Cyberinfrastructure: Opportunities and Challenges

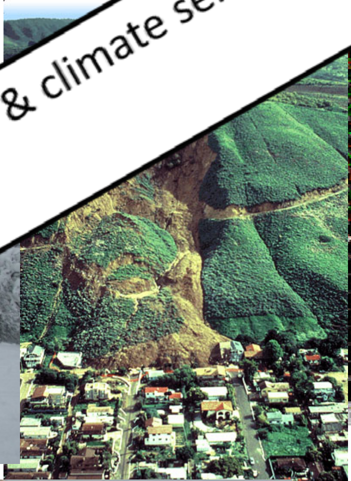
EarthCube Mesoscale Modeling Workshop

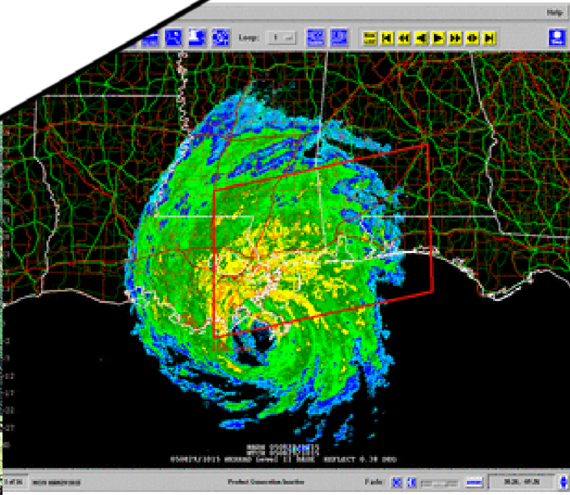
17 December2012

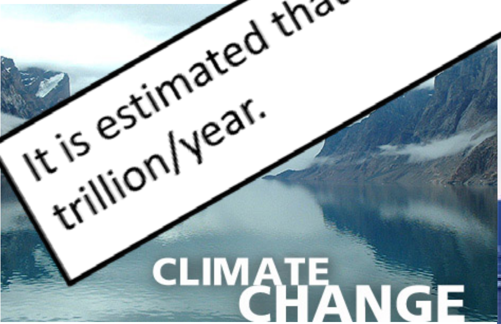
Boulder, CO

Mohan Ramamurthy
Unidata Program Center
UCAR Community Programs
Boulder, CO

Science and Society

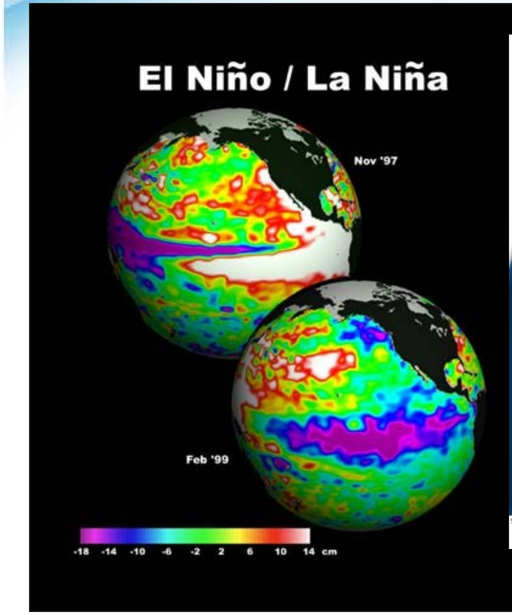


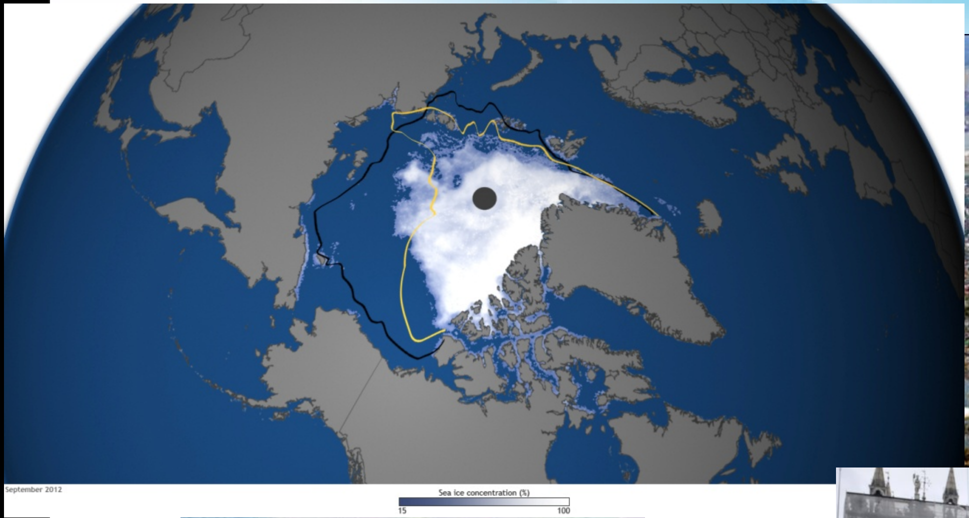


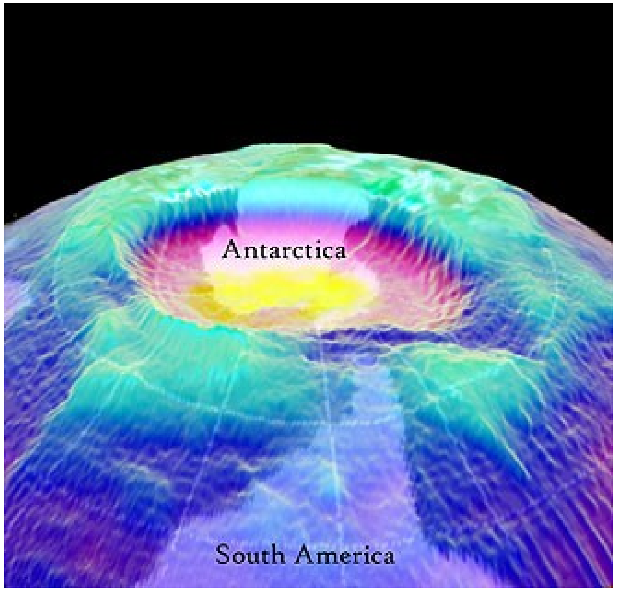


It is estimated that at least 1/3 of the U.S. GDP is weather & climate sensitive, a potential impact of $4 trillion/year.

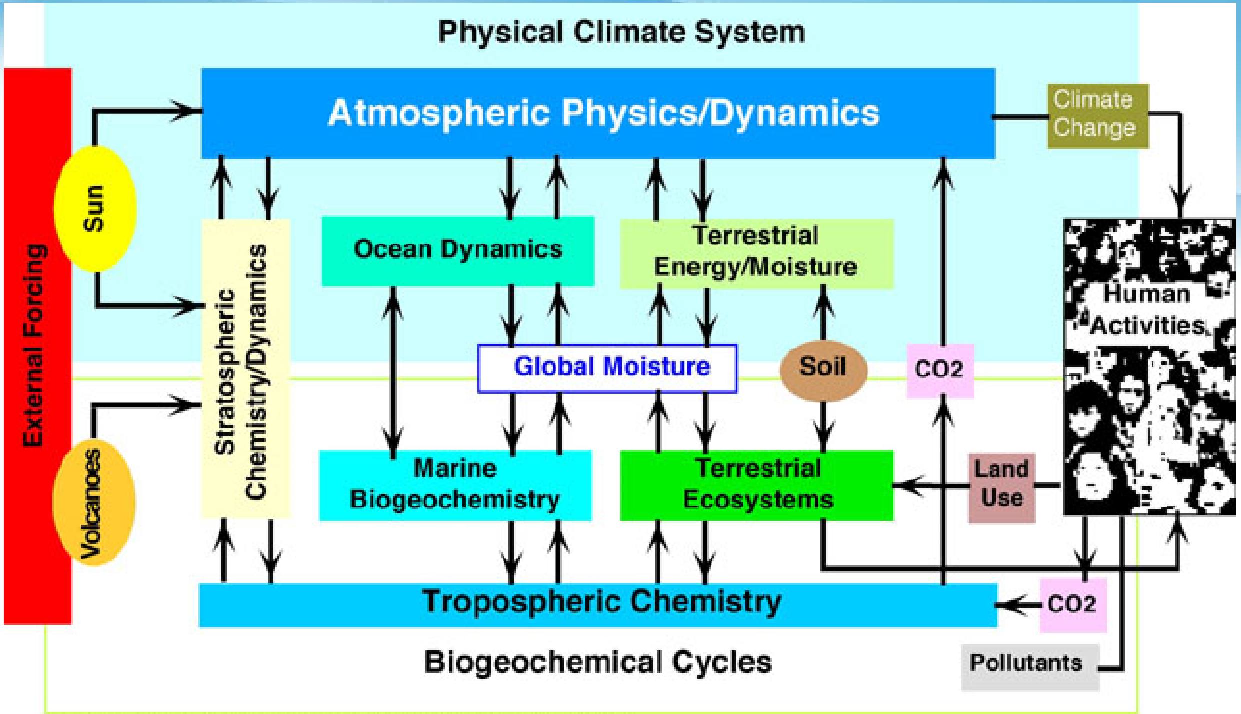
Grand Challenges are Global & Multidisciplinary







Earth System Science



(from Earth System Science: An Overview, NASA, 1988)

Need Cyberinfrastructure to support ESS thinking.

It requires data and information *integration* and knowledge *synthesis* across “systems” or domains.

Challenge: Providing the right data, in the right format, to the right application.

Networked Science

- Distributed knowledge communities working collaboratively as a virtual community to tackle problems never before possible.

TIGGE

THORPEX
A Global Atmospheric Research Programme

IPCC
INTERGOVERNMENTAL
PANEL ON
CLIMATE CHANGE

WMO UNEP

INTERNATIONAL POLAR YEAR
2007-2008

BIG DATA BUZZ

Fidelity
INVESTMENTS
BIG DATA

The Economist
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs

The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

GLOBAL PULSE

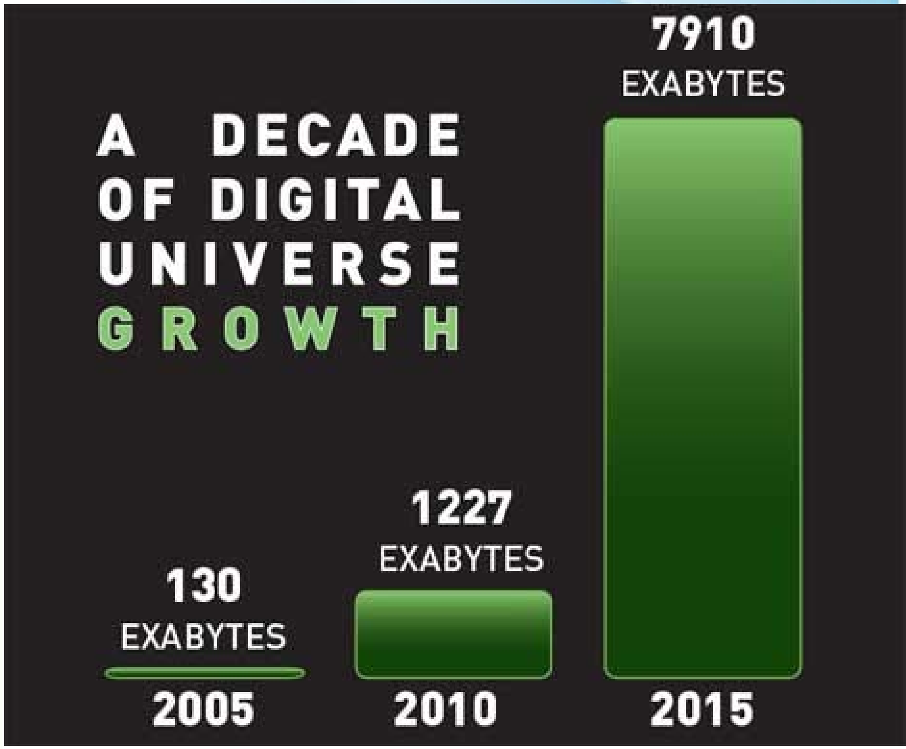
Big Data for Development:

Challenges & Opportunities

May 2012

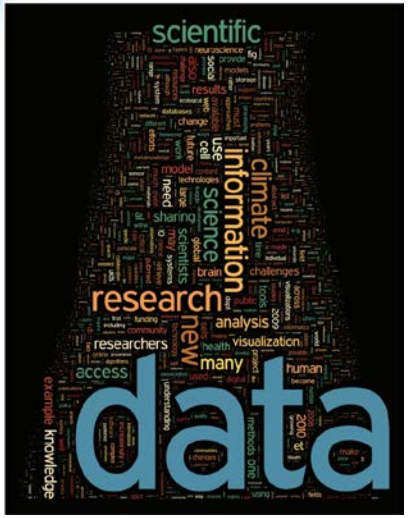
Digital Universe

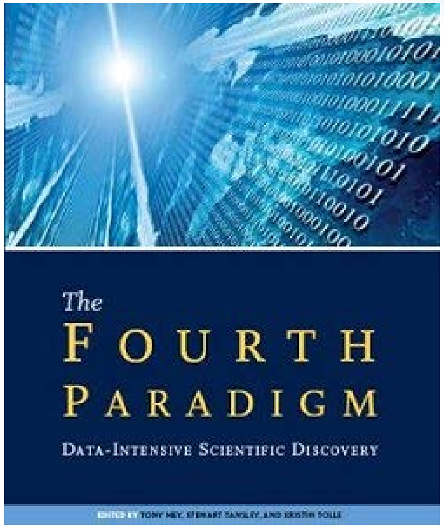
According to a study by IDC, 1.8 Zettabytes of information was created in 2011.

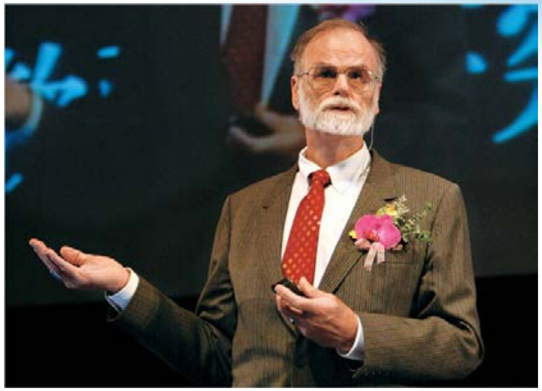


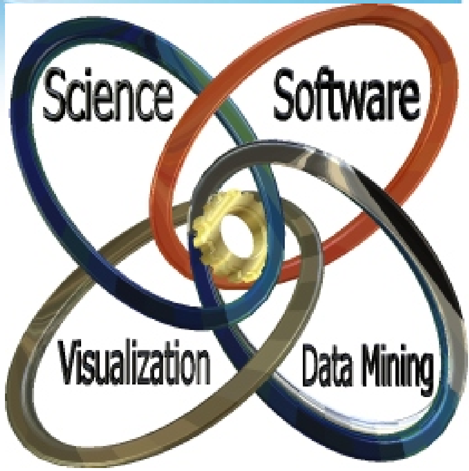
Source: IDC's Digital Universe Study, sponsored by EMC, June 2011

The Era of Data-Intensive Science



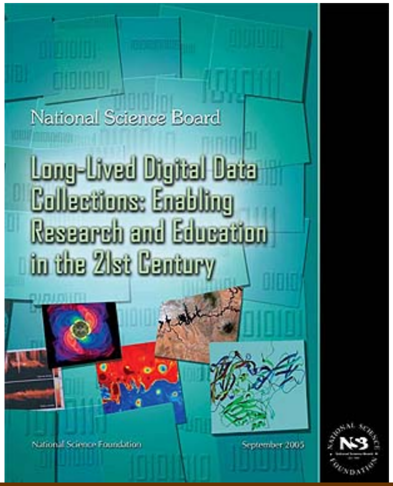

The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

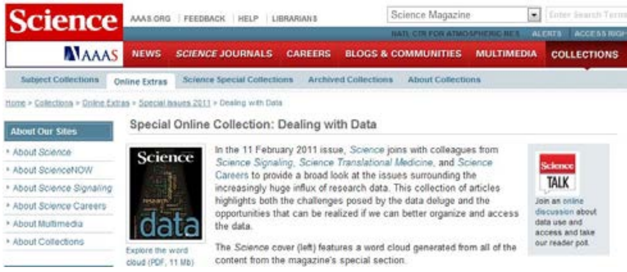




Scientific
COMPUTING
INFORMATION TECHNOLOGY FOR SCIENCE
LEADING EDGE COMPUTING
Application Frameworks for High Performance and Grid Computing

nature
THE BITTER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
SCIENCE IN THE PETABYTE ERA


National Science Board
Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century
National Science Foundation
September 2005



Special Online Collection: Dealing with Data

In the 11 February 2011 issue, *Science* joins with colleagues from *Science Signaling*, *Science Translational Medicine*, and *Science Careers* to provide a broad look at the issues surrounding the increasingly huge influx of research data. This collection of articles highlights both the challenges posed by the data deluge and the opportunities that can be realized if we can better organize and access the data.

The *Science* cover (left) features a word cloud generated from all of the content from the magazine's special section.

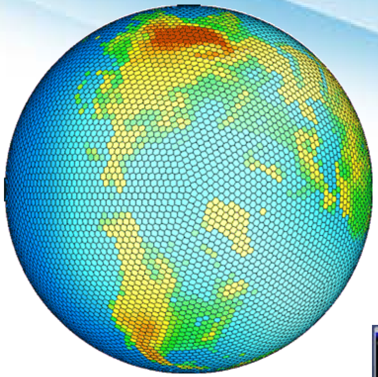
Explore the word cloud (PDF, 11 Mb)

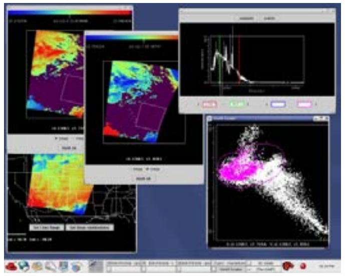
Join an online discussion about data use and access and take our reader poll.

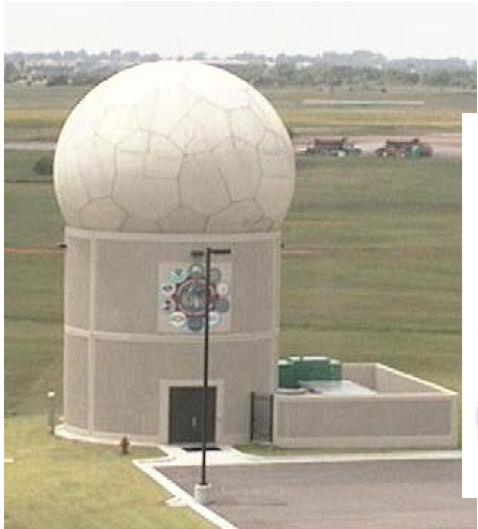
Data is the lifeblood of science, but we need to move from creating data to discovering knowledge.

“Sea of Data”

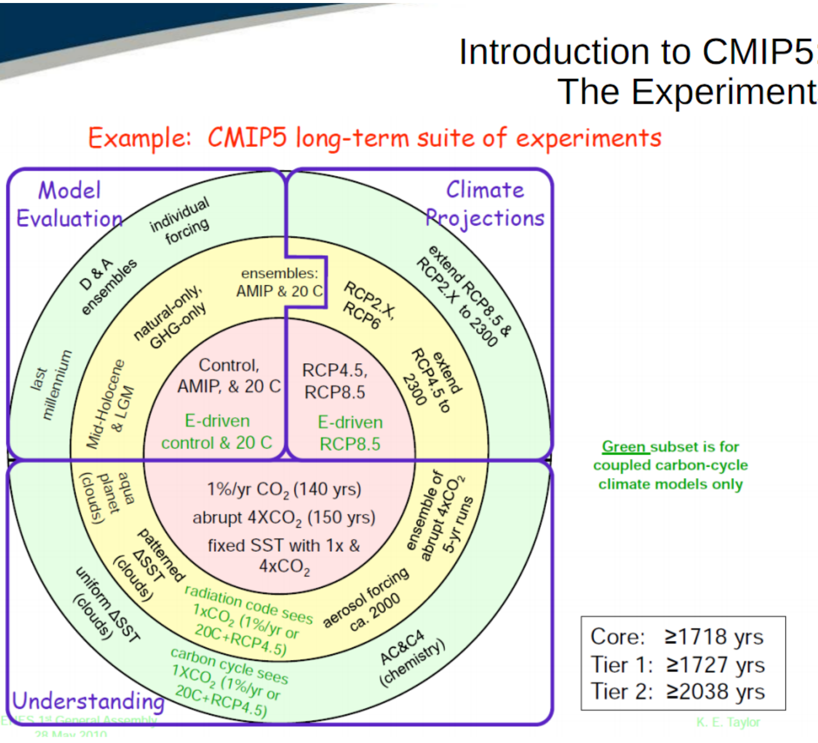
- GOES-R (2017)
- JPSS (2014)
  - ~3 Tb of data/day
- Phased Array Radar, with 20 to 30-second volume scans, compared with 5-7 mins. with current radars.
- Global, high-resolution coupled models integrated in ensemble mode from days to decades







CMIP-5 & IPCC Fifth Assessment



Introduction to CMIP5: The Experiments

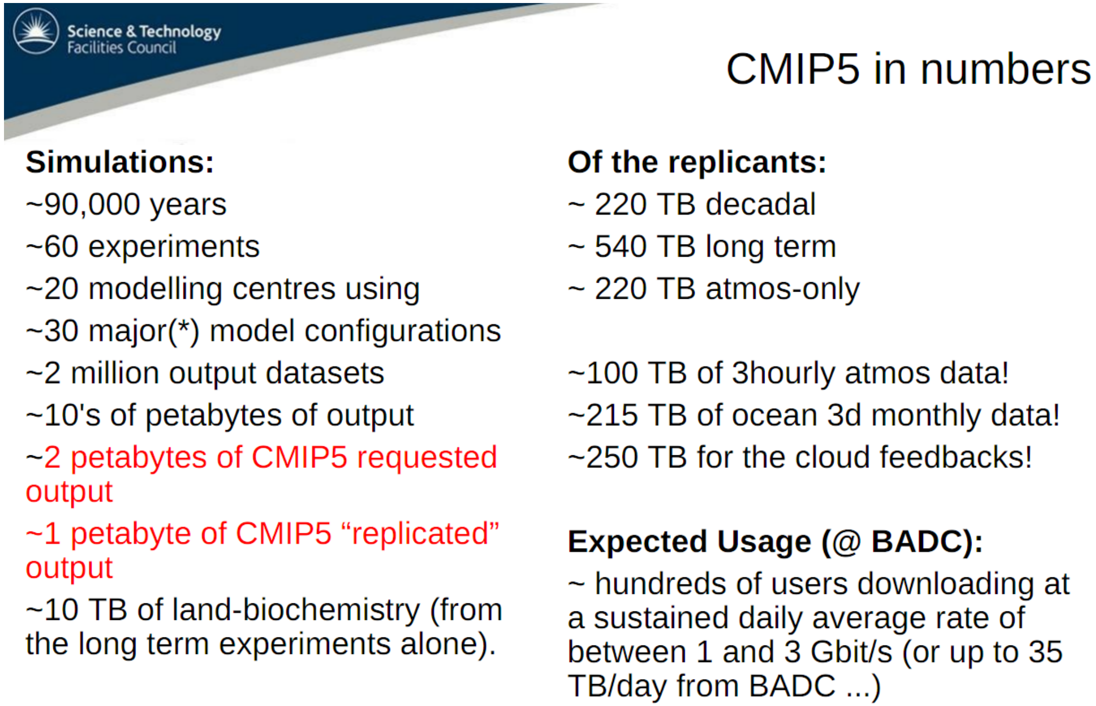
Example: CMIP5 long-term suite of experiments

Green subset is for coupled carbon-cycle climate models only

Core:	≥1718 yrs
Tier 1:	≥1727 yrs
Tier 2:	≥2038 yrs

ENES 1st General Assembly 28 May 2010

K. E. Taylor



Science & Technology Facilities Council

CMIP5 in numbers

Simulations:

~90,000 years

~60 experiments

~20 modelling centres using

~30 major(*) model configurations

~2 million output datasets

~10's of petabytes of output

~2 petabytes of CMIP5 requested output

~1 petabyte of CMIP5 "replicated" output

~10 TB of land-biochemistry (from the long term experiments alone).

Of the replicants:

~ 220 TB decadal

~ 540 TB long term

~ 220 TB atmos-only

~100 TB of 3hourly atmos data!

~215 TB of ocean 3d monthly data!

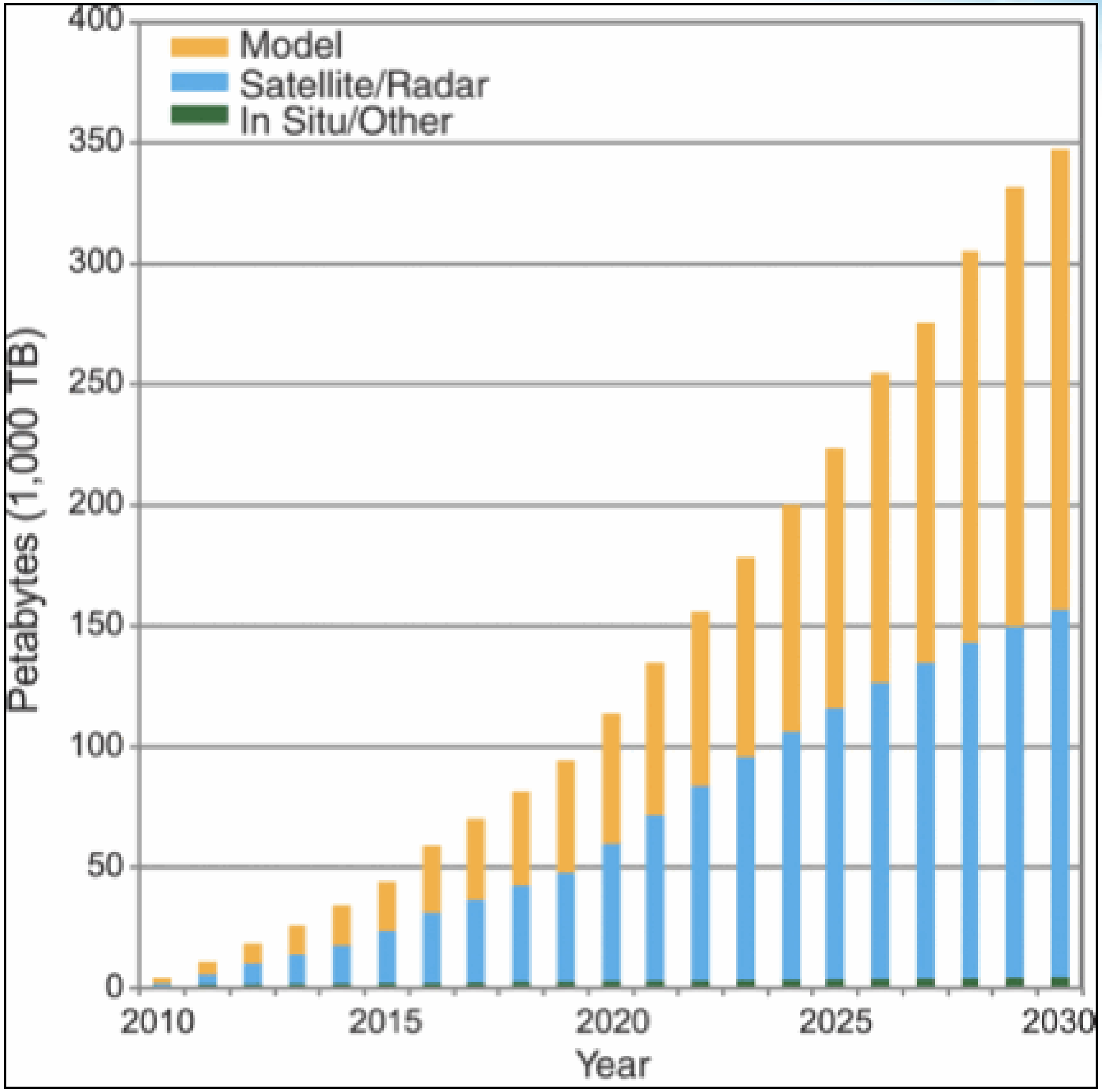
~250 TB for the cloud feedbacks!

Expected Usage (@ BADC):

~ hundreds of users downloading at a sustained daily average rate of between 1 and 3 Gbit/s (or up to 35 TB/day from BADC ...)

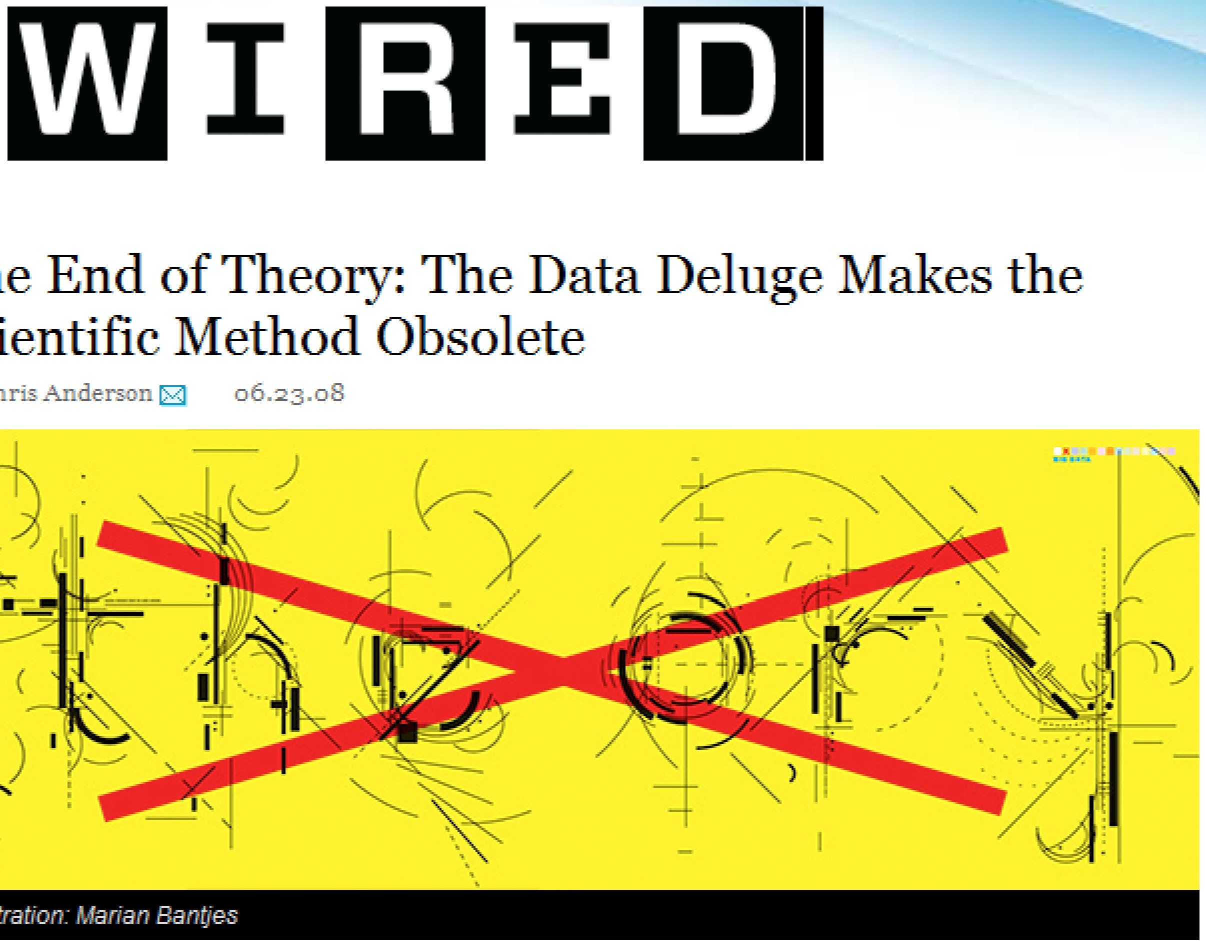
Source: Bryan Lawrence, British Atmospheric Data Centre

Expected Increase in Data Volume



Source: NCDC, NOAA

A Provocative Suggestion



WIRED

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson ✉ 06.23.08

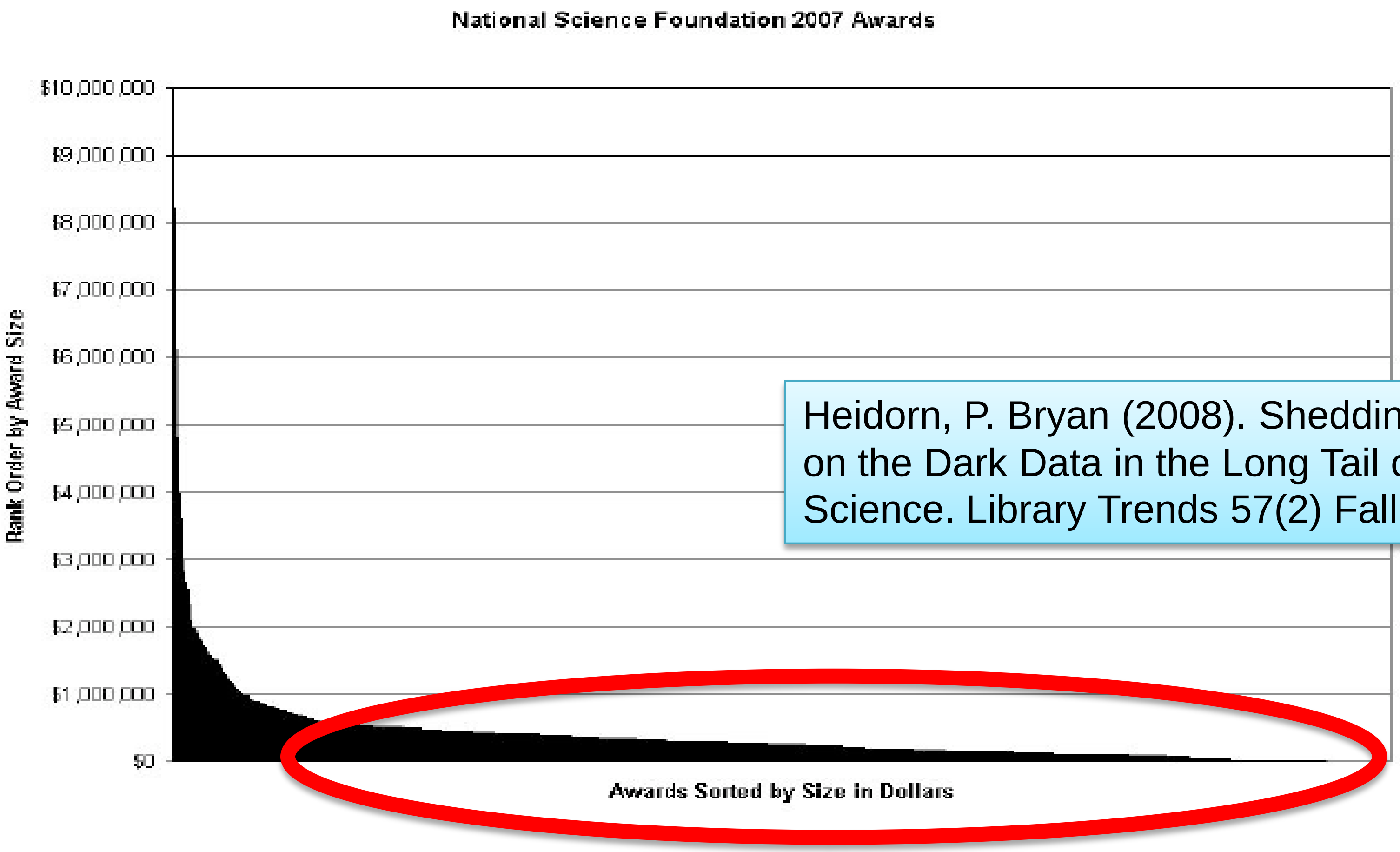
Illustration: Marian Bantjes

Wired, 23 June 2008 issue

Data Challenges: The Five "V"s

- **Volume:** Explosion of data
- **Variety:** Different types of data (e.g, multidisciplinary, societal information, etc.); interoperability
- **Velocity:** Speed of discovery, access, analysis, integration, and visualization;
- **Views:** Many consumers of data (e.g., researchers, educators, students, policy makers, social scientists, and the public); Diverse applications; Multiple devices;
- **Virtual Communities:** Globally distributed, different cultures, practices, and policies

The Long Tail Problem



Heidorn, P. Bryan (2008). Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57(2) Fall 2008 .

Slide courtesy of Kerstin Lehnert

The Long Tail Problem - contd

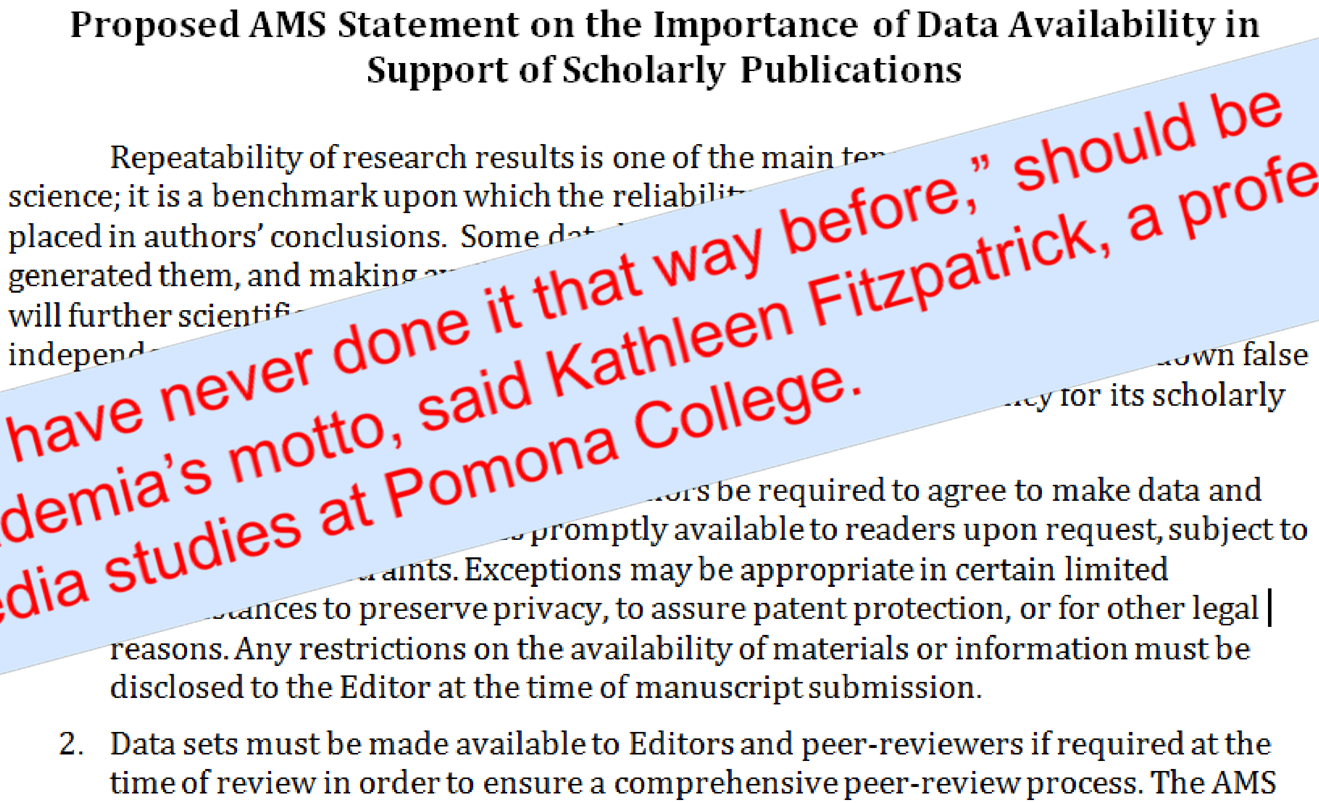
- By some estimates, only 5% of the data generated by individual PIs is shared
- There are many reasons for this – both sociological as well as technical
  - Lack of incentives
  - Inadequate resources
  - Additional burden or unfunded mandate
  - Protectiveness – PIs don't want to be "scooped"
  - Technical challenges
  - Absence of local or community data repositories
- Need to give PIs the tools required for sharing their data; also need tools for adding metadata
  - Need to create incentives
  - Need to change the culture

NSF Data Management Plan

- All proposals must include a Data Management Plan.
- Plan should describe how the proposal will conform to NSF policy on dissemination and sharing of research results.
- Plan will be reviewed as part of the intellectual merit and/or broader impacts of the proposal depending on the proposal intent.
- NSF will not permit submission of a proposal that is missing a data management plan.

Data Citation: The Next Frontier?

- Scientific publications should be accompanied by data, algorithms, models, and parameters – need comprehensive data citation. Need transparency. Important for reproducibility.
- This is not just a technical challenge, but it is also a major cultural and organizational challenge.



Proposed AMS Statement on the Importance of Data Availability in Support of Scholarly Publications

Repeatability of research results is one of the main ten

science; it is a benchmark upon which the reliabilit

placed in authors' conclusions. Some dat

generated them, and making av

will further scientif

independ

own false

cy for its scholarly

ors be required to agree to make data and

promptly available to readers upon request, subject to

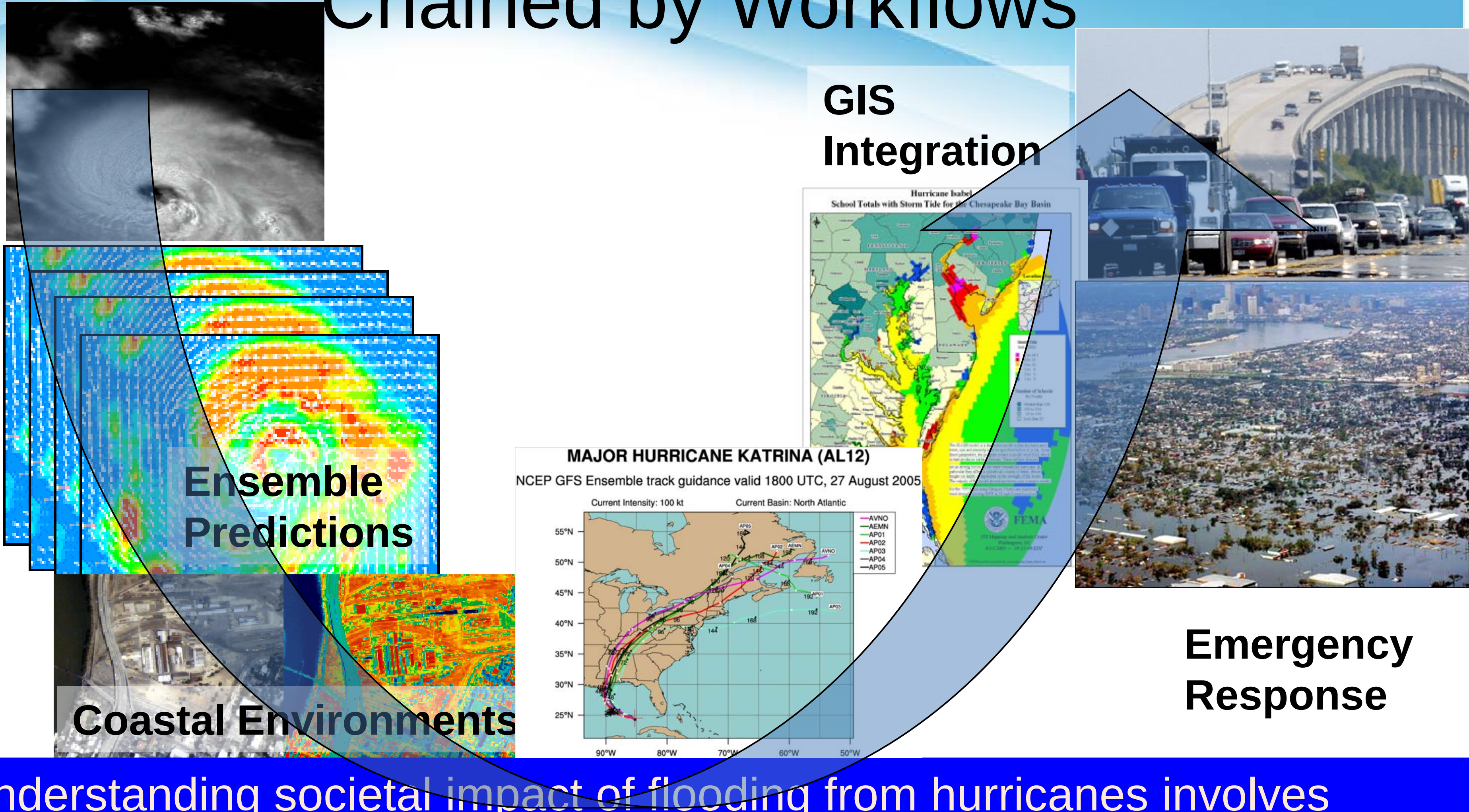
aints. Exceptions may be appropriate in certain limited

tances to preserve privacy, to assure patent protection, or for other legal reasons. Any restrictions on the availability of materials or information must be disclosed to the Editor at the time of manuscript submission.

2. Data sets must be made available to Editors and peer-reviewers if required at the time of review in order to ensure a comprehensive peer-review process. The AMS

- “We have never done it that way before,” should be academia's motto, said Kathleen Fitzpatrick, a professor of media studies at Pomona College.

End-to-End Data Services Chained by Workflows



Hurricane Isabel
School Totals with Storm Tide for the Chesapeake Bay Basin
FEMA
MAJOR HURRICANE KATRINA (AL12)
NCEP GFS Ensemble track guidance valid 1800 UTC, 27 August 2005
Current Intensity: 100 kt
Current Basin: North Atlantic
AVNO
AEMN
AP01
AP02
AP03
AP04
AP05
55°N
50°N
45°N
40°N
35°N
30°N
25°N
90°W
80°W
70°W
60°W
50°W

GIS Integration

Ensemble Predictions

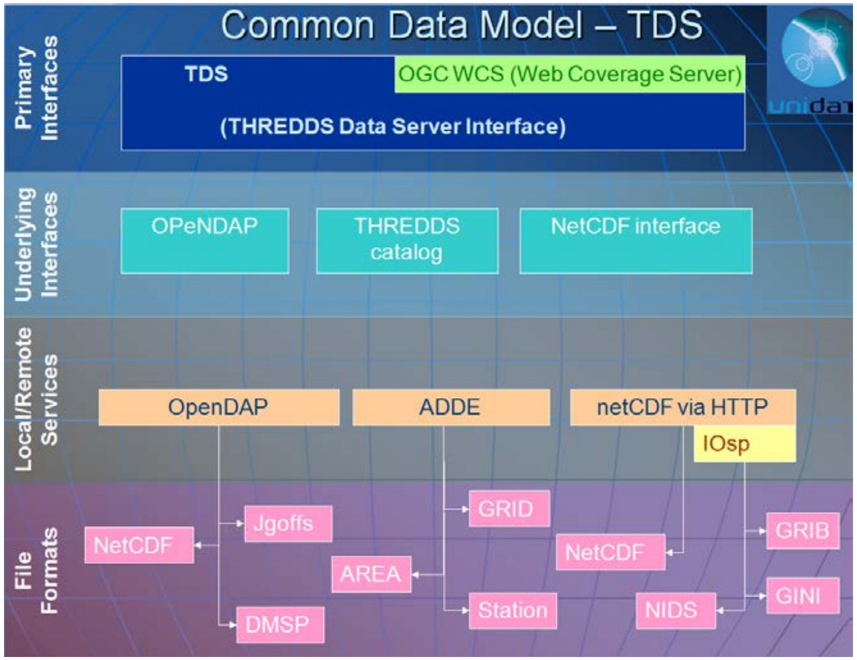
Coastal Environments

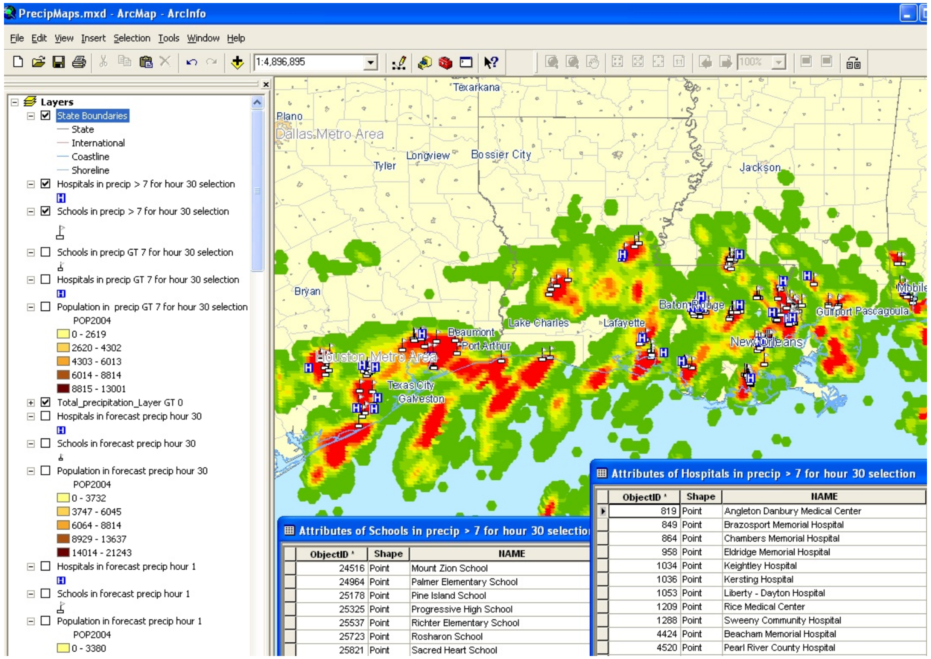
Emergency Response

Understanding societal impact of flooding from hurricanes involves integrating data from atmospheric sciences, oceanography, hydrology, geology, and social sciences and interfacing the results with decision support systems.

GIS Integration: An Enormous Opportunity

- Need geospatially-enabled cyberinfrastructure so that information can be integrated for location-based understanding of events, processes, interactions, and impacts. (May Yuan)
- GIS integration should not be an after thought. Scientific data systems need to directly enable GIS tools.





Layers
- ☑ State Boundaries
  - State
  - International
  - Coastline
  - Shoreline
- ☑ Hospitals in precip > 7 for hour 30 selection
- ☑ Schools in precip > 7 for hour 30 selection
- □ Schools in precip GT 7 for hour 30 selection
- □ Hospitals in precip GT 7 for hour 30 selection
- □ Population in precip GT 7 for hour 30 selection
  - POP2004
  - 0 - 2619
  - 2620 - 4302
  - 4303 - 6013
  - 6014 - 8814
  - 8815 - 13001
- ☑ Total_precipitation_Layer GT 0
- □ Hospitals in forecast precip hour 30
- □ Schools in forecast precip hour 30
- □ Population in forecast precip hour 30
  - POP2004
  - 0 - 3732
  - 3747 - 6045
  - 6064 - 8814
  - 8929 - 13637
  - 14014 - 21243
- □ Hospitals in forecast precip hour 1
- □ Schools in forecast precip hour 1
- □ Population in forecast precip hour 1
  - POP2004
  - 0 - 3380

Attributes of Schools in precip > 7 for hour 30 selectio

ObjectID *	Shape	NAME
24516	Point	Mount Zion School
24964	Point	Palmer Elementary School
25178	Point	Pine Island School
25325	Point	Progressive High School
25537	Point	Richter Elementary School
25723	Point	Rosharon School
25821	Point	Sacred Heart School

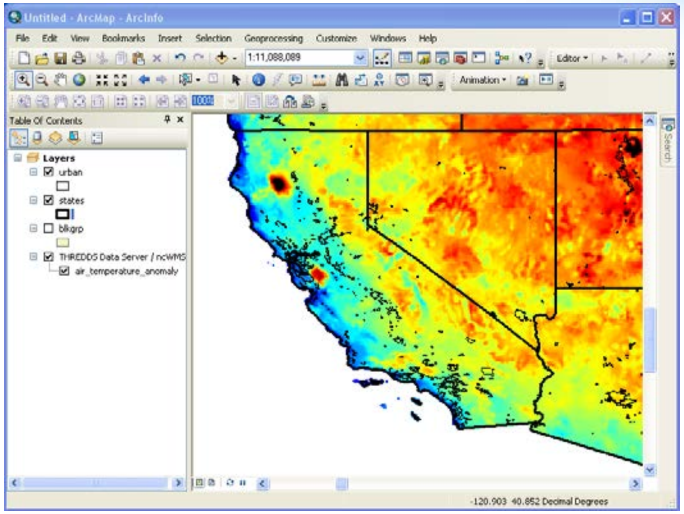
Attributes of Hospitals in precip > 7 for hour 30 selection

ObjectID *	Shape	NAME
819	Point	Angleton Danbury Medical Center
849	Point	Brazosport Memorial Hospital
864	Point	Chambers Memorial Hospital
958	Point	Eldridge Memorial Hospital
1034	Point	Keightley Hospital
1036	Point	Kersting Hospital
1053	Point	Liberty - Dayton Hospital
1209	Point	Rice Medical Center
1288	Point	Sweeny Community Hospital
4424	Point	Beacham Memorial Hospital
4520	Point	Pearl River County Hospital

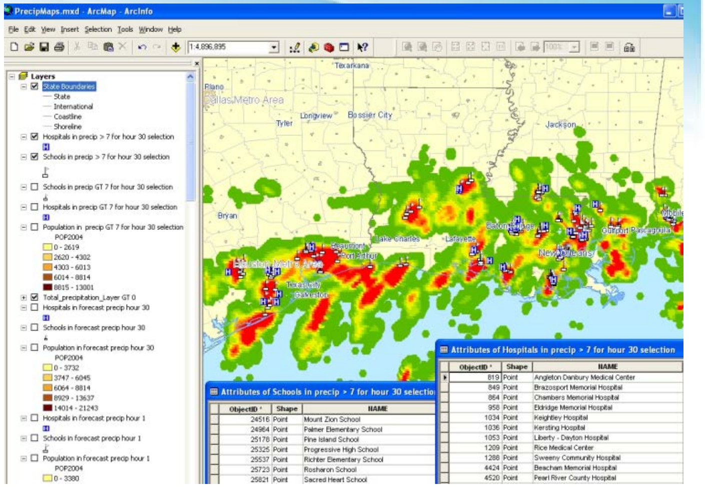
GIS and THREDDS Data Server working together

NCAR model air temperature anomaly

- Data from NCAR climate model
- TDS software from Unidata
- WMS software from University of Reading
- Server run by NCAR GIS program
- Map and analysis software from ESRI


Untitled - ArcMap - ArcInfo
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:11,088,089
Editor
Animation
Table Of Contents
Layers
urban
states
blkgrp
THREDDS Data Server / ncWMS
air_temperature_anomaly
-120.903 40.852 Decimal Degrees

NAM output displayed by ArcGIS


PrecipMaps.mxd - ArcMap - ArcInfo
File Edit View Insert Selection Tools Window Help
1:4,896,895
Layers
State Boundaries
State
International
Coastline
Shoreline
Hospitals in precip > 7 for hour 30 selection
Schools in precip > 7 for hour 30 selection
Schools in precip GT 7 for hour 30 selection
Hospitals in precip GT 7 for hour 30 selection
Population in precip GT 7 for hour 30 selection
POP2004
0 - 2619
2620 - 4302
4303 - 6013
6014 - 8814
8815 - 13001
Total_precipitation_Layer GT 0
Hospitals in forecast precip hour 30
Schools in forecast precip hour 30
Population in forecast precip hour 30
POP2004
0 - 3732
3747 - 6045
6064 - 8814
8929 - 13637
14014 - 21243
Hospitals in forecast precip hour 1
Schools in forecast precip hour 1
Population in forecast precip hour 1
POP2004
0 - 3380
Texarkana
Plano
Dallas Metro Area
Longview
Bossier City
Tyler
Jackson
Bryan
Lake Charles
Lafayette
Baton Rouge
New Orleans
Texas City
Galveston

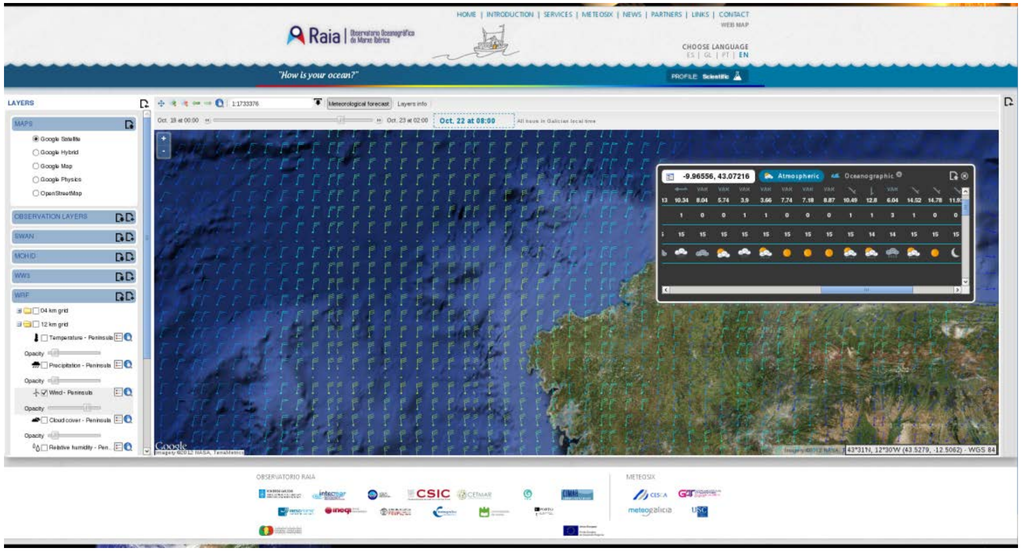
Attributes of Schools in precip > 7 for hour 30 selection

ObjectID *	Shape	NAME
24516	Point	Mount Zion School
24964	Point	Palmer Elementary School
25176	Point	Pine Island School
25325	Point	Progressive High School
25537	Point	Richter Elementary School
25723	Point	Rosharon School
25821	Point	Sacred Heart School

Attributes of Hospitals in precip > 7 for hour 30 selection

ObjectID *	Shape	NAME
819	Point	Angleton Danbury Medical Center
849	Point	Brazosport Memorial Hospital
864	Point	Chambers Memorial Hospital
958	Point	Eldridge Memorial Hospital
1034	Point	Keightley Hospital
1036	Point	Kersting Hospital
1053	Point	Liberty - Dayton Hospital
1209	Point	Rice Medical Center
1288	Point	Sweeny Community Hospital
4424	Point	Beacham Memorial Hospital
4520	Point	Pearl River County Hospital

WRF output displayed by Google Maps



“Cloud” Computing

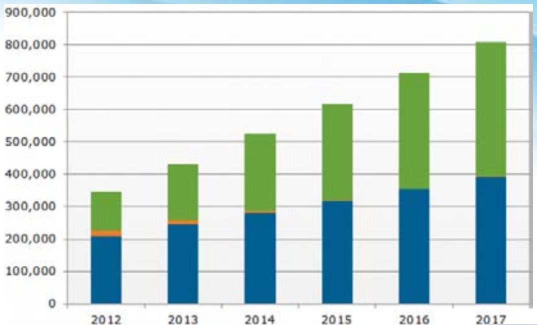
- Data volumes are too large to bring all of the data to your local environment
- Need to keep them close to the point of origin or dissemination
- Will need to move more of the processing, applications, and computations on to the server (e.g., GDS, NCO, etc.).
- Impractical to store/serve data in multiple formats – so need built-in translators, brokers and mediation services.
- Industry is well ahead of the science world in dealing with Big Data, Cloud Computing, and Virtualization (e.g., SaaS, Application Service Providers, MapReduce, Hadoop, etc.)

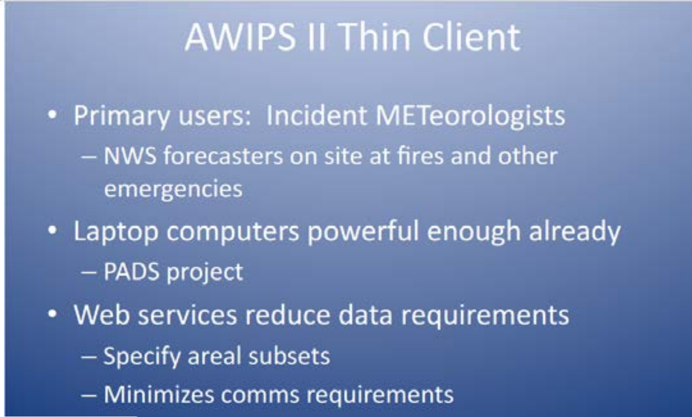
Virtualization

- **Virtualization** is a tool that has many technical uses, most of which have nothing to do with the cloud.
- Virtualization allows the use of a single piece of physical hardware, to perform the work of many. e.g., multiple operating system instances running on one hardware device are far more economical than a single piece of hardware for every server task.
- You can do virtualization on your laptop or desktop, although it is unlikely that either will be a cloud server.

Mobile Computing Devices

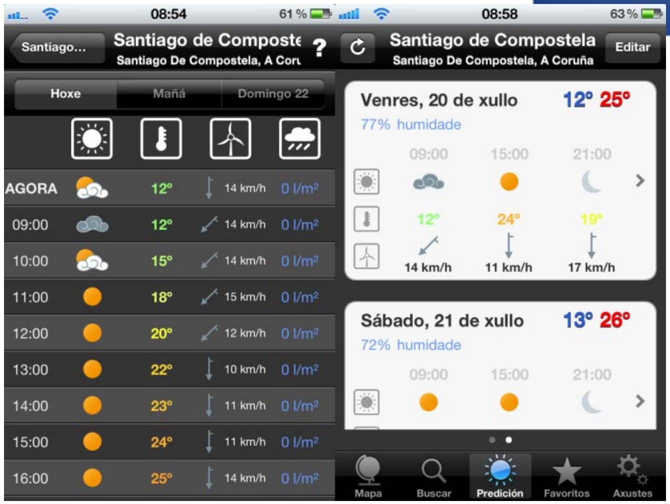
- Smartphones, tablets, laptops, etc.
- Ubiquitous access to network (anywhere, any place, etc.), with occasional offline needs
- Tablet applications, using thin client-server approaches, are currently being developed by ESRI, NWS, and others.





AWIPS II Thin Client

- Primary users: Incident METeorologists
  - NWS forecasters on site at fires and other emergencies
- Laptop computers powerful enough already
  - PADS project
- Web services reduce data requirements
  - Specify areal subsets
  - Minimizes comms requirements



08:54 61 %

Santiago... **Santiago de Composte** ?
Santiago De Compostela, A Coru

Hoxe | Mañá | Domingo 22

AGORA	12°	14 km/h	0 l/m²
09:00	12°	14 km/h	0 l/m²
10:00	15°	14 km/h	0 l/m²
11:00	18°	15 km/h	0 l/m²
12:00	20°	12 km/h	0 l/m²
13:00	22°	10 km/h	0 l/m²
14:00	23°	11 km/h	0 l/m²
15:00	24°	11 km/h	0 l/m²
16:00	25°	14 km/h	0 l/m²

08:58 63 %

Santiago de Compostela Editar
Santiago De Compostela, A Coruña

Venres, 20 de xullo 12° 25°
77% humidade

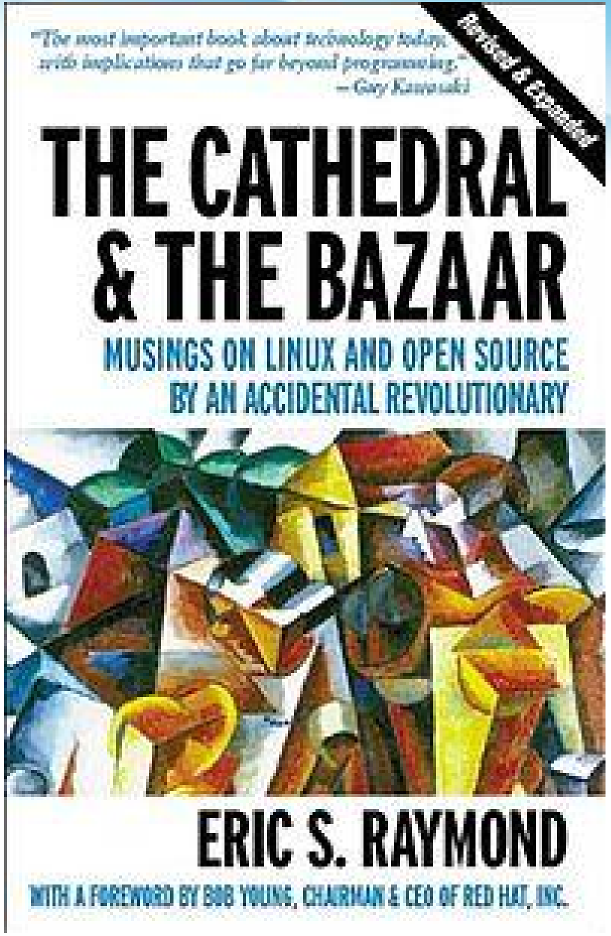
09:00	15:00	21:00
12°	24°	19°
14 km/h	11 km/h	17 km/h

Sábado, 21 de xullo 13° 26°
72% humidade

09:00	15:00	21:00

Open Source Software Development in the Geosciences

- The software community has many successful examples of Open Source Software Development (Linux, Mozilla, Apache,
- The atmospheric science community has had many decades of experience with community modeling efforts (NCAR GCM, CCSM, CESM, MM5, and WRF).
- Beyond that, there has been little attention paid to Open Source Development in our field.



"The most important book about technology today, with implications that go far beyond programming."
—Guy Kawasaki

Revised & Expanded

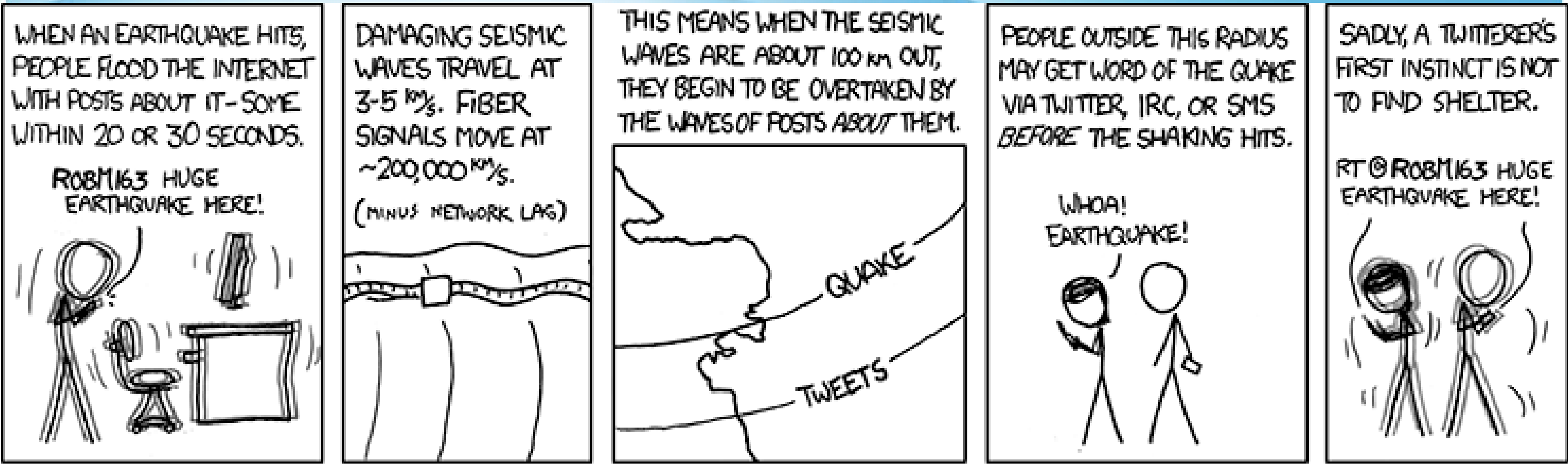
THE CATHEDRAL & THE BAZAAR

MUSINGS ON LINUX AND OPEN SOURCE BY AN ACCIDENTAL REVOLUTIONARY

ERIC S. RAYMOND

WITH A FOREWORD BY BOB YOUNG, CHAIRMAN & CEO OF RED HAT, INC.

Social Networking & Citizen Science


WHEN AN EARTHQUAKE HITS, PEOPLE FLOOD THE INTERNET WITH POSTS ABOUT IT—SOME WITHIN 20 OR 30 SECONDS.
ROBM163 HUGE EARTHQUAKE HERE!
DAMAGING SEISMIC WAVES TRAVEL AT 3-5 KM/S. FIBER SIGNALS MOVE AT ~200,000 KM/S.
(MINUS NETWORK LAG)
THIS MEANS WHEN THE SEISMIC WAVES ARE ABOUT 100 KM OUT, THEY BEGIN TO BE OVERTAKEN BY THE WAVES OF POSTS ABOUT THEM.
QUAKE
TWEETS
PEOPLE OUTSIDE THIS RADIUS MAY GET WORD OF THE QUAKE VIA TWITTER, IRC, OR SMS BEFORE THE SHAKING HITS.
WHOA! EARTHQUAKE!
SADLY, A TWITTERER'S FIRST INSTINCT IS NOT TO FIND SHELTER.
RT @ROBM163 HUGE EARTHQUAKE HERE!

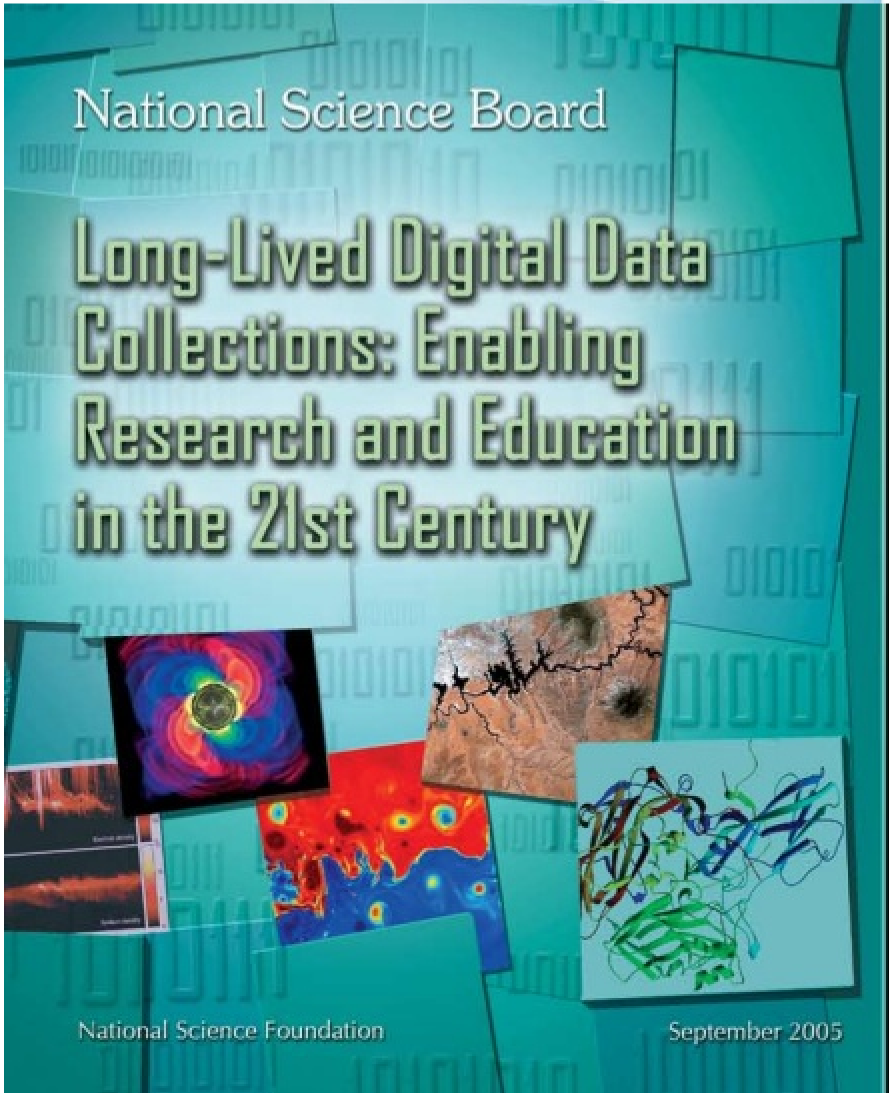
Wikis, Blogs, Twitter, Facebook, and Linked-In in increasing use

Mobile sensors and crowd sourcing and validation, using social network

Beyond being a communication tool, how do you use Social Networking for data collection, aggregation, and sharing?

Education, Training and Workforce Development

- Education and workforce requirements are not aligned;
- Computational, geospatially, and data literacy needs to be advanced;
- Far too few women and minorities in IT disciplines and in the workforce;



National Science Board

Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century

National Science Foundation

September 2005

Unidata Strategic Plan

Unidata 2020: Geoscience at the Speed of Thought *through accelerated data discovery, access, analysis, and visualization.*

Mission: To transform the geosciences community, research, and education by providing innovative data services and tools.

Reduce *“data friction”*, lower the barriers, and reduce “time to research”

Accelerate *user workflows (manual or automated)*

Contribute toward flipping the 4:1 ratio (the current 80-20 situation)

Concluding Remarks

- We live in an exciting era in which advances in computing and communication technologies, coupled with a new generation of geoinformatics, are accelerating scientific research, creating new knowledge, and leading to new discoveries at an unprecedented rate.
- Workshops like this play an ever more important role in bringing people together to understand our community needs, examine progress, look at opportunities, address challenges, and foster new partnerships.
- Thank you!