Improving Access to Open Data through NOAA's Big Data Project

NIST Big Data Workshop
Gaithersburg, MD
Friday June 2, 2017

Edward J. Kearns, Ph.D.

NOAA Chief Data Officer
Office of the Chief Information Officer

NOAA's Mission(s)

Life and Property

Aviation

Maritime

Space Operations

Forests

Emergency Management

Commerce

Ports

Energy

Hydropower

Reservoir Control

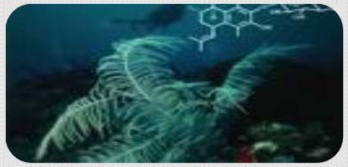
Infrastructure

Construction

Agriculture

Recreation

Ecosystems


Health

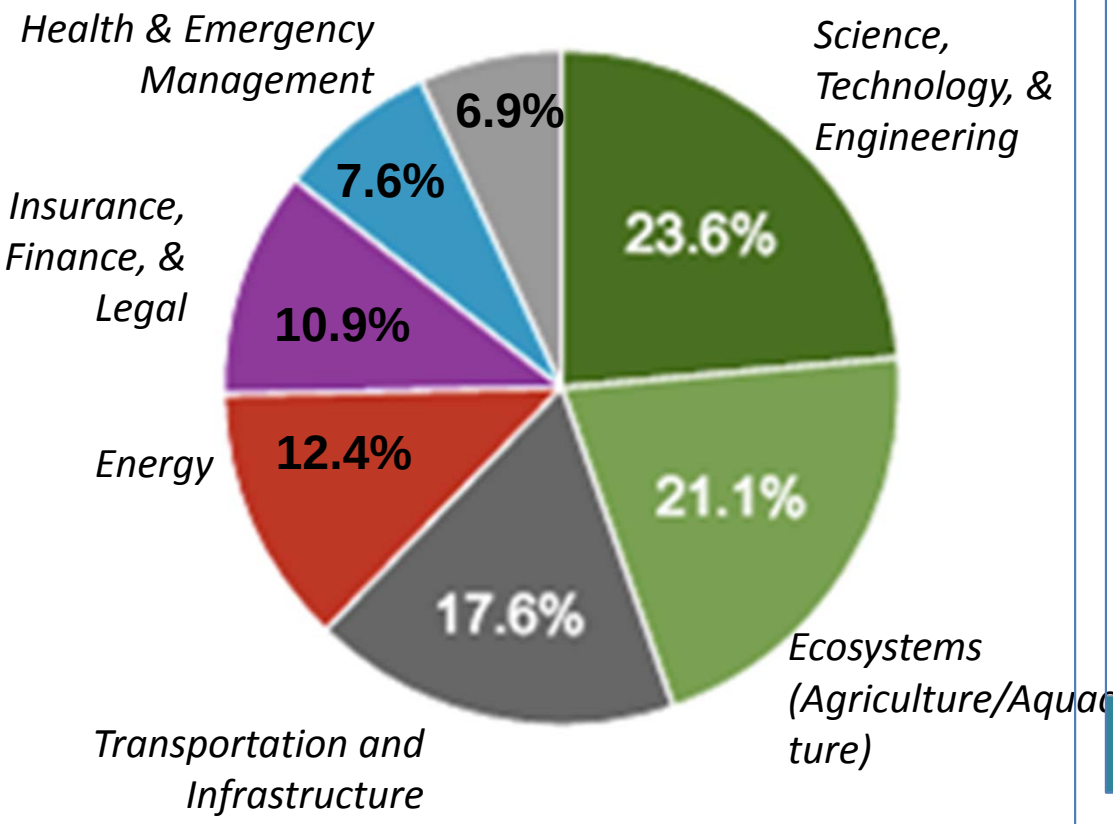
Environment

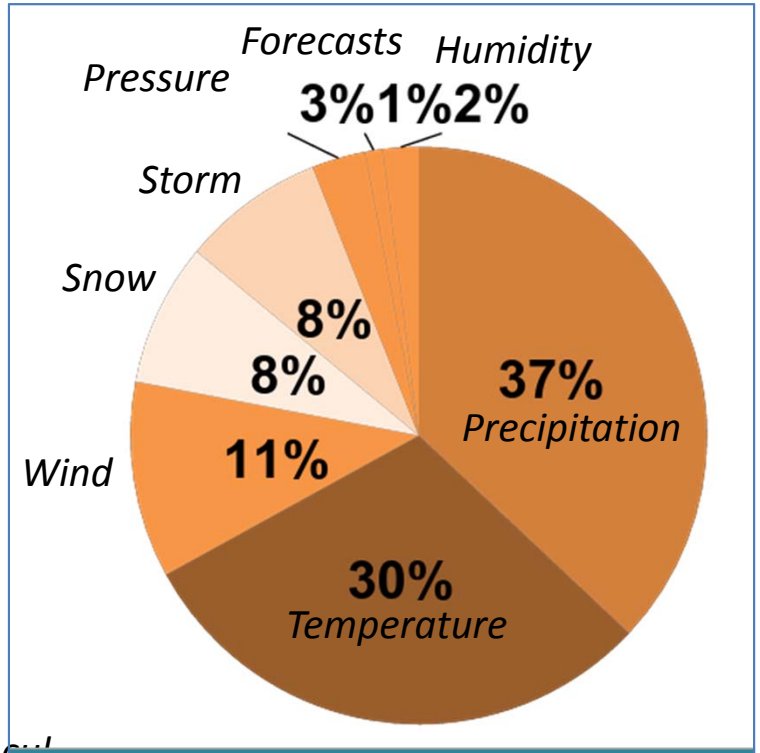
NOAA and Open Data

- NOAA is responsible for services related to Weather, Oceans, Coasts, Marine Ecosystems, Fisheries, Space Weather, ….
- Almost all of NOAA's data are "open data" – free to all, no limitations on use, in the public domain
  - International partners
  - Industry
- How do we share these data and enable their use in Big Data Applications, both for NOAA and others?

NOAA NCEI *Information Users*

By Targeted Sector





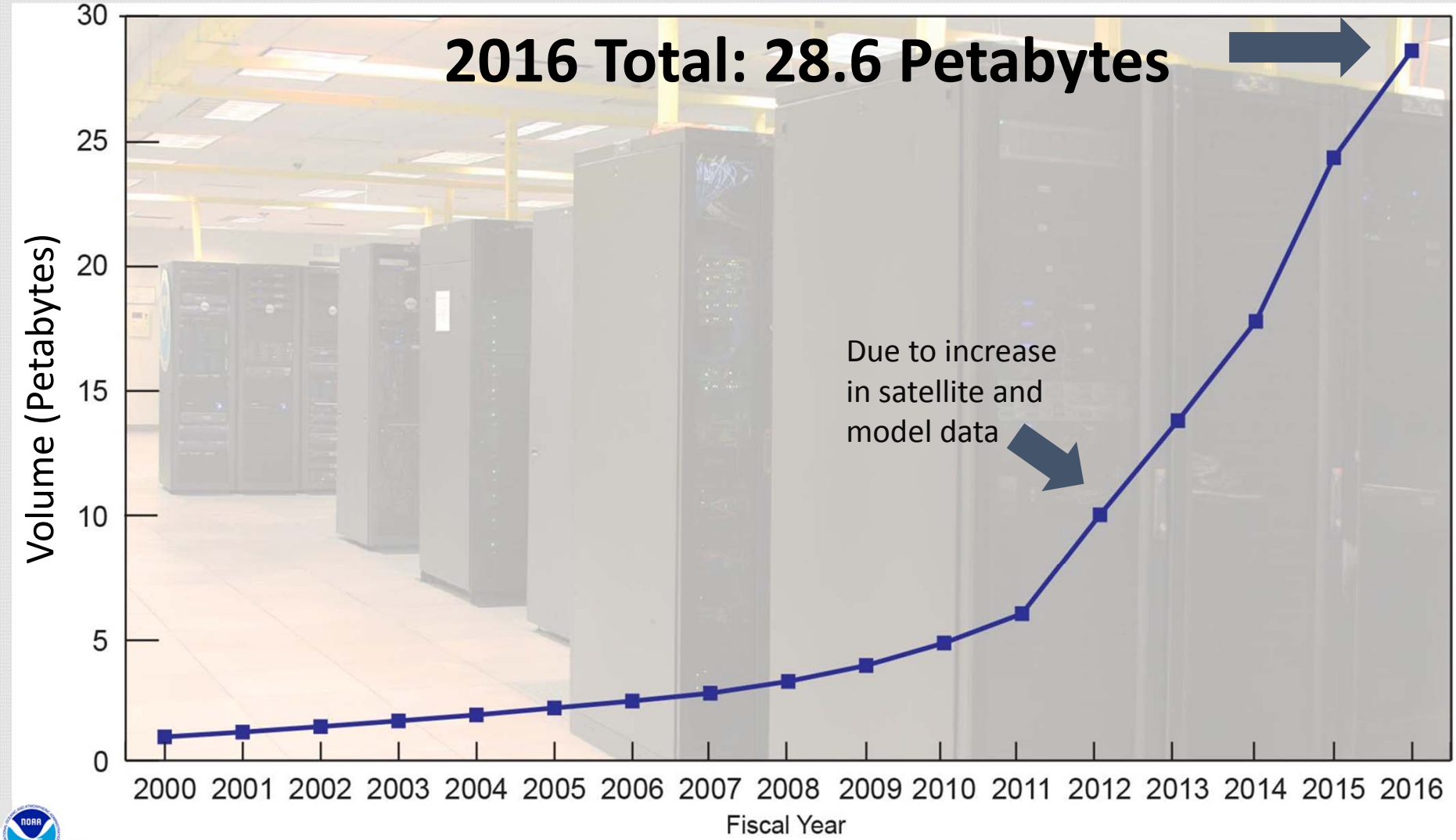
By Resolved Theme

NOAA NCEI's General User Community

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	Quantitative	Digital downloads machine to machine	Low	High

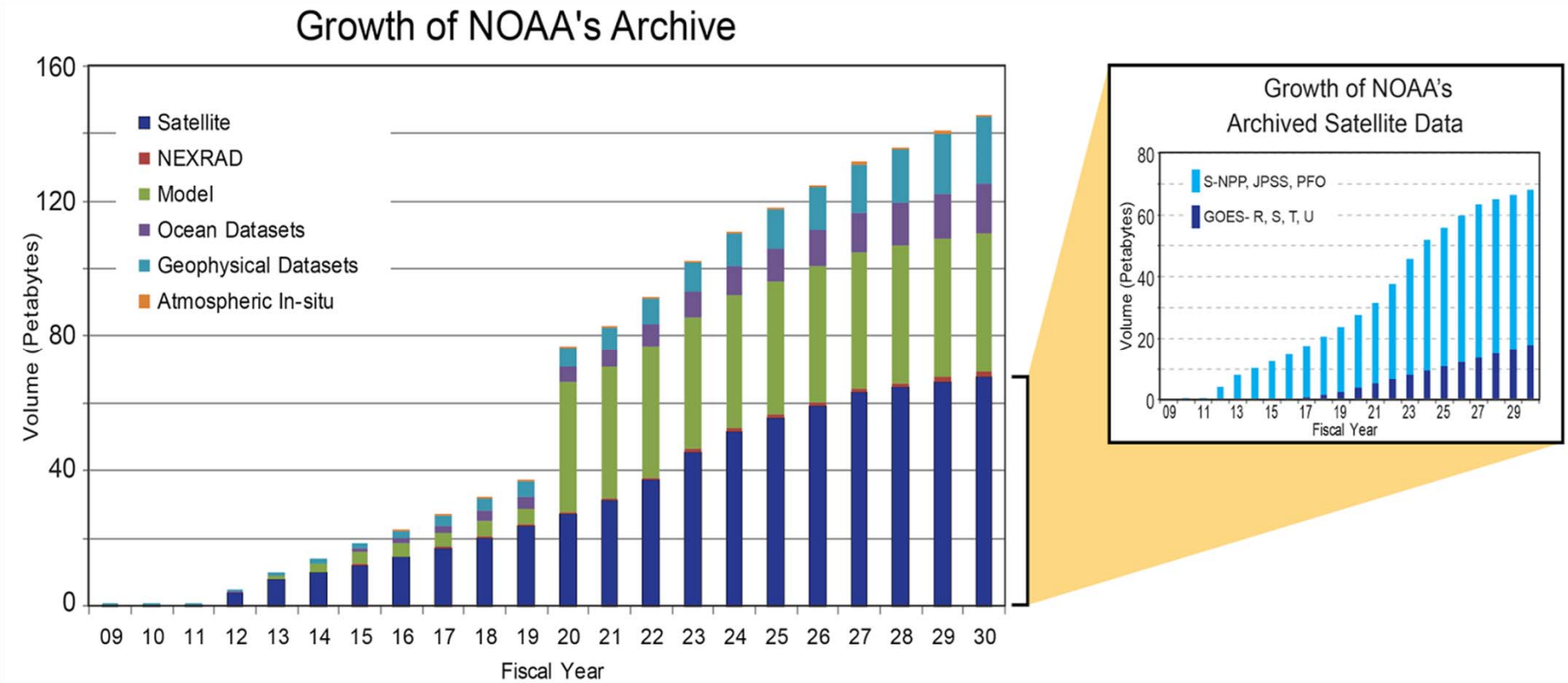
NOAA/NCEI's Environmental Data Archive

Increasing volumes from station, model, radar, & satellite data



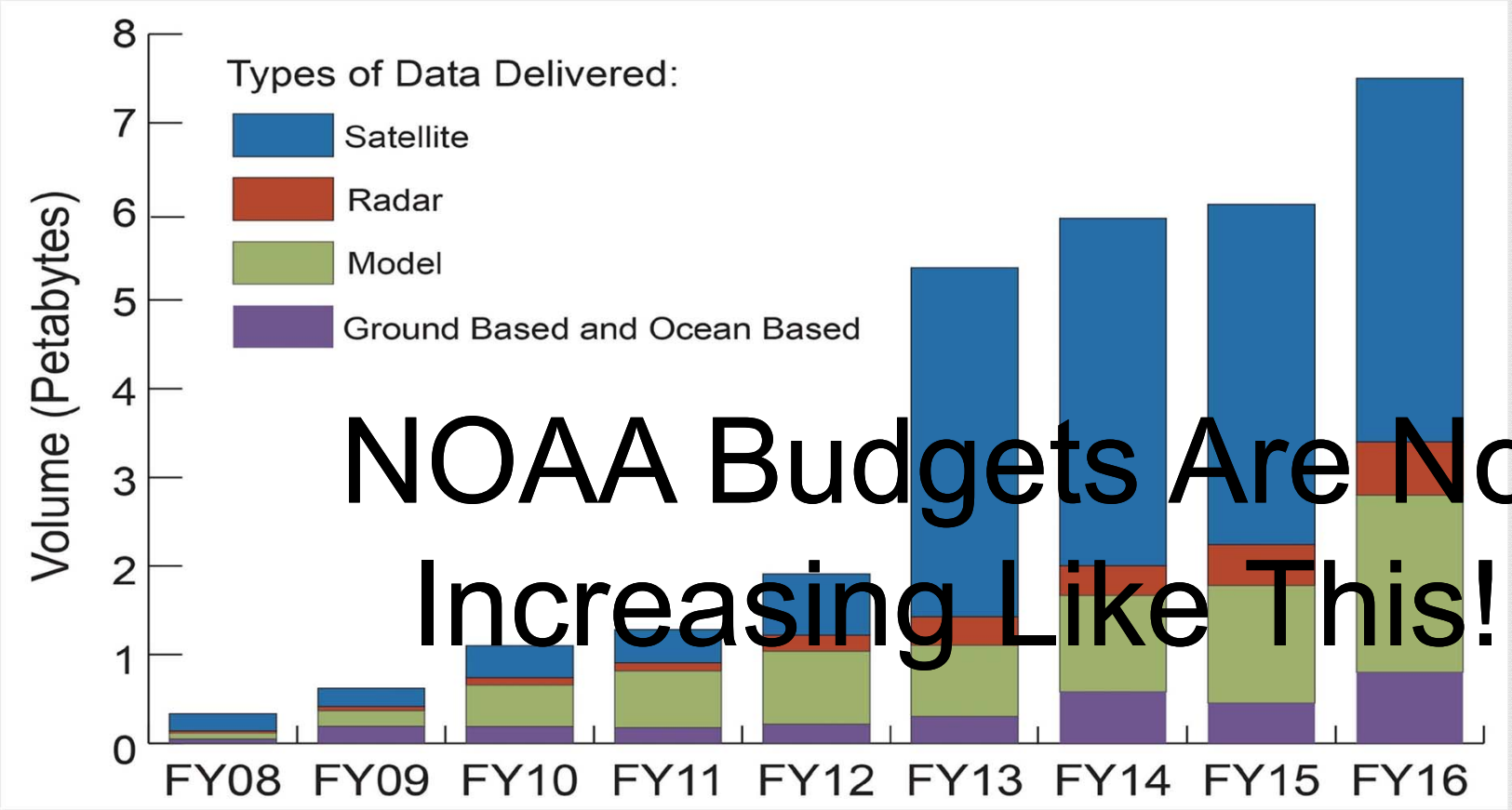
Data Volumes Rapidly Increasing

NCEI Projections for NOAA Data Archives



Demand for Access Increasing

NOAA NCEI Access by Volume



NOAA Budgets Are Not Increasing Like This!

8

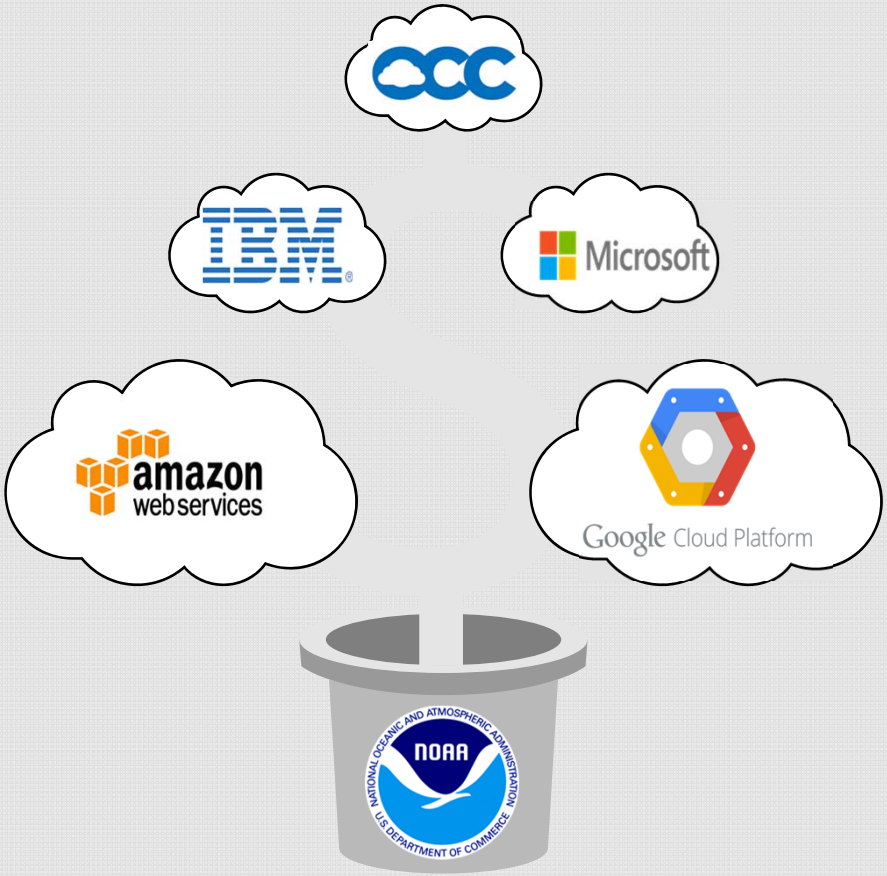
Why is NOAA so interested in Partnerships for Open Data?

- NOAA struggles to keep up with increasing public demand
  - Budgets for capacity and security aren't keeping pace with rapidly increasing data access costs.
- NOAA wants to learn about collaborative solutions
  - Seek opportunities for industry to support data access
  - Promote use and democratize data access
  - Utilize new technologies
  - Enable new economic opportunities for partners.

The Big Data Project (BDP)

Keys

- Bring users to the data
  - Not “just” about access
- CRADAs - research activity (2015)
- NOAA’s open data - freely available
- NOAA’s subject matter expertise
- Industry’s infrastructure expertise
- Level playing field
  - No privileged access
- Democratization of NOAA data
  - New opportunities for business



Leverage the value of NOAA’s data to increase their utilization

BDP Data Access Strategy

Leveraging Industrial Partners

Augment

Add Capabilities

Enable Big Data Applications

Amplify

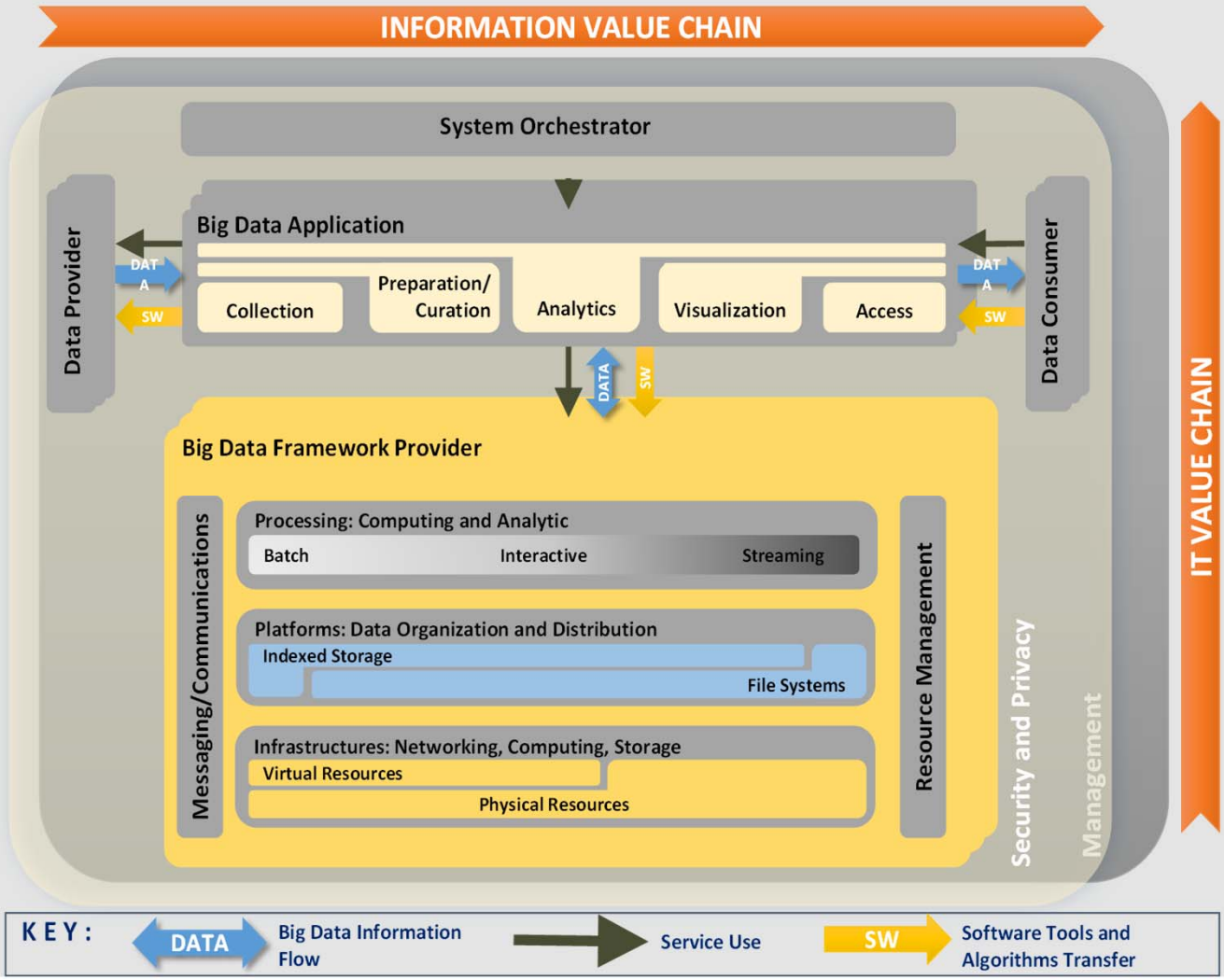
Add Capacity

Extend reach

BDP Elements

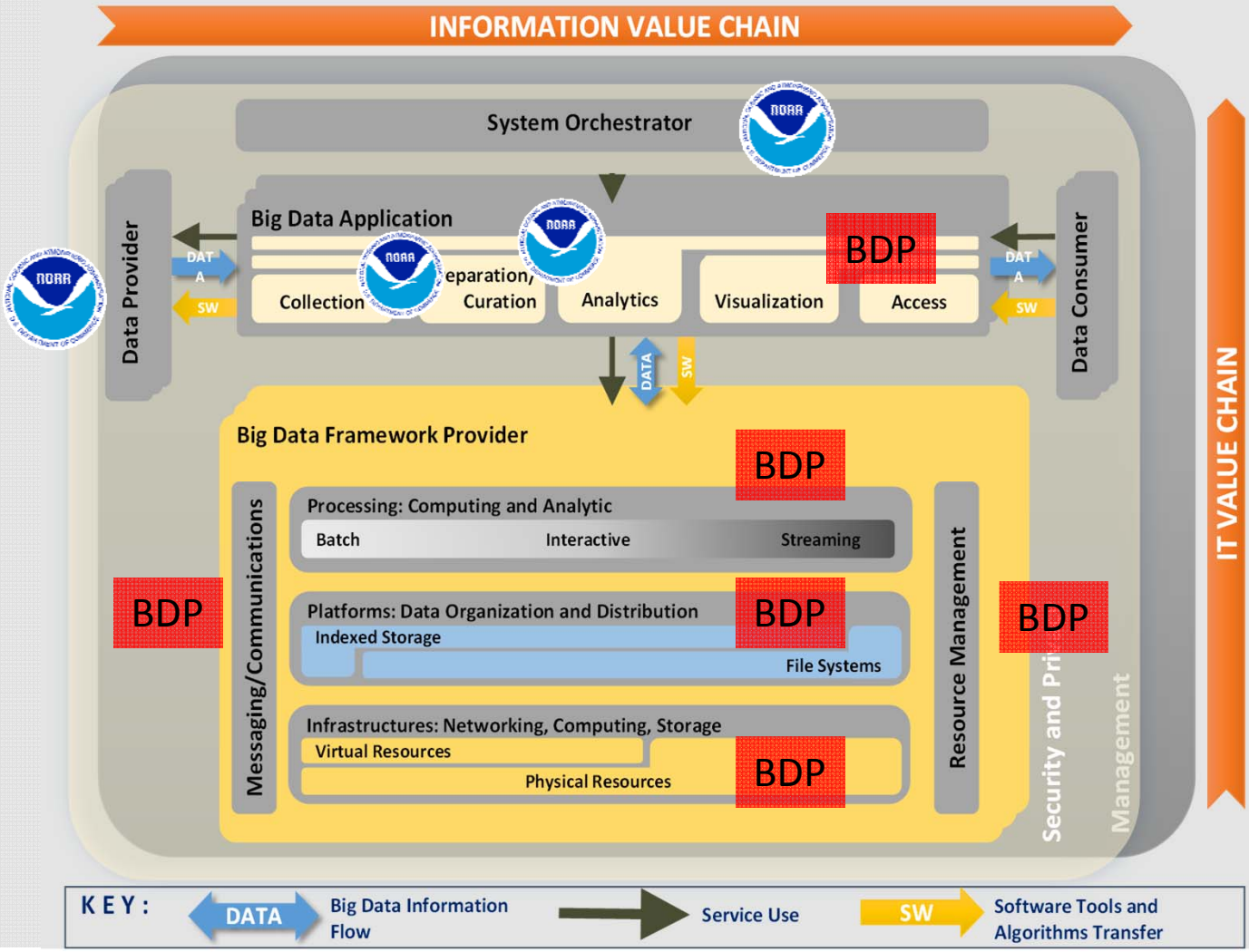
- **NOAA Basic Access - Limited capacity**
  - **Stewardship and service**
  - **Data Integrity and Authenticity**
- **Industrial Partners - Augment and Amplify**
  - **Capabilities, Technologies, Capacities**
  - **Scales to meet demand**
  - **Market pays for what it needs**

NIST Big Data Reference Architecture (NBDRA)



13

NIST Big Data Reference Architecture (NBDRA)

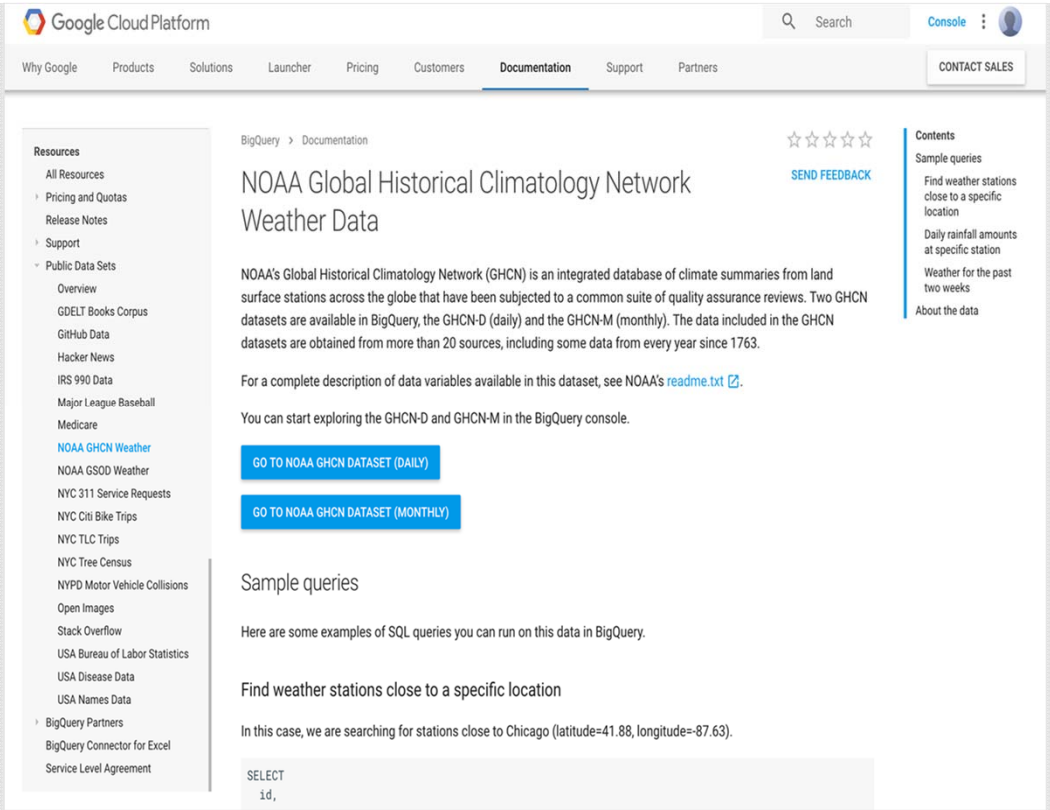


14

BDP Collaborator Services

- **AWS**
  - https://aws.amazon.com/noaa-big-data/nexrad/
- **Google Cloud Platform**
  - https://cloud.google.com/bigquery/public-data/
- **IBM**
  - https://noaa-crada.mybluemix.net/node/32
- **Microsoft**
- **Open Commons Consortium**
  - https://www.opensciencedatacloud.org/publicdata/?commons_type=Environmental

Example BDP Success: Google



- https://www.youtube.com/watch?v=OVc9GZzQu-4&index=7&list=PLIivdWyY5sqI6Jd0SbqviEgoA853EvDsq

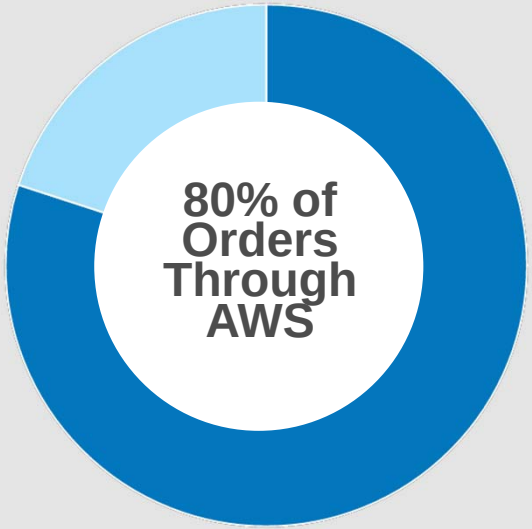
- **1.2 PB climate and weather data accessed through Big Query, from Jan-Apr 2017**
  - **Without "trying"**
  - **Joins, joins, joins**
  - **30-100x of NOAA deliveries in that time**

- **Images in Earth Engine**
  - **GOES-16 (June 2017)**
  - **National Water Model**
  - **Climate models**
  - **Climate data records**

Example BDP Success: AWS

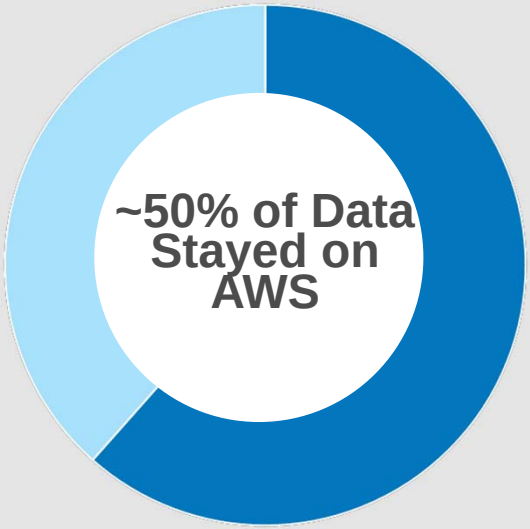
NEXRAD Weather Radar Data on AWS

NOAA Wins:
Reduced loads



AWS NCEI

AWS Wins:
Revenue generated



AWS Downloaded

End User Wins:
Easier access & production

AWS Job Time ~days
Through NCEI ~Years

17

BDP Challenges

- How do we manage public data access costs?
  - Egress versus “marginal cost of dissemination”
  - Importance of 3rd parties in understanding the market values
  - Do all NOAA’s data have a commercially-viable market?

- How to transfer knowledge of and curate many complex datasets?
  - How to ensure data integrity and authenticity?
  - How to best facilitate Big Data applications?
  - Real-time, e.g. satellites, weather observations, coastal data
  - Retrospective, e.g. reanalyses and observations, fisheries
  - Forecasts, e.g. CFSv2 and NMME climate models

- What comes next, after the CRADA expires?
  - Spin off new agreements or partnerships as April 2018 nears?
  - Have we learned enough yet? If not, extend CRADAs?

18

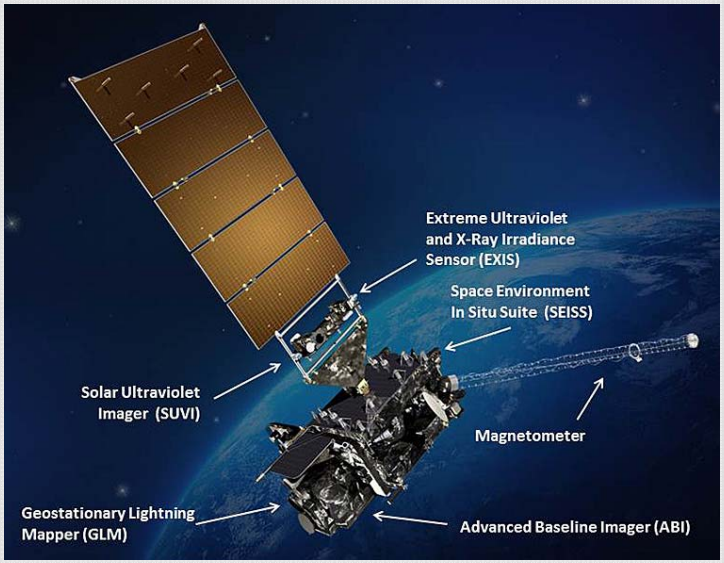
Acknowledgements

Many thanks to:
BDP Team: Andy Bailey, Shane Glass, Jeff de la Beaujardiere, Tony LaVoi, Jay Morris
NOAA: Brian Eiler, Zach Goldstein, Dave Michaud, Glenn Tallia, Derek Hansen, Amy Gaskins*, Alan Steremberg*, Maia Hansen*, Steve Ansari, Steve Del Greco*, Brian Nelson, Carlos Rivero*, Ken Casey, Rich Baldwin
NC State University / CICS-NC: Otis Brown, Scott Wilkins, Jonathon Brannock, Lou Vazquez, Scott Stevens, Paula Hennon, Andrew Buddenberg, Angel Li

NOAA's Big Data Collaborators and their partners (not an all inclusive list)
Amazon: Jed Sundwall, Arial Gold (now @DOT), Jeff Layton
Microsoft: Sam Khoury, Sid Krishna, Shannon Murphy
Google: Will Curran, Matt Hancher, Mike Hamberg, Eli Bixby, Tino Tereshko, Amy Unruh, Tanya Shastri, Ossama Alami, Valliappa "Lak" Lakshmanan (formerly @TCC)
Open Commons Consortium: Walt Wells, Maria Patterson, Zac Flamig
Unidata: Mohan Ramamurthy, Jeff Weber
IBM: James Stevenson, Stefani Jones, Mary Glackin, Peter Neilley, John Aviles
The Climate Corporation: Adam Pasch

19

Questions?



ed.kearns@noaa.gov

#NOAABigData

http://www.noaa.gov/big-data-project

20

More Datasets to Come…

- **In Progress**
  - Satellite data GOES-16
  - National Water Model
  - Climate models CFSv2/NMME
  - Fish catch/bycatch data
  - Marine Genomics
  - Surface Temp and Precip
  - Surface weather
  - Multi-Radar Multi-Sensor SST time series
  - Forecast database

- **Others in discussion**
  - Collections of realtime data
  - Collections of event data
  - Marine data around theme
  - Sandbox for experimentation
  - Weather model intermediate products