National Data Service

THE NATIONAL DATA SERVICE(S) & NDS CONSORTIUM

A Call to Action for Accelerating Discovery Through Data Services we can Build

Ed Seidel

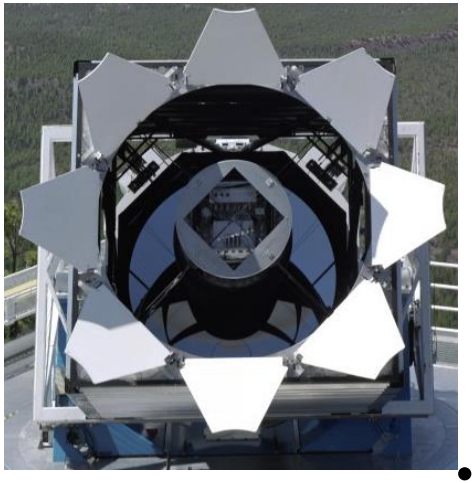
National Center for Supercomputing Applications

University of Illinois Urbana-Champaign

Data-enabled Transformation of Science

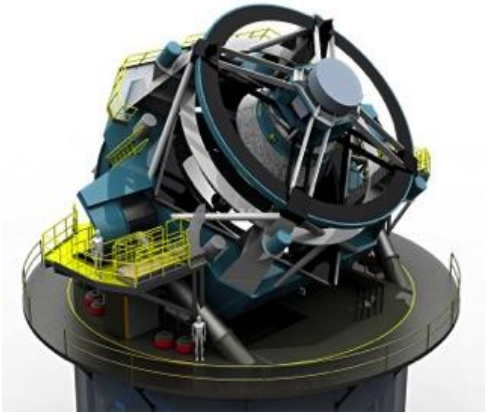
Astronomy 1500- 2000:

- Single scientist looks through telescope
- Record KB of data in notebook
- Require reproducibility



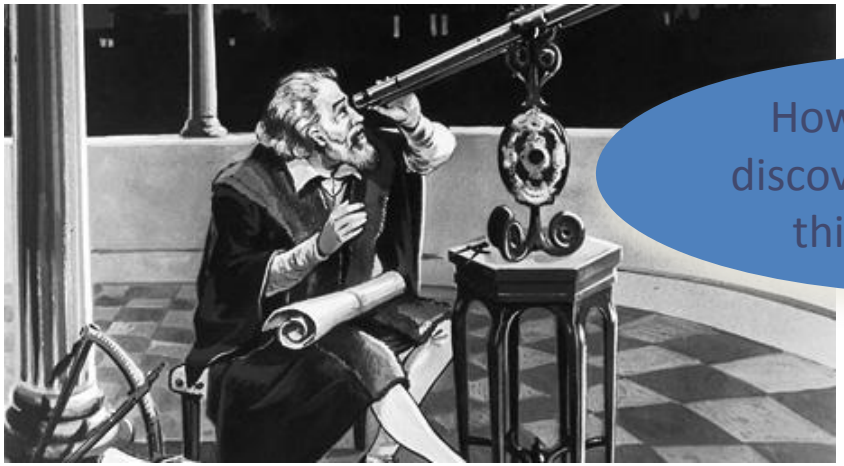
Sloan Digital Sky Survey 2000+

- Record data for decade (40TB)
- Serve to entire world
- Thousands of scientists work “together”

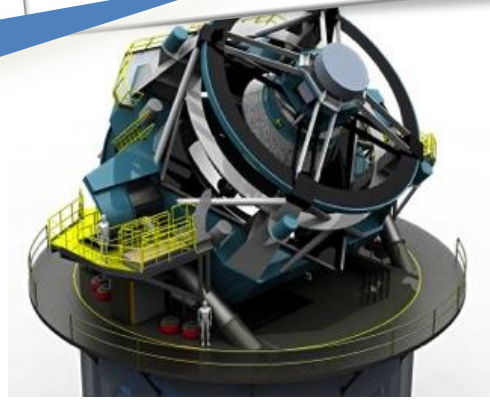


- DES (now)
  - 200GB/night
  - PB in decade
- LSST (6 years)
  - Record data for decade
  - SDSS/night!
  - 200 PB/decade

Data-enabled Transformation of Science



How can I publish, discover, verify data in this new world?



Astronomy 1500- 2000:

- Single scientist looks through telescope
- Record KB of data in notebook
- Require reproducibility

Sloan Digital Sky Survey 2000+

- Record data for decade (40TB)
- Serve to entire world
- Thousands of scientists work “together”

- DES (now)
  - 200GB/night
  - PB in decade
- LSST (6 years)
  - Record data for decade
  - SDSS/night!
  - 200 PB/decade

Big Data vs The Long Tail of Science

- Many “Big Data” projects are “special”
  - Highly organized, singular sources of data, professionally curated, a lot attention paid
- What about the “Long Tail” (the other 99%)?
  - 1000s of biologists sequencing communities of organisms
  - Thousands of chemists and materials scientists developing a “materials genome”
  - Characteristics:
    - Heterogeneous, perhaps hand generated
    - Not curated, reused, served, etc…

MATERIALS GENOME

Big Data vs The Long Tail of Science

- Many “Big Data” projects are “special”
  - Highly organized, singular sources of data, professionally curated, a lot attention paid
- What about the “Long Tail” (the other 99%)
  - 1[illegible] communities of organ[illegible]
  - Thousands of chemists and materials scientists developing a “materials genome”
  - Characteristics:
    - Heterogeneous, perhaps hand generated
    - Not curated, reused, served, etc…

Fundamental Observation: Scientists communicate by sharing data…

MATERIALS GENOME

A Data Scenario…

*In 2021 the **LIGO** gravitational wave observatory detects a "transient" burst event. An alert is issued….*

Communities share data, software, knowledge, in real time…

A Data Scenario…

Scientists (many of whom have never worked together) engage discovery services, find data from the IceCube neutrino observatory, isolating the source in the sky….

Communities share data, software, knowledge, in real time…

A Data Scenario…

Discovery services also connect researchers to data of DES and LSST to look for E&M precursors…

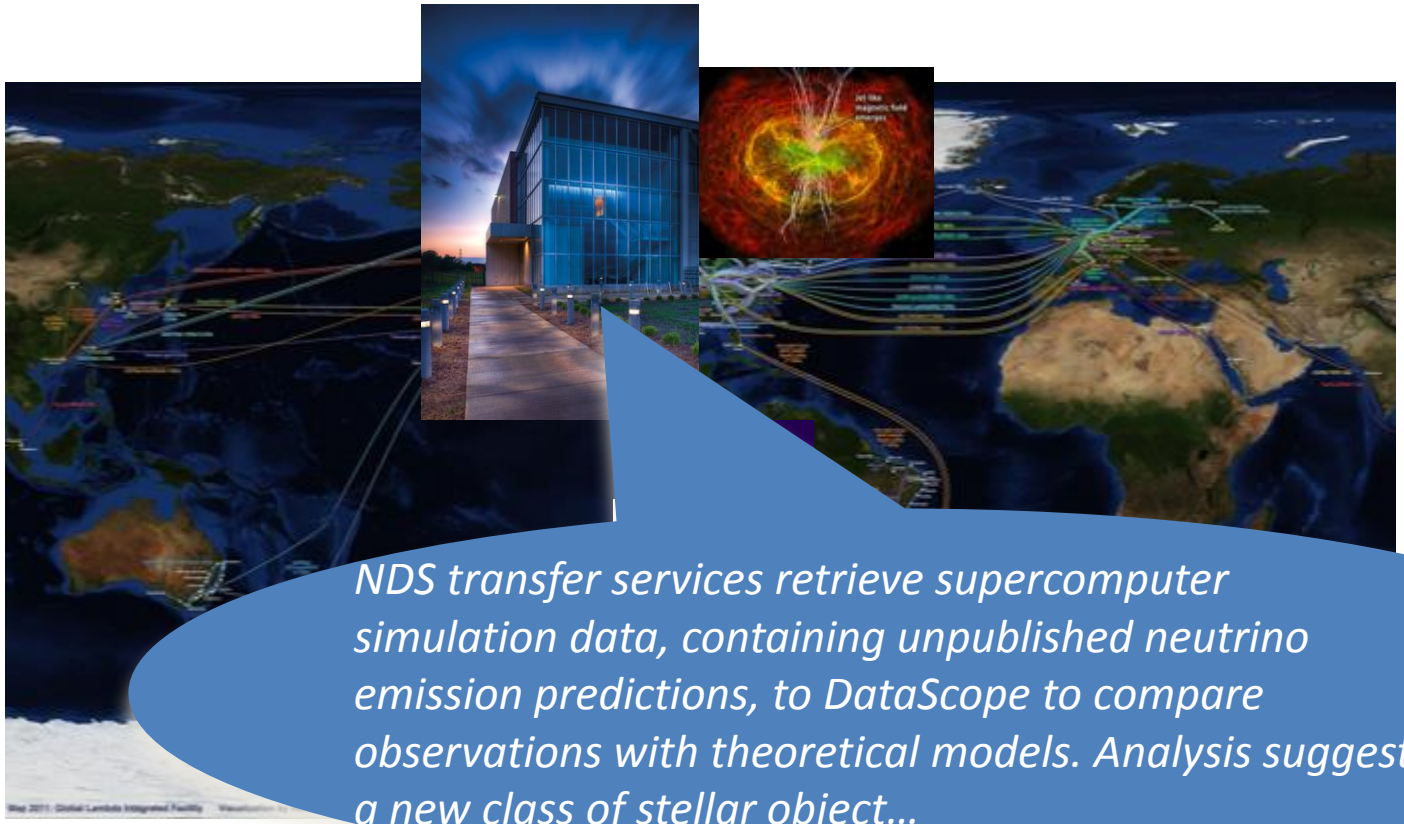
Communities share data, software, knowledge, in real time…

A Data Scenario…

Literature searches find publications describing similar detections; Elsevier, APS publications and arXiv preprints, supporting NDS data linking, lead them to the data underlying the analyses…

Communities share data, software, knowledge, in real time…

A Data Scenario…

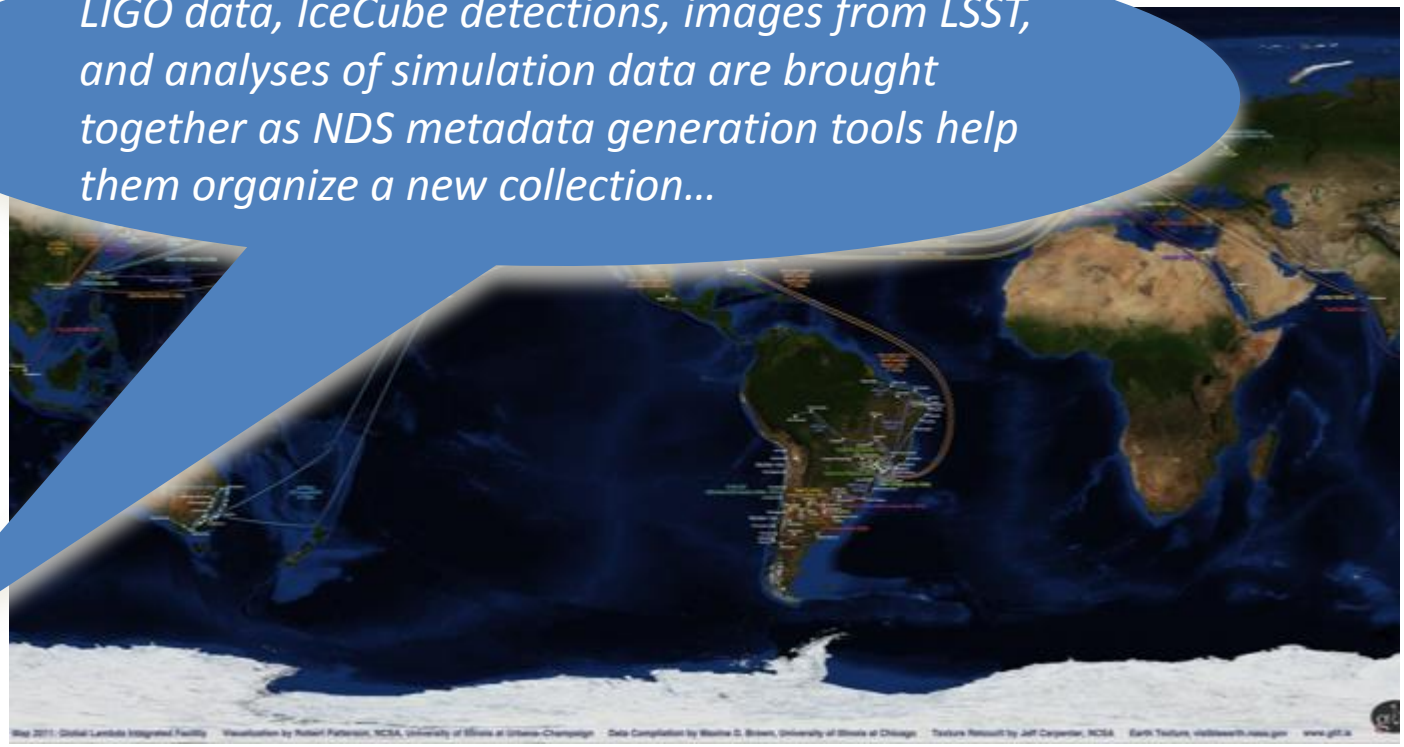


NDS transfer services retrieve supercomputer simulation data, containing unpublished neutrino emission predictions, to DataScope to compare observations with theoretical models. Analysis suggests a new class of stellar object…

Communities share

A Data Scenario…

LIGO data, IceCube detections, images from LSST, and analyses of simulation data are brought together as NDS metadata generation tools help them organize a new collection…



Communities share data, software, knowledge, in real time…

A Data Scenario

A publication is submitted to an OA journal, with identifiers for the new data collection included in the paper... Once the paper is accepted, the linked NDS data collection is sent to a campus repository for longer-term curated management...

PHYSICAL REVIEW X

committed to excellence

Communities share data, software, knowledge, in real time…

Readers have access to the underlying data, enabling them to verify and extend results. Data are further available to educators, who bring the discovery to a broad audience by updating astronomy e- textbooks.

PHYSICAL REVIEW X
committed to excellence

Communities share data, software, knowledge, in real time…

Digital Data

- Virtually all scien
  - But it is becomin
- Growing call for op
  - In many fields, scientist om on-line archives
  - NSF, NIH, other age s in US, world requiring more and more…
  - White House Open Data Policy supports sharable data at scale
- Despite this, there are no uniform mechanisms to store, share and re-use data

"We need to take steps to make scientific research data more liquid. The more we move towards open as the default for scientific research data, the more we will get out of the research enterprise. It is time to take deliberate steps to make that a reality." Mike Stebbins, White House OSTP

Inaccessibility of data

- Scientific results are shared through literature
  - *No uniform mechanisms to publish data that results are based on*
  - *Can't cite data just as readily*
- Some disciplines/communities are quite advanced in making data available
  - Discipline-specific federations address peculiar problems of metadata, data types, data formats
    - NEON, LTER, VAO, SEAD, EarthCube, HASTAC, …
    - We'll hear from three tomorrow
  - *Not for all data, nor for all disciplines*

“Long-tail” data problem

- Major surveys, missions, and projects provide standard archived products
- *Many more smaller collections remain inaccessible*
  - Datasets created by small research groups
  - Datasets (painstakingly) refined by further processing for specific science questions
  - New datasets created by synthesizing existing datasets
- *No universal tools to create, publish data collections*

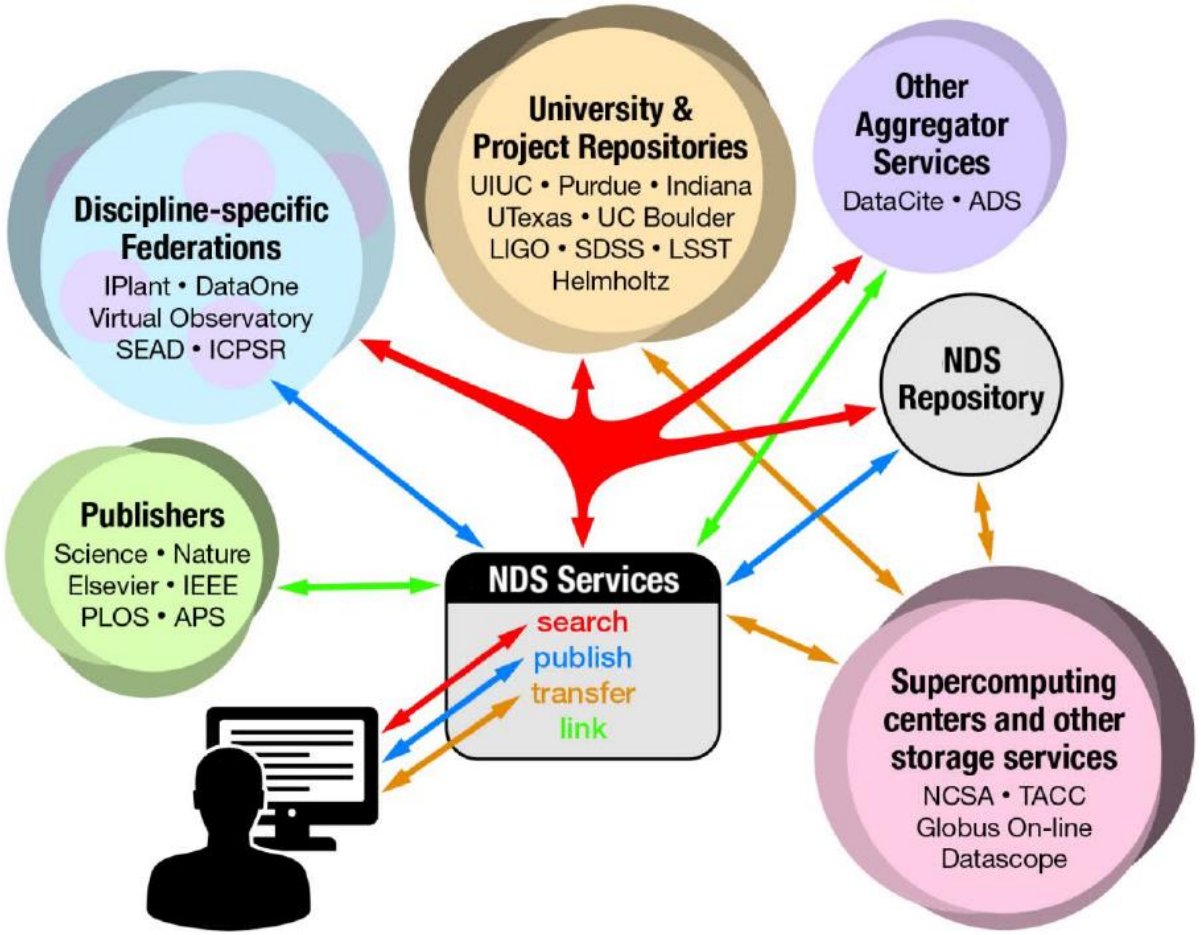
National Data Service(s)

Urgent need for national infrastructures for data

- Integrated set of national-scale services
  - Storing, sharing, finding, verifying, publishing, citing, reusing
- Open and federating architecture
  - Building on infrastructure at discipline/community level
  - Enable providers to make data accessible in national environment
  - Allow new and community-produced tools and resources to be plugged in
- Enable revolutionary new research
  - Rapid discovery and easy re-use of data
  - Unprecedented cross-disciplinary research
  - Greater collaboration
  - Make science verification more routine

The NDS Ecosystem

- Connects data resources into an interoperable web
  - Federations
  - University Libraries
  - Major Projects
    - MREFC, DIBBs, etc.
  - National and Regional Labs and Computing Centers
  - Publishers
- Connect/unify data capabilities:
  - Data discovery
  - Data publishing
  - Data movement
  - Data-Literature linking



What should NDS do for researchers?

- Help researchers ***find data***
  - Cross-disciplinary searching: federations, projects, archives, and other repositories
  - Find data related to a publication
  - Leverage specialized community-specific services
- Help researchers ***use data***
  - Download data, browse metadata, track provenance
  - Move data to processing platforms for specialized (re-)processing and analysis

What should NDS do for researchers?

- Help researchers ***share and publish data***
  - NDS Repository: platform for publication
    - Sharing privately with collaborators prior to publishing
    - Tools to help organize the data for publishing
    - Automatically ensure links to literature:
    - Assign DOIs, provide links to publishers, synchronize data publishing with papers
  - Recommend appropriate discipline/community repository for long-term preservation
  - NDS Repository as archive of last resort

NDS Consortium

- NDS vision requires collaboration of many institutions
- Consortium to guide the development of the NDS
  - Coordinate separately funded efforts to build components
    - Ensure interoperability
    - Integrate existing tools and resources
  - Forum for developing new partnerships and proposals
- A number of organizations already with us
  - DataONE, LTER, LIGO Lab, IceCube, …
  - Many others with strong interest, poised to join
- What should the NDS Consortium look like?
  - Break out session today, facilitated discussion tomorrow

NDS Relationship with RDA

- RDA & NDS Consortium are connected
  - No attempt to duplicate or compete with RDA activities
  - Work closely with RDA groups to implement NDS
- NDS focus: create infrastructure framework, build on/interoperating with existing activities
  - Narrow focus on functions, leveraging existing capabilities
  - Tools that can be built and deployed within a few years
  - National structures: HPC centers, XSEDE, Internet2
  - Libraries, publishers ensure appropriate links to literature
  - Specific communities & projects plug in

What's Next: Let's Get to Work!

- Much of this is possible now
  - Lacking is…
    - Framework to integrate
    - National integration structures to deploy
  - Need doers and visionaries both!
- This meeting: let's look at
  - What should an NDS do? How should it be built?
    - Inventory what is already there, what is missing
  - Demonstrations to build momentum: much can be done now!
  - How do we fund it?
    - NCSA, others fully committed
    - Organize coordinated proposals to build missing components of services!
- Form a consortium
- Future meetings