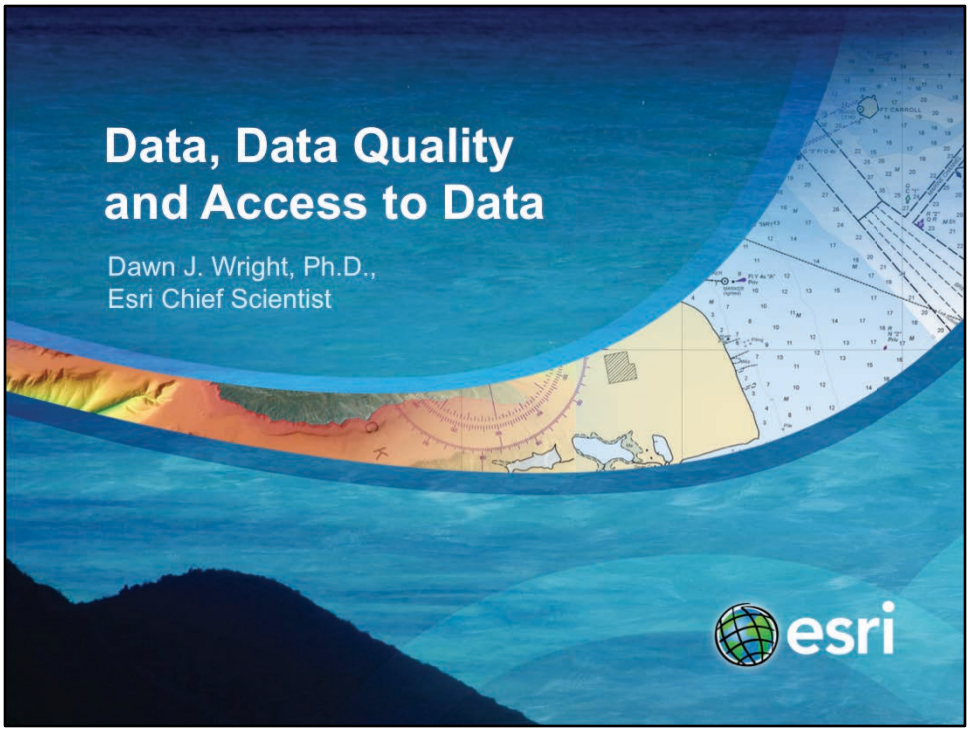


Invited “Marketplace of Ideas” presentation for the Ocean Exploration 2020 National Forum, July 19, 2013, Long Beach, CA
http://oceanexplorer.noaa.gov/oceanexploration2020/

We are in an era of regional- to global-scale observation and simulation of the Earth as exemplified by these large NSF and GEO programs. These big programs produce big data. “Big data [are] data that exceed the processing capacity of conventional database systems. The data [are] too big, move too fast, or don’t fit the strictures of your database architectures. To gain value from[these] data, you must choose an alternative way to process it.” Ed Dumbill in *Planning for Big Data*, O’Reilly Media

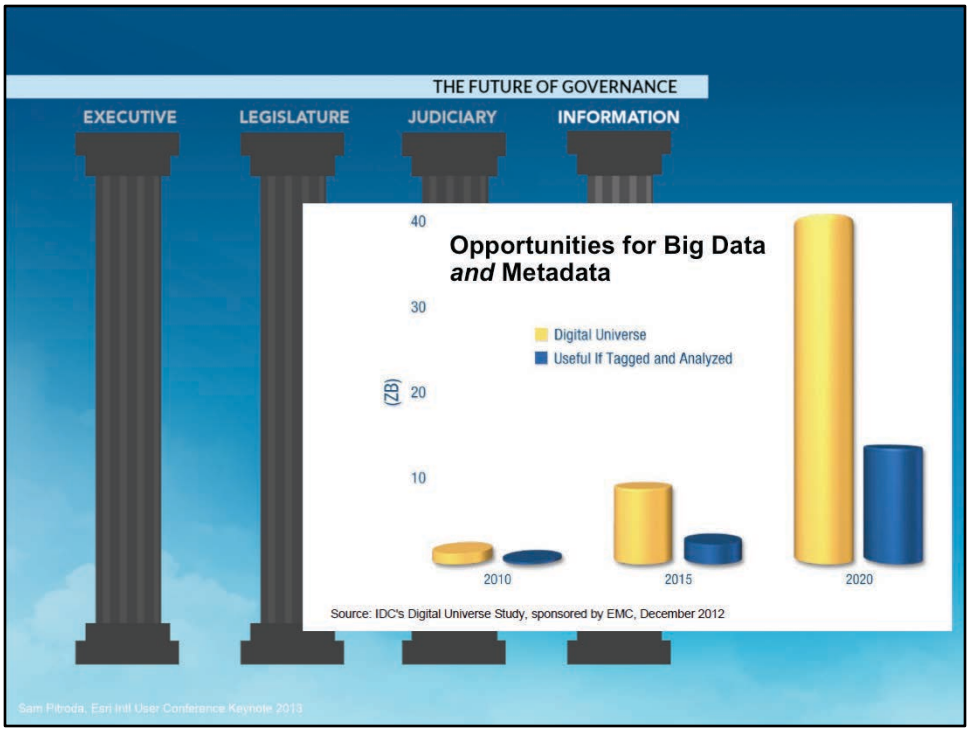
Critical Zone focuses on climate and land use, fluxes across watershed boundaries, Water, Carbon, Sediments, Nutrients
GEOSS – space-borne, airborne, in situ sensors, data mgmt architecture
EarthScope – seismometers, GPS on plate boundaries**, fault strain, crustal deformation**
OOI – cabled observatory – Endurance Array off Oregon will have fleet of gliders—T, Sal, Dissolved O2, Chlorophyll, backscatter-- surface buoys, seafloor instruments

Big data are leading to a new science paradigm, the new science of "big data" (the inundation of data from satellites, sensors, and other measuring systems and the issues associated with those large data sets)

Two recent special issues in Science and Nature speak to this most eloquently 2008 issue commemorated Google's 10th birthday and focused especially on semantics. (e.g., "The future of science depends in part on the cleverness of search engines being applied specifically to scientific data.")

And there is also the 2009 book **The Fourth Paradigm, which posits a new paradigm of scientific discovery beyond the existing 3 paradigms of EMPIRICISM, ANALYSIS, and SIMULATION to a 4th where insight is discovered through the manipulation and exploration of large data sets.**

The inspiration for the book's title comes from legendary Microsoft computer scientist Jim Gray, based on a lecture he gave at the National Academy of Sciences 3 weeks before he disappeared at sea.



MX: Investments in data rising. A zetabyte = 1 billion terabytes or 1 million times the content of the world's largest library

The market intelligence firm IDC (Intl Data Corporation) believes that by 2020, a third of the data in the digital universe (more than 13 zetabytes) will have Big Data value, but only if tagged and analyzed.
"Dark data"

The digital universe itself, of course, comprises data — all kinds of data. However, the vast majority of new data being generated is unstructured. This means that more often than not, we know little about the data, unless it is somehow characterized or tagged — a practice that results in metadata. Metadata is one of the fastest-growing subsegments of the digital universe (though metadata itself is a small part of the digital universe overall).

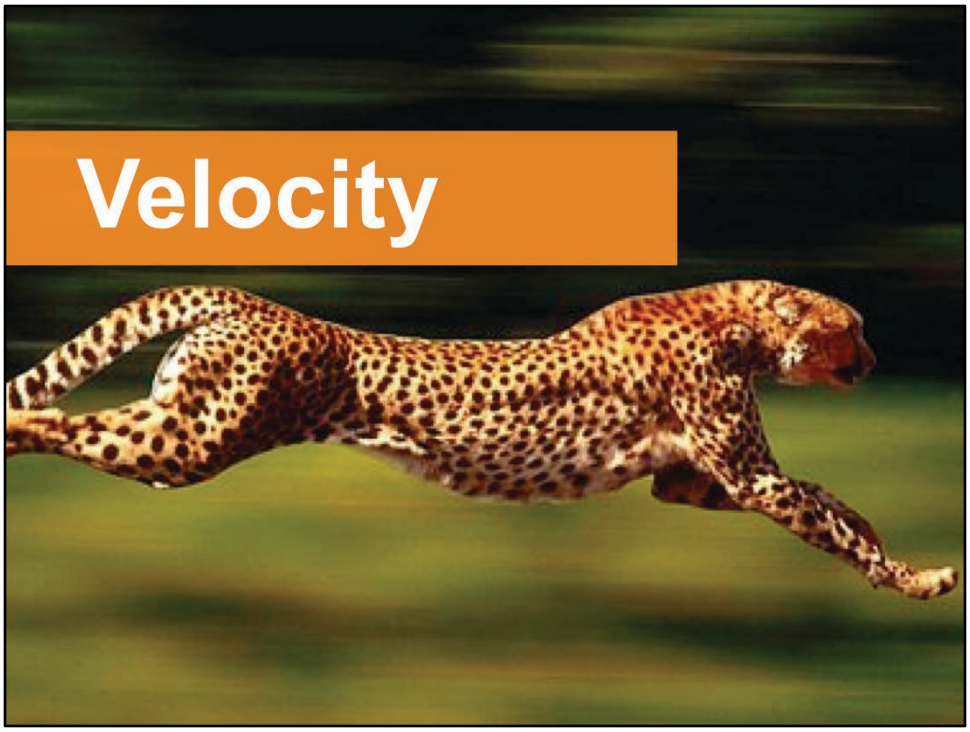
So I've used the term "big data," and we hear much about this term in the science literature, in the media, and at conferences.

But what do we really mean by "Big Data" particularly where ocean exploration is concerned?

There are several "tenets" that are being established in the commercial, government, and academic sectors that we will really need to pay attention to as we approach 2020.

Size and multidimensional nature of the data, e.g., with a modern dual head SONAR capturing bathy, backscatter and water column, that rate is 115GB/hr. That is approximately a 1.5 Billion to 1 ratio. The analogy is that if you take an ~8mm marble and multiply by the ratio, your marble is the size of the earth (approximately). **[HOLD UP MARBLE TO AUDIENCE]**
That's how MUCH data (shallow) we can capture with modern sonars.
Our sensor count of 2 million + will likely double every 5 years
So that oceans have ALWAYS been about big data, despite how much remains to be explored.

Next 3 images courtesy of Andrew Turner, geoIQ/Esri Washington, DCi R&D Center



So the Gb per HOUR example, leads us to the tenet of VELOCITY or the speed at which data are created and updated, often in near-real time. New challenges for **stream reasoning and rule systems**.
In the world of GIS and mapping we are talking about **30,000 features per SECOND**.
One important question may be which data do we keep?

The VARIETY or structural variability of the data may be the most DELICIOUS and compelling problem for the ocean exploration community.

These are data coming from **multiple sources and types** (photos, video, audio, text, scientific observations, scientific models)**, multiple perspectives** (governments, military, NGOs, etc.), which also have **various cultures of contributing data. This was reflected in the pre-forum survey results.**

A single oceanographic survey produces data in scores of different formats. Many physical oceanographic grids and models are **irregularly spaced!**

Not often discussed, but extremely important for the ocean exploration community is a fourth “V” – Veracity. Understanding data provenance, confidence in resulting information

One example = **CUBE** (Combined Uncertainty and **Bathymetric** Estimator) of Brian Calder et al. at UNH

How does one best model uncertainty in big data, especially from citizen science?

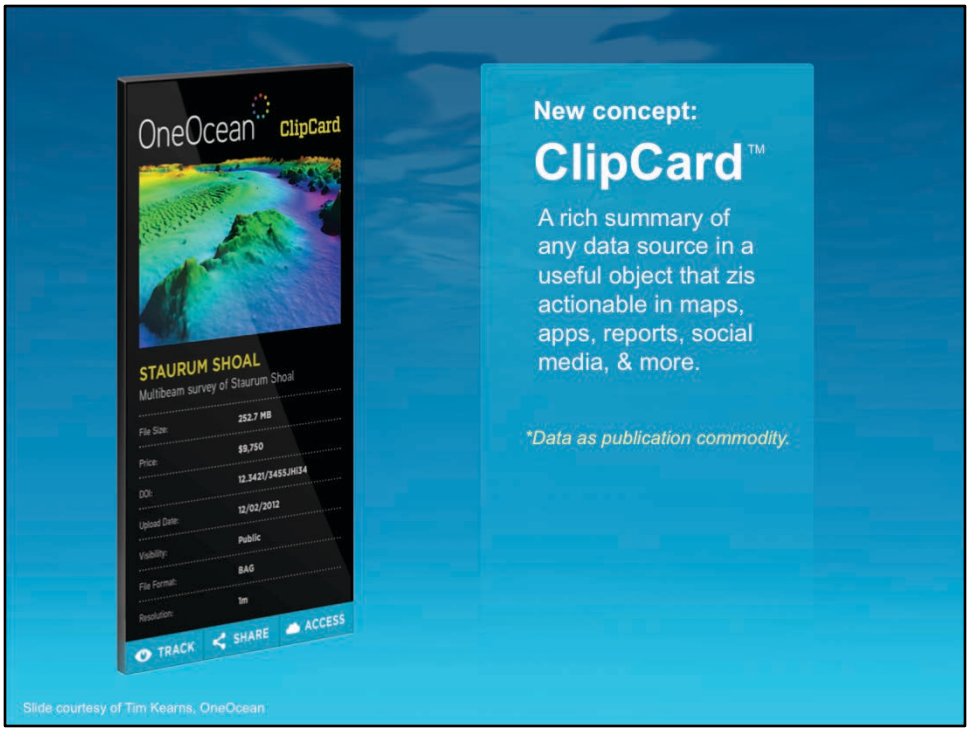
Issues of discovery, curation, provenance, organization, integrity.

All these tenets are part of the emerging discipline of Data Science, which works on solving these issues, on producing data products. It is the 4th paradigm of scientific discovery, the 4th paradigm of government

Marinexplore: “80% of decision-making processes in ocean science and business depend on data collection, management, processing, and distribution.”

There are great concerns with data quality, again as reflected in the pre-forum survey. Tell the audience not to forget this NRC Report!!!

From the report:
"Data Documentation, Curation, and Quality Assurance and Quality Control
In general, proper data storage includes supporting metadata, quality and fitness-for-use statements, and measurement error or uncertainty estimates. This is critically important for ocean research, as data are often collected in remote, hard-to-access areas. Data management facilities need to support efficient archival services and provide the capability to migrate data to different formats as computer technologies evolve (e.g., Miller et al., 2009). The Marine Metadata Interoperability (MMI) Project and Quality Assurance of Real Time Oceanographic Data (QuARTOD) are current examples of such efforts. Involving early career scientists in these and other activities will lead to better understanding of the fundamentals and implications of data reduction and quality management."



In terms of data quality, an innovative new concept for metadata, data, quality, and data publication is the ClipCard, by our new business partner OneOcean, a new startup in Seattle.

Links to source data but is only a fraction of its size
ClipCard. It's **lightweight**.

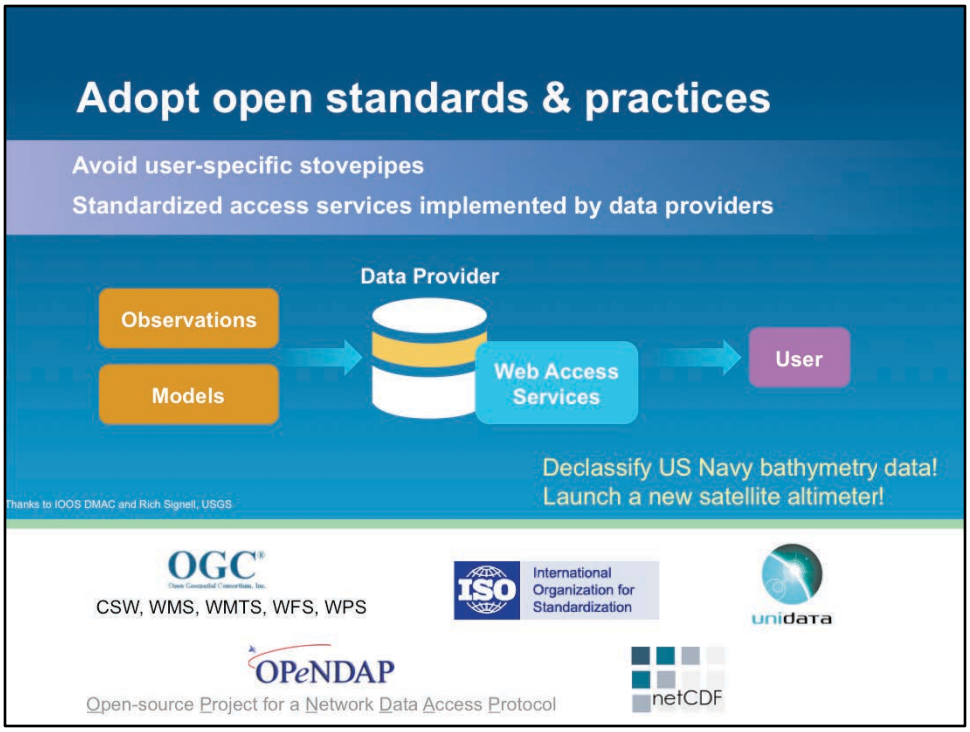
It's an **abstract representation** of the data – without being in touch with the data.
It's **independent from the source** – which means it can be shared.
It's **shareable and controlled** through a **user or organizational account.**
It's **multi-faceted –** meaning it has **multiple faces of content and information.**

Pre-forum survey favors partnerships – I'd like to share a specific idea on such. Ocean exploration has been in the realm of academics and federal agencies, but it will be critical to partner with industry in the DATA SCIENCE space. We are working on solving problems that you need to pay attention to. Be willing to partner with us. You will hear next from our colleague at **Google**, which academics are already well aware of partnering with.

But I just want to mention the new world of **ocean DATA industry**:
According to MX, the data acquisition market is currently $80 BILLION including ships, buoys, satellites, AUVS, ocean communication
The data management market will be $5 billion, including software and associated costs

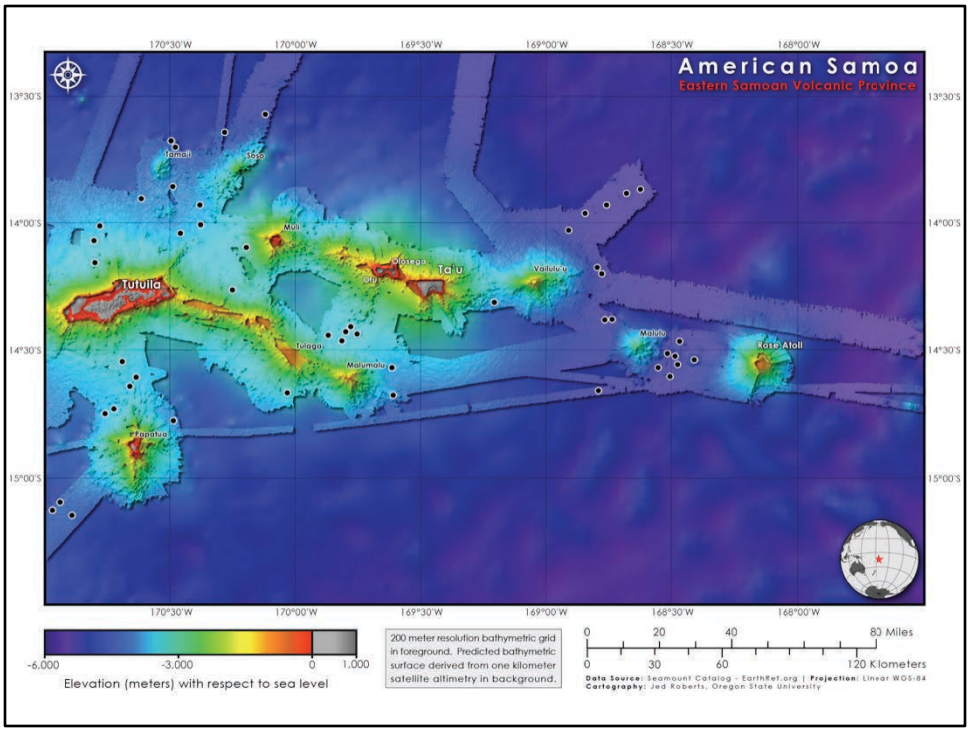
Several ocean IT companies including the industry partners here, are members of the World Ocean Council, a unique international BUSINESS alliance for corporate ocean responsibility, collaborating on issues of stewardship of the seas, including multisectoral voluntary ocean obs/ships of opportunity for science, CMSP, and of course improved data infrastructure, sharing for science. As partners, we are all working on different aspects of data science. MX is even building a marine operating system to streamline big data flows, as well as machine learning tools to improve data quality (e.g., automatically detecting sensor miscalibration, or automatically removing artifacts in ocean satellite data due to cloud cover).

RDA fostering public-private partnerships focusing on data use, data quality
From the NRC Critical Infrastructure report:
It would be beneficial for federal agencies to periodically examine and adopt data management practices that come from beyond the ocean sciences, as well as approaches to grow access to and use of community-wide facilities. Proven efforts from beyond the ocean sciences can be very informative and helpful.
Community-specific organizations that focus on data use and data quality will also be valuable to the ocean sciences (e.g., NSF EarthCube, AGU Earth and Space

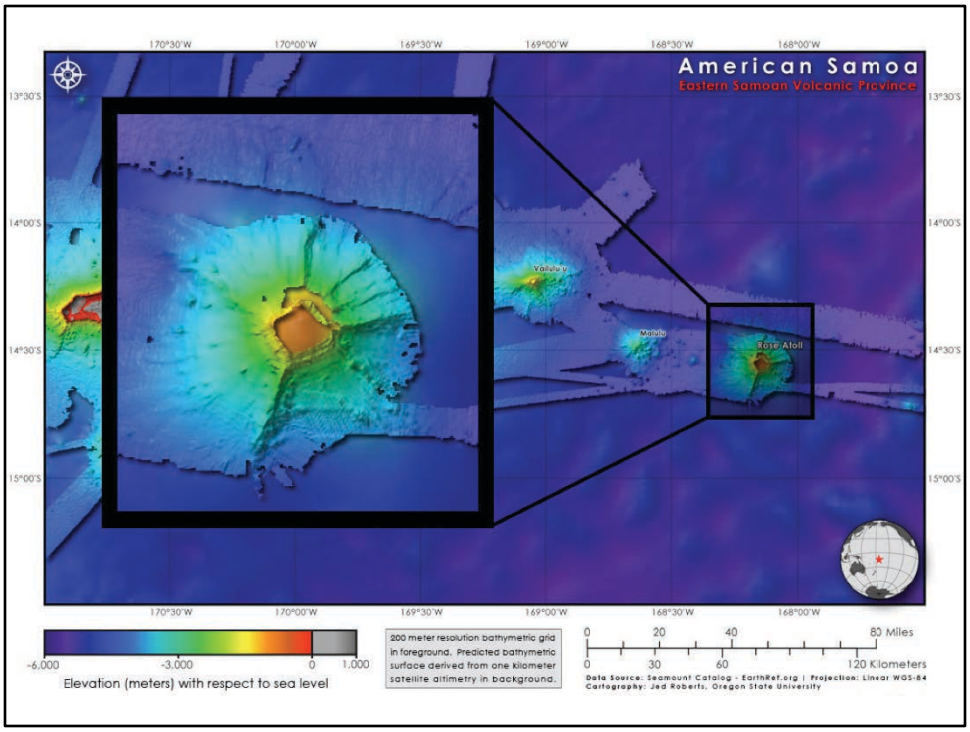


With all the data portals and organizations proliferating, the best approach is to stick to standards
To ensure that we are as nimble and effective as possible as we interoperate among **data formats, conventions, vocabularies, data files, data access portals, data catalogs, organizations.**
It ALSO wouldn't hurt to: Declassify US Navy bathy data! Launch a new satellite altimeter!

Open-source Project for a Network Data Access Protocol
WMTS = Web Map Tile Service with REST, SOAP, key-value-pairs encodings - http://en.wikipedia.org/wiki/Web_Map_Tile_Service



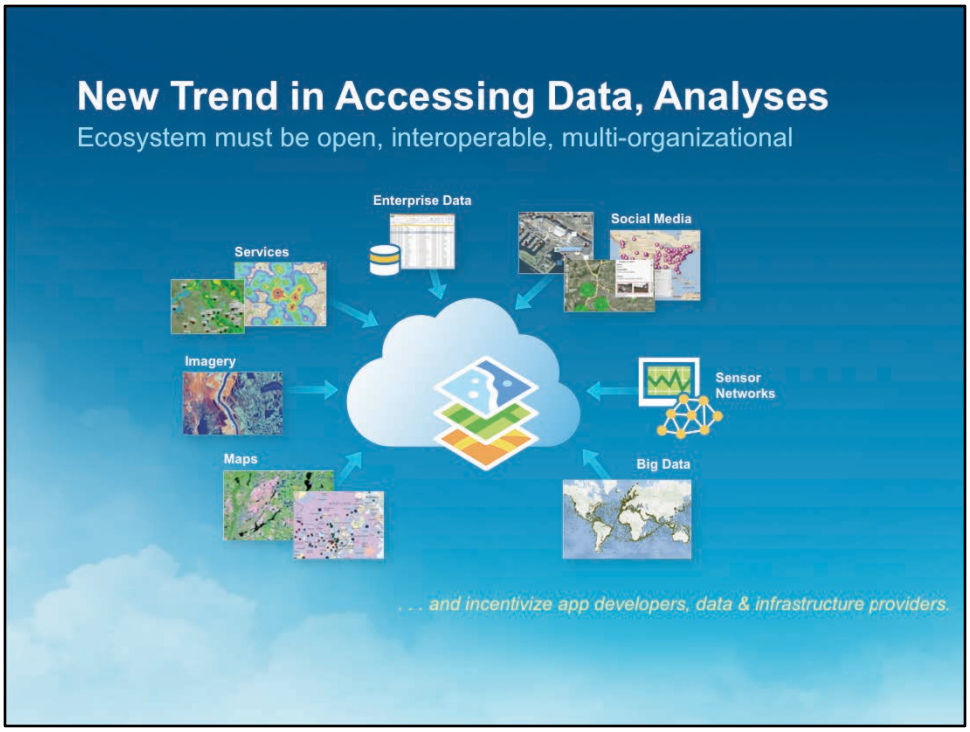
Eastern Samoa Bathy compilation showing 50 seamounts that were “discovered” circa 2007.



Eastern Samoa Bathy Compilation zoomed to Rose Atoll, where newly discovered seamounts are mapped, but we also want to know about their tectonic origin, if they are hotspots for biodiversity, other spatial features and processes in RELATIONSHIP to the seamounts.

In other words, "This is not about your eyeballs on a map, it's looking at the invisible rubber bands of mathematical manipulation of these different layers."

COUPLING of the appropriate data, analysis, and compelling design to effectively communicate the scientific results, benefits of exploration.



We are moving from an age where people are buying separate pieces, the hardware, the data, the software, etc., to a new trend of SaaS and other services

Stop moving data, push algorithms TO the data.

- subscription models are taking hold
- ecosystem must be open, interoperable, multi-organizational in its nature with multiple revenue streams that incentivize
    - app developers
    - data providers
    - infrastructure providers (e.g., Amazon, Microsoft)

http://esri.github.io/gis-tools-for-hadoop/

Currently no formal, accredited academic degree or curricula in ocean data management!

Given the importance of data science to ocean exploration, this is something that we need to fix! UNESCO IODE OceanTeacher provides resources as a start.

There are even ethical issues in data and information to be taught. A case study of this accident is included in my ethics of geographic information course at Oregon State U. http://dusk.geo.orst.edu/ethics and http://gisprofessionalethics.org

•Los Angeles class Submarine ran aground in route from Guam to Brisbane, Australia - 8 January, 2005

•One sailor killed, 120 injured

•Crash depth ~160 m, speed 33 kn, Sonar measured a depth of 2000 m, 4 minutes before crash

FINAL SLIDE with takeaway messages to summarize
How to best analyze, use, communicate ocean data.