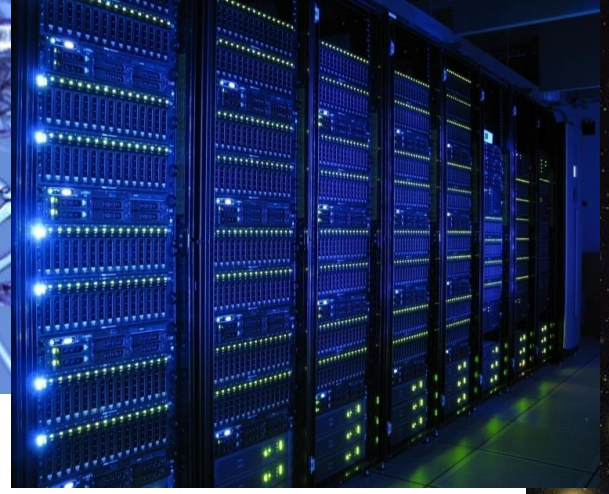
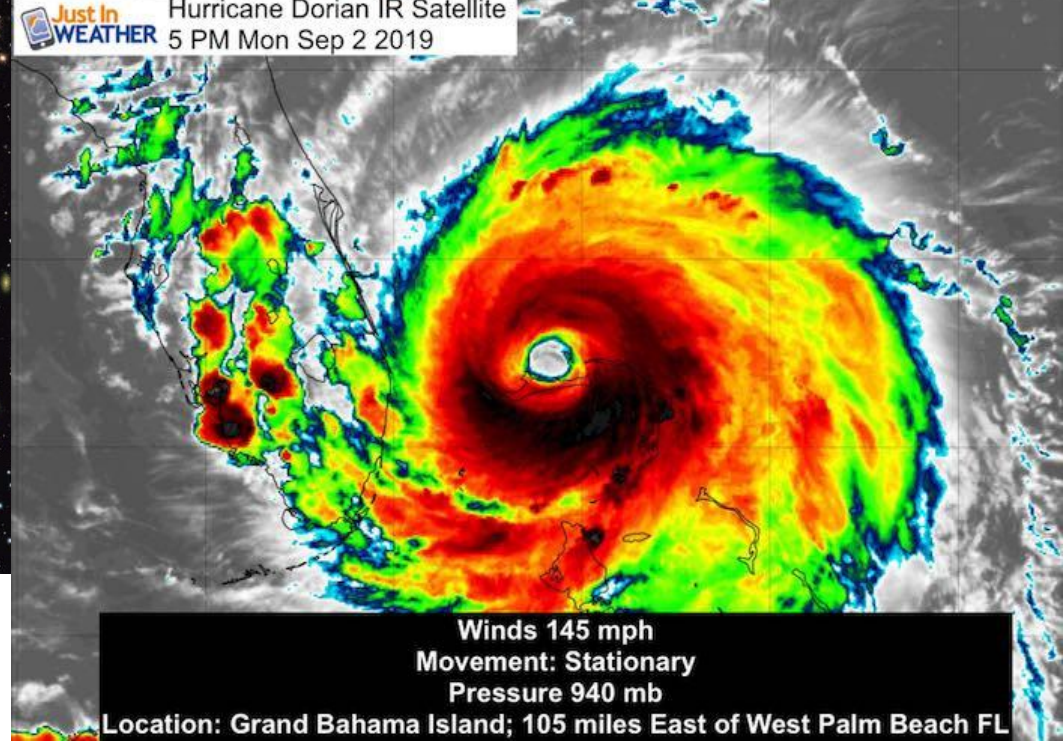





idies

# Lessons from the Sloan Digital Sky Survey

Alex Szalay
The Johns Hopkins University

# Big Data in Science

- Data growing exponentially, in all science
- All science is becoming data-driven
- This is happening very rapidly
- Data becoming increasingly open/public
- Non-incremental!
- Convergence of physical and life sciences through Big Data (statistics and computing)
- The “long tail” is important
- Scalability challenge
- A scientific revolution in how discovery takes place

  => a rare and unique opportunity

# Science is Changing

- We are moving from the era of “manufacture” to the “industrial revolution”
- Particle physics analogies:
  - *Van der Graaf -> cyclotron -> synchrotron -> National Labs -> SSC ☹ -> LHC ☺*
- Big Science projects are entering the multi-billion regime, their open data archives are analyzed by large communities

But there is a difference:

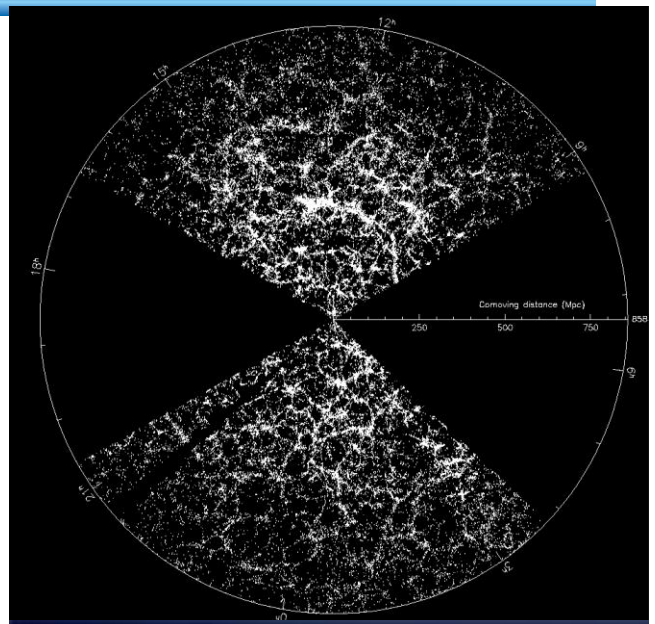
- In the past one experiment was followed by another in a short time, data sets had a short lifetime
- Today’s Big Science experiments (LIGO, LHC, LSST,OOI, NEON, IceCube) may not be surpassed by another in our lifetime

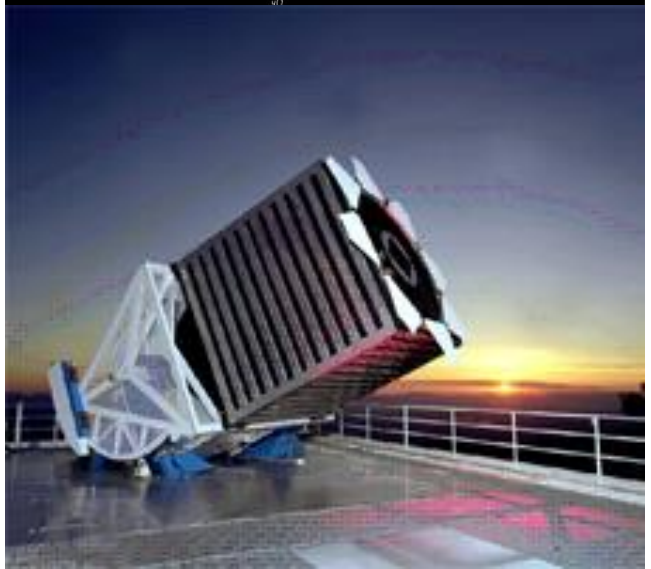
# Sloan Digital Sky Survey

## “The Cosmic Genome Project”

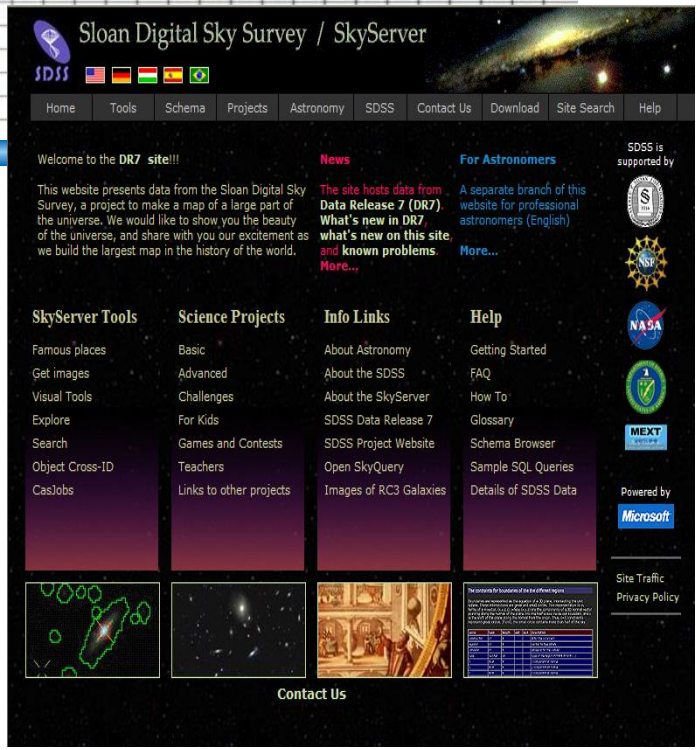
- Started in 1992, SDSS-II finished in 2008
- Data is public
  - 2.5 Terapixels of images => 5 Tpx of sky
  - 10 TB of raw data => 100TB processed
  - 0.5 TB catalogs => 35TB in the end

- Database and spectrograph built at JHU (SkyServer)
- Now SDSS-IV, data served from JHU
- SDSS-V starting soon



# Skyserver




Prototype in 21st Century data access

- *3.4B web hits in 12 years*
- *460M external SQL queries*
- *9,000 papers and 500K citations*
- *7,000,000 distinct users vs. 15,000 astronomers*
- *The emergence of the "Internet Scientist"*
- *The world's most used astronomy facility today*
- *Collaborative server-side analysis done by 9K astronomers*

# Why Is Astronomy Interesting?

Astronomy has always been data-driven….
now this is becoming more accepted in
other areas as well

*"Exciting, since it is **worthless**!"*

*— Jim Gray*

# The (Well Hidden) Long Tail

There is a lot of “open” data that never sees the light of day

- The “Long Tail” of a huge number of small data sets
  - *The integral of the “long tail” is big!*
  - *How do we integrate them with the large data collections?*
  - *Almost total failure so far*
- Facebook: bring many small, seemingly unrelated data to a single place and new value emerges
  - *What is the science equivalent?*
- The DropBox lesson
  - *Simple interfaces are more powerful than complex ones*
  - *API public*

# Data in HPC Simulations

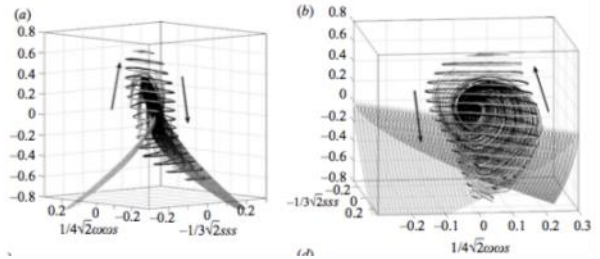
- HPC is an instrument in its own right
- Largest simulations approach petabytes today
  - *from supernovae to turbulence, biology and brain modeling*
- Need public access to the best and latest through interactive **Numerical Laboratories**

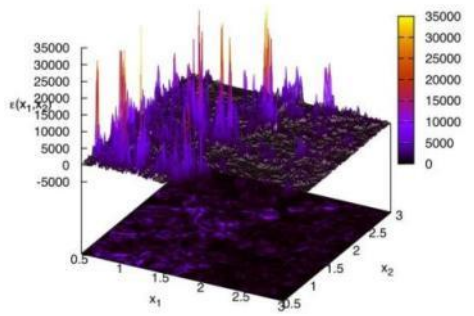
- Examples in turbulence, N-body
- Streaming algorithms (annihilation, halo finders)
- Exascale coming

# Immersive Turbulence

*"… the last unsolved problem of classical physics…" Feynman*

- **Understand the nature of turbulence**
  - *Consecutive snapshots of a large simulation of turbulence: 30TB*
  - *Treat it as an experiment, **play** with the database!*
  - ***Shoot test particles** (sensors) from your laptop into the simulation, like in the movie Twister*
  - *50TB MHD simulation*
  - *Now: channel flow 100TB, MHD 256TB*
  - *8K$^3$ simulation available*
  - *18K^3 in the works (PK Yang)*
- **New paradigm** for analyzing simulations!
- 78 Trillion points delivered in 5 years
- Total of 650TB of simulations accessible

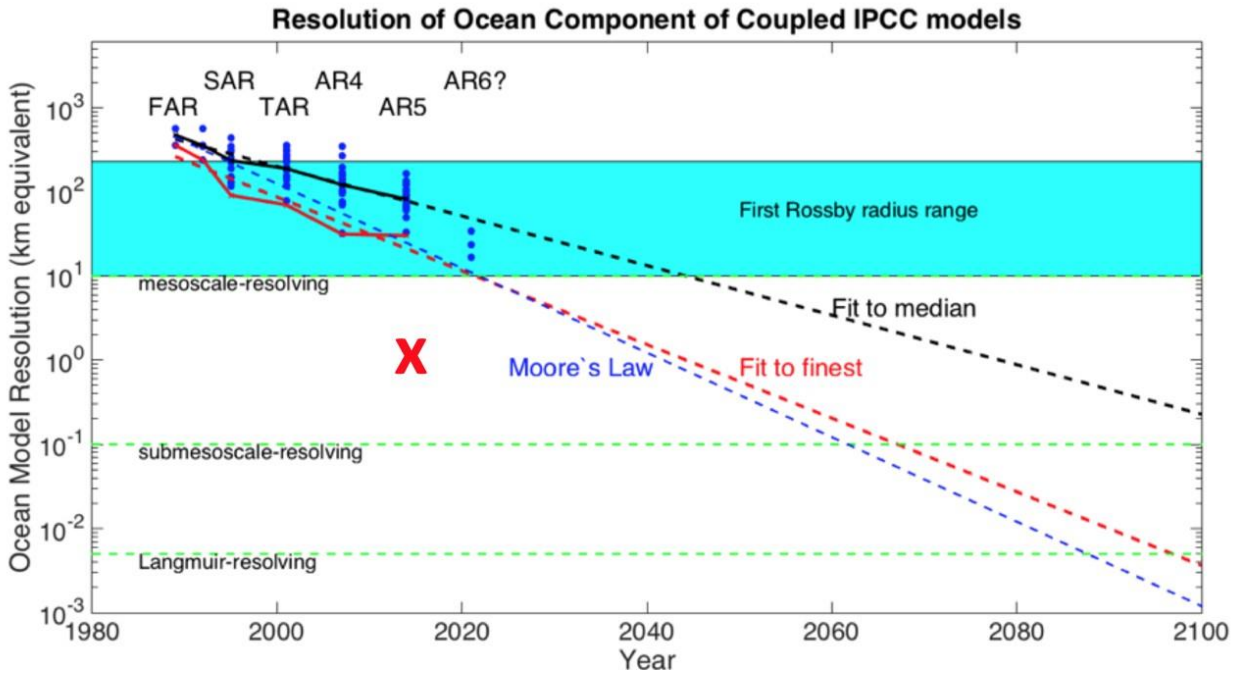




R. Burns, C. Meneveau, T. Zaki, G. Eyink, A. Szalay, E. Vishniac

# 2PB Ocean Laboratory

- 1km resolution whole Earth model, 1 year run
- Collaboration between JHU, MIT, Columbia
  - *T. Haine, C. Hill, R. Abernathy, R.Gelderloos, G. Lemson, A. Szalay, NSF $1.8M*



# From SkyServer to SciServer

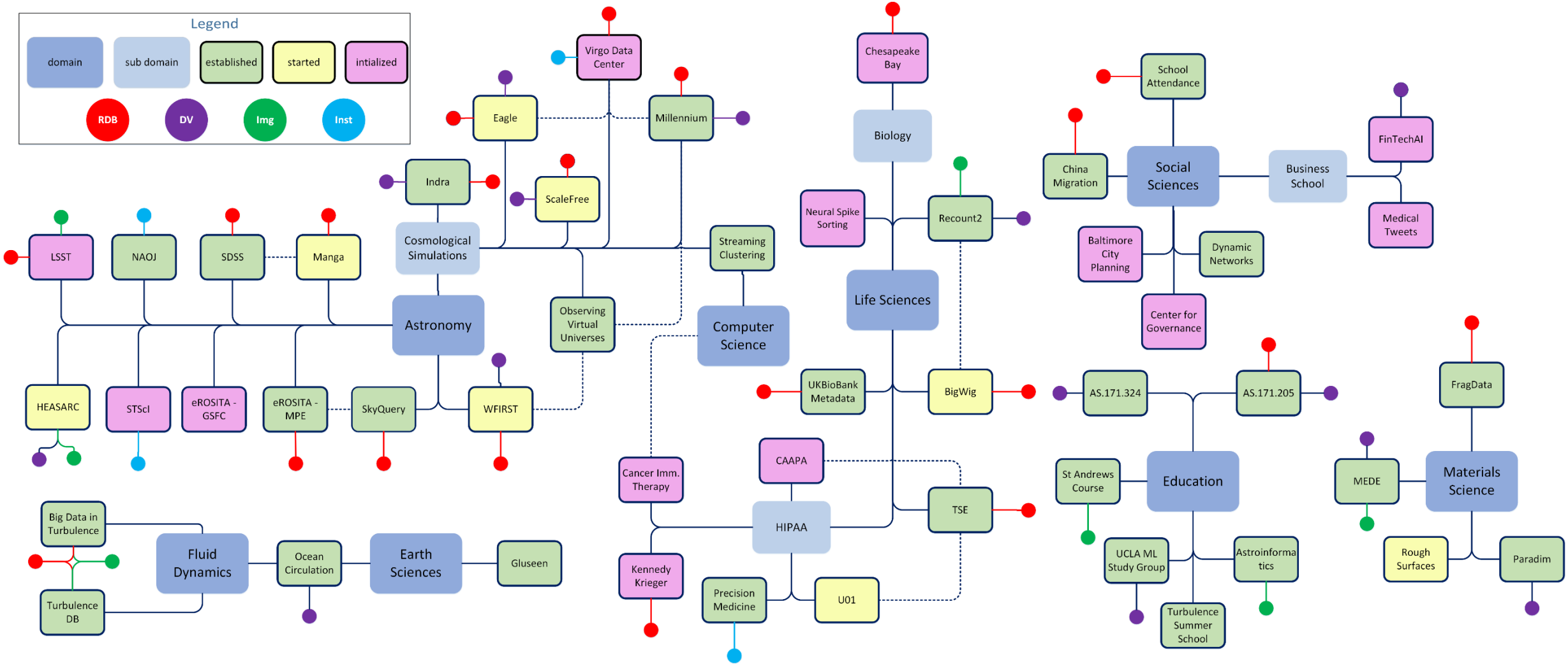
- Service oriented **smart** data
- More **collaborative** features added
- System captures **interactivity** of science well
- Read-only, secure core **– free-for-all in MyDB**
- Increasingly **complex analysis** patterns
- Extensive use by other disciplines
- But: signs of service lifecycle after 15 years

  *=> NSF DIBBS grant ($10M)*
- Comprehensive overhaul under the hood
- Now added iPython+R scripting, running on pool of VMs
- Single sign-on to CASJOBS + SciDrive (Dropbox+VOSpace)
- Currently about 3.5PB of live data, growing

# SciServer: Scalable Data Aggregator

- Difficult to aggregate large data sets: joint analysis requires co-location
- Most frequent mistake: trying to create the “mother of all databases”
  - *Building ontologies and data models is hard*
  - *We learned an enormous amount during the Virtual Observatory project*
- Real life uses require interactive exploration before big analysis
- The SciServer philosophy:
  - *Create Data Contexts, each with their own data model and ontology, self documenting*
  - *These are secure and read-only, under access control*
  - *User get their own databases and resources to create value added aggregations*
  - *These can be shared at will with authenticated users at owners discretion*
  - *We can bring in new datasets in isolation very quickly*

# Current SciServer Projects



# Lessons Learned

- Statistical analyses and collaboration easier with DB than flat files
- Adding own classes to DB core had dramatic impact on performance
- Automation is needed for statistical reproducibility at scale
- Scaling out was much harder than we ever thought
- Do not try to do “everything of everybody”, find the right tradeoffs using the 20 queries
- Find a common processing level that is “good enough” and earn the TRUST of the community
- Moving PBs of data is hard, importance of **smart data caching**
- Need **deep links** to the raw files

# Open Storage Network

- **NSF CC*: 150+ universities to connected at 40-100G**
- Ideal for a large national distributed storage system:
  - *Inexpensive (~$100K) storage (1.5PB) racks at each site (~200PB)*
  - *Store much of the NSF generated data*
  - *Provide significant shared storage for Big Data Hub communities*
  - *Distribute data from MREFC projects*
  - *Provide gateways/caches to XSEDE and cloud providers*
- Technology straightforward
  - *Automatic compatibility, ultra-simple standard API (S3)*
  - *Globus at the top layer (G-Connect, GlobusAuth)*
  - *Implement a set of simple policies*
  - *Enable sites to add additional storage blocks at their own cost*
  - *Variety of services built on top by the community*
- Estimated Cost: ~$20M for 100 nodes

SCHMIDT FAMILY FOUNDATION

NSF

- Current partnership: $1.8M NSF, $1M Schmidt Foundation
  - *Build a 6 node testbed and demonstrate feasibility*
  - *Establish wide community support, through the Big Data Hubs*

Rapidly establish the third main pillar for NSF science infrastructure

**COMPUTE** **NETWORKING** **DATA**

# The Future

- Long term data…
- Too much data…
- Machine learning, AI…

# How Do We Prioritize?

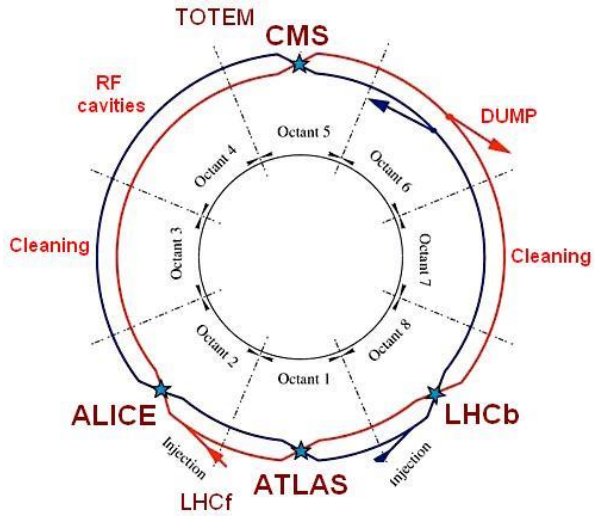
- It is becoming “too easy” to collect even more data
- Robotic telescopes, next generation sequencers, complex HPC simulations
- **How long can this go on?**

*“Do you have enough data or would you like to have more?”*

- No scientist ever wanted less data….
- But: Big Data is synonymous with Dirty Data
- How can we collect data that is ***more relevant*** ?
- We need to improve ideas on experiment design….

# LHC Lesson

- 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

# Long Term Lifecycles

- There is a **Data Lifecycle**:
  - *New data standards emerge*
  - *Metadata standards change*
  - *Usage patterns change*

- There is also a **Service Lifecycle**
  - *Survey data presented in Smart Services*
  - *Browsers change (HTML5)*
  - *OS change, DB change*
  - *Servers become obsolete, disks die*
  - *New software technologies (iPython emerging)*

# Failed Disks Over 18 Years


1885 drives

~2.5%/year
over 18 yrs

1.1 tons of Hard disks

# The Challenges for Long Term Data Sets

- How can we build a shared, common data ecosystem?
- FAIR (Findable, Accessible, Interoperable, Reproducible)
  - *Need more automation, manual approach cannot keep up*
- What happens to large, high-value data sets on the long term?
- Open/Free, Accessible and Self-sustaining?
  - ***Pick any two, and the third is determined!***
  - How can one ensure a steady, long-term support?
  - Who do we trust with all this irreplaceable data?
  - How can we decide what to preserve?
- Building TRUST is hard, losing it is easy

# What is the Value of Data?

- What is the value of (usable/accessible) survey data?
  - *Accelerates testing ideas, find targets for followup*
  - *Provides a platform for reproducible data*
  - *Not possible without a robust access*
  - ***Produces direct foundation to many papers***
- How much science funding does a data set attract?
  - *Typical NSF AST grant is $300K over 3 years*
  - *One refereed paper/year is good performance*
  - *Implies: $100K/paper is a reasonable metric*
    - *not $10K and not $1M*

# What is the Price of Data?

- How much did it take to produce?
  - *SDSS I-IV was probably around $200M total*
  - *Typical MREFC NSF project ~$1B*

- Price/Value/Output
  - *SDSS enabled 9,000 refereed papers so far:*
    *9K x 100K = $900M -> cost effective*

# What is the Cost of Data?

- **The cost is in long-term preservation**
- So far so good, data maintained by live projects
- This is about to change: all experiments will come to an end
- What happens to the data?
  - *Who will take ownership of it?*
  - *Who will remember what it is?*
  - *Who will pay for its maintenance?*
- Estimated cost for SDSS is ~$500K/year
  - *Either 5 papers, or 0.25% of the **price of data***
  - ***5% cost overrun in the survey would cover the data for 20 additional years***

# What Happens over a 30 Year period?

- Today we spend ~$B to acquire valuable data
- Much of these will not be superseded in the foreseeable future/decades
- At the end of projects the data sets will be handed off to someone
- We need an organization(s) that is
  - *With a long track record with a predictable future*
  - *Understands data preservation*
  - *Trusted by everyone*
  - *Technically capable*
  - *Can run under a sustainable model*
  - *Has no single points of failure*
- We need a long-term stable funding model for these high value data **=> Smithsonian model (private endowment + government)**

# The Future (in 8-10 years)

- Everything is commoditized, everything is accelerating
  (except universities)
- Everything is in the cloud…
  - *partly due to economies of scale*
  - *partly due to accelerated deployment*
- What will be the role of machine learning in science once the hype settles?
  - *Need to use physical insight, symmetries, simplify network design*
  - *Need explainable inference!*
  - *Discover sparse representations*
  - *Many experiments will be driven by AI*

# Summary

- Open Big Data is at the forefront of all science
- Convergence of physical and life sciences
- Major disruptive technological changes keep happening
- We made much of the data strategy up as it happened, by adopting to changes as they arose
- Funding much slower and mainly reactive
- We are spending billions on big experiments, yet no plan for long term data, we have to improvise again…

Questions for us scientists:

- How can we make our adaptation more agile?
- How can we steer towards long-term sustainability?
- Scientists will need a trusted intermediary to deal with the data
- We need to keep reinventing ourselves over and over…

"Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!"

— *Lewis Carroll,*
*Alice Through the Looking Glass (1865)*