

Big Science, Big Data, Big Challenges! Providing Open Access to Complex Data Sets Generated by Large-scale Physics Experiments

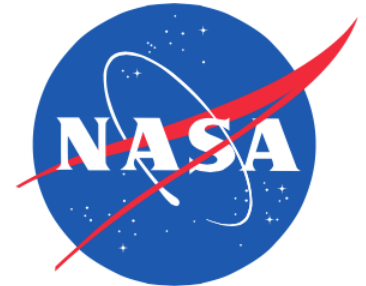
David Reitze
LIGO Laboratory
California Institute of Technology

Topics

- | Big Data from Big Science – Opportunities and Challenges
- | Views from Agencies and the Broader Research Community
- | The case of Big Physics with an Emphasis on LIGO
- | Big Science Data Implementation Issues and How LIGO Has Addressed Them

Making Data Publicly Available: Not a New Idea, but a Good Idea

- | Many excellent examples of successful data programs for releasing data to the broader research community and the public
 - | NASA missions, Sloan Digital Sky Survey, Human Genome Database, Protein Data Bank, ...
- | There are strong motivations to make data open
 - » New eyes and different perspectives on the data lead to new results
 - » Enables and facilitates the rise of Citizen Science
- | ***In a very real sense, this is a natural extension to the way science is already done!***
- | BUT ...
- | ... making their own data publicly available is an entirely new concept for many researchers
- | ... there are costs, both tangible and intangible, associated with making data public



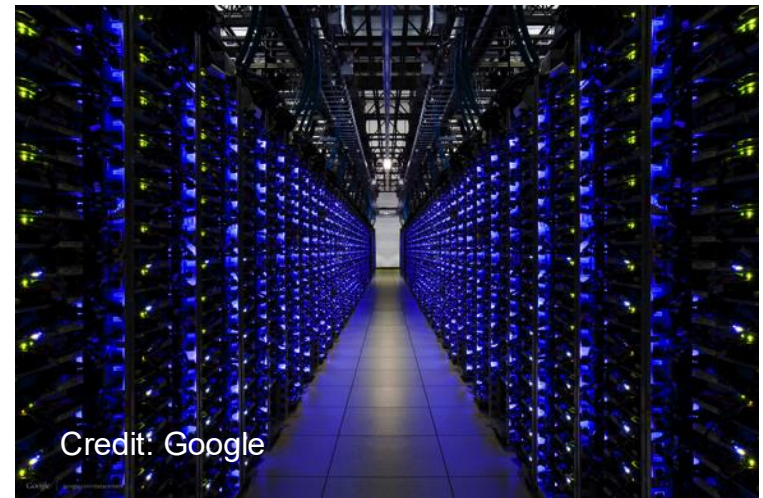
Credit: Halfblue, English language Wikipedia

Special Challenges Posed by 'Big Science' Data

- I **Immense Data Sets** (and getting bigger all the time...)
 - » In Physics, the big data producer is the Large Hadron Collider
 - collects 25 petabytes/year, equivalent to more than 5 million DVDs
 - » Coming in the next decade – the Square Kilometer Array
 - projected data rates of petabits/s (!)
- I **Complicated Data Sets**
 - » Nonstandard file formats
 - » homegrown file management systems
 - » simulated signals present in the data stream
 - » ...
- I **Old Data Sets**
 - » Long term data curation of older data is lacking
 - » Storage of old data has not migrated to current technology standards
 - » Software to read files incompatible with current operating systems
 - » ...



Credit: Reidar Hahn, Fermilab



Credit: Google

The Push from Above: View from the National Science Foundation

- | The National Science Foundation (via the National Science Board) has been ahead of the curve on pressing for public data access
 - » The NSF has long had a requirement for investigators to make data available for any NSF-funded projects
 - » **Most investigators rarely do so, or are called upon to do so.**
- | With respect to NSF Large Projects: In March 2008, the NSB approval of Advanced LIGO Project indicated that LIGO data would need to be publicly accessible
 - » “NSF management shall report back to the National Science Board within 12 months of this award what efforts have been undertaken and what provisions have been implemented to make the data obtained under this award available and useable to the broader research community.”
- | In 2009, LIGO worked with the NSF to develop a Data Management Plan containing a plan for making LIGO data publicly available.

The Push from Above: View from the National Science Foundation

- | In 2011, the NSB formally released ‘Digital Research Data Sharing and Management’ (NSB-11-79), its policy statement on access to digital data:

“The Board is committed to the development, implementation, and assessment of policies that promote efficient management of, and broad access to, digital research data that result from NSF-funded activities. This commitment includes sharing of results, data, physical collections, and other supporting materials created or gathered in the course of NSF-funded research.”

Key Challenge #2: As data collections expand in scale, scope, and complexity, successful data sharing and management require a change in research and institutional cultures.

www.nsf.gov/nsb/publications/2011/nsb1124.pdf

The Push from Above: View from the National Science Foundation

- | In 2011, the NSB formally released ‘Digital Research Data Sharing and Management’ (NSB-11-79), its policy statement on access to digital data:

“The Board is committed to the development, implementation, and assessment of policies that promote efficient management of, and broad access to, digital research data that result from NSF-funded activities. This commitment includes sharing of results, data, physical collections, and other supporting materials created or gathered in the course of NSF-funded research.”

Key Challenge #3: Data sharing requires the coordination of goals and efforts through international collaborations and activities.

www.nsf.gov/nsb/publications/2011/nsb1124.pdf

The Push from Above: View from the National Science Foundation

- | In 2011, the NSB formally released ‘Digital Research Data Sharing and Management’ (NSB-11-79), its policy statement on access to digital data:

“The Board is committed to the development, implementation, and assessment of policies that promote efficient management of, and broad access to, digital research data that result from NSF-funded activities. This commitment includes sharing of results, data, physical collections, and other supporting materials created or gathered in the course of NSF-funded research.”

Key Challenge #7: Data stewardship is critical to the longevity and sustainability of data sharing and management throughout the data lifecycle, but it is unclear where the responsibilities for this effort lie.

www.nsf.gov/nsb/publications/2011/nsb1124.pdf

The Push from Above: View from the National Science Foundation

- | In 2011, the NSB formally released ‘Digital Research Data Sharing and Management’ (NSB-11-79), its policy statement on access to digital data:

“The Board is committed to the development, implementation, and assessment of policies that promote efficient management of, and broad access to, digital research data that result from NSF-funded activities. This commitment includes sharing of results, data, physical collections, and other supporting materials created or gathered in the course of NSF-funded research.”

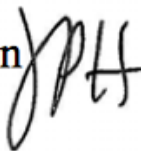
Key Challenge #10: Access to confidential data poses ethical and legal challenges.

www.nsf.gov/nsb/publications/2011/nsb1124.pdf

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren 
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

“The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and **digital data**.”

http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

“Bottom Up’ Pressure

- | Investigators want to look at data generated by others to further their own scientific interests
 - » Astronomy (particularly space-based astronomy) has developed a culture of releasing data, but also true of other fields
 - » Less true of highly coordinated Big Physics projects

- | Anecdotal evidence of bottom up pressure – from an astronomy blogger:
 - » “In the afternoon, _____ gave a seminar about gravitational radiation detection, particularly in the light of coincident electromagnetic signals. **At the end, discussion devolved into the insane data-secrecy policies of *LIGO*, which I abhor. *Occupy LIGO!*”**

Big Science Culture (a la Physics)

- | Large-scale physics projects typically involve 100s or 1000s of scientists and engineers who self-organize into collaborations:
 - » ADMX, Alice, Alpha, Argus, Atlas, ATRAP, BABAR, Bell, BESII, CDF, CDMS, CMS, COMPASS, ...
- | By and large, collaborations in a particular subfield or working on a specific physics target compete with each other, leading to attitudes such as:

“It’s my data, why should I share it with anyone else??”

Big Science Culture (a la Physics)

- | Large-scale physics projects typically involve 100s or 1000s of scientists and engineers who self-organize into collaborations:
 - » ADMX, Alice, Alpha, Argus, Atlas, ATRAP, BABAR, Bell, BESII, CDF, CDMS, CMS, COMPASS, ...
- | By and large, collaborations in a particular subfield or working on a specific physics target compete with each other, leading to attitudes such as:

“I’m the expert; who else could possibly analyze my data?! Except the competition, and I can’t have that!!”

Big Science Culture (a la Physics)

- | Large-scale physics projects typically involve 100s or 1000s of scientists and engineers who self-organize into collaborations:
 - » ADMX, Alice, Alpha, Argus, Atlas, ATRAP, BABAR, Bell, BESII, CDF, CDMS, CMS, COMPASS, ...
- | By and large, collaborations in a particular subfield or working on a specific physics target compete with each other, leading to attitudes such as:

“I’m a busy scientist, I shouldn’t have to dedicate some of my precious time to making my data available to others.”

Major Implementation Issues Facing Big Physics Public Data Release Programs

- | *What is the nature of the data, the size of the data stream, and what analysis tools are needed to manipulate and analyze the data?*
- | *When should the data be released?*
- | *Who are the broader research community of potential users?*
- | *What is the added cost of releasing the data and who pays?*
- | *What are the implications regarding intellectual property?*
- | *What are the implications of making data publicly available which is 'owned' by international collaborations?*

LIGO: the world's leading facility for searching for gravitational waves

Gravitational waves – propagating fluctuations of space-time itself; predicted by Einstein in 1916

Gravitational waves provide a unique way to probe the most violent astrophysical events in the universe

- including colliding black holes and neutron stars, supernovae, possibly the Big Bang



Once detected, gravitational waves will open a completely new window onto the universe

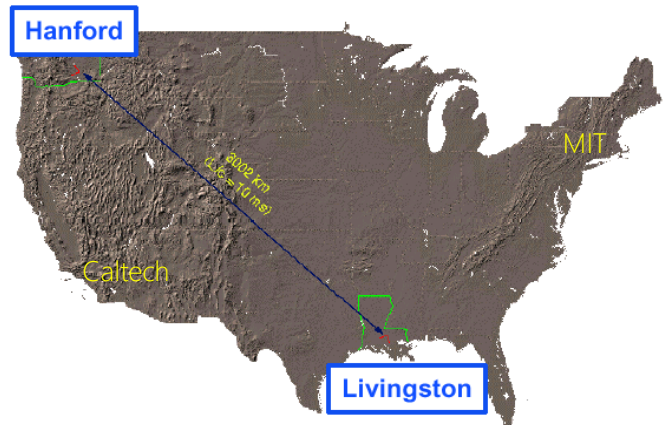
LIGO Hanford Observatory, WA, USA



LIGO: the world's leading facility for searching for gravitational waves

- LIGO uses precision laser interferometry with 4 km long arms in one of the world's largest vacuum systems
- Passing gravitational waves stretch and compress the distance in the arms of the interferometer
- LIGO is sensitive to distance changes smaller than 10^{-18} m

LIGO Laboratory is jointly operated by the California Institute of Technology and the Massachusetts Institute of Technology

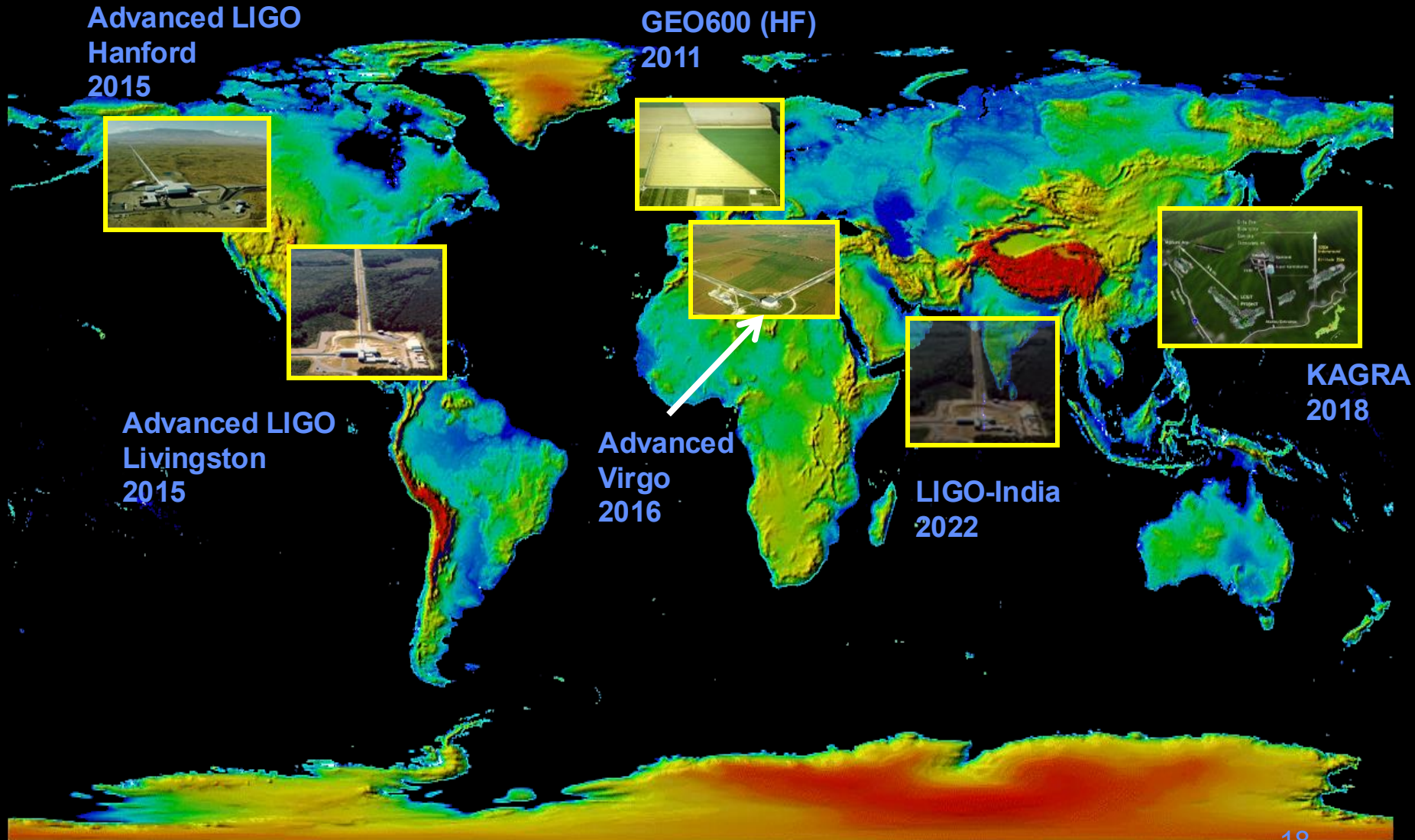


LIGO Livingston Observatory, LA, USA

- Currently in a major upgrade phase to Advanced LIGO
 - Beginning science operations in 2015



The advanced gravitational-wave detector network



Gravitational-wave Science Collaborations Are (Somewhat) Different

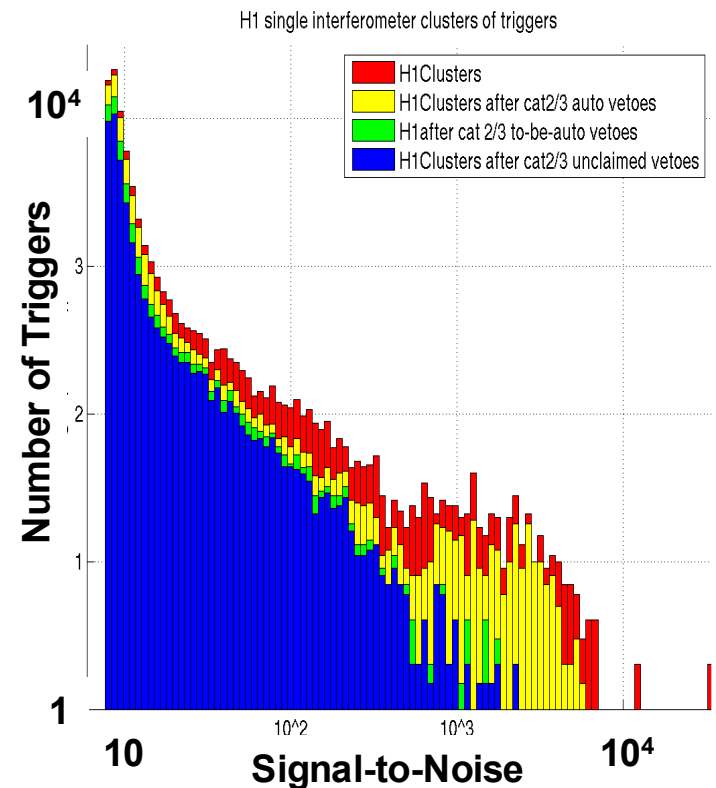
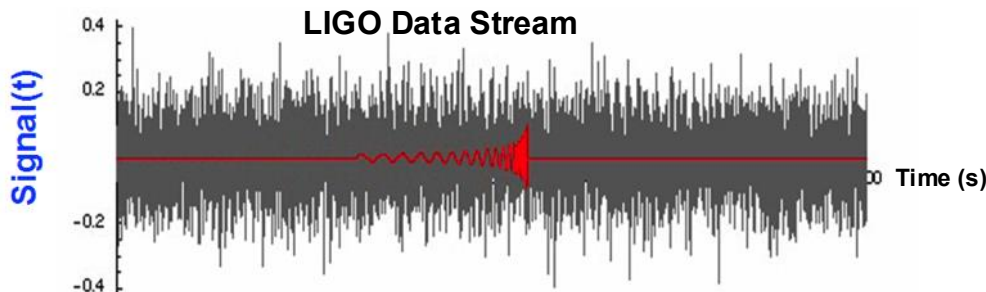
- | Gravitational-wave science is *enhanced* by having multiple detectors working collaboratively
 - » It leads to more statistical confidence in detections, better physical parameter extraction, higher accuracy in localizing gravitational-wave sources
- | **All international gravitational-wave detector collaborations share data with each other**
 - » The scientific benefits of sharing data outweigh the benefits of competing with each other

But...

- | Even within gravitational-wave detector collaborations, releasing data engenders strong feelings and reservations
 - » How to preserve 'first detection' for the collaborations?
 - » Will an open data policy dis-incentivize collaboration members from being part of the collaboration?
 - » Will we have to defend against false detection claims?

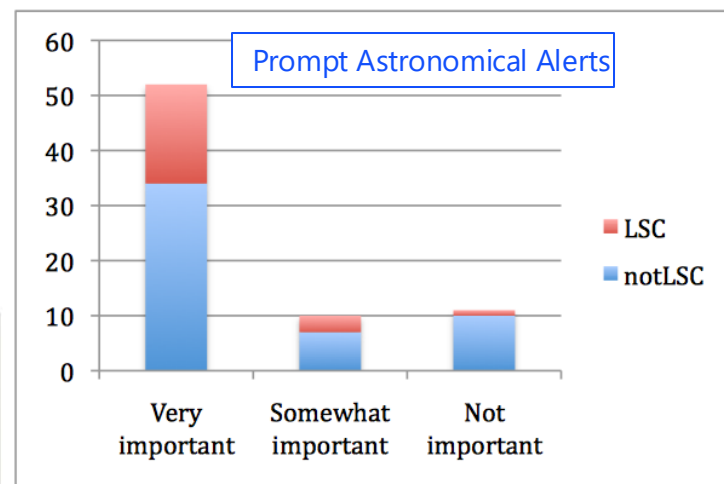
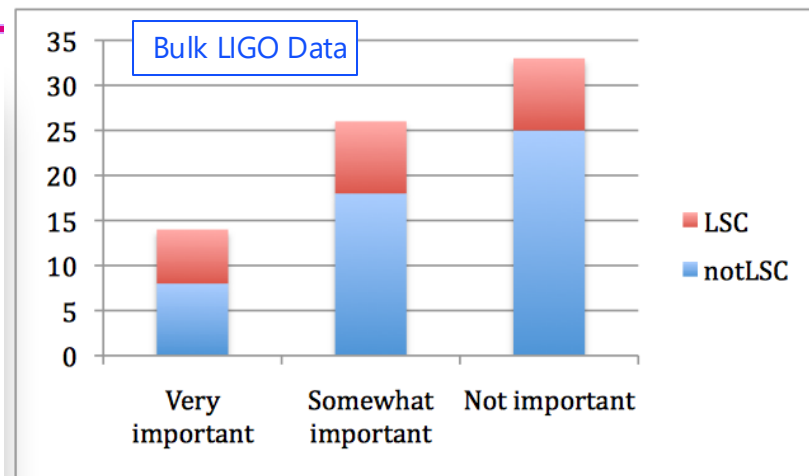
There's data, and then there's data

- | LIGO data comes in different flavors
 - » Gravitational wave data: 10 TB/yr
 - » Auxiliary channel data: 1 PB/yr
 - » Metadata: small
- | Raw LIGO data is messy and ugly → data quality verification is time-consuming and imperfect
 - » There are millions of false positives in each detector in a typical LIGO data run
- | Roughly 50 LIGO scientists work to clean LIGO data and assure data quality
- | Cleaning the data takes time
 - » The process isn't perfect, and artifacts remain in the data, requiring careful evaluation of high confidence triggers
 - » This has implications for the latency of the data release



“If you build it, will they come?”

- | An understanding of what types of data are most useful to the broader research community is crucial to implementing a successful strategy
- | In LIGO’s case, there are broadly three data classifications
 1. Full LIGO GW and ancillary data streams
 2. Small data segments associated with gravitational wave or electromagnetic events (such as supernovae, gamma ray bursts)
 3. Astronomical alerts: times, sky locations, error boxes, and detection confidence associated with events (to be provided within minutes to an hour of the event)
- | We conducted a ‘market survey’ to determine which types of data were most interesting



'There ain't no such thing as a free data lunch...'

POLICYFORUM

SCIENCE PRIORITIES

Who Will Pay for Public Access to Research Data?

Francine Berman¹ and Vint Cerf²

9 AUGUST 2013 VOL 341 SCIENCE www.sciencemag.org



- | **Serving large, complex data sets to the broader research community and the public costs money**
- | For large projects, costs/estimates vary widely
 - » Approaches 10% of operating costs for some NASA missions
 - » LHC data preservation efforts estimated at 1% of operating costs
- | LIGO's budget is 1% of operating costs, about \$400k/yr
 - » Limited by overall budget constraints, we could do more if we had more
 - » Covers salary for 1.5 full time equivalent staff, servers, travel

LIGO Who owns the Intellectual Property that public data generates?

- | Q: If I work a company/university and receive US government funding for my research, who owns the IP that I generate with data I've taken?
A: I do!
- | Q: If I work a company/university and receive US government funding for my research, who owns the IP that someone else generates with data I've taken?
A: I do (?) Do I?
- | Policies are needed at the agency level which provide clear guidelines for awardees
 - » From NSB 11-79: “new data licensing mechanisms can preserve intellectual property rights and provide researchers with incentives to make their data public.”

- | Two phases/eras of data release
 - » Phase 1 ‘Discovery Era’– release of small data segments (detections or segments near noteworthy electromagnetic detections) upon publication in peer-reviewed journal
 - Phase 1 starts as soon as Advanced LIGO is operational
 - » Phase 2 ‘Observational Era’ – release of full sets of cleaned gravitational-wave data in 6 months cadence with a latency of two years
 - Phase 2 can be triggered by calendar, by astrophysical threshold (‘space-time volume’) or by discretion of the LIGO Scientific Collaboration
- | Also provides for long term curation of all LIGO data
- | What about alerts?
 - » For first four detections, released through collaborative agreements with astronomer partners; after that, public release
 - » Dictated by agreements with Italian-French Virgo detector



LIGO Open Science Center

LIGO is operated by California Institute of Technology and Massachusetts Institute of Technology and supported by the National Science Foundation of the United States.

Getting Started

Data

Bulk Data

Event Lists

Timelines

Toolbox

Tutorials

Software

My sources

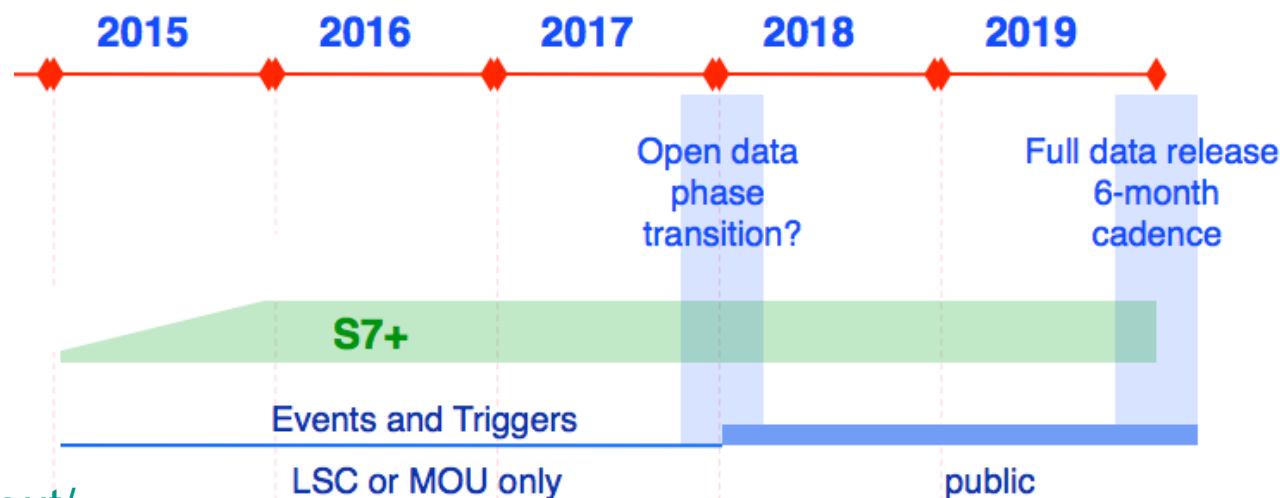
GPS ↔ UTC

Learn More

About LIGO

Contact

- | LIGO Open Science Center is the portal through which LIGO data will be released
- | Currently under development and internal review
- | Work under way to release older data sets accumulated during initial LIGO science runs during 2005 - 2010
- |



<https://losc.ligo.org/about/>



LIGO Big Science, Big Data, Big Challenges?Big Opportunities!

- | Big Science is everywhere -- Astronomy, Biology, Chemistry, Geology, Medicine, Physics, ...
- | And it is growing - the continuing move toward interdisciplinarity and the cost of doing science will push all scientific communities in the direction of virtual collaborations with a 'Big Science' character
 - See 'A Sciences Services Marketplace', Elizabeth Iorns in 'Session Outsourcing Science: Will the Cloud Transform Research?' from Friday's session
- | Making 'Big Science' data broadly available will both enhance existing collaborative science and foster new opportunities
- | The methods, costs, and implications must be carefully addressed

Thanks to: Stuart Anderson, Roy Williams, Alan Weinstein, Albert Lazzarini (Caltech), Gabriela Gonzalez (LSU), the LIGO Scientific Collaboration



www.ligo.org

Support: National Science Foundation



Thanks!