

## Scalable Data Aggregation for Science

**Alex Szalay**  
**The Johns Hopkins University**

# The Era of Surveys

- We are moving from the era of “manufacture” of scientific data in small-scale experiments to the “industrial revolution” of Big Science
- Particle physics analogy:

Van der Graaf -> Cyclotron -> Synchrotron -> National Labs

SSC ☹️

LHC 😊

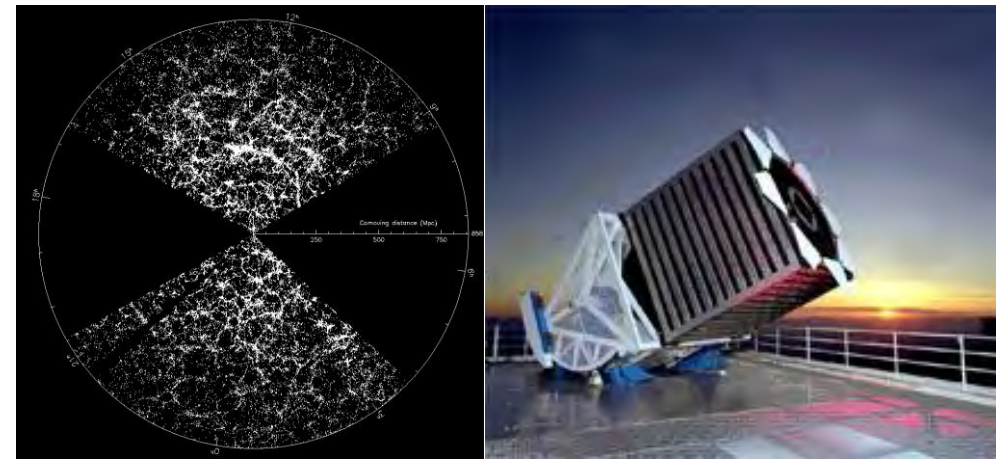
Big Science projects are entering the multi-billion regime, their open data archives analyzed by large communities

## This is a big difference

- Past: Experiments rapidly followed one another, data sets had a short lifetime
- Today: 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 entirely 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
- Now SDSS-IV data served from JHU
- SDSS-V starting soon

## Prototype in 21st Century data access

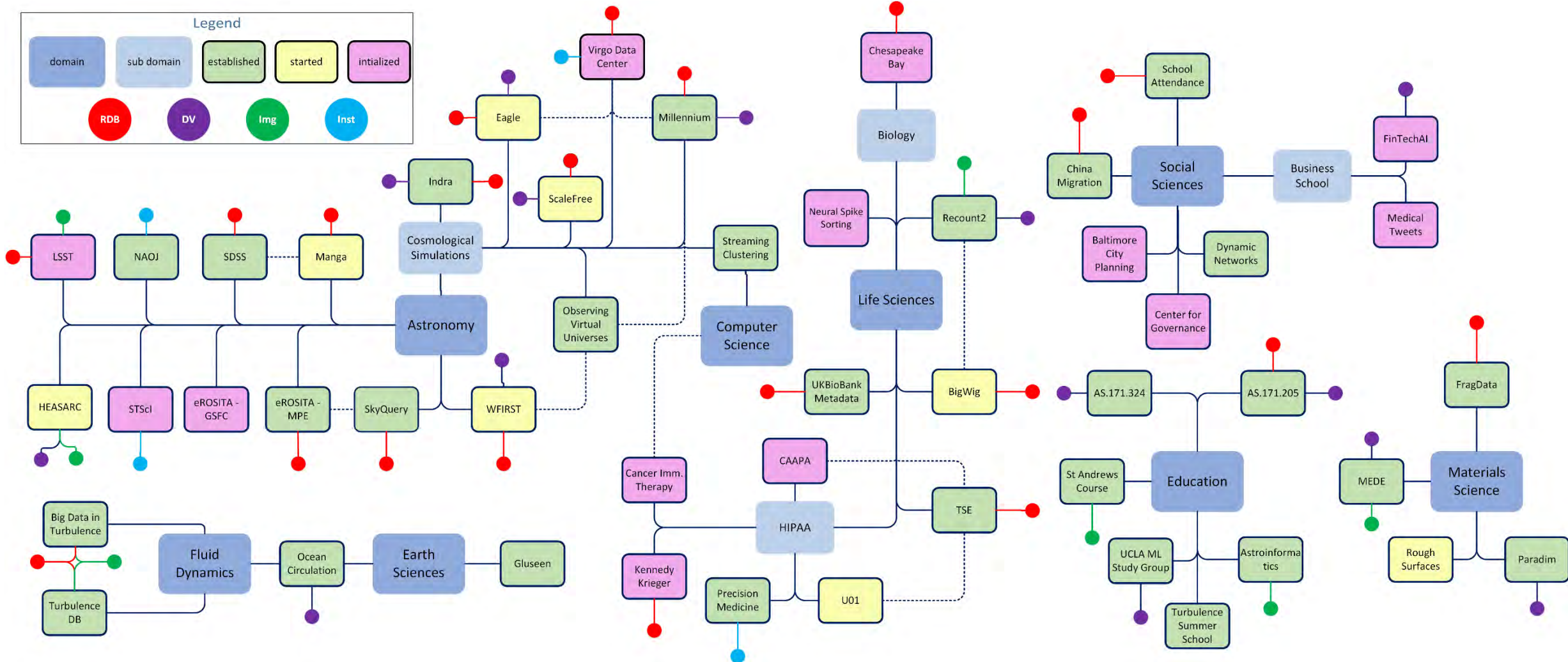
- 3.5B web hits in 18 years
- 480M external SQL queries
- 8,000+ papers and 400K+ citations
- 7,000,000 distinct users vs. 15,000 astronomers
- The emergence of the “Internet Scientist”
- The world’s most used astronomy facility today
- Collaborative server-side analysis done by 9K astronomers

# 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, just ended*



# Current SciServer Projects



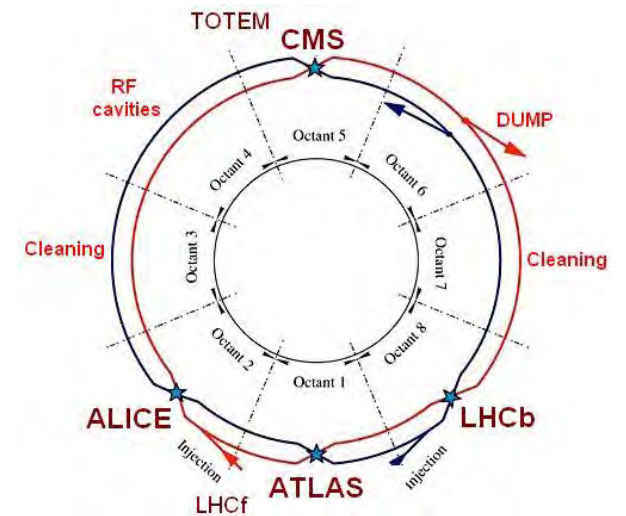
# SciServer: Scalable Data Aggregator

- 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 exponentially hard ( $n!$  connections)
- Real life uses require **interactive exploration** before big analysis with **workflows**
- The SciServer philosophy:
  - Read-only Data Contexts, separate data model and ontology, self documenting
  - User owned databases and resources to create value added aggregations, shared at will with others, at owners discretion
  - Can aggregate new datasets in isolation very quickly
  - Detailed user logs, dependencies, potentially turned into scripts automatically (MST)
- iPython+R scripting, running on pool of VMs + ML packages
- Additional API for resource sharing, like streaming data ingest (PARADIM)
- Single sign-on to all resources
- Currently about 3.5PB of live data, growing

# Prioritizing for Relevant Data

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

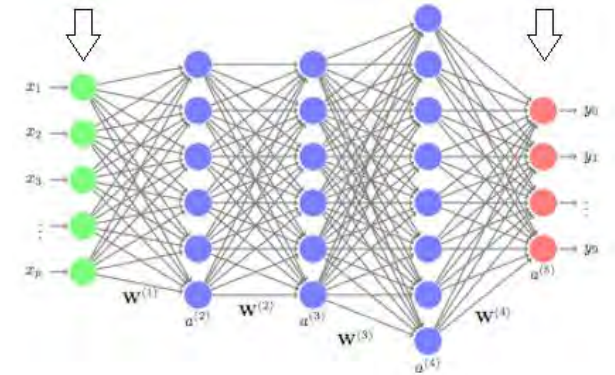
- Delicate tradeoff between the scientific value and the cost of preservation
  - One extreme – store everything, go bankrupt!
  - Other extreme – too little data!
- LHC lesson: single data source, \$\$\$\$\$
  - Multiple experiments tap into the beamlines
- **In-situ** hardware triggers to filter data
  - Only 1 in 10M events are saved (9999999:1)
- Resulting “small subset” analyzed many times **off-line**
  - This is still 10-100 PB
  - Keeps the whole community busy for a decade or more



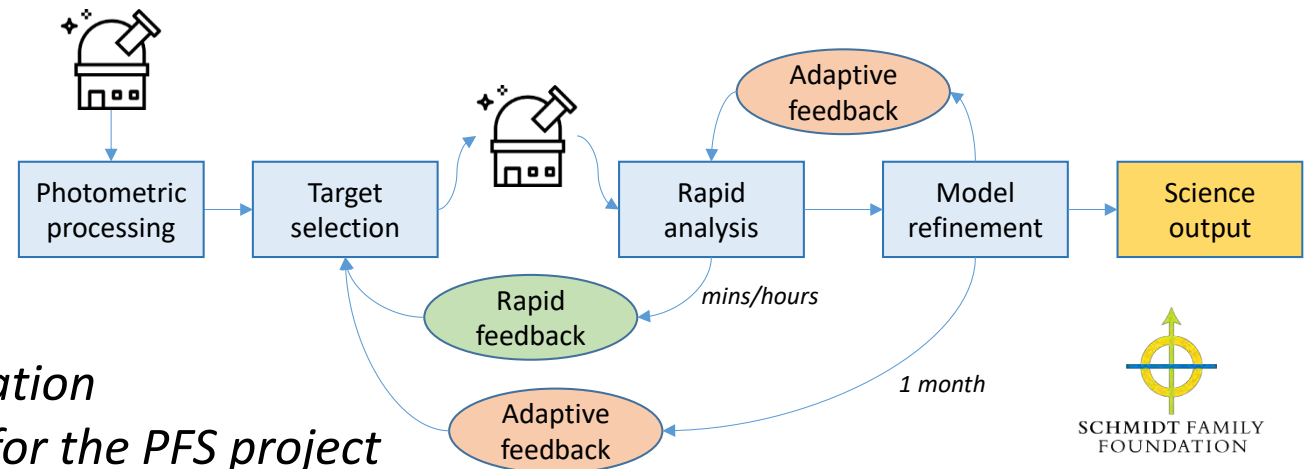
*Follow 90-10 rule: 90% of the science value is carried by 10% of the data (or 999-1?)*

# Experimental Design by AI

- Instead of more data we need MORE RELEVANT DATA
- Need to dramatically improve experiment design....
- Artificial Intelligence in large-scale experiments:  
    use AI before we collect the data
- Example: Next Generation Astronomical Surveys
  - Observing spectra is 1,000 times more expensive than imaging
  - Use feedback from observed targets and improve target selection via reinforcement learning



- Put the telescope in the reinforcement loop!



*Supported by the Schmidt Family Foundation  
at JHU and Princeton: Use AI Feedback for the PFS project*

# Scalable Data Access / Analysis

- Traditionally we “threw our data over the wall”
  - Scientific publishers caught it and handled it (for a fee...)
- This no longer works
  - Data is too much for the journal publishers
  - Active curated services needed
  - Nobody is willing to sign data rights over to commercial entities
- Traditional publishing models are collapsing as well
- The evolution of the music industry:

LP/CD      => iTunes      => Spotify/Pandora

- What are the data equivalents?

Download all data

=> Queries to project servers

=> Run in the cloud, stream result



# Cloud vs Local?



- Everything computational is increasingly commoditized, accelerating exponentially (except universities – this is good!!)
- Everything will be in the cloud...
  - Due to economies of scale
  - Due to accelerated pace of deployment
- What is the role Data  $\Leftrightarrow$  Machine Learning in science (once the hype settles)?
  - Much of the Deep Learning software will be commodities
  - Experiments driven by AI
  - Users want “AI-ready” data sets
  - Science needs explainable inference!
  - Scientists can add physical insight, symmetries
  - Discover sparse representations, build faster proxy simulations

# Long Term Access to Data

- FAIR (Findable, Accessible, Interoperable, Reproducible)
  - Total failure on capturing the many small data sets (“long-tail”)
  - Need more automation, manual approach cannot keep up
- How about **Free and Sustainable**?
  - Increasing expectations for open/free access to scientific data
  - What happens to large, high-value data sets when they are completed?
- Free, Accessible and Sustainable?
  - **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?
- We need a new trusted intermediary
- *Where is the “Smithsonian for Digital Data?”*



“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)*