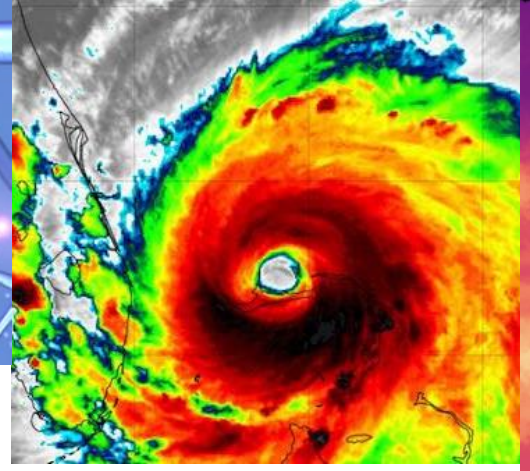
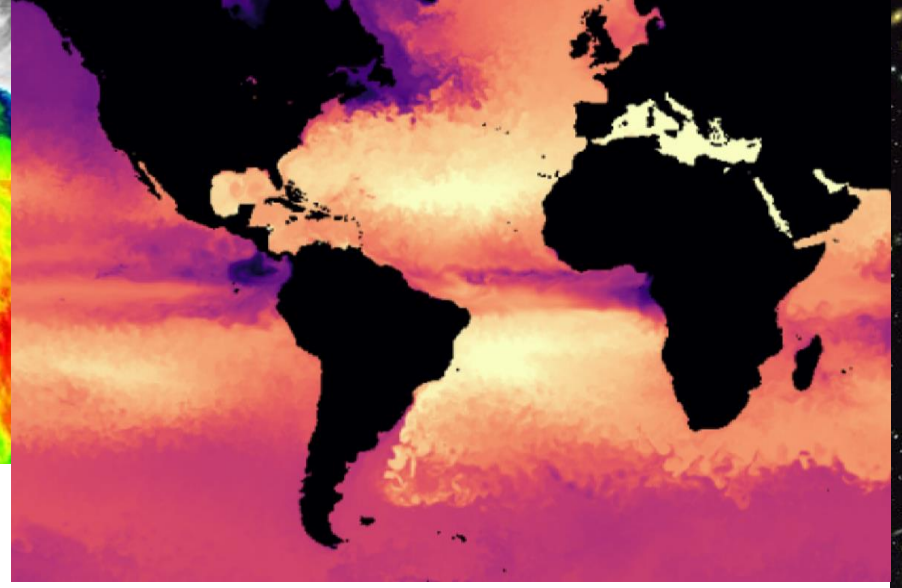




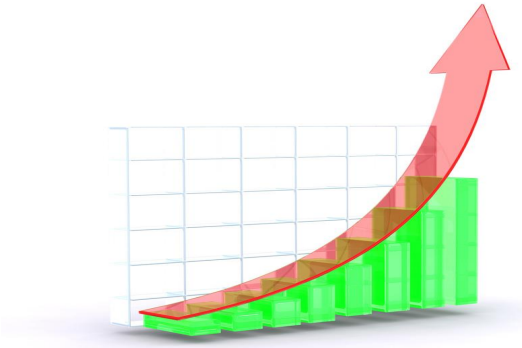

Science in the Era of AI

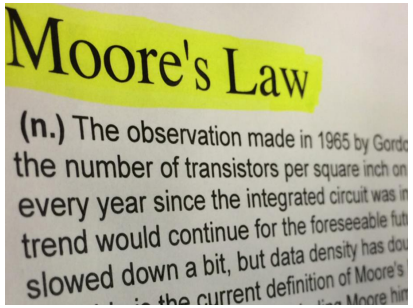
Alex Szalay

The Johns Hopkins University

Introduction

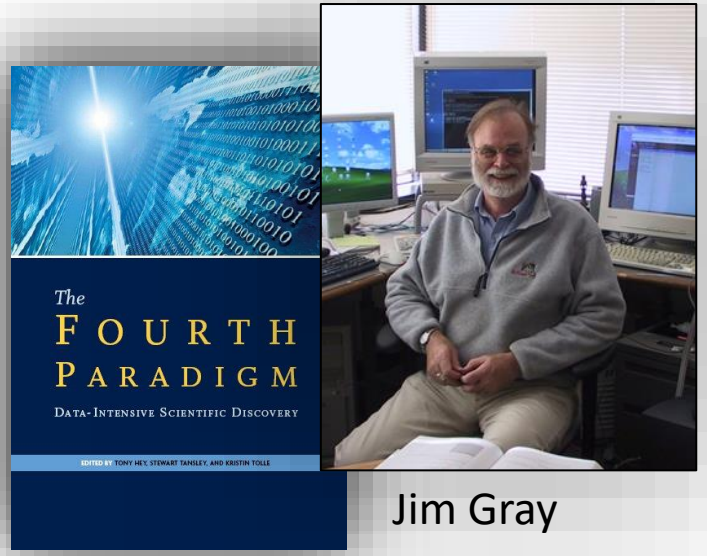
There is a new revolution in science, transforming everything





Enabled by the exponential growth in our computational technologies, using massively parallel “scale-out”

The Fourth Paradigm of Science emerged, driven by Open Data



Jim Gray

Agenda

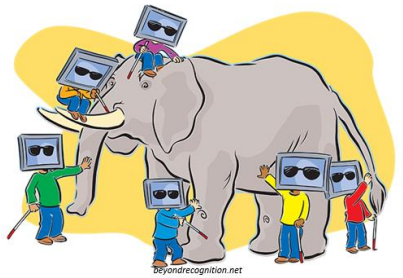
The Exponential Evolution of Science

The Changing Granularity of Science

Aggregating Data

The Emergence of AI

The Challenges Are Not Technical



The Exponential Evolution of Science

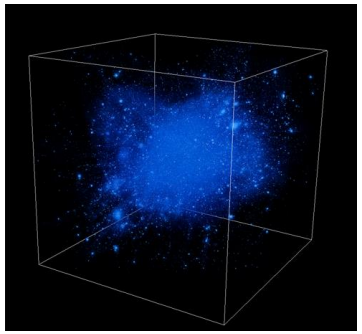
Science is Changing Exponentially

THOUSAND YEARS AGO
science was empirical
describing natural phenomena

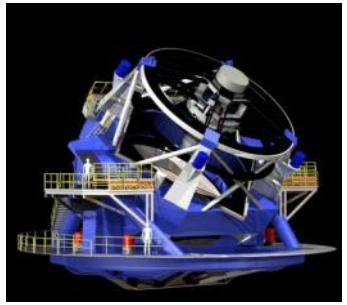
LAST FEW HUNDRED YEARS
theoretical branch using models, generalizations

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

LAST FEW DECADES
a computational branch simulating complex phenomena

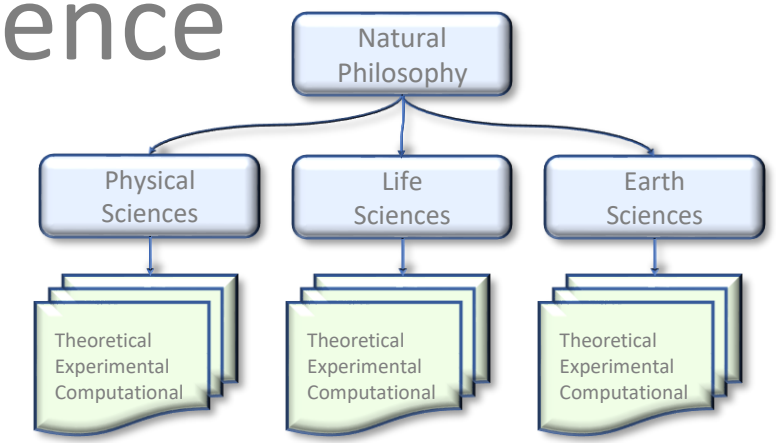


LAST FEW YEARS
data and AI driven, synthesizing theory, experiment and computation with statistics
►new way of thinking required! AI coming



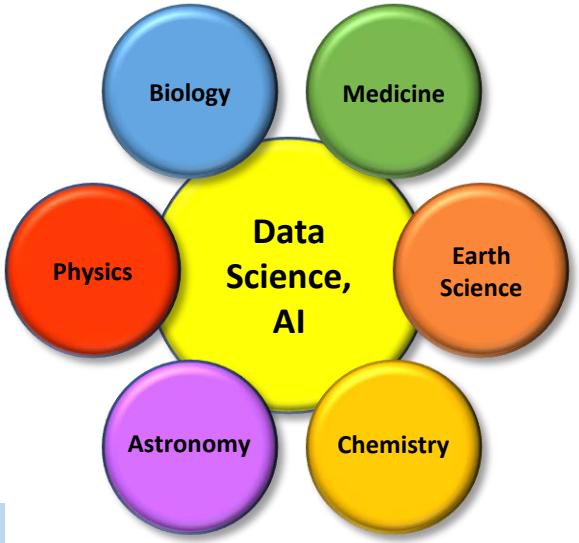
Science: From Fractal to Convergence

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/AI methods and approaches



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

Tomorrow's Scientists are Multi-Disciplinary

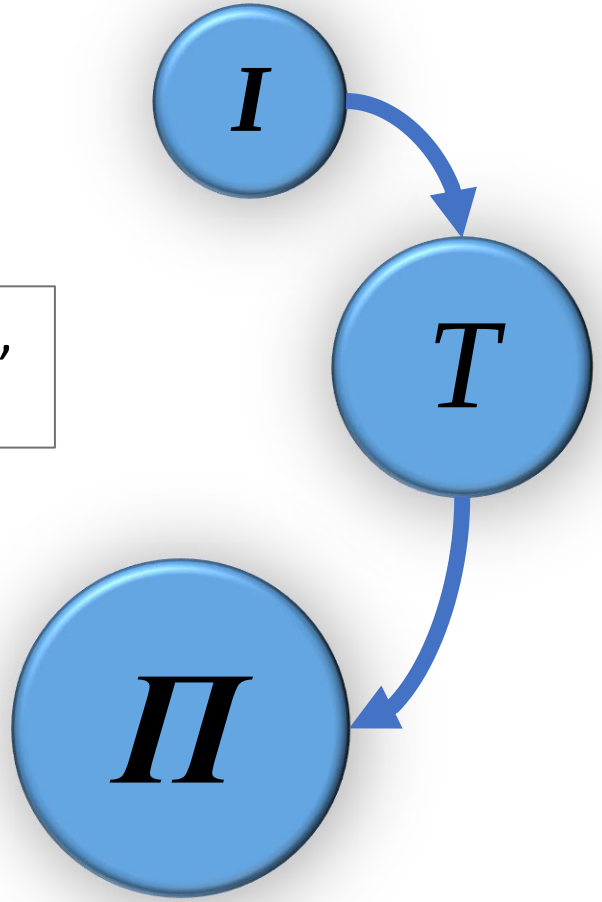
Our higher education is training deep but narrow people, ***I-shaped***

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

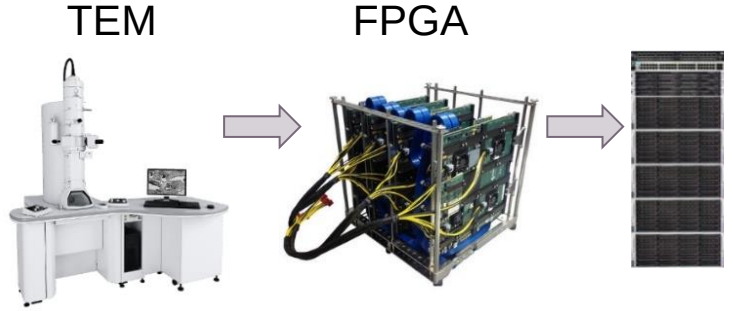
Van der Graaf -> Cyclotron -> Synchrotron -> National Labs

SSC ☹

LHC ☺

The data is here to stay for decades…

Today's Hot 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, simulation…)
  - Use cutting edge technology, take risks, push budgets to the limit, maximize science, generate petabytes of data, use AI to analyze
  - ***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

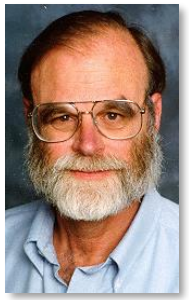
  => generating petabytes
- Innovative uses of AI will optimize experiments and discover new patterns
- This requires the data sets to be “AI-ready”

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

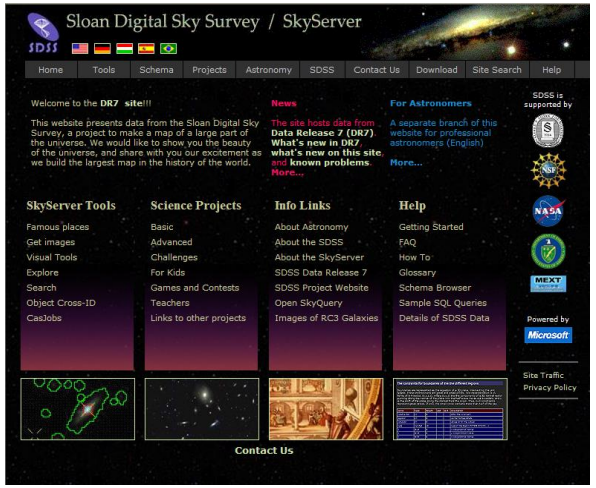
Mid-Scale Example: Sloan Digital Sky Survey

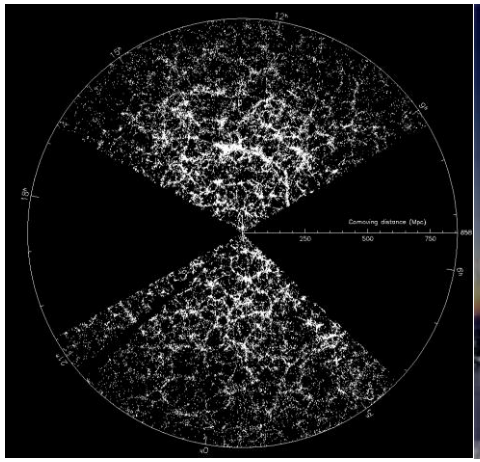
"The Cosmic Genome Project"

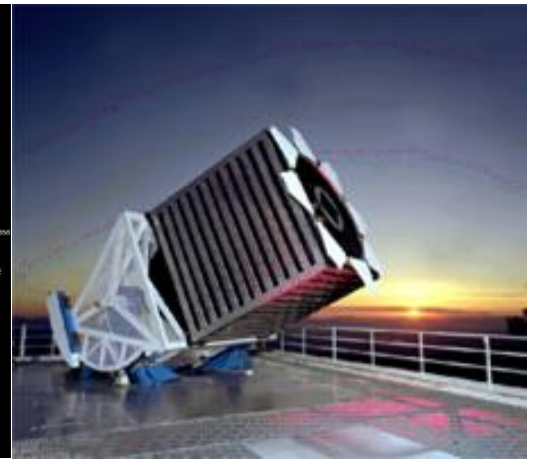
- Started in 1992, SDSS-II finished in 2008
- Data is entirely ***public, open and free***
- Database built at JHU
- Project marked a transition in astronomy
  - From manufacturing to mass production



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!

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 (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”

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)

Aggregating Data

Implications of Big Data

- 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”)

Statistical and Computational Challenges

- Data volume and computing power double every year,
  - no polynomial algorithm can survive, only N log N
- Minimal variance estimators scale as $N^{3,}$
  they also optimize on the wrong thing
- The problem today is not the statistical variance
  - systematic errors => optimal subspace filtering (PCA)
  - If it is so large that it is obvious => calibration
- We need incremental algorithms, where computing is part of the cost function:
  - the best estimator in a minute, day, week, year?
  - … like training a neural network

SkyServer > 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

Turning Lessons into Practice

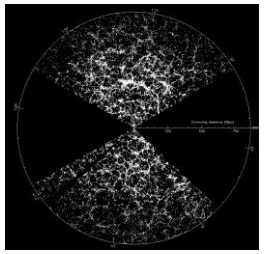
- We saw many **repeating patterns** for data intensive projects in different domains
- Now we are trying to turn these into **processes** that can be replicated
- Need to strike the right **tradeoff** of protecting the valuable data, while allowing creative (**disruptive**) innovations
- Invest more 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

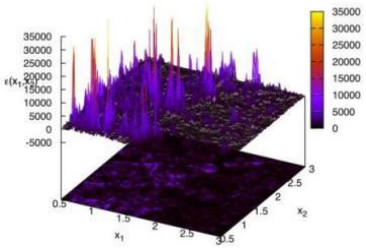
All of these require ongoing commitments, not one-shot investments

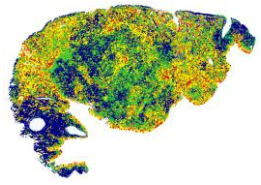
Mid-Scale Science => “Game Changing” Data

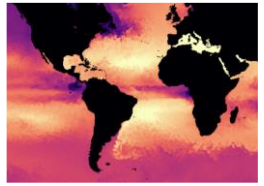
Leapfrog – “non-incremental” – but still Mid-Scale Science – Similarities

- (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

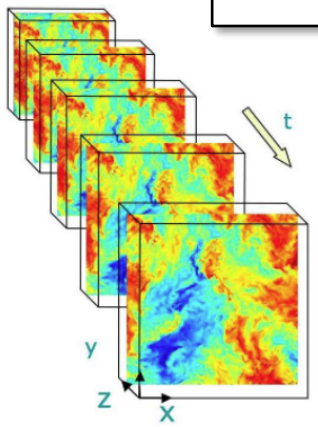
Johns Hopkins Turbulence Databases

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/



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 their client platforms. Calls to fetch subsets of the data can be made directly from within a program being executed on the client's platform. Manual queries for data at individual points and times via web-browser are also supported. Evaluation of velocity and pressure at arbitrary points and time is supported using interpolations executed on the database nodes. 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 a

296,940,985,463,805 points queried

s

s

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 idies. JHTDB data may also be accessed via SciServer resources.

Questions and comments? turbulence@lists.johnshopkins.edu

187,955,501,619,752 points queried

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 neig hbors, 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
- Increasing integration with AI tools for tissue annotation, segmentation and classification, biomarker inference

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 7 Tpixels!)
- Additional 100+ slides already scanned with multiple tumor types

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 Large Language Models
  - They can recite much of the literature
  - ChatGPT – beware of “hallucinations”

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

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?

Explainable AI

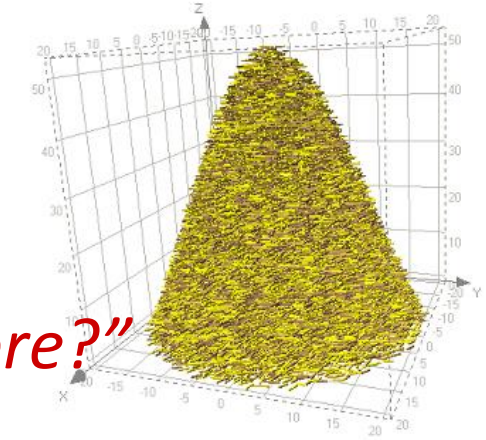
- Scientists do not like Black Boxes
- We need to know what is happening inside
- The Physics of AI is emerging in interpreting the evolution of the complex networks (has its roots in spinglasses)
- The initially random weights of a network develop long range correlations during learning – like a phase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

Finding Systematic Errors?

- Why can’t we feed also all our instrumental parameters to a Neural Network?
- If there are instrument-dependent biases, we should be able to find them in a systematic fashion
  - TB example…
- The removal of systematics, finding optimal calibrations could be automated instead of “black art”

AI will be probably just as important in our data acquisitions as it is in the analyses!

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**

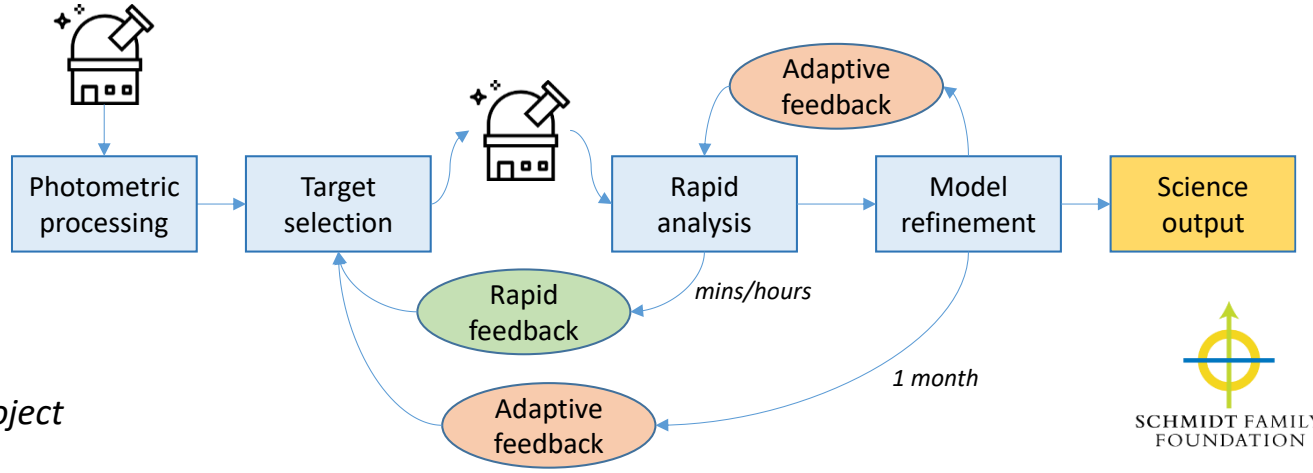
- **In-situ** (AI) hardware filters data, optimizing for "new science"
  - *Only 1 in 10M events saved (9999999:1)*
  - *Pattern recognition and filtering near the detectors:*
    - *"first meter problem" at the edge*
- Resulting "small subset" is still 10-100 PB

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

Collect More RELEVANT Data!

- Need to dramatically improve our experimental design…
- Artificial Intelligence in large-scale experiments:
  use AI **before** and **while** we collect the data
- 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!

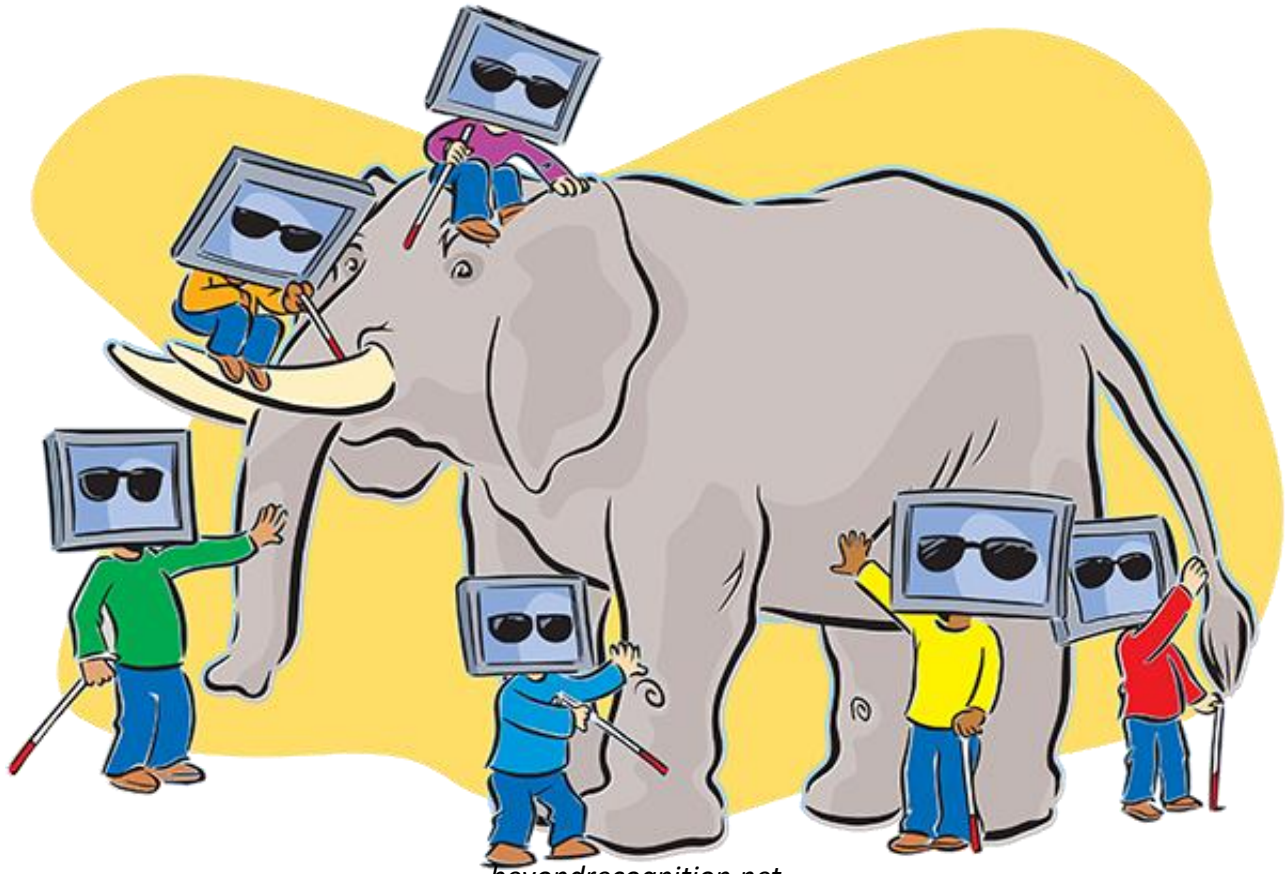


Supported by the Schmidt Family Foundation at JHU and Princeton: Use AI Feedback for the PFS project

SCHMIDT FAMILY FOUNDATION

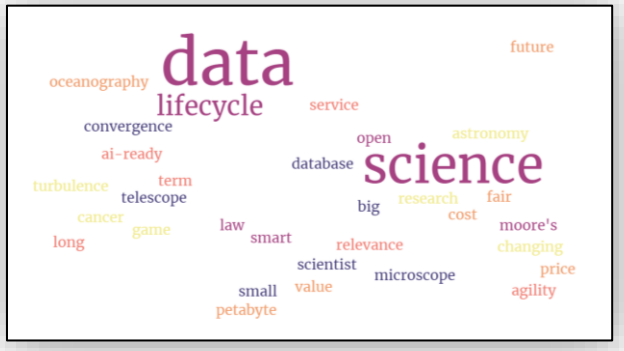
If an AI algorithm can drive our cars, why cannot it run our microscopes?

The Challenges Are Not Technical

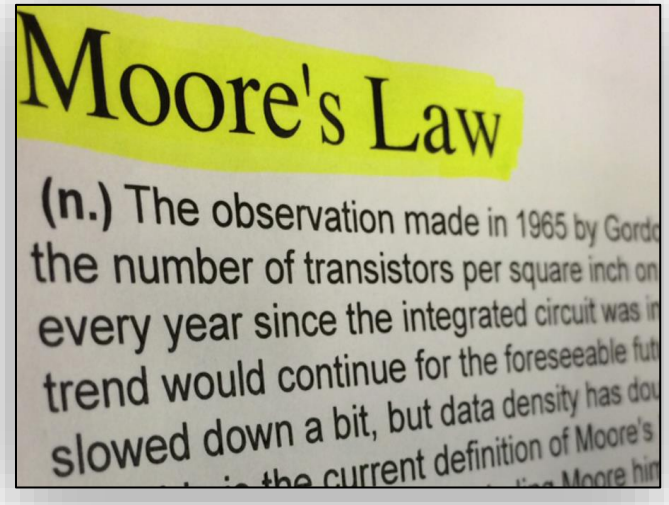


beyondrecognition.net

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… (NIST RDaF…)

- 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 -> AI?

Organization of science is changing

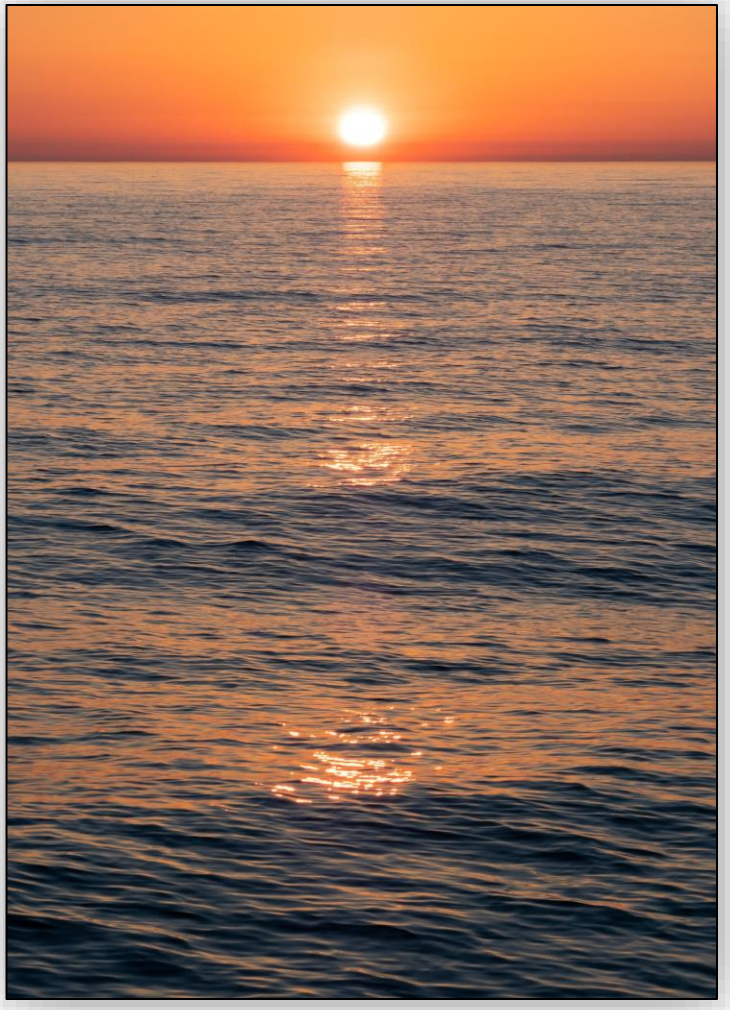
- Granularity of science (small → bimodal → mid-scale)
- Data sharing & long-lived data → Accelerating the change

Changing relationship between scientists and data

- Data only in papers → Now big datasets → Curation responsibility
- A trusted data intermediary → Empowers sharing and reuse
- Future: Automation is the key, AI everywhere
- Creation of high value data sets is crucial for success

AI is changing Science

- We only see the outlines of the future
- Every aspect of science will be changed by AI
- AI may even take control of our experiments
- Lots of challenges ahead, how to compete, who to compete with
- Universities will be transforming, including how we train students



"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)