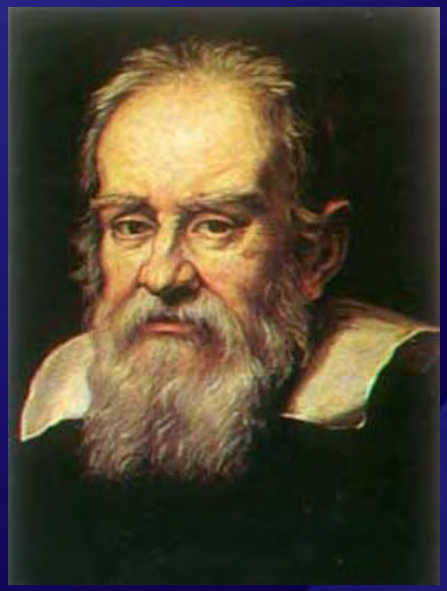
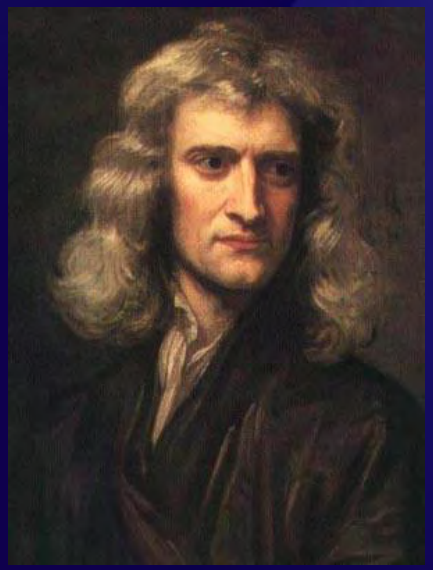
NSF Perspective on the Data Deluge

Philip Bogden
Program Director
NSF Office of Cyberinfrastructure

Profound Transformation of Science
Gravitational Physics



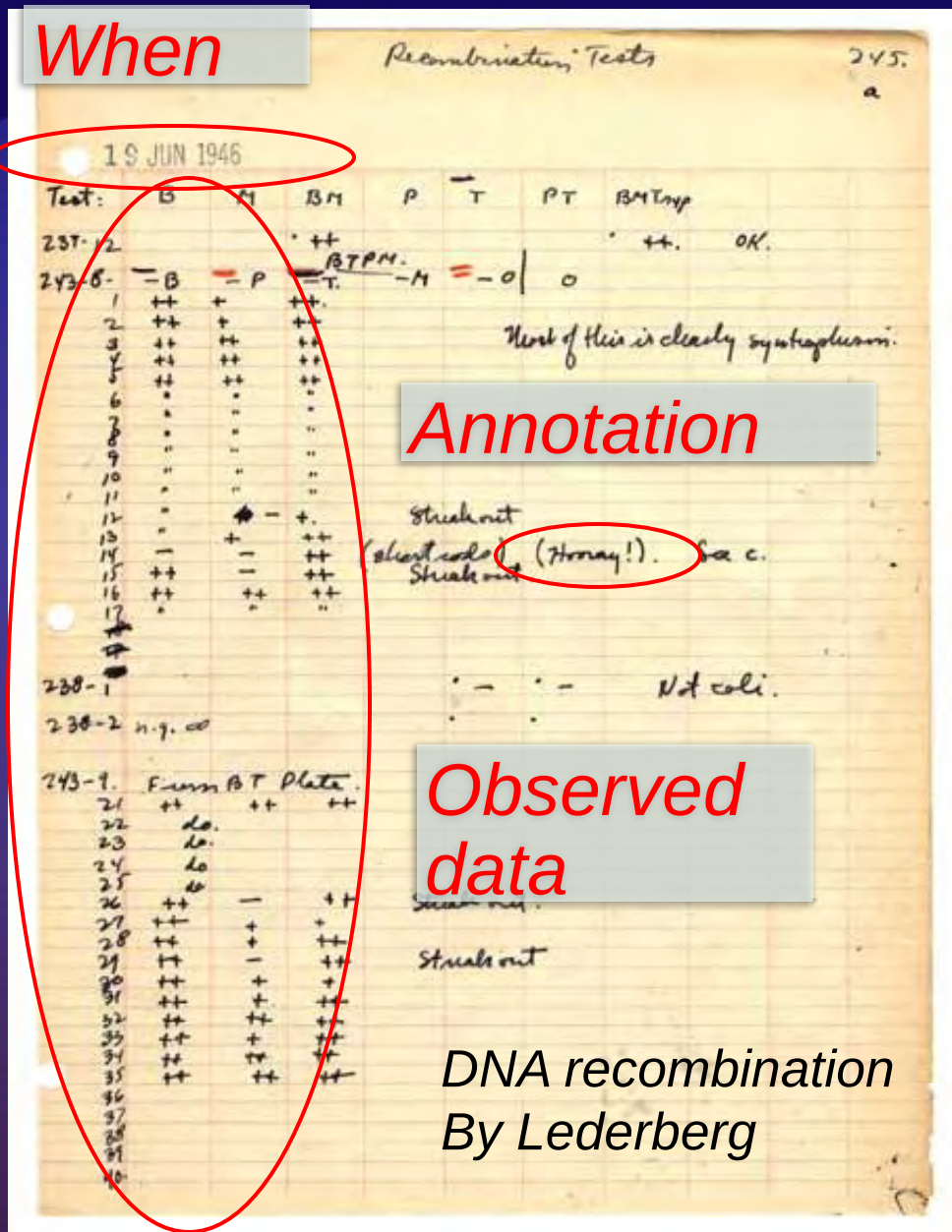


- Galileo, Newton usher in birth of modern science: c. 1600
- Problem: single "particle" (apple) in gravitational field (General 2 body-problem already too hard)
- Methods
  - Data: notebooks (Kbytes)
  - Theory: driven by data
  - Computation: calculus by hand (1 Flop/s)
- Collaboration
  - 1 brilliant scientist, 1-2 student

Source: Juliana Freire, U of Utah

Provenance in Science

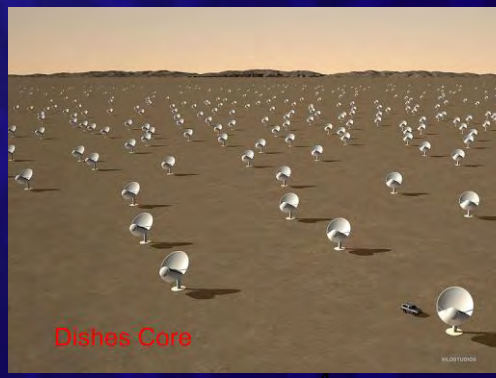
- *Provenance is as important as the result*
- Not a new issue
- Lab notebooks have been used for a long time
- What is new?
  - Large volumes of data
  - Complex analyses—computational processes
- Writing notes is no longer an option...
- *GC Communities require open, sharable data, standards, metadata*



Explosive Trends in Data Growth

- Comparative Metagenomics
  - DNA sequencing of entire families of organisms
  - Already hundreds of TB, thousands of users
- HD Collaborations and Optiportals
  - Multichannel HD, gigapixel visualizations
- Petascale-Exascale simulation
  - They generate peta-exabytes per simulation!
- Square Kilometer Array
  - 3000 radio receivers, 1 km^2 area!
  - 19 countries! Possibly beginning in 201X, operational 202X
  - Data: exabyte per week! Analysis: Exaflops!

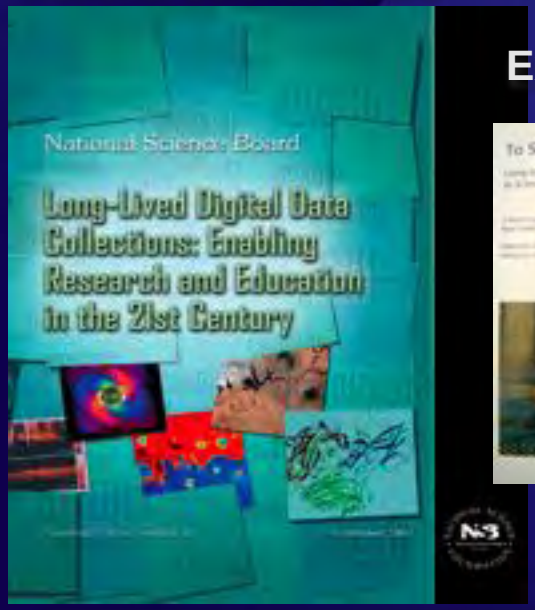
camera
Marine Microbial Ecology



Data Crisis: Information Big Bang

PCAST Digital Data

National Science Board



Expert Study

Industry

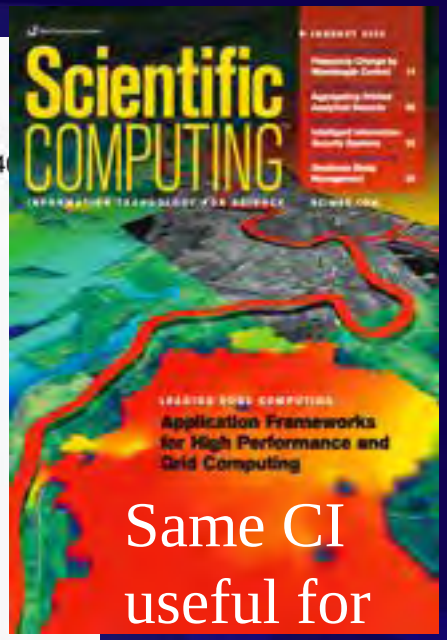
Storage Networking Industry Association (SNIA) *100 Year Archive Requirements Survey Report*

"there is a pending crisis in archiving… we have to create long-term methods for preserving information, for making it available for analysis in the future." 80% respondents: >50 yrs; 68% > 100 yrs

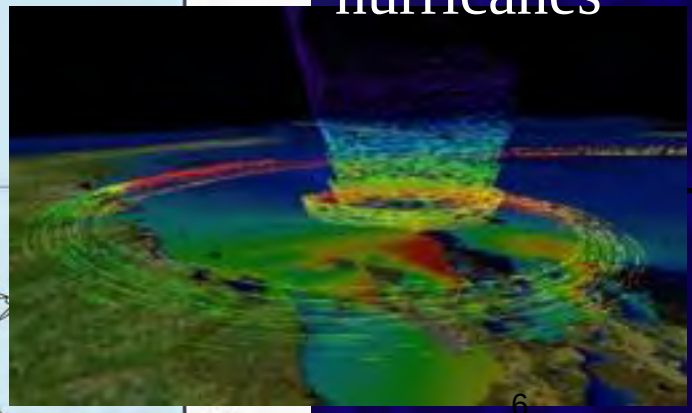
Wired, Nature

Grand Challenge Communities Combine it All...

Where is it going to go?



Same CI useful for black holes, hurricanes



The Shift Towards Data
Implications

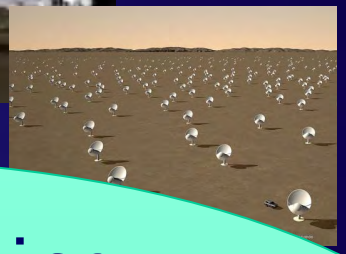
- All science is becoming data-dominated
  - Experiment, computation, theory
- Totally new methodologies
  - Algorithms, mathematics
  - All disciplines from science and engineering to arts and humanities
- End-to-end networking becomes critical part of the ecosystem
- How do we train "data-intensive" scientists?
- Data policy becomes critical!

NSF Vision & National Cyberinfrastructure Blueprint

Education Crisis: I need all of this to start to solve my problem!

neon
National Ecological Observatory Network, Inc.

Nets



DataNet

DataNet

Track 2

Science is becoming unreproducible in this environment. Validation?Provenance? Reproducibility?

Campus

Campus

Campus

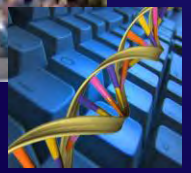
Campus

Campus

Campus

Campus

Campus



"Data Deluge" Drives Change at NSF

- Data Issues resonate the mo
- DataNet – $100M inves
  Access Partners – dev
  network of interactiv
  grand challenges, in
- INTEROP – Communit
  interdisciplinary, commu
  re-use data in ways not e
- Data-intensive computin SDSC Gordon facility
- NSF Data Policy – The "Data Working Group" (NSF-wide group of Program Directors) working to assure that data are shared within and across disciplines

There is a major shift in science towards data-intensive methods. NSF is responding...

Recent NSF Activities on Data Policy and Implementation

Changes Coming for Data!

- Long-standing NSF Policy on Data (Proposal & Award Policies & Procedures Guide)
  - "*Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data... created or gathered in the course of work under NSF grants*"
- NSF will soon require a Data Management Plan (DMP), subject to peer review; criterion for award
  - The DMP will be in the form of a 2-page supplementary document to the proposal
  - It will not be possible to submit proposals without such a document
  - Customization by discipline, program necessary

Upcoming Implementation of NSF Data Policy

- Directorate-Specific Issues & Peer Review
- Many details are implemented/enforced via peer review and Program Officer discretion, including things like embargo period, standards, etc.
- The challenge at NSF is that "no one size fits all" so each Directorate will be responsible for its own recommendations for DMP content, appropriate institutional repositories, etc.
- This does not address Open Access as applied to publications

Fundamental poi
ublicat

Where is it placed? Author web site? Library? NSF sites?

Who pays? The NSF? The Institution? What is the cost model? What is reasonable?

ntifi
and s
pla
it

What data must be made available? Raw
ta? Peer reviewed?

How long is it made
? 6
fter

There is great variability in requirements across science communities: peer review can help guide this process.

Closing Remarks

This is Step 1

National Science Board has a newly formed Task Force on Data Policy

Discussions within NSF acknowledge that no one-size fits all, focus on incentives, consider software and publications (knowledge)