

# The Promise of Big Data

## An NSF Perspective

Howard Wactlar  
National Science Foundation



CENDI Meeting  
July 2013

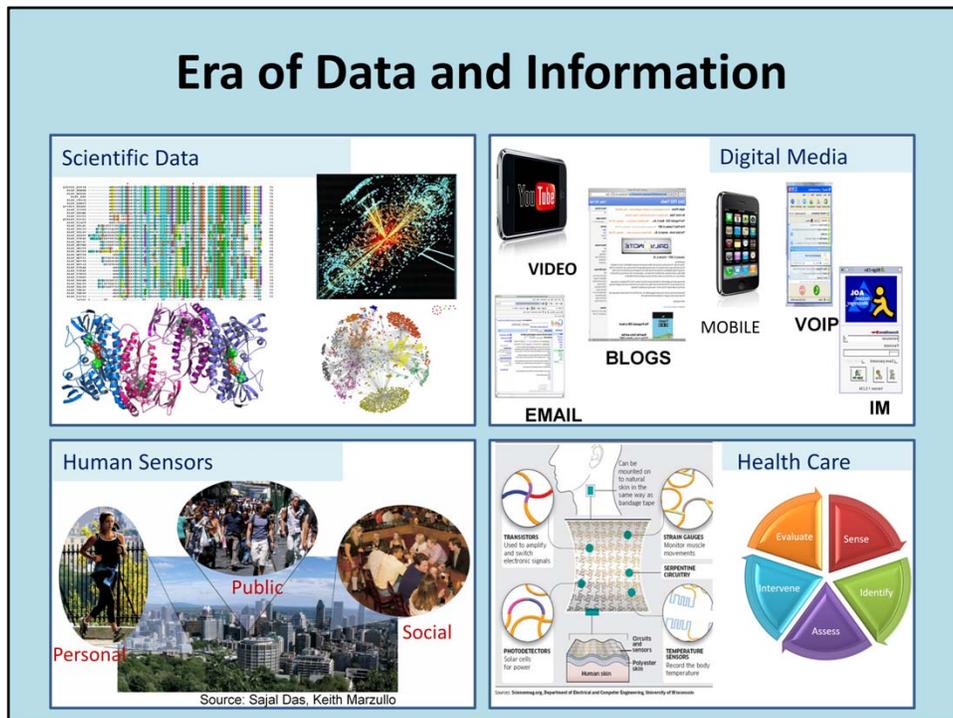
Good morning. It is a pleasure to be with you here today to talk about the value and promise of Big Data.

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



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

# Era of Data and Information



We are now living in the “Era of Data and Information” ... enabled by

- modern experimental methods and observational (**longitudinal**) studies,
- large-scale simulations,
- scientific instruments such as telescopes and particle accelerators,
- Internet transactions,
- email, videos, images, click streams...
- Not to mention ... ubiquitous widespread deployment of **sensors** everywhere
- in the environment,
- in our critical infrastructure such as bridges and smart grids,
- in our homes, and even on our clothing!

Sloan Digital Sky Survey in 2000, collected more data in its 1<sup>st</sup> few weeks than had been amassed in the entire history of astronomy

Approximately 90% of the data in the world today were created in the last two years alone. However, when we talk about Big Data, it is

important to emphasize not only the enormous **volume** of data being generated, but also the **velocity, heterogeneity** and **complexity** of the data that now confront us.



## Common Misconceptions

- If you can store the data on your laptop, you don't have "Big Data"
- Analyzing Big Data just requires a big computer (and maybe HADOOP)
- Big Data are just lots of small data

## Big Data Challenges

Big Data is not just about “rate” and “size”, but about “complexity”:

*Error rates and types, heterogeneous data, missing data*

Methods that have high accuracy on small datasets may have poor accuracy on “bigdata”.

“Scaling up” of existing methods will not be not enough!

There is a trade-off between data quality and quantity. Having lots of data available doesn't necessarily mean you should use them all.

Requires fundamentally new methods from Computer Science, Mathematics, and Statistics.

## 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



## Why is Big Data Important?

Big Data is important to all facets of the discovery and innovation ecosystem, including the Nation's academic, government, industrial, entrepreneurial, and investment communities.

**First**, insights and more accurate predictions from large and complex collections of data are creating opportunities in new markets, driving the creation of IT products and services, and boosting the productivity of businesses.

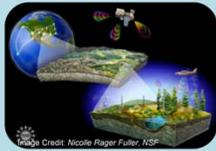
**Second**, advances in our ability to store, integrate, and extract meaning and information from data are accelerating the pace of discovery in almost every science and engineering discipline, and informing the development of new strategies for effective learning and education.

**Third**, Big Data has the potential to solve some of the

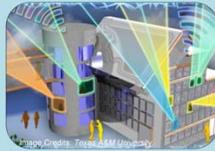
# Nation's most pressing challenges



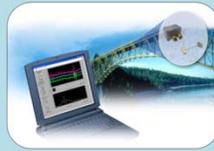
## Data-driven Discovery and Innovation Address Societal Challenges



Environment & Sustainability



Broadband & Universal Connectivity



Manufacturing, Robotics, & Smart Systems



Emergency Response & Disaster Resiliency



Secure Cyberspace



Health & Wellbeing



Transportation & Energy



Education and Workforce Development

As data gathers at an ever-increasing rate across all scales and complexities, there are enormous opportunities

- to harness data,
- to **extract knowledge** from them,
- to provide powerful new approaches to **drive discovery** and **decision-making**,
- to make increasingly **accurate predictions** based on data, and
- to gain a deeper understanding of **causal relationships** based on advanced data analysis.

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 **correlating disparate data streams** through text mining, image analysis, and face recognition, we can enhance public safety and security.

By correlating disparate sources of information, we can enable a cyber world that is safe, secure, and resilient.

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.

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



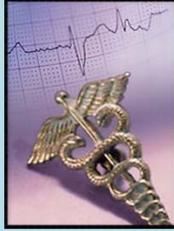
“By 2018 the United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data.”<sup>1</sup>

<sup>1</sup>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](http://www.mckinsey.com/insights/MGI/Research/Technology_and_innovation/Big_data_The_next_frontier_for_innovation)

And not only will Big Data help us to address education, but we also will need people with the skills to analyze, understand and make decisions based on big data.

Read QUOTE from McKinsey.

At the end of the day, cyberinfrastructure is all about people; enabling them to do what they have not been able to do before



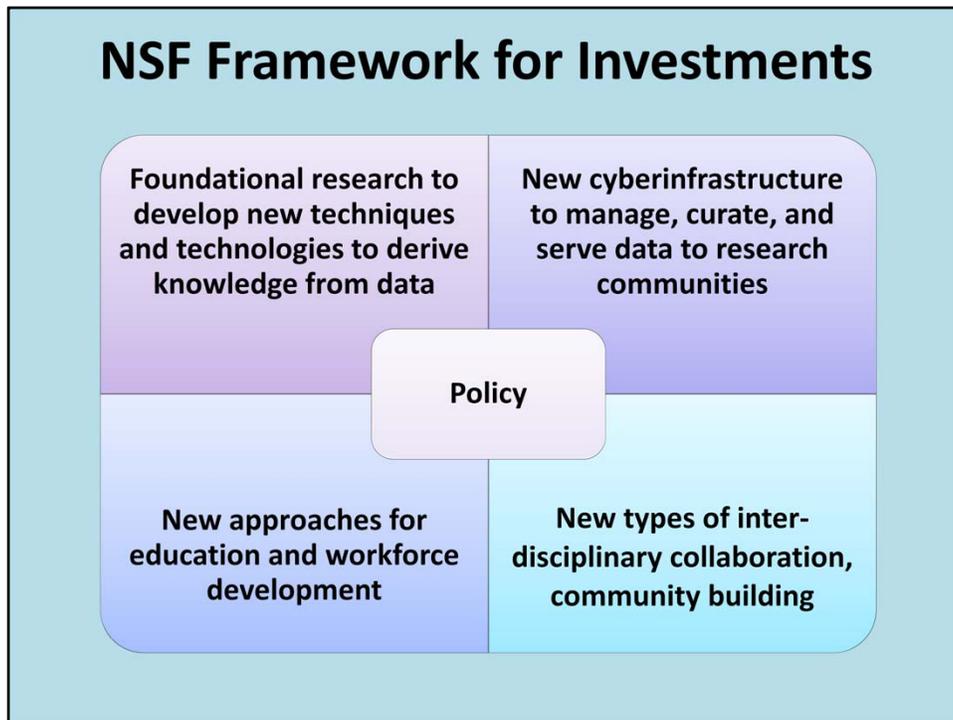
Classifying Breast Cancers  
via Image Analysis

Energy Savings  
in the Home



Reducing Traffic Congestion  
in Urban Areas

[Discuss specific examples or skip]



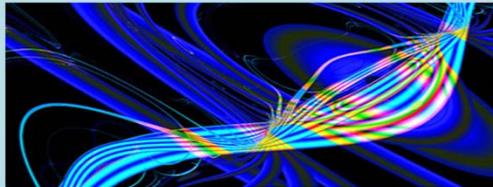
The National Science Foundation has identified four major investment areas that address current challenges and promise to serve as the foundations of a comprehensive, long-term agenda.

They are:

- Investments in **fundamental research** to advance big data techniques and technologies.
- Support **for building new multidisciplinary research communities.**
- Investments in **education and workforce development.**
- Development and deployment of **cyberinfrastructure** to capture, manage, analyze, and share digital data.

## Complex Policy Setting

- Practitioners and researchers want data.
- Public policy requires access to data.
- Public policy also requires protection of privacy, intellectual property, and other sensitive information.
- Policy and implementation plan for data sharing and open access are in progress. (WH OSTP Feb. 22<sup>nd</sup> memo on public access)



Policy considerations are also intertwined in our efforts.

- Researchers and practitioners want data.
- Public policy requires access to data.
- Public policy also requires protection of privacy, intellectual property, and other sensitive information.
- Policy and implementation plan for data sharing and open access are in progress.

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

So why does Big Data matter now?

A confluence of social, technical and policy interests place us on new terrain. The horizon is clear enough from where we stand today to provide us a glimpse of the future.

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

**Why Now? Confluence of Social, Technical and**

- Decades of advances in technology
- Data is no longer regarded as static:
  - now a raw material of business, potentially used to create new products and services
- Scalability: collecting, organizing, storing and analyzing data
- Increasing transparency of democratic governance (open government)
- Public access to high value datasets (data.gov)
- Democratization of data and tools

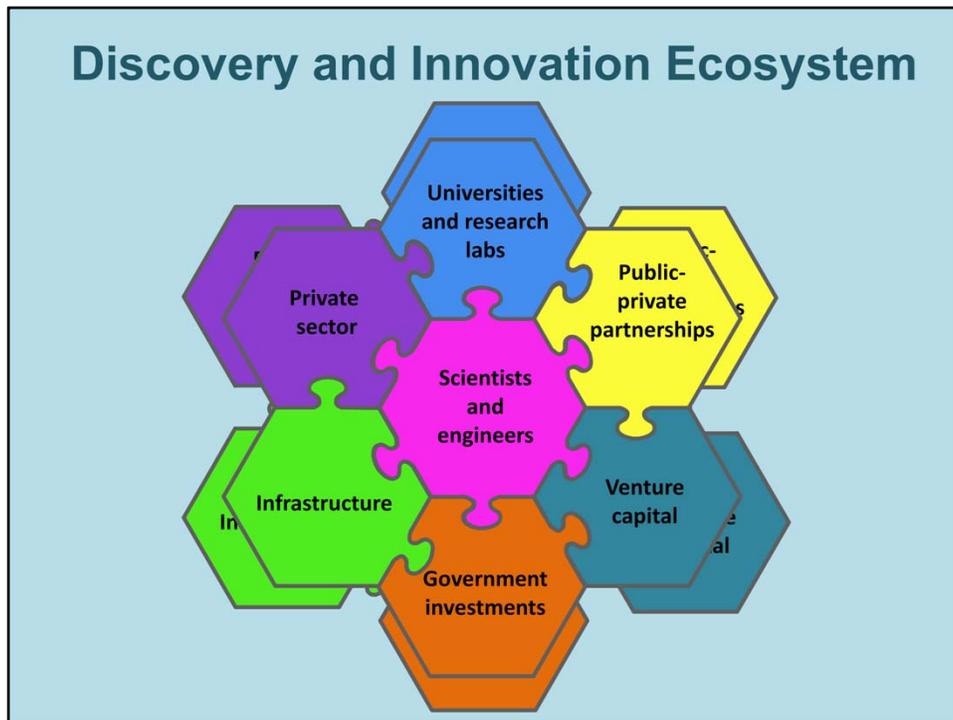


- Moore's Law**
- Kryder's Law**
- Pervasive Sensors**
- Data Mining**
- Machine Learning**
- NL Understanding**
- Info Retrieval**
- Computer Vision**
- Video Analytics**
- Data Visualization**
- Crowd Sourcing**
- Social Networks**
- ...

So why does Big Data matter now?

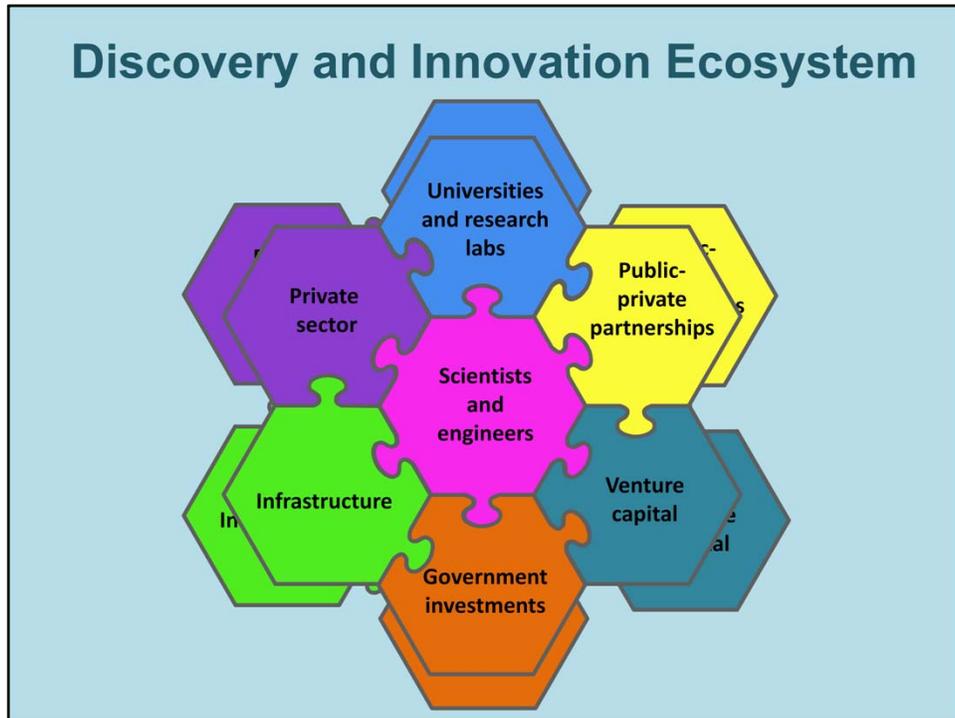
A confluence of social, technical and policy interests place us on new terrain. The horizon is clear enough from where we stand today to provide us a glimpse of the future.

Walmart (retailing) and Capital One (banking)



Realizing the enormous potential of Big Data requires a long-term bold, sustainable and comprehensive approach, not only by NSF, and not only by the Federal Government.

- Our success in harnessing Big Data will depend on a thriving discovery and innovation ecosystem that includes:
- Leading-edge universities and research labs;
- Scientists and engineers in a flexible talent-rich labor market;
- Government investments in research and education - the building blocks of discovery and innovation;
- A vibrant private sector catalyzed by American entrepreneurial spirit;
- Public-private partnerships – that promote competitive markets that spur productive **entrepreneurship** through promotion of innovation communities;
- An integrated and sustainable infrastructure;
- Venture capital



And, the extraordinarily productive interplay of these parts – of federally funded university research, federally and privately funded industrial research, and entrepreneurial companies founded and staffed by people who moved back and forth between universities and industry— is **the *principal* reason for the dramatic advances in information technology and the subsequent increase in innovation and productivity.** This flow of ideas and people, stimulated by investments in research, is a critical element of the IT R&D ecosystem.



*Thanks*

[hwactlar@nsf.gov](mailto:hwactlar@nsf.gov)

The work we do today will help lay the groundwork for the future ...

for new enterprises, promote economic growth, improve our citizens' quality of life, and fortify the foundations for U.S. competitiveness for decades to come.

Thank you again for your attention and for your contributions. I wish you a productive day!