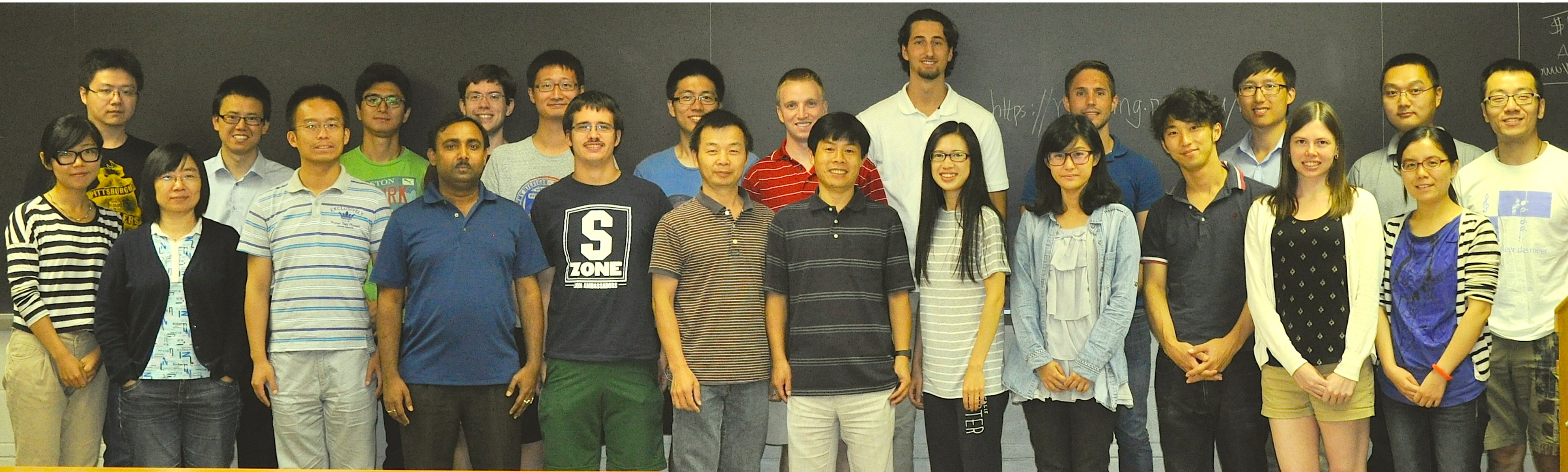
Penn State Center for
Advanced Data Assimilation and Predictability Techniques

PENNSTATE

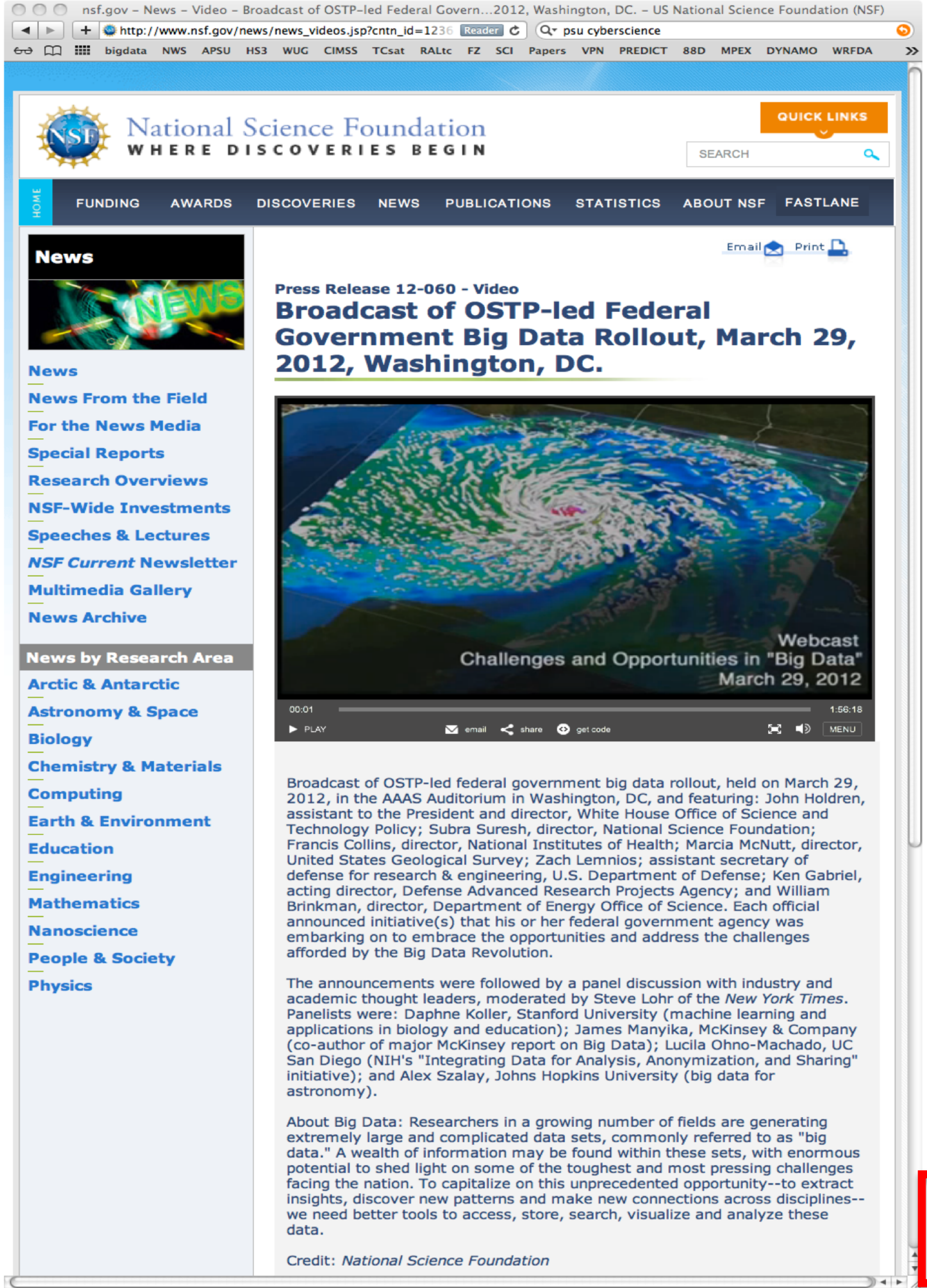
“Data” “Issues” for an Active Researcher

Fuqing Zhang
Pennsylvania State University

Professor of Meteorology and Professor of Statistics
Director, Penn State ADAPT Center



Penn State hurricane research in NSF “Big Data” rollout


nsf.gov – News – Video – Broadcast of OSTP-led Federal Govern…2012, Washington, DC. – US National Science Foundation (NSF)
http://www.nsf.gov/news/news_videos.jsp?cntn_id=1236
Reader
psu cyberscience
bigdata NWS APSU HS3 WUG CIMSS TCsat RALtc FZ SCI Papers VPN PREDICT 88D MPEX DYNAMO WRFDA
National Science Foundation
WHERE DISCOVERIES BEGIN
QUICK LINKS
SEARCH
HOME
FUNDING
AWARDS
DISCOVERIES
NEWS
PUBLICATIONS
STATISTICS
ABOUT NSF
FASTLANE
News
NEWS
Email
Print
00:01
1:56:18
PLAY
email
share
get code
MENU

News
News From the Field
For the News Media
Special Reports
Research Overviews
NSF-Wide Investments
Speeches & Lectures
NSF Current Newsletter
Multimedia Gallery
News Archive

News by Research Area

Arctic & Antarctic
Astronomy & Space
Biology
Chemistry & Materials
Computing
Earth & Environment
Education
Engineering
Mathematics
Nanoscience
People & Society
Physics

Press Release 12-060 - Video

Broadcast of OSTP-led Federal Government Big Data Rollout, March 29, 2012, Washington, DC.

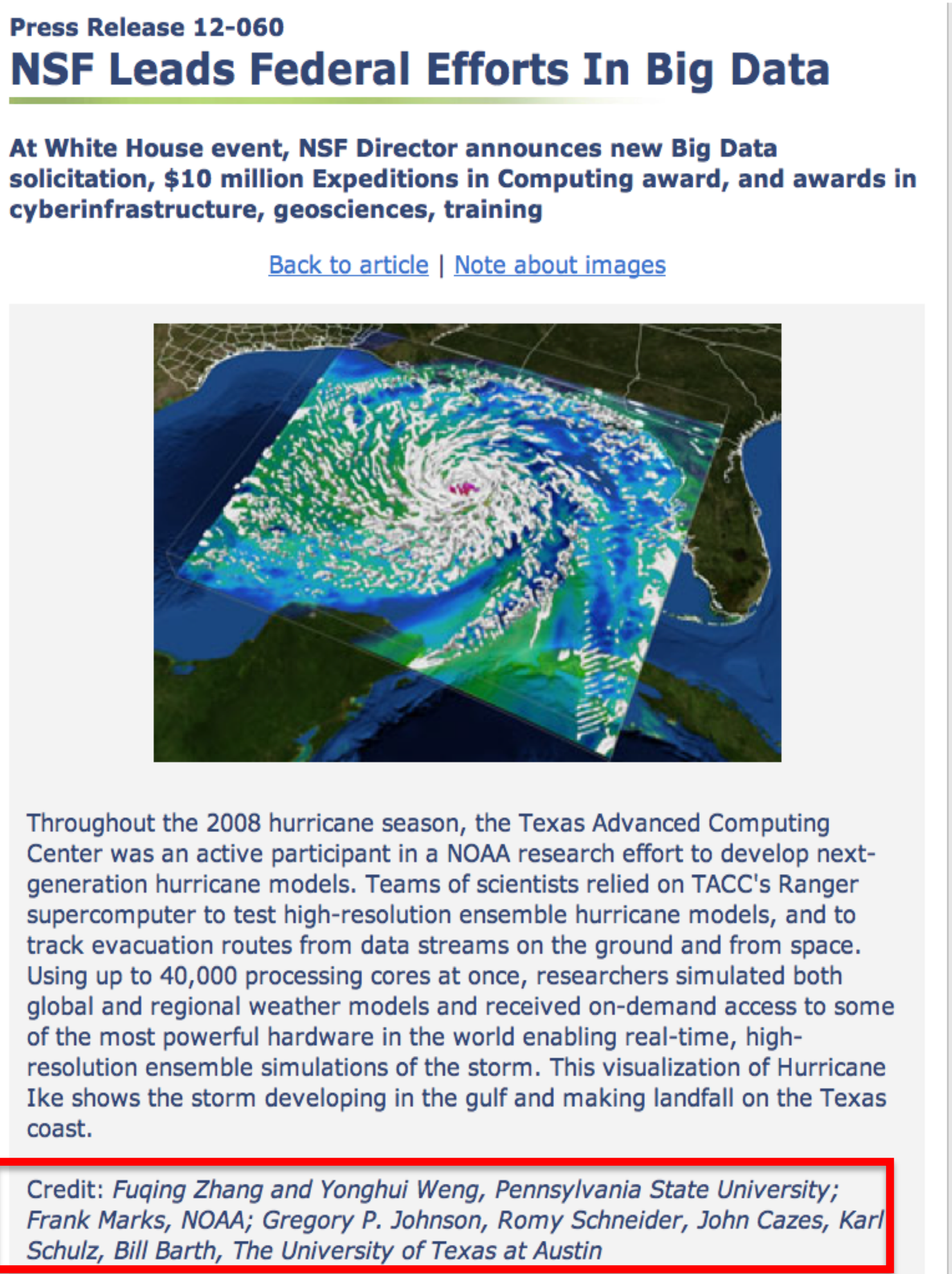
Webcast
Challenges and Opportunities in "Big Data"
March 29, 2012

Broadcast of OSTP-led federal government big data rollout, held on March 29, 2012, in the AAAS Auditorium in Washington, DC, and featuring: John Holdren, assistant to the President and director, White House Office of Science and Technology Policy; Subra Suresh, director, National Science Foundation; Francis Collins, director, National Institutes of Health; Marcia McNutt, director, United States Geological Survey; Zach Lemnios; assistant secretary of defense for research & engineering, U.S. Department of Defense; Ken Gabriel, acting director, Defense Advanced Research Projects Agency; and William Brinkman, director, Department of Energy Office of Science. Each official announced initiative(s) that his or her federal government agency was embarking on to embrace the opportunities and address the challenges afforded by the Big Data Revolution.

The announcements were followed by a panel discussion with industry and academic thought leaders, moderated by Steve Lohr of the *New York Times*. Panelists were: Daphne Koller, Stanford University (machine learning and applications in biology and education); James Manyika, McKinsey & Company (co-author of major McKinsey report on Big Data); Lucila Ohno-Machado, UC San Diego (NIH's "Integrating Data for Analysis, Anonymization, and Sharing" initiative); and Alex Szalay, Johns Hopkins University (big data for astronomy).

About Big Data: Researchers in a growing number of fields are generating extremely large and complicated data sets, commonly referred to as "big data." A wealth of information may be found within these sets, with enormous potential to shed light on some of the toughest and most pressing challenges facing the nation. To capitalize on this unprecedented opportunity--to extract insights, discover new patterns and make new connections across disciplines--we need better tools to access, store, search, visualize and analyze these data.

Credit: *National Science Foundation*



Press Release 12-060

NSF Leads Federal Efforts In Big Data

At White House event, NSF Director announces new Big Data solicitation, $10 million Expeditions in Computing award, and awards in cyberinfrastructure, geosciences, training

Back to article | Note about images

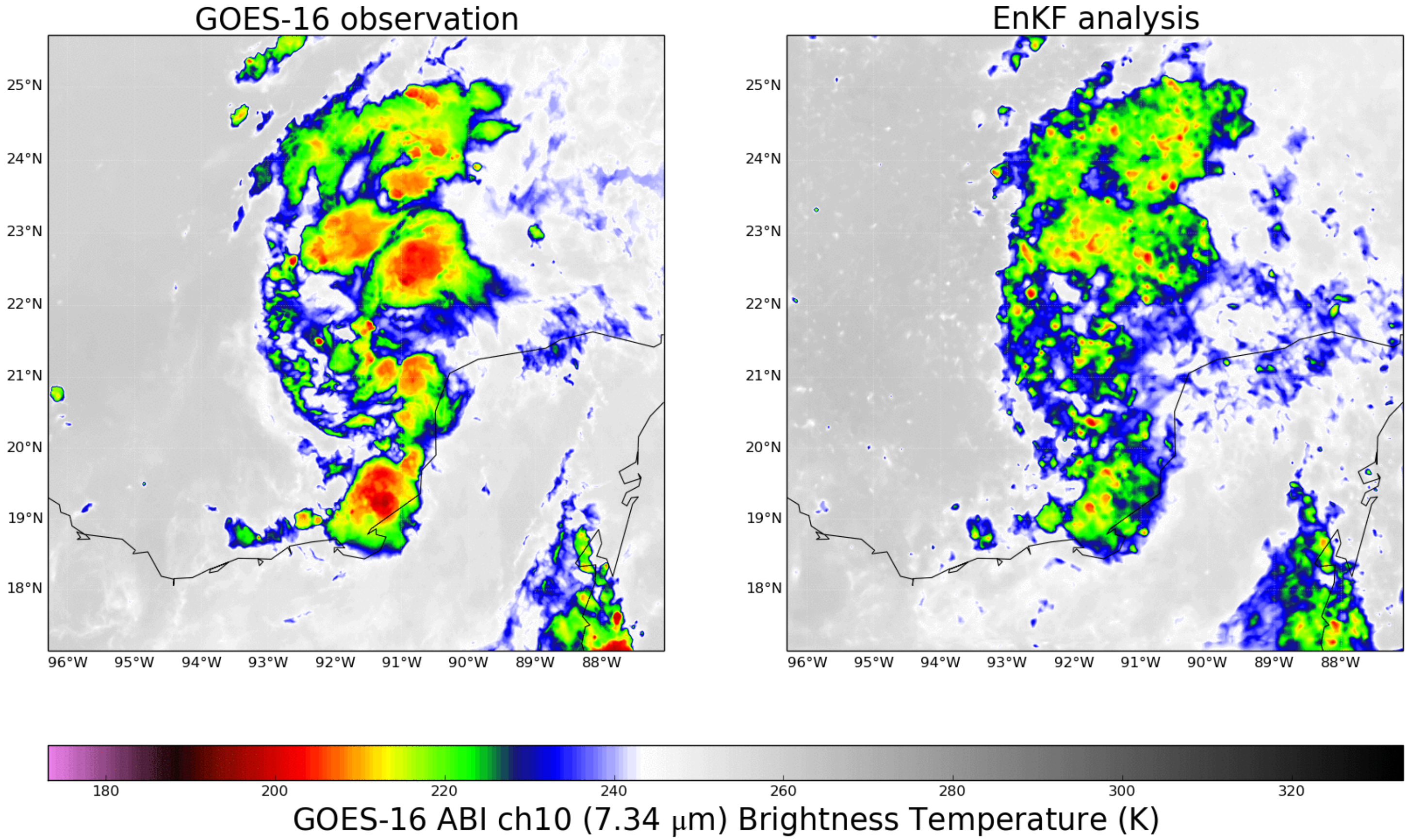
Throughout the 2008 hurricane season, the Texas Advanced Computing Center was an active participant in a NOAA research effort to develop next-generation hurricane models. Teams of scientists relied on TACC's Ranger supercomputer to test high-resolution ensemble hurricane models, and to track evacuation routes from data streams on the ground and from space. Using up to 40,000 processing cores at once, researchers simulated both global and regional weather models and received on-demand access to some of the most powerful hardware in the world enabling real-time, high-resolution ensemble simulations of the storm. This visualization of Hurricane Ike shows the storm developing in the gulf and making landfall on the Texas coast.

Credit: *Fuqing Zhang and Yonghui Weng, Pennsylvania State University; Frank Marks, NOAA; Gregory P. Johnson, Romy Schneider, John Cazes, Karl Schulz, Bill Barth, The University of Texas at Austin*

Assimilating All-sky GOES-R Radiances: Harvey (2017)

PSU WRF-EnKF, Δx=3km, ensemble size=60, channel 8, every 1h

Independent observations vs. EnKF analysis of channel 10



[2017-08-23_12:00]

What external data sources we have used?

- Observations and reanalysis from big trusted data centers: NCAR (RDA, field catalog, ...), NOAA, NASA, Navy, …, ECMWF, JMA, CMA*, …

  ----they are the easiest sources to handle

- Some international datasets from collaborative authorship or personal connections: CMA, PAGADA, HKO, Taiwan, …

  ---tricky but at least permanent archive exists

What external data sources we have used?

- Some more research-focused big modeling output from big modeling centers:

  ECMWF model output, NOAA experimental or operational forecasts, Panasonic Weather*, …

  These data either from operational models not routinely saved, or from partnership but unlikely these data will or have been archived by the providers

- Other individual researchers or collaborators

  ad-hoc but best effort in both sides

What are the venues our research group has used for storage, backup and/or archive so far?

- *Local group owned tape or disk spaces (~ 100TB)*
  - reliable but aging; many university PI gets from startup but hard to replace afterwards
  - Individual group members using cloud service to backup codes, small data
- *Share of routine supported and routinely backup-ed departmental servers (~100GB)*
  - varies from dept to dept, univ to univ
- *University discounted data center service*
  - limited allocation time cycle (Chuck, can you chime in?)
- *NCAR HPC computing (some data all the way to my postdoc life)*
  - allocation application renewal cycle, timeout issues?
- *NSF HPC center (Texas Advanced Computing Center)*
  - most data stored but not sure what is the limit; big lost of data due to TACC tape/disk failure or file corruption
- *NOAA Jet clusters*
  - allocation based, ad hoc renewal, reliable and free at present but uncertain pending on NOAA policy and regulations

What are our mandates in backing up data as a scientist for our published work using data?

Who is eventually responsible for the data cost and service after a few years?