Data sharing in astronomy

Željko Ivezić, Department of Astronomy, University of Washington

With contributions from: Andy Connolly, Bob Hanisch, David Hogg, Mario Jurić, Andy Lawrence, Robert Lupton, Mathias Steinmetz, Michael Strauss, Alex Szalay, Tony Tyson, Roy Williams

The Case for International Sharing of Scientific Data:
A Focus on Developing Countries
An International Symposium, Washington D.C., April 18, 2011

The Sloan Digital Sky Survey Telescope
Apache Point Observatory, NM

Outline

- **What astronomy does, why and how?**
  Data sharing example:
  Sloan Digital Sky Survey
- **Cost/benefit analysis for data sharing**
  Examples of analysis:
  Large Synoptic Survey Telescope
  Laser Interferometer Gravitational Wave Observatory

``The big questions" in astronomy

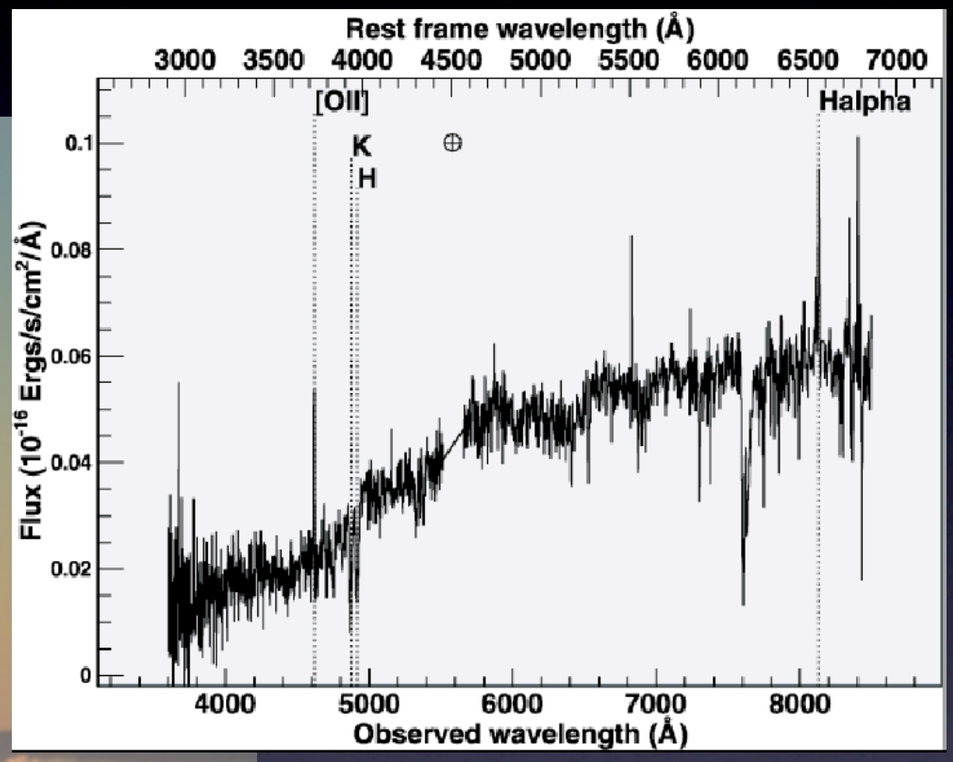
- How and when did the Universe begin, and what is it made of?
- How did the structure (planets, stars, galaxies) form and evolve?
- Is our planetary system unique? Are we alone?

Alternatively: does "our" physics work everywhere (e.g. black holes, thermonuclear explosions in supernovae)? Can we learn more physics by studying the heavens (e.g. neutrinos, dark matter and dark energy)?

Tools and methods: rapid progress on many fronts over the last decade, led by sensor development and, above all, **information technology** **the advent of massive digital astronomical sky surveys as a leading research method**

Context: three modern observational methods in astronomy and astrophysics:

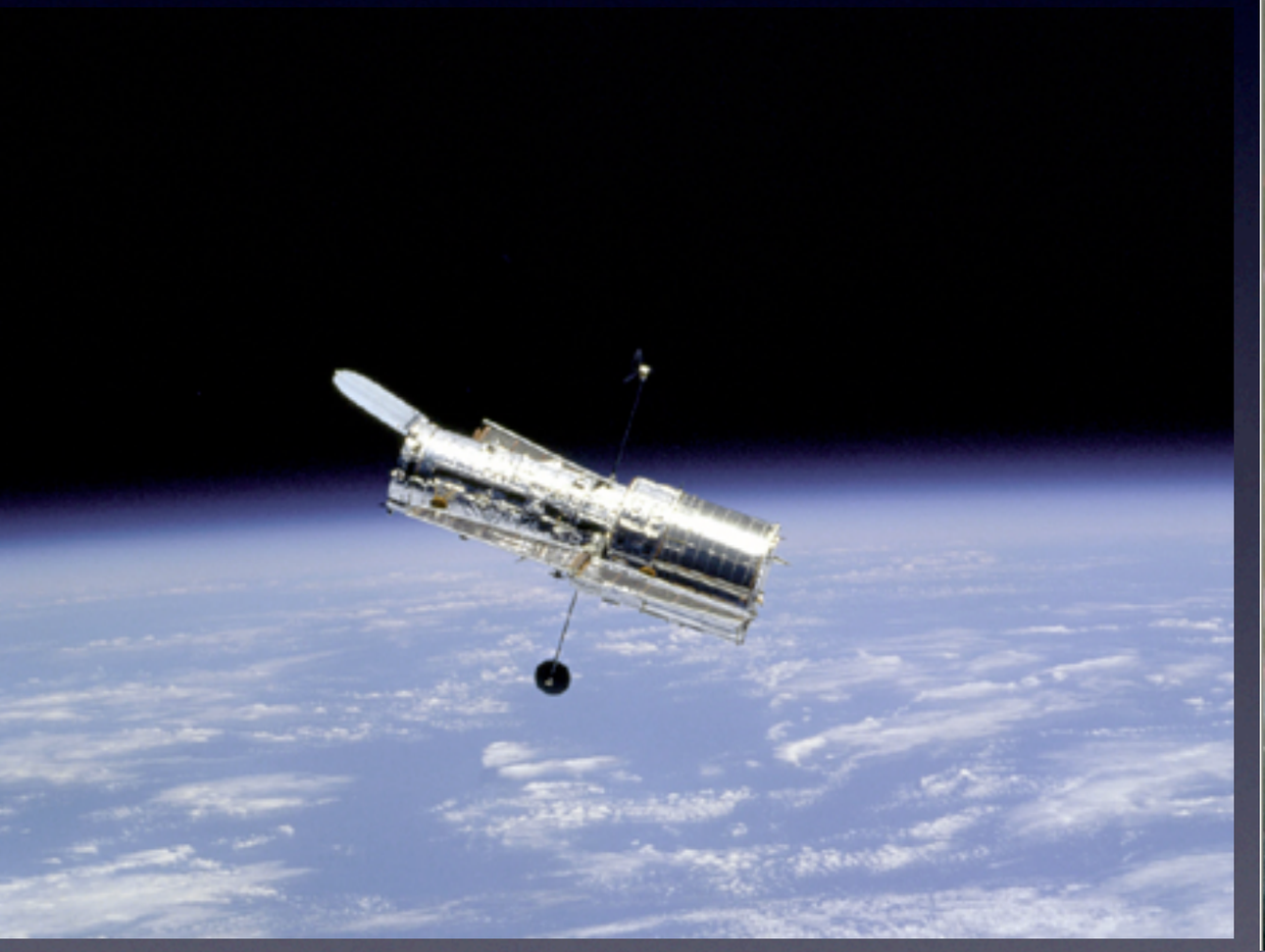
- Large telescopes (~10m): faint objects, especially spectroscopy



The Keck telescopes on Mauna Kea (Hawaii)

Context: three modern observational methods in astronomy and astrophysics:

- Telescopes above the atmosphere: high angular resolution (e.g., the Hubble Space Telescope) and other wavelength regions (X-ray, radio, infrared)



The HST in orbit and an example of a galaxy image

Context: three modern observational methods in astronomy and astrophysics:

- Large telescopes (~10m): faint objects, especially spectroscopy
- Telescopes above the atmosphere: high angular resolution (e.g., the Hubble Space Telescope) and other wavelength regions (X-ray, radio, infrared)
- Large sky surveys: accurate digital data for hundreds of millions of astronomical sources enables uniquely powerful statistical analysis

Key development: modern sky surveys make all their data (images and catalogs) **publicly available**

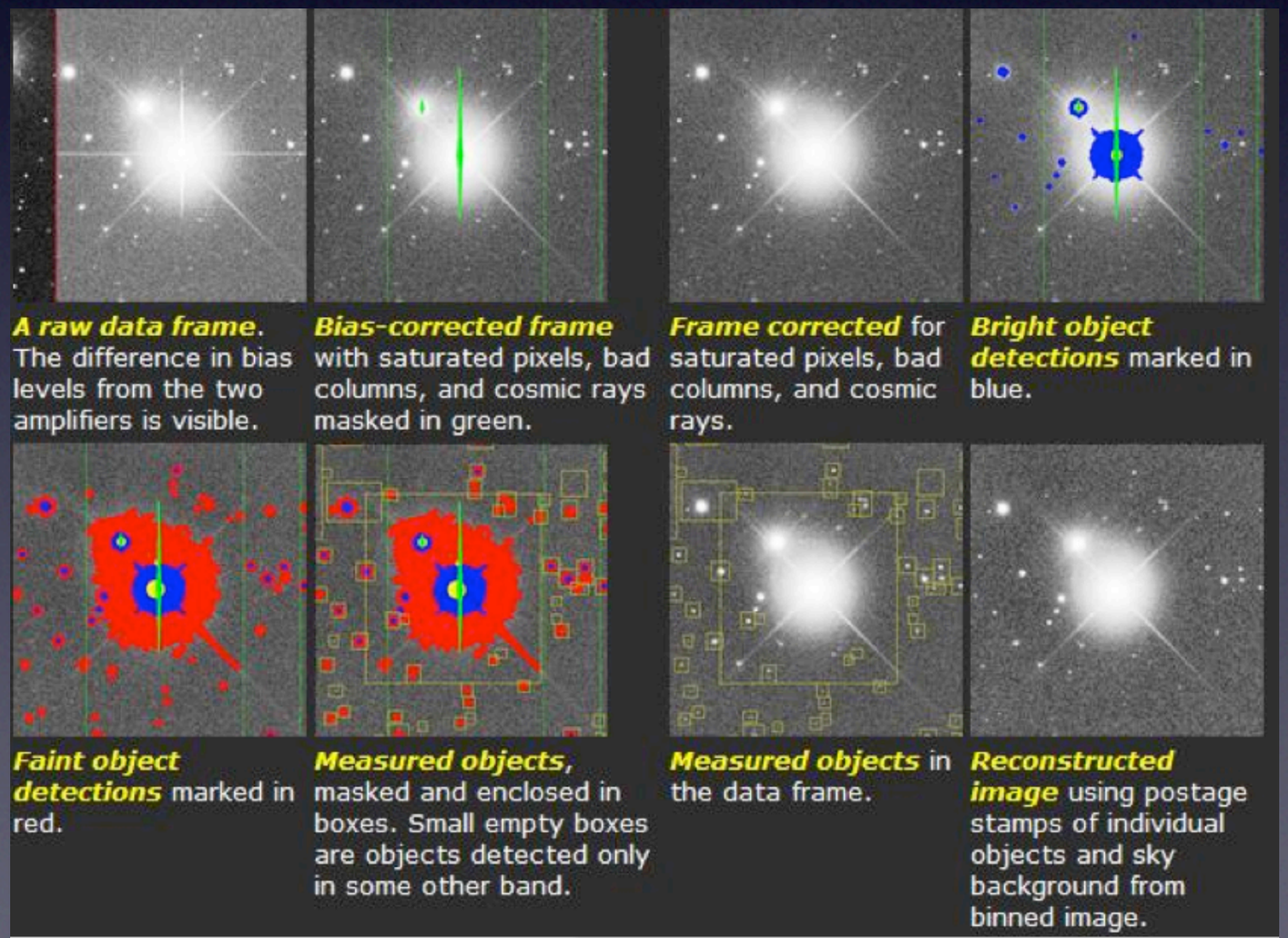
What are sky maps and catalogs?
Why are sky maps useful?

- Sky maps are abstracted into catalogs:
  - a list of all detected objects (stars, galaxies, ...)
  - measured parameters (size, color, brightness,...)

Basic steps in astronomical image processing (example: Sloan Digital Sky Survey):

All these (complicated) steps are already done:

"science-ready database"



What is a sky map? Why are sky maps useful?

- Sky map, abstracted into catalogs:
  – a list of all detected objects (stars, galaxies, ...)
  – measured parameters (size, color, brightness,...)

- The utility of sky maps:
  Discoveries of new objects: "Is this a new asteroid, or is it already cataloged?"
  Object classification: "What types of galaxies exist?"
  Statistical population studies: "Do quasars change their properties with time?"
  Search for unusual objects: "Is this star very weird?"
  Cosmological measurements: "How fast does the Universe expand?"

"Science-ready database": measurements can be (simply) analyzed without the need for (complex) image processing

Sloan Digital Sky Survey (SDSS):

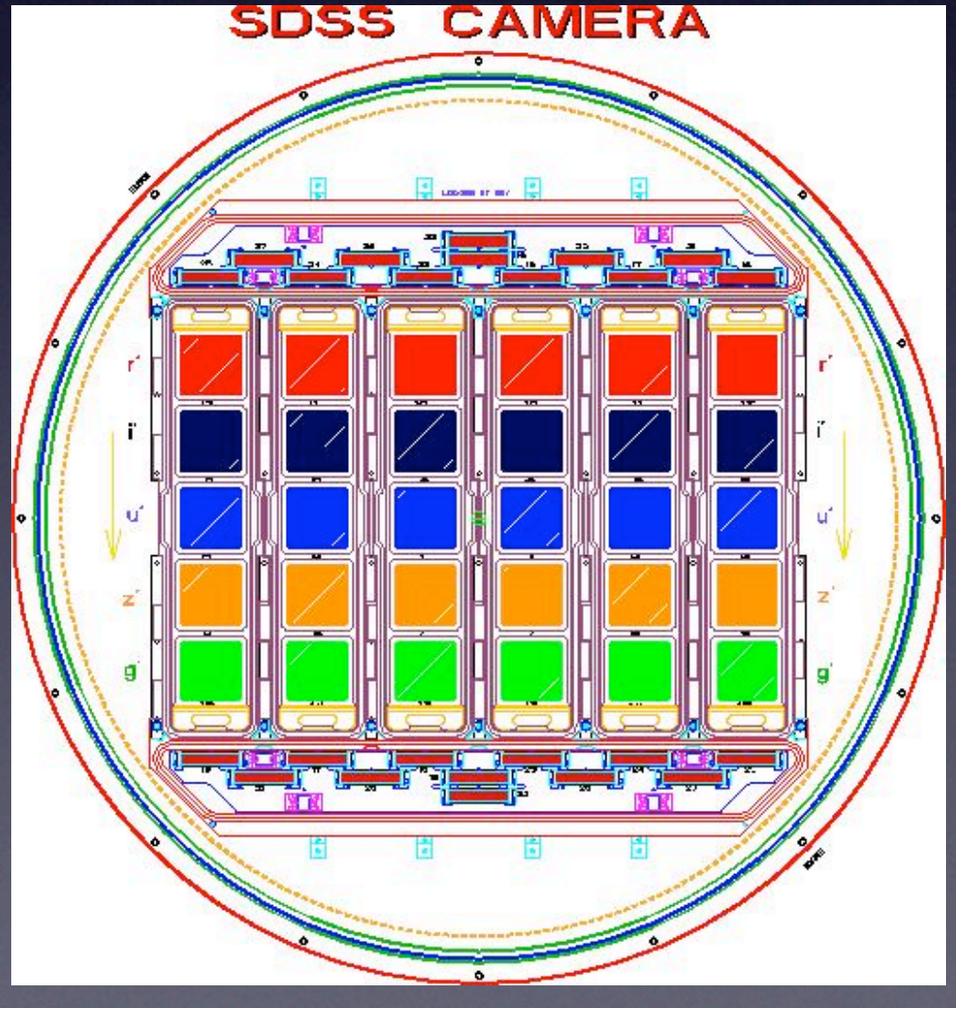
the first massive digital color map of the night sky

Many other sky surveys: 2MASS, GALEX, FIRST, NVSS, WISE...

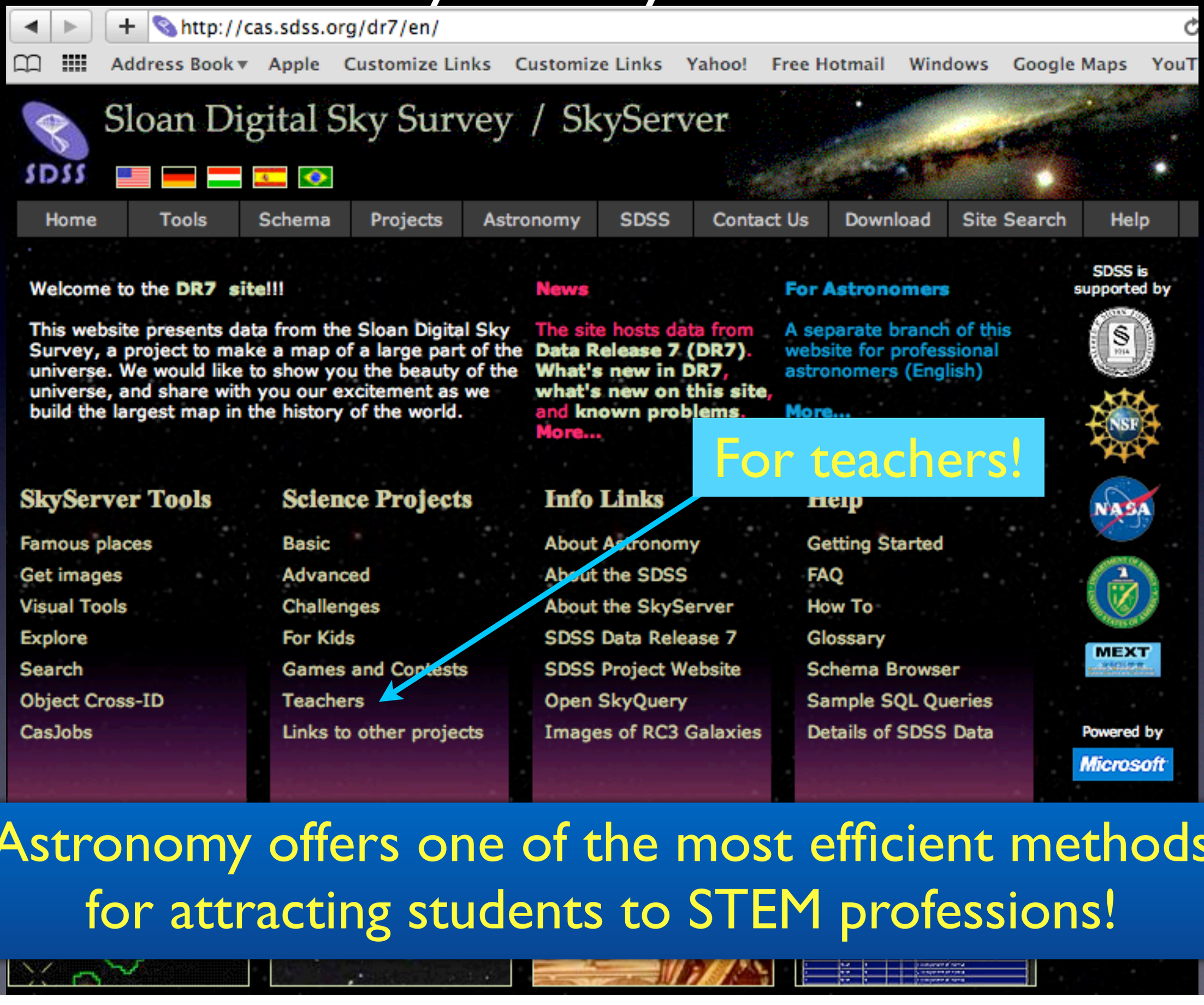
Apache Point Observatory, New Mexico

The last decade: SDSS as an example

- Digital sky survey with a 120 Megapix CCD camera
- Precise measurements for 400,000,000 objects
- Revolution in astronomy: public databases released before a substantial fraction of analysis was done by the project team



Astronomy "from your armchair"



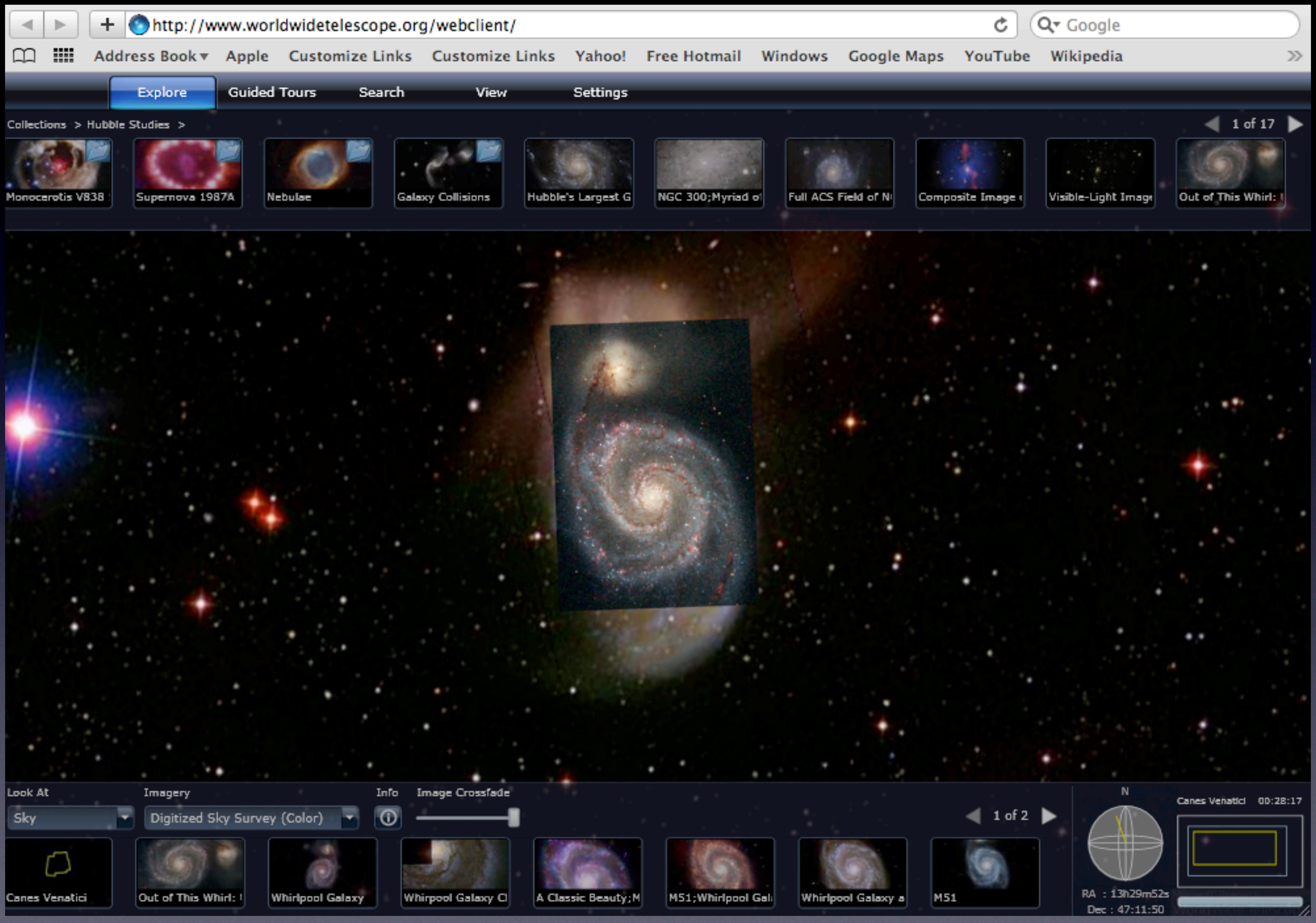
Astronomy offers one of the most efficient methods for attracting students to STEM professions!

As a result of SDSS public data releases:

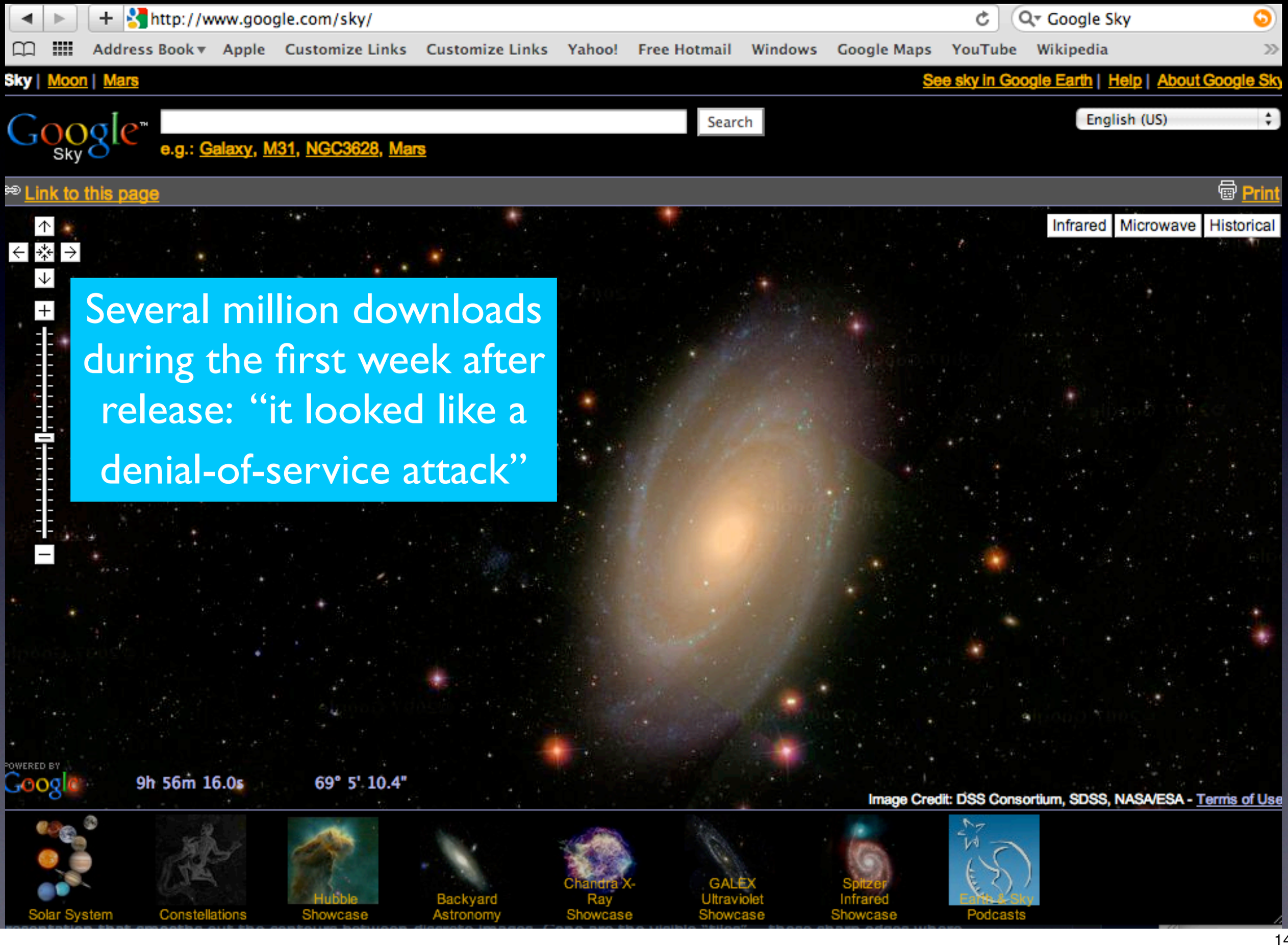
- Several thousands of refereed papers, a majority authored by scientists not associated with SDSS
- Delivered >100 times the total data volume
- Over 300,000,000 web hits in 6 years with over a million unique users (vs. 10,000 astronomers)

Prof. James E. Gunn accepts a National Medal of Science

Sharing data with everyone: World Wide Telescope



Sharing data with everyone: Google Sky



What did astronomers learn about data sharing?

Data sharing comes with costs and risks:

1) requires **higher standards** than for internal use, including publications describing data formats, provenance and metadata (insiders' "know-how")

2) **the cost of curation** (servers, help desk, etc.)

3) risk of being "scooped" (larger for very focused/specialized data streams, more likely for an experiment than for a survey)
But: "Does releasing data weaken collaboration?" Not really, indeed Steve Ritz: "The Fermi collaboration became even stronger after the public data release." and a lot of additional evidence from other surveys

Data sharing is not free: need a cost/benefit analysis

``The top 10" benefits of data sharing in astronomy

Caveat: benefits vary with discipline, and within a discipline with time

1) Early data releases greatly improve the final product

- **quality assurance:** deficiencies can be discovered and mitigated while still taking the data (to discover subtle systematic errors cutting-edge science analysis is often required) and **before it's too late**

- more people "looking" at the data increases the chance of finding subtle problems (especially important for space missions with finite lifetime, e.g. the ESA's Gaia mission)

- evidence: even imperfect early data releases are well received by the broad community

- building the open data science center and its documentation is a resource not just for the broader community, but also for the collaboration members (especially for the new members, such as graduate students and postdocs)

``The top 10" benefits of data sharing in astronomy

2) Early data releases enable coeval science

- especially important in astronomy where "experimental setup" cannot be controlled (observations vs. experiments)

- there are transient sources (e.g. supernovae explosions, comets, and other rare events)

- instruments can have finite lifetime (especially true for space missions)

- there are regretful cases in astronomy when good science was lost due to inadequate and delayed data sharing

``The top 10" benefits of data sharing in astronomy

3) More science is extracted from the same dataset

- true for all astronomical surveys: **more users yield more science**

- SDSS: more papers were written by outsiders (a total of several thousand papers) than by project members, resulting in significant impact boost

- diversity of ideas: many of the most visible SDSS results were unanticipated in the original project proposal (probably true for other surveys)

- many examples in history of a single iconoclast voice being correct: real outsiders should be allowed to see the data!

``The top 10" benefits of data sharing in astronomy

4) Enables reproducibility of science results

- easy verification of published results if data are readily available

- preserving data for posterity (especially important for astronomy, e.g. historical data for supernovae and other transient sources)

- processing code should be released together with the data!

``The top 10" benefits of data sharing in astronomy

5) Synergy between different datasets enables science that cannot be done with individual datasets

- cross-correlation of multi-wavelength astronomical data (X-ray, optical, infrared, radio)

- open data encourages uniformity and interoperability of datasets, resulting in more cross-fertilization

``The top 10" benefits of data sharing in astronomy

6) Enables cross-disciplinary science

- good example: astronomy/statistics/computer science

- foster exchange of ideas, tools and methods (such as data mining and visualization of massive datasets)

``The top 10" benefits of data sharing in astronomy

7) Sometimes the only way to secure scarce resources

- "easy things" (e.g. those that can be put together by a small number of groups/institutions) have been done in the last century; the "road ahead" requires more substantial merging of research resources

- strong argument for international collaboration

- expensive and limited resources in astronomy require broad community support to secure them: **everybody gets the data**

- good examples from astronomy: HST Deep Field, UKIDSS, LSST

``The top 10" benefits of data sharing in astronomy

8) Results in more citations and prestige to the team who produced data

- "yes, we are all driven by curiosity, but we also need to feed our children"

- career benefits, especially for those in early stages of their careers

- insiders with know-how (especially postdocs) become a hot market commodity (practically all postdocs from the first phase of SDSS hold faculty-level positions today)

``The top 10" benefits of data sharing in astronomy

9) Education and public outreach

- Galaxy Zoo project based on SDSS data recruited 200,000 volunteers!

- astronomy offers one of the most efficient methods for attracting students to STEM professions: hands-on research and learning about the scientific method

``The top 10" benefits of data sharing in astronomy

10) Ethics and "broader impact"

- last but not least!

- "sharing is nice" (as per kindergarten teachers)

- taxpayers have paid for most of research projects and they should see the results at every level of understanding

- world-wide democratization of scientific research

- developing countries: "small" teams can do "big" science

Additional considerations for the data sharing cost/benefit analysis

- who will use the data? (will "customers" indeed come?)

- will they will be able to use data (e.g. easy-to-use astronomical catalogs vs. data streams from high-energy physics accelerators)

- security/proprietary issues (national security, e.g. PanSTARRS/AirForce collaboration, commercial gains for "foreign" competition?)

- Jim Gray (Microsoft): "astronomical data are worthless"

- funding details for public data release (e.g. who will pay? does everyone get the same bandwidth for data access and the same computing resources?)

``The top 10" benefits of data sharing in astronomy

1) Early data releases greatly improve the final product

2) Early data releases enable coeval science

3) More science is extracted from the same dataset

4) Enables reproducibility of science results

5) Synergy between different datasets

6) Cross-disciplinary science

7) Sometimes the only way to secure scarce resources

8) More citations and prestige to the team

9) Education and public outreach

10) Ethics and broader impact

Is there a practical value to this list? Can it be used to perform cost/benefit analysis?

Two examples from astronomy and physics

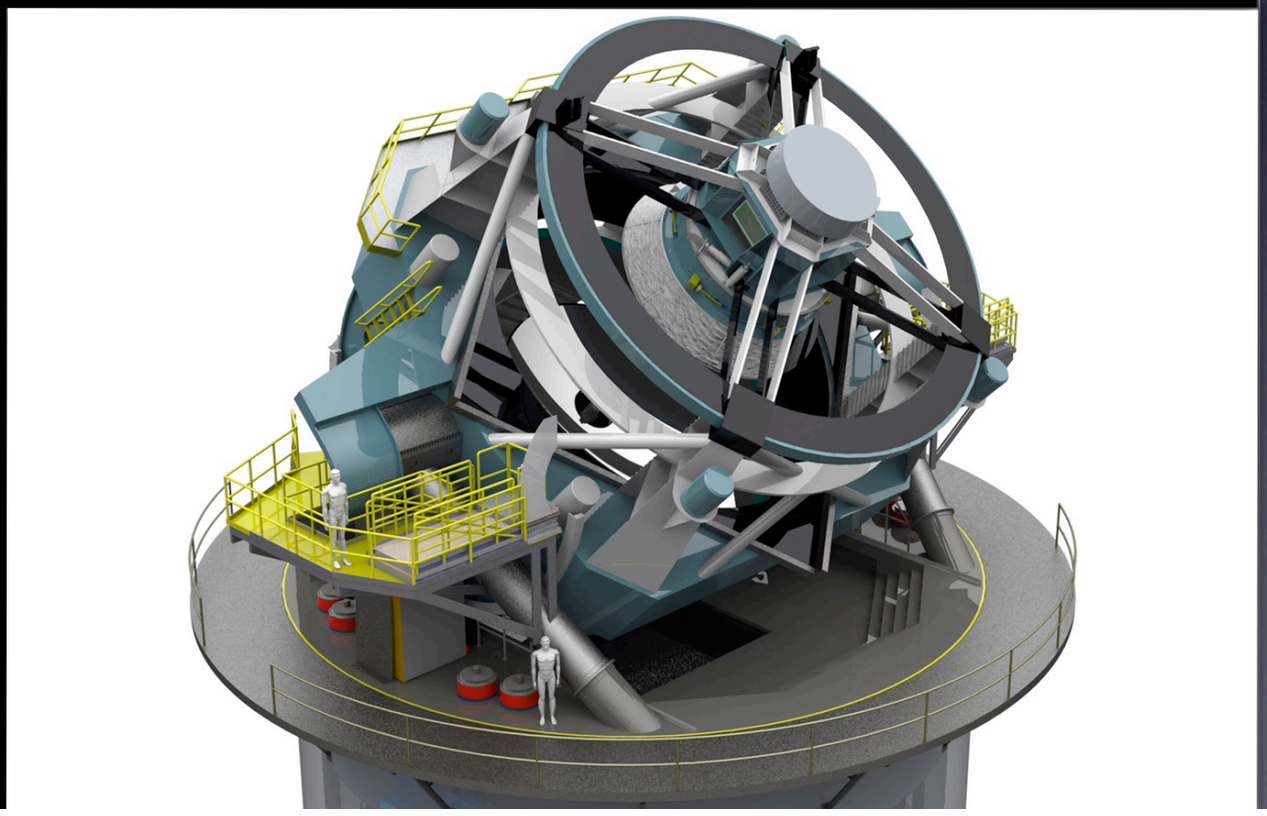
A peek into the future: the Large Synoptic Survey Telescope

SDSS:
a digital color
snapshot of the
night sky

LSST:
a digital color
movie of the sky

SDSS: one US Library of Congress worth of data

LSST: one SDSS per night, or all the words ever printed!



The Data Challenge

- ~3 Terabytes per hour that must be mined in real time.
- 20 billion objects will be monitored for important variations in real time.

All 10 benefits apply to LSST dataset:

The LSST data, all >100,000 TB, will be available to everyone in 2020s, just like 40 TB of SDSS data are today!

Laser Interferometer Gravitational Wave Observatory

Gravitational waves are ripples in the fabric of space and time produced by accelerating masses (much as electromagnetic waves are produced by accelerating charges).

LIGO: a facility dedicated to the detection of cosmic gravitational waves and the measurement of these waves for scientific research: **a detection will represent a fundamental physics accomplishment**

Two US sites (LA, WA) (and similar facilities in Europe: VIRGO, GEO)

Left: the LIGO site in Hanford, WA

``The top 10" benefits of data sharing in astronomy

1) Early data releases greatly improve the final product

2) Early data releases enable coeval science

3) More science is extracted from the same dataset

4) Enables reproducibility of science results

5) Synergy between different datasets

6) Cross-disciplinary science

7) Sometimes the only way to secure scarce resources

8) More citations and prestige to the team

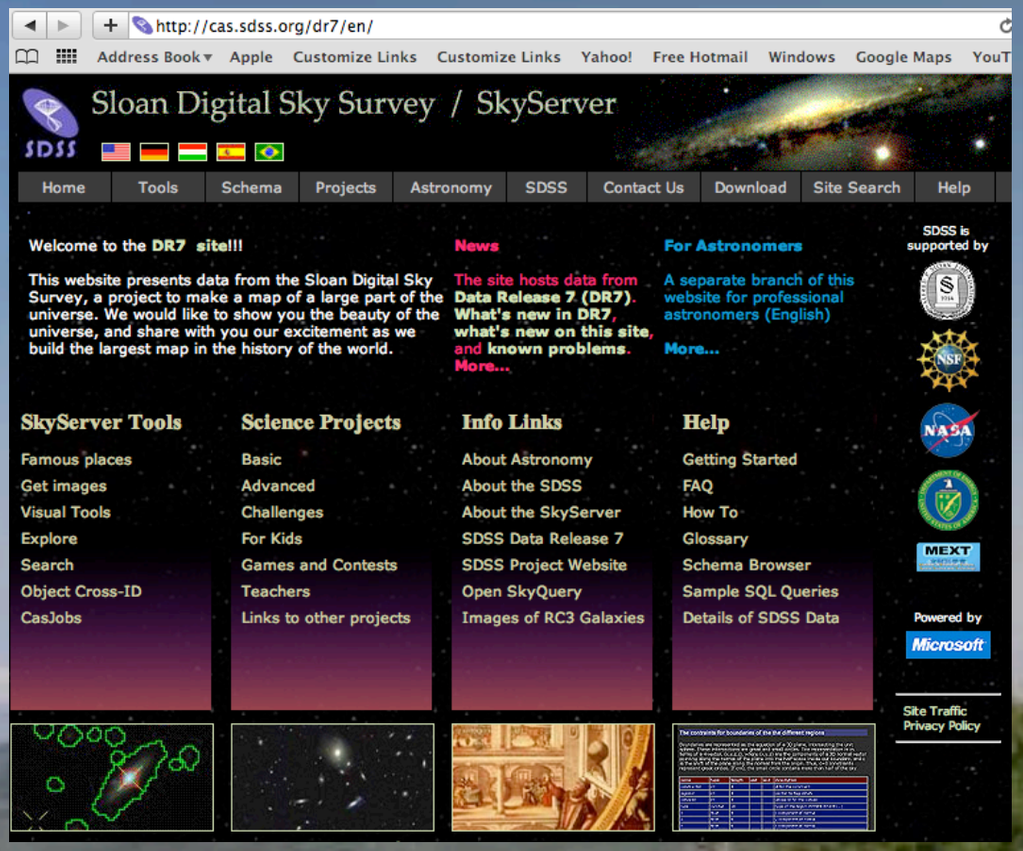
9) Education and public outreach

10) Ethics and broader impact

LIGO: benefits 2, 4, 5, 6, 8, 9 and 10 definitely applicable, and 1, 3 probably (7: LIGO is already fully funded).

(yes, the LIGO team is preparing a public data release)

Data sharing: in astronomy, and elsewhere



On the one hand, it's a matter of cost/benefit analysis, but on the other hand, we don't seem to have a choice but to do it.

1) Early data releases greatly improve the final product
2) Early data releases enable coeval science
3) More science is extracted from the same dataset
4) Enables reproducibility of science results
5) Synergy between different datasets
6) Cross-disciplinary science
7) Sometimes the only way to secure scarce resources
8) More citations and prestige to the team
9) Education and public outreach
10) Ethics and broader impact