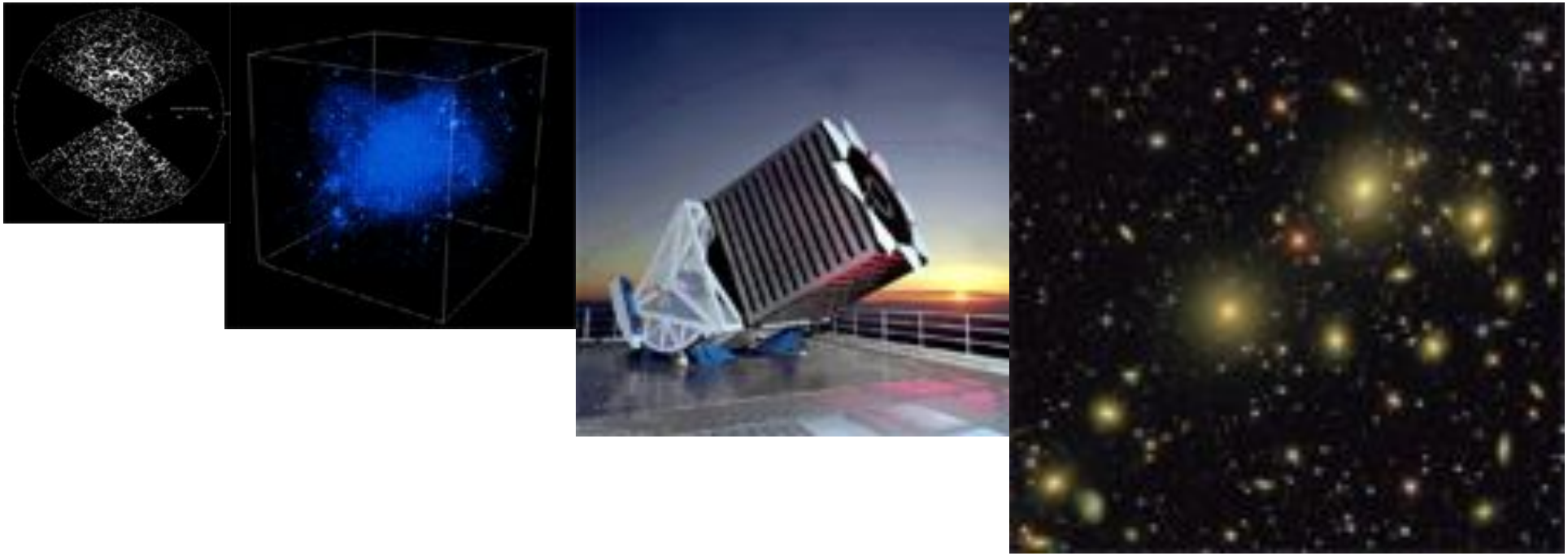


Data Driven Discovery in Science: The Fourth Paradigm

Alex Szalay
The Johns Hopkins University

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
- 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

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Search | Global | DealBook | Markets | Economy | Energy

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

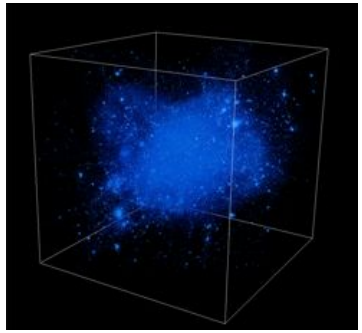
Science is Changing

THOUSAND YEARS AGO
science was empirical
describing natural phenomena

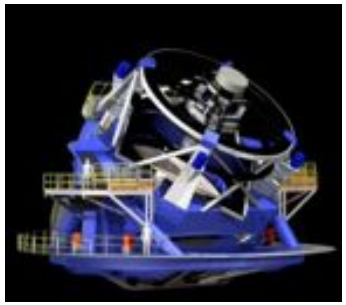
LAST FEW HUNDRED YEARS
theoretical branch using models, generalizations

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

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!



Non-Incremental Changes

- Multi-faceted challenges
- New computational tools and strategies

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

Astronomy has always been data-driven…. now becoming more generally accepted

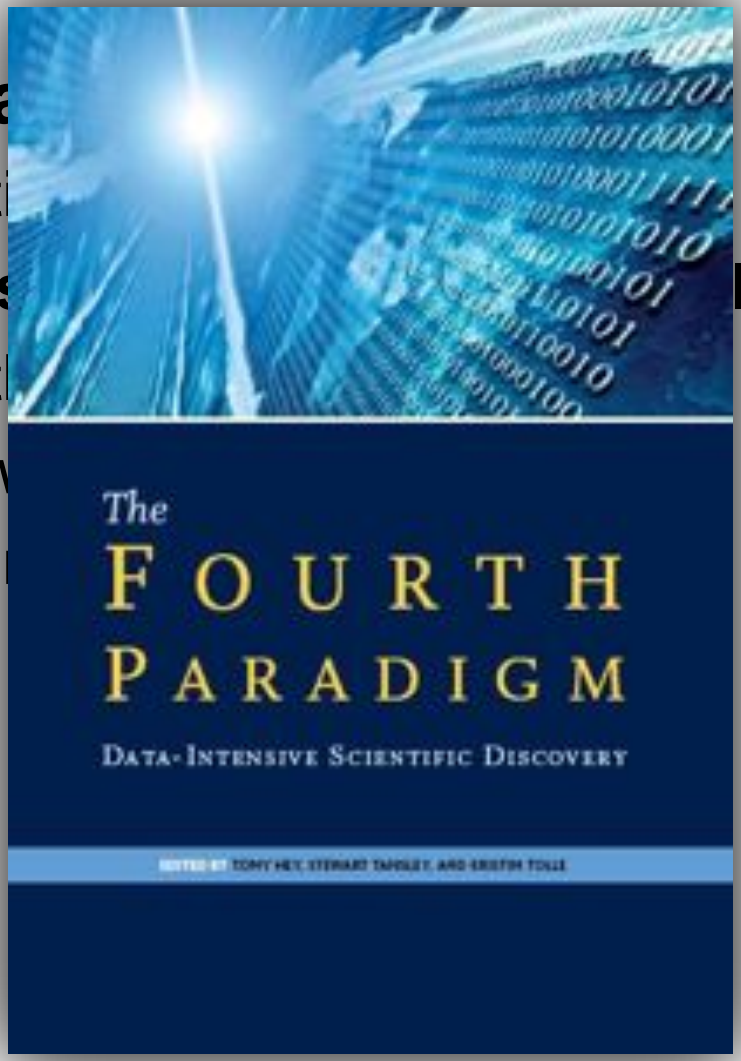
Scientific Data Analysis Today

- Scientific data is doubling every year, reaching PBs
- Not only Big Data, small data growing as well
- Data is everywhere, never will be at a single location
- Architectures increasingly CPU-heavy, IO-poor
- Scientists spend 90% of time on data management
- Most data analysis on small/midsize BeoWulf clusters
- Universities hitting the “power wall”
- Soon we cannot even store the incoming data stream
- **Not scalable, not maintainable…**

Gray's Laws of Data Engineering

Jim Gra

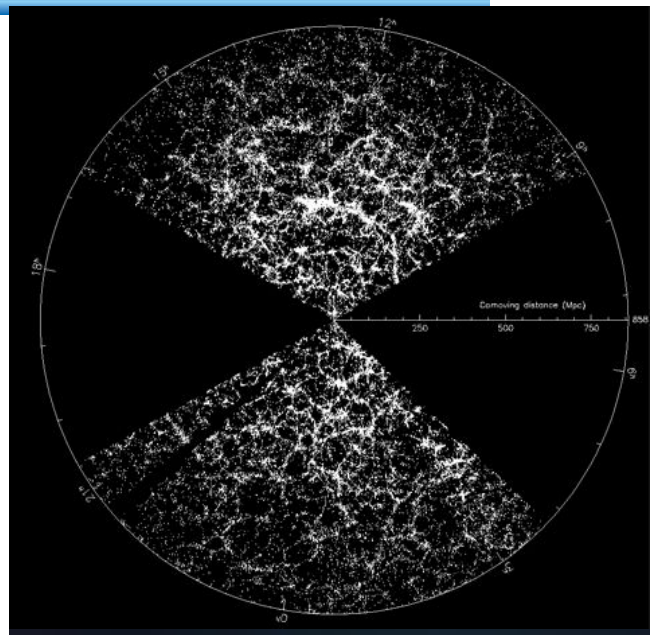
- Scienti around **data**
- Need s lysis
- Take t
- Start w
- Go fro



Sloan Digital Sky Survey

- "**The Cosmic Genome Project**"
- Two surveys in one
  - Photometric survey in 5 bands
  - Spectroscopic redshift survey
- Data is public
  - 2.5 Terapixels of images => 5 Tpx
  - 10 TB of raw data => 120TB processed
  - 0.5 TB catalogs => 35TB in the end
- Started in 1992, finished in 2008
- Extra data volume enabled by
  - *Moore's Law, Kryder's Law*

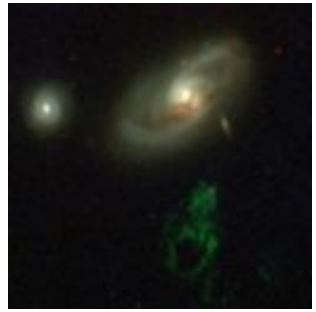
Partnership of Sloan Foundation, NSF, DOE, NASA, universities



Skyserver

- Prototype of 21st Century discovery engine
  - *990 million web hits in 10 years*
  - *The world's most used astronomy facility today*
  - *1,000,000 distinct users vs. 15,000 astronomers*
  - *The emergence of the "Internet scientist"*

- GalaxyZoo (Lintott et al)
  - *40 million visual galaxy classifications by the public*
  - *Enormous publicity (CNN, Times, Washington Post, BBC)*
  - *300,000 people participating, blogs, poems…*
  - *Amazing original discoveries (Voorwerp, Green Peas)*



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

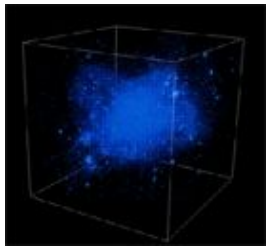
The researchers then totted up all the citations and produced a top 10 ranking (see table). Way out in front with 1892 citations is the SDSS, which has been running since 2000 and uses the 2.5 m telescope at Apache Point in New Mexico to obtain images of more than a quarter of the sky. NASA's Swift satellite,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

Sociology

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



Data in HPC

- Not only experimental data, HPC is a new instrument
- Largest simulations approach petabytes
  - *from supernovae to turbulence, systems biology and brain modeling*
- Need public access to the best and latest through interactive numerical laboratories
- 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 (virtual sensors, immersive analysis)*
  - *How to analyze (algorithms, scalable analytics)*

Summary

- Science is increasingly driven by data (large and small)
- Large data sets are here, solutions are not
- Changing sociology
- From hypothesis-driven to data-driven science
- We need new instruments: “microscopes” and “telescopes” for data
- There is also a problem on the “long tail”
- Same problems present in business and society
- Data changes the not only science, but society
- A new, Fourth Paradigm of Science is emerging…

A convergence of statistics, computer science, physical and life sciences…..