Astroinformatics
Proceedings IAU Symposium No. 325, 2016
M. Brescia, S.G. Djorgovski, E. Feigelson,
G. Longo & S. Cavuoti, eds.

doi:10.1017/S1743921317000825

The European perspective for LSST

Emmanuel Gangler

Laboratoire de Physique Corpusculaire,
Universit Clermont-Auvergne, Universit Blaise-Pascal, CNRS/IN2P3
Clermont-Ferrand, France.
email: `gangler at clermont.in2p3.fr`

Abstract. LSST is a next generation telescope that will produce an unprecedented data flow. The project goal is to deliver data products such as images and catalogs thus enabling scientific analysis for a wide community of users. As a large scale survey, LSST data will be complementary with other facilities in a wide range of scientific domains, including data from ESA or ESO. European countries have invested in LSST since 2007, in the construction of the camera as well as in the computing effort. This latter will be instrumental in designing the next step: how to distribute LSST data to Europe. Astroinformatics challenges for LSST indeed includes not only the analysis of LSST big data, but also the practical efficiency of the data access.

Keywords. Surveys, methods: data analysis, astronomical data bases: miscellaneous

1. LSST as a data factory

1.1. *Brief overview of LSST*

LSST[†] is a next generation telescope that will acquire data starting 2021. Located at Cerro Pachón (Chile), it will map the entire visible sky every 3 night under an airmass limit of the order of 1.5. This is enabled by a 3.2 Gpixel CCD camera with a 9.6 deg^2 field of view mounted through the secondary mirror of a 8.4m diameter telescope. Each pixel is capable of 0.2" sampling which optimizes the sensitivity. The camera is equipped with a 6 band filter system, *ugrizy*. With a 40s average visit time comprising 2 exposures of 15s, it will observe the 18000 deg^2 of the wide fast deep survey area with at least 888 visits summed over all filters for any given pointing over the 10 years of the survey. The median number of visits in the *ugrizy* bands is (62,88,199,201,180,180) respectively (LSST 2016). With a median number of visit per night of 816, the amount of data produced will be of the order of 15 to 30 TB per night, leading to a total of 90 PB of raw data after 10 year of operations. The reduced data stored in catalogs will comprise over 15 PB of data. All those numbers have to be considered as provisional, experience has shown that the data problem is usually widely underestimated before a survey actually starts.

The effort to build the LSST is a partnership between public and private organizations. Financial support for LSST comes from the National Science Foundation, the Department of Energy, and funding raised by the LSST Corporation, a non-profit corporation formed in 2003. The NSF-funded LSST Project Office for construction was established in 2012 as an operating center under management of the Association of Universities for Research in Astronomy (AURA). The DOE-funded effort to build the LSST camera is managed by the SLAC National Accelerator Laboratory (SLAC). The LSST project organizational chart relies on 4 main workpackages : Telescope and Site, Camera, Data Managment, and Education and Outreach.

[†] https://www.lsst.org

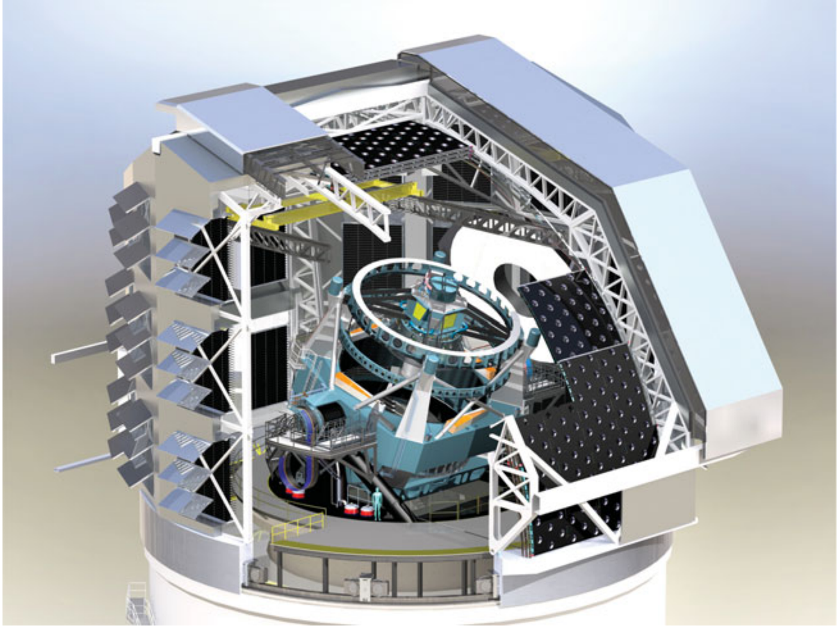

Figure 1. LSST dome and telescope.

While LSST Project Office is responsible for the construction of LSST, the scientific analysis of LSST data is not part of the project and will be conducted by independent collaborations whose members shall have the right to access LSST data. The role of the LSST Corporation is to enable the full science exploitation of the unique LSST data set through coordination and preparation of the member institutions, the broader physics and astronomy community and international contributors. The LSST Corporation includes two categories of partners: institutional members and international contributors. Institutional members are institutions who brought substantive contributions to the development, construction, and the scientific exploitation of the LSST. It comprises 37 institutions, 5 of them being outside US and 3 in Europe: the Institut de Physique Nuclaire et de Physique des Particules (IN2P3 — France), the University of Portsmouth, Institute of Cosmology & Gravitation (UK), and The Institute of Physics of the Academy of the Czech Republic (Czech Republic). The international contributors are institutions which agreed to share in the annual operating costs of the LSST in exchange for data rights for a specified list of their principal investigators during LSST operations and commissioning. There are 2 countries counting as International Contributors (US and Chile) and 25 institutions, 11 of them coming from european countries: Instituto de Astrofisica de Canarias (IAC — Canary Islands), Ruer Bokovi Institute (RBI — Croatia), Institut National de Physique Nucleaire et de Physique des Particules (IN2P3 – France), Ludwig-Maximilians-Universitt (LMU — Germany), Max Planck Institute for Astrophysics (MPA — Germany), Max Planck Institute for Astronomy (MPIA — Germany), Max Plank Institute for Extraterrestrial Physics (MPE — Germany), Max Plank Institute for Extraterrestrial Physics (MPE — Hungary), Konkoly Observatory (Hungary), Istituto Nazionale di Astrofisica (INAF — Italy), Nano Center (Serbia), University of Nova Gorica (UNG — Slovenia), Barcelona-Madrid Consortium (BCN-MAD — Spain), Stockholm University Department of Physics (SU-Physics — Sweden), Eidgenoessische Technische Hochschule Zuerich (ETH Zurich), Institute for Astronomy (Switzerland), Science and Technology Facilities Council (STFC) - UK LSST Consortium (UK). About

https://doi.org/10.1017/S1743921317000825 Published online by Cambridge University Press

900 scientists around the world are expected to have privileged access to LSST data, 300 of them expected to come from European countries. Outside these institutions, data access will be possible only trough the open data access policy of LSST, which means that data will be available provided that no additional cost incurs to the project and after a proprietary period.

1.2. *Scientific program enabled by LSST*

The LSST science goals covers 4 major themes within the astronomical community: the nature of Dark Matter and understanding of Dark Energy, cataloging the Solar System, exploring the changing sky, Milky Way structure and formation (Izevic 2008, Abell 2009). In 2008, eleven separate quasi-independent science collaborations were formed to focus on these topics, and after some reshaping there are currently eight active LSST Science Collaborations.

- Galaxies: this collaboration will address topics as demographics of galaxy populations, dwarf galaxies, galaxy mergers and merger rates, galaxy morphology, tidal tails and streams, wide-area multi-band searches for high redshift galaxies. It comprises currently 46 members, including 3 members from UK and 1 from Croatia.
- Stars, Milky Way, and Local Volume : this collaboration addresses variable stars, star clusters, Magellanic clouds, near field cosmology, galactic buldge, the Solar neighborhood and the galactic structure and ISM. It comprises currently 118 members, including 4 members from UK and 2 members from Germany.
- Solar System : this collaboration adresses the science enabled by the LSST catalog of over 5 million asterods from the different families : Main Belt asteroids, Jupiter Trojans, NEOs, and TNOs. It does not release the list of its members.
- Dark Energy : the Dark Energy Science Collaboration (DESC) is by far the largest scientific collaboration of LSST. It addresses all cosmological topics related to LSST capabilities including weak lensing, large scale structure, supernovae, clusters, strong lensing as well as photometric redshifts, theory and joint probes analyses. It is also organized around technical and computation and simulation workpackages such as cosmological simulations, survey simulations, software and computing infrastructure, sensor anomalies and photometric corrections (DESC 2012). It counts 565 members, 25% of which being non US-based. This includes 70 members from UK, 65 from France, 2 from Czech Republic and 1 from Spain.
- Active Galactic Nuclei : this collaboration counts 36 members, including 1 member from UK and 1 member from Serbia.
- Transients/variable stars : this collaboration addresses many different topics, such as cosmological use of transients, transient classification and characterization, the distance scale, fast transients, galactic transients, gravitational waves counterparts, interacting binaries, magnetically active stars, microlensing, multiwavelength characterization, non-degenerate eruptive variables, pulsating variables, supernovae, tidal disruption events and transiting planets. It counts more than 104 members, including 2 members from Germany, 1 member from France, Italy and UK each.
- Strong Lensing : the activities of this group have a large fraction in common with the interests of the DESC.
- Informatics and Statistics : this collaboration comprises scientists devoted to developing tools for use with large astronomical surveys. It includes astronomers, statisticians, computer scientists, and machine learning researchers. Currently it counts 60 members, 1 being from UK.

All the numbers aforementioned are in rapid evolution, and represent only a snapshot at the time of the IAU symposium.

https://doi.org/10.1017/S1743921317000825 Published online by Cambridge University Press

1.3. *LSST synergies with other projects*

With a catalog of 37 billion objects and 2 million transient alerts raised each night, LSST data will act as a discovery machine for the astrophysics and physics communities. While many discoveries will be made using LSST data alone, maximizing the science from LSST will require ground-based optical-infrared (OIR) supporting capabilities, e.g., observing time on telescopes, instrumentation, computing resources, and other infrastructure. In a recent report (Najita 2016), NOAO and LSST have identified supporting capabilities that will be needed from an US perspective. Expectedly, this includes a set of follow-up instrument such as a highly multiplexed, wide-field optical multi-object spectroscopic capability on an 8m-class telescope, preferably in the Southern Hemisphere, a broad wavelength coverage, moderate-resolution (R = 2000 or larger) OIR spectrograph on Gemini South, single-object, multi-color imaging on $<$ 5m facilities and single-object R = 100–5000 spectroscopy on 3–5m facilities. But interestingly, it also includes facilities related to the data management and distribution issues, such as the development and early deployment of an alert broker, scalable to LSST and the support for an OIR system infrastructure developments that enable efficient follow-up programs. It also recommends to study and prioritize the needs for computing, software, and data resources, including the development and deployment of data analysis and exploration tools that work at the scale of LSST; training for scientists at all career stages in LSST-related analysis techniques and computing technologies; and cross-disciplinary workshops that facilitate the cross-pollination of ideas and tools between astronomy and other fields. This highlights the importance of astroinformatics in enabling the use and reuse of LSST data.

From the European perspective, 2 agencies, ESA an ESO will provide complementary data with respect to LSST. Being located in the southern hemisphere, ESO has a complete set of instruments adapted to LSST follow-up, including multi-objects spectrographs such as 4MOST, and the European Extra LargeTelescope whose first light is expected in 2024. As a multi-instrument facility, ESO will adapt to the future by implementing a new strategy, as discussed in a recent workshop†. The changes for the community will be that ESO will move towards coherent programs, where the time allocation will be tailored on the long run and if needed on different instruments to focus on solving a scientific problem in one go. ESO also highlights the importance of archives and the combination of archival data and new observations from different facilities, a concerned shared by the distribution of LSST data.

Regarding ESA, LSST has scientific complementarity with existing and future programs. For instance, it will complement the capabilities of the GAIA satellite regarding photometry, astrometry and proper motions of objects of r magnitude between 20 and 24 (27.5 for photometry). For the Dark Energy science, the complementarity will be with the Euclid and WFIRST satellites (Jain 2015). LSST and Euclid will share around 5000 deg^2 of sky observations, with different band coverage, Euclid complementing LSST observations in YJH while providing highly accurate shape measurements in its extended Vis filter. Having three different surveys with their own characteristics will allow to mitigate the systematic uncertainties. For instance, the joint analysis of large scale structures would increase the sensitivity to the sum of the neutrino masses by a factor > 3 with respect to the best limit obtained by an individual program. The best approach to reach these improved results will be achieved by linking the data from the different surveys and providing a common access point for interrogating the data in a user-friendly way with the appropriate tools. In this regard, institutions belonging to more than one of these ambitious surveys will have an edge when it will come to the joint analysis.

† https://www.eso.org/sci/meetings/2015/eso-2020.html

https://doi.org/10.1017/S1743921317000825 Published online by Cambridge University Press

2. Getting ready for LSST in Europe

2.1. *A sizable construction effort*

The challenge raised by the LSST project range in 3 categories : building the telescope, the camera, and preparing for the data deluge. If European subcontractors participate as vendors for the telescope construction, the most important hardware investment is made by France regarding the camera construction, an effort which dates back in 2007. The IN2P3 contribution to the camera focuses on two key elements, the CCDs and filter exchange system, with accompanying effort given to camera control, commissioning and calibration. This amounts to 64 engineers and technical staff in addition to the active scientists. The early contribution to the camera R&D allowed France to reach its status as LSST core member. Apart from France, the Czech Republic has a contribution to the qualification of the sensors.

The other sizable effort in Europe is regarding the Data Management infrastructure, where the Computing Center of IN2P3 will play a major role. As a Satellite Data Release Processing Site — the controlling center being NCSA — CC IN2P3 will produce 50% of the annual data release and host a full copy of LSST raw data and catalogs, thus being also a Data Archive Center. The concept of satellite data release processing was tested and proved successful in 2013 with a joint processing of SDSS data coming from stripe 82, employing LSST data reduction software. Since this test, one of the goals for the qualification of LSST reduction software is to process verification datasets coming from other projects such as DES, HSC and CFHTLS with LSST stack. Thanks to the deep knowledge of the Megacam instrument by the IN2P3 scientists involved in SNLS, the French LSST computing group is now fully responsible of the CFHTLS reprocessing at CC-IN2P3.

In addition to the infrastructure effort, European scientists are involved in the LSST data reduction software, with an expertise on image subtraction pipeline and on image processing for faint surface brightness object analysis. As an example, the ongoing work on the sensor characterization also led to the unravelling of the brighter-fatter effect (Guyonnet 2015), which now has to be properly corrected at the pixel level for cosmological analyses.

Finally, the European scientists are also part of the data distribution effort, with the Serbian development of Alertsim, and the French and UK effort on the Qserv database under development at SLAC. The purpose of AlertSim is to assure that the external brokers that will serve as the primary delivery mechanism of LSST transient data to the public, will be capable of receiving and processing the LSST event stream (Aleksic 2016). Qserv is a distributed database system designed to deliver LSST data products and will be detailed in the last section.

2.2. *Towards an european data access center*

While the Archive Centers will be the main repository feeding the distribution of LSST data to the community, a network of data access centers (or DACs) are envisioned for broad user access, according to a tiered access model where the tiers define the capacity and response available. There are two project-funded data access centers co-located with the base facility in Chile and the archive center in NCSA. These centers replicate all of the LSST data to facilitate disaster recovery and provide open and public interfaces to the LSST data products. In the meantime, LSST is encouraging the other partners to host additional data access centers to make LSST data available to the broadest possible community of end users and help amortize observatory operations costs. A European

https://doi.org/10.1017/S1743921317000825 Published online by Cambridge University Press

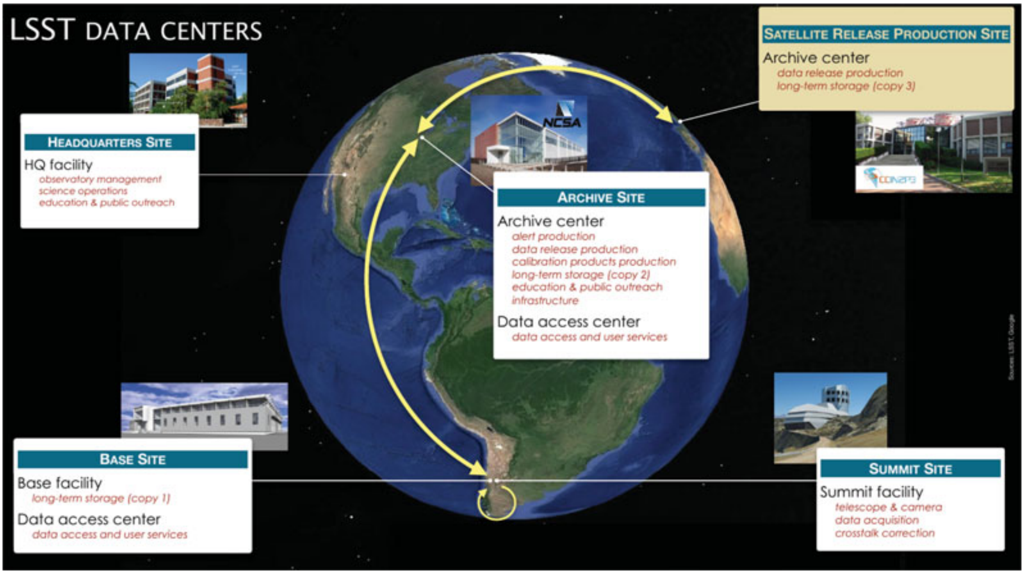

Figure 2. LSST data centers around the world.

Data Access Center will thus be welcomed by LSST provided the adequate funding can be gathered by the European partners.

As the IN2P3 Computing Center will be a Data Archive Center, hosting the 3rd full copy of LSST data, it would be the base infrastructure to serve the data in the scope of a data access center: hosting the data is indeed not enough to provide an efficient access and serve the data to the community, there is the need to offer an infrastructure adequate for science analysis, which includes not only the delivery of the catalogs, but also should enable the production of the Level 3 scientific data products. This analysis infrastructure will have an investment cost as well as annual running costs. France is getting ready for a French data access center, which would combine the requirements of a standard LSST DAC and an analysis facility. There are currently two options regarding this French DAC. The minimum option is to guarantee the scientific return of the hardware investment that was made by France. In such a case, the DAC would emphasis the analysis facility to be focused on the relevant topics of interests for French researchers, that is, infrastructure needed for DESC. The other ambitious option is to build an european data access center to serve stakeholders at the European level. This would need some level of integration of European efforts, both to secure the funding of this project within European calls for infrastructure, and to design the technical drivers of this DAC, which can be a multisite facility. The natural schedule to complete this ongoing discussion with European partners is the European calls expected to be published by the end of 2017. UK and France are now actively participating in enabling the DAC architecture, within the implementation of a prototype data access center

3. New challenges for astroinformatics

As good algorithms for data analysis tend to be specialist, a successful data system must allow the development and incorporation of a wide variety of algorithms. In order to facilitate their implementation, LSST data release complies with a scheme adapted for astroinformatics (Borne 2009). One of the key concept is that the data products will

https://doi.org/10.1017/S1743921317000825 Published online by Cambridge University Press

come with everything needed to perform the end user analysis, with minimal exposure of the raw pixel images. There will be three different data production pipelines. The alert production will be an online process running on images right after their acquisition. The data release production and the calibration pipeline production will be run annually, with full reprocessing each year during the 10 years of the project. The data products issued by the nightly and annual releases enter three different categories: images available as files, catalogs stored in database, and alerts available through a broker.

The annual release catalog consist of more than 100 tables, the most important of which being the object catalog summarizing for each physical source all information acquired during the project lifetime, and the source catalog giving access to each individual measurement data of a single object on a single exposure. The object catalog will contain around $40{\times}10^9$ records of individual objects, with of the order of 500 attributes per record, for a total volume around 100 to 200 TB. The source catalog will contain 5×10^{12} records of individual measurements for the objects, with more than 100 attributes per record, for a total volume of 3 to 5 TB, and the forced source catalog will hold up to 32×10^{12} records, for a total of 1 to 2 TB.

Building the catalogs is thus only the starting point of LSST data analysis challenge: any statistical analysis of machine learning task will first need an efficient access to LSST data.

3.1. *Distributing LSST data*

The LSST baseline for the database architecture is to support massive user queries in a massively parallel relational database composed of a single-node non-parallel DBMS, a distributed communications layer, and a master controller, all running on a shared-nothing cluster of commodity servers with locally attached disk drives. All large catalogs such as Object, Source and Forced Source are horizontally partitioned into materialized chunks according to their spatial coordinates, and the remaining catalogs are replicated on each server. There is a further partitioning into sub-chunks materialized on the flight when needed: this provides a dramatic reduction in execution time when dealing with cone-like searches where the predication is performed on pairs of neighboring objects. The chunks will be distributed automatically across all nodes, without exposure to users. The system will also use a few critical indexes to speed up spatial searches, time series analysis, and simple but interactive queries. It will also provide shared scans. The architecture is driven by the variety and complexity of anticipated queries, ranging from single object lookups to complex O(n2) full-sky correlations over billions of elements. As no off-the-shelf reasonably priced solution meets LSST requirements nowadays, LSST developed a prototype of the baseline architecture, called Qserv, based on open-source software (Becla 2013).

A prototype test platform on Qserv has been deployed at IN2P3 Computing Center thanks to a partnership with DELL. This test bench is made of 50 physical nodes for a total of 400 cores and 800 GB RAM memory. The available disk space is 500 TB. Performance tests were run on a data volume of 35 TB, made of reprocessed data coming from SDSS stripe 82 which were replicated in order to achieve the desired volume (Jammes 2015). This data volume is representative of about 10% of the first year of LSST dataset. A full range of queries were run on a large scale test in summer 2015, from short queries such as single object selection up to full table joins. Results were more promising than initially expected in terms of performance, while exhibiting some limitations regarding the concurrent scans. These test were instrumental for the Qserv development in order to assess the current status of the software and to design future improvements. They have shown that a test platform is a necessity in order to build complex systems, and the

https://doi.org/10.1017/S1743921317000825 Published online by Cambridge University Press

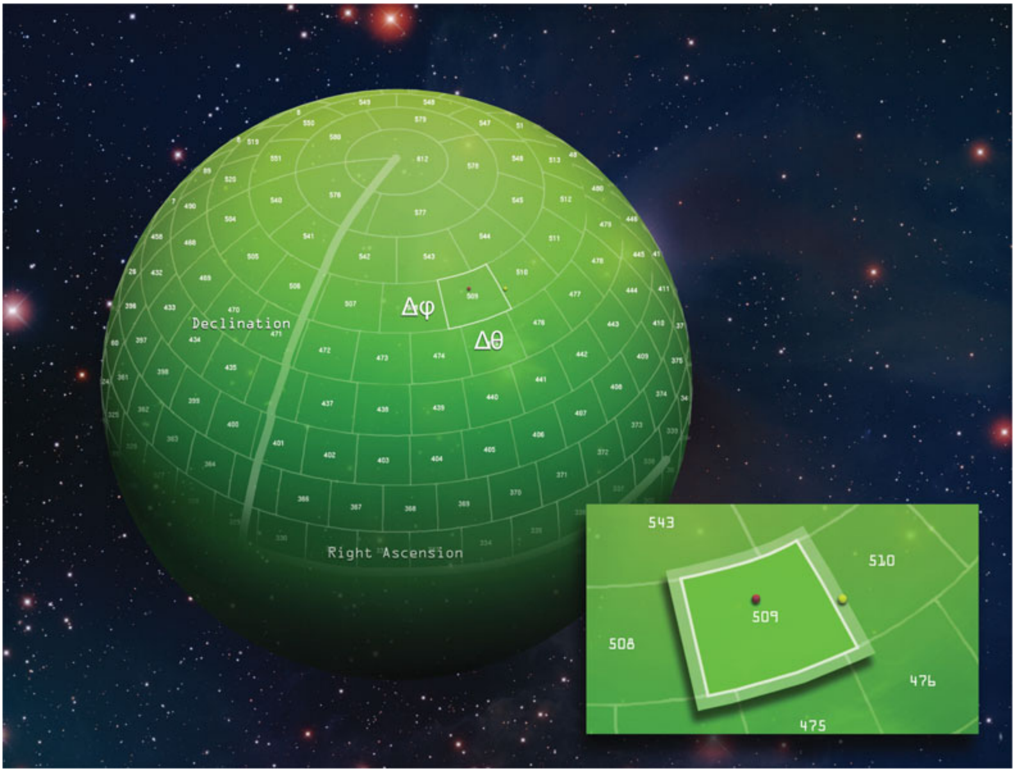

Figure 3. Qserv partitioning scheme on the sphere.

IN2P3 Computing Center facility Qserv R&D platform is the only available platform for integration tests. The experience learned is now moving towards establishing a running instance of the service as a prototype data access center at NCSA.

In addition to Qserv, the challenges raised by LSST have attracted interdisciplinary collaboration with computer scientists such as the PetaSky project in France†. The current design of Qserv makes it an ad-hoc optimized solution to serve LSST data, while the optimization of queries in distributed database systems is an active research field. In other words, Qserv can be seen as a specific solution to a more generic problem, and keeping up to speed with the research in database systems can open new insights into the LSST data distribution. One interesting research path is the interaction of machine learning techniques on top of a database: astronomers indeed tend to still use the"ftp-grep" model where they decouple fetching the data form the analysis. With data volumes such as LSST's, this will turn out to be impractical for some tasks, for instance when analyzing the time domain data, and methods which directly handle computations within the database system can come up an edge in performance.

3.2. *Analysing LSST data*

While it is too early to make an extensive review of machine learning tasks which will successfully be applied to LSST data, there are some data analysis problems that LSST has to solve where the solution goes through machine learning. There is some European expertise some of those problems. The best way to show improvement is by the analysis

† http://com.isima.fr/Petasky

of mock data sets, and of precursor surveys such as DES. Most of the problems that the LSST analyst will face has already to be tackled by these surveys. For instance, the photometric redshift issue, or the photometric supernova selection can be studied already, and while the designated DESC working groups come up with an implementation adapted to LSST data, the scientific study of the problem can be performed today.

Beyond the traditional use of machine learning techniques to solve tasks for the data currently at hand, LSST come up with other perspectives. One of them regards the optimization of the cadence: as LSST is the first survey of its kind with large-scale production of time-domain data on a wide field project, the current optimization scheme relies on known phenomena. This means a more or less regular cadence, which will fit most of the needs. However, there is room to improve this cadence, which has to be performed beforehand. The other perspective is regarding the development of advanced machine learning tools : research in computer science develop new techniques to analyze data which are proven to scale with the volume. Finding the right way to use these techniques in order to gather relevant new information on the data is challenging indeed. Thus, training new scientists to new data analysis techniques and bridging astronomy and informatics communities, such as in the Europena COST BigSkyEarth project† is one of the key aspect of the European perspective for future surveys.

4. Conclusion and perspectives

LSST will provide unprecedented astronomical data, bring astronomy to the Big Data scale. The project will address major science topics covering the solar neighborhood, the Milky Way and extragalactic studies, and cosmology. It will also sytematically address the time domain with the transient sky. Overall, 300 european scientists from 9 countries are expected to analyze LSST data. With its data set, many science topics will get advantage to combine LSST data with other facilities, including data coming from ESO or ESA programs. The oldest european institution to join LSST was France in 2007. There has been since sizable investments in the camera construction and in the development of a full computing infrastructure in Europe, including a processing and a data archive center. The next goal is to enable an european access to LSST data through an european data access center. With its impressive data flow, LSST has also attracted the attention of the computer science community, both from database experts and machine learning sides. Despite a lot of expertise in Europe on astroinformatics, there is now only one european official member of the LSST astroinformatics and astrostatistics science collaboration. With the steadily growing interest of the European community for LSST, we can target for a much improved participation to this working group in the coming years.

References

A. Abell *et al.* 2009, *arXiv* 0912.0201

J. Aleksic 2016 in: M. Brescia, S. G. Djorgovski, E. Feigelson, G. Longo & S. Cavuoti (eds.) *Atroinformatics 2016* Proc. IAU Symposium No. 325 (Sorrento)

J. Becla *et al.* 2013 *LSST document LDM-135*

K. Borne *et al.* 2010 *arXiv* 0909.3892

LSST Dark Energy Science Collaboration 2012, *arXiv* 1211.0310

F. Jammes *et al.* 2015 in: *JRES 2015*, Proc. JRES 2015 (Montpellier)

† http://bigskyearth.eu

https://doi.org/10.1017/S1743921317000825 Published online by Cambridge University Press

A. Guyonnet t al. 2015 *A&A* 575, 41.
Z. Izevic *et al.* 2008, *arXiv* 0805.2366
B. Jain *et al.* 2015 *arXiv* 1501.07897
LSST Project Science Team and LSST Simulations Team 2016, *LSST Document-19427*
J. Najita *et al.* 2016, *arXiv* 1610.01661

https://doi.org/10.1017/S1743921317000825 Published online by Cambridge University Press