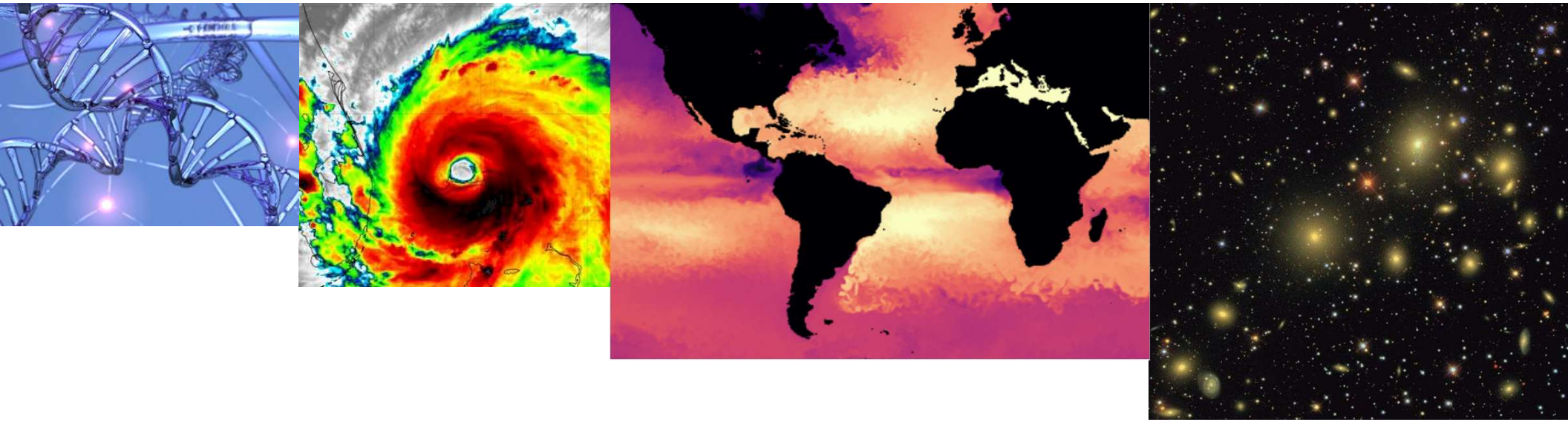
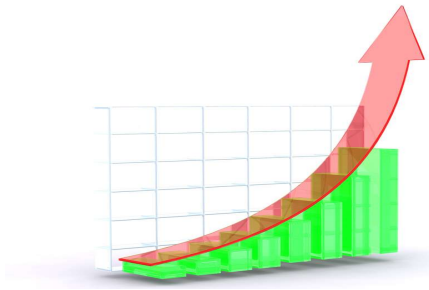


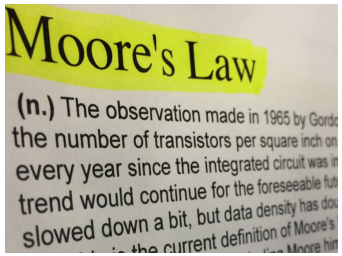
Aggregating and Harmonizing Data

Alex Szalay
The Johns Hopkins University

Introduction

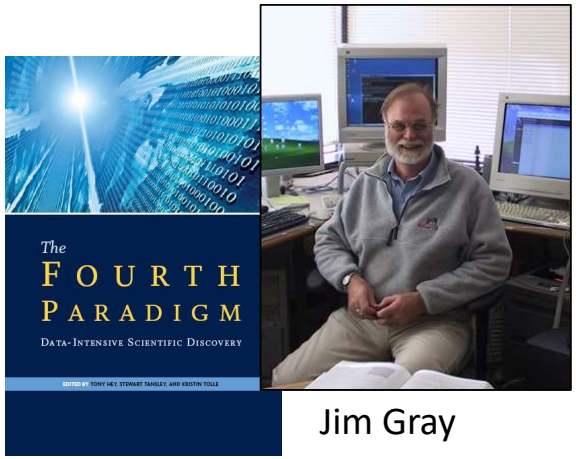
Open Data is bringing a new revolution in science, transforming everything => Open Science





Enabled by the exponential growth in our computational technologies

The Fourth Paradigm of Science emerged, driven by Open Data



Jim Gray

Agenda

The Exponential Evolution of Science

The Changing Granularity of Science

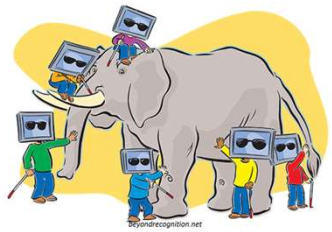
Aggregating Data

AstroPath: From Stars to Cells

ASTROPATH

Future Proofing Open Data

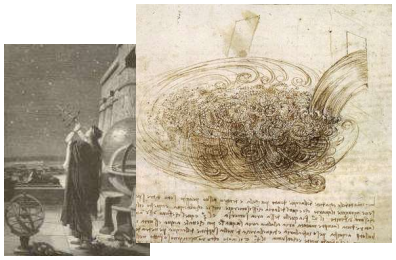
The Challenges Are Not Technical



The Exponential Evolution of Science

Science is Changing Exponentially

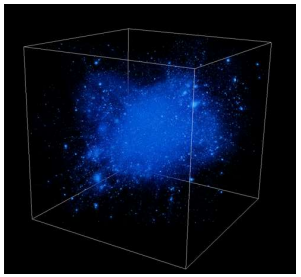
THOUSAND YEARS AGO
science was empirical
describing natural phenomena



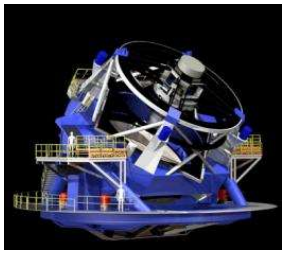
LAST FEW HUNDRED YEARS
theoretical branch using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

LAST FEW DECADES
a computational branch simulating complex phenomena

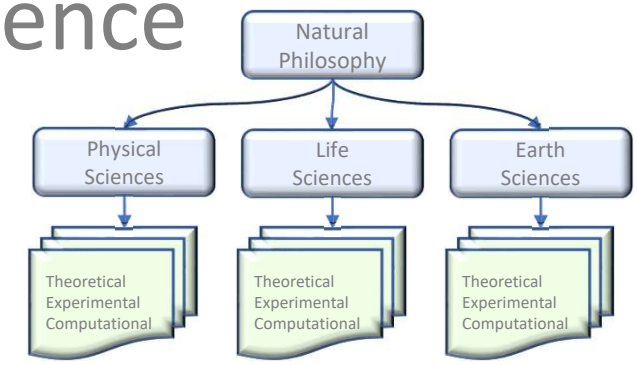


TODAY
data intensive science, synthesizing theory, experiment and computation with statistics
►new way of thinking required!



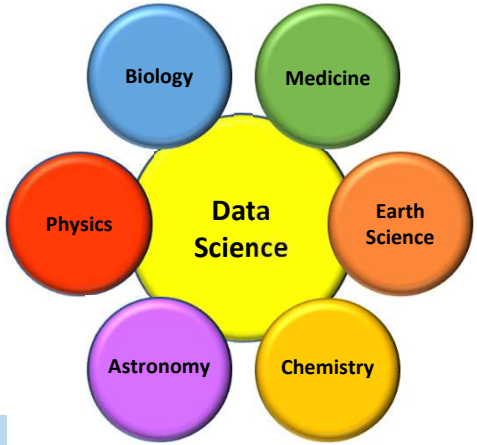
Science: From Fractal to Convergence

Historically science was fragmenting into narrower and narrower sub-disciplines



Today we see a CONVERGENCE!

All Physical and Life Science domains share common data science methods and approaches



Data Science is becoming the "New Math", the shared language of science!

Tomorrow's Scientists are Multi-Disciplinary

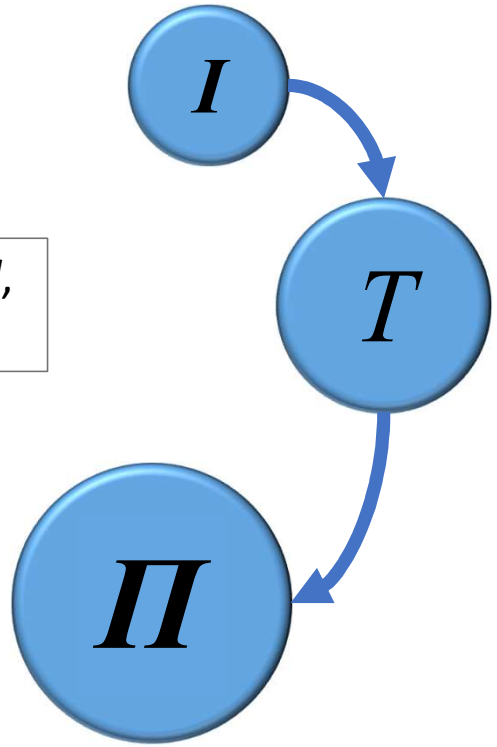
Our higher education is training deep but narrow people, ***I-shaped***

As we get older, we become ***T-shaped***, with a shallow but broad layer on top

New disciplines emerge when two domains intersect

=> Watson and Crick (physicist+ornithologist) => genomics

Scientists need to become ***Π-shaped***, grow a deep leg in data science as well



We need to train Π-shaped people …

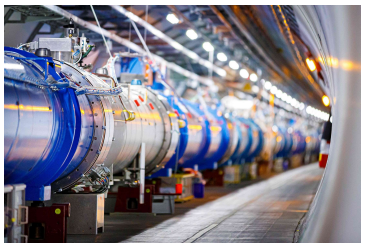
The Changing Granularity of Science

The Emergence of Big Science

- From "***manual production***" of scientific data to the "***industrial revolution***"
- 1920-50 : Small experiments by few individuals, slowly growing
- 1960-: Big Science, costing $1B+, take decades, very risk-adverse, thousands of people

This is a big difference

- Past: Experiments rapidly followed one another, data sets had a short life
- Today: Big Science experiments (LIGO, LHC, SKA, LSST, OOI, NEON,…) may not be surpassed by another variant in our lifetime

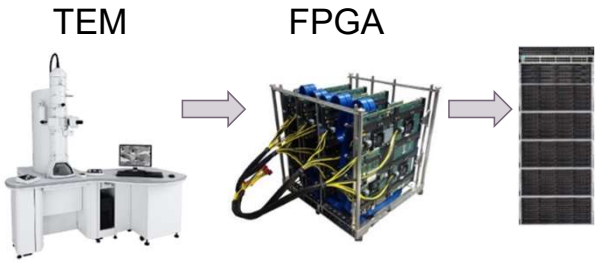


Van der Graaf -> Cyclotron -> Synchrotron -> National Labs

- SSC ☹
- LHC ☺

The data is here to stay for decades…

Today's Science is Mid-Scale



- The optimum scale of science is changing today
  - – more in the ***middle***
    - NSF MSRI, NIH U01, public-private partnerships
      => Sky Surveys Human Genome … $10-100M
- Create a unique instrument (microscope, telescope,…)
  - Use cutting edge technology, take risks, push budgets to the limit, maximize science, generate petabytes of data
  - ***Agility*** – important because of the exponential technology growth
  - ***Highly automated, robotic experiments*** – the next step in scientific data acquisition

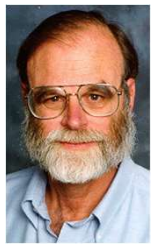
Enormous fresh creative energy liberated, the "sweet spot" for science!

Even smaller groups can generate petabytes of open data using advanced technology!

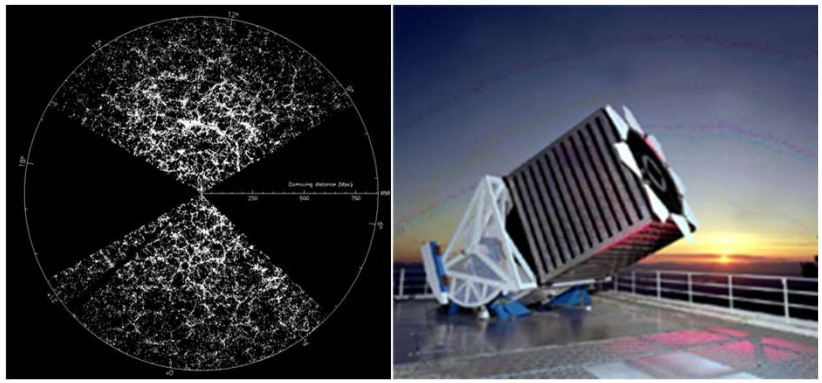
Mid-Scale Example: Sloan Digital Sky Survey

"The Cosmic Genome Project"

- Started in 1992, SDSS-II finished in 2008
- Data is entirely ***public, open and free***
- Database built at JHU
- Project marked a transition in astronomy
  - From manufacturing to mass production



Jim Gray



SkyServer: Prototype in 21st Century data access

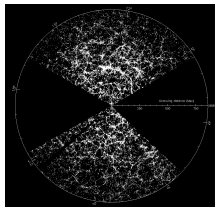
- Visual interface integrated with object-relational DB
- Remarkably fast adaptation by the community
- 10M distinct users vs. 15,000 astronomers
- The emergence of the *"Internet Scientist"*
- Collaborative server-side analysis

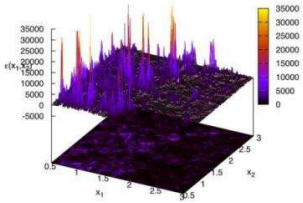
Scientists are becoming publishers and curators of large data!

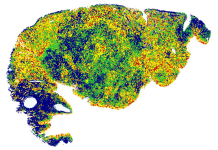
Mid-Scale Science => “Game Changing” Data

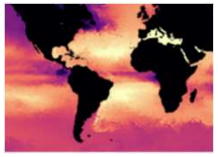
Leapfrog – “non-incremental” – but still Mid-Scale Science – Similarities

- (2001-) **Sloan Digital Sky Survey (SDSS)** – grew data by a **factor of 100**,
  still the world’s most used astronomy facility,
  4.6B web hits, 713M SQL queries, 10M users, 10K papers, 500K citations
- (2006-) **Turbulence database (JHTDB)** the **world's largest simulations,**
  the "virtual observatory" of turbulence,
  1.5PB of data, 200 trillion points delivered to the world
- (2016-) **AstroPath (JHMI)** – **1000-fold increase** in data for cancer immunotherapy,
  astronomy => pathology, soon Open Cancer Cell Atlas with 1B+ cells
  16T pixels, 500M cells
- (2017-) **POSEIDON (JHU/MIT/Columbia)** building **the world's largest** ocean circulation model, 10x higher resolution, open petascale interactive laboratory
  2.5PB of data on its way









Using similarities to the SDSS, we are able to create unique leapfrog projects over and over

Aggregating Data

Astronomy Adopted to Open Data Very Fast

Astronomy was always data-driven….

We cannot experiment with stars in the lab…

*"Very exciting to work with astronomers, since their data is '**worthless**'!"*

— Jim Gray to Bill Gates (~2003)

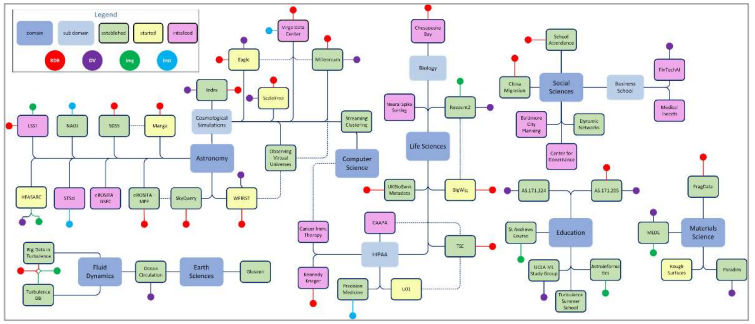
Scientific data must not only be OPEN and FAIR, but FREE and SUSTAINABLE!

IDIES: Open Science with Interactive Petabytes

Institute for Data Intensive Engineering and Science

idies

- Provide “disruptive assistance” -- from “patterns to processes”
- Institutionalize “lessons learned” in a multidisciplinary setting
  - Science engagements have distinctive “phases of maturity”
  - Critical mass of interdisciplinary postdocs and software engineers
- Convergent, multidisciplinary engagements (70+ ongoing projects)
  - Hosted on the SciServer – collaborative platform for petabytes of data
  - Collaborations with national labs, federal agencies (NASA, NIST, DOE), Max Planck, Japan, RAL
- Broad innovative educational and outreach program
- Leverage our scalable open infrastructure
  - Currently 30PB+, 200 servers
  - 10M casual users, 10K+ power users
  - Mostly built with previous large NSF investments
  - Operating at very good economies of scale



JHU has a unique expertise to do this (the SciServer platform)

SciServer: Scalable Data Aggregator

- The main challenge in creating big data sets is **DATA AGGREGATION**
- Difficult to aggregate large data sets, yet the joint analysis requires co-location
- Most frequent mistake: trying to create the “mother of all databases”
  - Building integrated ontologies/data models is hard, becomes combinatorically complex
- Real life uses require interactive exploration before big analysis
- **Our goal is to enable interactive, collaborative use of Petabyte-scale data**
- The JHU SciServer philosophy is “keeping in simple”
  - Store all the data together for the best economies of scale as distinct Data Contexts
  - Users have their own databases and resources to create value added aggregations (links)
  - These can be shared at will with collaborators **at owners’ discretion**
- We can add new datasets/modalities in isolation very quickly => linear complexity

The SciServer is uniquely capable of managing many Petabytes of data, and supporting data-intense collaborations

Lessons Learned

- Statistical analyses and collaboration **easier with DB** than flat files
- **Collaborative features** essential
- Need to **go beyond SQL** scripting => Jupyter and Deep Learning
- Everything is **spatial**
- **Multiple access patterns** (visualization, interactive and batch analyses)
- **Automation** is needed for statistical reproducibility at scale
- **Scaling out** was much harder than we ever thought
- Always need **deep links** to the raw files
- 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**

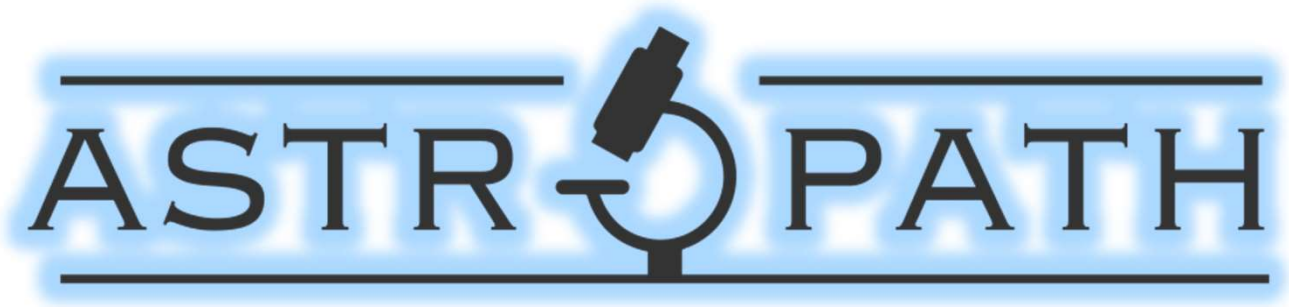
Find the right tradeoffs -- do not try to do “everything for everybody”

Turning Lessons into Practice

- We saw many **repeating patterns** for data intensive projects in different domains
- Now we are trying to turn these into **processes** that can be replicated
- Need to strike the right **tradeoff** of protecting the valuable data, while allowing creative (**disruptive**) innovations
- Invest more in critical **sustainable human infrastructure**
  - Key roles: *Architects, Implementers, Disruptors, Trainers*
- Create a sustainable model for massive data and compute infrastructure
  - The right balance between local and cloud resources
  - Active support in the creation of novel data resources/databases
- **Innovate in data collection**
  - Use AI in optimizing next-generation quasi-autonomous experiments
- Build sustainable funds for preserving high-value data
  - Cost is <0.25%/year of price of the experiment

All of these require ongoing commitments, not one-shot investments

AstroPath: From Stars to Cells

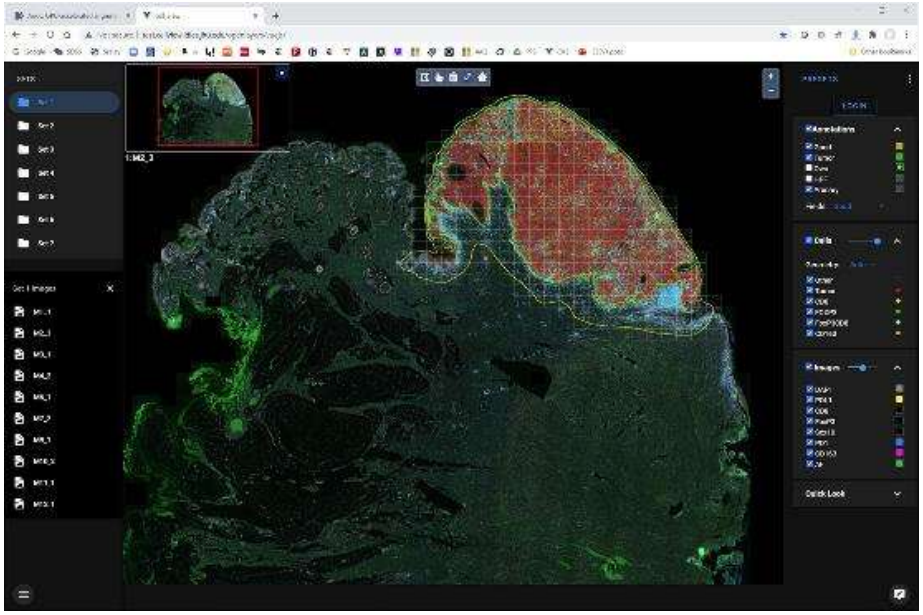


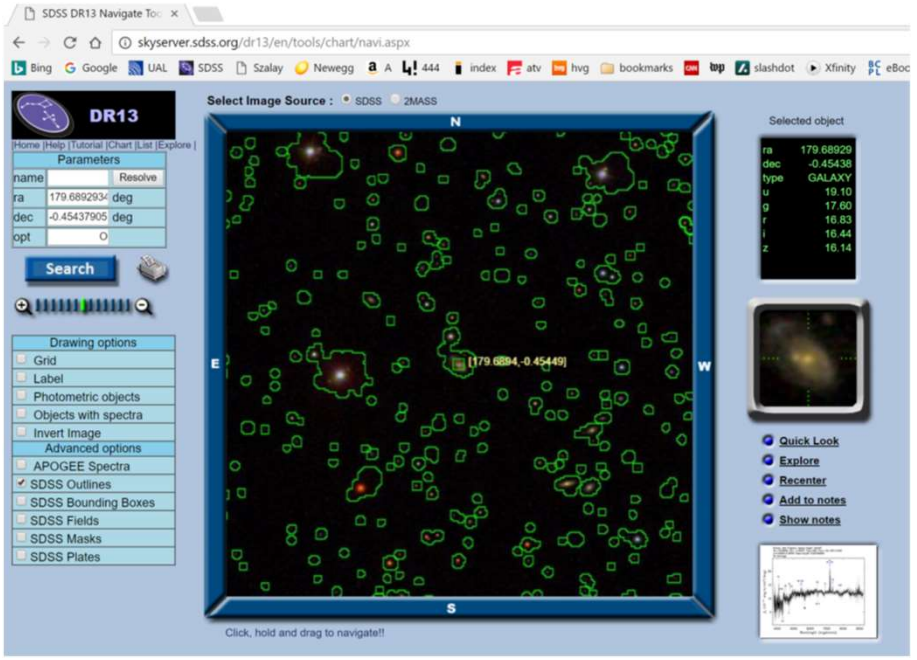
From Stars to Cells

- Strong parallels between medicine today and astronomy 25 years ago
  - "Disruptive assistance" from astronomy to pathology
  - Using techniques astronomers learned the hard way (flat field, unwarp, calibrate)
- Stars and galaxies are like the cells in pathology
  - Multicolor photometry, image segmentation, locality
  - Spatial relations, spatial searches, outlines
- Transitioning to the "industrial revolution"
- We had to rebuild the whole workflow and data acquisition protocol

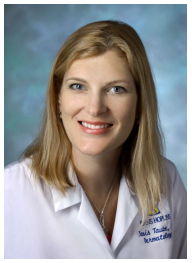
This requires

- Automated approaches to cell segmentation, image analysis, and data management
- Focus on scaling out the workflow to parallel execution





AstroPath: Atlas of Cancer Cells

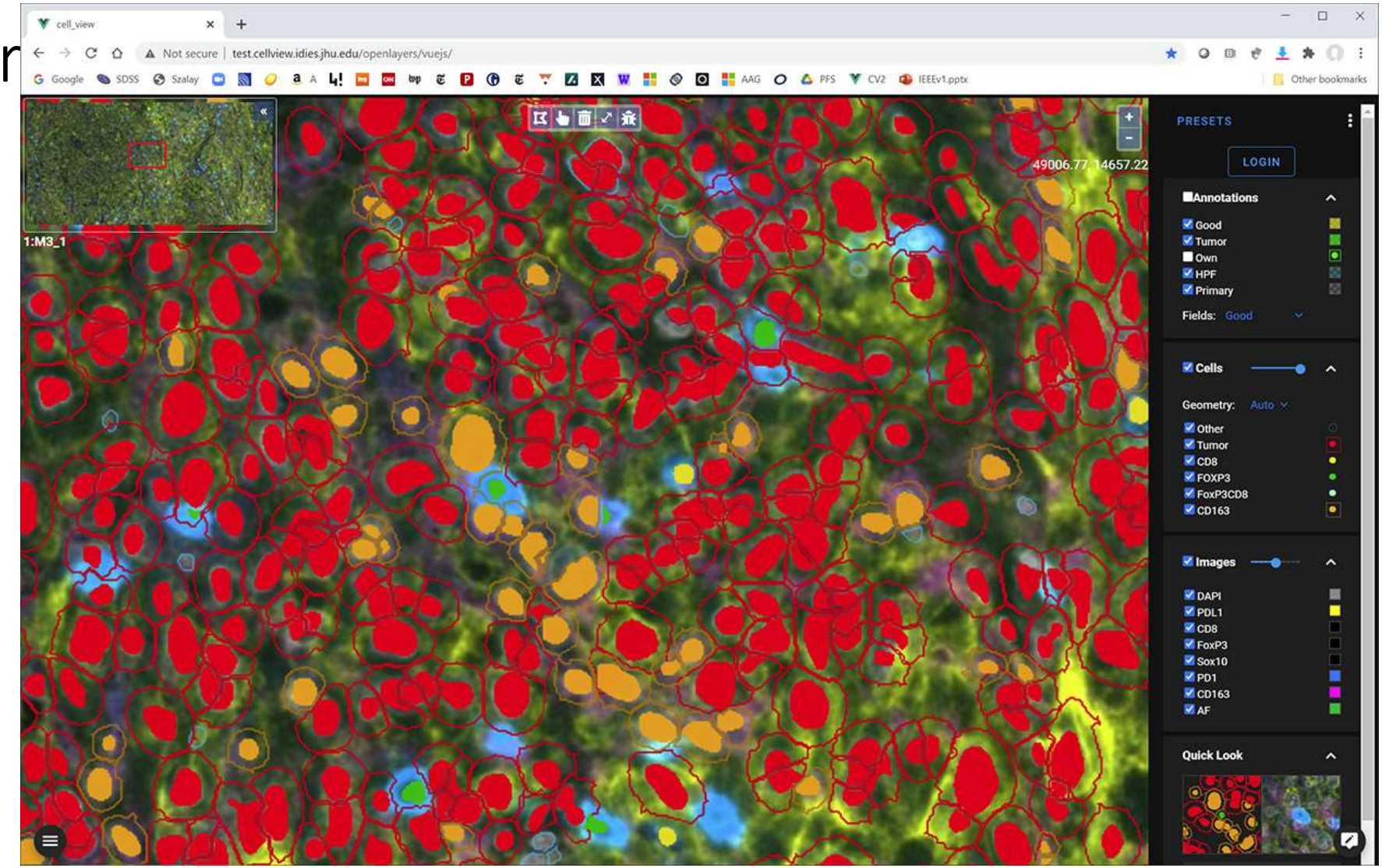


- **Astro**nomy meets **Path**ology
  - with Prof. Janis Taube (JHMI BKI)
- Studying the tumor microenvironment to understand cancer immunotherapy
  - Spatial interactions of activated T cells and tumor near the tumor boundaries
- Goal: increase data collection by a factor of >1,000
  - 400GB mosaic of 35-band multiplex images/slide (from 10 to 2000 images/slide)
  - 7 markers (lineage + PD-1, PD-L1), more markers via additional panels
  - Use a farm of automated microscopes => 2PB/year
  - Heavy use of parallel processing
- Tumor boundaries, cell geometries represented as GIS polygons
- Dynamic computation of nearest neighbors, spatial relations
- Interactive viewer like the SkyServer, or Google Maps
- Processing workflows mostly automated
- Working on validating a large enough training set for Deep Learning
- Databases linked to SciServer, collaborative Jupyter, Keros/TensorFlow, R
- Collaboration with Akoya BioSciences (microscopes)

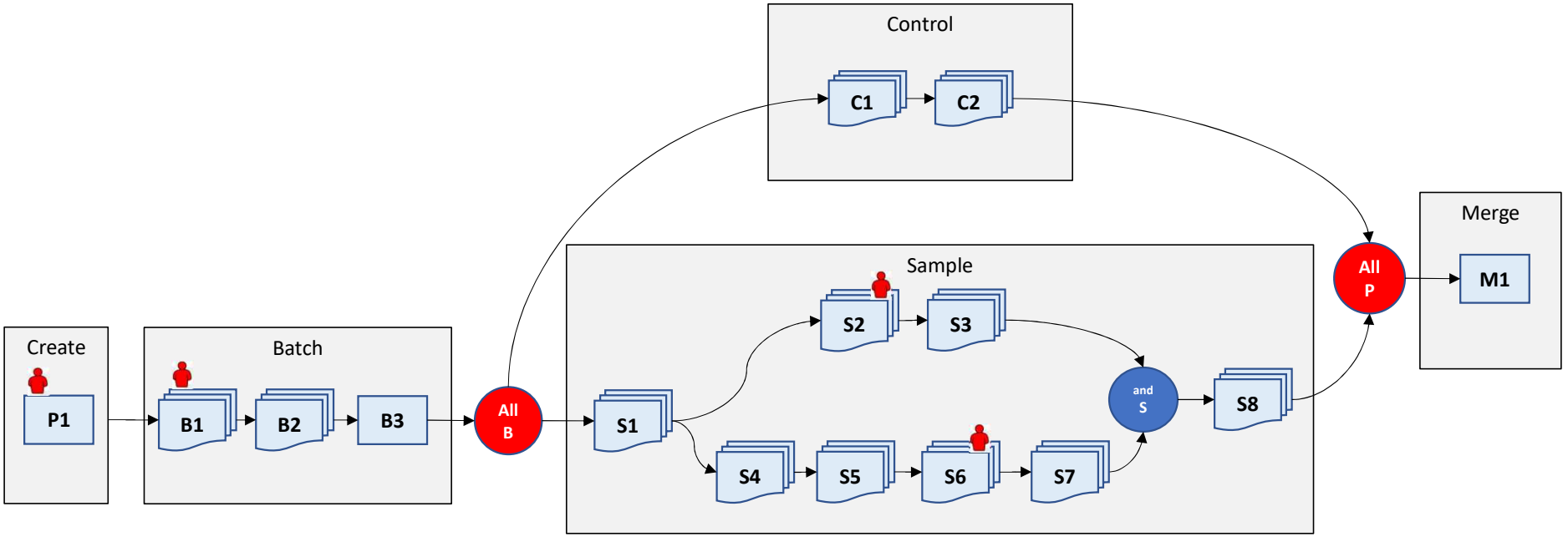
Current data in the database

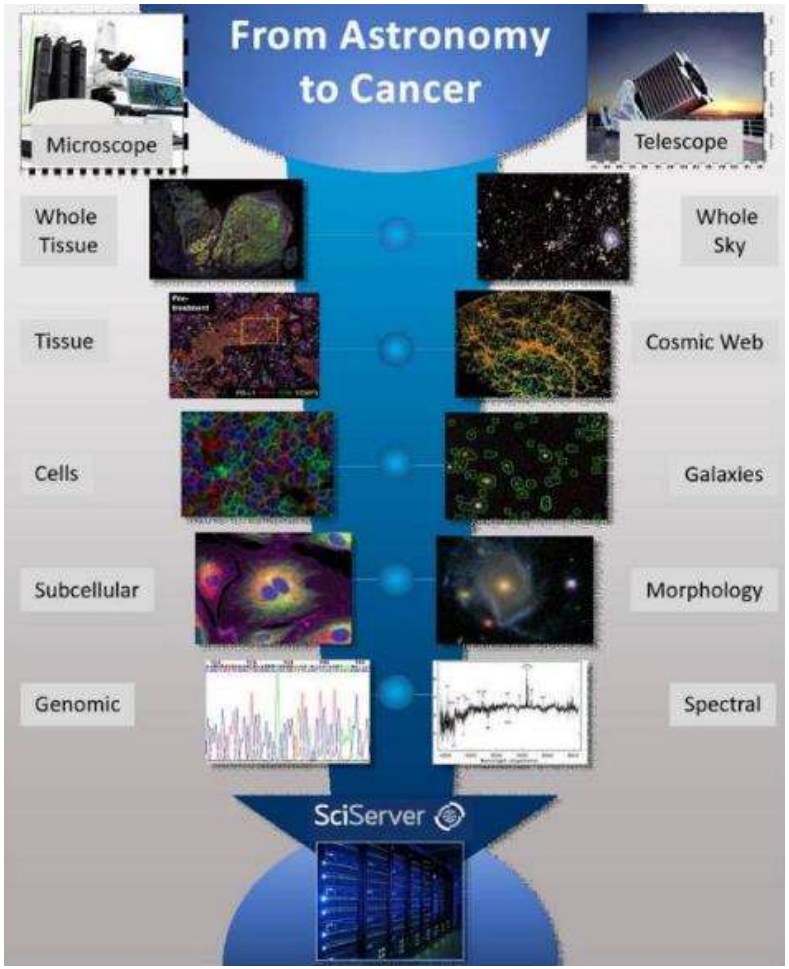
- 8 Cohorts, 692 slides
- 365,023 High Powered Fields
- 564M detected cells
- 257M unique cells
- 10B cell pairs
- 22 trillion pixels (whole SDSS was 5 Tpixels!)
- Additional 100+ slides already scanned with multiple tumor types

Inter


cell_view
Not secure
test.cellview.idies.jhu.edu/openlayers/vuejs/
Google
SDSS
Szalay
AAG
PFS
CV2
IEEEv1.pptx
Other bookmarks
1:M3_1
49006.77, 14657.22
PRESETS
LOGIN
Annotations
Good
Tumor
Own
HPF
Primary
Fields: Good
Cells
Geometry: Auto
Other
Tumor
CD8
FOXP3
FoxP3CD8
CD163
Images
DAPI
PDL1
CD8
FoxP3
Sox10
PD1
CD163
AF
Quick Look

Full Astropath Workflow





ASTROPATH

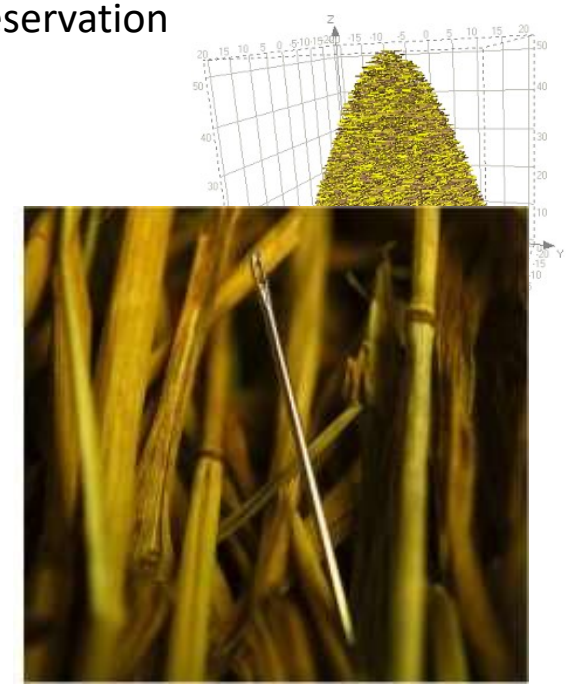
- Applies techniques from astronomy image acquisition and analysis of stars and galaxies of large fields of cells in a pathologic specimen.
- Developed a unique, scalable facility to produce petabytes of robust tissue imaging data on par with large sky surveys.
- Found a predictive biomarker for immuno- therapy using AstroPath, which was validated in a separate patient cohort (Science, June 11, 2021)
- Both a discovery tool and a path forward towards a new standard of diagnostics in pathology.

Future Proofing Open Data

Prioritizing for Relevance

"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 – collect too little data, not enough for the science!



- **LHC lesson**

- I**n-situ** hardware filters data, optimizing for "new science"
  - *Only 1 in 10M events saved (9999999:1)*
- Resulting "small subset" is still 10-100 PB

Tradeoffs are essential: cannot do everything for everybody (9-1, 99-1 or 999-1?)

Collect More RELEVANT Data!

- Need to dramatically improve our expe
- Artificial Intelligence in large-scale expe
  use AI **before** and **while** we collec
- It is already happening at CERN, materi
- Maybe this will be the **Fifth Paradigm**, algorithms control our experiments
  => also make intelligent, real-time decisions

If an AI algorithm can drive our cars, why cannot it run our microscopes?

Agility vs Tenacity – How can We Compete?

- Extremely **agile** changes in the industry (particularly in AI)
  - Google, Facebook, Amazon, Microsoft
- Universities cannot compete with the industry in agility
  - Faculty hires are for 40 years…
- **But we can compete in tenacity and high-value data!**

- More mid-scale projects emerging at Universities

  => generating petabytes
- Innovative uses of AI will optimize experiments and discover new patterns
- This requires the data sets to be "AI-ready"

The breakthroughs will always come from a unique data set (Human Genome,SDSS, ImageNet) – combined with a disruptive idea

The Evolving Data Analysis

The evolution of the music industry is a good example:

LP/CD 

=> iTunes

=> Spotify/Pandora

What are the data equivalents?

Download all data — ***Send tapes, disk, sneakernet***

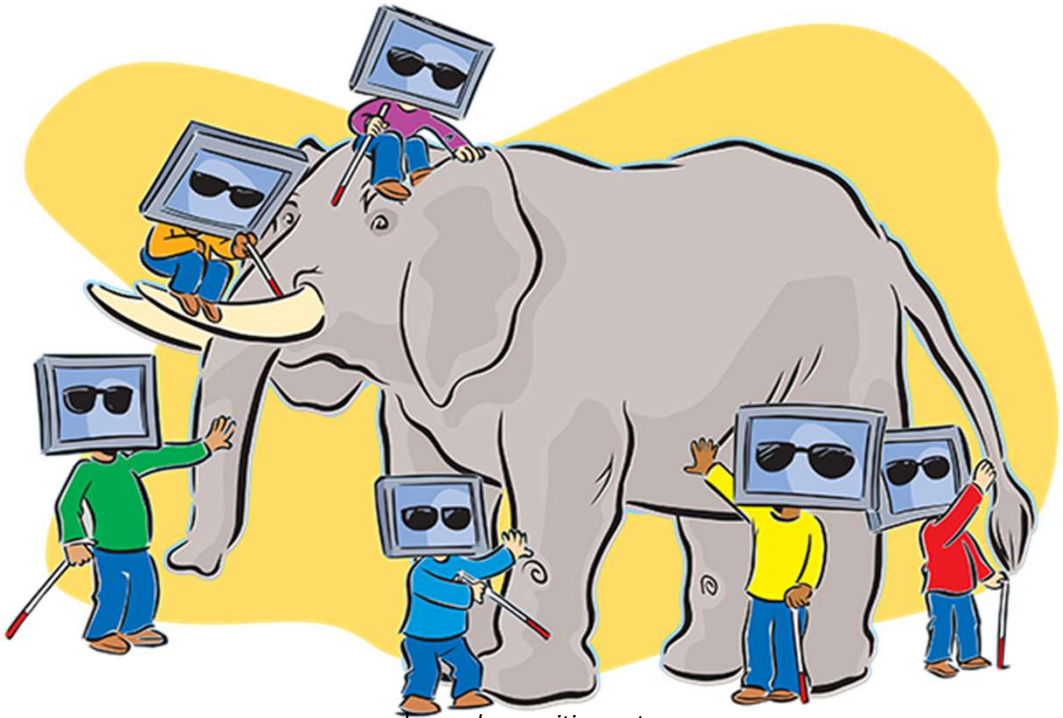
=> Run queries at project servers — ***Astronomy archives, SkyServer, IVOA, MAST, NED,…***

=> Run in the cloud, view the result — ***Google Colab, SciServer***

Scientific software needs to be Analysis Ready and Cloud Optimized (ARCO)

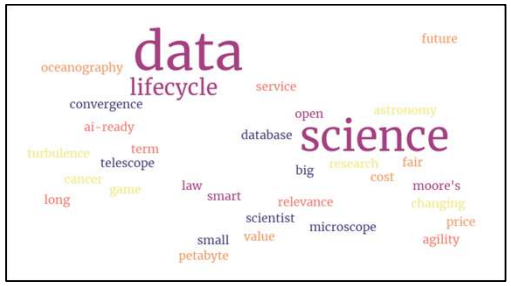
Ryan Abernathey (Columbia)

The Challenges Are Not Technical

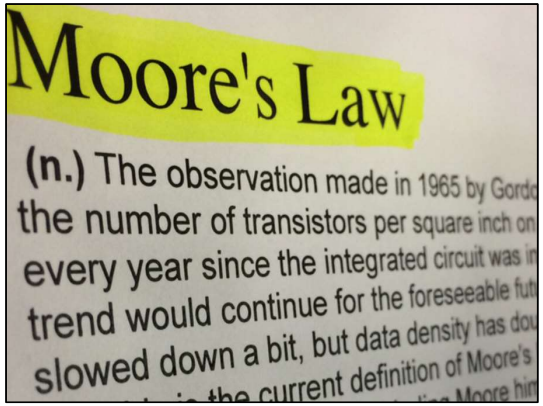


beyondrecognition.net

Data Lifecycle => Service Lifecycle



- The value of our national investments in science is the **DATA!**
- The high-value open data sets will live for decades
- Results in much more data reuse



- There is also a **Service Lifecycle**
- The data is becoming smarter
- Smart platforms need to be maintained for decades

Smart data platforms are constantly evolving – following the technology

The Economics of Long-Term Data

- $100B+ investments => Today's Open Science data
  - National Treasure **=> must be preserved**

- Conflict: Short term federal funding cycle
  vs long term data preservation
- Different federal agencies have different strategies
  NASA Data Centers, NIH Data Commons, NSF MREFC,
  DOE National Labs, NOAA, NCAR, EPA…
  Coherence/convergence is yet to emerge..

- The Smithsonian is hosting physical specimen from historical scientific discoveries => private-public partnership

Where is the Smithsonian of Data?

The Challenges are Non-Technical

The Four Paradigms of Science

- Empirical → Theoretical → Computational → Data Driven

Organization of science is changing

- Granularity of science (small → bimodal → mid-scale)
- Data sharing & long-lived data → Accelerating the change

Changing relationship between scientists and data

- Data only in papers → Now big datasets → Curation responsibility
- A trusted data intermediary → Empowers sharing and reuse

OPEN DATA → legal framework for sharing

- FAIR Data (Findable, Accessible, Interoperable, & Reproducible)
- Science needs Free & Sustainable data

The big dilemma is funding and coordination

- Objective metric for the "usefulness of data" → Value, price, cost
- Trusted agent?
- Need National strategy for preserving OPEN DATA