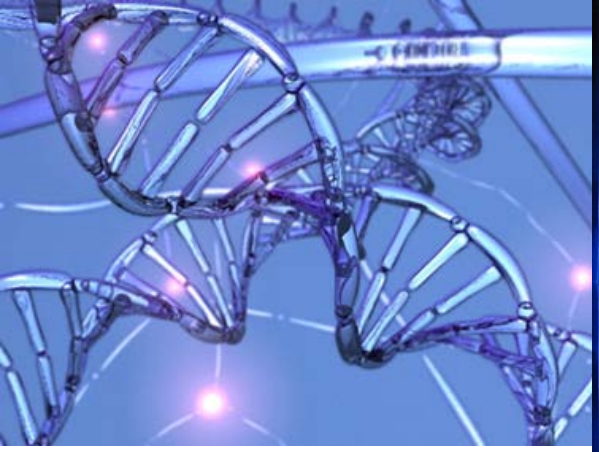


The Fourth Paradigm: How Big Data is Changing Science

Alex Szalay
The Johns Hopkins University

idies

Big Data in Science

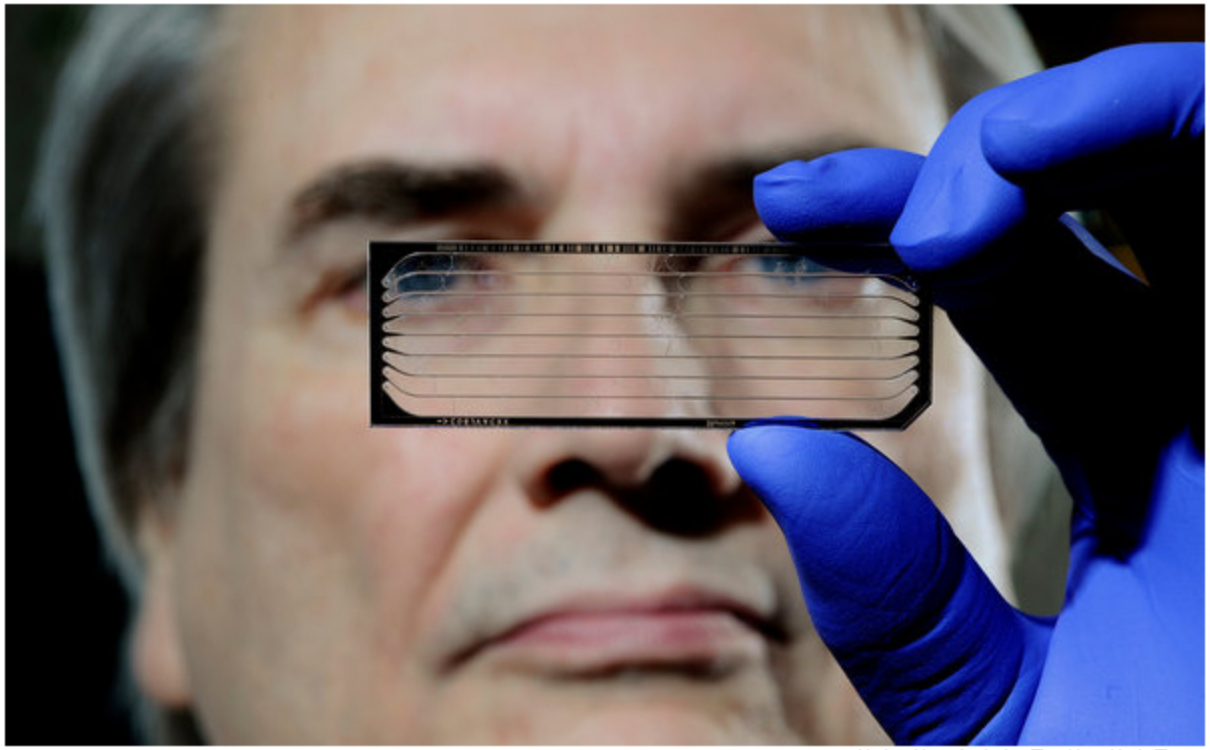
- Data growing exponentially, in all science
- All science is becoming data-driven
- This is happening very rapidly
- Data becoming increasingly open/public
- Non-incremental!
- Convergence of physical and life sciences through Big Data (statistics and computing)
- The “long tail” is important
- A scientific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

  => a rare and unique opportunity

The New York Times

Business Day

DNA Sequencing Caught in Deluge of Data



Kathy Kmonicek for The New York Times

W. Richard McCombie, a professor of human genetics at the Cold Spring Harbor Laboratory, examining DNA samples.

By ANDREW POLLACK
Published: November 30, 2011

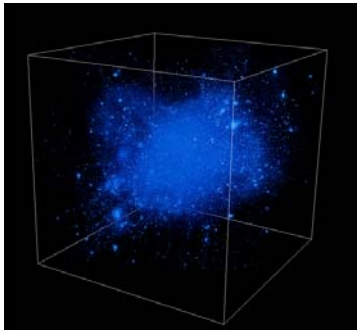
Science is Changing

THOUSAND YEARS AGO
science was empirical
describing natural phenomena

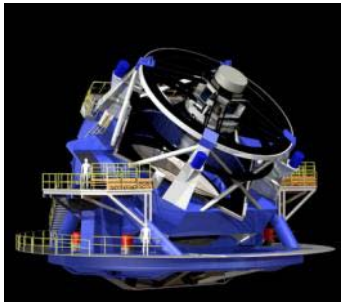
LAST FEW HUNDRED YEARS
theoretical branch using models, generalizations

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

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!



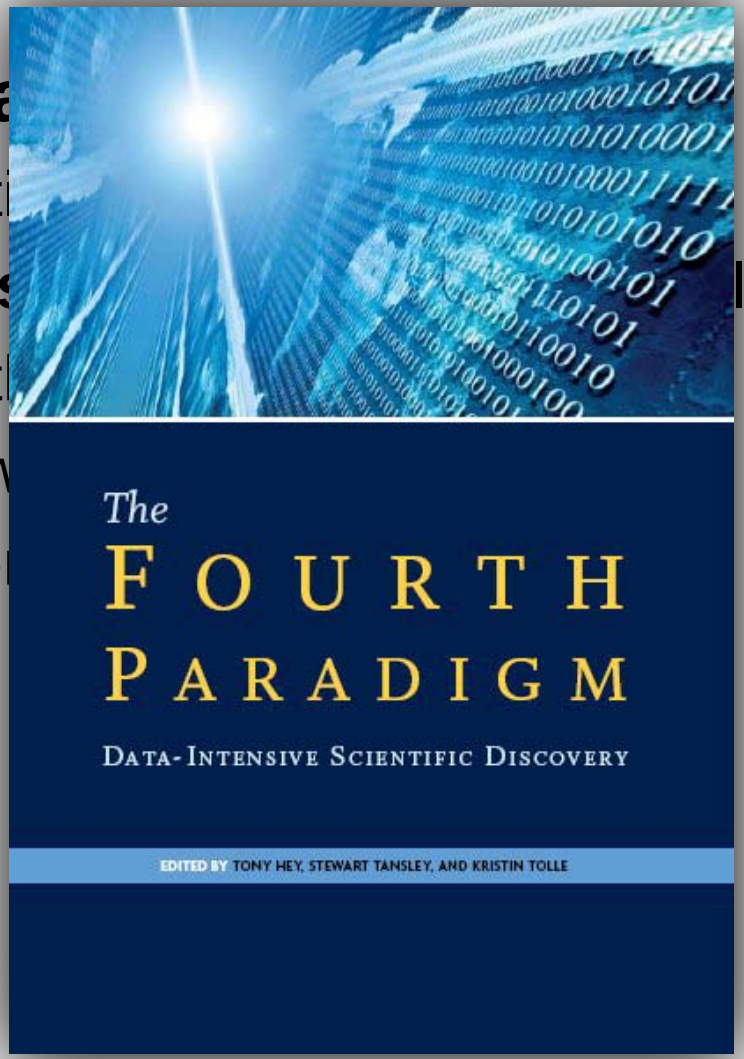
Scientific Data Analysis Today

- Data is everywhere, never will be at a single location
- Data grows as fast as our computing power
  - *Need randomized, incremental algorithms*
  - *Best result in 1 min, 1 hour, 1 day, 1 week*
- Statistical vs systematic errors
- Both “exploratory” and “confirmatory” searches needed
- Architectures increasingly CPU-heavy, IO-poor
- Most scientific data analysis done on small to midsize BeoWulf clusters, from faculty startup, in broom closets
- Universities hitting the “power wall”
- **Not scalable, not maintainable…**

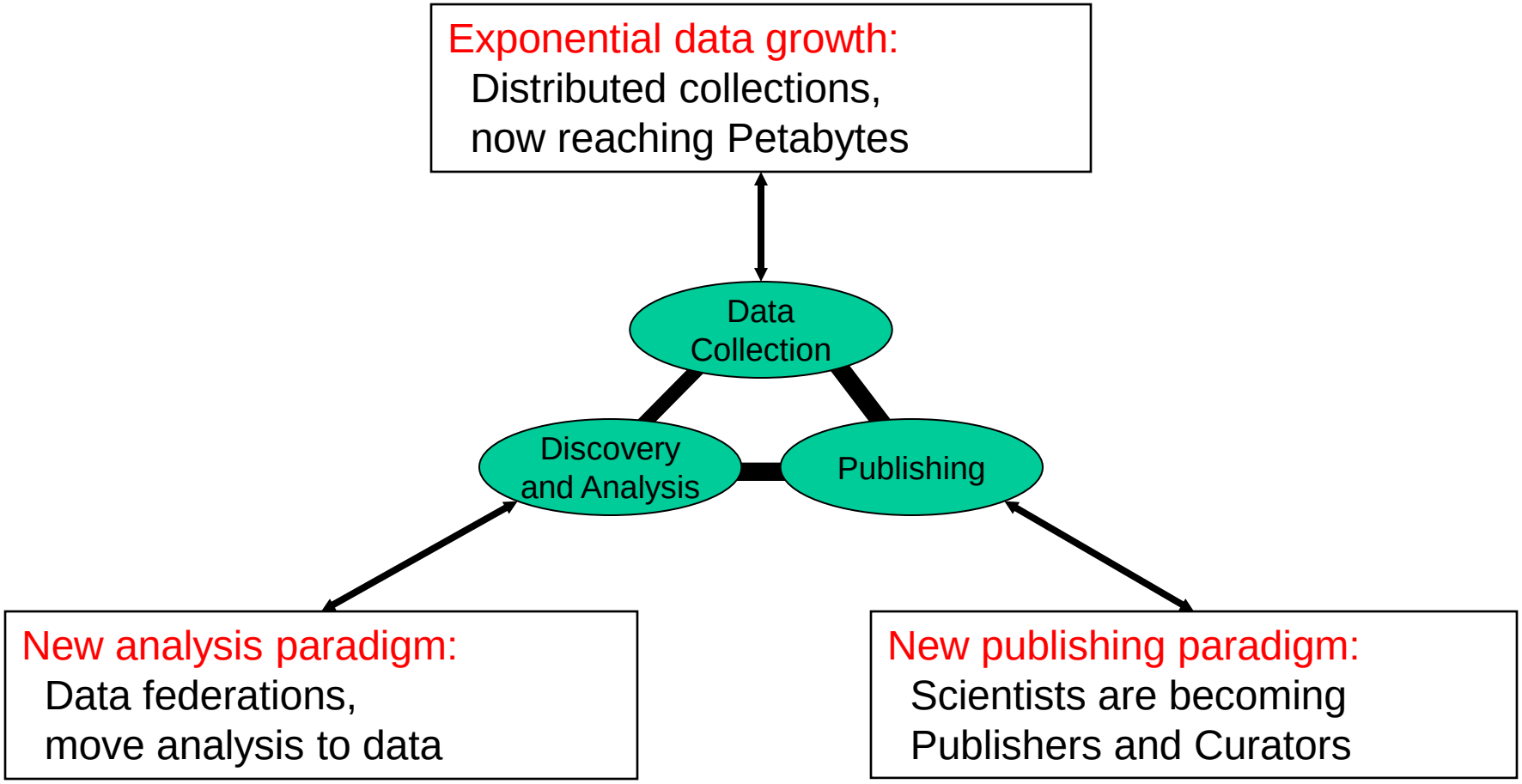
Gray's Laws of Data Engineering

Jim Gra

- Scienti around **data**
- Need s lysis
- Take t
- Start w
- Go fro



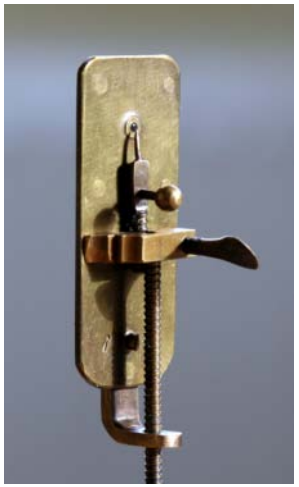
The Challenges



Non-Incremental Changes

- Multi-faceted challenges in the analysis as well
- New computational tools and strategies

  … not just statistics, not just computer science, not just astronomy, not just genomics…
- Science is moving increasingly from hypothesis-driven to data-driven discoveries

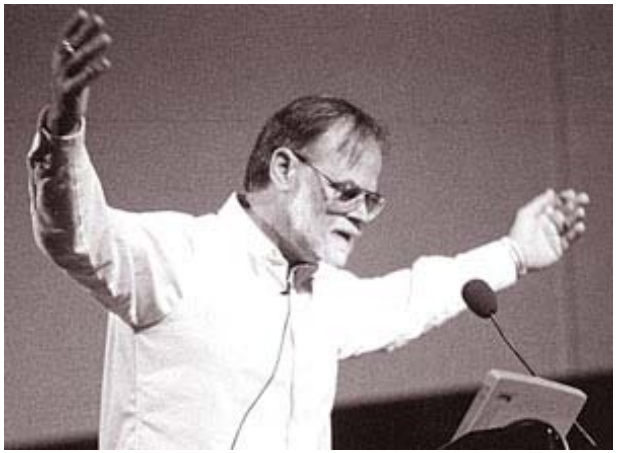


Why Is Astronomy Interesting?

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

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

— Jim Gray

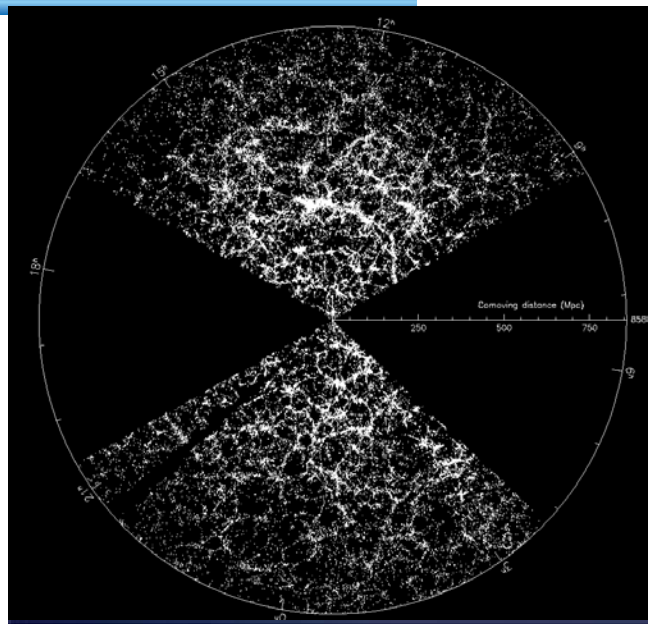


Sloan Digital Sky Survey

"The Cosmic Genome Project"

- Started in 1992, 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-3, data served from JHU



Skyserver

Prototype in 21st Century data access

- *1.2B web hits in 12 years*
- *200M external SQL queries*
- *5,000 papers and 200K citations*
- *4,000,000 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 7K astronomers*

Impact of Sky Surveys

Astronomy

Sloan Digital Sky Survey tops astronomy citation list

NASA's Sloan Digital Sky Survey (SDSS) is the most significant astronomical facility, according to an analysis of the 200 most cited papers in astronomy published in 2006. The survey, carried out by Juan Madrid from McMaster University in Canada and Duccio Macchetto from the Space Telescope Science Institute in Baltimore, puts NASA's Swift satellite in second place, with the Hubble Space Telescope in third (arXiv:0901.4552).

Madrid and Macchetto carried out their analysis by looking at the top 200 papers using NASA's Astrophysics Data System (ADS), which charts how many times each paper has been cited by other research papers. If a paper contains data taken only from one observatory or satellite, then that facility is awarded all the citations given to that article. However, if a paper is judged to contain data from different facilities – say half from SDSS and half from Swift – then both facilities are given 50% of the citations that paper received.

Top 10 telescopes

Rank	Telescope	Citations	Ranking in 2004
1	Sloan Digital Sky Survey	1892	1
2	Swift	1523	N/A
3	Hubble Space Telescope	1078	3
4	European Southern Observatory	813	2
5	Keck	572	5
6	Canada–France–Hawaii Telescope	521	N/A
7	Spitzer	469	N/A
8	Chandra	381	7
9	Boomerang	376	N/A
10	High Energy Stereoscopic System	297	N/A

The researchers then totted up all the citations and produced a top 10 ranking (see table). Way out in front with 1892 citations is the SDSS, which has been running since 2000 and uses the 2.5 m telescope at Apache Point in New Mexico to obtain images of more than a quarter of the sky. NASA's Swift satellite, which studies gamma-ray bursts, is second with 1523 citations, while the Hubble Space Telescope (1078 citations) is third.

Although the 200 most cited papers make up only 0.2% of the references indexed by the ADS for papers published in 2006, those 200 papers account for 9.5% of the citations. Madrid and Macchetto also ignored theory papers on the basis that they do not directly use any telescope data. A similar study of papers published in 2004 also puts SDSS top with 1843 citations. This time, though, the European Southern Observatory, which has telescopes in Chile, comes second with 1365 citations and the Hubble Space Telescope takes third spot with 1124 citations.

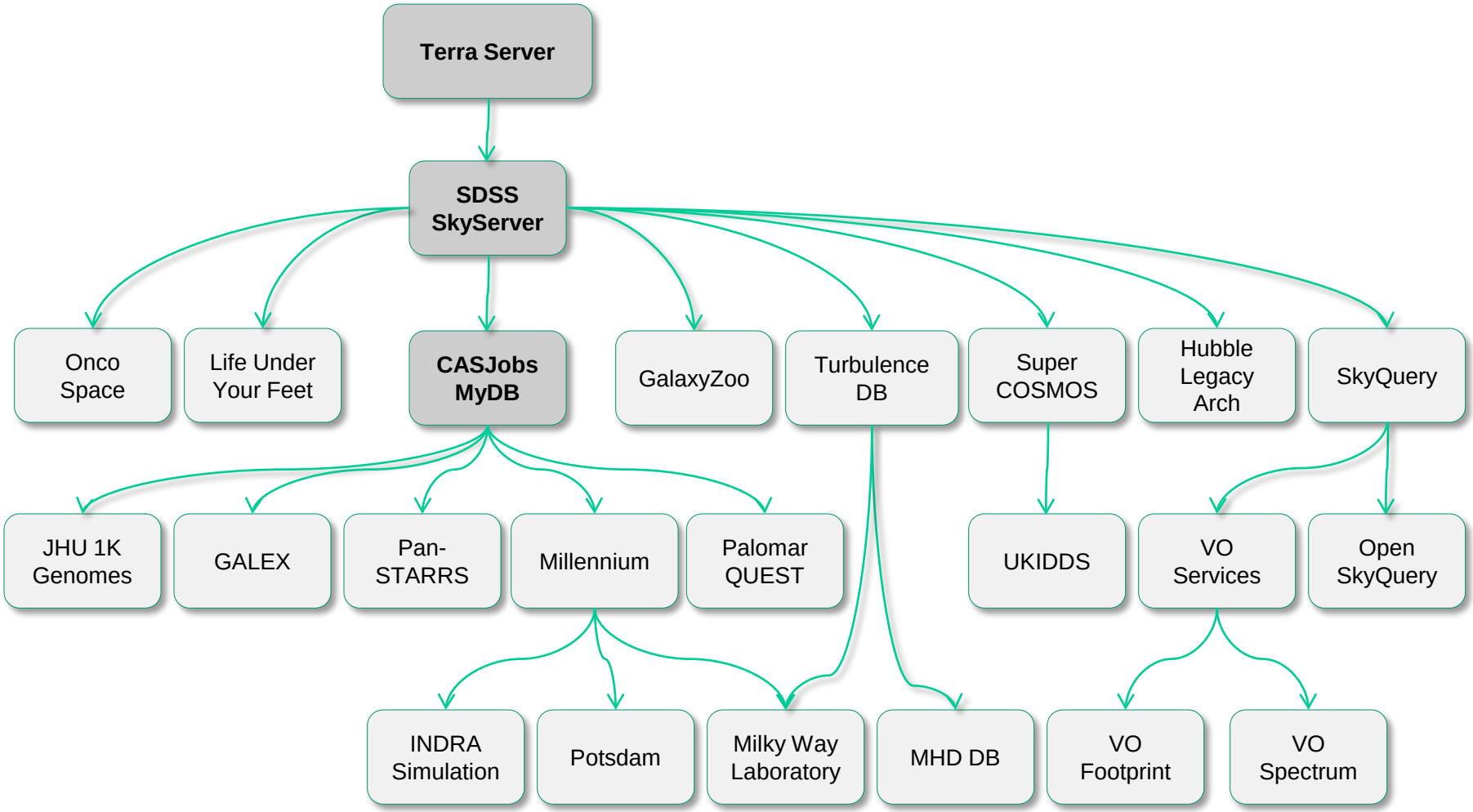
Michael Banks

GalaxyZoo

- 40 million visual galaxy classifications by the public
- Good publicity (CNN, Times, Washington Post, BBC)
- 300,000 people participating, blogs, poems…
- Original discoveries by the public
  (Voorwerp, Green Peas)

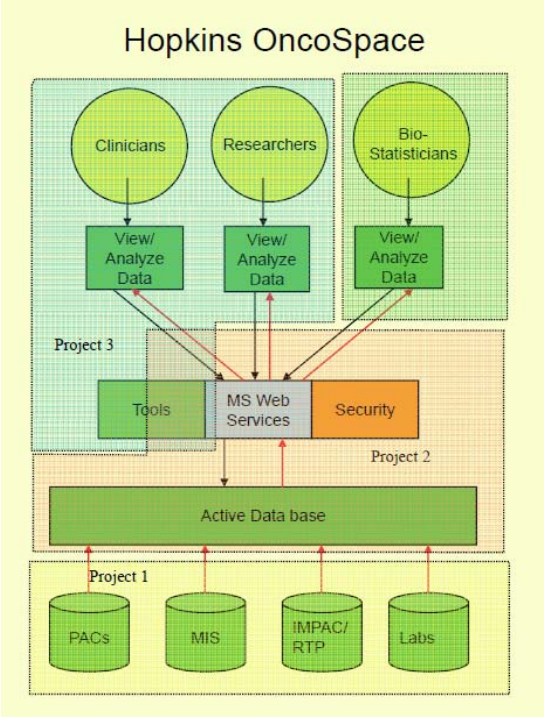
Chris Lintott et al

The SDSS Genealogy



Oncospace

Todd McNutt, John Wong, JHU Radiation Oncology



OncoSpace: Adapting the SkyServer Approach

- ***Active Databases***
- There is too much data to move around,
  take the analysis to the data!
- Do all data manipulations at database
  - ***Build custom procedures and functions in the database***
- Established Web-service for broad access
  - Query across multiple databases
- Automatic parallelism guaranteed

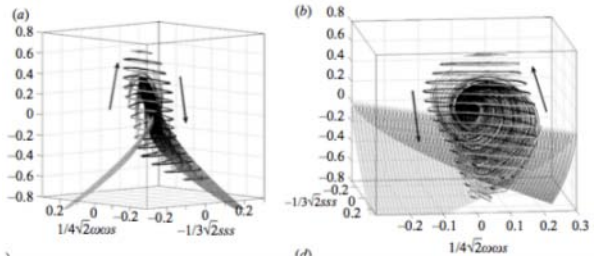
Data in HPC Simulations

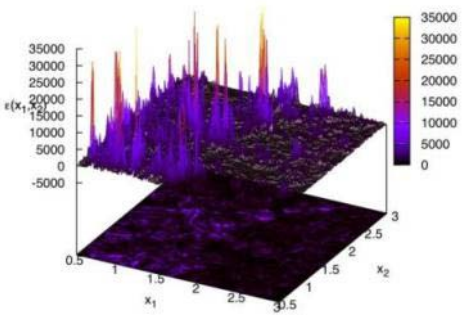
- HPC is an instrument in its own right
- Largest simulations approach petabytes
  - *from supernovae to turbulence, biology and brain modeling*
- Need public access to the best and latest through interactive numerical laboratories
- Creates new challenges in
  - *How to move the petabytes of data (high speed networking)*
  - *How to look at it (render on top of the data, drive remotely)*
  - *How to interface (smart sensors, immersive analysis)*
  - *How to analyze (value added services, analytics, …)*
  - *Architectures (supercomputers, DB servers, ??)*

Immersive Turbulence

"… the last unsolved problem of classical physics…"

- **Understand the nature of turbulence**
  - *Consecutive snapshots of a large simulation of turbulence: 1024^4 => 30 Terabytes*
  - *Treat it as an experiment, **play** with the database!*
  - ***Shoot test particles** (sensors) from your laptop into the simulation, like in the movie Twister*

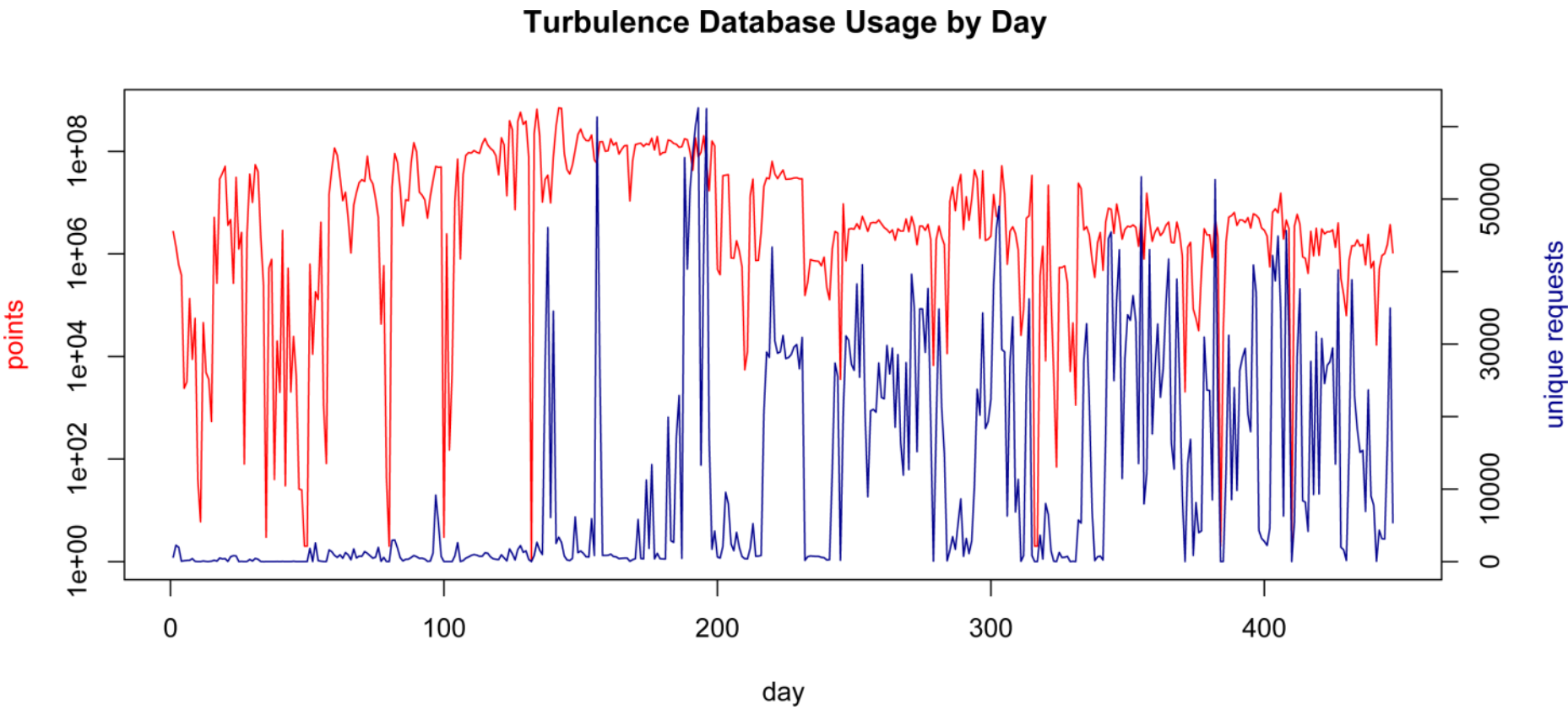




- **New paradigm** for analyzing simulations!

with C. Meneveau, S. Chen (Mech. E), G. Eyink (Applied Math), R. Burns (CS)

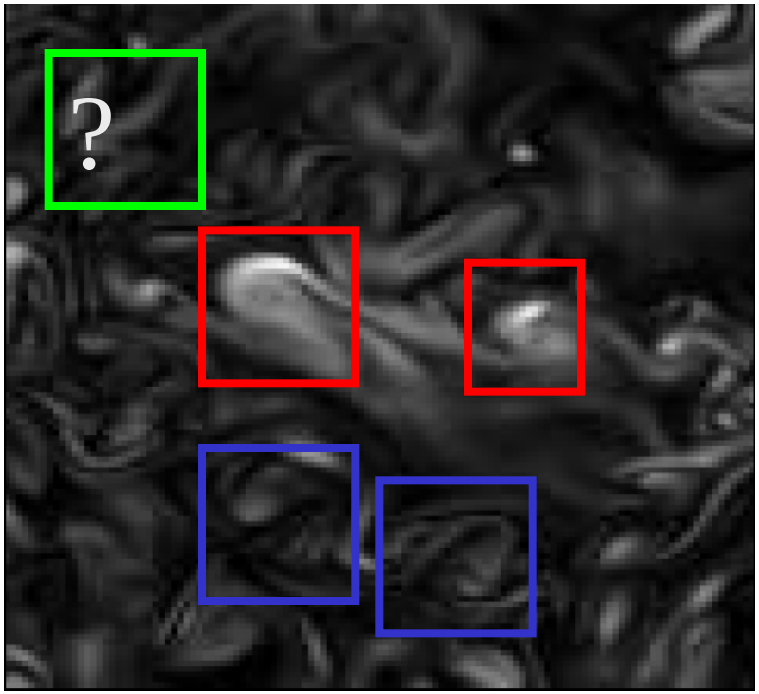
Daily Usage



More than 12 trillion “sensors” delivered to the community

Applications of ML to Turbulence

Renyi divergence



Vorticity

Similarity between regions ⇒

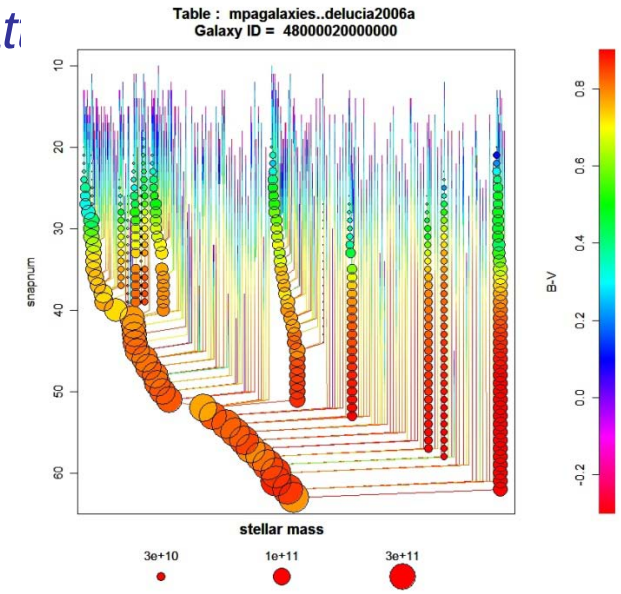
- clustering,
- classification,
- anomaly detection

J. Schneider, B. Poczos, CMU

Cosmology Simulations

- Millennium DB is the poster child/ success story
  - *Built by Gerard Lemson (now at JHU)*
  - *600 registered users, 17.3M queries, 287B rows*

    http://gavo.mpa-garching.mpg.de/Millennium/
  - *Dec 2012 Workshop at MPA: 3 days, 50 people*
- Data size and scalability
  - *PB data sizes, trillion particles of dark mat*
- Value added services
  - *Localized*
  - *Rendering*
  - *Global analytics*



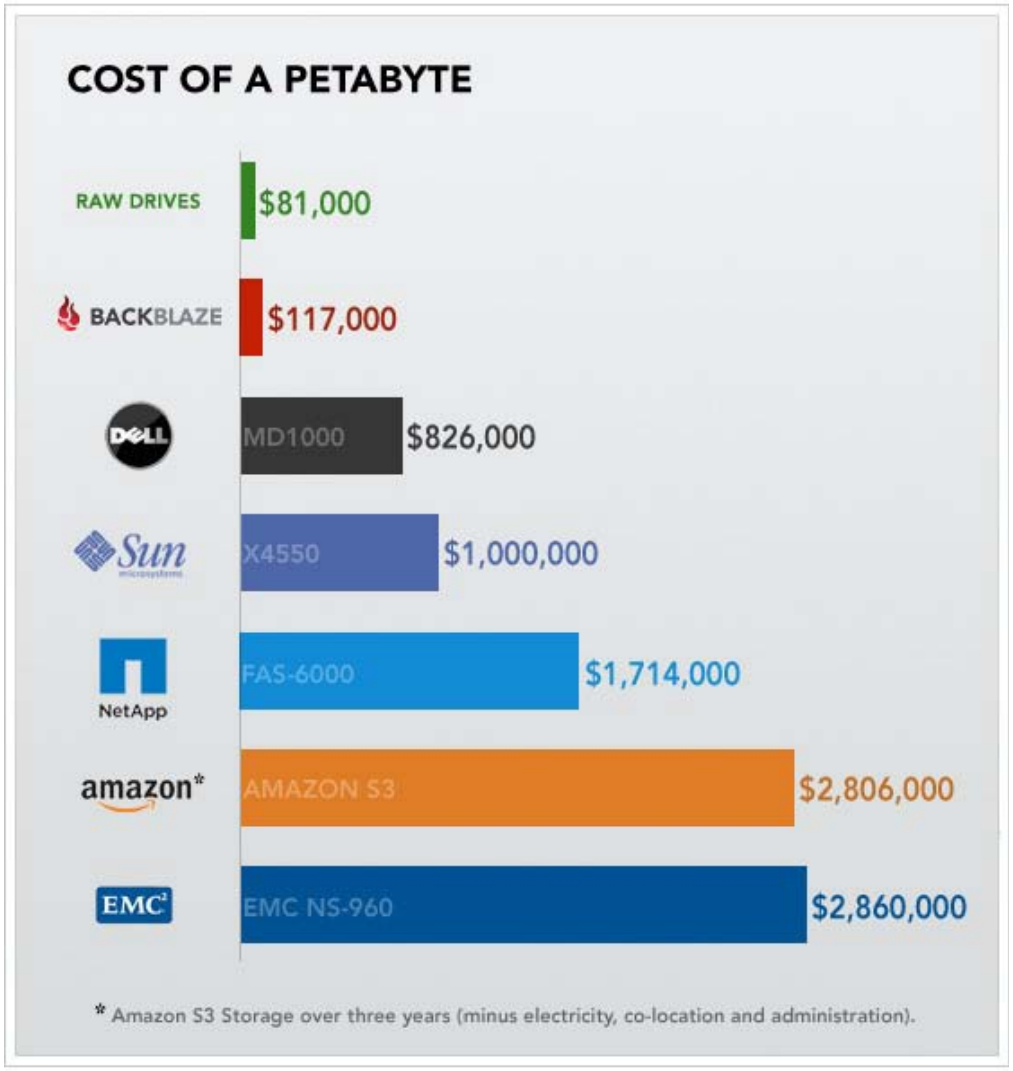
Simulations in the DB

- Amazing progress in 7 years
- Millennium and turbulence are showcases
- Community is now using the DB as a musical instrument
- New challenges emerging:
  - *Petabytes of data, trillions of particles*
  - *Increasingly sophisticated value added services*
  - *Need a coherent strategy to go to the next level*
- Not just storage, but integrate access and computation
- Filling the gap between DB server and supercomputer

Scalable Data-Intensive Analysis

- Large data sets => data resides on hard disks
- Analysis has to move to the data
- Hard disks are becoming sequential devices
  - *For a PB data set you cannot use a random access pattern*
- Both analysis and visualization become streaming problems
- Same thing is true with searches
  - *Massively parallel sequential crawlers (MR, Hadoop, etc)*
- Spatial indexing needs to be maximally sequential
  - *Space filling curves (Peano-Hilbert, Morton,…)*

Cost of a Petabyte



From backblaze.com
Aug 2009

Result: we are still building our own!

JHU Data-Scope

- Funded by NSF MRI to build a new ‘instrument’ to look at data
- Goal: ~100 servers for $1M + about $200K switches+racks
- Two-tier: performance (P) and storage (S)
- Large (6.5PB) + cheap + fast (500GBps), but …
  . ..a special purpose instrument
- 100G connectivity to the outside world

	o					
	1P	*1S*	*All P*	*All S*	*Full*	
servers	1	1	90	6	102	
rack units	4	34	360	204	564	
capacity	24	720	2160	4320	6480	TB
price	8.8	57	8.8	57	792	$K
power	1.4	10	126	60	186	kW
GPU*	1.35	0	121.5	0	122	TF
seq IO	5.3	3.8	477	23	500	GBps
IOPS	240	54	21600	324	21924	kIOPS
netwk bw	10	20	900	240	1140	Gbps

The Long Tail

- The “Long Tail” of a huge number of small data sets
  - *The integral of the “long tail” is big!*
- 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*
  - *Interface is open, public*
- SciDrive: JHU project (funded by the Sloan Foundation)
  - *Enable people to drag and drop (and share) their data*
  - *No metadata required*
  - *We are trying to figure it out from the data itself + papers*

SciDrive

From Stars to Genes

- Strong parallels between genomics today and astronomy 20 years ago
- All based on files and manually ran scripts
- Metadata stored in file headers and long filenames
- Everybody is obsessed with running their own aligners
- This works with a small number of genomes, but will break with thousands or millions
- Astronomy lessons:
  - *for statistical processing and collaboration you need a DB, not flat files*
  - *Find a common processing that is “good enough”*

Technology+Sociology+Economics

- Technology is changing very rapidly
  - *Google, tags, sensors, Moore's Law*
  - *Trend driven by changing generations of technologies*
- Sociology is changing in unpredictable ways
  - *YouTube vs. MySpace*
  - *In general, people will use a new technology if it is*
    - Offers something entirely new
    - Or substantially cheaper
    - Or substantially simpler
    - Or all our friends are using it
- Funding is essentially level

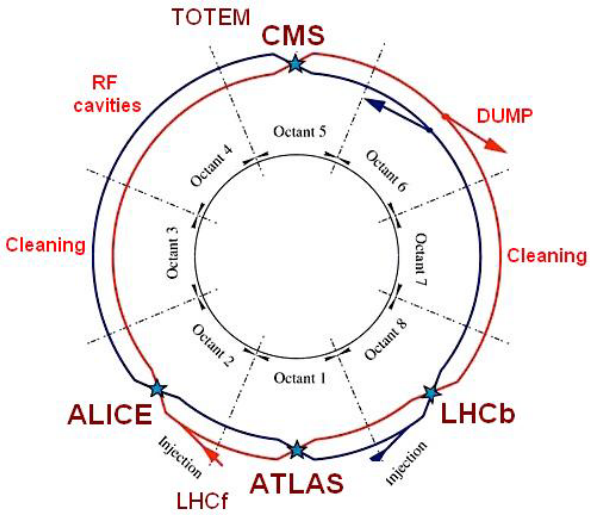
How Do We Prioritize?

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

- “Do I have enough data or would I like to have more?”
- No scientist ever wanted less data….
- But: Big Data is synonymous with Dirty Data
- How can we decide how to collect data that is ***more relevant*** ?

LHC Lesson

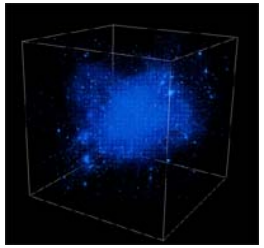
- LHC has a single data source, $$$$$
- Multiple experiments tap into the beamlines
- They each use **in-situ** hardware triggers to filter data
  - *Only 1 in 10M events are stored*
  - *Not that the rest is garbage, just sparsely sampled*



- Resulting “small subset” analyzed many times **off-line**
  - *This is still 10-100 PBs*
- Keeps a whole community busy for a decade or more

Trends

- Broad sociological changes
  - *Convergence of Physical and Life Sciences*
  - *Data collection in ever larger collaborations*
  - *Virtual Observatories: CERN, IVOA, NCBI, NEON, OOI,…*
  - *Analysis decoupled, off archived data by smaller groups*
  - *Emergence of the citizen/internet scientist (GalaxyZoo…)*
- Need to start training the next generations
  - *Π-shaped vs I- and T-shaped people*
  - *Early involvement in “Computational thinking”*



Summary

- Science is increasingly driven by data (big and small)
- Changing sociology – surveys analyzed by individuals
- From hypothesis-driven to data-driven science
- We need new instruments: “microscopes” and “telescopes” for data
- There is a major challenge on the “long tail”
- A new, Fourth Paradigm of Science is emerging…
- SDSS has been at the cusp of this transition

“*If I had asked people what they wanted, they would have said faster horses…*”

—*Henry Ford*