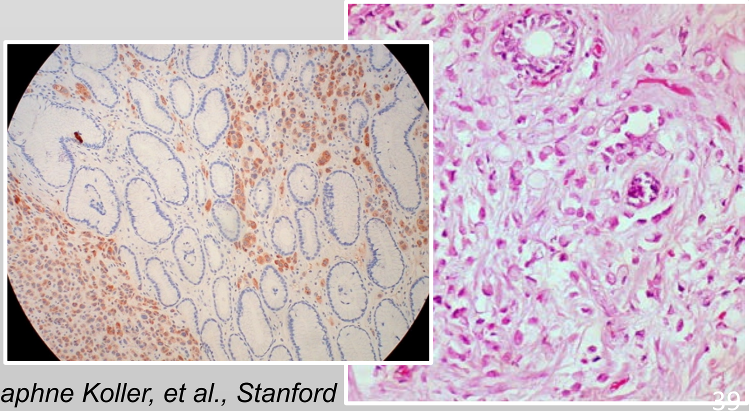
- Applying **data mining and machine learning** techniques to EKGs has identified new "computational biomarkers."
- These markers can help determine **heart abnormalities and defects**, leading to significant improvement in identifying and treating at-risk patients.



Zeeshan Syed, University of Michigan and John Guttag, MIT

Predicting Survival Time of Breast Cancer Patients

- Applying **image analysis techniques** to breast cancer biopsy images has identified a small subset of cellular features (out of 6,000 possible) predictive of survival time among breast cancer patients.
- This feature set that is the best predictor were not from the cancer tissue itself, but rather from adjacent tissue – something that previously **had gone undetected by pathologists and clinicians**.



Daphne Koller, et al., Stanford

Transformative Implications for Commerce

UPS Uses Data Analytics to Deliver Faster and More Sustainably

- Telemetry and data analysis system enables peak efficiency, saving millions of dollars:
  - Eliminated 5.3 M miles from its routes
  - Reduced engine idling time by almost 10 million minutes
  - Saved 650,000 gallons of fuel
  - Reduced carbon emissions by more than 6,500 metric tons

Helping Small Businesses

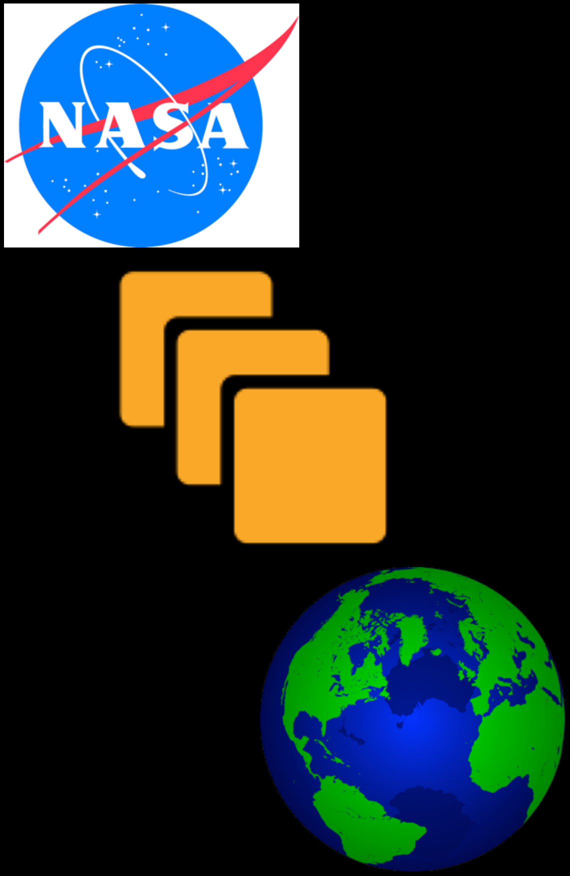
- The **NYC Mayor's Office of Data Analytics (MODA)** is working across city government to use city data to improve daily operations, help to prepare for and respond to disasters, and support economic growth

- Now working with NYC New Business Acceleration Team (NBAT) to **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 reduce a new business' time to open

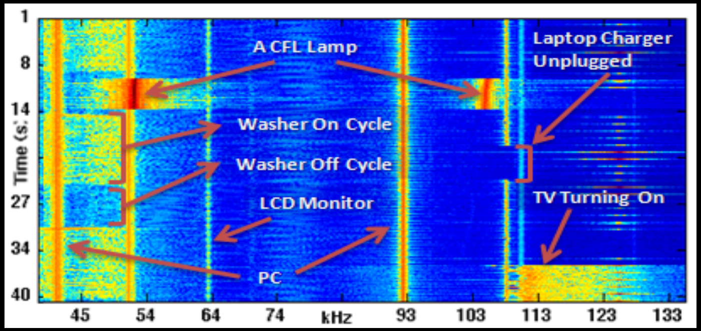
Big Data for a Sustainable Future

Bringing NASA Data Down to Earth



- Amazon Web Services (AWS) and NASA are providing a significant amount of NASA's Earth science data and models to the public
- 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)

ElectiSense



From *The Human Face of Big Data (Rick Smolan/Peter Menzel)*

Images courtesy of Shwetak Patel, UW

- Just two sensors can detect the use of electrical devices in the home
- Pattern matching and machine learning provides new insights to electrical use

Surprising Insights!

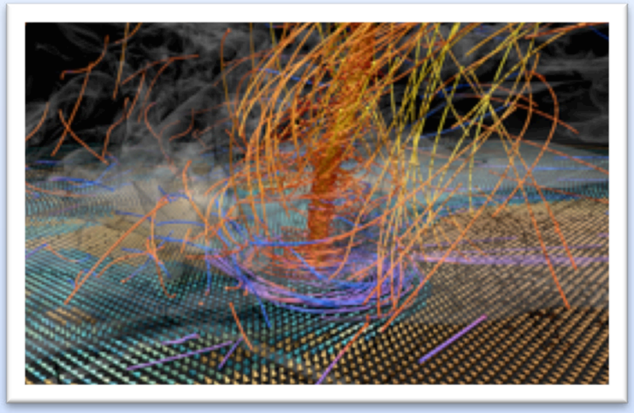
- All kitchen appliances: < 5%
- TVs: 10%
- Pool pump: 30%
- DVR: 13%
- Lighting: 15%

Accelerating the Pace of Discovery in Science and Engineering

Extracting Knowledge from Data

Image Credit: *Sigrid Knemeyer*

New Tool for Extracting Knowledge from Large Data Sets: A new statistical tool, part of a suite called MINE, can tease out multiple patterns hidden in health information from around the globe, statistics amassed from major league baseball, data on bacterial biodiversity, and much more. (Michael Mitzenmacher, Harvard with researchers from the Broad Institute)



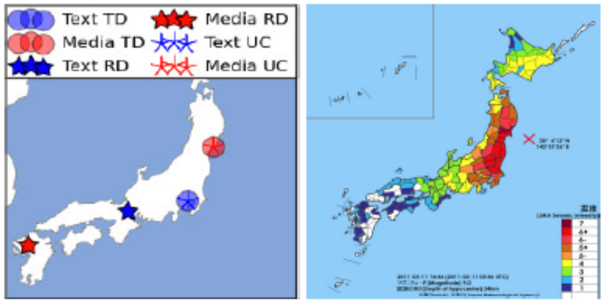
Forecasting Tornadoes: Parallel computing, data mining, and meteorology are being used to determine tornado formation and more reliable tornado forecasting. (Amy McGovern and Kevin Droegemeier, University of Oklahoma)

Image Credit: Bob Wilhelmson, NCSA and the University of Illinois at Urbana-Champaign; Lou Wicker, National Oceanic and Atmospheric Administration's National Severe Storms Laboratory; Matt Gilmore and Lee Cronce, University of Illinois atmospheric science department. Visualization by Donna Cox, Robert Patterson, Stuart Levy, Matt Hall and Alex Betts, NCSA

Social Media and Big Data

Social Media Provides Rapid Insight into the Extent of Damage Following a Catastrophic Event

- Researchers at Texas A&M **analyzed and modeled** the spatial coverage of social media following the 2011 Tokhuku earthquake in Japan and the the 2011 Christchurch earthquake in New Zealand.
- Analyzed tweet density, re-tweet density, and tweet count to estimate the **epicenter** and model **intensity** attenuation of each earthquake.

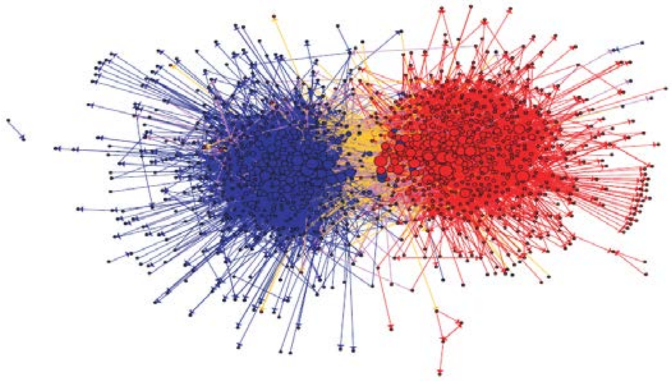


(a) Estimated Epicenter (b) Actual Epicenter

Yuan Liang, James Caverleeand John Mander, Texas A&M

Understanding Political Communities

- Measuring the degree of interaction between liberal and conservative blogs uncovered differences in the **structure of the two communities.**
- Found **differences in the behavior** of liberal and conservative blogs, with conservative blogs linking to each other more frequently and in a denser pattern.



From "The Political Blosphere and the 2004 US Election: Divided They Blog" by Lada Adamic and Natalie Glance

Barriers, Challenges and Opportunities

Barriers, Challenges …

- Technological Solutionism
- Big data security
- Big data and privacy
- Dangers of predictive analytics
- Education and Workforce Development

Technology *alone* will not solve all of society's challenges. Must consider economic, social and cultural barriers to adoption and use of solutions.

Big Data and Security

Security of Big Data

Data Analytics for Securing Cyber Space

Image Credit: CCC and SIGACT CATCS

Evolution of Cyber Threats

Future security challenges will follow technology and Internet adoption patterns

- Proliferation of mobile devices and wireless networks exposes new vulnerabilities.
- Social media platforms open new avenues for hackers.
- Protecting cloud infrastructure has become key to long-term adoption.
- Increasingly cyber-enabled systems expands the scope of attacks to physical infrastructure
  - manufacturing, energy production, healthcare and transportation.

Big Data and Security

- Big data is a big security target:
  - Apple has 400 million customer credit cards on file
  - Facebook has more than 1Billion registered users who share more than 1Billion pieces of contents each day
  - …
- No longer about protecting internal data
- Impact of platforms and ecosystems:
  - More than 1,000,000 apps in the AppStore
  - How secure are partner platforms and third-party apps?

Cloud and Virtualization

- **What's a perimeter?** With **cloud computing** and proliferation of mobile devices, an organization's information is no longer stored and accessed within its walls or perimeter. Information entirely created and stored on the web. **Expect rise in *insider threats!***
- Systems and resources, including networks, hosts, storage, data centers and applications are **increasingly virtualized and distributed**, and commonly under the control of the end-users themselves.
- Confidential information and intellectual property are increasingly flowing from back-end systems that the organization **doesn't control**, through networks that it **doesn't control**, to endpoints and end-users that it **doesn't control**.

Big Data and Privacy

Big Data and Privacy

- Amazon monitors our shopping preferences
- Google tracks our browsing habits
- Facebook captures our social interactions and more
- Twitter tracks what on our minds in real-time
- Wireless service operators track our connections, and who is nearby
- …

Big Data and Privacy

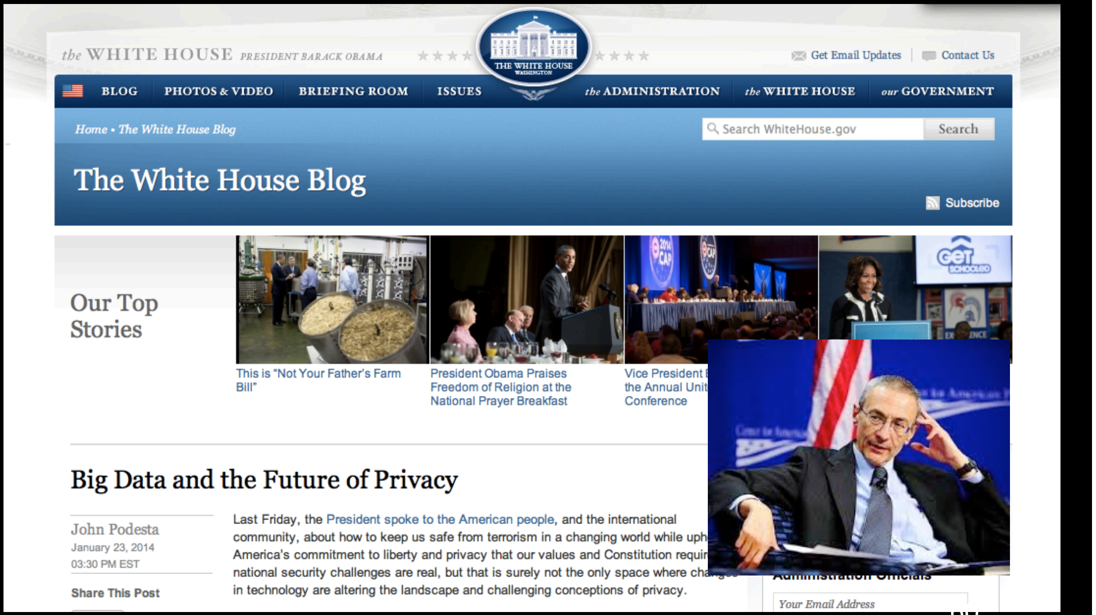
- Informed consent, opting out, and anonymization are not as effective to ensure privacy with big data:
  - How can companies provide notice for a purpose that has yet to be exist?
  - How can one give informed consent if new secondary uses of information haven't even been imagined?
  - What if anonymized data can be mined with other (public) datasets to identify private information?

Big Data and the Future of Privacy

January 17, 2014

"A comprehensive 90-day review of the way that *big data* will affect the way we live and work; the relationship between government and citizens; and how public and private sectors can spur innovation and maximize the opportunities and free flow of this information while minimizing the risks to privacy."

-President Obama



The White House Blog

Our Top Stories

This is "Not Your Father's Farm Bill"

President Obama Praises Freedom of Religion at the National Prayer Breakfast

Vice President
the Annual Unit
Conference

Big Data and the Future of Privacy

John Podesta
January 23, 2014
03:30 PM EST

Share This Post

Last Friday, the President spoke to the American people, and the international community, about how to keep us safe from terrorism in a changing world while uph
America's commitment to liberty and privacy that our values and Constitution requir
national security challenges are real, but that is surely not the only space where ch
in technology are altering the landscape and challenging conceptions of privacy.

Predictive Analytics

The Power and Dangers of Predictive Analytics

- Data mining and pattern recognition to make predictions:
  - Flu outbreaks throughout the world
  - Predicating election outcomes with high accuracy
  - Suggestions on customer preferences for books
  - Predicting customer behavior and managing inventory
  - Detecting insurance fraud

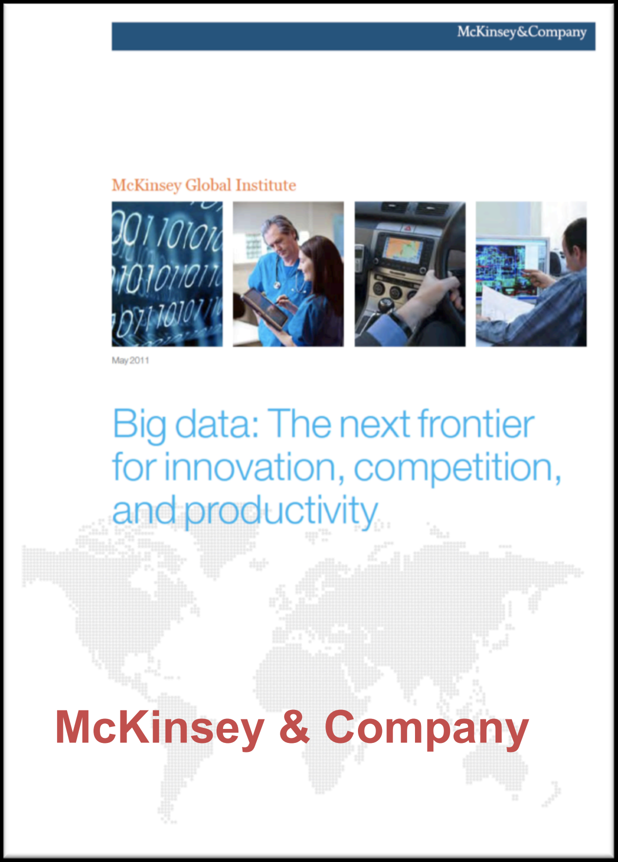
- Who would object to Preventing unhealthy or dangerous behavior?
- Who would object to improving quality of customer experience?
- Who would object to improving business efficiency?
- Who would object to efficient use of resources in crime prevention?

Dangers: Predictions Based on Correlations to Make Causal Decision

- Correlation does not imply causation
- "Fooled by randomness" as suggested by Nassim Taleb
- Crossing the line from improvement to profiling
- Crossing the line from prevention to penalizing
- Idea of the "presumption of innocence" is foundational to our legal system

Education and Workforce Development

Education, Learning, Workforce Development, Computational and Data-enabled Science



McKinsey&Company

McKinsey Global Institute

May 2011

Big data: The next frontier for innovation, competition, and productivity

McKinsey & Company

"By 2018 the United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data."[1]

[1] McKinsey&Company (May 2011), "Big data: The next frontier for innovation, competition, and productivity." Available at: http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

“Data Science: The Sexiest Job of the 21st Century”

- Harvard Business Review, October 2012

- The number of private-sector economists surged 57% between 2009 and 2012 as U.S. firms looked for help digesting data.
  - Even the New York Times has hired a chief data scientist.

NSF Research Traineeship (NRT)

Preparing professionals in emerging STEM fields vital to the nation

Priority research theme: Data-enabled science and engineering

- **Purpose:** create and promote new, innovative, effective, and scalable models for STEM graduate student training and prepare scientists and engineers of the future, particularly in emerging STEM fields vital to the nation.
- **Anticipated award amount:** up to $3M over 5 yrs.

NSF-wide Initiative

"Paradox of Innovation: no one knows how an invention will impact the world until it is widely used, leading to unintended consequences”

Imagine a Day…

- By integrating biomedical, clinical, and scientific data, we can predict the **onset of diseases and identify unwanted drug interactions.**
- By coupling roadway sensors, traffic cameras, and individuals' GPS devices, we can **reduce traffic congestion and generate significant savings** in time and fuel costs.
- By **accurately predicting natural disasters** such as hurricanes and tornadoes, we can employ life-saving and preventative measures that mitigate their potential impact.
- By integrating emerging technologies, such as MOOCS and inverted classrooms, with knowledge from research about how people learn, we can **transform formal and informal education**.
- By **correlating disparate data streams** through text mining, image analysis, and face recognition, we can enhance public safety and security.

Big Opportunities for the Future

- Transformative implications for commerce and economy
- Critical to accelerating the pace of discovery and innovation
- Enhancing quality of life and societal wellbeing

Thanks!

fjahania@nsf.gov

Credits

- Copyrighted material used under Fair Use. If you are the copyright holder and believe your material has been used unfairly, or if you have any suggestions, feedback, or support, please contact: ciseitsupport@nsf.gov.
-
- The inclusion of a logo does not express or imply the endorsement by NSF of the entities' products, services, or enterprises.