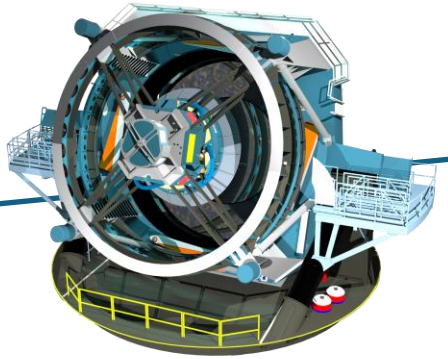
This Laptop is Inadequate

(and intro to data organization)

Bryan Scott

LSST-DA Data Science Fellowship Program Session 21

University of Illinois, Urbana-Champaign

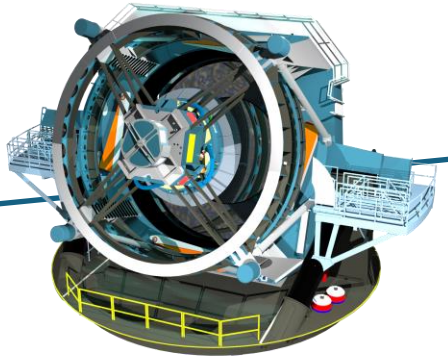


Rubin-LSST in 3 numbers

37 billion sources

1000 Observations

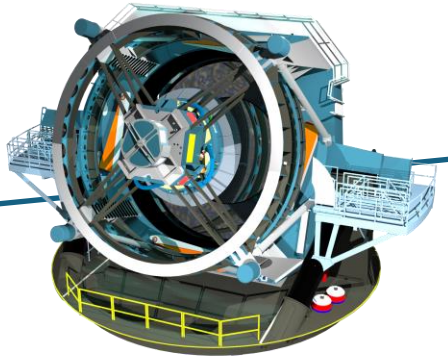
10 years



Rubin-LSST in 1 number

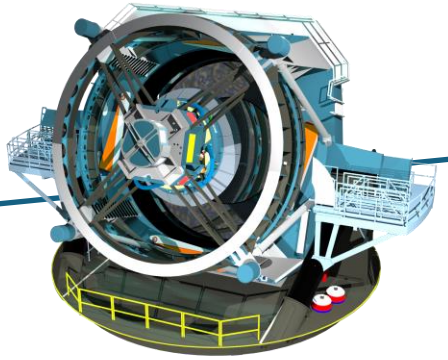
37,000,000,000,000

(observations)



Suppose you could describe every source detected by LSST with a single number. Assuming you are on a computer with a 64 bit architecture, to within an order of magnitude, how much RAM would you need to store every LSST source within your laptop's memory?

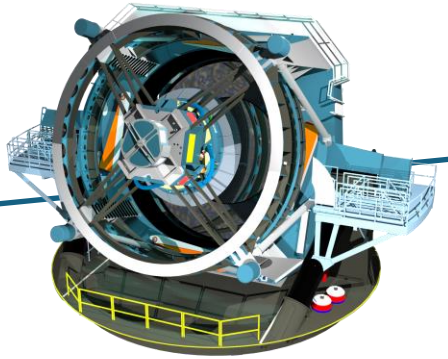
Bonus question - can you think of a single number to describe every source in LSST that could produce a meaningful science result?



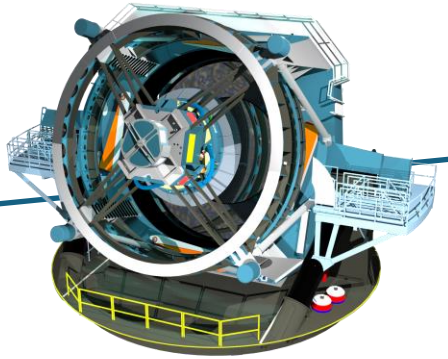
Suppose you could describe every source detected by LSST with a single number. Assuming you are on a computer with a 64 bit architecture, to within an order of magnitude, how much RAM would you need to store every LSST source within your laptop's memory?

Bonus question - can you think of a single number to describe every source in LSST that could produce a meaningful science result?

$$\frac{64\,\text{bit}}{1\,\text{source}} \times \frac{1\,\text{GB}}{8 \times 10^9\,\text{bit}} \times 3.7 \times 10^{10}\,\text{sources} \approx 296\,\text{GB}$$



But that raises the question - how should you analyze LSST data?



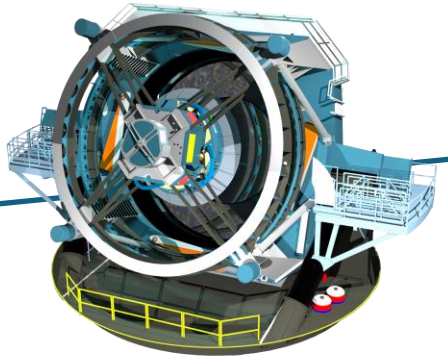
But that raises the question - how should you analyze LSST data?

By buying a large desktop? (impractical to ask of everyone working on LSST)

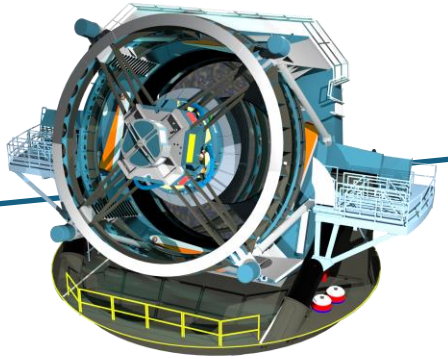
On a local supercomputer? (not a bad idea, but not necessarily equitable)

In the cloud? (AWS is expensive)

On computers that LSST hosts/maintains? (probably the most fair, but this also has challenges)



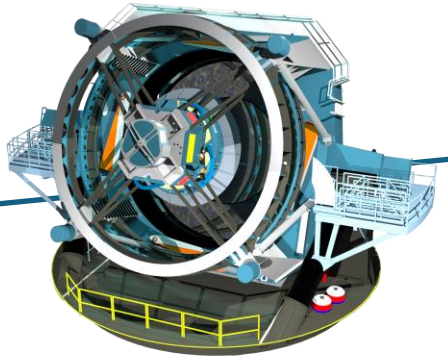
So far we have been focused on only a single aspect of computing: storage.
(your laptop sucks at storage)



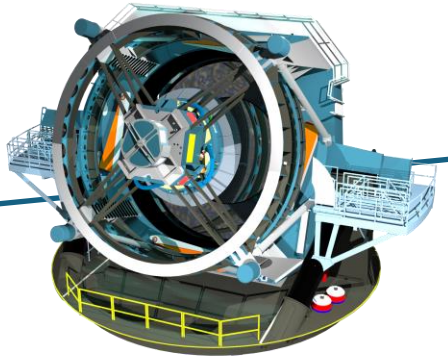
So far we have been focused on only a single aspect of computing: storage. (your laptop sucks at storage)

But here's the thing - your laptop is also incredibly slow.

Supposing for a moment that you could hold all (or even a significant fraction) of the information from LSST in memory on your laptop, you would still be out of luck, as you would die before you could actually process the data and make any meaningful calculations.

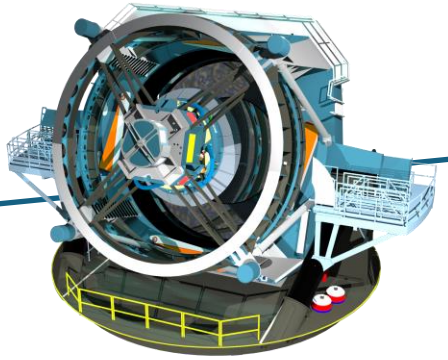


How long would it take to perform basic processing of all of LSST on your laptop? The bare minimum for image processing includes bias (subtraction) and flat-field (division) corrections. Assume your laptop has a single 3 GHz processor that requires 1 tick to perform a single addition operation and 4 ticks to perform a single multiplication operation.



How long would it take to perform basic processing of all of LSST on your laptop? The bare minimum for image processing includes bias (subtraction) and flat-field (division) corrections. Assume your laptop has a single 3 GHz processor that requires 1 tick to perform a single addition operation and 4 ticks to perform a single multiplication operation.

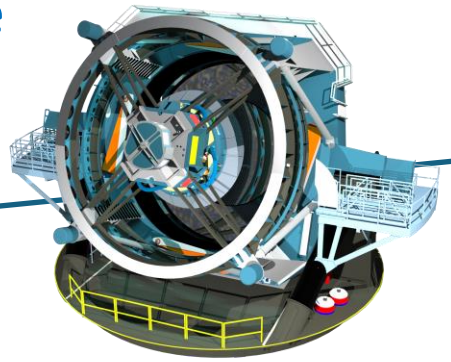
$$\frac{3.2 \times 10^9\,\mathrm{pix}}{\mathrm{field\,obs}} \times \frac{\mathrm{field}}{10\,\mathrm{deg}^2} \times 20{,}000\,\mathrm{deg}^2 \times \frac{5\,\mathrm{ticks}}{\mathrm{pix}} \times \frac{\mathrm{s}}{3 \times 10^9\,\mathrm{ticks}} \times 1000\,\mathrm{obs} \approx 4\,\mathrm{months}$$



How long would it take to perform basic processing of all of LSST on your laptop? The bare minimum for image processing includes bias (subtraction) and flat-field (division) corrections. Assume your laptop has a single 3 GHz processor that requires 1 tick to perform a single addition operation and 4 ticks to perform a single multiplication operation.

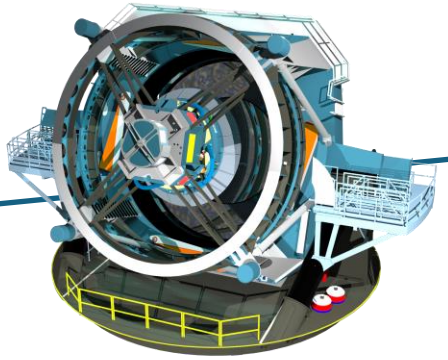
$$\frac{3.2 \times 10^9\,\mathrm{pix}}{\mathrm{field\,obs}} \times \frac{\mathrm{field}}{10\,\mathrm{deg}^2} \times 20{,}000\,\mathrm{deg}^2 \times \frac{5\,\mathrm{ticks}}{\mathrm{pix}} \times \frac{\mathrm{s}}{3 \times 10^9\,\mathrm{ticks}} \times 1000\,\mathrm{obs} \approx 4\,\mathrm{months}$$

A more detailed calculation shows it takes ~30 s to fully process (bias, flat-field, astrometry, photometry, image subtraction…) 1M pixels (much of this is tied up in I/O). Using the same numbers from the previous example, LSST will take ~200 yr to process.



You are in luck, however, as you need not limit yourself to your laptop. You can take advantage of multiple computers, also known as parallel processing.

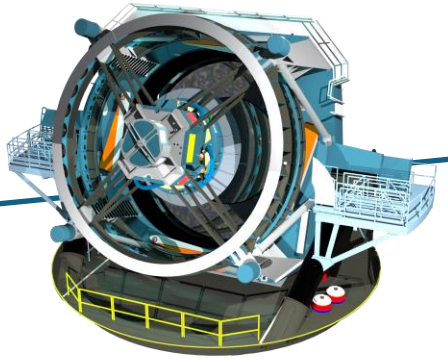
Robert Lupton, one of the primary developers of the LSST photometric pipeline, when asked "How many CPUs are being used to process LSST data?" replied, "However many are needed to process everything within 1 month."



You are in luck, however, as you need not limit yourself to your laptop. You can take advantage of multiple computers, also known as parallel processing.

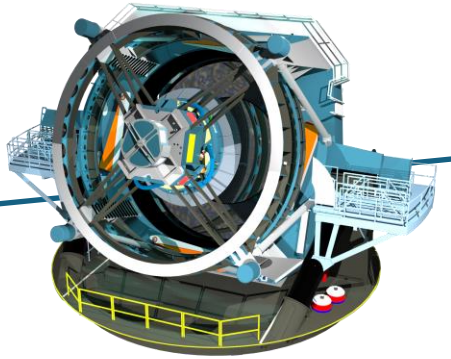
Robert Lupton, one of the primary developers of the LSST photometric pipeline, when asked "How many CPUs are being used to process LSST data?" replied, "However many are needed to process everything within 1 month."

The critical point here is that if you can figure out how to split a calculation over multiple computers, then you can finish any calculation arbitrarily fast with enough processors (to within some limits, like the speed of light, etc). We will come back to this in Lehman's lectures at the end of the week.



What is data?

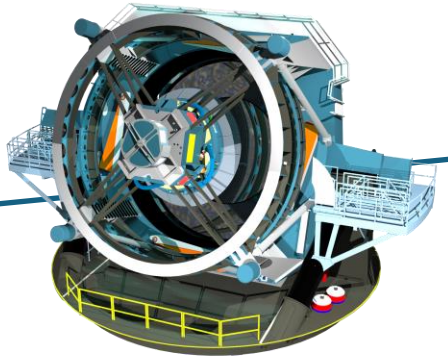
Take a minute to discuss with your partner



What is data?

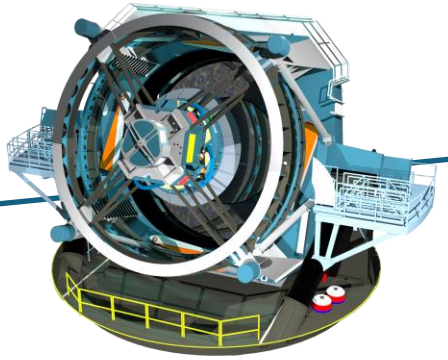
Take a minute to discuss with your partner

Data are constants.



This leads to another question:
Q - What is the defining property of a constant?
A - They don't change.

If data are constants, and constants don't change, then we should probably be sure that our data storage solutions do not alter the data in any way.

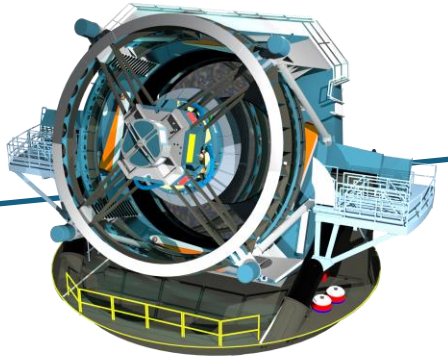


Text files have some advantages:

- anyone, anywhere, on any platform can read text files
- text files are easily inspected (and corrected) if necessary
- special packages are needed to read/write in binary
- binary files, which are not easily interpretable, are difficult to use in version control (and banned by some version control platforms)

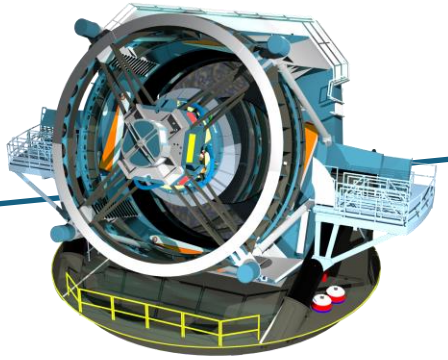
To summarize, here is my advice: think of binary as your (new?) default for storing data.

But, as with all things, consider your audience: if you are sharing/working with people that won't be able to deal with binary data, or, you have an incredibly small amount of data, csv (or other text files) should be fine.

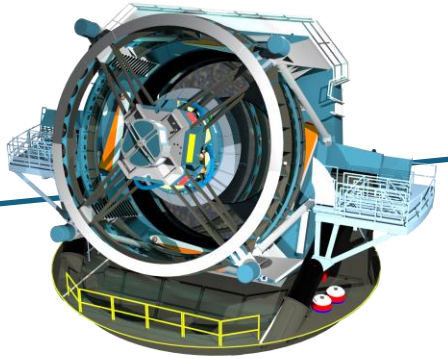


How would you organize the following: (a) 3 very deep images of a galaxy, (b) 4 nights of optical observations (~50 images night−1) of a galaxy cluster in the ugrizY filters, (c) images from a 50 night time-domain survey (~250 images night−1) covering 1000 deg2?

Take a minute to discuss with your partner

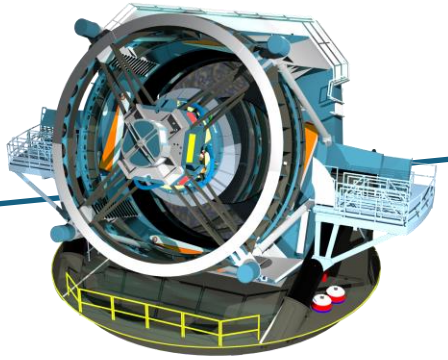


Keeping in mind that there are several suitable answers to each question – a few thoughts: (a) the 3 images should be kept together, probably in a single file directory.



Keeping in mind that there are several suitable answers to each question – a few thoughts: (a) the 3 images should be kept together, probably in a single file directory.

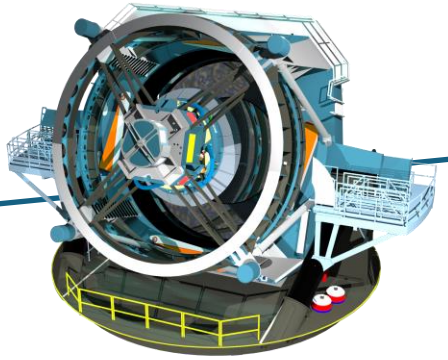
(b) With 200 images taken over the course of 4 nights, I would create a directory structure that includes every night (top level), with sub-directories based on the individual filters.



Keeping in mind that there are several suitable answers to each question – a few thoughts: (a) the 3 images should be kept together, probably in a single file directory.

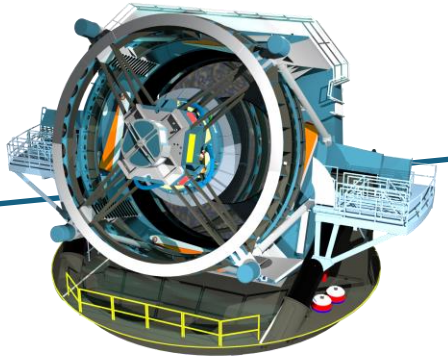
(b) With 200 images taken over the course of 4 nights, I would create a directory structure that includes every night (top level), with sub-directories based on the individual filters.

(c) Similar to (b), I'd create a tree-based file structure, though given that the primary science is time variability, I would likely organize the observations by fieldID at the top level, then by filter and date after that.

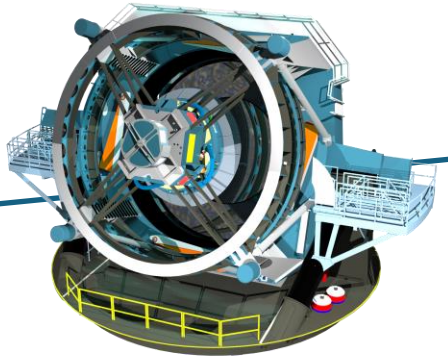


Similarly, how would you organize: (a) photometric information for your galaxy observations, (b) photometry for all the galaxies in your cluster field, (c) the observations/light curves from your survey?

Take a minute to discuss with your partner

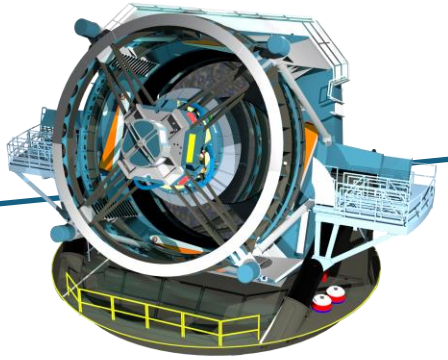


a) For 3 observations of a single galaxy, I would use… a text file (not worth the trouble for binary storage)



a) For 3 observations of a single galaxy, I would use… a text file (not worth the trouble for binary storage)

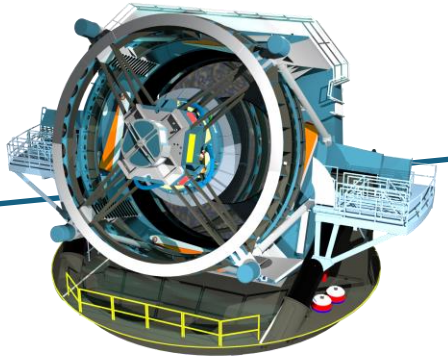
b) Assuming there are 5000 galaxies in the cluster field, I would store the photometric information that I extract for those galaxies in a *table*. In this table, each row would represent a single galaxy, while the columns would include brightness/shape measurements for the galaxies in each of the observed filters. I would organize this table as a pandas DataFrame (and write it to an hdf5 file).



a) For 3 observations of a single galaxy, I would use… a text file (not worth the trouble for binary storage)

b) Assuming there are 5000 galaxies in the cluster field, I would store the photometric information that I extract for those galaxies in a *table*. In this table, each row would represent a single galaxy, while the columns would include brightness/shape measurements for the galaxies in each of the observed filters. I would organize this table as a pandas DataFrame (and write it to an hdf5 file).

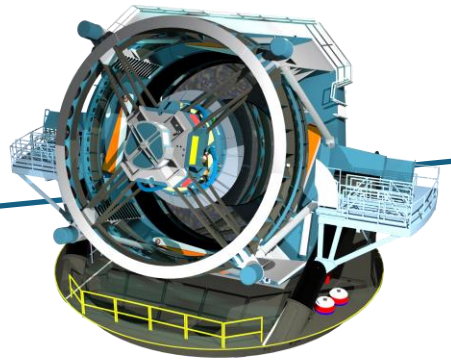
c) For the time-domain survey, the organization of all the photometric information is far less straight forward.



c) For the time-domain survey, the organization of all the photometric information is far less straight forward.

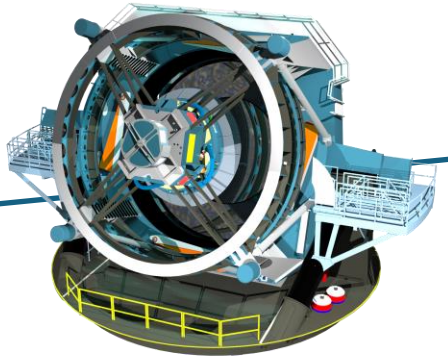
objID	RA	Dec	mean_mag	mean_mag_unc
0001	246.98756	-12.06547	18.35	0.08
0002	246.98853	-12.04325	19.98	0.21
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

objID	JD	filt	mag	mag_unc
0001	2456785.23465	r	18.21	0.07
0001	2456785.23469	z	17.81	0.12
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
0547	2456821.36900	g	16.04	0.02
0547	2456821.36906	i	17.12	0.05
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.



c) For the time-domain survey, the organization of all the photometric information is far less straight forward.

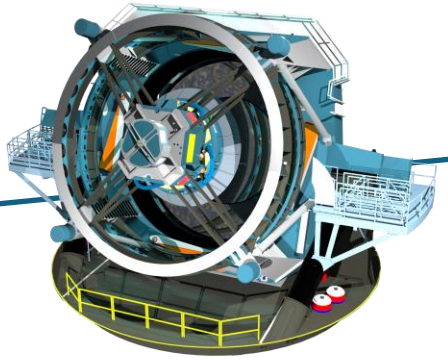
The critical thing to notice about these tables is that they both contain objID. That information allows us to connect the tables via a "join." This table, or relational, structure allows us to easily connect subsets of the data as a way to minimize storage (relative to having everything in a single table) while also maintaining computational speed.



Some concluding comments:

Typically, when astronomers (or data scientists) need to organize data into several connected tables capable of performing fast relational algebra operations they use a database. We will hear a lot more about databases over the next few days, so I won't provide a detailed introduction now.

One very nice property of (many) database systems is that provide an efficient means for searching large volumes of data that cannot be stored in memory . Whereas, your laptop, or even a specialized high-memory computer, would not be able to open a csv file with all the LSST observations in it.



In the problem notebook...