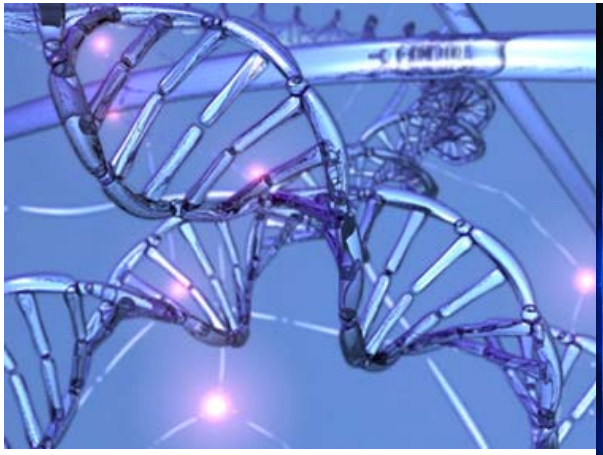


The Ecosystem of Scientific Data

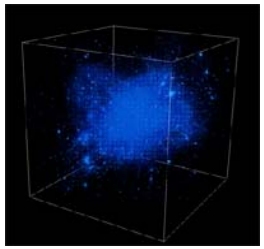
Alex Szalay
Institute for Data-Intensive Engineering and Science
The Johns Hopkins University

idies

Goals of this Workshop

- Build a community and establish TRUST
- How to optimize the use of shared cyberinfrastructure and decide what is currently missing
- Understand the implications and tradeoffs required for a sustainable data ecosystem
- Agree on a common vision how the legacy of our data is preserved

Trends



- Broad sociological changes
  - *Convergence of Physical and Life Sciences*
  - *Data collection in ever larger collaborations*
  - *Virtual Observatories: CERN, IVOA, NCBI, NEON, OOI,…*
  - *Analysis decoupled, off archived data by smaller groups*
  - *Emergence of the citizen/internet scientist (GalaxyZoo…)*
- How do we start new collaborations?
- How do people change the way they do things?

"The best collaborators are the desperate ones"

-- *Jim Gray*

Lifecycles Everywhere

- Data lifecycle, service lifecycle, instrument lifecycle

- Several phases in the life of a big project
- Growth, Consolidation, Legacy

- Each phase requires different skills and solutions

Phase 1: Organic Growth

- Everybody has good reasons to build the systems they have
- We cannot impose/force solutions on others
- In order to attract talent, we need to offer the ability to create something new
- This first phase of the data systems is driven by (domain) function and innovation

Phase 2: Consolidation

- This phase is about efficiency, stability and maturity
- We share technologies and data only after trust is established
- We discover that there are lots of similarities deep down across different domains
- There is a convergence emerging, even between physical and life sciences

Phase 3: Legacy

- This comes after the instruments are turned off
- The main issue is survival and legacy of the data
- We must not lose the bytes, but have to prioritize
- The hard thing is to ensure domain specific curation and corporate memory
- What are the tradeoffs, who decides?
- Who owns the data?
- What is the funding model?

How Do We Prioritize?

Data Explosion

- Science is becoming data driven
- It is becoming “too easy” to collect even more data
- Robotic telescopes, next generation sequencers, complex simulations
- Supercomputers are not just consumers of Big Data, they are also becoming the **source of Big Data**

How long can this go on?

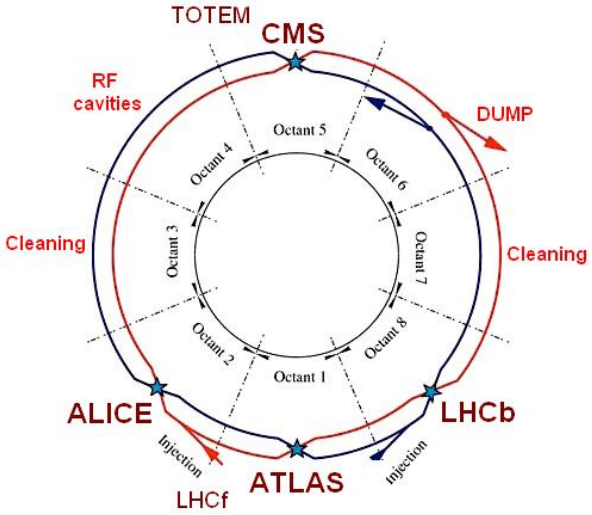
- “Do I have enough data or would I like to have more?”
- No scientist ever wanted less data….
- How can we decide how to collect data that is *more relevant?*

Numerical Laboratories

- On **Exascale** everything will be a Big Data problem
- Memory footprint will be >2PB
- With 5M timesteps => 10,000 Exabytes/simulation
  - *Impossible to store*
- Doing all in-situ limits the scope of science
- File sharing is not possible due to size
- Danger of losing support if the data is not accessible for the broad community
- This may be the biggest challenge on Exascale!
- Very relevant with NSCI

LHC Beamlines

- LHC has a single data source, $$$$$
- Multiple experiments tap into the beamlines
- They each use **in-situ** hardware triggers to filter data
  - *Only 1 in 10M events are stored*
  - *Not that the rest is garbage, just sparsely sampled*



- Resulting “small subset” analyzed many times **off-line**
  - *This is still 10-100 PBs*
- Keeps a whole community busy for a decade or more
- An excellent example of tradeoffs we will have to make

Needs of a Stable Ecosystem

- Must be distributed to be fault tolerant
- The NSF has been spectacularly successful in building a high speed network (CC-NIE)
- It is now possible to move large amounts of data (at least on the backbone)
- It is possible to process data at supercomputers
- Can one grow this to the next phase, by adding distributed storage?
- And the hardest: who will “own” and curate the data?
- How do we ensure a stable, sustained funding?
- Is there a role of libraries?

Goals of this Workshop

- Build a community and establish TRUST
- Agree on a common vision how the legacy of our data is preserved
- How to optimize the use of shared cyberinfrastructure and decide what is currently missing
- Understand the implications and tradeoffs required for a sustainable data ecosystem

- *Use breakouts to create an open discussion*
- *Trying to get people out of their comfort zone (slightly)*
- *Best and worst outcomes…*