Date: Thu, 07 Oct 2010 13:22:54 -0400
To: "Dr. Baruch Fischhoff - Chair, National Academy Committee on Improving Intelligence" <baruch@cmu.edu>, "Dr. Richard Atkinson - Chair - NRC/DBASSE" <rcatkinson@ucsd.edu>, "Dr. Kenneth Prewitt" <kp2058@columbia.edu>
From: Lloyd Etheredge <lloyd.etheredge@policyscience.net>

Subject: 162. Wolfram 5: Data-Empowered, Rapid-Learning Behavioral Science. Cross-disciplinary comparisons. The Global Content Analysis System (GCAS)

Dr. Fischhoff and Colleagues:

A New Model of Rapid-Learning Science

A new model of rapid-learning science, developed in the biomedical field, is RL = R + C + D, where R (great researchers) are combined with C (very fast computers) and - now - D (databases - a national system of high quality, pre-designed, pre-built, and pre-populated databases for RL).

You might want to recommend the same model to the DNI and related organizations (e.g., NSF) with responsibility for, or potential benefits from, rapid-learning behavioral science. [Global content analysis, to take one example.] In this model of Rapid-Learning science, it is the responsibility of government, in partnership, to develop the large, high quality, pre-designed, pre-built, and pre-populated online databases.

The BioInformatics Comparison

- Behavioral sciences continue to lag. For example the new NIH-Kaiser Permanente-RWJF national biobank/bio-repository is underway to have 500,000 patient EHR's (electronic health records) with full genetic and environmental data and biospecimens for future analysis as theoretical and measurement progress pose new questions. There will be a national system of "gap-filling" registries and networks for sub-populations usually not included in clinical trials: children, pregnant women, seniors, minorities, persons with multiple chronic conditions, rare diseases, surgery, etc. There also will be a focused, specific, RL system for cancer building on the NCI's Biomedical Informatics Grid. There are new generations of fast-discovery software for the new "in silico" R&D. The Obama Administration also has committed $ 1.1 billion for comparative effectiveness research, and plans national electronic health record system ($40+ billion) and an FDA Sentinel Network with 100 M electronic patient records, etc.

The GCAS Project

This is the first truly rational, information-age, system for rapid learning science that I have seen.<1> Major pieces also are available to come together for the equivalent Global Content Analysis System, rapid-learning capability, if you will recommend it and the DNI will migrate major components of the system into the public domain. BBC ontology systems (# 160) also are available for public affairs/events and other data can be linked via DALLAS (for example).

AstroInformatics

There are some people who believe that the GCAS capability is too big a data project. But this is a technical assessment/calibration of the imagination from the 20th century. You might be interested in the attached Wolfram presentation re the data systems that are becoming available for fast discovery 21st century astronomy, with federal funding. [I.e., and without national security implications.] The new Pan-STARRS system at 40 Petabytes is just starting and the next Large Synoptic Survey Telescope will generate 100 Petabytes for real-time analysis across a decade.

best regards,
Lloyd E.

<1> A brief disclaimer: Lynn Etheredge directed the project that launched the concept in the special issue of Health Affairs in January 2007. Institutionally, the RWJ Foundation, the Institute of Medicine, NIH, and many others have provided leadership to build the system; and the Obama Administration has provided funding.

Dr. Lloyd S. Etheredge - Director, Government Learning Project
Policy Sciences Center
URL: www.policyscience.net
301-365-5241 (v); lloyd.etheredge@policyscience.net
lloyd.etheredge@aya.yale.edu(email)

[The Policy Sciences Center, Inc. is a public foundation that develops and integrates knowledge and practice to advance human dignity. Its headquarters are 127 Wall St., Room 322 PO Box 208215 in New Haven, CT 06520-8215. It may be contacted at the office of its Chair, Michael Reisman (michael.reisman@yale.edu), 203-432-1993. Further information about the Policy Sciences Center and its projects, Society, and journal is available at www.policysciences.org.]

The GCAS Project

This is the first truly rational, information-age, system for rapid learning science that I have seen.<1> Major pieces also are available to come together for the equivalent Global Content Analysis System, rapid-learning capability, if you will recommend it and the DNI will migrate major components of the system into the public domain. BBC ontology systems (# 160) also are available for public affairs/events and other data can be linked via DALLAS (for example).

AstroInformatics

There are some people who believe that the GCAS capability is too big a data project. But this is a technical assessment/calibration of the imagination from the 20th century. You might be interested in the attached Wolfram presentation re the data systems that are becoming available for fast discovery 21st century astronomy, with federal funding. [I.e., and without national security implications.] The new Pan-STARRS system at 40 Petabytes is just starting and the next Large Synoptic Survey Telescope will generate 100 Petabytes for real-time analysis across a decade.

best regards,
Lloyd E.

<1> A brief disclaimer: Lynn Etheredge directed the project that launched the concept in the special issue of Health Affairs in January 2007. Institutionally, the RWJ Foundation, the Institute of Medicine, NIH, and many others have provided leadership to build the system; and the Obama Administration has provided funding.

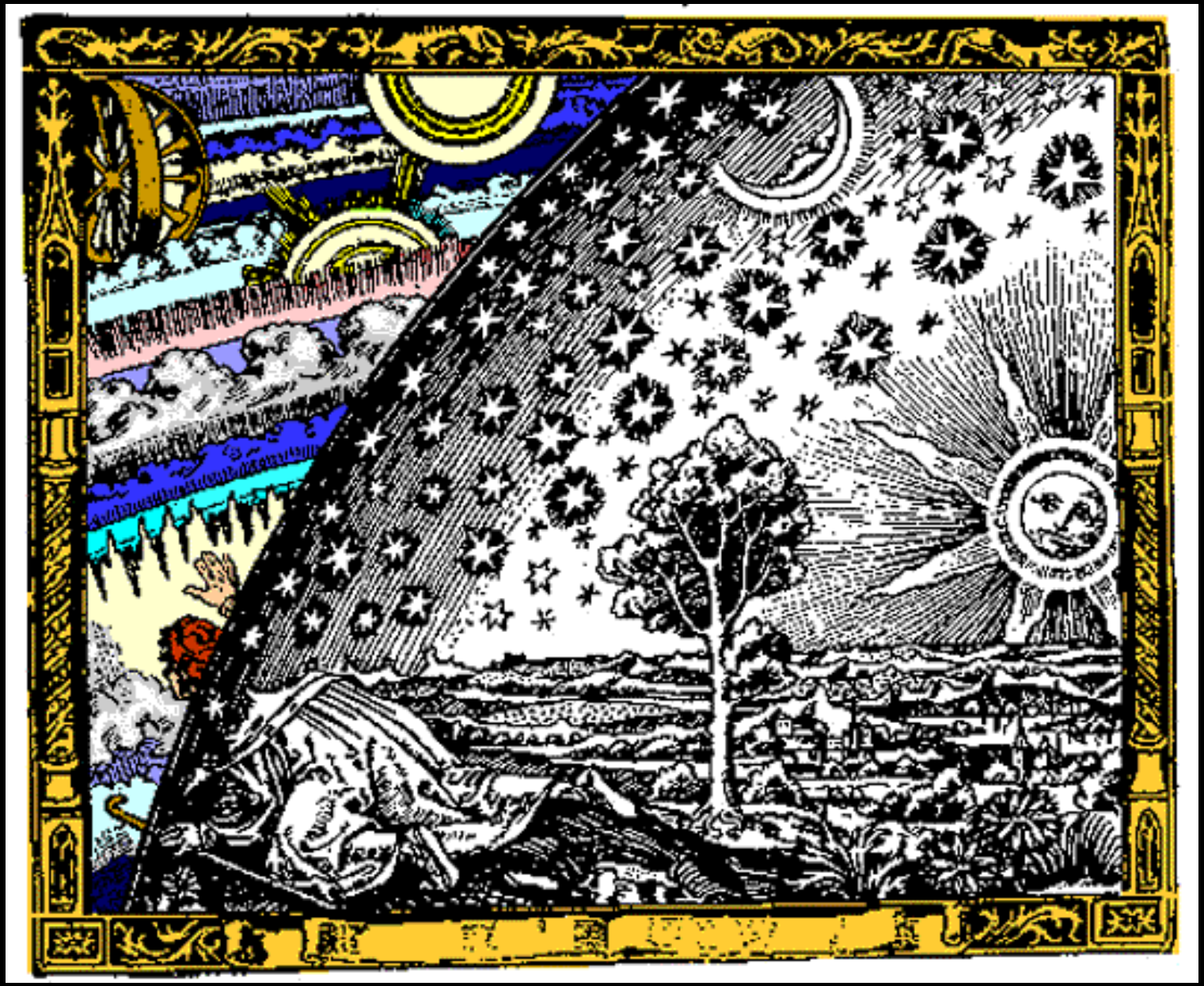
Dr. Lloyd S. Etheredge - Director, Government Learning Project
Policy Sciences Center
URL: www.policyscience.net
301-365-5241 (v); lloyd.etheredge@policyscience.net
lloyd.etheredge@aya.yale.edu(email)

[The Policy Sciences Center, Inc. is a public foundation that develops and integrates knowledge and practice to advance human dignity. Its headquarters are 127 Wall St., Room 322 PO Box 208215 in New Haven, CT 06520-8215. It may be contacted at the office of its Chair, Michael Reisman (michael.reisman@yale.edu), 203-432-1993. Further information about the Policy Sciences Center and its projects, Society, and journal is available at www.policysciences.org.]

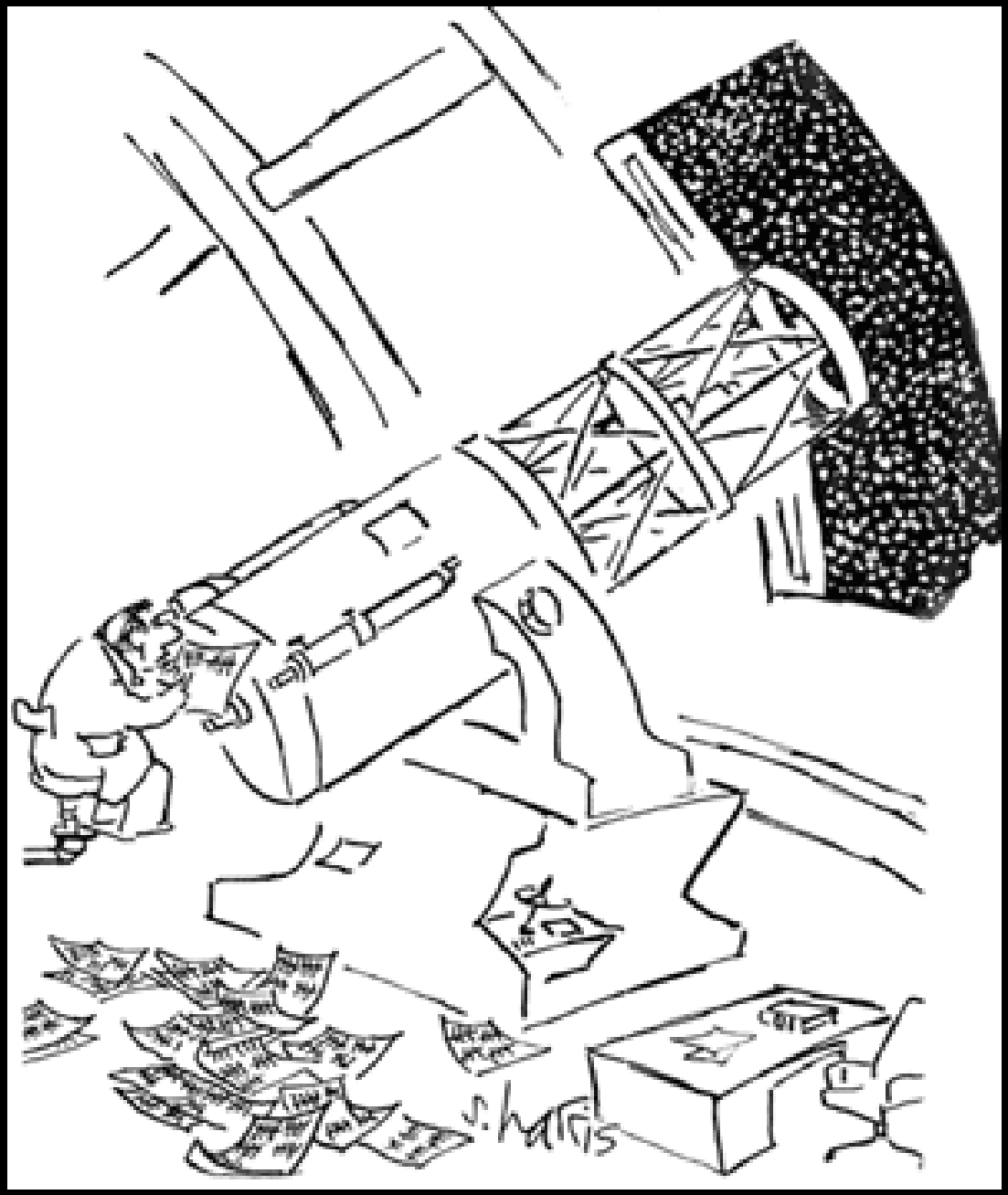
Astroinformatics: massive data research in Astronomy

Kirk Borne
Dept of Computational & Data Sciences
George Mason University
kborne@gmu.edu , http://classweb.gmu.edu/kborne/

Ever since humans first gazed into the heavens …



… Astronomy has been a Data-Driven Science



From Data-Driven to Data-Intensive

- Astronomy has always been a data-driven science
- It is now a data-intensive science
- It will become even more data-intensive in the coming decade(s) ….
- Hence, a new field of data-oriented research and education in Astronomy is emerging:

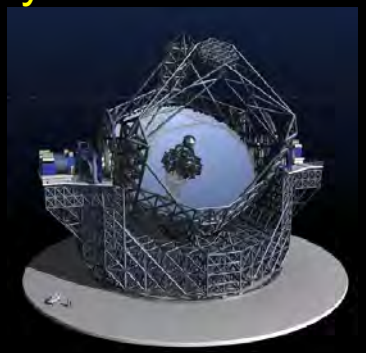
Astroinformatics

Vermeer's "Astronomer & Geographer"

(2 mappers, collaborating on Astroinformatics and Geoinformatics)

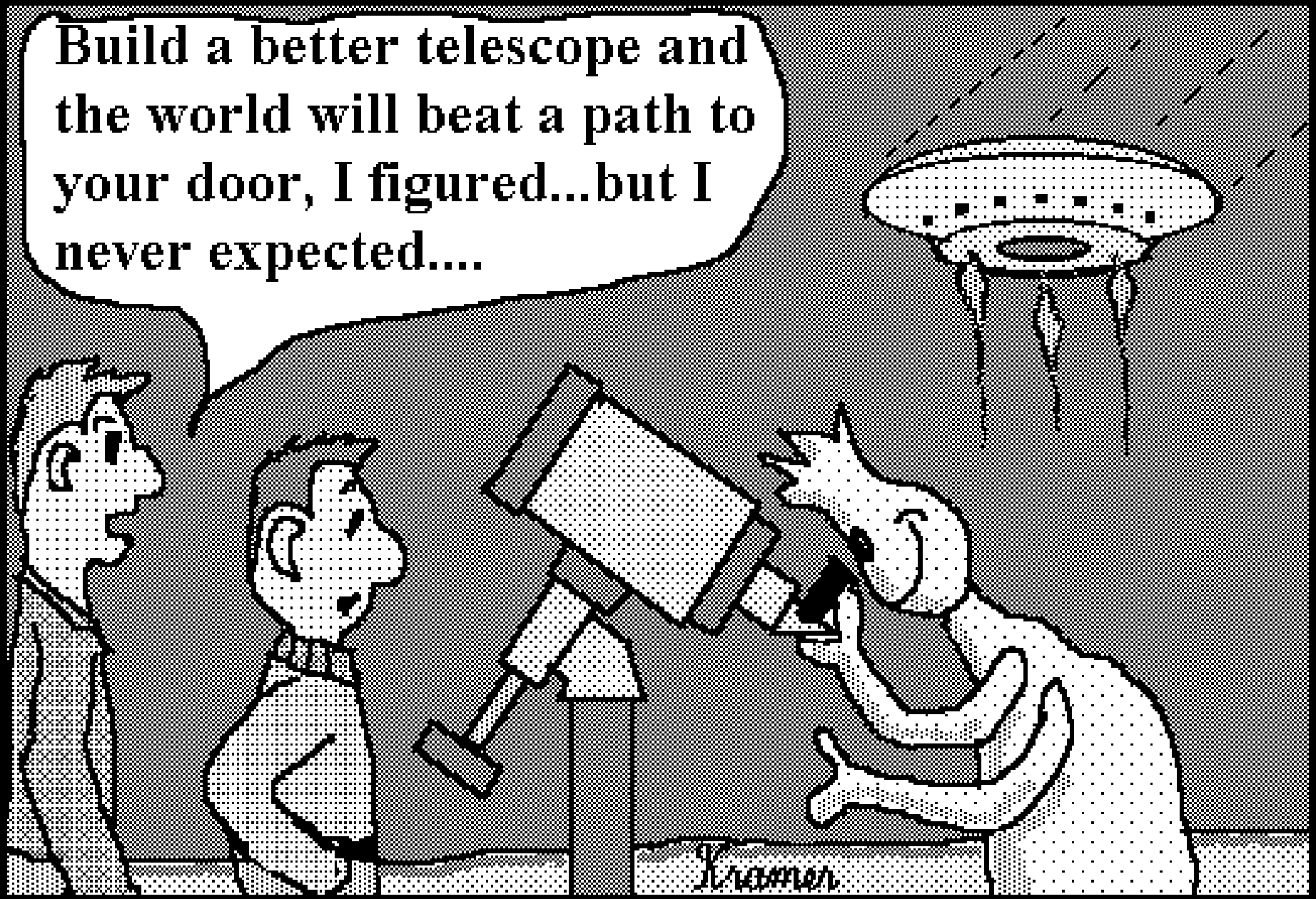
Astronomy Data Environment: Sky Surveys

- To avoid biases caused by limited samples, astronomers now study the sky systematically = **Sky Surveys**
- Surveys are used to measure and collect data from all objects that are contained in large regions of the sky, in a systematic, controlled, repeatable fashion.
- These surveys include (... this is just a subset):
  - MACHO and related surveys for dark matter objects: ~ 1 Terabyte
  - Digitized Palomar Sky Survey: 3 Terabytes
  - 2MASS (2-Micron All-Sky Survey): 10 Terabytes
  - GALEX (ultraviolet all-sky survey): 30 Terabytes
  - Sloan Digital Sky Survey (1/4 of the sky): 40 Terabytes
  - and this one is just starting: Pan-STARRS: 40 **Peta**bytes!



- **Leading up to the big survey next decade:**
  - **LSST (Large Synoptic Survey Telescope):** 100 Petabytes!

The LSST : a better sky survey

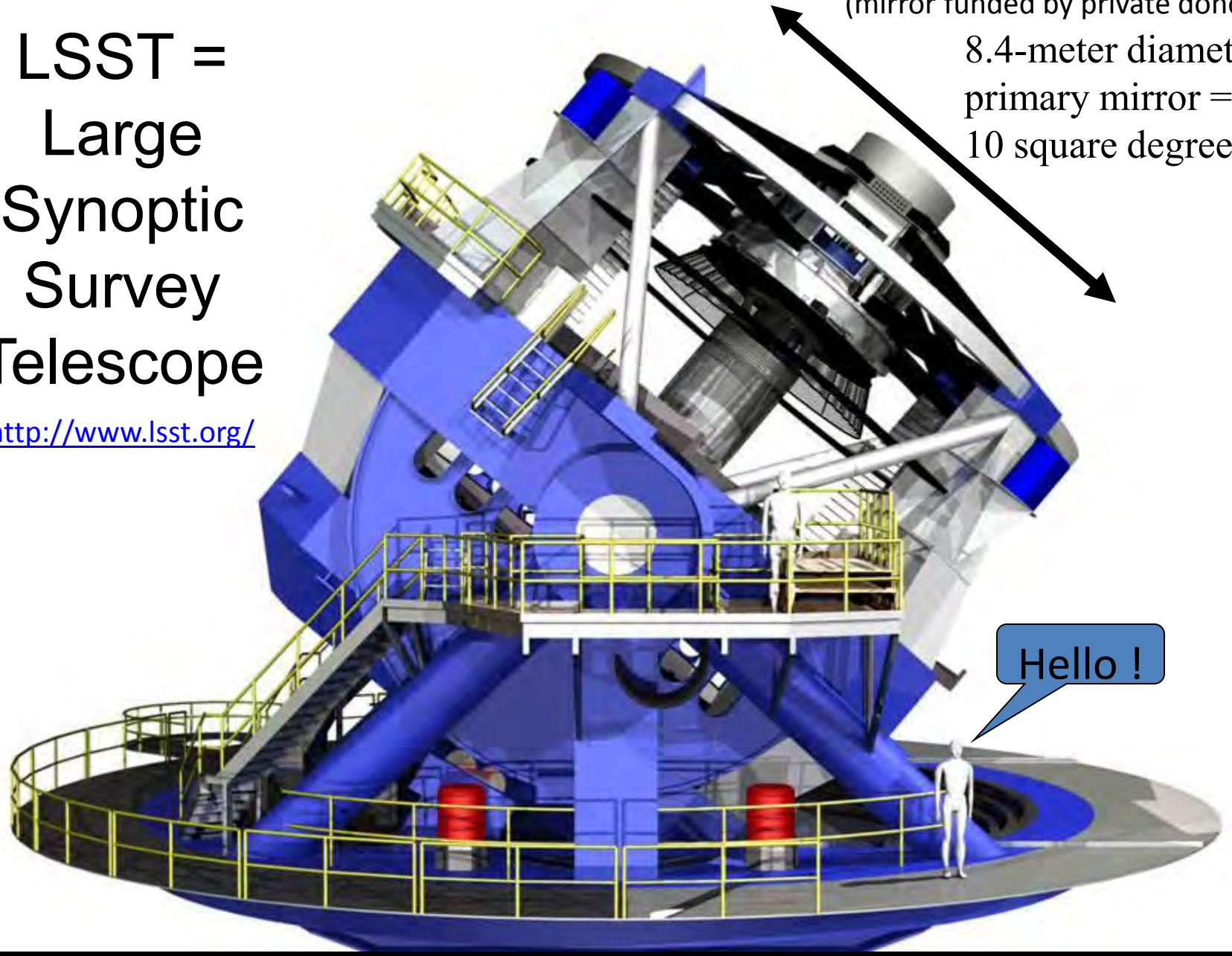


LSST = Large Synoptic Survey Telescope

http://www.lsst.org/

(mirror funded by private donors)

8.4-meter diameter primary mirror = 10 square degrees!



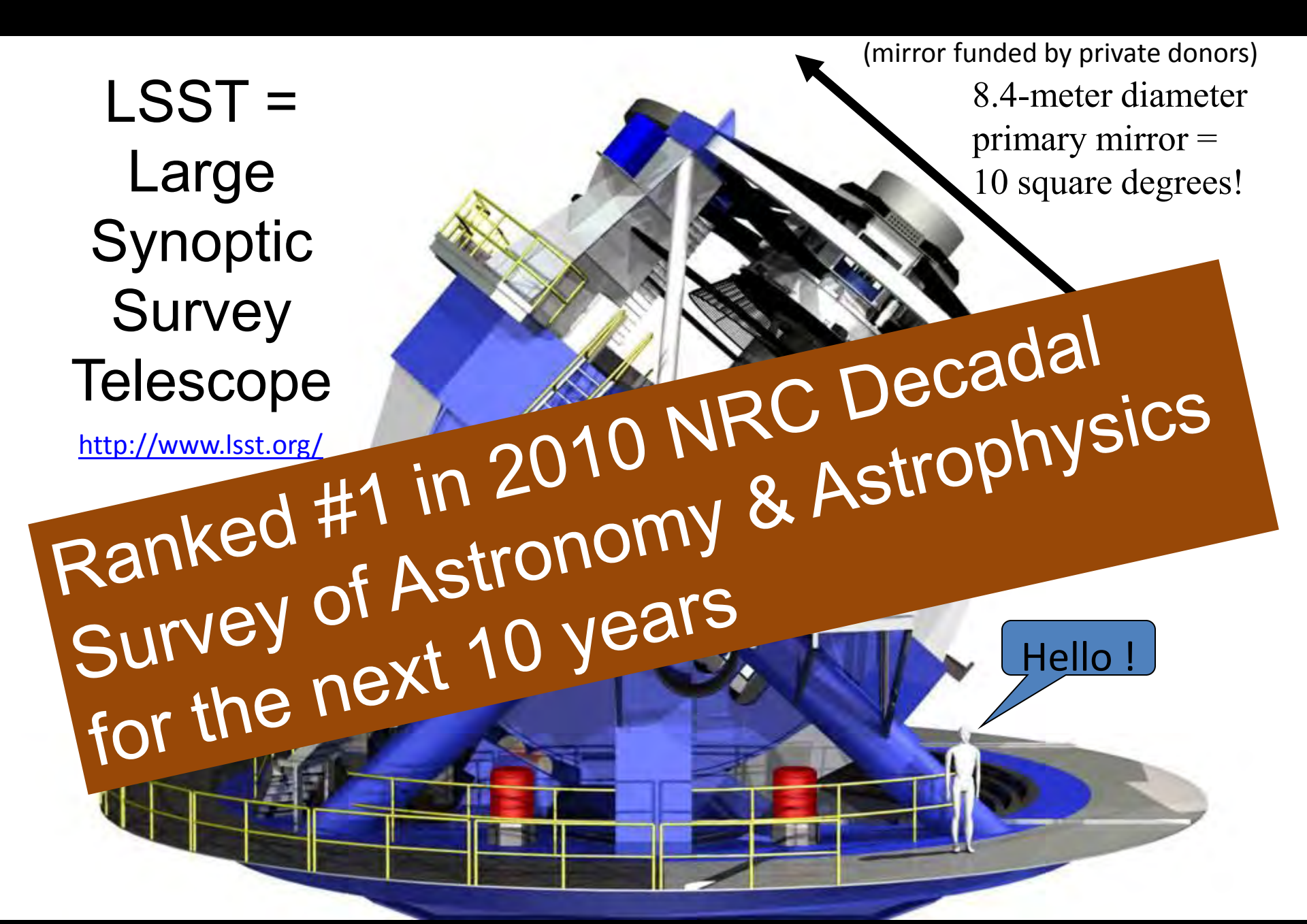
(design, construction, and operations of telescope, observatory, and data system: NSF) (camera: DOE)

LSST = Large Synoptic Survey Telescope

http://www.lsst.org/

(mirror funded by private donors)

8.4-meter diameter primary mirror = 10 square degrees!



Ranked #1 in 2010 NRC Decadal Survey of Astronomy & Astrophysics for the next 10 years

(design, construction, and operations of telescope, observatory, and data system: NSF) (camera: DOE)

LSST Key Science Drivers: Mapping the Universe

– Solar System Map (moving objects, NEOs, asteroids: census & tracking)
– Nature of Dark Energy (distant supernovae, weak lensing, cosmology)
– Optical transients (of all kinds, with alert notifications within 60 seconds)
– Galactic Structure (proper motions, stellar populations, star streams)

South America

Chile

Region de Coquimbo

Summit of Cerro Pachon -

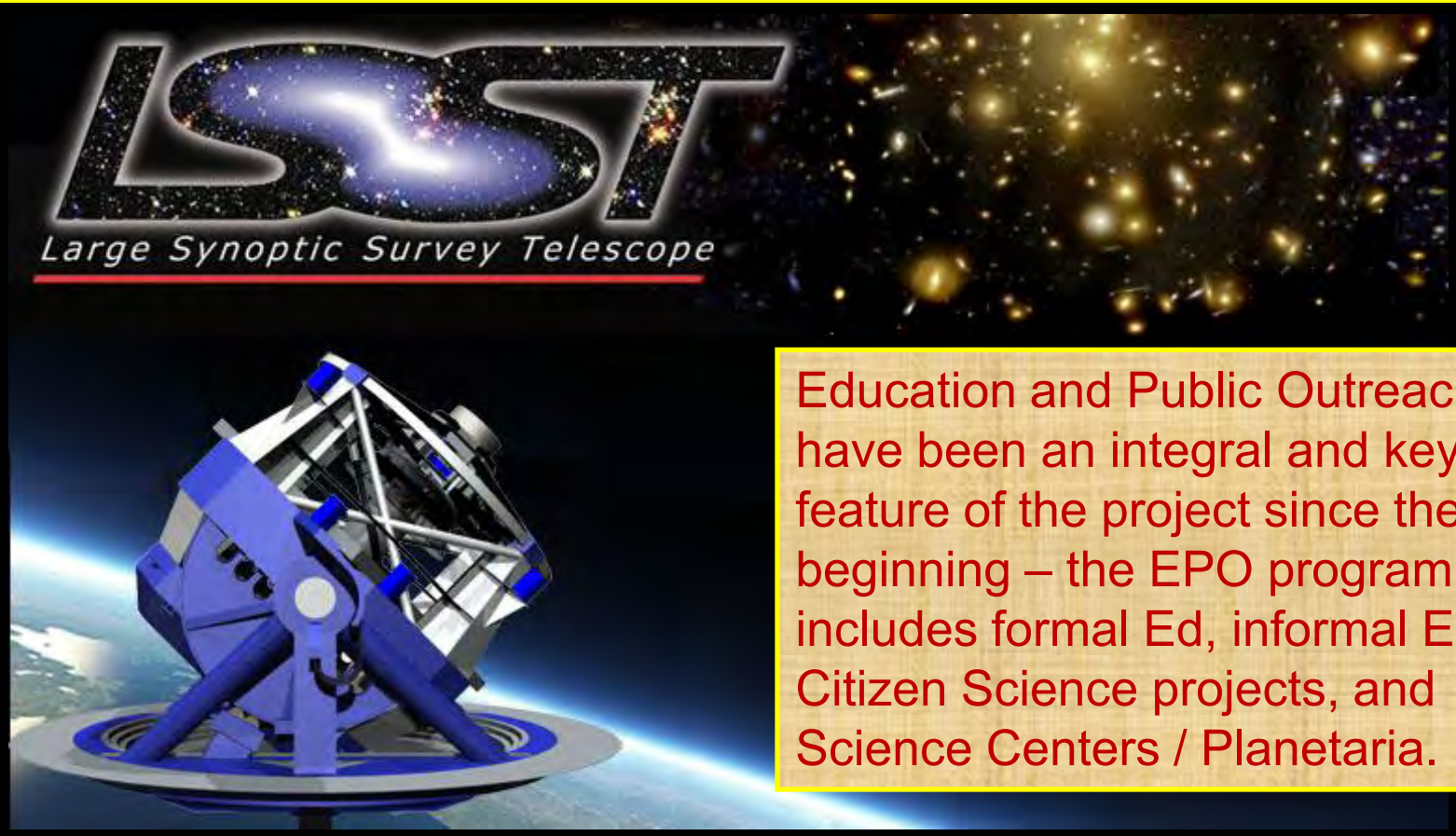
Model of LSST Observatory

<u>LSST in time and space:</u>

– When? 2016-2026
– Where? Cerro Pachon, Chile

Observing Strategy: One pair of images every 40 seconds for each spot on the sky, then continue across the sky continuously every night for 10 years (2016-2026), with time domain sampling in log(time) intervals (to capture dynamic range of transients).

- **LSST (Large Synoptic Survey Telescope):**
  - Ten-year time series imaging of the night sky – mapping the Universe !
  - **100,000 events each night** – *anything that goes bump in the night !*
  - ***Cosmic Cinematography! The New Sky! @ http://www.lsst.org/***



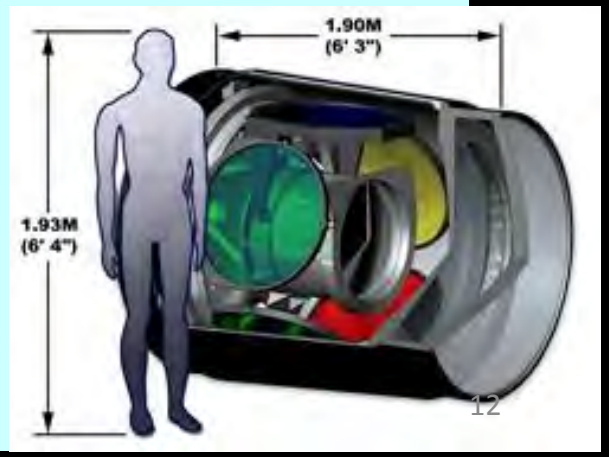
Education and Public Outreach have been an integral and key feature of the project since the beginning – the EPO program includes formal Ed, informal Ed, Citizen Science projects, and Science Centers / Planetaria.

The LSST focal plane array

Camera Specs: (pending funding from the DOE)
201 CCDs @ 4096x4096 pixels each!
= 3 Gigapixels = 6 GB per image, covering 10 sq.degrees
= ~3000 times the area of one Hubble Telescope image

LSST Data Challenges

- Obtain one 6-GB sky image in 15 seconds
- Process that image in 5 seconds
- Obtain & process another co-located image for science validation within 20^s (= 15-second exposure + 5-second processing & slew)
- Process the 100 million sources in each image pair, catalog all sources, and generate worldwide alerts within 60 seconds (e.g., incoming killer asteroid)
- Generate 100,000 alerts per night (VOEvent messages)
- Obtain 2000 images per night
- Produce ~30 Terabytes per night
- Move the data from South America to US daily
- Repeat this every day for 10 years (2016-2026)
- Provide rapid DB access to worldwide community:
  - **100-200 Petabyte image archive**
  - **20-40 Petabyte database catalog**



The LSST Data Challenges

- Massive data stream: ~2 Terabytes of image data per hour that must be mined in real time (for 10 years).
- Massive 20-Petabyte database: more than 50 billion objects need to be classified, and most will be monitored for important variations in real time.
- Massive event stream: knowledge extraction in real time for 100,000 events each night.

Informatics-based Science Education

- Informatics enables transparent reuse and analysis of scientific data in inquiry-based classroom learning (http://serc.carleton.edu/usingdata/).
- Students are trained:
  - to access large distributed data repositories
  - to conduct meaningful scientific inquiries into the data
  - to mine and analyze the data
  - to make data-driven scientific discoveries
- The 21st century workforce demands training and skills in these areas, as all agencies, businesses, and disciplines are becoming flooded with data.
- Numerous Data Sciences programs now starting at several universities (GMU, Caltech, RPI, Vanderbilt, Michigan, Cornell, …).
- CODATA **ADMIRE** initiative: ***A****dvanced* ***D****ata* ***M****ethods and* ***I****nformation technologies for* ***R****esearch and* ***E****ducation*

Citizen Science

- Exploits the cognitive abilities of **Human Computation**!
- Citizen Science = Volunteer Science = Participatory Science
  - **Crowdsourcing for science !**
  - *e.g.,* Galaxy Zoo @ http://www.galaxyzoo.org/
- Citizen science refers to the involvement of volunteer non-professionals in the research enterprise.
- The Citizen Science experience …
  - must be engaging,
  - must work with real scientific data/information (all of it),
  - must not be busy-work (all clicks must count),
  - **must address authentic science research questions** that are beyond the capacity of science teams and enterprises, and
  - must involve the scientists.

ZOONIVERSE
REAL SCIENCE ONLINE

The Zooniverse: http://zooniverse.org/

Advancing Science through User-Guided Learning in Massive Data Streams

- Building a framework for new Citizen Science projects
- Includes user-based research tools
- Science domains:
  - Astronomy (Galaxy Merger Zoo)
  - The Moon (Lunar Reconnaissance Orbiter)
  - The Sun (STEREO dual spacecraft)
  - Egyptology (the Papyri Project)
  - and more (… accepting proposals from community)

Concluding comment: Why do this? Ans: to respond to the science data flood

- ***X-Informatics*** (*e.g.,* X = Bio, Geo, Astro, …) (*in silico*):
  - addresses the scientific data lifecycle challenges in the era of data-intensive science and the data flood
  - defines lightweight ontologies, semantics, taxonomies, concepts, content descriptors for a science domain
  - for the purpose of organizing, accessing, searching, fusing, integrating, mining, and analyzing massive data repositories.
- ***Citizen Science*** (user-guided, informatics-powered):
  - Human computation (*e.g.,* tagging, labeling, classification)
    - characterized by enormous cognitive capacity and pattern recognition efficiency (**carbon-based computing**)
  - Semantic e-Science **and** Volunteer Citizen Science
  - Tagging everything, everywhere: ***Analytics in the Cloud***