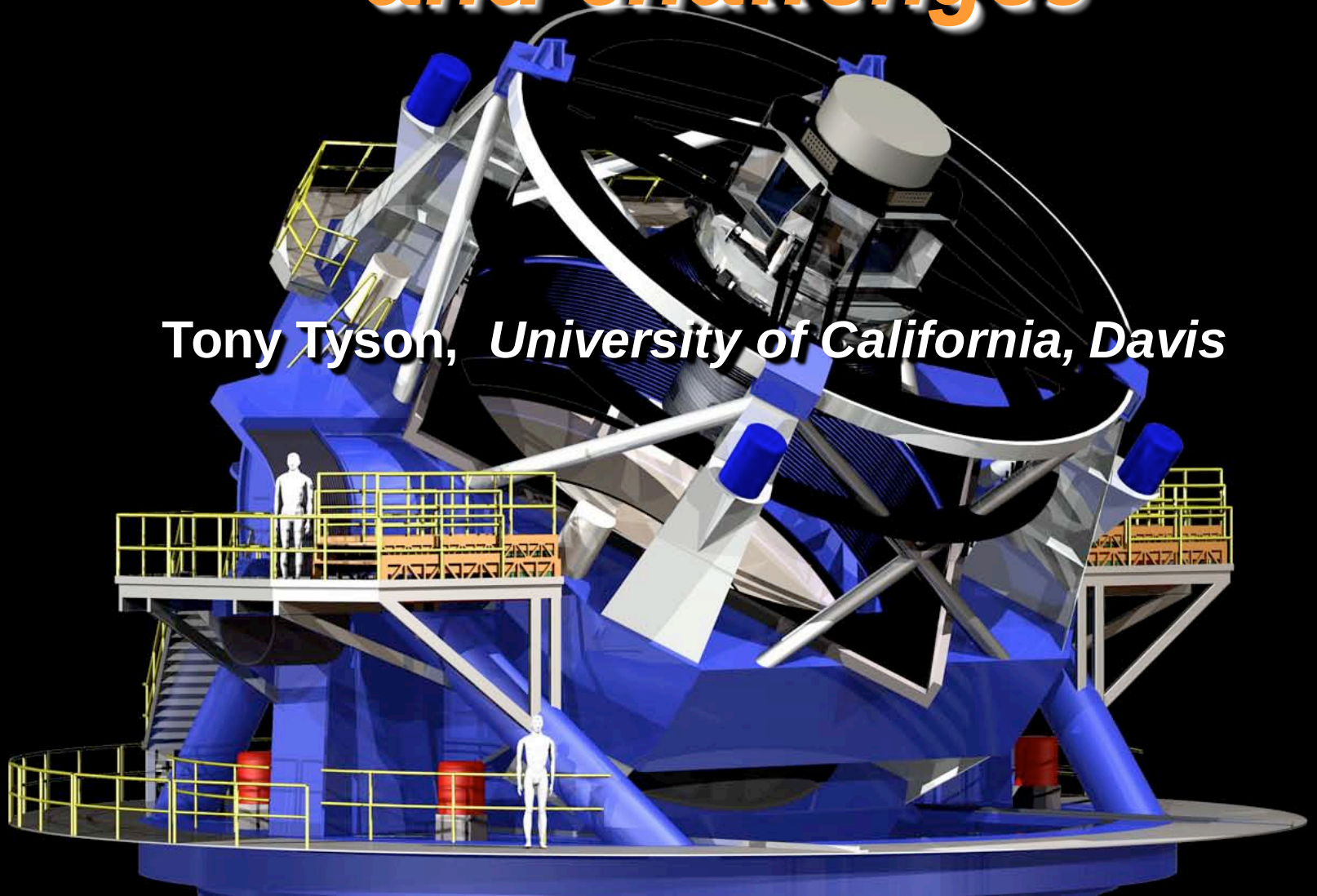
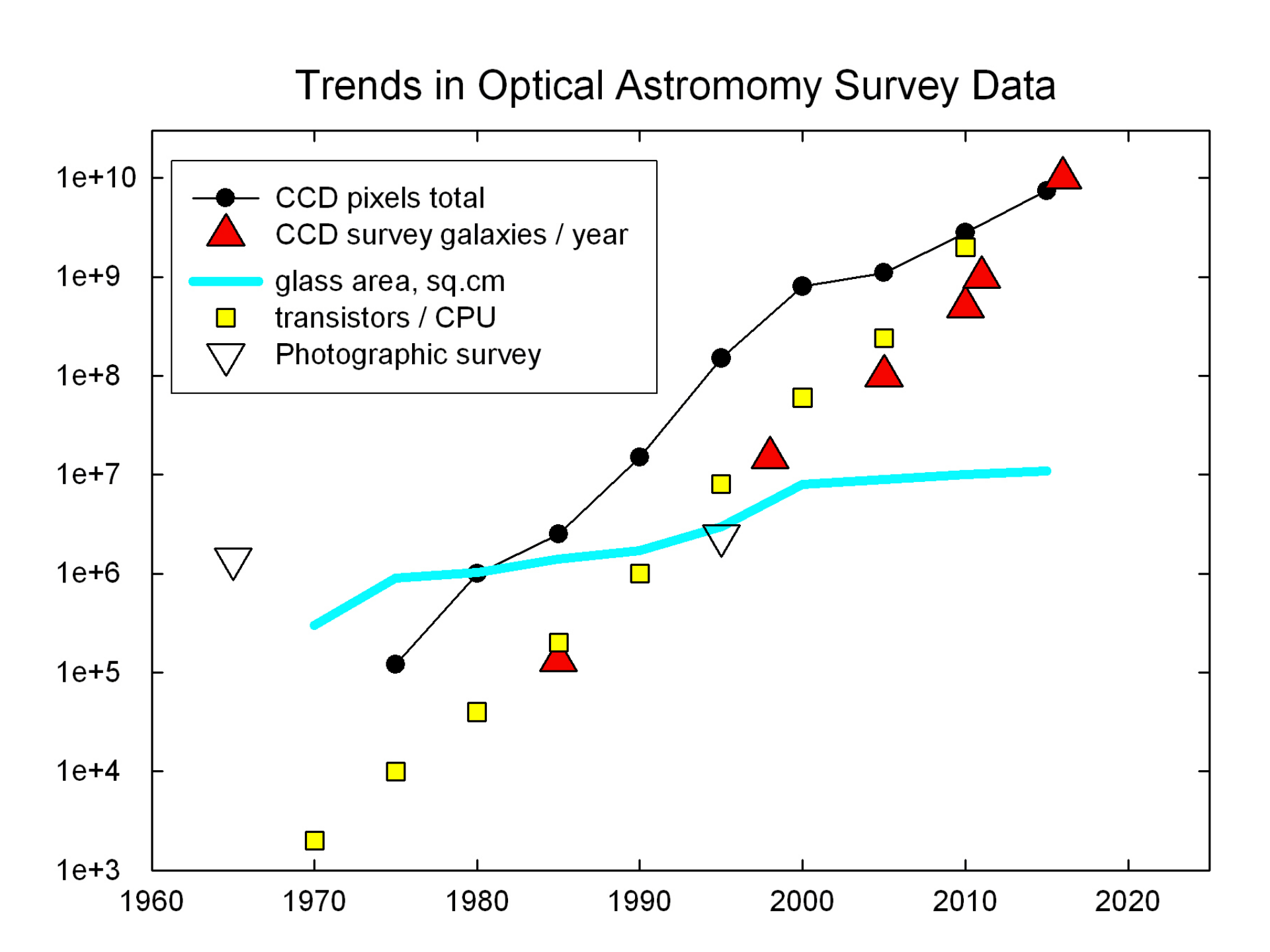
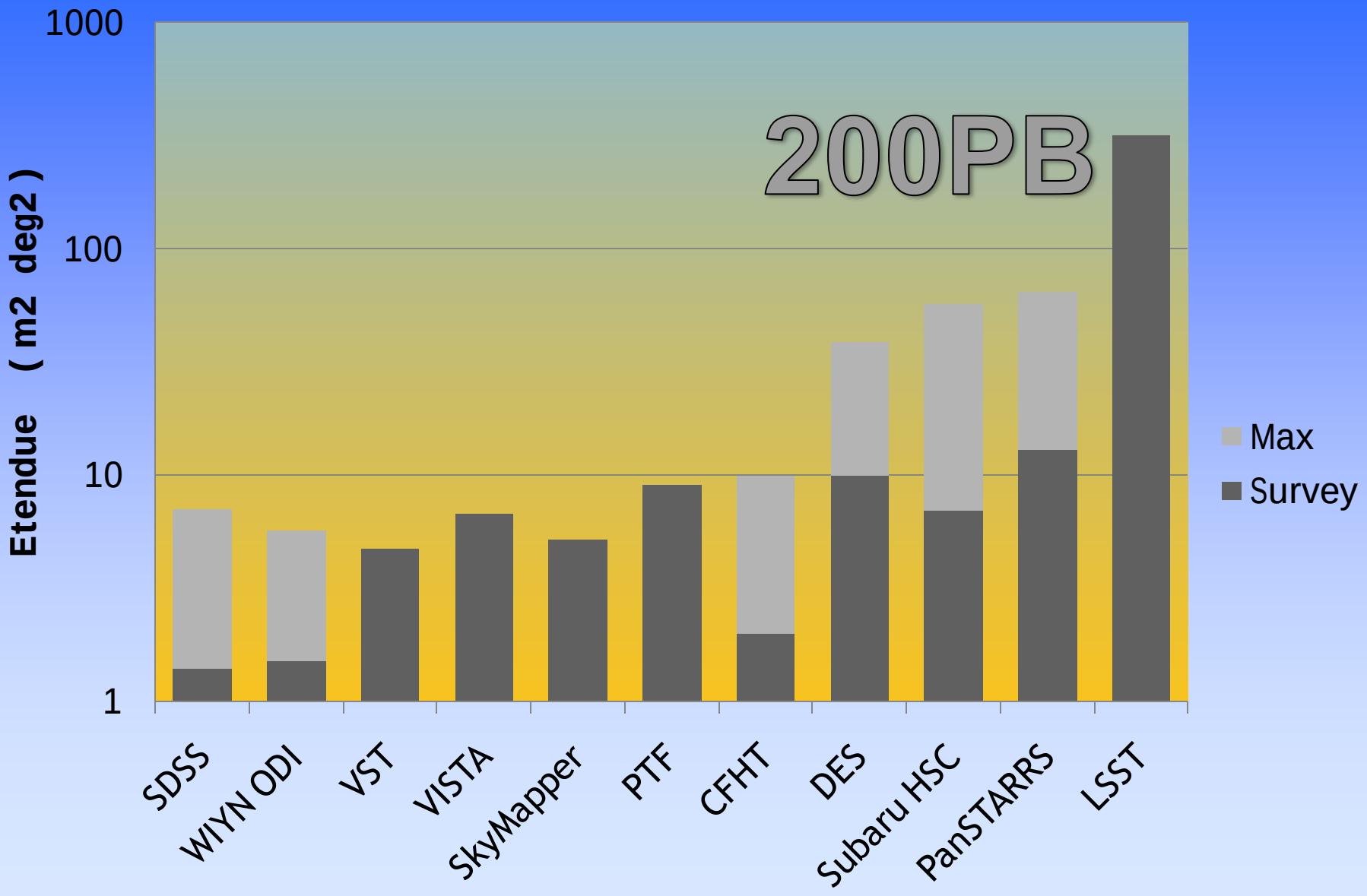
LSST: Petascale opportunities and challenges



Tony Tyson, *University of California, Davis*

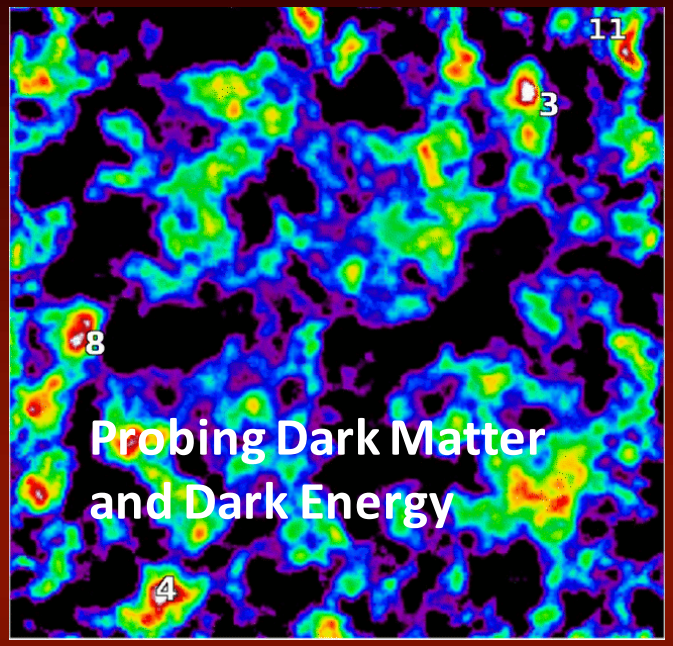

Trends in Optical Astromomy Survey Data
CCD pixels total
CCD survey galaxies / year
glass area, sq.cm
transistors / CPU
Photographic survey
1e+10
1e+9
1e+8
1e+7
1e+6
1e+5
1e+4
1e+3
1960
1970
1980
1990
2000
2010
2020

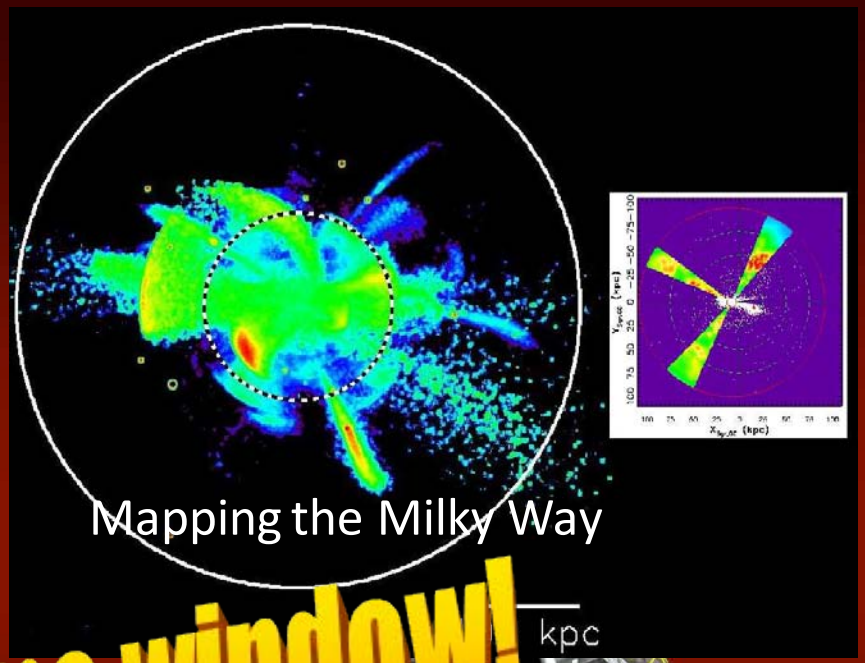
Relative data volume from survey telescopes & cameras



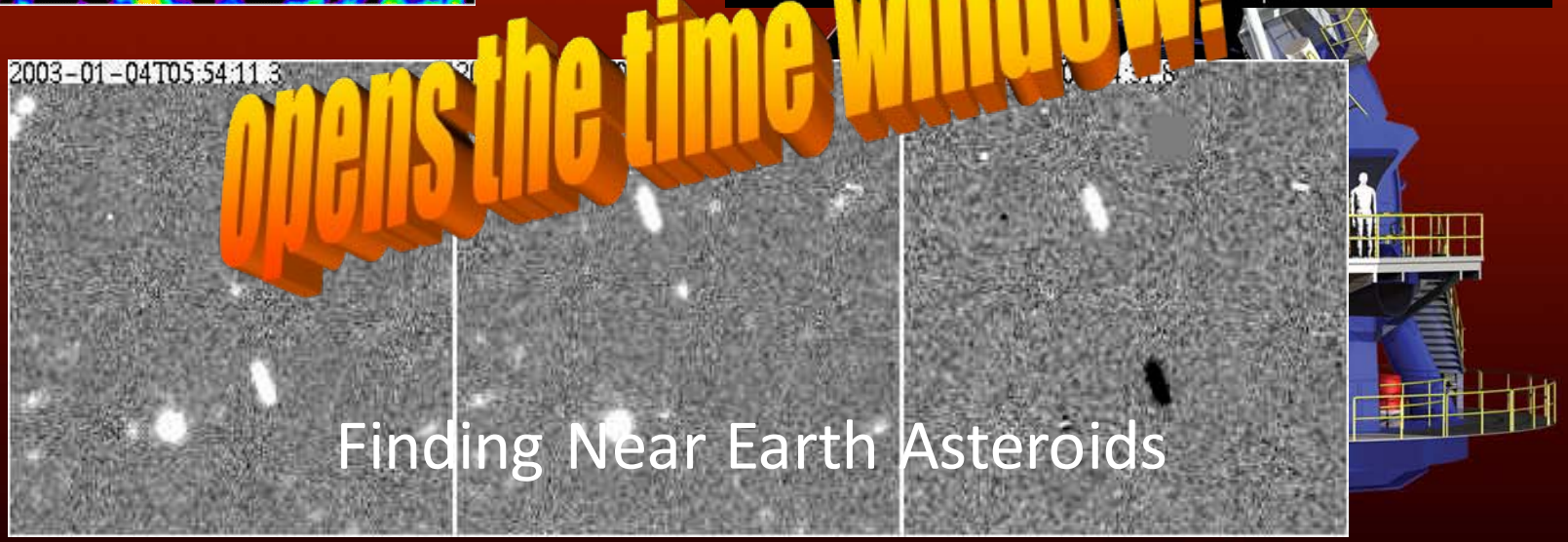
The new sky

Large Synoptic Survey Telescope

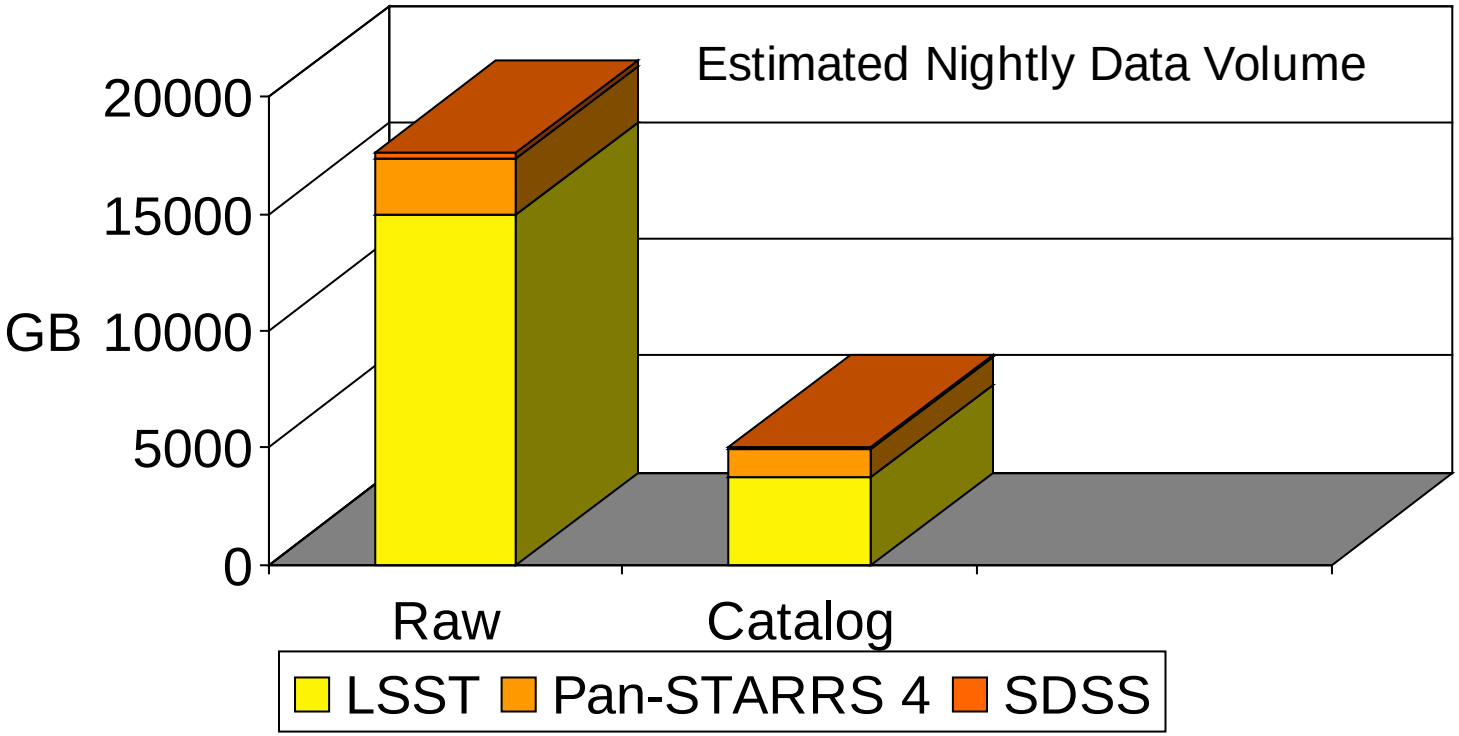




opens the time window!

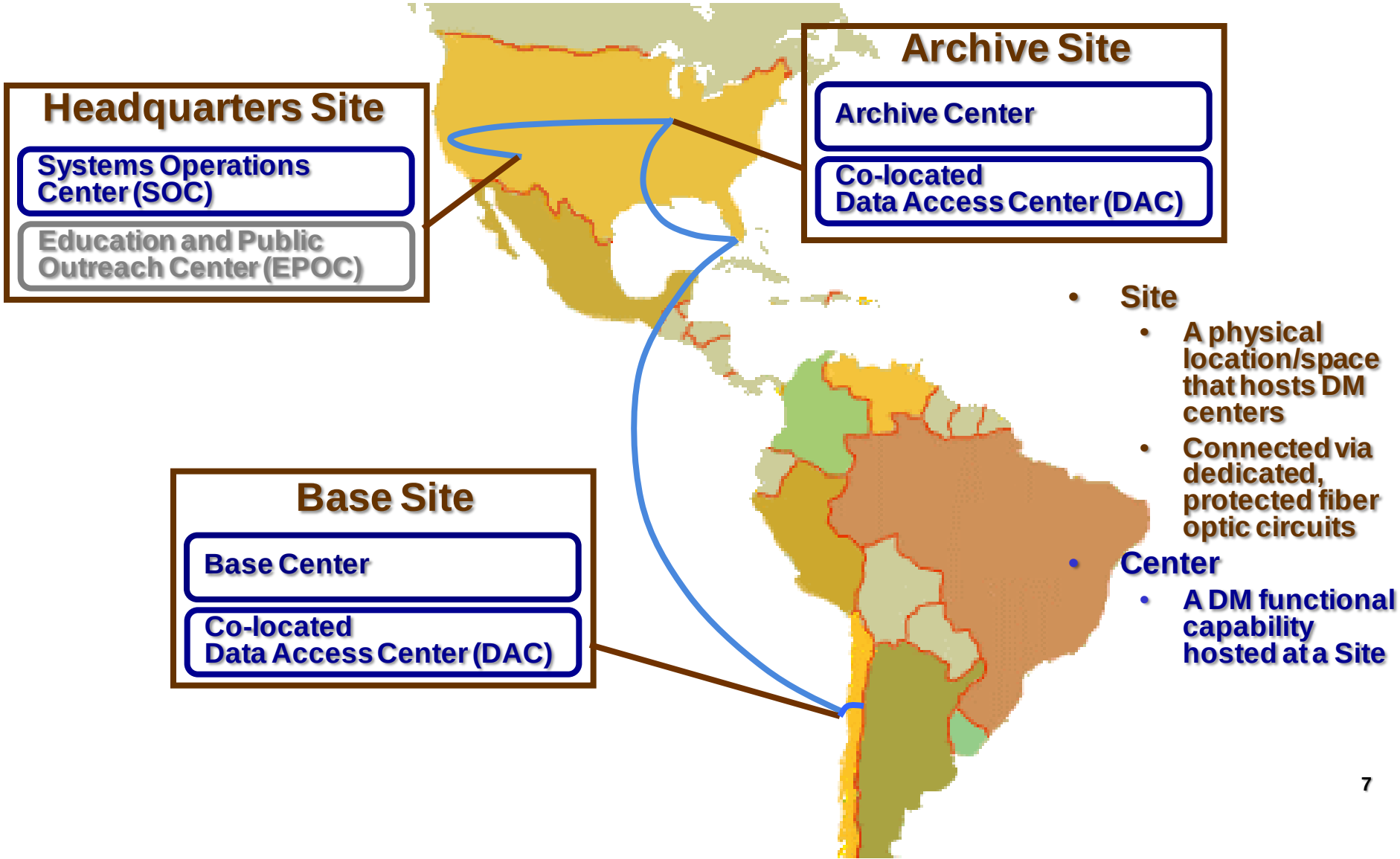


Data volumes & rates are unprecedented in astronomy



LSST will make tens of trillions photometric observations of tens of billions of objects

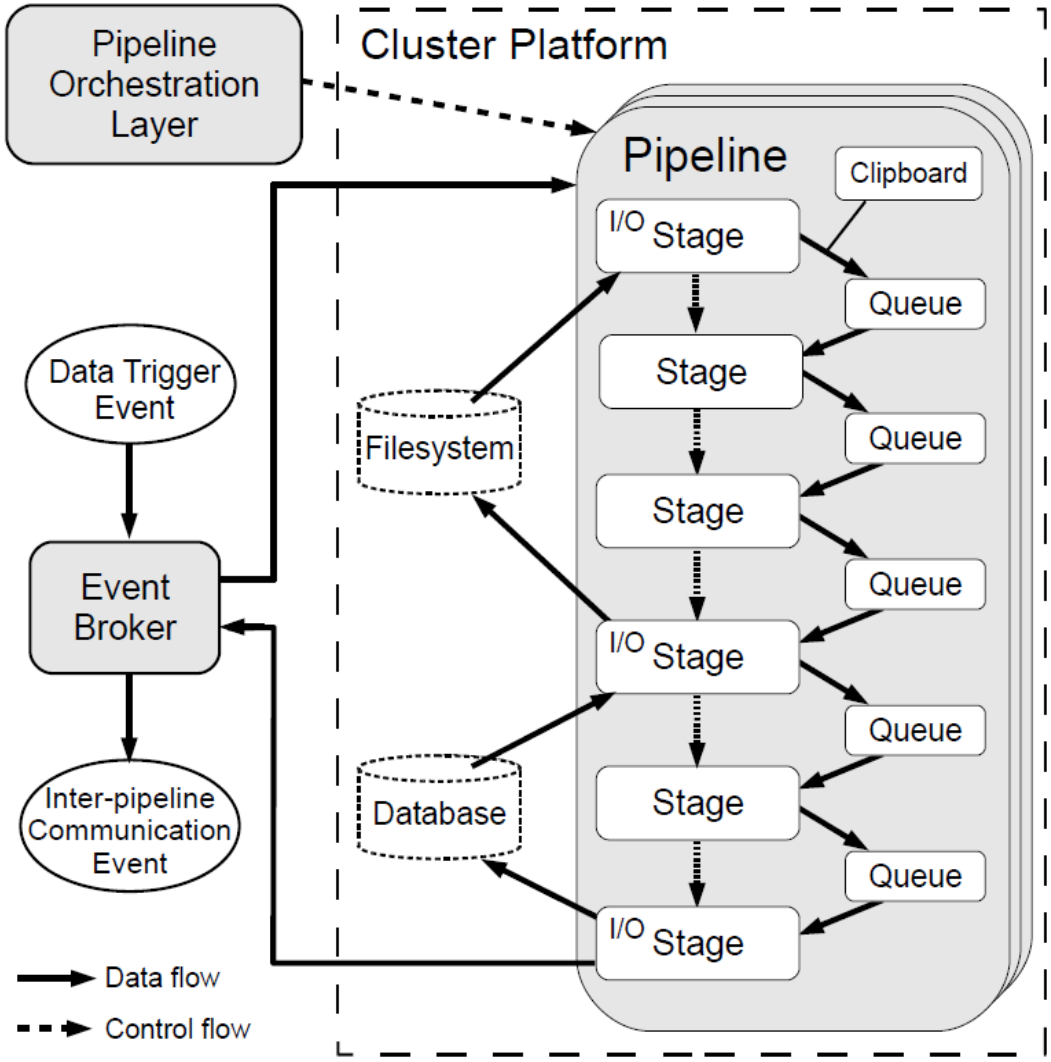
DM System is widely distributed



- Site
  - A physical location/space that hosts DM centers
  - Connected via dedicated, protected fiber optic circuits
- Center
  - A DM functional capability hosted at a Site

DM System relies on large-scale computational parallelism

- With few exceptions, LSST pipeline processing is "embarassingly parallel"
  - 3024 parallel image readouts
  - O(10^8) sky tiles
  - O(10^9) objects
- Computational clusters are well matched to the available parallelism
  - 5000 cores at Base
  - 12000 (yr1) – 33000 (yr10) cores at Archive
- Middleware implements flexible pipeline/production model of parallelism

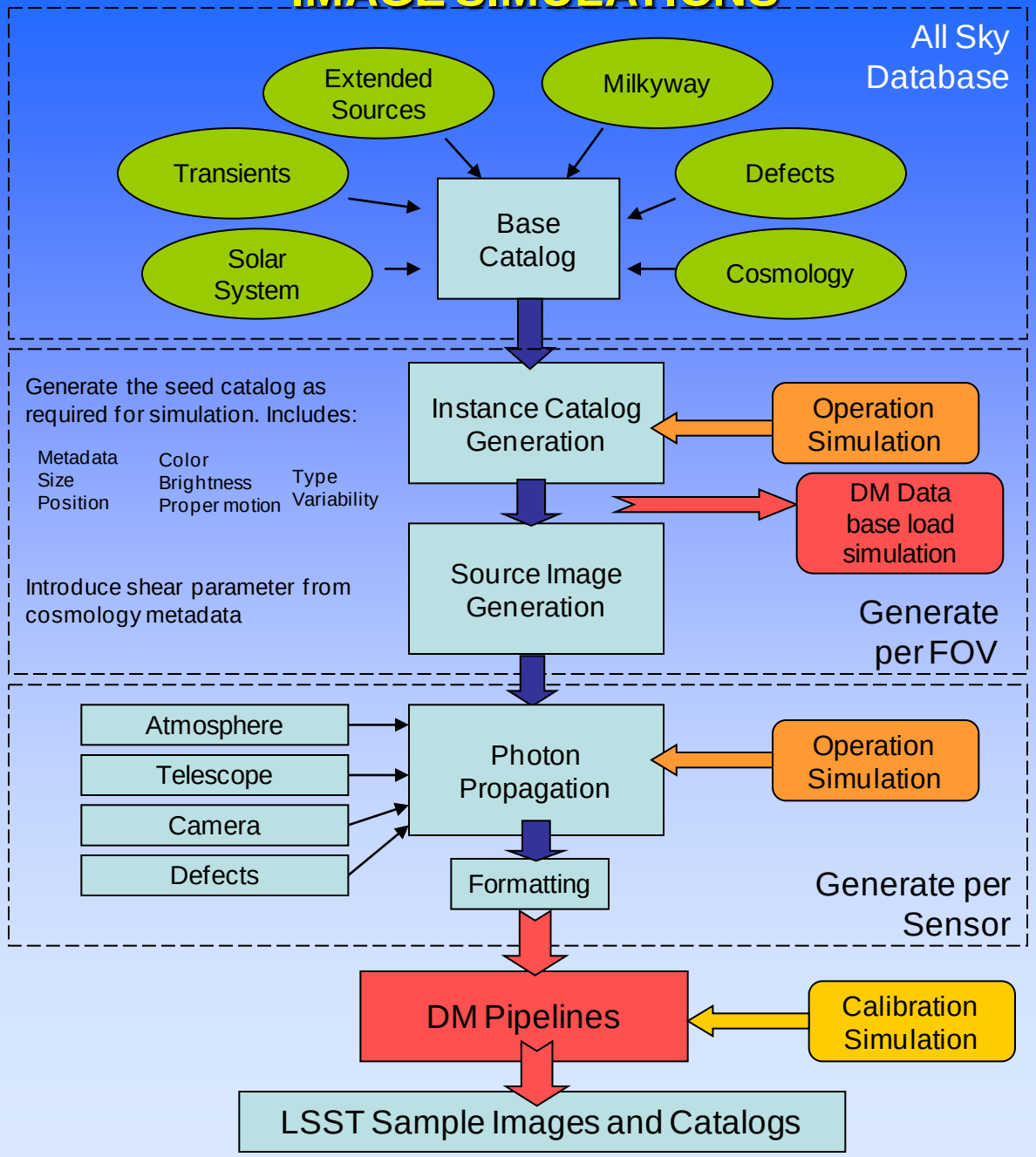


DATA PRODUCTS

Application Layer -
Generates open, accessible data products with fully documented quality

Processing Cadence	Image Category (files)	Catalog Category (database)	Alert Category (database)
Nightly	Raw science image Calibrated science image Subtracted science image Noise image Sky image Data quality analysis	Source catalog (from difference images) Object catalog (from difference images) Orbit catalog Data quality analysis	Transient alert Moving object alert Data quality analysis
Data Release (Annual)	Stacked science image Template image Calibration image RGB JPEG Images Data quality analysis	Source catalog (from calibrated science images) Object catalog (optimally measured properties) Data quality analysis	Alert statistics & summaries Data quality analysis

IMAGE SIMULATIONS

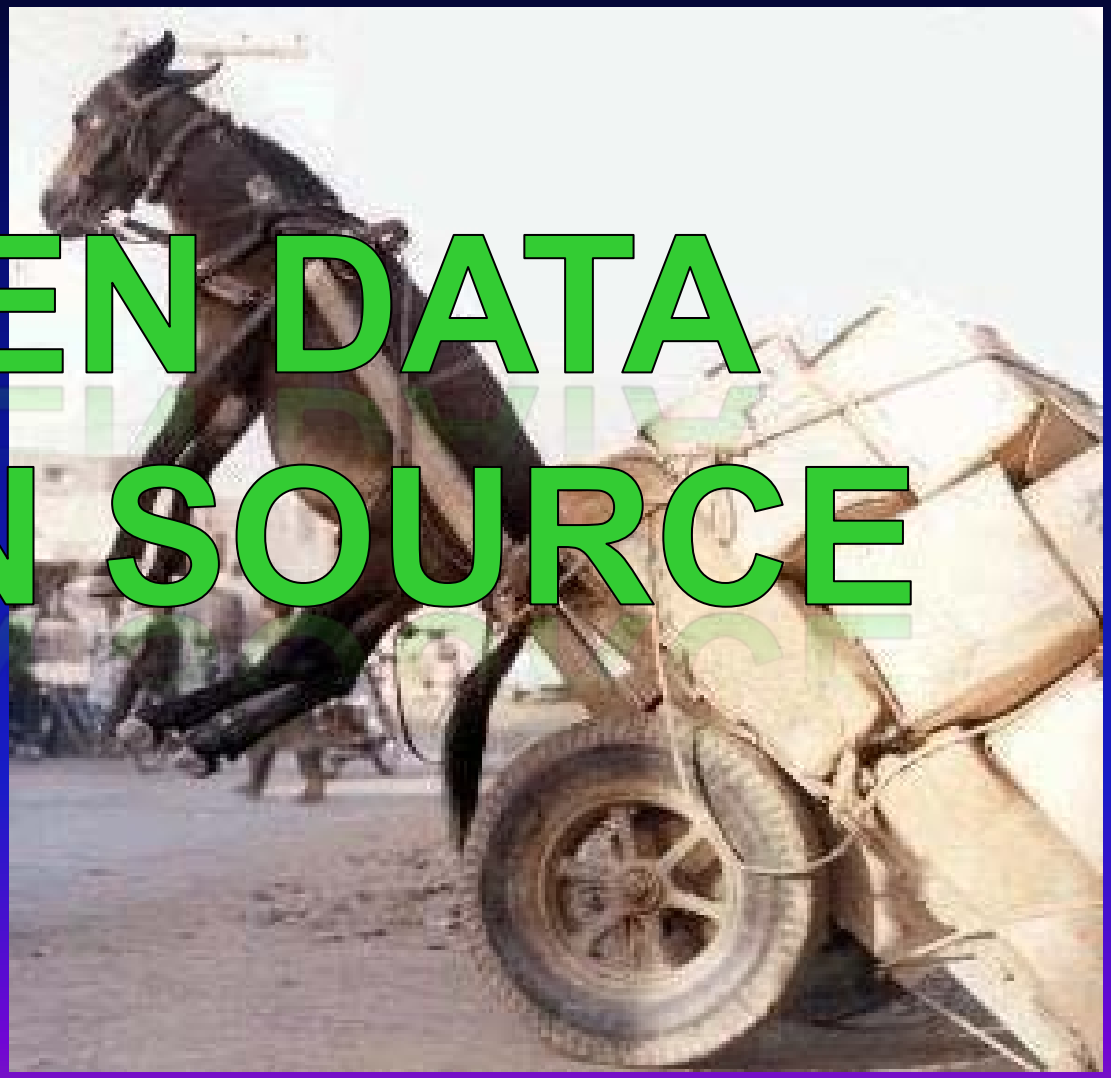


The Data Challenge

- 3 Terabytes per hour that must be mined in real time.
- 20 billion objects will be monitored for important variations in real time.
- A new approach must be developed for knowledge extraction in real time.

The Data Challenge

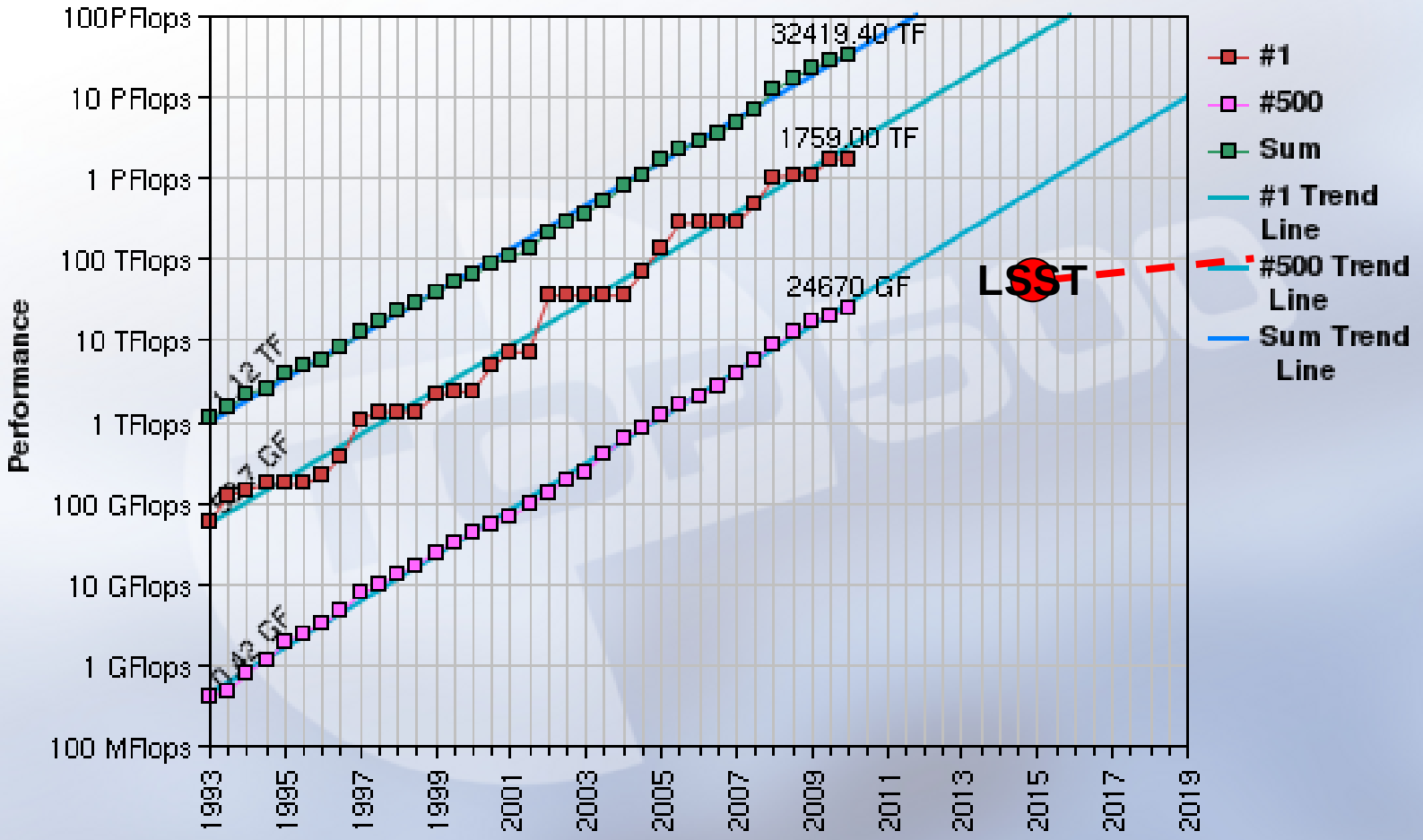
- ~3 Terabytes per hour that must be mined in real time.
- 20 billion objects will be monitored for important variations in real time.
- A new approach must be developed for knowledge extraction in real time.



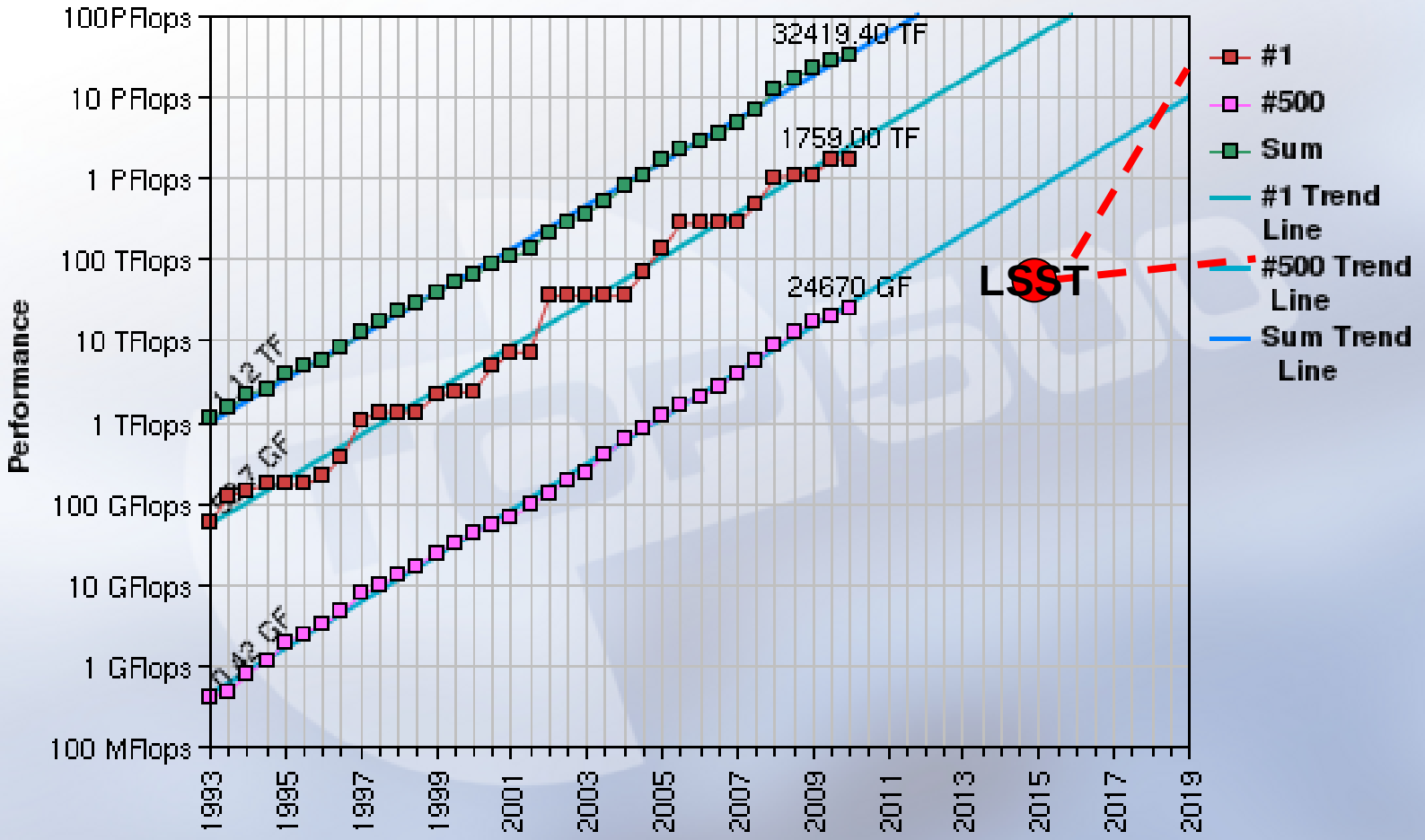
OPEN DATA

OPEN SOURCE

Projected Performance Development



Projected Performance Development



Analytics

- **Complex computations**
  - 100s of attributes per query
- **Iterative, successively more restrictive**
- **Curiosity driven questions**
- **3 major query types**
  - Needle in haystack
  - Correlations
  - Time series

Science at the Limit

- Much of the breakthrough science using surveys (imaging or spectroscopy) have occurred at the limits of the surveys
  - Sample incompleteness
  - Systematic errors

LSST Wide-Fast-Deep survey

- 4 billion galaxies with redshifts

- Time domain:
  1 million supernovae
  1 million galaxy lenses
  1 billion moving objects
  new phenomena

UNPRECEDENTED STATISTICAL PRECISION

LSST Wide-Fast-Deep survey

- 4 billion galaxies with redshifts

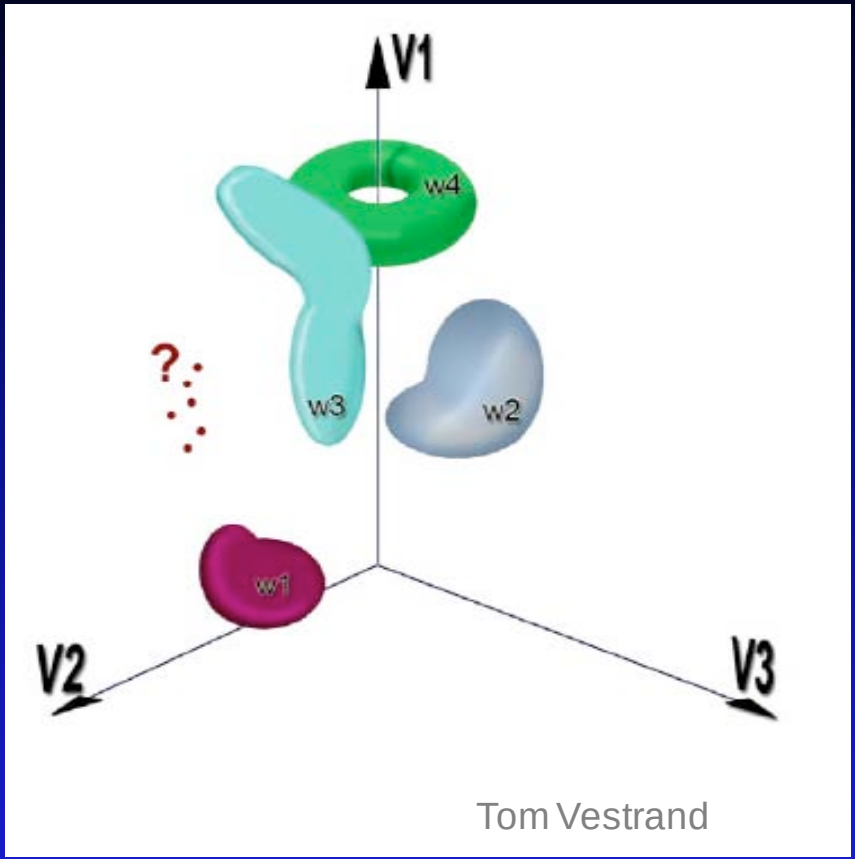
- Time domain:
  - 1 million supernovae
  - 1 million galaxy lenses
  - 1 billion moving objects
  - new phenomena

UNPRECEDENTED SYSTEMATICS CONTROL

Major opportunity and challenge:

Discovering The Unexpected

- Characterize the known (clustering)
- Assign the new (classification)
- Discover the unknown (outlier detection)



Benefits of very large data sets:

- **best statistical analysis of "typical" events**
- **automated search for "rare" events**

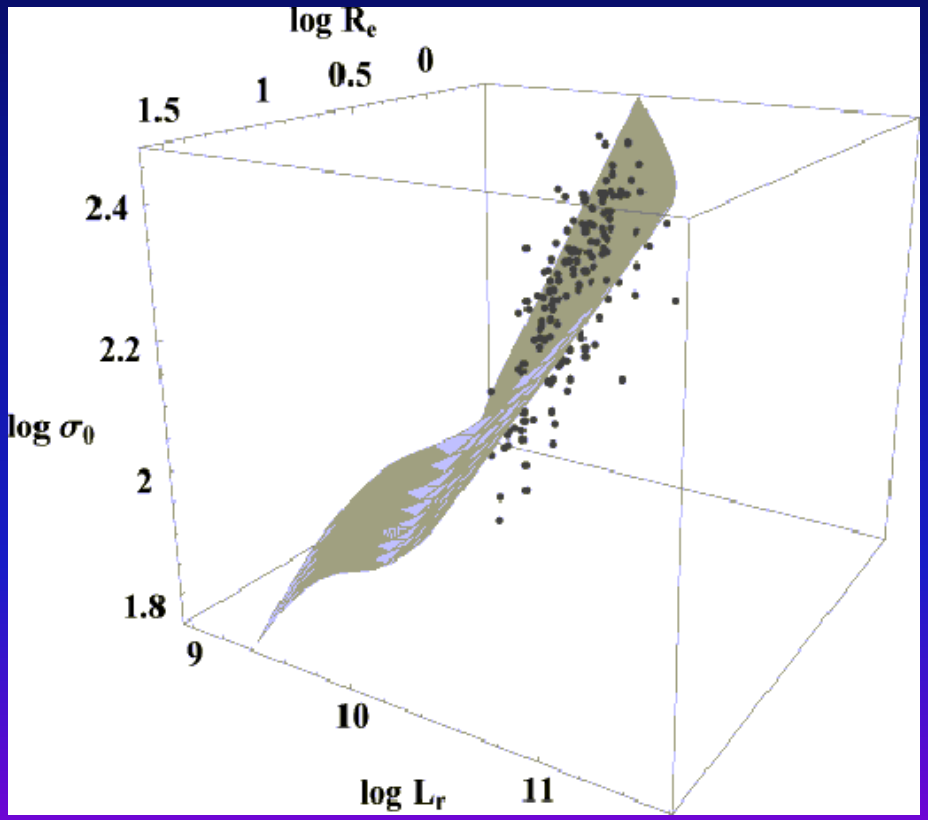
The dimension reduction problem:

Finding correlations and "fundamental planes" of parameters

- **The Curse of High Dimensionality !**

– Are there combinations (linear or non-linear functions) of observational parameters that correlate strongly with one another?

– Are there eigenvectors or condensed representations (e.g., basis sets) that represent the full set of properties?



Automated discovery

Data exploration

DISCOVERING THE UNEXPECTED

This is required also for automated Data Quality Assessment

How To Learn More / Get Involved?

➢ LSST lsst.org

- Check out LSST database trac at http://dev.lsstcorp.org/trac/wiki/LSSTDatabase

➢ XLDB

- XLDB4 (Oct 6-7@SLAC)
- Read past XLDB reports http://www-conf.slac.stanford.edu/xldb
- Share your use cases, join the community

Open conference starting this year

➢ SciDB

- Check out http://scidb.org
- Try it out

1st public release