Challenges and Opportunities with NOAA's Big Data Partnership

Dr. Edward J. Kearns
NOAA's National Centers for Environmental Information

SECOORA Stakeholder Meeting
NCSU, Raleigh NC 19 May 2016

Outline

- What's the problem(s) NOAA is trying to solve? Some Big Data Challenges
- Origins and progress of the Big Data Partnership research activity at NOAA
- Some lessons learned about the Challenges from the recent Big Data activities
  - Marine data challenges
- Questions and Discussion

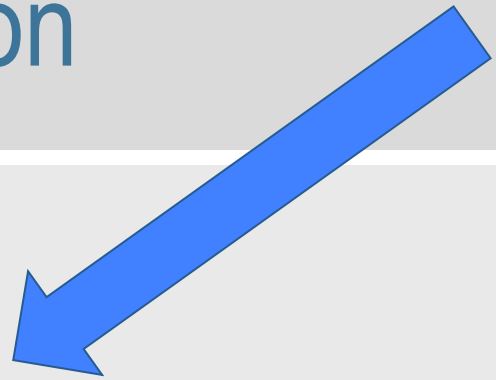
Challenges : Preservation

- Collection and preservation of data?
  - Rapidly increasing volume is not the problem…
  - Supporting the labor necessary to steward and curate collections of data is a problem.
  - Recent advances in standards, formats, descriptions are very helpful.
  - Data security in Earth Sciences to be a future problem?
- Access to the data?
- Utilization of the data?

Challenges : Access

- Collection and preservation of data?
- Discovery of and access to the data?
  - Access for “designated communities” is probably sufficient for their research
  - Network bandwidth is a problem -- if we’re going to keep moving the data when we need to use them.
  - Discovery and access outside of a desginated community is a problem. Catalogs? Free text search?
- Utilization of the data?

Challenges : Utilization



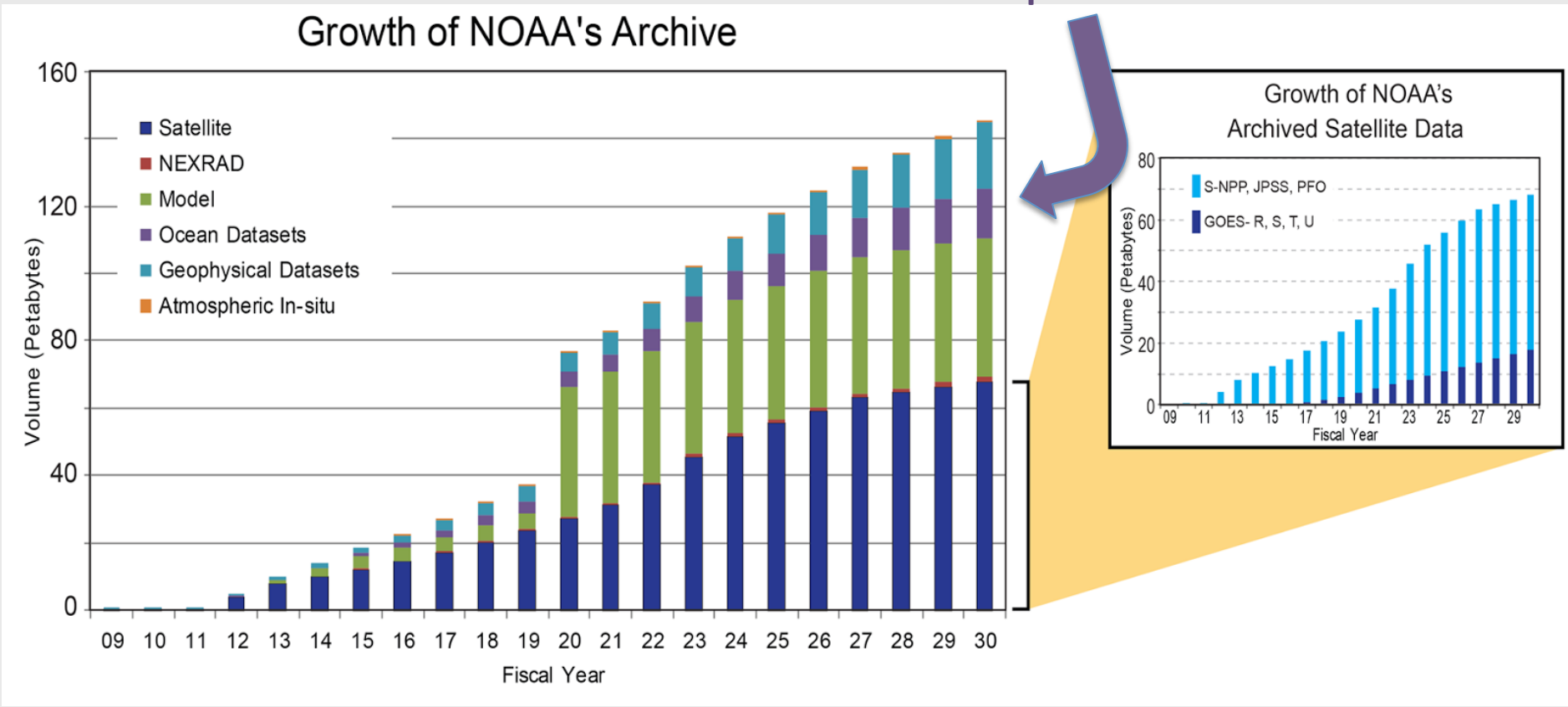
- Collection and preservation of data?
- Access to the data?
- Utilization of the data?
  - The true Big Data problem is about ease of use.
  - Portals versus Platforms
  - Applications require expert interpretation.
  - Big Data applications across fields or industries require diverse expertise – how is this expertise provided?
  - Open Data versus Proprietary Data issues: do the data or services hold the most value?
    - Pay for data? Or pay for access & services?
      - Value of curated collections, e.g. Music Services
  - Growth of platforms to use data to answer questions

National Centers for Environmental Information's User Profile : Most want "The Answer"

Fraction (%)	Typical User	Data or Info Need	Preferred Format	Access Volume	Access Frequency
70	General business, media, public	Qualitative	Point-and-click, graphics, assessments	Low	High
15	Researchers, business consultants	Quantitative	Digital downloads	High	Low
15	Value-added Providers (database scrapers)	Quantitative	Digital downloads machine to machine	Low	High

Archival Storage Projections for NOAA data

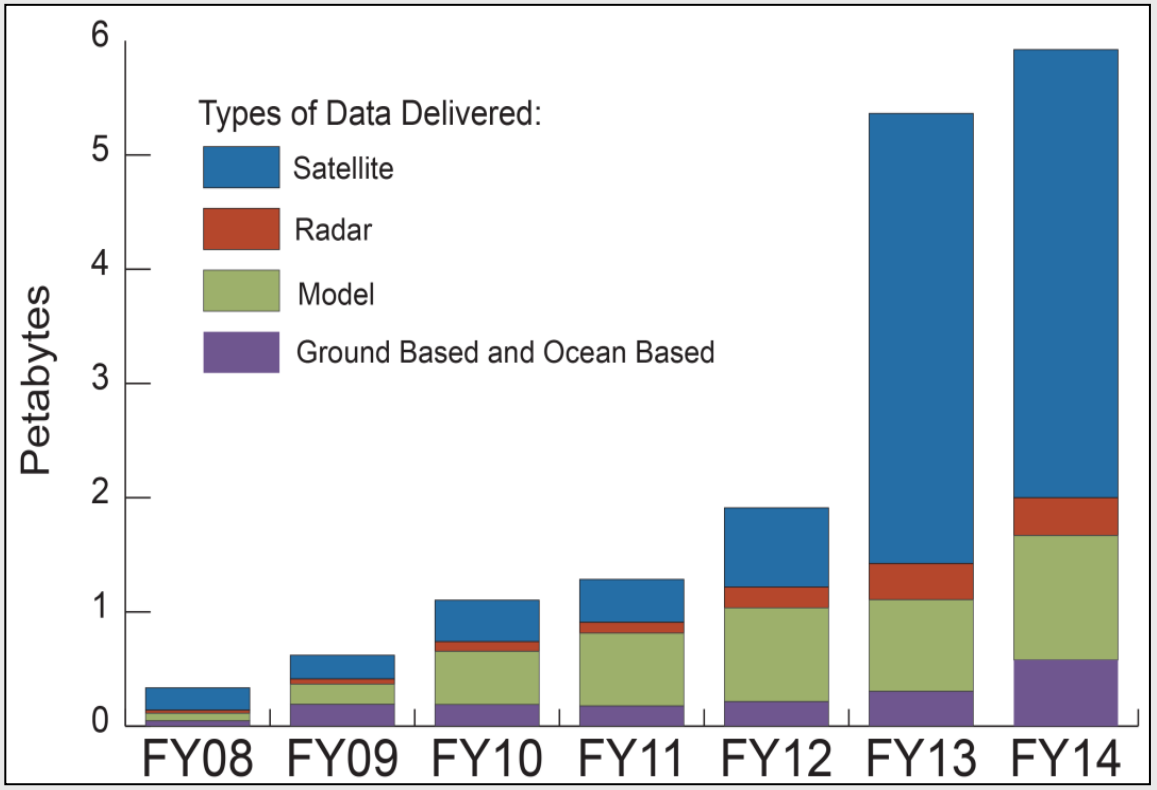
Ocean-specific datasets



Courtesy Steve Del Greco/Ken Casey, NCEI

Accelerating User Demand for NOAA data

- NOAA's National Centers for Environmental Information (NCEI) alone is now serving >8 PB of data annually
- Servicing over 20,000 personal contacts across many sectors
- 2.6 billion web hits in FY14 with 19 million users
- **Significant load on NOAA infrastructure ($$$)**



Intro to NOAA Big Data Partnership (BDP)

1. NOAA has a lot of data, which is expensive to store and disseminate.

2. Due to a variety of accessibility issues, much of NOAA's environmental data are also **under-utilized** – especially beyond the expert community.

3. **<u>There is untapped economic value in all that data.</u>**

4. That value can be leveraged to improve accessibility and pay for staging of that data on the public cloud, where people and organizations of all kinds can innovate as part of a market ecosystem.

April 2015: Partnership Announcement

Unusual, no-net-cost proposition

- Could the value inherent in NOAA's datasets support the cost of their distribution?

Enthusiastic, cross-industry response

- Interest from 200 companies
- 70+ responses to NOAA's "request for information"

<u>NOAA is using Cooperative Research and Development Agreements (CRADAs), not typical acquisitions</u>

Selection/announcement made on 21 Apr 2015

COMMERCE.GOV

Apr 21 2015

Home

U.S. Secretary of Commerce Penny Pritzker Announces New Collaboration to Unleash the Power of NOAA's Data

NOAA joins Amazon Web Services, Google Cloud Platform, IBM, Microsoft Corp., and the Open Cloud Consortium for big data project

Environment | Data | American Meteorological Society | America's Data Agency | Open Cloud Consortium

Posted at 10:28 AM

FOR IMMEDIATE RELEASE
Tuesday, April 21, 2015

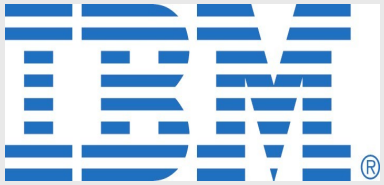
Office of Public Affairs
202-482-4883
publicaffairs@doc.gov

U.S. Secretary of Commerce Penny Pritzker today announced a big data project with Amazon Web Services, Google Cloud Platform, IBM, Microsoft Corp., and the Open Cloud Consortium to explore ways of bringing the Department closer to its goal of unleashing its vast resources of environmental data and delivering on one of the Department's key priorities – transforming Department data capabilities and supporting a data-driven economy. Secretary Pritzker made the announcement during a keynote address at the American Meteorological Society's Washington Forum.

BDP CRADA Collaborators

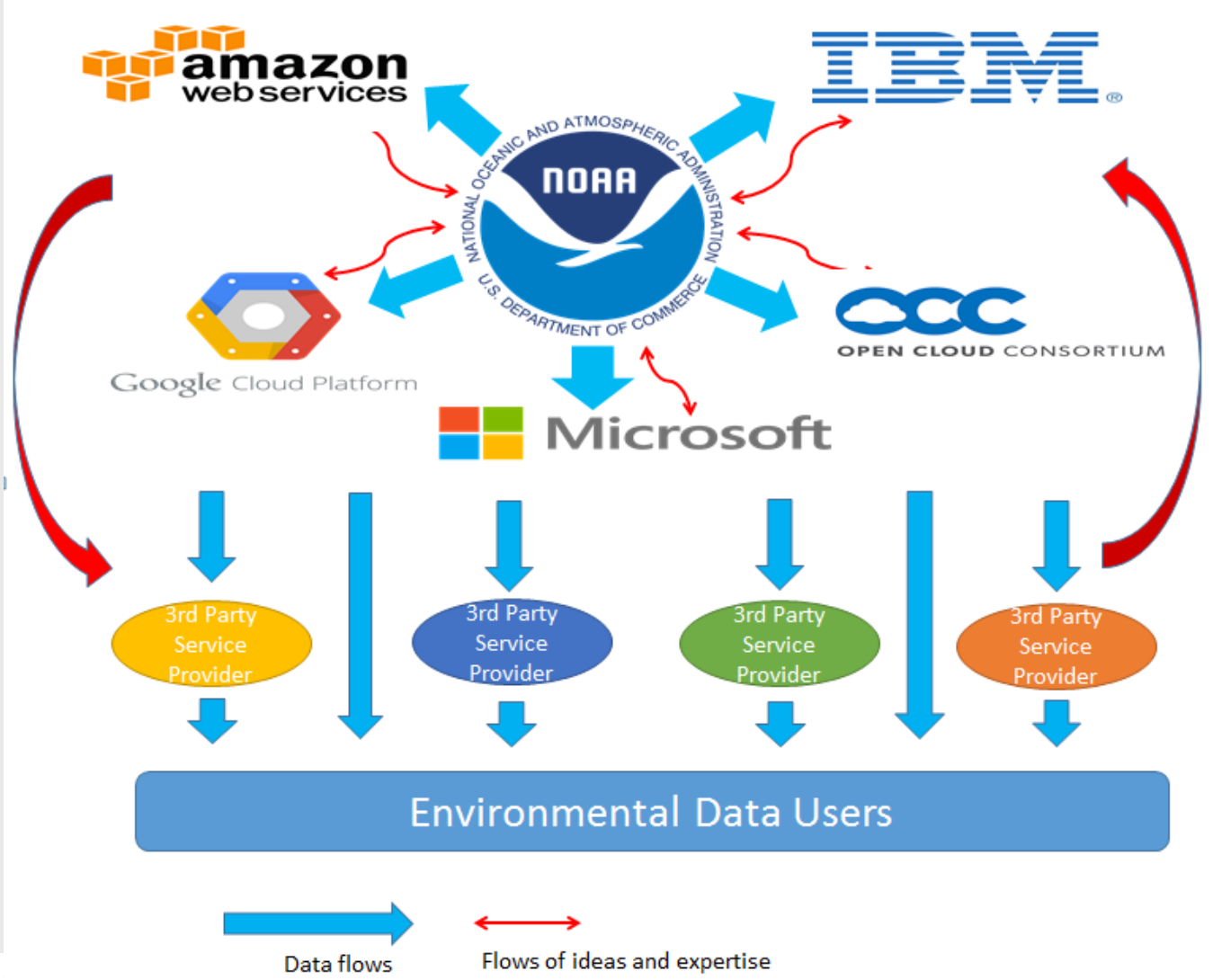
- IaaS companies to serve as project anchors
- These CRADA "Collaborators" are nuclei for data alliances and markets
- Members of industry, research, and academia may join these alliances
- NOAA receives data requests from the Collaborators…

Google Cloud Platform

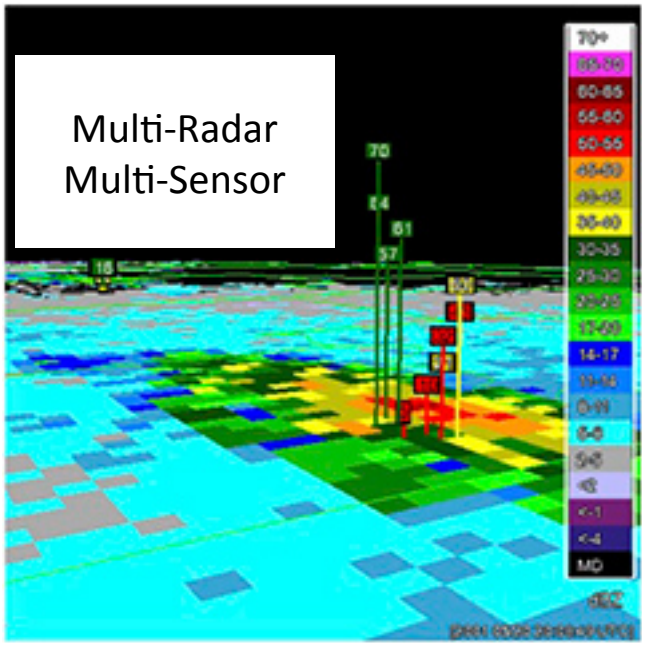


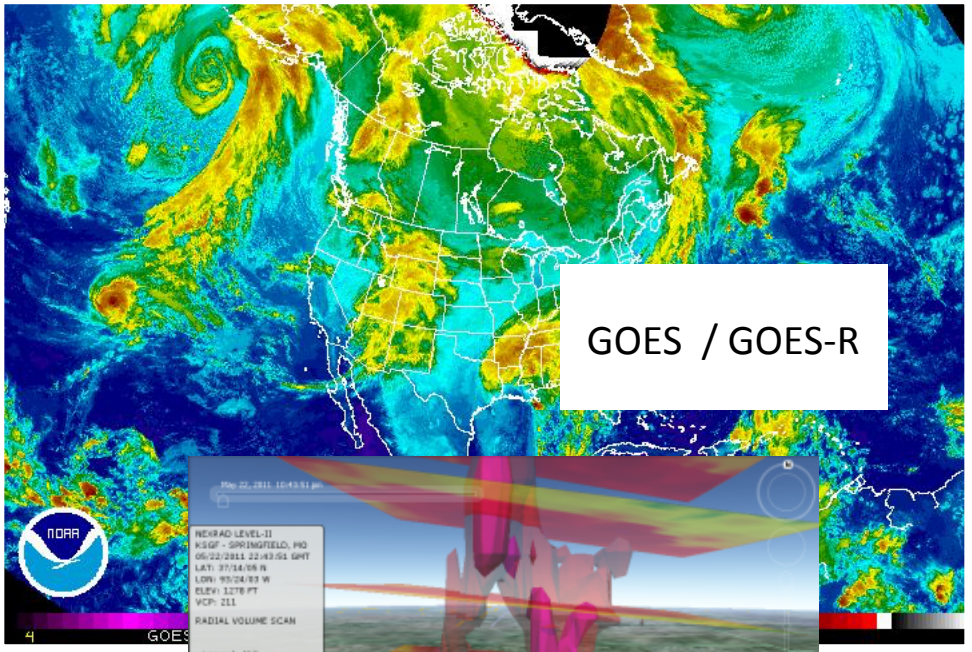
Microsoft

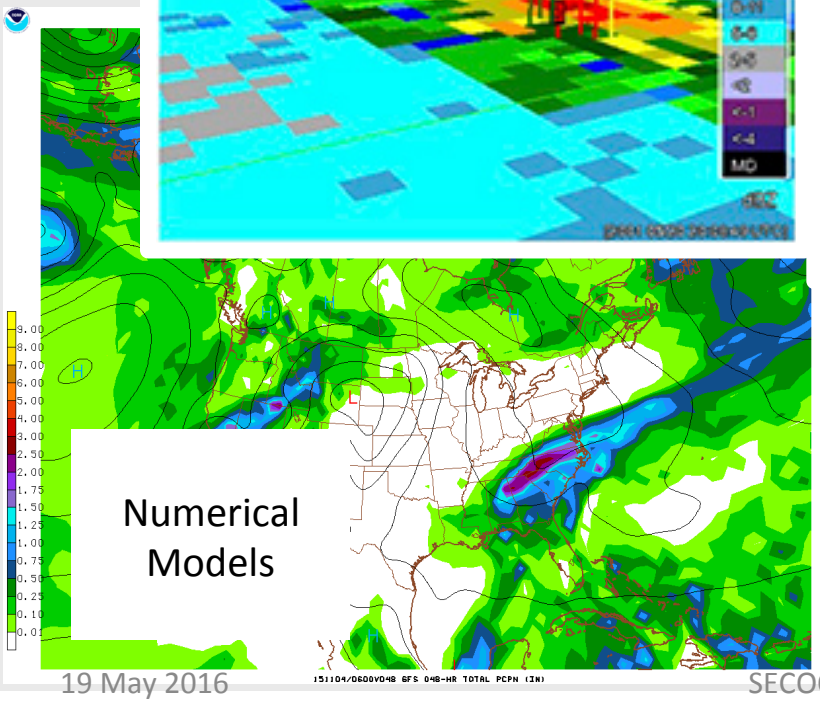
Conceptual Relationship for BDP

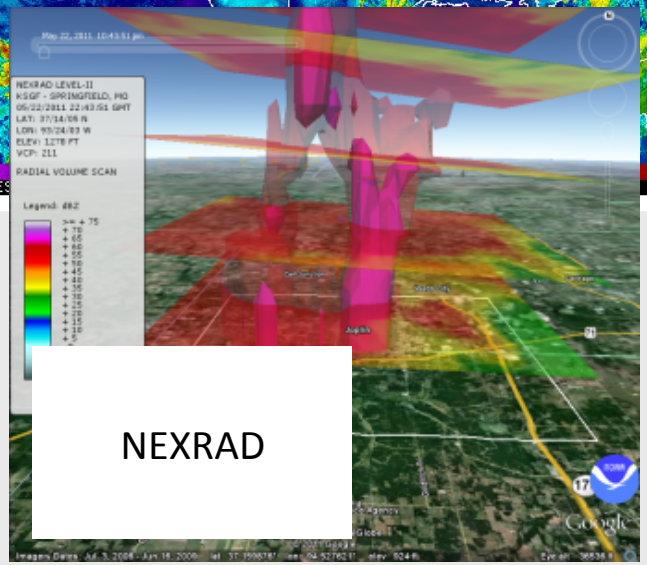


Requested Initial Data Included
…not much ocean data









Why was NEXRAD Weather Radar selected first?

- Archived NEXRAD data are optimized for *preservation,* not access

- All NEXRAD data are publicly available, but difficult to use
  - unwieldy size (270 TB for compressed Level II alone)
  - specialized format (radial volume scans)
  - resides on NOAA's NCEI tape archive, relatively slow to access

- Highly popular dataset for use in industry
  - Many industry users of realtime and archived NEXRAD data
  - Multiple derivative uses possible - hail, rain, snow, tornados, etc.

- The utilization of entire NEXRAD archive never before realized

- NOAA had recently reprocessed 2001-2012 NEXRAD Level II
  - half of the dataset still resided on disk at the Cooperative Institute for Climate and Satellites in NC (CICS-NC)
  - easy to jump-start initial delivery

NEXRAD #2 in US National Observation Impact

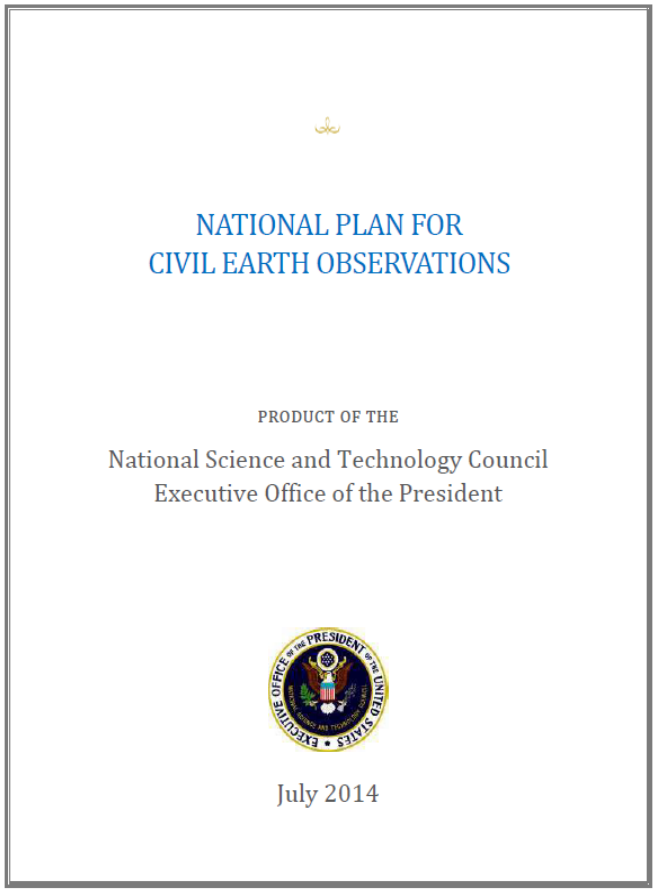
Annex I: 2012 EOA Results

This annex provides results for the 145 high-impact observation systems identified from the 362 observation systems assessed by the 13 SBA teams of approximately 300 Federal subject-matter experts. These 145 observation systems are listed in two tiers in the tables below. Impact is indicated with respect to each of the 13 societal themes (12 SBAs and reference measurements), as described in Section 2.2.

Table 1: Tier 1 High-Impact Observation Systems (Ranked Order)

Observation System (Ranked Order)	Agency	Ag&Frst	BioDiv	Climate	Disasters	Ecosys	Energy	HumanHlth	Ocn&Cstl	Space Wx	Trans	WaterRes	Wx	RefMeas
1. Global Positioning System (GPS) satellites	DOD/USAF													
2. Next Generation Weather Radar (NEXRAD)	DOC/NOAA								*					
3. Landsat satellite	DOI/USGS, NASA										*			
4. Geostationary Operational Environmental Satellite System (GOES-NOP)	DOC/NOAA			*		*								
5. National Agriculture Imagery Program (NAIP)	USDA/FSA													
6. Airborne LIDAR	DOC/NOAA, DOD/USACE, DOI/USGS, NSF													
7. Forest Inventory and Analysis (FIA)	USDA/USFS							*						

Impact: * Contributes; Moderate; High; Very High; Highest



National Plan for Civil Earth Observations (2014)

AWS Access to NEXRAD Level II Data

Oct 27, 2015 rollout

- **https://s3.amazonaws.com/noaa-nexrad-level2**

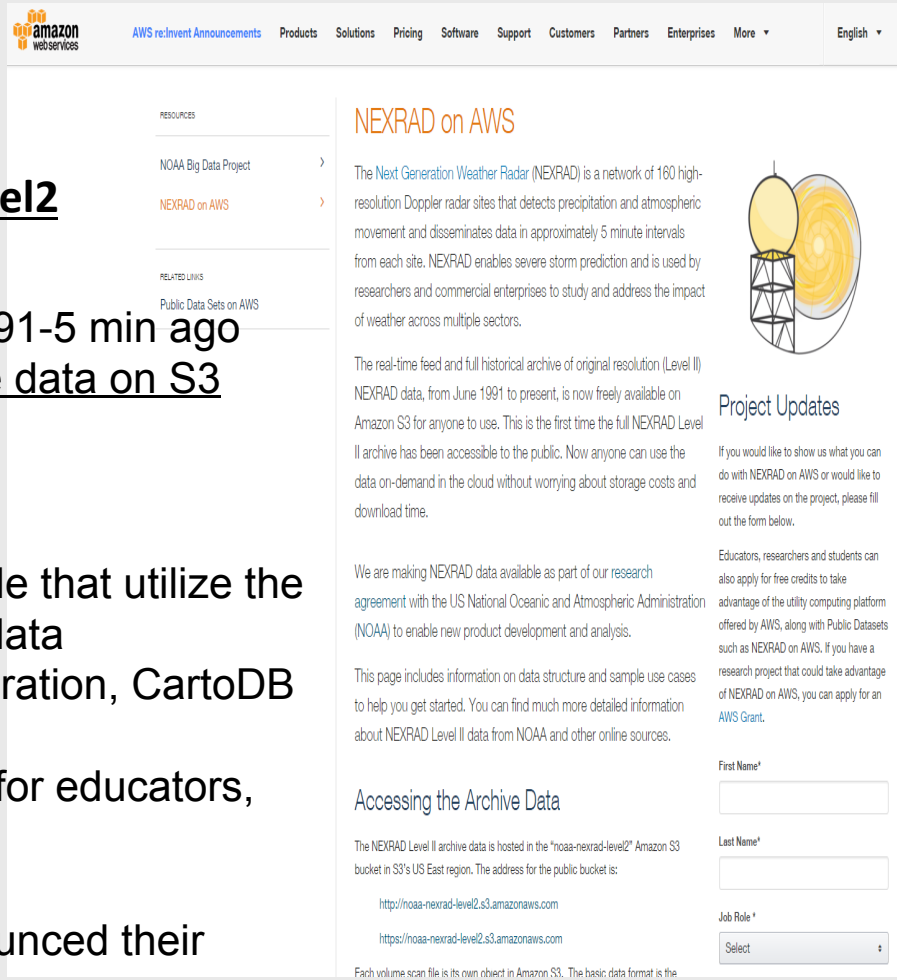
- AWS now serving all NEXRAD Level II from 1991-5 min ago
- Single point of access for archived and realtime data on S3

- Free data download available to users

- User-contributed tools and services are available that utilize the AWS platform, e.g. THREDDS installed by Unidata
- Tutorials available from Unidata, Climate Corporation, CartoDB

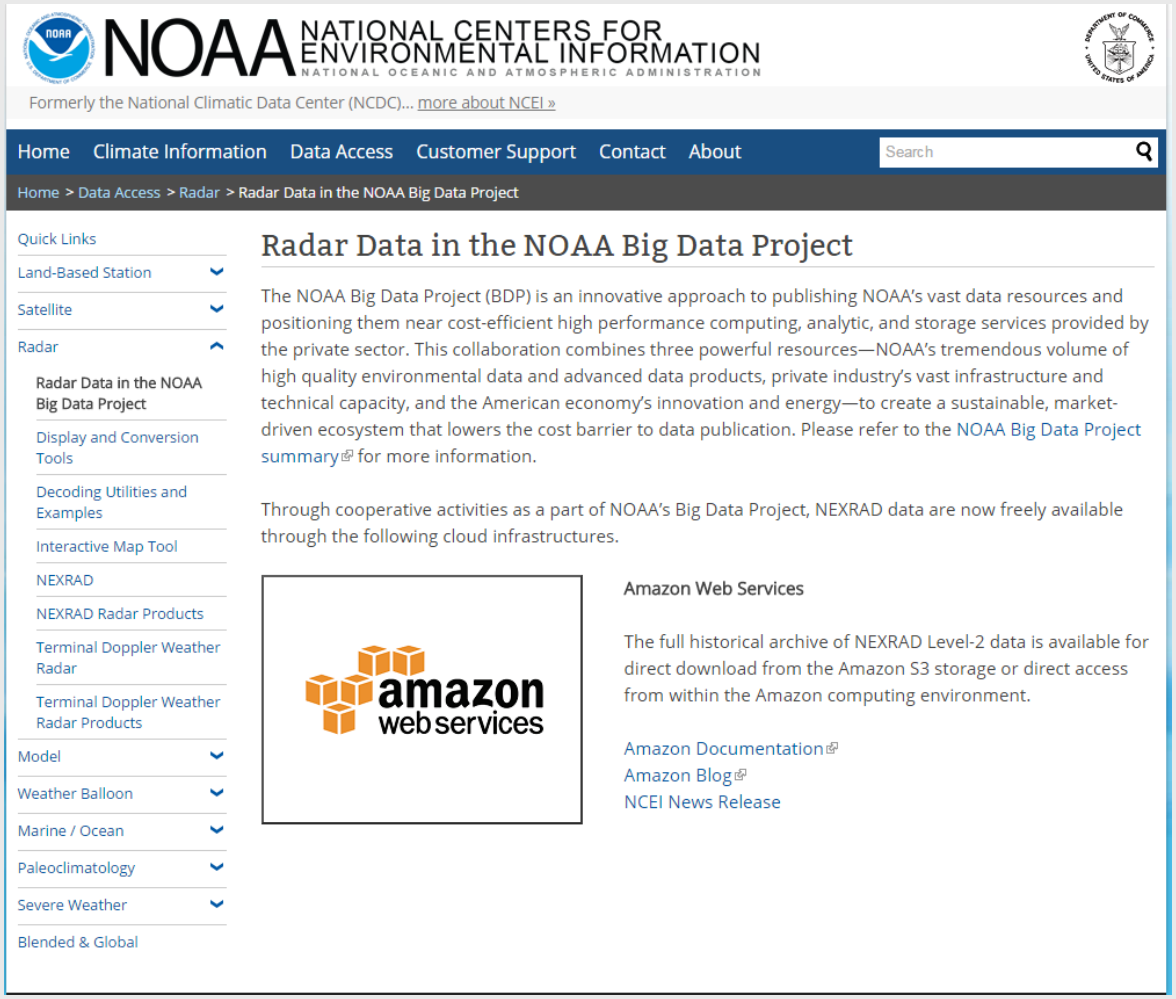
- Free AWS credits and grants may be available for educators, researchers, and students

- Google, Microsoft, and OCC have not yet announced their access plans for the NEXRAD Level II.

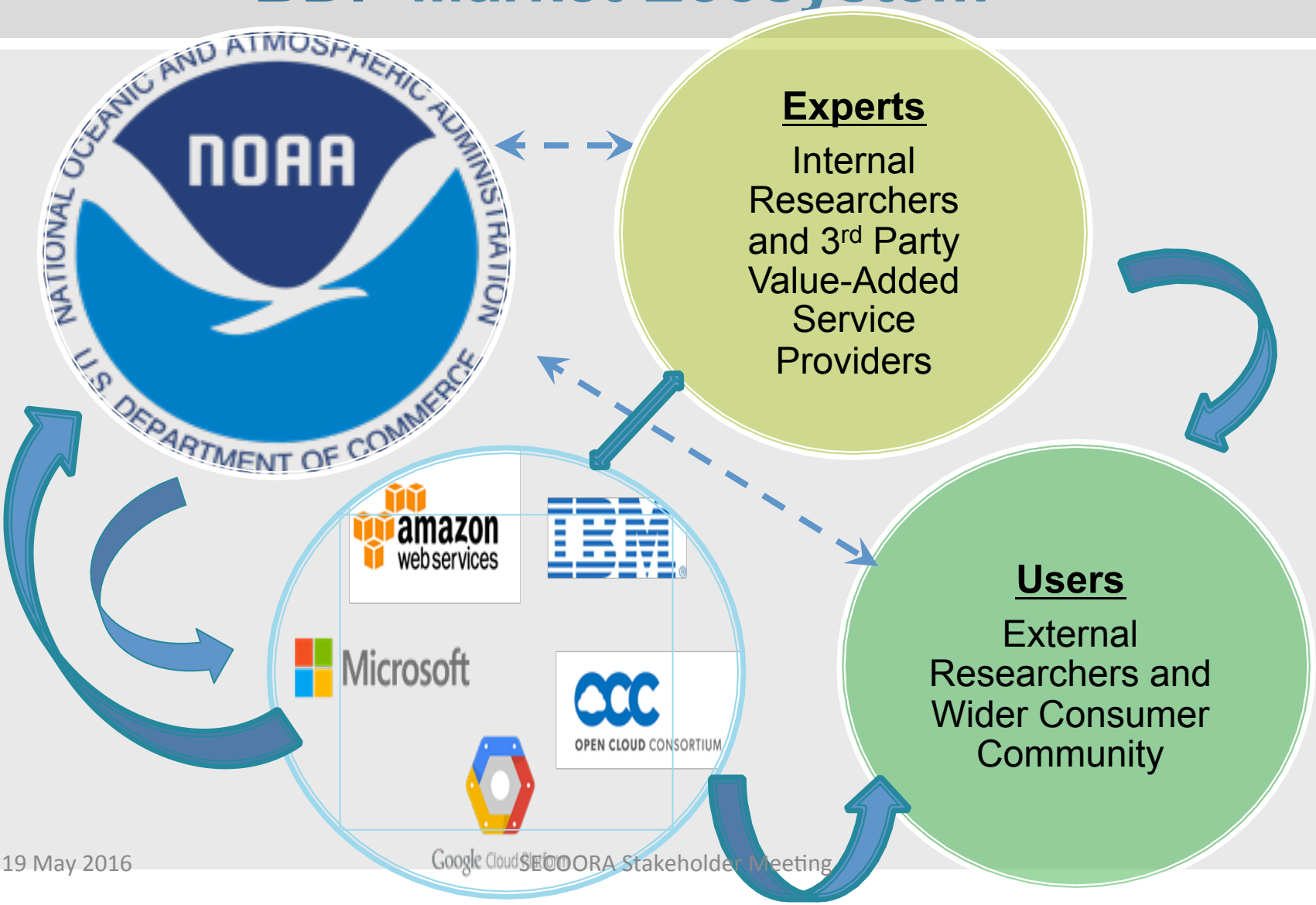


Steering Users to New Services

- New services are being integrated from within NOAA's NEXRAD sites.
- Archive data ordering system
  - 30-60% decrease in anticipated data orders from NOAA/ NCEI (Feb-Apr 2016, S. Ansari)
- Integration into NOAA's free "Weather and Climate Toolkit" application



The Elements of a Successful BDP Market Ecosystem



So What?

- NOAA's Big Data Partnership is showing that data access and research may be facilitated, at reduced cost, through leveraging effective partnerships with industry.

- Seamless access available to NOAA's holdings across time - users find both historical and realtime NEXRAD Level II data in the same place, in the same way.
  - Historical data provides context for new realtime observations – important for decision-making

- New business and research opportunities are being created

- New applications can be developed FASTER using less bandwidth
  - data are co-located with the processing capacity
  - Quicken the pace of app development and time-to-market
  - **Faster pace of innovation and scientific discovery**
    - NOAA's NEXRAD recent reprocessing (of 11 years' data) took *years*
    - Same volume of processing today could now take ~~*weeks*~~ *days*

Other Lessons Learned in BDP, so far…(1/2)

- Expertise need to be effectively involved throughout the process.
  - "Push" and "Pull" at the same time works the best
  - Sometimes the Collaborators acquire the expertise internally…
- Almost all of NOAA's open data are available, digitally, today…if you know who to ask, or where to look.
- True Big Data applications now limited by their business cases
  - Marine Data have proven difficult for building new business cases
  - Single valuable datasets, versus truly diverse and large datasets
  - Without easy access to Big Data, how do you know what it will bring?

Other Lessons Learned in BDP, so far…(2/2)

- Easy access to archived data is valuable for dynamic provision of the historical/statistical context for new data or realtime data flows.
- Traditional data-centered economic value models are not necessarily easily adaptable to the Big Data methodology
- There may be more social problems than technical problems.
- How can NOAA effectively steward data held by others?

Other Opportunities Ahead

- The rise of the Platform
  - Platform = data + processors + storage + apps
  - Scientists used to go to the data. Then the data came to them. Now they all can coexist in a shared, virtual, computing environment.
  - Has Platform utilization become “simply” an acquisitions and governance problem? Or do social problems need to be addressed first?

- The US Digital Service and its related entities (e.g. the Commerce Data Service, the new Technology Transformation Service (GSA’s 18F))
  - Focusing on enabling data discovery and use
  - E.g. the Commerce Data Usability Project

Summary

- The full utilization of Earth Science data is a Big Data challenge since it requires the provision of adequate expertise, involves many datasets, and is challenging conventional value models.
- "Big Data" success requires not just access to the data, but the expertise (algorithms, workflows, interpretive skill) as well.
- Government and Industry partners may be able to facilitate research activities by leveraging the inherent value in Earth Science data.
  - Need help identifying value for industry for NOAA Marine and other non-weather data
- Through the BDP CRADA, NOAA has moved the NEXRAD Level II dataset first, based on market need and data opportunity. AWS experience Oct 2015 appears to be heading towards success so far…
  - New applications created, increased use of AWS infrastructure, reduced load on NOAA systems
- NOAA and the Collaborators have not really gotten deep into Big Data territory yet, so far mainly enabling wider use of unwieldy, hard to access and high volume datasets.

Acknowledgements

Many thanks to:

- NOAA: Amy Gaskins, Alan Steremberg, Maia Hansen, Steve Ansari, Steve Del Greco, Jeff de la Beaujardiere, Brian Nelson, Tony LaVoi, Jay Morris, Carlos Rivero, Ken Casey, Ken Knapp
- NC State University / CICS-NC: Otis Brown, Jonathon Brannock, Lou Vazquez, Scott Stevens

NOAA's Big Data Collaborators and their partners involved in the NEXRAD project

- Amazon: Arial Gold (now @DOT), Jed Sundwall, Jeff Layton
- Climate Corporation: Adam Pasch
- Unidata: Jeff Weber
- Microsoft: Sam Khoury, Sid Krishna
- Google: Eli Bixby, Tino Tereshko, Amy Unruh, Tanya Shastri, Ossama Alami, Valliappa "Lak" Lakshmanan (formerly @ClimateCorp)
- Open Commons Consortium: Maria Patterson, Walt Wells

Thank You - Discussion

Questions?

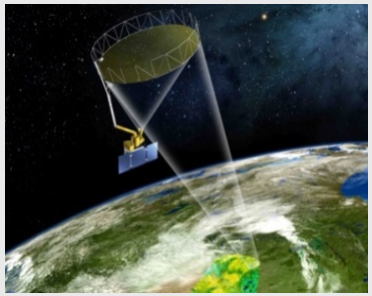
Big Data Partnership "Rules"

- **Level Playing Field, for all interested companies via Collaborators**
  - **All NOAA Data, available equally, with no privileged access**
  - **NOAA-sourced data are "free and open" - only cost recovery allowed**
- **5 individual CRADAs with NOAA, 3 year terms + 2 one-year options**
- **CRADAs can be ended early with appropriate notice**
- **No funds may move from NOAA to Collaborators**

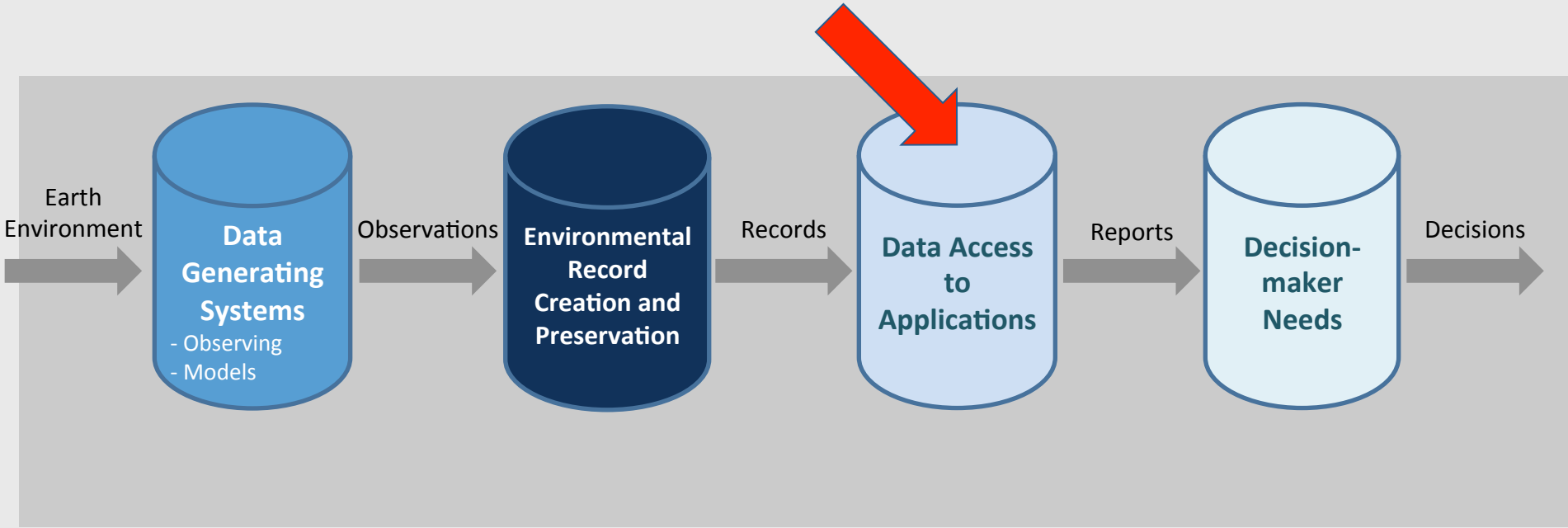
Big Data Partnership Methodology

- Collaborators & their partners identify datasets of interest
- Demonstrate business use case(s), with no restrictions on open distribution of data
- Develop a strategy for data sharing/delivery from NOAA to BDP Collaborator(s)
- Engage NOAA subject matter experts, BDP Collaborators, and their Associates/Partners for technical interchanges
- Collaborators and their Partners create applications
- NOAA continues all of its existing data services
  - No interruption of services to customers
  - BDP activities are an augmentation of existing services

Impact on the Information Lifecycle



- NOAA makes foundational investments in environmental information production and preservation
- **BDP promotes increased data access and encourages application development across private and public sectors.**



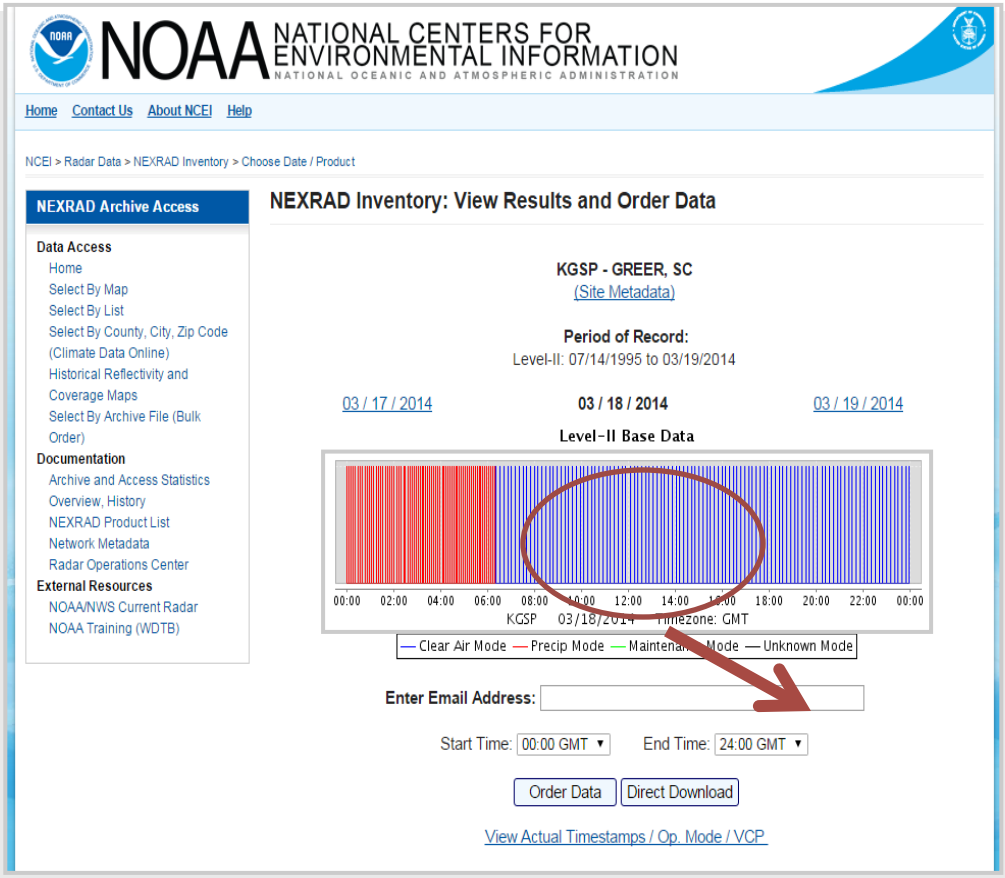
NEXRAD Data Transfer

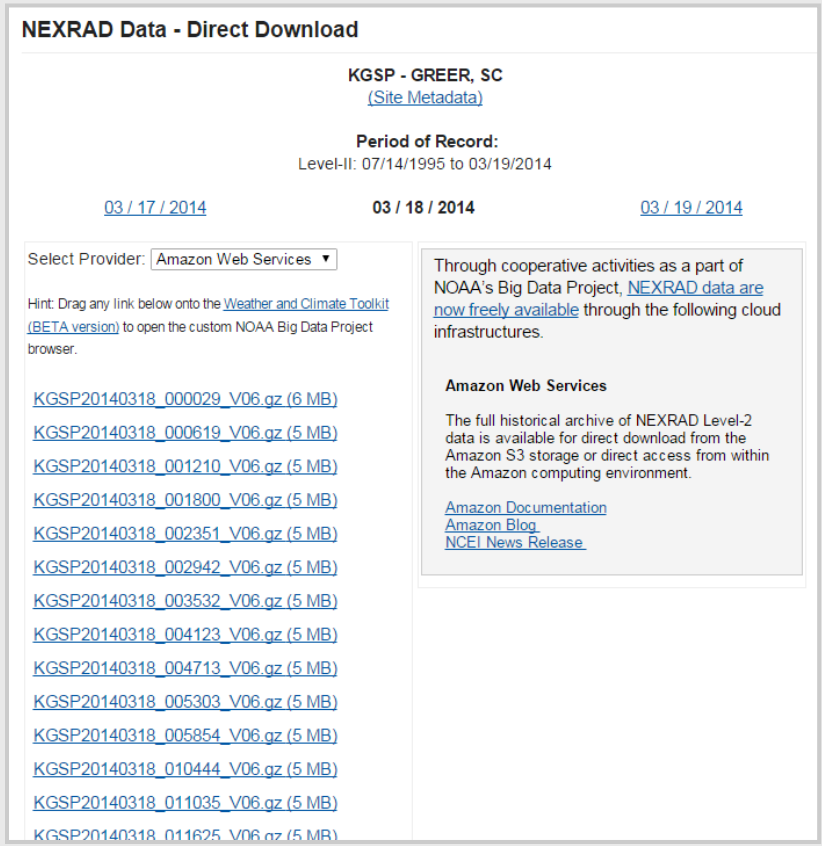
- Archived Level II data from NESDIS' National Centers for Environmental Information (NCEI)
  - Over 270 TB for compressed Level II volume scan file (>1 PB uncompressed)
  - Move from the NCEI tape archive to disks at CICS-NC
  - CICS-NC acting as middleman allowed minimal impact to NOAA's data operations
  - Updates from NCEI archive operations as new data are pushed to the archive

- Realtime Level II data from NOAA's National Weather Service
  - Established through Unidata, part of the Open Commons Consortium (OCC)

- Data moved to 4 of the 5 Big Data Partnership Collaborators
  - Amazon Web Services (entire archive plus realtime)
  - Microsoft (entire archive)
  - Google (entire archive)
  - Open Commons Consortium (2015 plus realtime)

Steering Users to New Services





Other BDP Collaborators' sites will be added as they roll out their data services

Role of NOAA in the BDP Market Ecosystem

- NOAA's data are widely available, with "free and open" access
- NOAA's role will be to provide objective scientific expertise, ensure long-term preservation and sound data management (scientific data stewardship)

- Data + Expertise + Need = Opportunity

https://data-alliance.noaa.gov