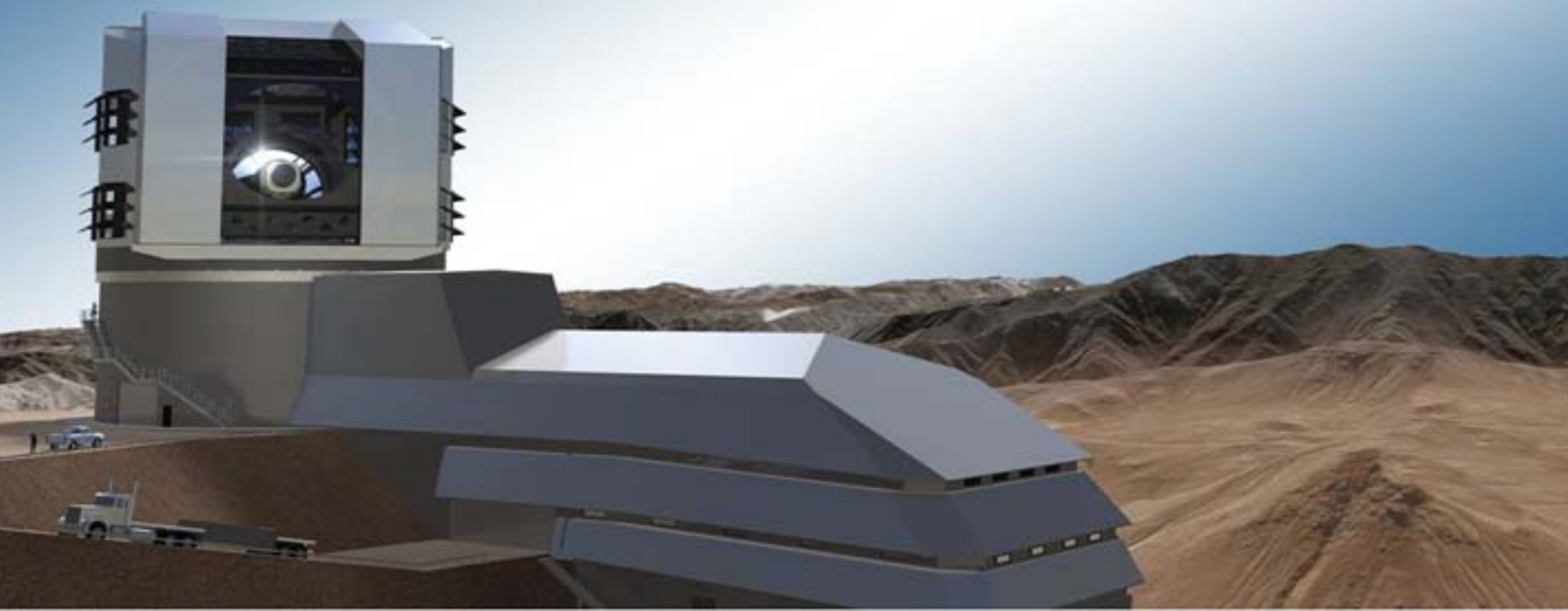
LSST: The Data is Just the Beginning of the Challenge

Tim Axelrod
LSST Data Management Scientist



A Lightning Quick Look at the LSST

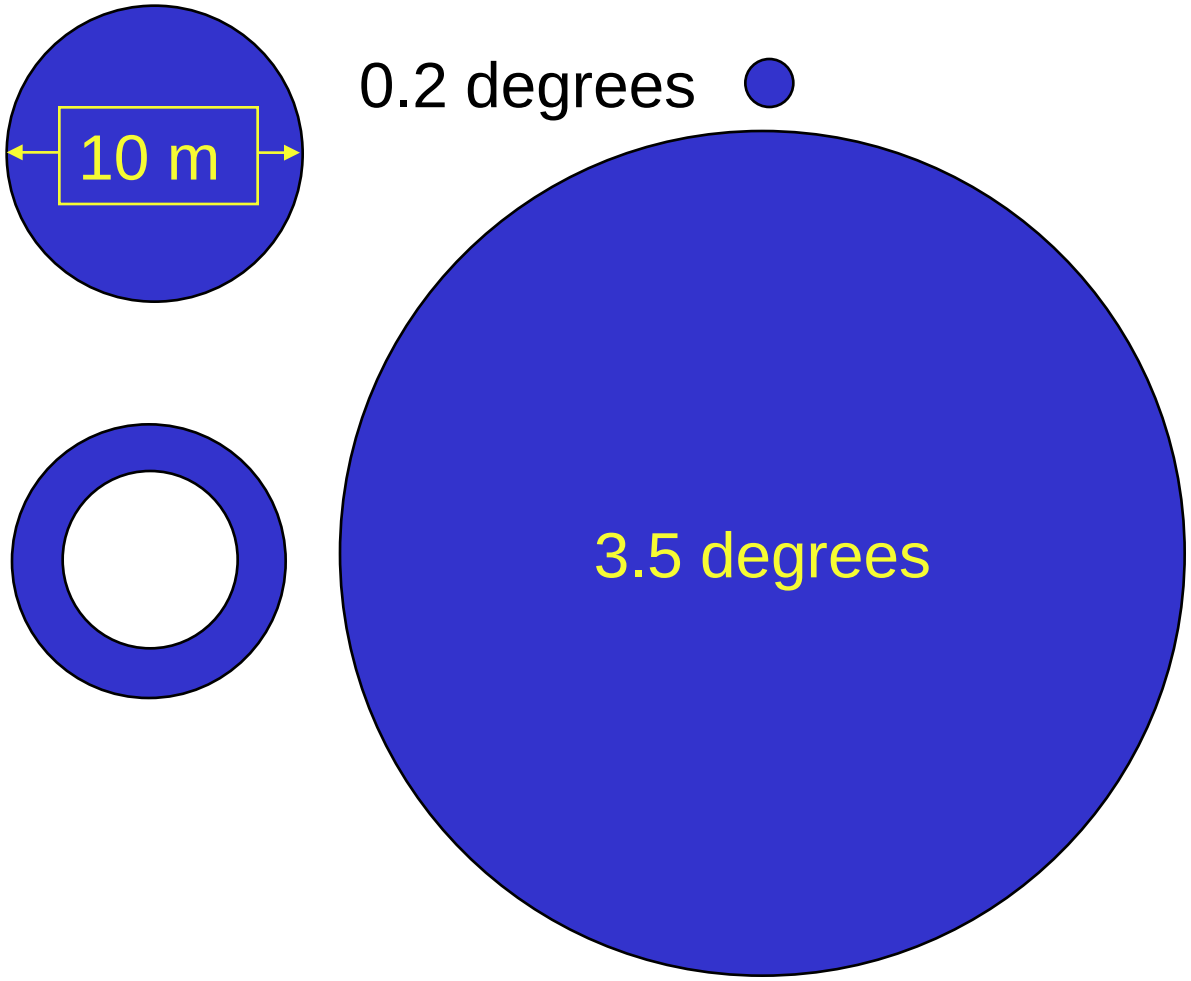
Large Synoptic Survey Telescope: Wide+Deep+Fast

Keck Telescope

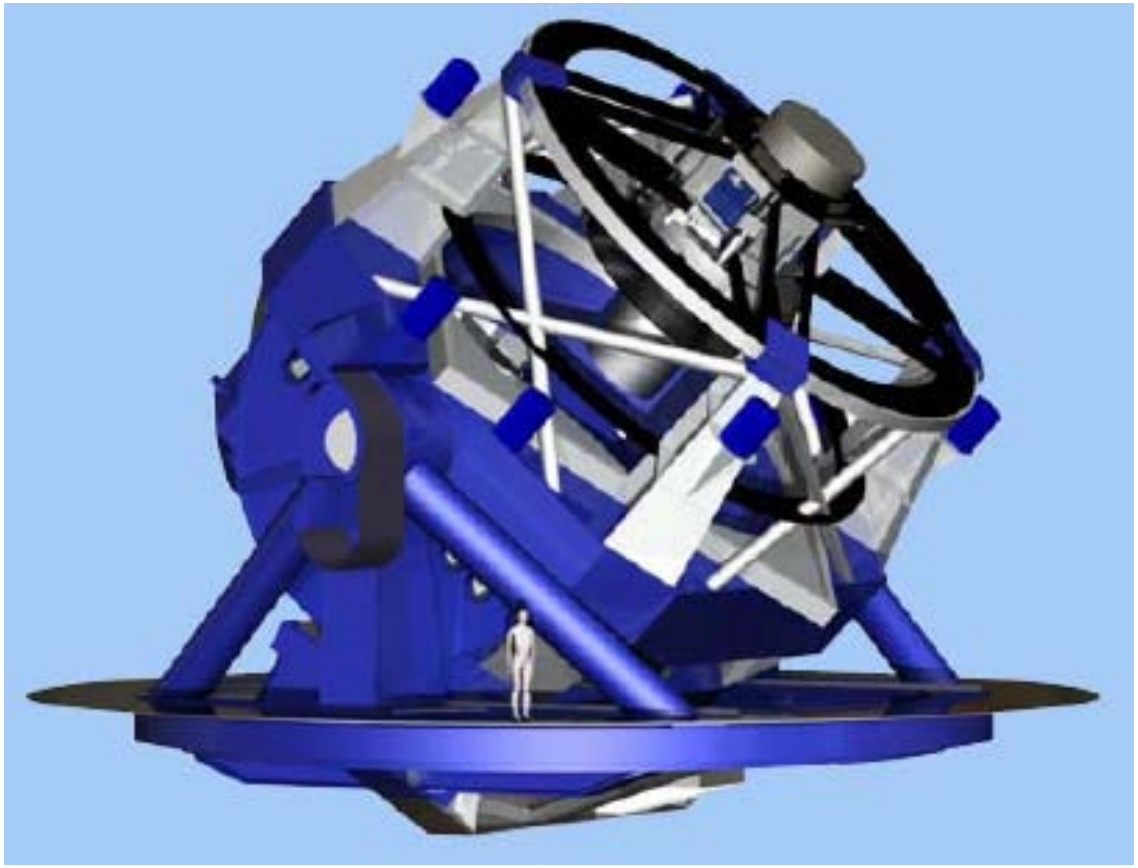
LSST

Primary mirror diameter

Field of view



LSST - Essential Statistics

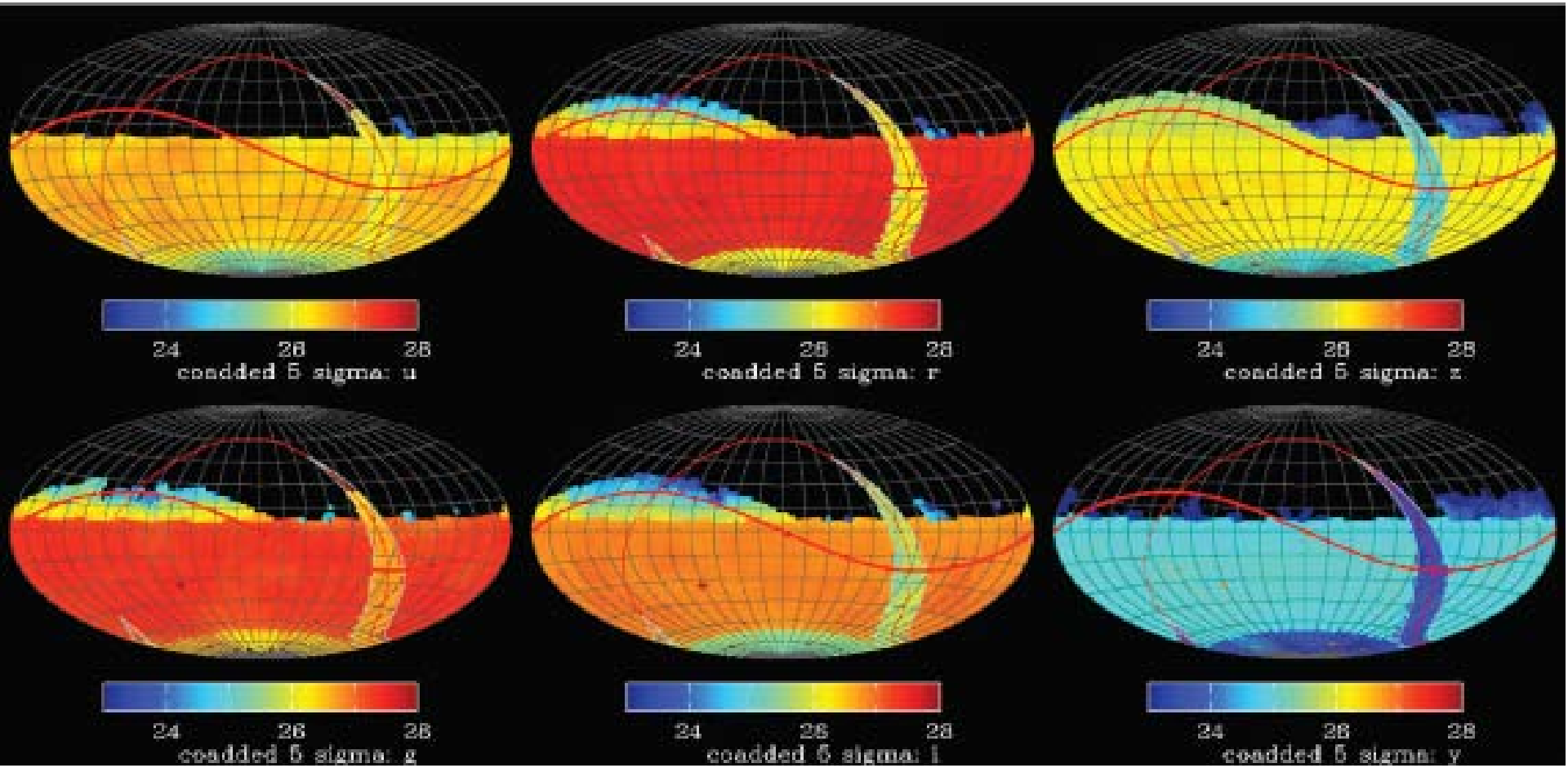


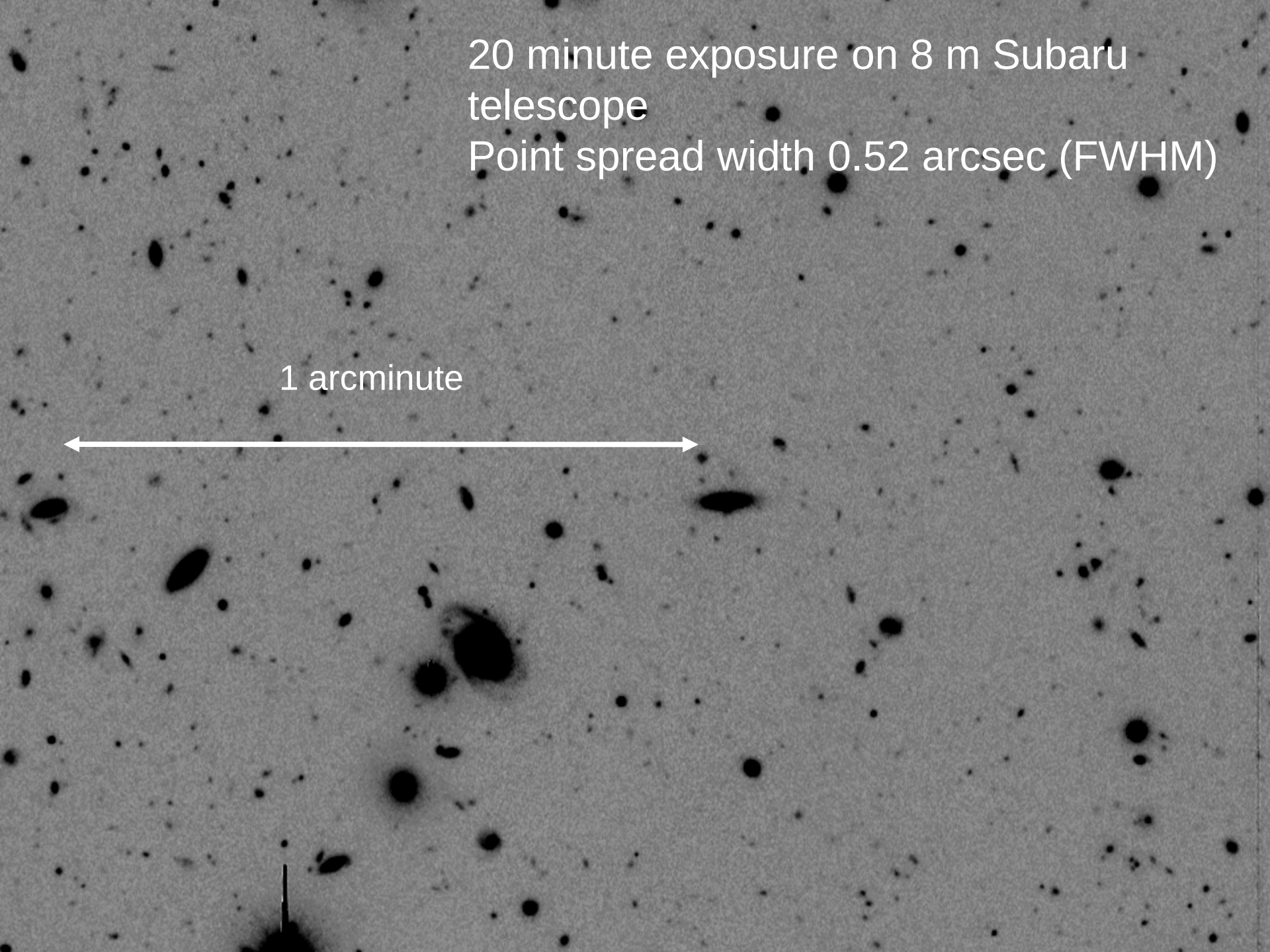
- Aperture diameter: 8.4m
- Effective aperture: 6.7m
- FOV: 3.5 deg
- Filters: u, g, r, i, z, y
- 3.2 gigapixels
- 2 sec, 5 electron noise readout
- Observing mode: pairs of 15 sec exposures, separated by 5 sec slew
- Single exposure depth: ~24.5
- Repetitively scan 20000 sq deg
- Site: Cerro Pachon, Chile
- Data flows at 0.5 GB/sec – all night
- 18 TB / night
- First light: ~2020

One Survey – Many Science Programs

- **The LSST Observatory will produce a data stream which the Data Management System turns into data products.**
  - **0.5 GB/sec all night, every night for 10 years**
  - **104 PB of images at survey end**
  - **2.5 PB science database at survey end**
- **Many science programs are supported by the same data products**
  - **Weak lensing**
  - **Supernovae & transient astrophysics**
  - **Milky Way structure**
  - **Solar System inventory**
  - **Many more in individual science collaborations**

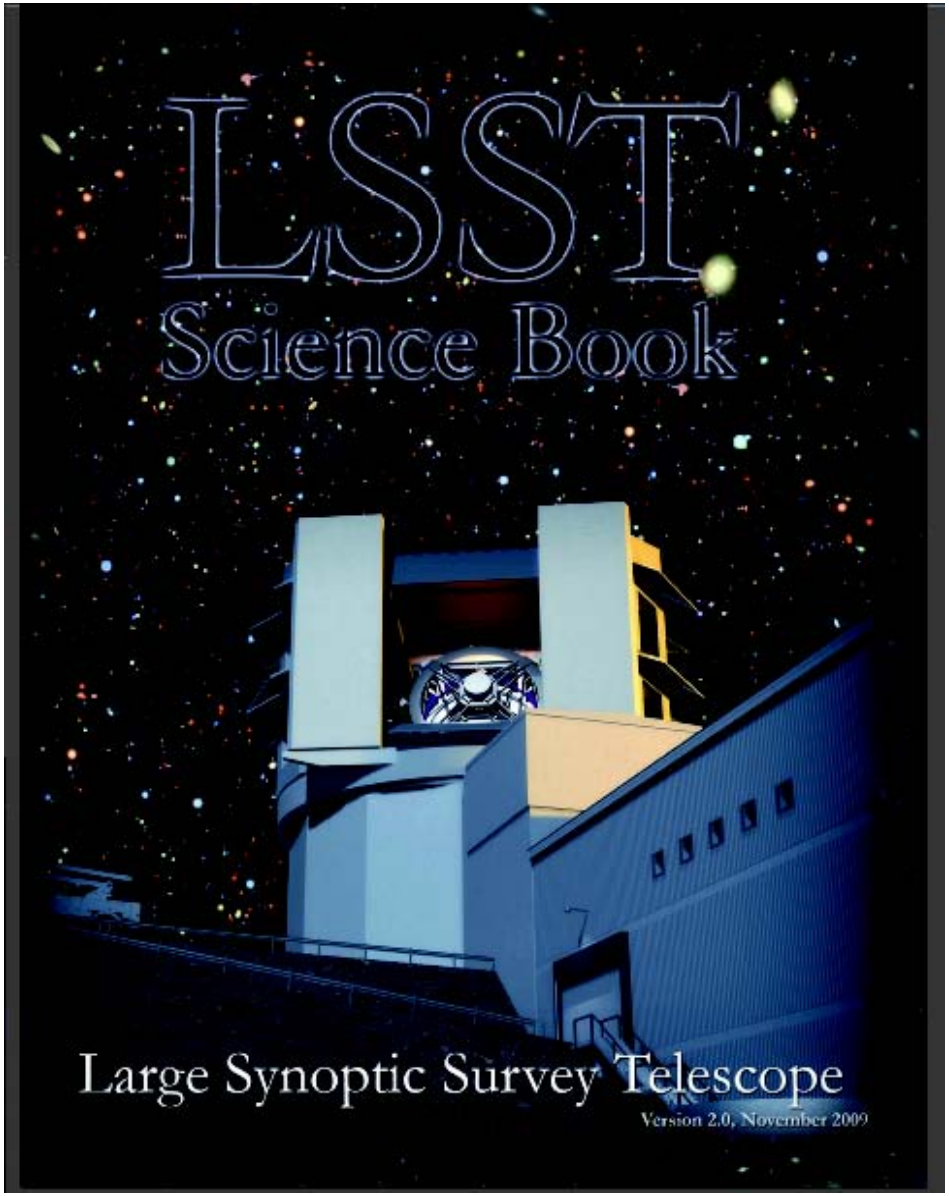
Simulated Results of 10 yr Survey
5.3M Exposures



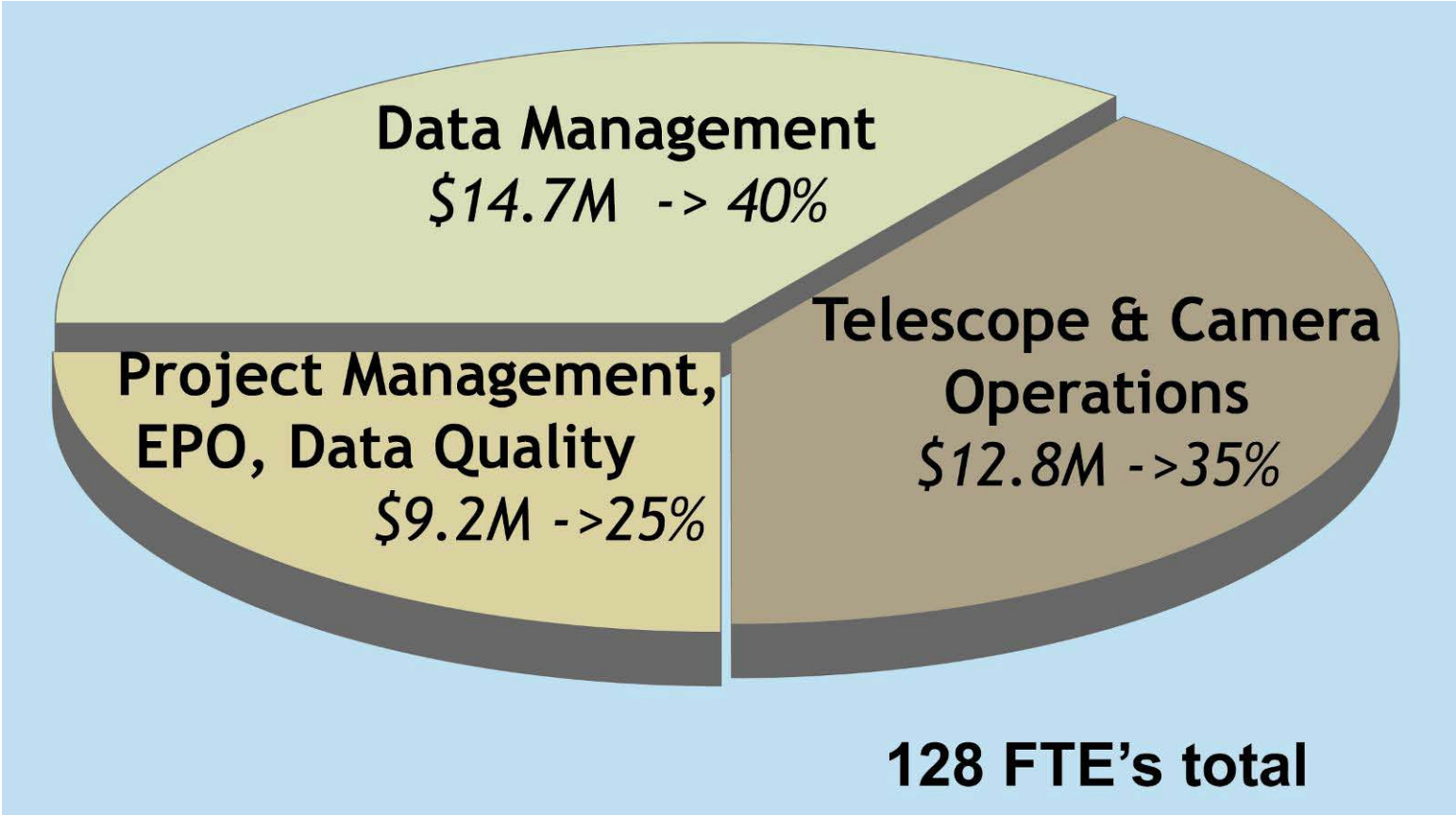

20 minute exposure on 8 m Subaru telescope
Point spread width 0.52 arcsec (FWHM)
1 arcminute

LSST Science Book

- Systematically explores science that can be done with LSST data
- 600pp, 245 authors...
- ArXiv:0912.0201
- This science requires precise control of data systematics – of which more later...



Operations and data budget



Managing All That Data

LSST Data Management System

Data Access Centers
U.S. (3) and Chile (1)
45 TFLOPS, 87 PB

Archive Center
NCSA, Champaign, IL
100 to 250 TFLOPS, 75 PB

Long-Haul Communications
Chile - U.S. & w/in U.S.
2.5 Gbps avg, 10 Gbps peak

Mountain Summit/Base Facility
Cerro Pachon, La Serena, Chile
10x10 Gbps fiber optics
25 TFLOPS, 150 TB

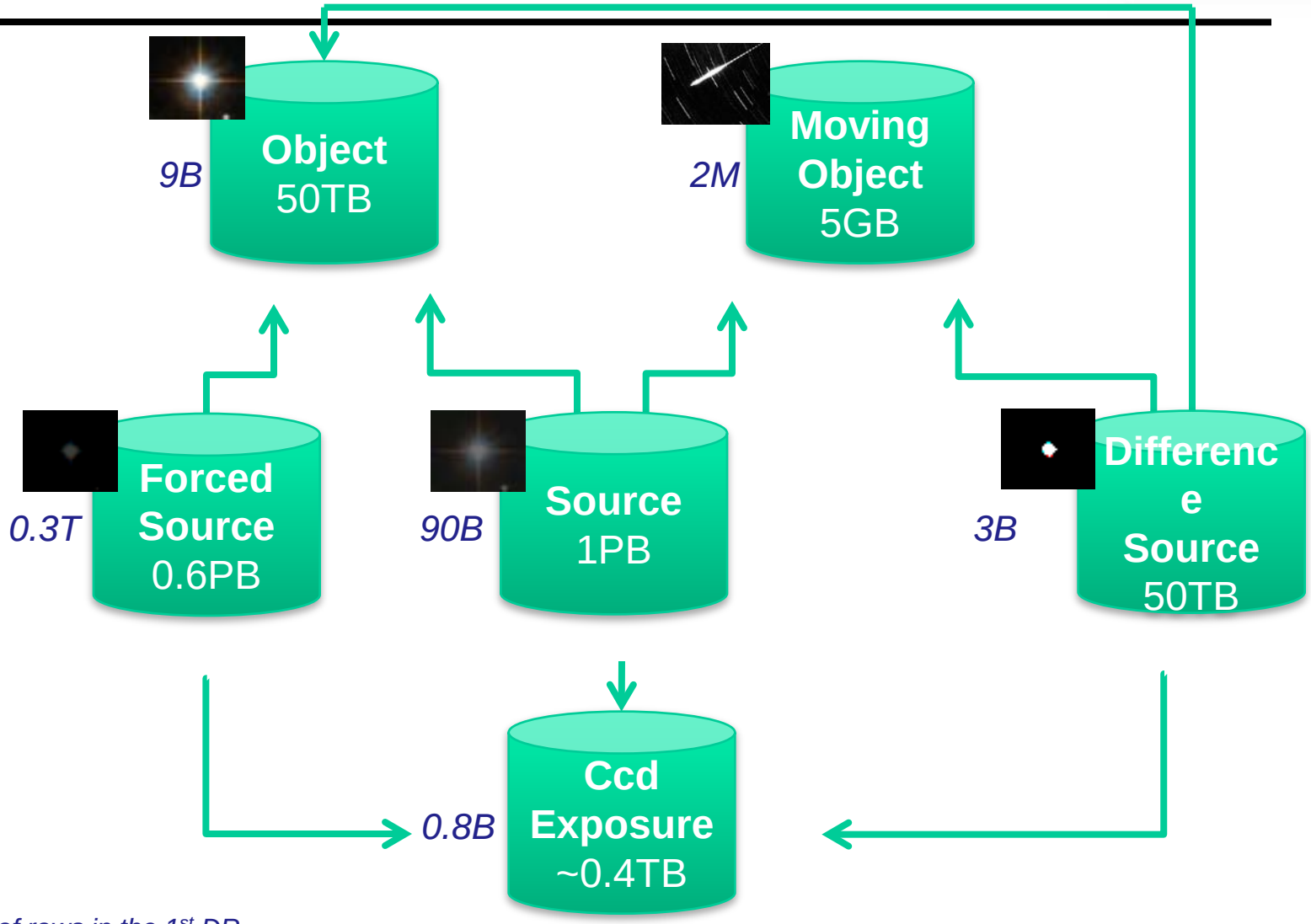
1 TFLOPS = 10^12 floating point operations/second

1 PB = 2^50 bytes or ~10^15 bytes

Level 1, 2, and 3 Data Products

- **Level 1 data products are produced continuously during an observing night, and include catalogs and images. Examples:**
  - **Raw Images**
  - **Transient Alerts**
- **Level 2 data products are produced as part of a Data Release, and include catalogs and images. Examples:**
  - **Coadded images**
  - **Optimally measured small galaxy shapes**
- **Level 3 data products are produced from L1 and L2 data products to support science goals that are not part of LSST's requirements. Examples:**
  - **Classification of variable stars**
  - **Morphological classification of large galaxies**

Catalogs are Implemented as Database Tables



Number of rows in the 1st DR

So Let’s Just Assume We’ve Done All That…

The Real Challenge Has Just Begun!

We must assess and control the quality of the data

- **Wide range of time scales, from a few minutes to years**
- **Wide range of science perspectives on quality**
  - Per-image pixel statistics at one extreme
  - Entire catalog in the middle
  - Consistency between LSST catalog and catalogs from other surveys at the other extreme

That's been done before, but…

- **Data volumes have grown enormously**
- **Operating budgets (=people) have not**
- **Data / person has therefore grown enormously as well**
- **IMPLICATION: Most data will never be directly looked at by a human. This is uncharted territory**

But Wait! There's More, Much More

- **Growth in number of objects in catalog, and precision of measurements, means statistical errors become potentially small**
  - **New science is enabled, but…**
  - **More subtle systematic errors are now visible in the data – if you know where to look**
  - **If they are not found, the science can be compromised**
- **Essentially, we are looking for correlations in some high dimensional space**
  - **Example: LSST Object catalog has about 20 billion rows, each with about 200 attributes**
  - **Humans are pretty good at finding correlations only in low dimensional spaces, when assisted by visualization tools, eg**
    - HR diagram
    - Fundamental plane for elliptical galaxies
  - **Big mismatch!**

Example: LSST Object catalog

- **The LSST survey will have about 20 billion Objects in its catalog**
  - **Each Object is represented by a row in the Object table**
  - **Each row has ~200 attributes which describe its mean properties**
    - 28 attributes are filter independent, eg name, mean ra and dec positions and proper motions, parameterization of variability
    - 35 attributes for each of 6 filters describe wavelength-dependent properties, eg mean calibrated magnitude, shape of galaxies
- **Each Object is associated with ~600 individual measurements**
  - **Each measurement is described by a row in the Source table**
  - **Each row has ~100 attributes**
  - **Among other things, gives the Object's time-dependent behavior**

How to deal with THAT?

What Tools do we have?

- Dimensionality reduction is clearly the biggest need
- Statistical anomaly detection can be built on top of dimensionality reduction
- Astronomers are lacking good tools at present
- Elsewhere, though, there are many
  - Network intrusion detection
  - Financial transactions
  - Monitoring of web services
  - Monitoring of water systems
  - …
- None of these tools are likely to be directly usable, but central ideas may carry over

Extending beyond a single survey

- **Astronomy will increasingly be done by combining information from multiple surveys**
  - **Extend wavelength coverage**
  - **Combine spectra with imaging**
  - **Etc**
- **Cross-matching is challenging, and brings in possible systematics of its own**
- **Dimensionality becomes even higher**

Conclusions and Topics for Discussion

- Gathering the raw data and pushing it through sophisticated pipelines is just the beginning
- The major challenges occur after that
  - How to make effective use of small numbers of people to monitor the quality of the data
  - Detecting low-amplitude systematics in data of very high dimensionality
  - Reliable sorting of the wheat from the chaff in the time domain
- Meeting these challenges will doubtless require importing ideas from other fields, and sustained intellectual effort within the astronomical community
- Computational resources required will ultimately dwarf those required for data acquisition and pipeline processing