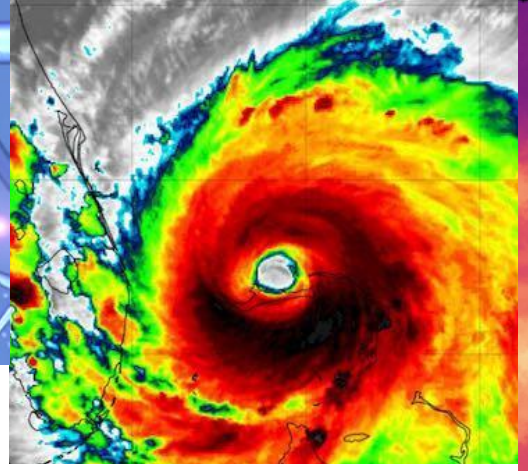
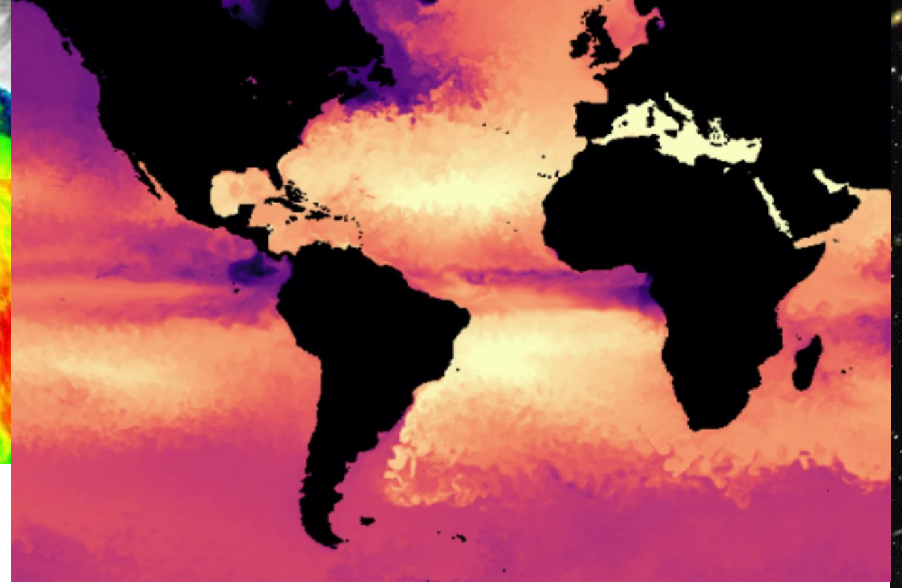




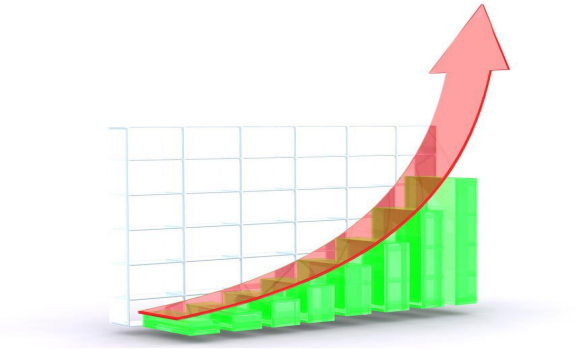

AI-ready data in Astrophysics (from Sensors to Tensors…)

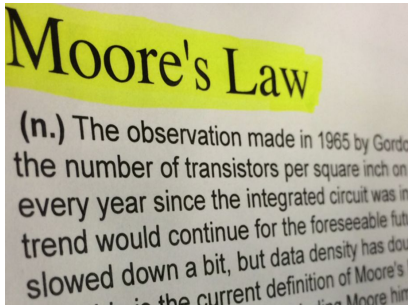
Alex Szalay

The Johns Hopkins University

Introduction

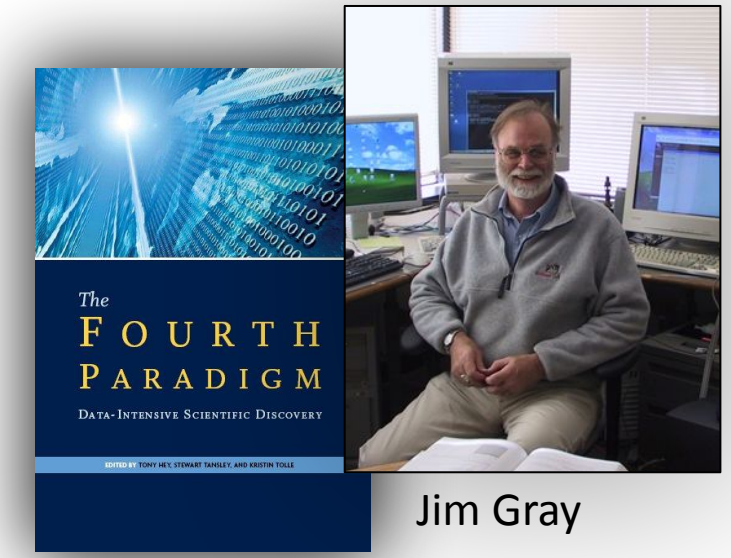
Open Data is bringing a new revolution in science, transforming everything
=> Open Science





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

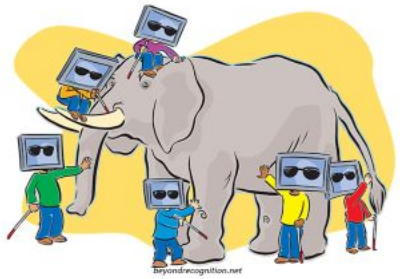
The Changing Granularity of Science

Measuring the Usefulness of Data

The Emergence of AI

More Data Becomes Too Much Data

The Challenges Are Not Technical



The Exponential Evolution of Science

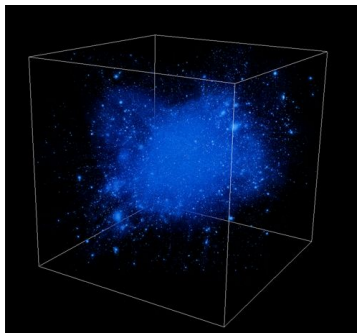
Science is Changing Exponentially

THOUSAND YEARS AGO
science was empirical
describing natural phenomena

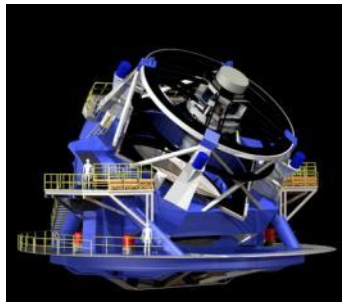
LAST FEW HUNDRED YEARS
theoretical branch using models, generalizations

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

LAST FEW DECADES
a computational branch simulating complex phenomena

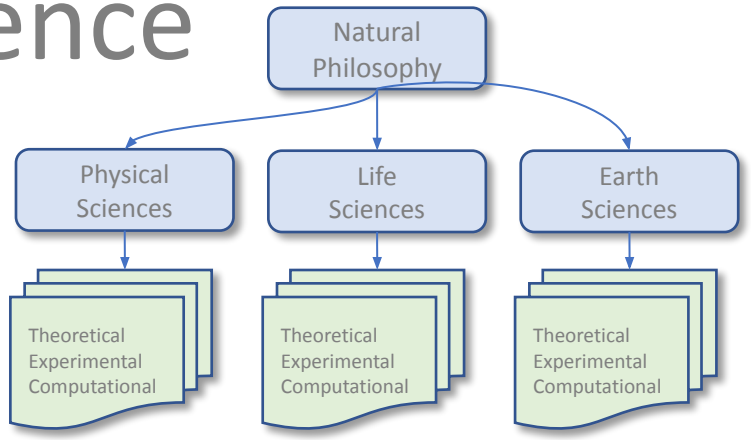


TODAY
data intensive science + AI, synthesizing theory,
experiment and computation with statistics
►new way of thinking required!



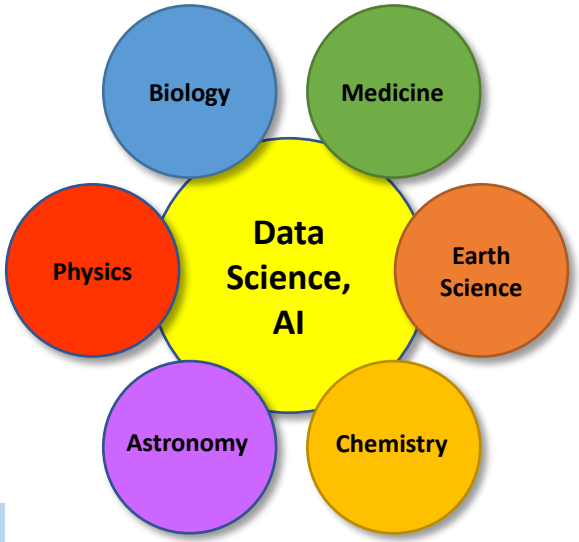
Science: From Fractal to Convergence

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

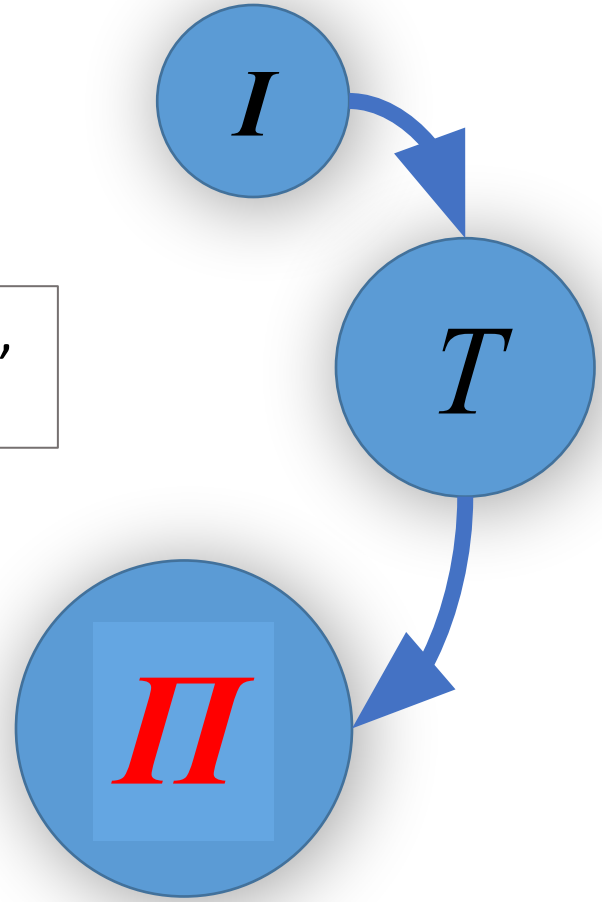
Our higher education is training deep but narrow people, ***I-shaped***

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/AI as well



We need to train Π-shaped people …

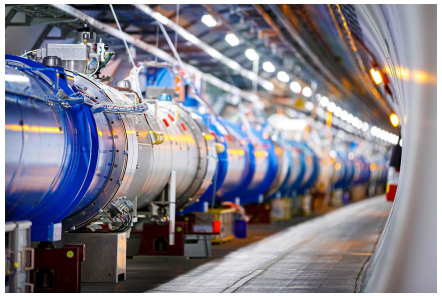
The Changing Granularity of Science

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



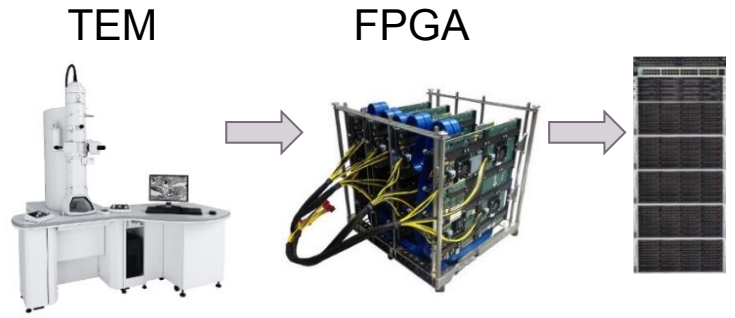
Van der Graaf -> Cyclotron -> Synchrotron -> National Labs

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!

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

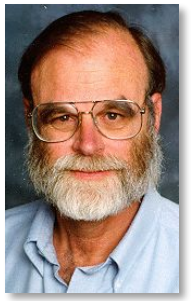
  => generating petabytes
- Innovative uses of AI will optimize experiments and discover new patterns
- This requires the data sets to be “AI-ready”
- The breakthroughs came from unique data sets (SDSS, AlexNet/CIFAR, Human Genome) – combined with a disruptive idea

Creating such datasets offer the best chances to remain competitive!

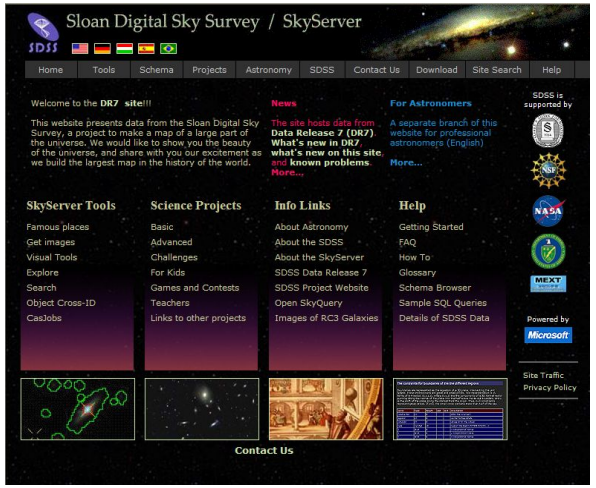
Mid-Scale Example: Sloan Digital Sky Survey

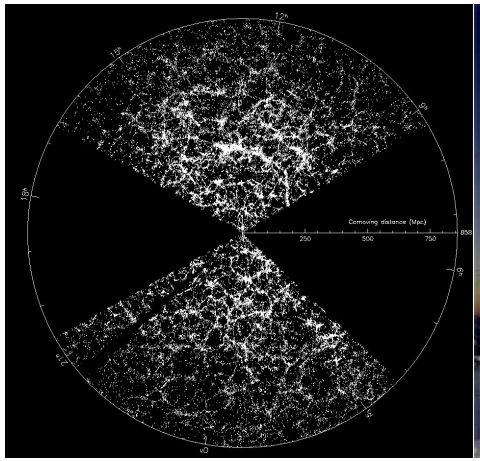
"The Cosmic Genome Project"

- 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 become publishers and curators of large data!

Main Concepts in the SDSS Design

- Requirements definition via “20 queries”
- Metadata encapsulated as comments into DDL -> autogenerate docs
- Capture and transform data (ETL -> ELT)
- Preserve hyperlink to raw data
- Annual versioning (DR*, accompanied by paper in journal)
- DB integrated with a visual, interactive interface
- SQL backdoor enabled

Always go from working to working!

Lessons Learned (Patterns to Processes)

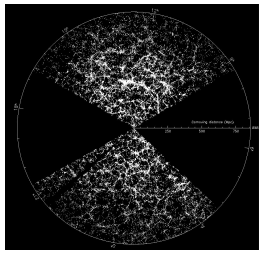
- Statistical analyses and collaboration **easier with DB** than flat files
- **Collaborative features** essential
- Need to **go beyond SQL** scripting => Jupyter and Deep Learning
- Everything is **spatial**
- **Multiple access patterns** (visualization, interactive and batch analyses)
- **Automation**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

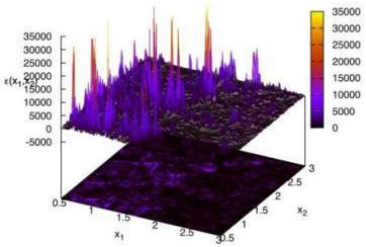
Find the right tradeoffs -- do not try to do “everything for everybody”

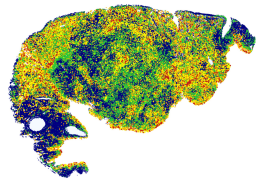
Mid-Scale Science => “Game Changing” Data

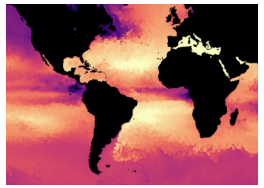
Leapfrog – “non-incremental” – Mid-Scale Science projects at JHU

- (2001-) **Sloan Digital Sky Survey (SDSS)** – grew data by a **factor of 100**,
  still the world’s most used astronomy facility,
  5.1B web hits, 800M SQL queries, 10M users, 13K papers, 770K citations
- (2006-) **Turbulence database (JHTDB)** the **world's largest simulations,**
  the "virtual observatory" of turbulence,
  1.5PB of data, 566 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
  28T pixels, 1B 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









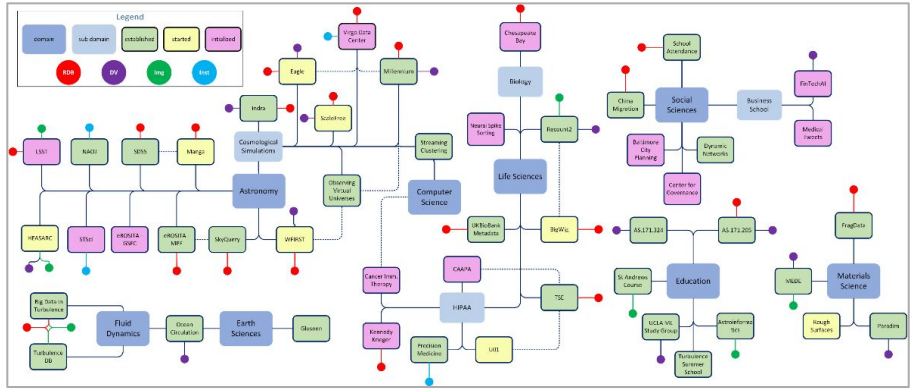
Using similarities to the SDSS, we are able to create unique leapfrog projects over and over

IDIES: Open Science with Interactive Petabytes

Institute for Data Intensive Engineering and Science

idies

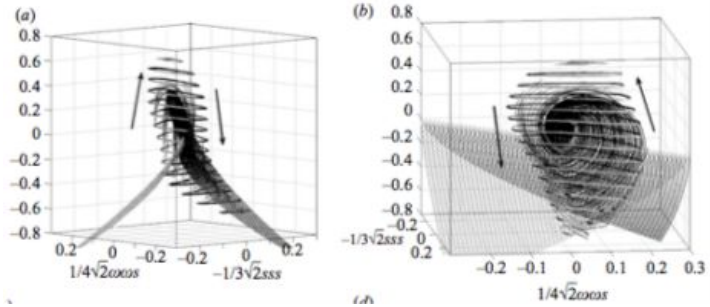
- Provide “disruptive assistance” -- from “patterns to processes”
- Institutionalize “lessons learned” in a multidisciplinary setting
  - Science engagements have distinctive “phases of maturity”
  - Critical mass of interdisciplinary postdocs and software engineers
- Convergent, multidisciplinary engagements (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
  - Increasing use of AI tools

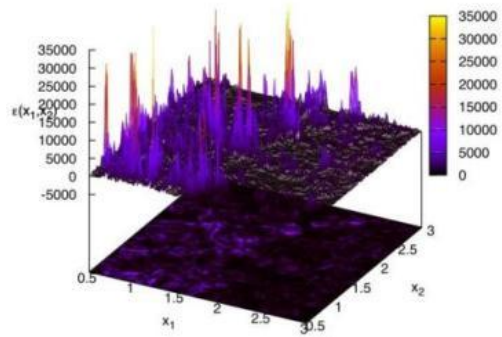


Immersive Turbulence

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

- **Understand the nature of turbulence**
  - *First: 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 (2005-) like in the movie Twister*





- **New paradigm** for analyzing simulations!

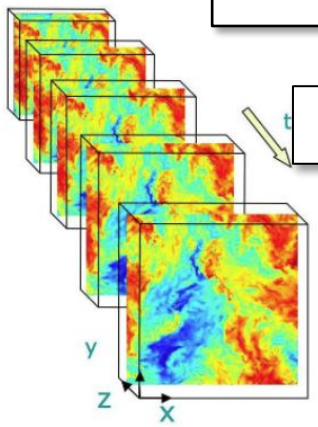
Johns Hopkins Turbulence Databases

Home | Database Access | Documentation | Links | Visualizations | About

NOTICE: Jul-27-2021. Servers are functioning normally. For past announcements, please click here

Welcome to the Johns Hopkins Turbulence Database (JHTDB) site

http://turbulence .pha.jhu.edu/



565,743,570,786,292 points queried

Access to the data is facilitated by a Web services interface that permits numerical experiments to be run across the Internet. We offer C, Fortran and Matlab interfaces layered above Web services so that scientists can use familiar programming tools on the ion of Spatial differentiation using various order approximations (up to 8th order) and filtering are also supported (for details, see documentation page). Particle tracking can be performed both forward and backward in time using a second order accurate Runge-Kutta integration scheme. Subsets of the data can be downloaded in hdf5 file format using the data cutout service.

To date the Web-services-accessible databases contain a space-time history of a direct numerical simulation (DNS) of isotropic turbulent flow in incompressible fluid in 3D (100 Terabytes), a DNS of the incompressible magneto-hydrodynamic (MHD) equations (50 Terabytes), a DNS of forced, fully developed turbulent channel flow at Re_τ=1000 (130 Terabytes), a DNS of homogeneous buoyancy driven turbulence (27 Terabytes), and a transitional boundary layer flow (105 Terabytes). Also available are individual snapshots (spatially but not temporally resolved data) of 4096^3 DNS of isotropic turbulence (1 snapshot), 8192^3 DNS of isotropic turbulence (6 snapshots at higher Reynolds number), rotating stratified turbulence (5 snapshots, 5 Terabytes), and channel flow at Re_τ=5200 (11 snapshots, 20 Terabytes). Basic characteristics of the data sets can be found in the datasets description page. Technical details about the database techniques used for this project are described in the publications.

The JHTDB project is funded by the US National Science Foundation. JHTDB operations is also supported by the Institute for Data Intensive Engineering and Science. JHTDB data may also be accessed via SciServer resources.

Questions and comments? turbulence@lists.johnshopkins.edu

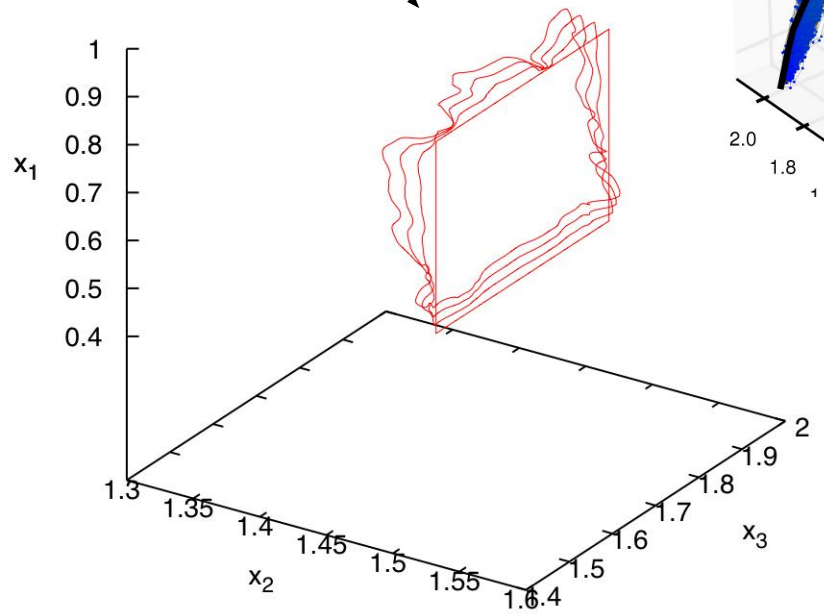
187,955,501,619,752 points queried

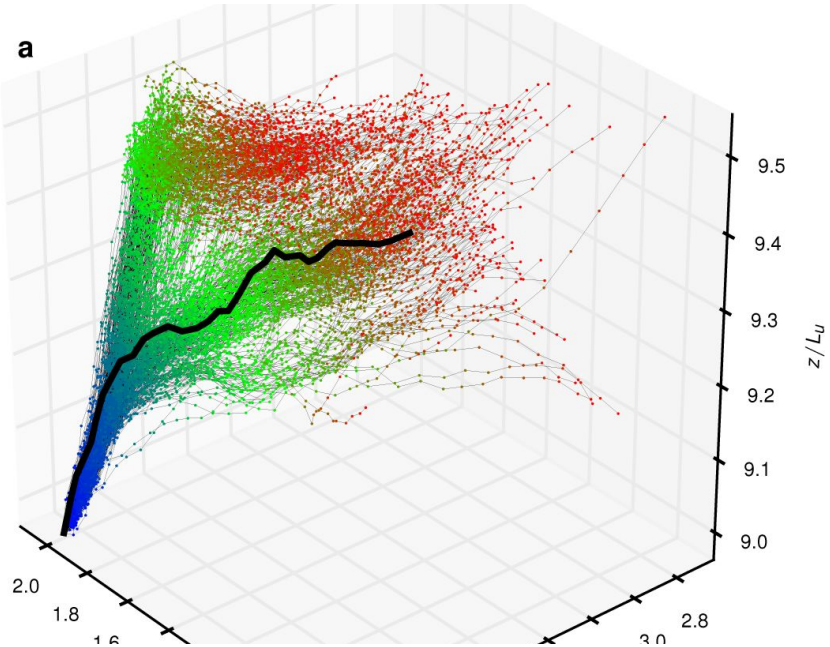
Move backwards in time

```
!
 do it = 1,15,1
!
   print *,'time = ',time
   time=time+deltat
!
   CALL getvelocity(authkey, dataset1, time, Lagrangian6thOrder , PCHIPInterpolation, 4*n, points, dataout)
 !
  do i=1,4*n
  do k=1,3
     points(k,i)=points(k,i)+dataout(k,i)*deltat
  end do
  end do
!
  if (it.eq.5.or.it.eq.10.or.it.eq.15) then
   do i=1,4*n
     write(10,*) points(1,i),points(2,i),points(3,i)
   end do
     write(10,*) points(1,1),points(2,1),points(3,1)
  endif
  write(10,*) '   '
  end do
!
  endif
```

minus

advect backwards in time !

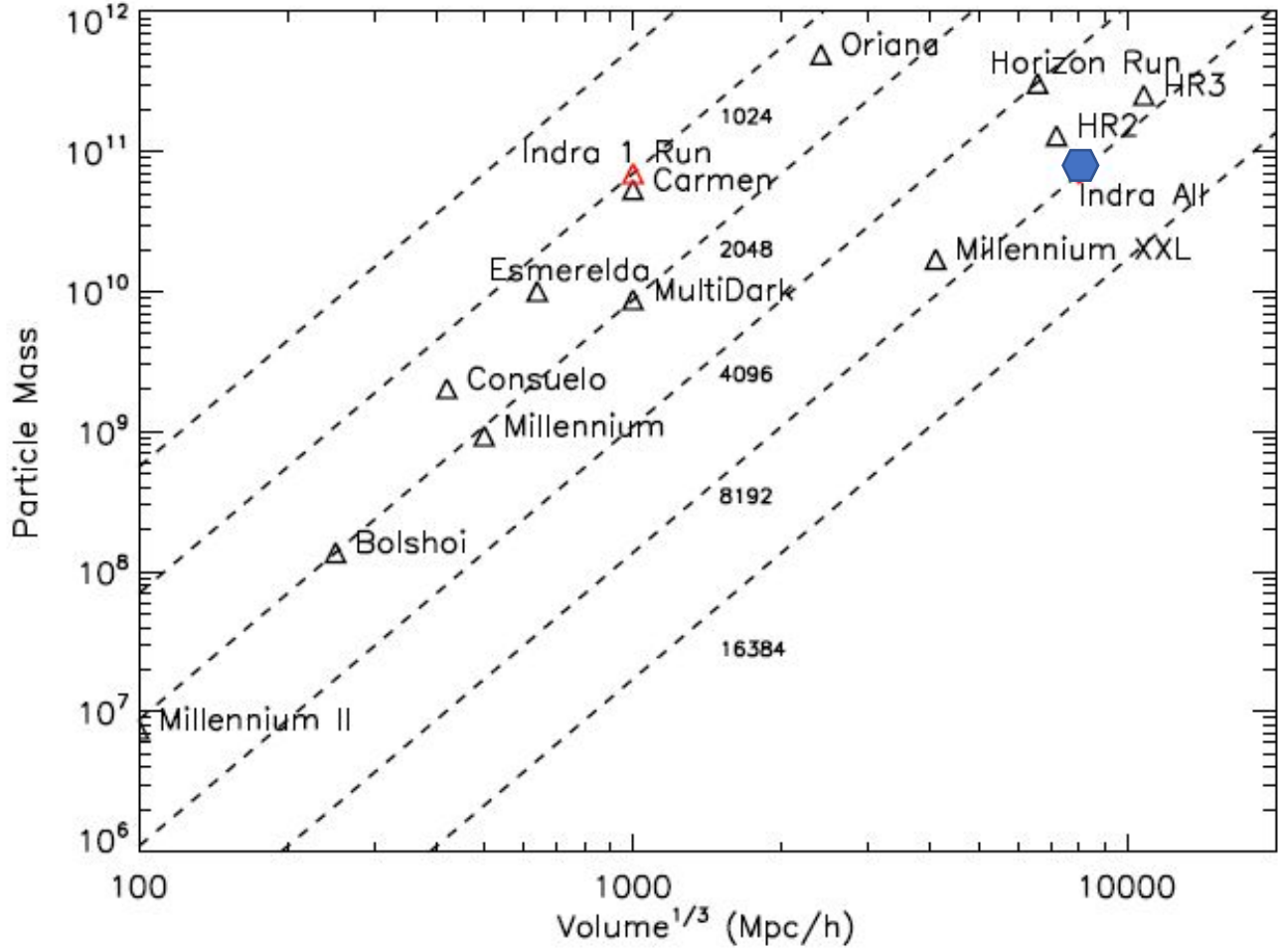




Eyink et al (2013), Nature

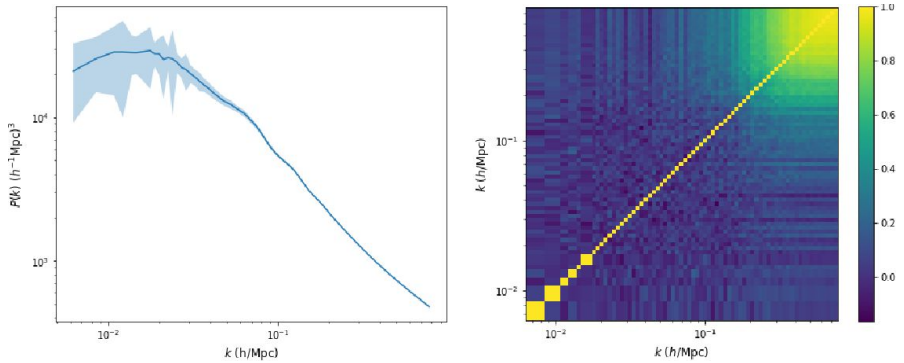
Not possible during DNS

Cosmlogical N-body Simulations



The Indra Simulations

- Suite of dark matter N-body simulations
  - 512 different random instances, WMAP7 cosmology
  - each 1 Gpc/h-sided box
  - 10243 particles per simulation
  - About 1 PB of data
  - All data loaded into a SQL database
  - Available to the public



Bridget Falck etal 2021

- Particle data:
  - All particle positions and velocities for all 64 snapshots of each simulation run
- Halo catalogs:
  - Standard Friends-Of-Friends (and others), linked to particles
- Fourier modes:
  - Density grid for 512 time steps of each run

Recently used for training Deep Learning, reconstructing the peculiar velocity field from positions and redshifts

Chen et al 2023

The Evolving Data Analysis

The evolution of the music industry is a good example (it is happening, like it or not…)

LP/CD 

=> iTunes

=> Spotify/Pandora

What are the data equivalents?

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

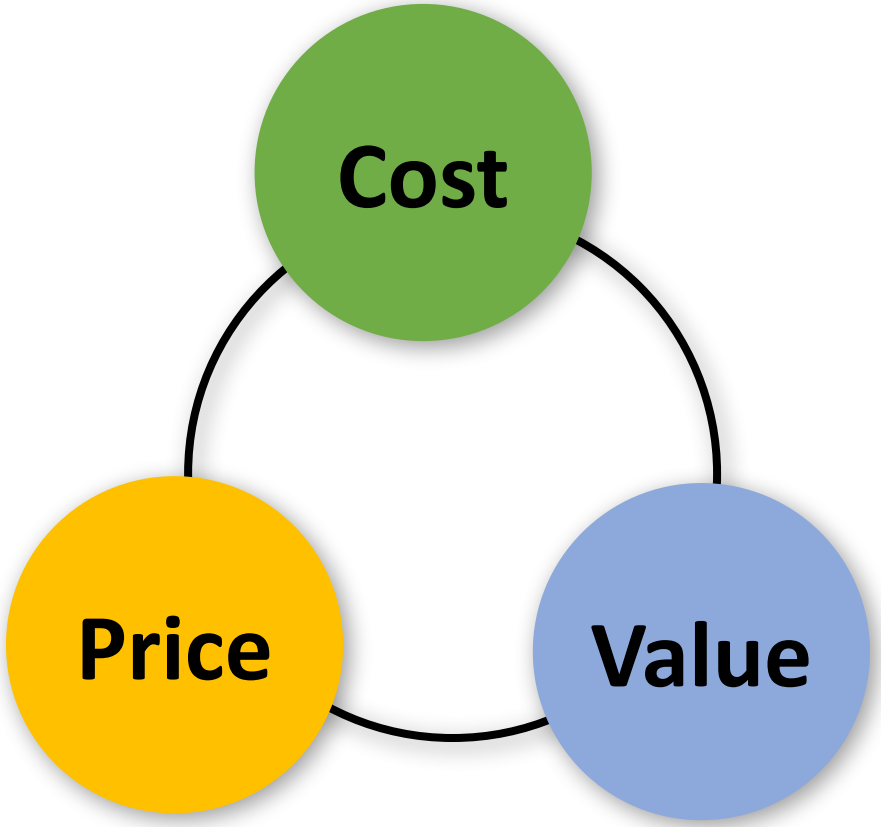
=> Run queries at project servers — ***Astronomy archives, SkyServer, IVOA, MAST, NED,…***

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

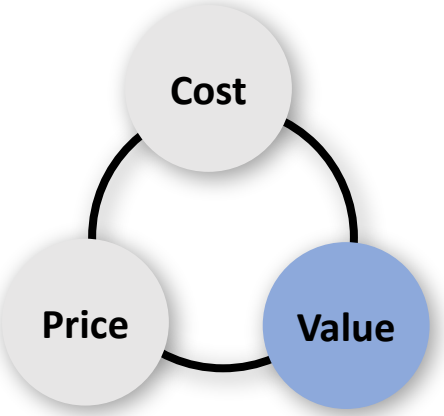
Scientific software needs to be Analysis Ready and Cloud Optimized (ARCO)

Ryan Abernathey (Columbia)

Measuring the Usefulness of Data



The **VALUE** of Scientific Data

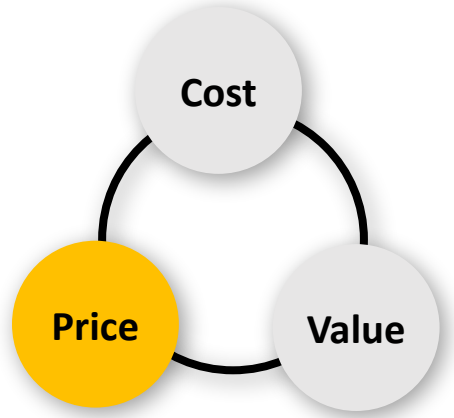


- What is the VALUE of data?
  - Accelerates testing ideas, find targets for followup
  - Provides the basis for reproducible science
  - **Direct foundation to many publications**

- ***Metric: how much science funding does a data set attract?***
  - Typical NSF Astronomy grant is $300K over 3 years
  - 1-2 refereed paper/year implies: ~$100K/paper (not $10K and not $1M)
  - Scientists **spend $100K of their research money** to work on a data set

Every paper is a $100K vote on the VALUE of a data set!

The **PRICE** of Scientific Data



- What is the **PRICE** of data?
  - How much did it take to build and run the experiment
  - Spent over a decade during the lifetime of the experiment
  - About 50-50 split between construction and operations

    - New electron microscope $10M-$50M
    - For the Sloan Digital Sky Survey around $200M
    - Rubin/LSST over $1B
    - CERN LHC, Hubble, James Webb over $10B

All that remains for this investment at the end of the experiment is the DATA!

The **COST** of Scientific Data

- The annual **COST** is maintaining the data
  => mostly in people, not the disks
- So far so good, while data owned by live projects

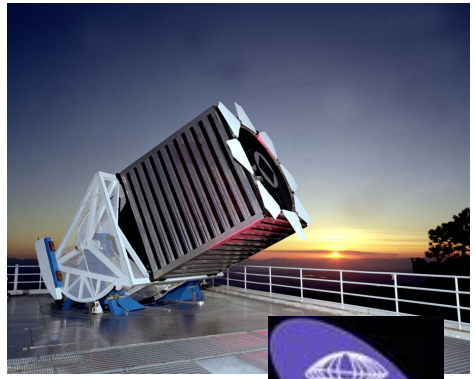
- But: experiments will be shutting down…
- Data is offered through smart services
  - What happens to the data?
  - Who will remember what it is?
  - **Who will take ownership of it?**
  - Who will pay for it?

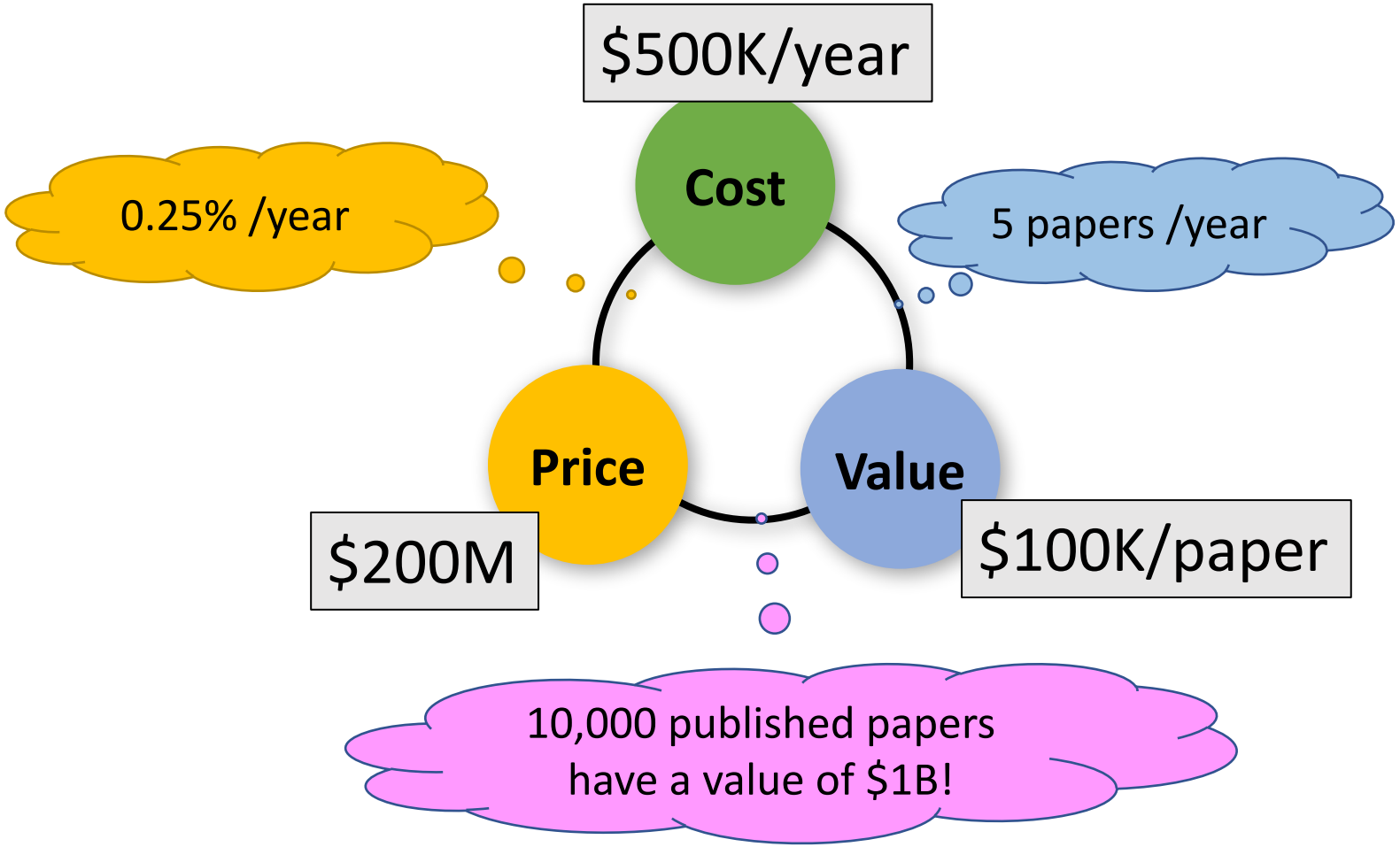
- 1885 dead drives in SDSS
- 2.5%/year over 18 yrs
- 1.1 tons of failed HDD

But: this is the EASY part!!!

There must be a "trusted intermediary" between the data and the scientists…

Comparison for the SDSS





5% of the survey price would cover the data for 20 years!

The Emergence of AI

AI in Science Today

- Much related to posterior analyses of existing data
  - Proxy simulations (turbulence, cosmology, cloud formation)
  - Recognizing patterns (image segmentation, Alpha fold, denoising)
  - Compression, discovering correlations
  - Anomaly alerts
- Recent developments with LLMs
  - They can recite much of the literature
  - ChatGPT – beware of “hallucinations”
  - But zero-shot, fine-tuning
    - why and how does it work???

Using LLMs

- Solve the “Long Tail” problem
  - Most scientific data sets are small, and appear as tables in papers
  - Publishing them in a reusable digital form very hard
  - Efforts to capture this have been a total failure
- But: we could (and should) use the LLMs to harvest data
  - We have the digital text of the surrounding information in the paper
  - We also have to list of coauthors and their papers for broader context
  - The AI framework can extract not just the data but their meaning and context
- Fine-tuning (LoRA)
  - Easy to build generative models
  - Use the LLM as a generic pattern recognition engine
  - This may work because of the “long-tail” of natural processes: 1/f everywhere

Automatic Code Generation

- We have now LLMs trained on github etc (Copilot)
- They are quite successful in writing code from scratch
- Science is interactive: we often explore data in a hit and miss fashion
  - We start with a smaller subset of data, try many things
  - Lots of scattered dead-end
  - We still do a manual cleanup of our attempts to write a clean script in the end
  - Wouldn’t it be nice to have a button on top of a Jupyter notebook that would generate a clean script from my attempts?

Nature is Sparse (and non-Gaussian)

- Many natural processes are dominated by a few processes and described by a sparse set of parameters
- They are also full of sharp, non-Gaussian features (edges…)
- Compressed Sensing has emerged to identify in high dimensional data sets the underlying sparse representation (Candes, Donoho, Tao, etal)
- This enables signal reconstruction with much less data!
- The resolution depends not on the pixel count but on the information content of an image…
- Sampling very skewed distributions is hard, but the layers of neural nets are acting like random projections -> Gaussian sampling…

Explainable AI

- Scientists do not like Black Boxes
- We need to know what is happening inside
- The Physics of AI is emerging in interpreting the evolution of the complex networks (has its roots in spinglasses)
- The initially random weights of a network develop long range correlations during learning – like a phase transition
- Identify symmetries in the problem
  - Latent layers of the autoencoders
  - Interpretable autoencoders emerging (Regev et al)

Big Data: From Sensors to Tensors

- Two kinds of errors: statistical and systematic
- Statistical errors decrease with $1/\sqrt{N}$

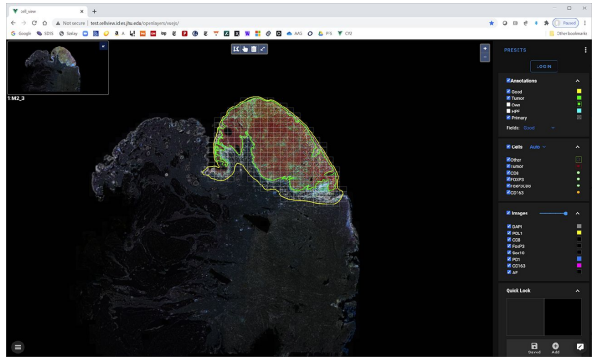
- Big Data needs parallelism: many similar, inexpensive devices
- This scale-out is everywhere, like cloud computing
- Same in experiments, many similar cheap sensors
  - phones, wearables, CubeSat…

- However, **similar is not identical**!
  - **Systematic errors**: subtle instrumental biases
  - If obvious, we call it calibration, and do it
  - If not, it remains often undetected

- In most scale-out projects the biggest challenges are the systematic errors
- But: these can be corrected in software, much cheaper overall!
- Particularly important for AI training sets (“garbage in, garbage out”)

AI can tackle the hardest problems in experimental scaleout!

AstroPath: Atlas of Cancer Cells



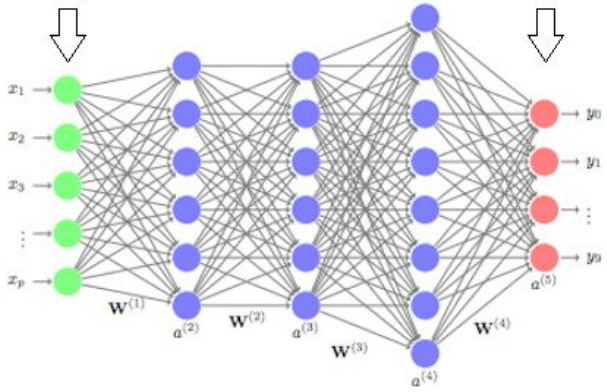
- **Astro**nomy meets **Path**ology
  - Project started by Janis Taube (JHMI BKI) and Alex Szalay (JHU IDIES)
- Studying the spatial interactions of activated T cells and tumor near the tumor boundaries
- Parallels sky surveys (as of 20 years ago)
  - "Disruptive assistance" from astronomy to pathology
  - Using techniques astronomers learned the hard way (flat field, unwarp, calibrate)
- Transitioning to the "industrial revolution" of data acquisition
- Goal: increase data collection by a factor of >1000
  - 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 => Petabytes/year
  - Heavy use of parallel processing, automation, scale-out (and AI tools)
- Databases linked to SciServer, collaborative Jupyter, PyTorch, Keras/TensorFlow, R
- Goal is to build a significant spatial atlas with billions of cancer cells

Current census

Slides	759
Cells	1,074,449,553
Images	539,331
Pairs	11,368,538,836

Great example of a 21st Century experiment, transformed by AI

Replacing the Human with AI



- Currently three steps of human involvement
  - Staining/Scanning
    - This will require a human for the foreseeable future,
    - Aided by largely automated scanners
  - Tissue annotations
    - We are ready to deploy a U-Net-based segmenter, using ViT
    - Tissue outlines and torn tissue masks already functional
  - Cell segmentation/classification
    - We have Mesmer + transfer learning, already better than existing commercial product
    - Expert validation over 100k cells just finished
- Todo:
  - Need to replace unmixing and denoising algorithms
    - Planning to complete this by the end of 2024

Digital Pathology at JHU

- JHU is producing 600K cancer slides/ year
- Fully annotated during surgery
- Whole medical history known
- Collected in a warehouse since 1980

- We can generate 20M scanned slides at a fraction of $1 each
- We can train a LLM on the annotations, and use the cell data for training on many different types of cancer

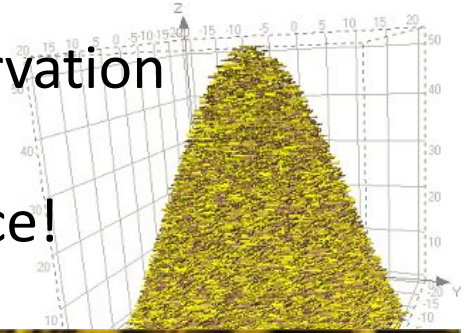
AI is becoming indispensable in running scalable experiments!

More Data Becomes Too Much Data

Prioritizing for Relevance

"Do you have enough data or would you like to have more?"

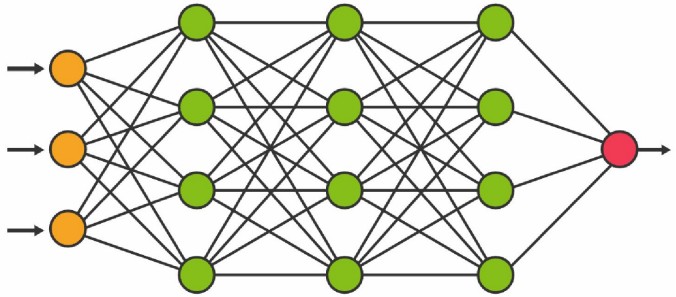
- Delicate **tradeoff** between the scientific value and the cost of preservation
  - One extreme – store everything, go bankrupt!
  - Other extreme – collect too little data, not enough for the science!
  - **We are approaching a critical point**



- 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?)

Use AI to Collect More RELEVANT Data!



- Use of Generative AI to learn the corpus of known science
- Build autoencoders to recognize if incoming data is inside the affine hull of the known parameter space
- Use this to downsample, and create well-stratified data sets that represent the **UNKNOWN** domain without breaking our budgets

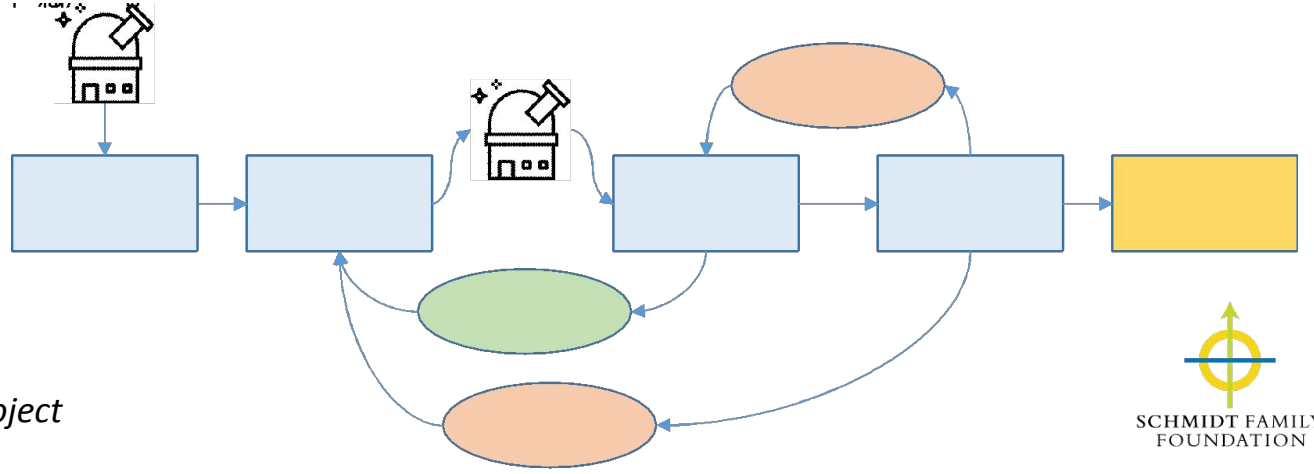
- ***This can have a much bigger impact on science than the posterior analysis of existing data sets!***

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

AI in Experimental Design

- Need to dramatically improve our experimental design…
- Machine learning is already used in various areas:
  - Adaptive target selection, active learning, Gaussian processes
  - It is already happening at CERN, material science, drug design, astronomy
- Maybe this will be the **Fifth Paradigm**, algorithms control our experiments
  => also make intelligent, real-time decisions

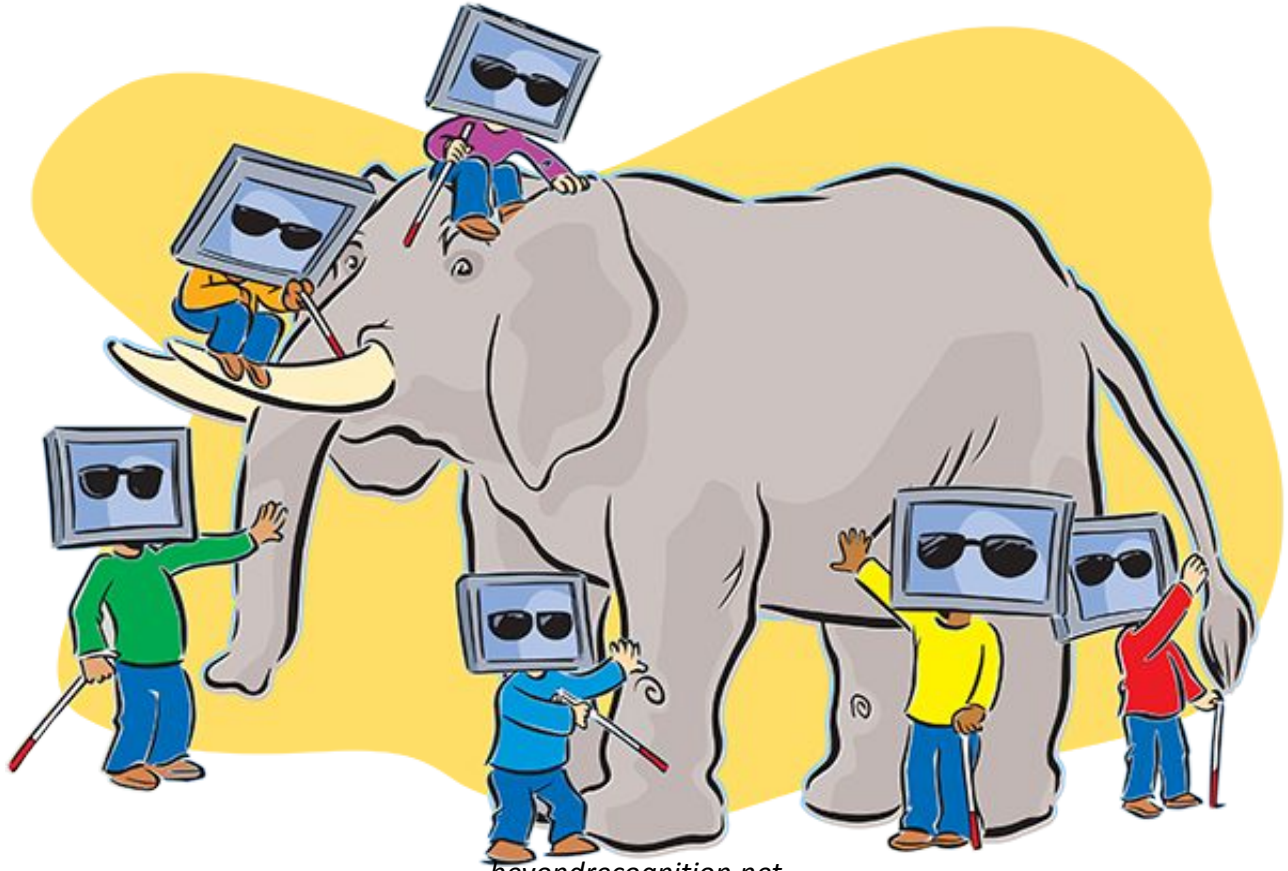
Put the telescope in the reinforcement loop!



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

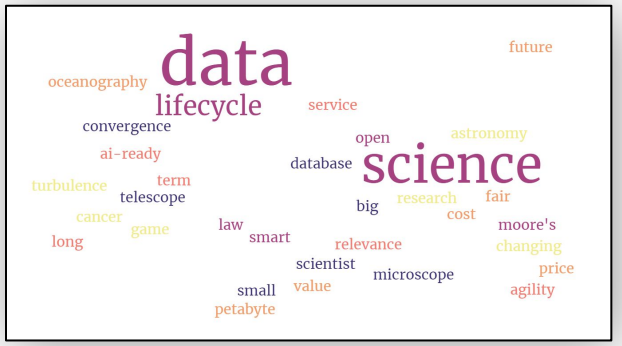
SCHMIDT FAMILY FOUNDATION

The Challenges Are Not Technical

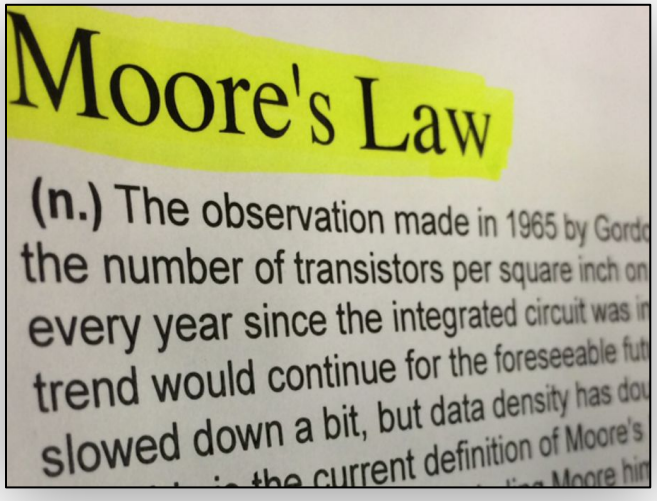


beyondrecognition.net

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

The value, price and cost of scientific data

- Preserving digital data is incredibly inexpensive

The emerging AI is changing everything

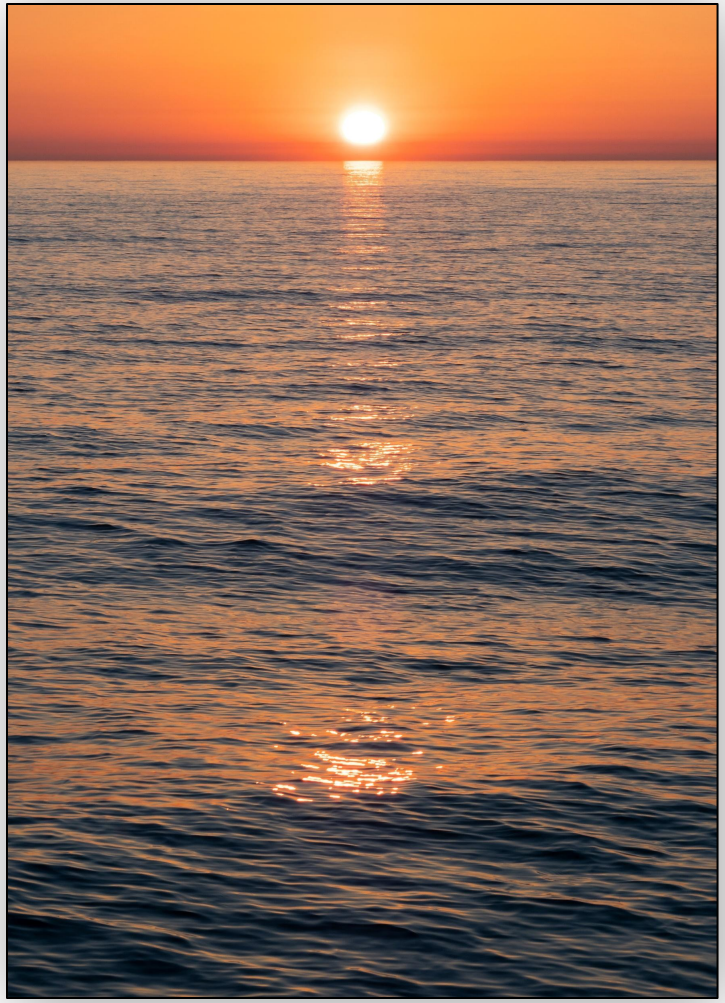
- We need explainable AI

We need to collect more relevant data

- How can we use AI to subsample the “known science”?

All of today’s science ends up as digital data

- This is the only legacy of the experiments
- Yet no coherent policy to preserve it for the long haul



Summary

- Use of Generative AI to learn the corpus of known science for a given experiment, using fine-tuning/LoRa
- Build AI tools recognize if incoming data is inside the affine hull of the known parameter space
- Use this to downsample, or compress using the parameters of the generative model and create well-stratified data sets that represent the **UNKNOWN** domain without breaking our budgets
- Recognize and correct/calibrate for systematic errors
- Optimize experimental strategy to maximize science goals
  - Target selection, active learning, denoising, …
  - Realize Ross King's vision of Robotic Scientist, but now driven by AI

AI can have a much bigger impct on our experiments than on our analyses!

“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)