Everything I’d like to do with LSST data, but don’t know (yet) how

Željko Ivezić
Department of Astronomy
University of Washington

Local Group Astrostatistics Conference, June 4, 2015
The era of surveys...

"Ask Not What Data You Need To Do Your Science, Ask What Science You Can Do With Your Data."
A catalog of 20 billion stars and 20 billion galaxies with exquisite photometry, astrometry and image quality!


**LSST in one sentence:**
An optical/near-IR survey of half the sky in ugrizy bands to r~27.5 based on ~1000 visits over a 10-year period:

A catalog of 20 billion stars and 20 billion galaxies with exquisite photometry, astrometry and image quality!
and while dignitaries are celebrating…
some are happily doing real work!
Outline

• Data analysis challenges ahead of us
  – large data set
  – complex analysis
  – aiming for small systematics

• Rapid tour of LSST
  - multi-color time-resolved faint sky map
  - 20 billion stars (all the way to the edge of the Milky Way)

• Examples of astrostatistics and astroinformatics in LSST era
  – “astrostatics and astroinformatics are the new calculus!”
Data analysis challenges in the era of Big Data

1) Large data volume (petabytes)

2) Large numbers of objects (billions)

3) Highly multi-dimensional spaces (thousands)

4) Unknown statistical distributions

5) Time-series data

6) Heteroscedastic errors, truncated, censored and missing data

7) Unreliable quantities (e.g. unknown systematics and random errors)

The bottleneck will not be (is not any more?) data availability but instead the ability to extract useful and reliable information from data.
Some tools and methods...

- Correlation coefficients (many dimensions, missing data)
- The bootstrap and the jackknife methods
- Maximum Likelihood Method
- The goodness of fit and model selection
- **Bayesian statistics**
- Markov Chain Monte Carlo methods
- Regression (“fitting”, LSQ, outliers, regularization)
- Density estimation (“multi-dimensional histograms”)
- Clustering (kernel, parametric)
- Classification (supervised and unsupervised, active learning)
- Dimensionality Reduction (PCA, ICA, LLE and friends)
- Time-series analysis (periodogram, stochastic processes)
AstroML: Machine Learning and Data Mining for Astronomy

AstroML is a Python module for machine learning and data mining built on numpy, scipy, scikit-learn, and matplotlib, and distributed under the 3-clause BSD license. It contains a growing library of statistical and machine learning routines for analyzing astronomical data in python, loaders for several open astronomical datasets, and a large suite of examples of analyzing and visualizing astronomical datasets.

The goal of astroML is to provide a community repository for fast Python implementations of common tools and routines used for statistical data analysis in astronomy and astrophysics, to provide a uniform and easy-to-use interface to freely available astronomical datasets. We hope this package will be useful to researchers and students of astronomy. The astroML project was started in 2012 to accompany the book *Statistics, Data Mining, and Machine Learning in Astronomy* by Zeljko Ivezic, Andrew Connolly, Jacob VanderPlas, and Alex Gray, to be published in late 2013. The table of contents is available here: [here](https://www.astroml.org).
LSST Science Themes

- Dark matter, dark energy, cosmology (spatial distribution of galaxies, gravitational lensing, supernovae, quasars)
- Time domain (cosmic explosions, variable stars)
- The Solar System structure (asteroids)
- The Milky Way structure (stars)

LSST Science Book: arXiv:0912.0201
Summarizes LSST hardware, software, and observing plans, science enabled by LSST, and educational and outreach opportunities

245 authors, 15 chapters, 600 pages
Basic idea behind LSST: **a uniform sky survey**

- 90% of time will be spent on a uniform survey: every 3-4 nights, the whole observable sky will be scanned twice per night

- after 10 years, half of the sky will be imaged about 1000 times (in 6 bandpasses, ugrizy): a digital color movie of the sky

- ~100 PB of data: about a billion 16 Mpix images, enabling measurements for 40 billion objects!

**LSST in one sentence:**
An optical/near-IR survey of half the sky in ugrizy bands to r~27.5 (36 nJy) based on 825 visits over a 10-year period: **deep wide fast.**

Left: a 10-year simulation of LSST survey: the number of visits in the r band (Aitoff projection of eq. coordinates)
SDSS vs. LSST comparison: \( \text{LSST} = \frac{d(\text{SDSS})}{dt}, \text{LSST} = \text{SuperSDSS} \)

3x3 arcmin, gri

3 arcmin is 1/10 of the full Moon's diameter.

Deep Lens Survey \((r \sim 26)\)

(almost) like LSST depth (but tiny area)

SDSS, seeing 1.5 arcsec

Subaru, seeing 0.8 arcsec

20x20 arcsec; lensed SDSS quasar (SDSS J1332+0347, Morokuma et al. 2007)
Not just point source depth: faint surface brightness limit

3x3 arcmin, gri

\( r \sim 26 \)

(almost) like LSST depth (but tiny area)

Gawiser et al
The field-of-view comparison: Gemini vs. LSST

Gemini South Telescope

Primary Mirror Diameter

8 m

Field of View

0.2 degrees

3.5 degrees

(Full moon is 0.5 degrees)

LSST
The largest astronomical camera: 2800 kg, 3.2 Gpix
The ultimate deliverable of LSST is not just the telescope, nor the camera, but the fully reduced science-ready data as well.

LSST Data Products: see http://ls.st/dpdd
LSST simulations: a virtual system
End-to-end modeling creates a virtual prototype of the LSST system: learning and preparing prior to first light!

**A simulated sky**

- Galaxies (de Lucia et al 2006)
- Stars (Juric et al 2008)
- Asteroids (Grav et al 2007)

**Observing sequence simulation**

- Pointing, Filter, Airmass, Time and Atmosphere from Op Sim

**Producing a simulated image**

- 10^{10} photons per CCD
- Separate amplifiers
- 2.5 hours per CCD
The halo is especially interesting because gravitational potential becomes dominated by the dark matter halo.

- Components trace the DM dominated potential
- They are a product of Milky Way formation and evolution
- Thin/thick disk
- Galactic bulge
- Stellar halo
The Milky Way structure: 20 billion stars, time domain massive statistical studies!

Compared to SDSS: LSST can “see” about 40 times more stars, 10 times further away and over twice as large sky area.

Distance and [Fe/H]:

Main sequence stars

SDSS RR Lyrae

100 kpc
The Milky Way is complex and big!
And interesting and informative!

Only 2.5 deg wide:
<1% halo volume!
“We need more data”
(at least 2 mag deeper than SDSS, and time-resolved)
The large blue circle: the $\sim 400$ kpc limit of future LSST studies based on RR Lyrae

The large red circle: the $\sim 100$ kpc limit of future LSST studies based on main-sequence stars (and the current limit for RR Lyrae studies)

6D information from LSST: 3D spatial, 2 velocities, [Fe/H]

The small insert: $\sim 10$ kpc limit of SDSS and future Gaia studies for kinematic & [$Fe/H$] mapping with MS stars
The large blue circle: the $\sim 400$ kpc limit of future LSST studies based on RR Lyrae

The large red circle: the $\sim 100$ kpc limit of future LSST studies (and the current SDSS limit)

200 million stars from LSST!

montage from B. Willman
Gaia vs. LSST comparison

- **Gaia**: excellent astrometry (and photometry), but only to $r < 20$
- **LSST**: photometry to $r < 27.5$ and time resolved measurements to $r < 24.5$
- Complementarity of the two surveys: photometric, proper motion and trigonometric parallax errors are similar around $r=20$

The Milky Way disk “belongs” to Gaia, and the halo to LSST (plus very faint and/or very red sources, such as white dwarfs and LT(Y) dwarfs).

Dwarfs in LSST

**White dwarfs:** LF is age probe

~400,000 halo white dwarfs from LSST (10 million total):

**L / T dwarfs:** L dwarfs are dime a dozen: 200,000 in LSST with proper motion and trigonometric parallax measurements

Simulations predict 2400 T dwarfs with >5σ proper motion and parallax measurements

Compared to UKIDSS, 5 times larger sample of T dwarfs, with parallaxes and 10-20 times more accurate proper motions

(~100 Y dwarfs [model based])
Statistical analysis of a massive LSST dataset

- A large (100 PB) database and sophisticated analysis tools: for each of 40 billion objects there will be about 1000 measurements (each with a few dozen measured parameters)

Data mining and knowledge discovery

- 10,000-D space with 40 billion points
- Characterization of known objects
- Classification of new populations
- Discoveries of unusual objects

Clustering, classification, outliers
Due to size and complexity of LSST dataset, and susceptibility of many of its major science goals to systematics in both measured quantities and astrophysical predictions, **substantial preparatory work is required to fully exploit LSST dataset.**

**LSST data analysis and astro-statistics will be closely related, and will open numerous opportunities for people with "Big Data" skills**

**Main work areas:**
- advanced astronomical digital image processing
- statistical modeling and analysis
- data mining and machine learning
- high performance computing
Selected topics (where work is needed)

1) Interpretation of spectral energy distributions (SEDs)

2) Spatial correlations

3) Moving objects

4) Variable objects

5) Systematic measurement uncertainties

6) Astrophysical simulations and astrophysical systematics

7) LSST System Enhancements

8) New algorithms in LSST
1) **Interpretation of spectral energy distributions (SEDs)**

- efficient and robust interpretation of time-resolved multi-band photometry for “billions and billions” of objects

Because of integration over broad bandpasses, forward modeling using a trial SED is superior to “correcting data" (fluxes, positions, sizes)

a) **Photo-z algorithms**: observed SED depends on the redshift of an intrinsic SED (expansion of the universe, source evolution, intergalactic extinction)

b) **photometric parallax for stars** (will greatly benefit from Gaia parallaxes!)

c) **photometric metallicity for stars** (trained using spectroscopic metallicities)
1) Interpretation of spectral energy distributions (SEDs)

- efficient and robust interpretation of time-resolved multi-band photometry for “billions and billions” of objects

Open/interesting issues:

- machine learning vs. SED templates for galaxies vs. cross-correlation of samples
- the impact of heteroscedastic noise, priors, truncated and censored data
- how much would per visit processing help (due to varying effective bandpasses)?
- posterior pdfs vs. likelihoods, optimal compression, etc
- covariances (also for pretty much everywhere else below)
2) Spatial correlations

examples:

a) large-scale distribution of galaxies (e.g. distinguish GR vs. modified gravity)

b) matched filters for dwarf galaxies and streams/tidal tails

c) LMC/SMC, Virgo overdensity, great circle streams morphologies

Open/interesting issues
- how to scale up to billions of points across the sky?
- can this be done directly in db?
3) Moving objects
- cadence optimization: do we really need two visits per night? Or perhaps as many as four!
- how robust and efficient would be full Bayesian approach?
- how hard is it to do shift-and-coadd for KBOs and more distant objects on scale of LSST?

4) Variable objects
a) regular vs. irregular
b) short vs. long timescales
c) robust detection
d) classification

Open/interesting issues:
- machine learning vs. light curve templates
- heteroscedastic noise, priors, truncated/censored data
- can this be done directly in db?
5) Systematic measurement uncertainties

e.g. across the sky, vs. seeing, sky brightness, etc.

a) astrometry
b) photometry
c) galaxy shapes (cosmic shear: as small as $10^{-6}$)

Open/interesting issues:
- unknown SEDs!
- the impact of atmosphere (variable seeing and transmissivity, DCR, stochastic)
- multiplicative and additive errors in galaxy shear
- systematics in photo-z
- billions of objects measured a thousand times: does $\sqrt{N}$ still work in this regime?
6) Astrophysical simulations and astrophysical systematics

e.g.
- growth of cosmic structure
- formation and evolution of galaxies
- formation and evolution of Solar System
- the ISM in the Milky Way

Open/interesting issues:
- is GR correct?
- what are feedback mechanisms in galaxy formation?
- nonlinear galaxy bias
- intrinsic alignments of galaxy shapes with the density field
- baryonic effects on dark matter halo profiles
- gravity vs. hydro simulations issues (gas dynamics, star formation, feedback, etc)
7) LSST System Enhancements

observing strategy (e.g. angular and temporal sampling functions, dithering patterns)
other filters?
improved algorithms (for image differencing, calibration, etc)
Level 3 (a.k.a. specialized processing as well as everything we didn’t think of)

Open/interesting issues:
- cadence optimization
- cadence evolution
- shift-n-add for arbitrary space-time trajectories
- complex galaxy models (e.g. tidal tails)
- transient classification
8) New algorithms in LSST

some mature, some not even started…

- Multifit (forward modeling on per visit basis)
- psf depends on time, position, instrument state, and color (more precisely on in-band SED shape)
- image differencing
- SED is unknown, especially for transients…
- incorporating other surveys in data processing, e.g. how to best benefit from Gaia, Euclid, WFIRST, etc.
- how much can you do in 60 seconds?

Data Release Processing:
assume 1000 cores and 8 months run time, so 2e10 core-seconds; with 20 billion objects: 1 sec/object

** What can you compute in about 1 second? **
8) New algorithms in LSST

- how to automate tradeoffs between false positive rate (contamination) and the false negative rate (completeness)

- robust detection of extremely rare events and real-time intelligent event filtering for follow-up efforts

- unsupervised, semi-supervised and supervised clustering/classification on massive datasets in real time (in db?)

- joint morphological and color-magnitude based probabilistic object classification (e.g. “galaxy”, with its type and its photo-z vs. “star” and its photometric parallax and metallicity)
Selected topics: which one will make *you* famous?

1) Interpretation of spectral energy distributions (SEDs)
2) Spatial correlations
3) Moving objects
4) Variable objects
5) Systematic measurement uncertainties
6) Astrophysical simulations and astrophysical systematics
7) LSST System Enhancements
8) New algorithms in LSST
LSST data analysis and astro-statistics will be closely related, and will open numerous opportunities for people with "Big Data" skills.