Lecciones en Astroinformática Avanzada (Semester 1 2023)

# Automatic Classification of Variable Stars (II)

**Nina Hernitschek**
Centro de Astronomía CITEVA
Universidad de Antofagasta

May 28, 2024

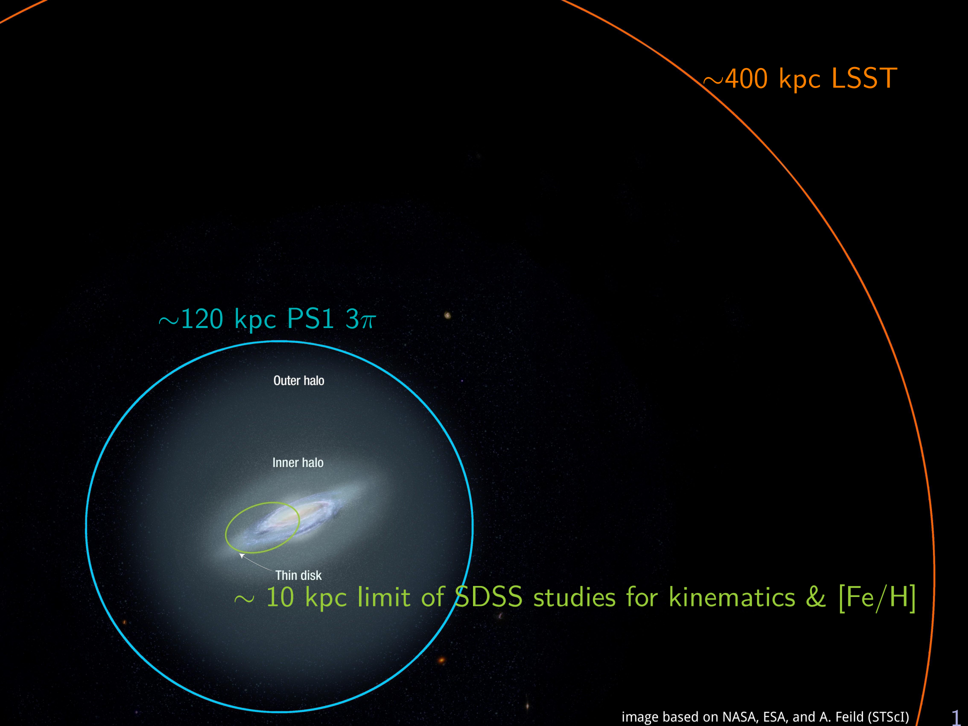
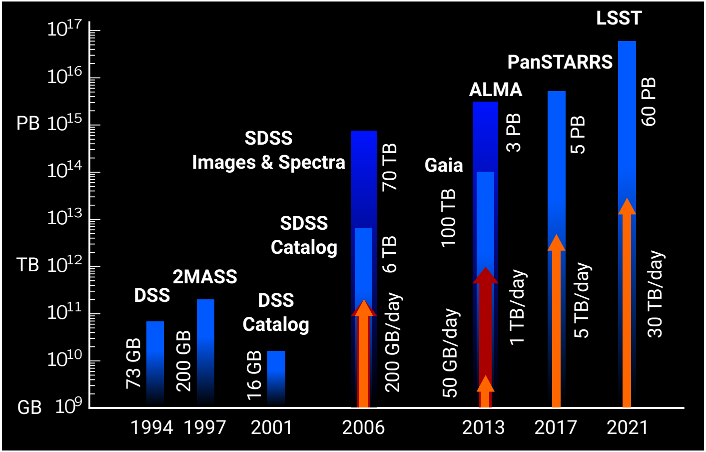



image based on NASA, ESA, and A. Feild (STScI)

# Recap: Large Astronomical Surveys

increasing data volume in astronomical surveys



# Big Data: Challenge and Chance



upcoming large all-sky surveys like (but not only) LSST

**Challenge:**

- enormous data volume, $<$ 60 PB total, 15 - 30 TB/ night
- follow-up opportunities should be identified immediately

**Chance:**

large data volume enables for

population studies

finding rare 'one-in-a-million', 'one-in-a-billion' events, often called *anomalies*

# Big Data: Challenge and Chance



upcoming large all-sky surveys like (but not only) LSST

**Challenge:**

- enormous data volume, $<$ 60 PB total, 15 - 30 TB/ night
- follow-up opportunities should be identified immediately

**Chance:**

large data volume enables for

- population studies
- finding rare 'one-in-a-million', 'one-in-a-billion' events, often called *anomalies*

⇒ **tools for discoveries**: software, computing resources, research projects to make use of astronomical survey data

# Challenges in Data Handling

**BIG Challenges:**

- often infeasible to download (all) data for analysis
- sometimes infeasible to store all data $\Rightarrow$ crucial to efficiently analyze data and to produce a response $\sim$ real time
- additional observations needed to fulfill science goals (example: precisely timed spectroscopic observations to get stellar velocities)

# Challenges in Data Handling



**BIG Challenges:**

- often infeasible to download (all) data for analysis
- sometimes infeasible to store all data $\Rightarrow$ crucial to efficiently analyze data and to produce a response $\sim$ real time
- additional observations needed to fulfill science goals (example: precisely timed spectroscopic observations to get stellar velocities)

Big Data Analytics & Machine Learning are **transforming** how discoveries are made

# The Data

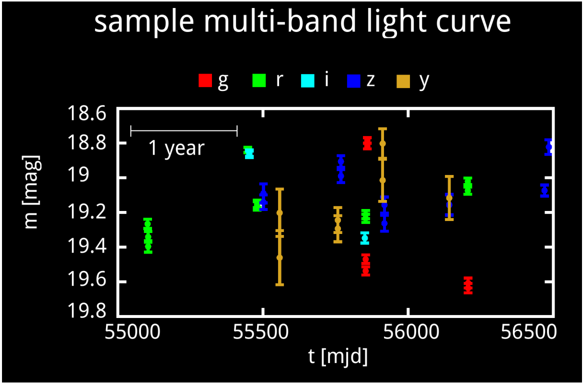
**The Pan-STARRS1 $3\pi$ Survey in one sentence:**

An optical/near-IR survey covering 3/4 of the sky in non-simultaneous *grizy* to $r\sim21.8$ with $\sim70$ visits over 5.5 years.

# The Data

**The Pan-STARRS1 $3\pi$ Survey in one sentence:**

An optical/near-IR survey covering 3/4 of the sky in non-simultaneous *grizy* to $r{\sim}21.8$ with ~70 visits over 5.5 years.



# The Data

Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 $3\pi$

Big Data

Fitting a Model

Machine Learning

The LSST Survey

goal: a catalog of variable sources in PS1 $3\pi$

# The Data

goal: a catalog of variable sources in PS1 $3\pi$

to **model a survey**, tools are needed for

- describing data quality $\rightarrow$ outlier might fake or hide true variability
- describing light curve characteristics $\rightarrow$ "features" with scientific relevance
- classifying sources $\rightarrow$ catalogs others can use
- finding substructure $\rightarrow$ clumps, overdensities, ... the science we want to do

# The Data

goal: a catalog of variable sources in PS1 $3\pi$

to **model a survey**, tools are needed for

- describing data quality $\rightarrow$ outlier might fake or hide true variability
- describing light curve characteristics $\rightarrow$ "features" with scientific relevance
- classifying sources $\rightarrow$ catalogs others can use
- finding substructure $\rightarrow$ clumps, overdensities, ... the science we want to do

**challenge:**
processing $\sim 10^9$ rather sparse, noisy light curves

# The Methods

**Classification of variable sources** relies fundamentally on algorithms quantifying different aspects of variability.

**feature extraction:**

light curve $\xrightarrow{\text{signal processing}}$ numbers

$\Rightarrow$ features should be as discriminative and informative as possible

# The Methods

**Classification of variable sources** relies fundamentally on algorithms quantifying different aspects of variability.

**feature extraction:**

light curve $\xrightarrow{\text{signal processing}}$ numbers

$\Rightarrow$ features should be as discriminative and informative as possible

challenges:

- non-simultaneous multi-band data
- noise & uncertainties
- foreground effects
- not all variables are periodic: QSOs, supernovae...
- time-sampling can act as window function (hiding variability)
- many period-finders are computationally very expensive: pre-selection

# The Methods



**multi-band structure-function variability model**:

describe light curves as stochastic processes: how much should you expect a multi-band source to vary within $\Delta t$?

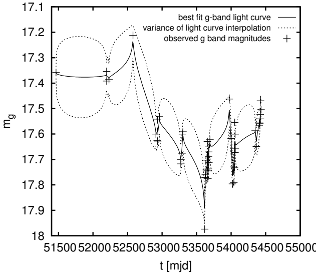
$$\mathrm{V}(|\Delta t|) \equiv \mathrm{E}[(m(t) - m(t+\Delta t))^2]$$

assume functional form

$$V(\Delta t) \overset{\text{model}}{\equiv} \omega_i(\lambda_i)\omega_j(\lambda_j)\left(1 - \exp\left[-\frac{|\Delta t|}{\tau}\right]\right)$$

with $\tilde{m}_\lambda(t) = m_\lambda(t) - \bar{m}_\lambda$, $\omega_k(\lambda_k) = \omega_r \left(\frac{\lambda_k}{\lambda_r}\right)^\alpha$

(Hernitschek et al. 2016)



Recap: Large Astronomical Surveys
Overview
Classifying Pan-STARRS1 $3\pi$
Big Data
Fitting a Model
Machine Learning
The LSST Survey

# The Methods

**multi-band structure-function variability model**:

describe light curves as stochastic processes: how much should you expect a multi-band source to vary within $\Delta t$?

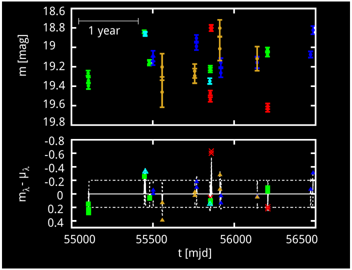
$\Rightarrow$ fit $\Rightarrow$ characteristic variability timescale & amplitude

# The Methods

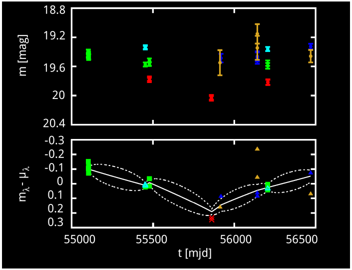
**multi-band structure-function variability model**:

describe light curves as stochastic processes: how much should you expect a multi-band source to vary within $\Delta t$?

$\Rightarrow$ fit $\Rightarrow$ characteristic variability timescale & amplitude




RR Lyrae, $\omega_r$=0.3, $\tau$=1.5 days




QSO, $\omega_r$=0.13 , $\tau$=560 days

# The Methods

**period fitting**:

RR Lyrae period crucial for distance determination:
Period-Luminosity-Metallicity (PLZ) relation

$L$=f($P$,$Z$), $D$=f(L,$m$) $\Rightarrow$ $D$ = f($m$,$P$,$Z$)

$\Rightarrow$ goal: 3D map of Milky Way's RR Lyrae
from measured magnitude $m$ and derived period $P$ (and $Z$)

# The Methods

Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 $3\pi$

Big Data

Fitting a Model

Machine Learning

The LSST Survey

**period fitting**:

RR Lyrae period crucial for distance determination:
Period-Luminosity-Metallicity (PLZ) relation

$L$=f($P$,$Z$), $D$=f(L,$m$) $\Rightarrow$ $D$ = f($m$,$P$,$Z$)

$\Rightarrow$ goal: 3D map of Milky Way's RR Lyrae
from measured magnitude $m$ and derived period $P$ (and $Z$)

However:

- sparse light curves
- computationally expensive

$\Rightarrow$ apply methods suitable for sparse and unevenly sampled multi-band data to **pre-selected sources**

Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 3π

Big Data

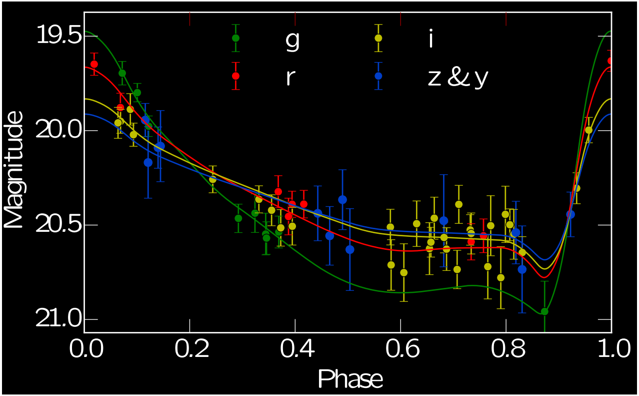
Fitting a Model

Machine Learning

The LSST Survey

## Period and Phase from Template Fitting

example: RR Lyrae period/phase fitting, using light curve templates from SDSS Stripe 82 (Sesar et al. 2010)



# Results

$10^9$ light curves from PS1 $3\pi$

outlier cleaning using a machine-learning method

feature extraction: structure functions, mean magnitudes

first machine-learning classification

**RR Lyrae**

$1.5 \times 10^5$ RR Lyrae candidates
(80% purity, 80% completeness)

feature extraction: template fitting

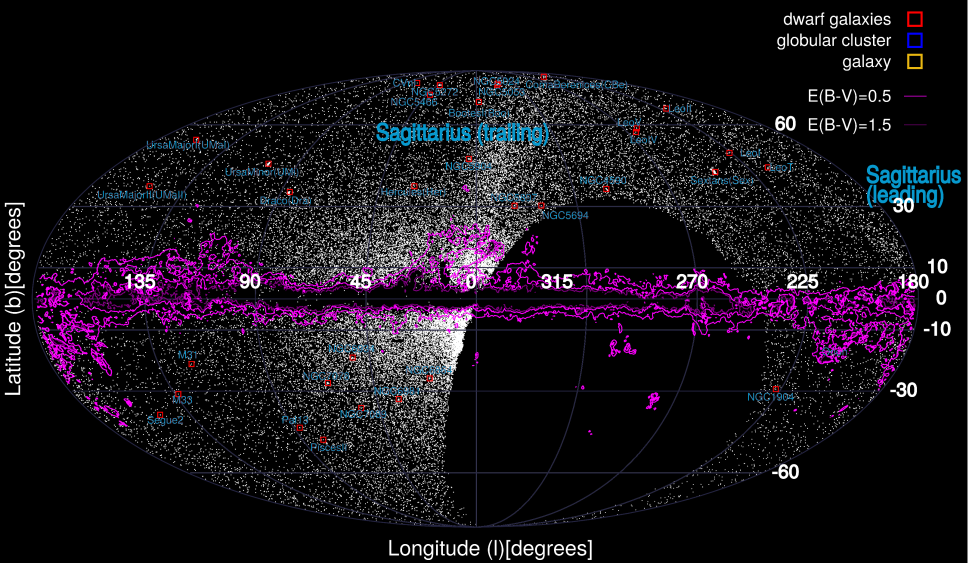
second machine-learning classification

$4.4 \times 10^4$ RR Lyrae stars (90% purity, 80% completeness), **distances** up to $\sim$140 kpc, $\sigma = 3\%$

**QSO**

$3.8 \times 10^6$ QSO candidates
(85% purity, 85% completeness)

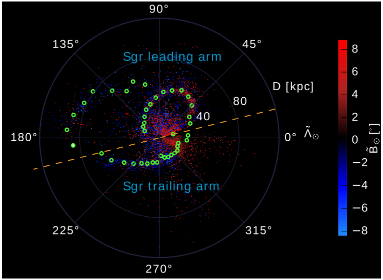
# The Results

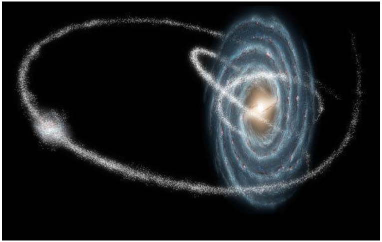

dwarf galaxies
globular cluster
galaxy
E(B-V)=0.5
E(B-V)=1.5
Sagittarius (trailing)
Sagittarius (leading)
60
30
10
0
-10
-30
-60
135
90
45
0
315
270
225
180
NGC5694
NGC1904
M31
M33
Segue2
Pal13
LeoT
Latitude (b)[degrees]
Longitude (l)[degrees]

# The Results

**Sagittarius stream: an example for structure finding**

globular cluster or dwarf galaxy $\rightarrow$ torn apart and stretched out along its orbit by tidal forces $\rightarrow$ stellar stream



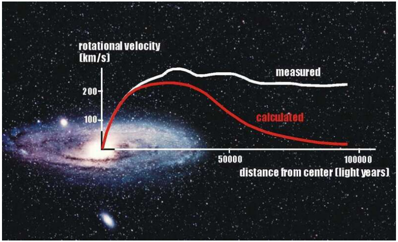
artistic image,
www.spitzer.caltech.edu

Automatic Classification of Variable Stars (II)
Recap: Large Astronomical Surveys
Overview
Classifying Pan-STARRS1 $3\pi$
Big Data
Fitting a Model
Machine Learning
The LSST Survey

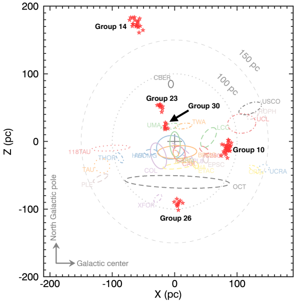
# The Follow-Up Survey

Milky Way dynamics: get 3D velocities

Dark Matter




comoving groups and clusters in the Milky Way

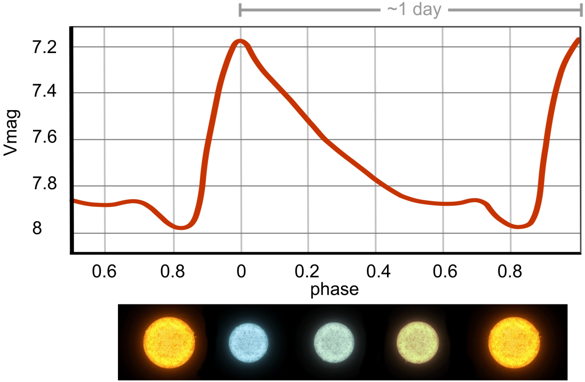



Faherty+2018

# The Follow-Up Survey



crucial for RR Lyrae (and pulsators in general): **timing**

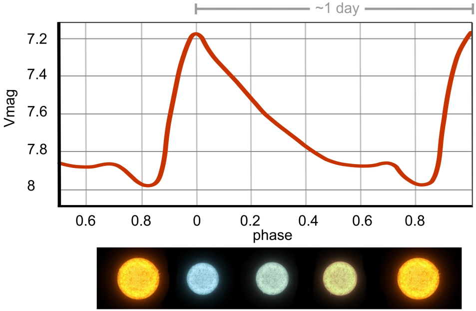



lack of hydrostatic equilibrium drives pulsation and thus periodic change in brightness

# The Follow-Up Survey



crucial for RR Lyrae (and pulsators in general): **timing**




$$v_{\mathrm{obs}} = v_{\mathrm{systemic}} + v_{\mathrm{photospheric}}$$

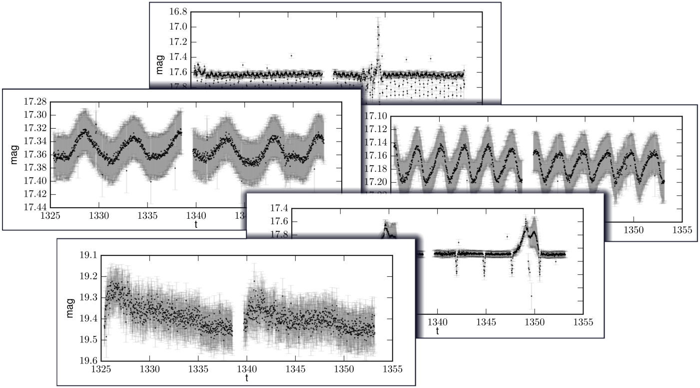
from pulsation models:

observe at $\phi = 0.37$ where $v_{\mathrm{photospheric}} \sim 0$

# Another Survey



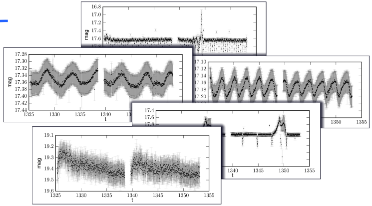
… more possible from better sampled TESS* light curves

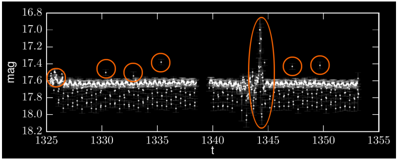



* An all-sky satellite survey taking single-band light curves with a 30 minute cadence and 27 day baseline.

# Another Survey

... more possible from better sampled TESS* light curves



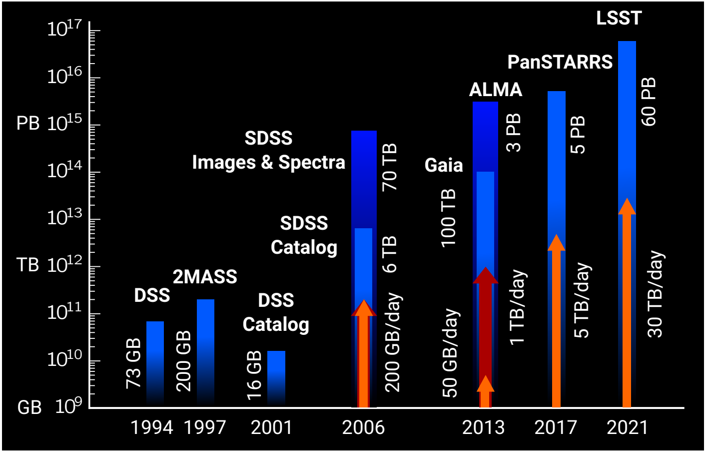



reliable outlier cleaning and detrending is extremely crucial for automated data analysis

* An all-sky satellite survey taking single-band light curves with a 30 minute cadence and 27 day baseline.

# Challenges in Data Handling

increasing data volume in astronomical surveys



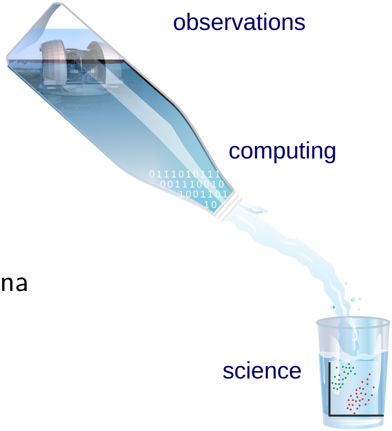
Recap: Large Astronomical Surveys
Overview
Classifying Pan-STARRS1 $3\pi$
Big Data
Fitting a Model
Machine Learning
The LSST Survey

# Challenges in Data Handling

astronomy is largely determined by computational capacity

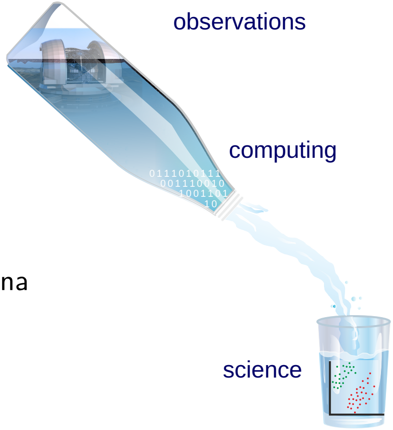
$\Rightarrow$ telescopes & instruments as front-ends for data processing systems & follow-up telescopes

$\Rightarrow$ challenge and chance: understanding complex phenomena requires complex data



# Challenges in Data Handling

astronomy is largely determined by computational capacity




⇒ telescopes & instruments as front-ends for data processing systems & follow-up telescopes

⇒ challenge and chance: understanding complex phenomena requires complex data

**Big Data** is transforming how and which discoveries are made.

Recap: Large Astronomical Surveys
Overview
Classifying Pan-STARRS1 $3\pi$
Big Data
Fitting a Model
Machine Learning
The LSST Survey

# Big Data



Recap: Large Astronomical Surveys
Overview
Classifying Pan-STARRS1 $3\pi$
Big Data
Fitting a Model
Machine Learning
The LSST Survey

Laney et al. 2001: data growth challenge is **three-dimensional**

Big Data is data with at least one big dimension:

- volume
- velocity: bandwidth, response speed
- variety: number and size of individual assets

**shifting use cases:**
As data becomes big data, finding the *right* data has become more important.

# Big Data

Laney et al. 2001: data growth challenge is **three-dimensional**

Big Data is data with at least one big dimension:

- volume
- velocity: bandwidth, response speed
- variety: number and size of individual assets

**shifting use cases:**
As data becomes big data, finding the *right* data has become more important.

⇒ powerful astrostatistical & machine-learning tools are needed to derive scientific insights

# Big Data

**shifting use cases:**

As data become more plentiful, finding the *right* data has become more important.

$\Rightarrow$ powerful astrostatistical & machine-learning tools are needed to derive scientific insights

Individual measurements giving way to **statistics, clustering, patterns** in the data.

Data processing needs to be **highly automatized**.
Analysis growing more exploratory rather than pre-defined/scripted.

# Big Data

## Examples:

Finding and classifying variable stars in PS1 $3\pi$ required processing of $10^9$ sparse light curves $\Rightarrow$ 44,000 RRab stars.

Dark matter + dark energy through weak-lensing requires combination of WFIRST / LSST / CMB, organized in multiple archives.

Transient science (gravitational wave follow-up, GRBs, unknowns from LSST) requires rapid access to data sets of what is already known, anywhere on sky.

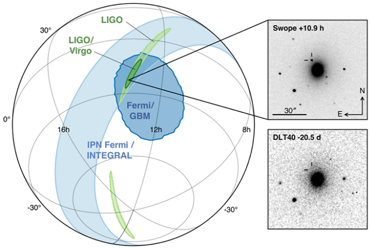



image credit: LIGO

# Statistical Data Analysis

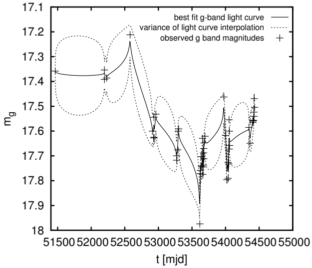
As we transition to a regime of large amounts of data, **data-driven models** can improve the precision of noisy data possibly **without the biases of physical models**.

# Statistical Data Analysis

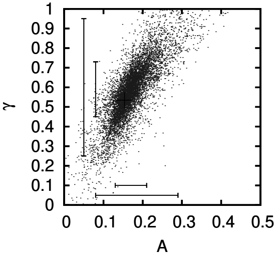


As we transition to a regime of large amounts of data, **data-driven models** can improve the precision of noisy data possibly **without the biases of physical models**.

Statistical methods can reliably **quantify information** embedded in scientific data and score the relevance of theoretical models.




MCMC sampling of log $P_{posterior}$



# Statistical Data Analysis

As we transition to a regime of large amounts of data, **data-driven models** can improve the precision of noisy data possibly **without the biases of physical models**.

Statistical methods can reliably **quantify information** embedded in scientific data and score the relevance of theoretical models.

**Requirements:**

- find the right method(s): modern statistics is vast in its scope and methodology
- scientific inferences should not depend on arbitrary choices in methodology and variable scale
- correct interpretation of the meaning of a statistical result w.r.t. the scientific goal: (astro-)statistics and machine learning are only tools!

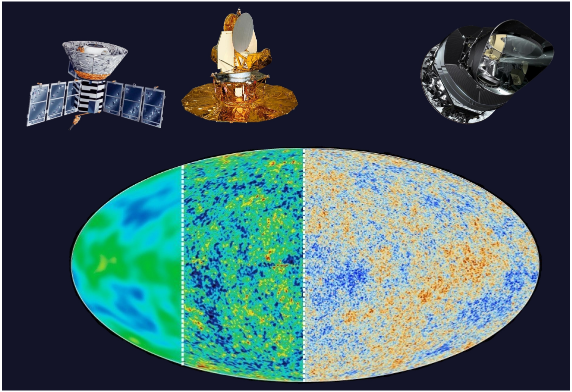
# Model Selection - typical Questions in Astronomy & Cosmology

- Are the fluctuations in the Cosmic Microwave Background best fit by Big Bang models with dark energy or with quintessence?

WMAP (2003)
resolution: 0.3°

COBE (1992)
resolution: 7°

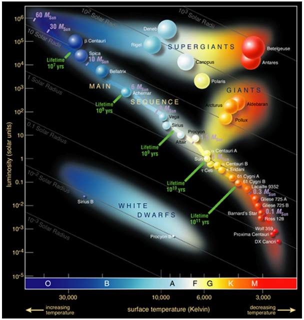
PLANCK (2013)
resolution: 0.07°

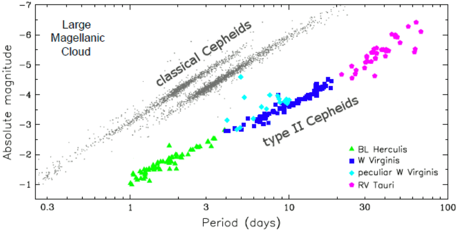


# Model Selection - typical Questions in Astronomy & Cosmology



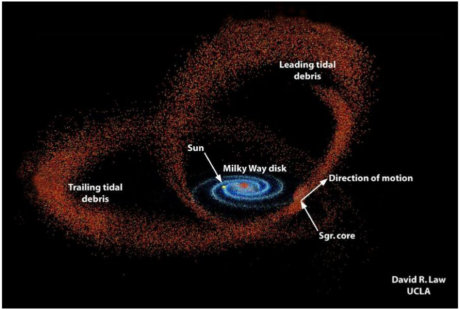
- Are the fluctuations in the Cosmic Microwave Background best fit by Big Bang models with dark energy or with quintessence?
- Are there interesting (cor-)relations among properties?



# Model Selection - typical Questions in Astronomy & Cosmology

- Are the fluctuations in the Cosmic Microwave Background best fit by Big Bang models with dark energy or with quintessence?
- Are there interesting (cor-)relations among properties?
- Interpreting the radial velocity variations of a large sample of stars. This can lead to the discovery of orbiting systems.



# Model Selection

Criteria for a good model:

- model simplicity
- conform fitted model to the data (goodness of fit)
- easily generalizable
- not *under*-fit (that would exclude key variables)
- not *over*-fit (that would produce an unnecessarily complex model with extra variables)

# Model Selection

Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 $3\pi$

Big Data

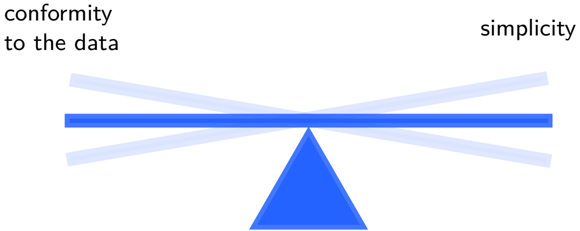
Fitting a Model

Machine Learning

The LSST Survey

Criteria for a good model:

- model simplicity
- conform fitted model to the data (goodness of fit)
- easily generalizable
- not *under*-fit (that would exclude key variables)
- not *over*-fit (that would produce an unnecessarily complex model with extra variables)



# Likelihood-based Model Selection - The Likelihood



**notation:**

observed data $D$; a set of $N$ data values $D = \{X_i\}$ is a sample

$M_j = M_1, ..., M_k$ are models for $D$, $\theta_j$ is a $k$-dim. parameter vector; the parameter space is the space of possible values of $\theta_j$; more general term: hypothesis space

**example:**

likelihood function $\mathcal{L}_j = f(D \,|\, \theta_j; M_j)$, $\log \mathcal{L}_j = \log f(D \,|\, \theta_j; M_j)$

$f(D \,|\, \theta_j; M_j)$ is the probability density function evaluated on $D$

assume: $D = (X_1, ..., X_n)$, $X_i$ independent & dist. $\sim N(\mu, \sigma^2)$

$$\mathcal{L} = f(D \,|\, \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \right\}$$

# Likelihood-based Model Selection - Bayesian Statistical Inference



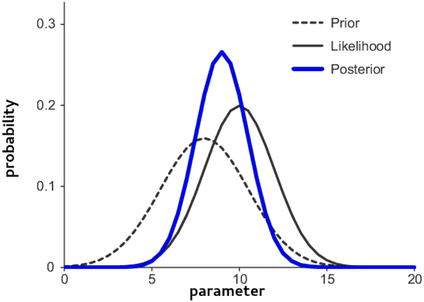
**the posterior probability distribution - Bayesian statistical inference:**

turn the likelihood of the data into a belief of the true parameters using **prior information**:

posterior probability $\sim$ likelihood $\mathcal{L}$ $\times$ prior probability

$$p(\theta_j \mid D_i, \sigma_{D,i}) = \frac{p(D_i, \sigma_{D,i} | \theta_j)}{p(D_i, \sigma_{D,i})} p(\theta_j)$$

with $M_j = M_1, ..., M_k$ are models for $D$ with parameters $\theta_j = \theta_1, ..., \theta_k$



# Likelihood-based Model Selection - Bayesian Statistical Inference



goal:

**Best-Fit parameters**

$$\theta^* = \arg\max_x p(\theta_j \mid D_i, \sigma_{D,i})$$

**Confidence regions** (Bayesian: Creditable regions)

We may be uncertain about $j$ (model uncertainty) or $\theta_j$ (parameter uncertainty). Credible region $\Delta$ of probability $C$:
$C = p(\theta \in \Delta \mid D, M) = \int_\Delta \mathrm{d}\theta\, p(\theta \mid D, M)$

# Likelihood-based Model Selection - Bayesian Statistical Inference

Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 $3\pi$

Big Data

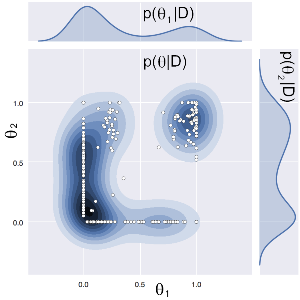
Fitting a Model

Machine Learning

The LSST Survey

optimizing **Posterior Probability Distribution** $p(\theta_j \mid D_i, \sigma_{D,i})$

- on a grid in $\theta_j$
- using rejection sampling methods such as Markov chain Monte Carlo (MCMC)




MCMC constructs a biased random walk to explore the posterior distribution

# Machine Learning

… is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed
(Arthur Samuel, 1959)

# Machine Learning

... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959)

$\Rightarrow$ allows to **uncover hidden correlation patterns** through iterative learning by sample data

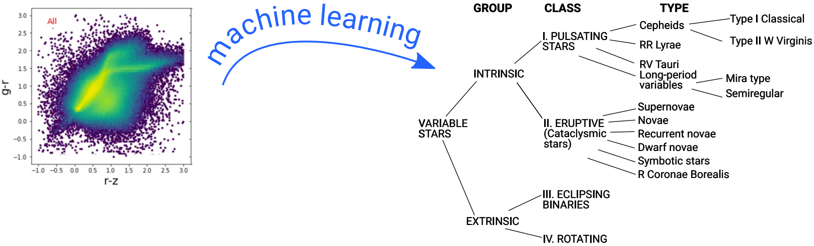
# Machine Learning

… is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959)

$\Rightarrow$ allows to **uncover hidden correlation patterns** through iterative learning by sample data

parameter space of measurements

parameter space of astrophysical objects



# Machine Learning

… is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959)

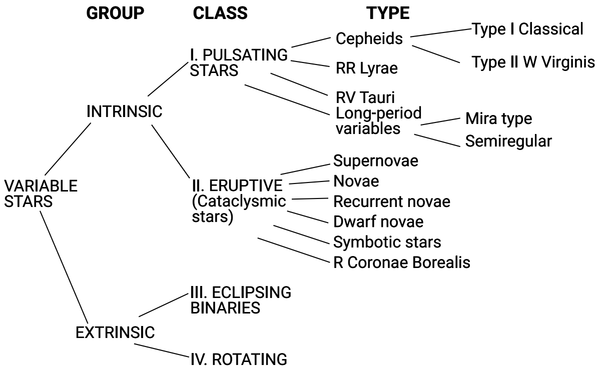
$\Rightarrow$ allows to **uncover hidden correlation patterns** through iterative learning by sample data

$\Rightarrow$ allow "to model a survey":

- describing data quality $\rightarrow$ outlier
- describing light curve characteristics $\rightarrow$ "features"
- classifying sources $\rightarrow$ catalogs
- finding substructure $\rightarrow$ clumps, overdensities, …

# Classification Methods

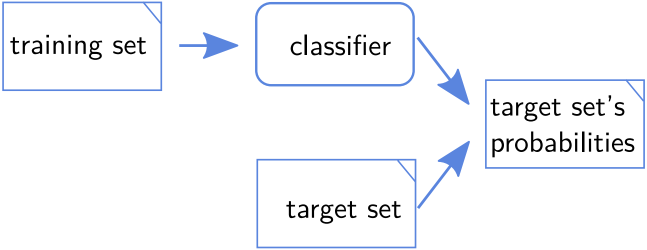





**Classification:**

Use *a priori* group labels to assign new observations to a particular group or class $\Rightarrow$ *supervised learning* or "learning with labels".

# Concepts of Supervised Classification

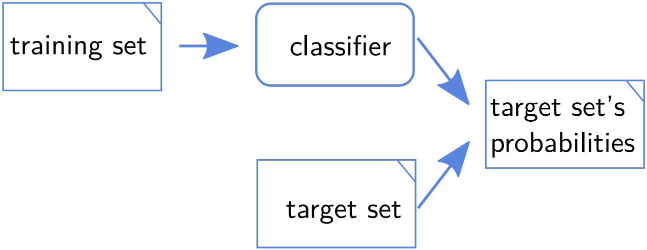



training set:

- set of sources inside/outside category we are looking for
- same data quality as found in target set

# Concepts of Supervised Classification






training set:

- set of sources inside/outside category we are looking for
- same data quality as found in target set

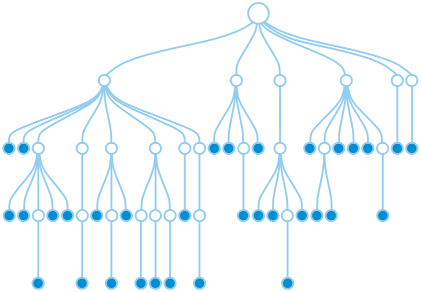
**What's happening internally?**

# Concepts of Supervised Classification

The learning process ("training"):

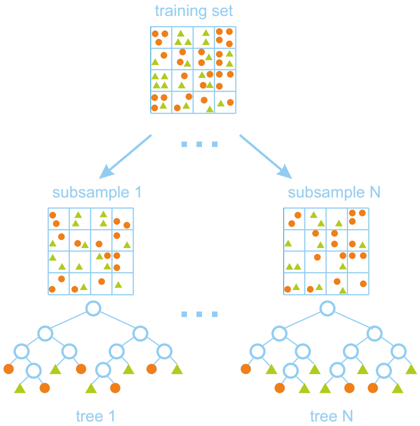
To build a decision tree, the set is divided into smaller and smaller subsets by **splitting** w.r.t. a single **feature** at a time.

Split criteria: select feature and split point to produce the smallest impurity in the two resultant nodes based on the **training set**.



# Supervised Classification - Ensemble Methods

**Random Forest Classifier** as ensemble method: many trees are grown from subsets of the training set



Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 $3\pi$

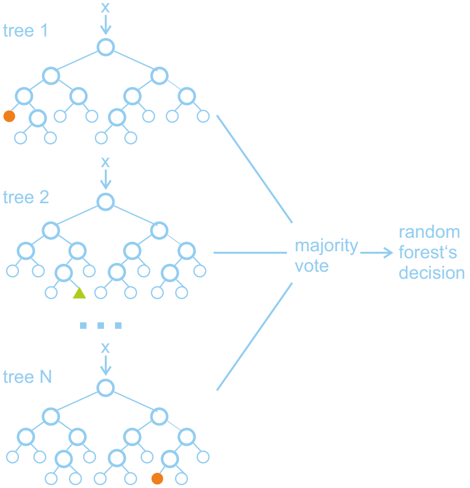
Big Data

Fitting a Model

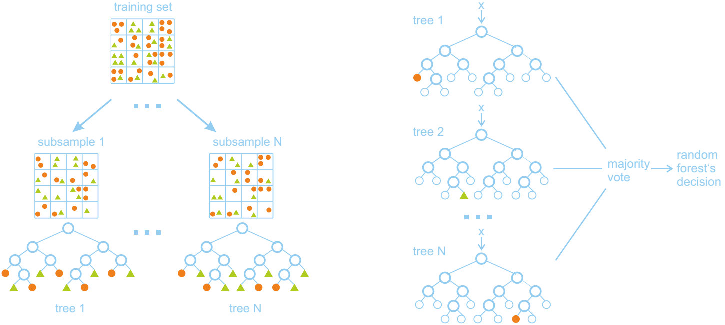
Machine Learning

The LSST Survey

**Random Forest Classifier** as ensemble method: ... and are "voting" for classification



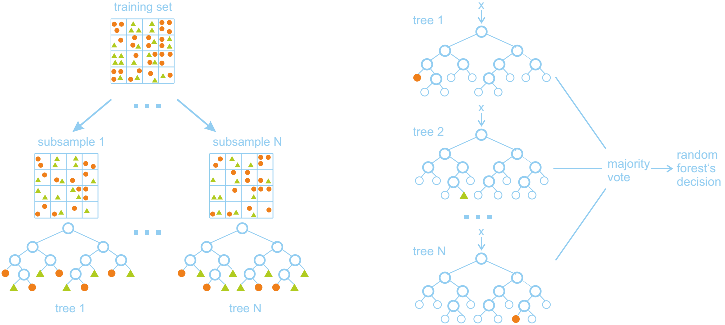
# Supervised Classification - Ensemble Methods




**divide-and-conquer approach** improves classification performance

- less sensitive to training set variances
- training and classification can be parallelized

# Supervised Classification - Ensemble Methods




**divide-and-conquer approach** improves classification performance

- less sensitive to training set variances
- training and classification can be parallelized

$\Rightarrow$ **ideal for big data**

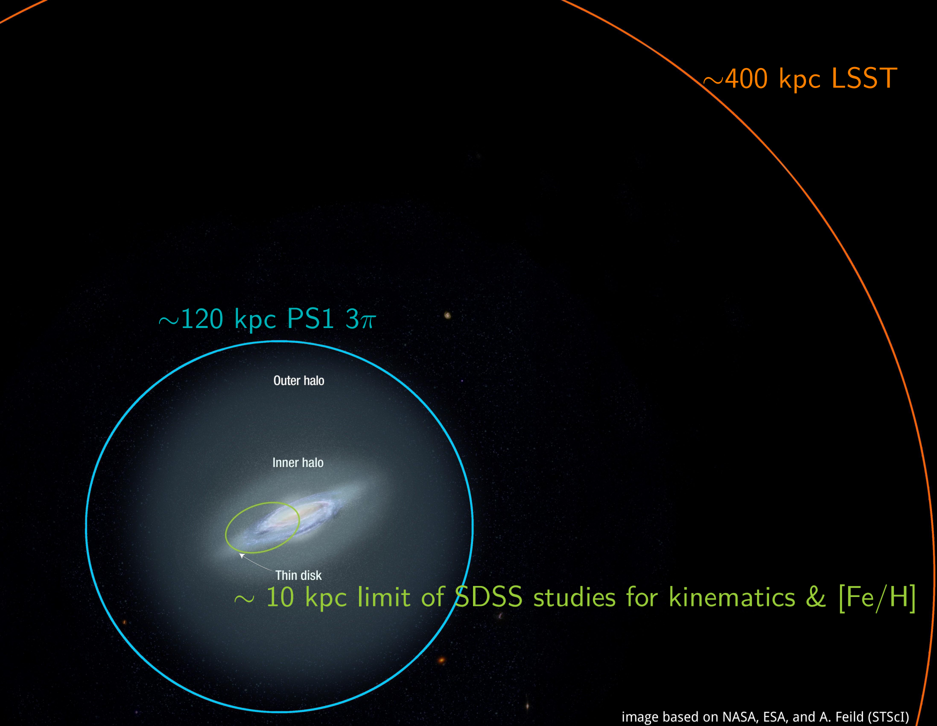


image based on NASA, ESA, and A. Feild (STScI)

# LSST/ Rubin Observatory

Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 $3\pi$

Big Data

Fitting a Model

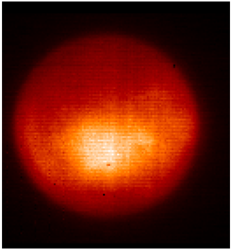
Machine Learning

The LSST Survey

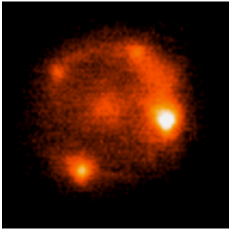
different telescopes study different things

some telescopes let scientists look at specific parts of the sky in high resolution to study the fine details of objects, or for a long time to collect more light to see fainter objects: **targeted observations**

example: Keck, Magellan



moon Europa from the Keck Observatory, credit: Mike Brown



near-IR image of gravitationally lensed Type Ia SN, Keck Observatory

Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 $3\pi$

Big Data

Fitting a Model

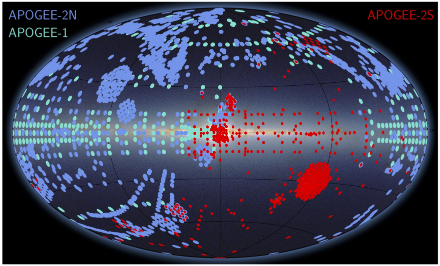
Machine Learning

The LSST Survey

different telescopes study different things

other telescopes let scientists study lots of objects in wide areas of the sky, but at lower resolution: **survey telescopes**

example: SDSS, Pan-STARRS, Rubin Observatory




APOGEE-2: A stellar spectroscopic survey of the Milky Way, composed of a northern survey with Apache Point Observatory (APOGEE-2N), and a southern survey with the 2.5m du Pont Telescope at Las Campanas (APOGEE-2S).

# LSST/ Rubin Observatory

**Survey telescopes**, like at Rubin Observatory, map the night sky by scanning and taking pictures of all parts of the sky instead of taking pictures of one specific object or set of objects.

# LSST/ Rubin Observatory

Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 $3\pi$

Big Data

Fitting a Model

Machine Learning

The LSST Survey

**Survey telescopes**, like at Rubin Observatory, map the night sky by scanning and taking pictures of all parts of the sky instead of taking pictures of one specific object or set of objects.

**Rubin Observatory** is a unique survey telescope

# LSST/ Rubin Observatory



**Survey telescopes**, like at Rubin Observatory, map the night sky by scanning and taking pictures of all parts of the sky instead of taking pictures of one specific object or set of objects.

**Rubin Observatory** is a unique survey telescope

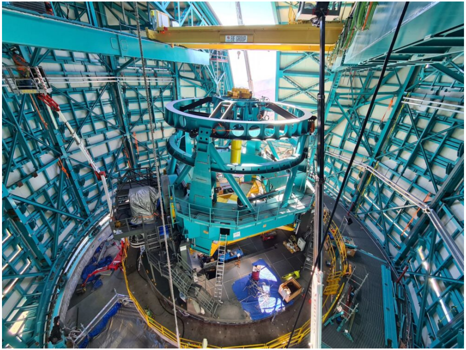
it is specially designed to:

- quickly take huge pictures of the entire Southern hemisphere sky
- repeat those pictures every few nights for ten years
- take those pictures in super-high detail while also being able to see very faint objects

# LSST/ Rubin Observatory



Rubin Observatory[1] will conduct the Legacy Survey of Space and Time (LSST)

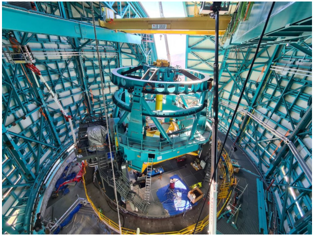


credit: www.rubinobservatory.org

# LSST/ Rubin Observatory



Rubin Observatory[1] will conduct the Legacy Survey of Space and Time (LSST)



credit: www.rubinobservatory.org

? HOW?

# LSST/ Rubin Observatory



? HOW?

Every night for ten years, Rubin Observatory will take hundreds of images of the Southern Hemisphere sky producing about 20 terabytes of data every night.
By the end of the survey, the resulting data set will be enormous: about 60 petabytes!

# LSST/ Rubin Observatory


Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 $3\pi$

Big Data

Fitting a Model

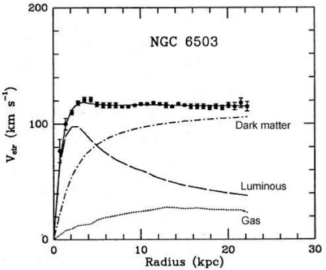
Machine Learning

The LSST Survey


the namesake:

Dr. Vera C. Rubin was an American astronomer who made essential contributions to the study of dark matter by recognizing that galaxy rotation curves show some "missing matter":

credit: Vassar College Library

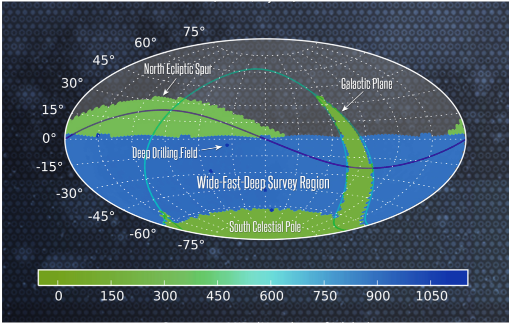



K.G. Begeman, A.H. Broels, R.H. Sanders. 1991. Mon.Not.RAS 249, 523.

stars at the outer edges move just as fast as those towards the center - high velocities caused by some invisible mass holding the galaxies together

# The LSST Survey

- 10-year photometric *ugrizy* survey (near-UV, optical, near-IR)
- depth of r$\sim$27.5 mag
- 1000 images/night = 15 TB/night, 10 million transients/night
- start of operations: 2024




LSST survey strategy, number of visits incl. sub-surveys. (credit: www.lsst.org)

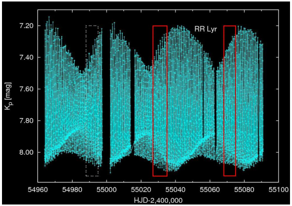
# Research Questions



large data volume of LSST (and other all-sky surveys) enables for

population studies
e.g.: larger samples of RR Lyrae stars to understand the Blazhko effect

finding rare 'one-in-a-million', 'one-in-a-billion' events, often called *anomalies*

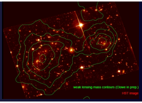



e.g.: extremely low mass (ELM) white dwarf
(El-Badry et al. 2021)

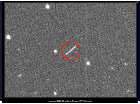
# Science with the LSST Survey

LSST is designed to address four science areas:

Probing Dark Energy and Dark Matter



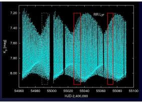

Cataloging the Solar System



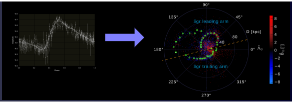
Exploring the Variable/ Transient Optical Sky

TVSSC

Transients and Variable Stars Science Collaboration




Mapping the Milky Way



# LSST/ Rubin Observatory



Rubin Observatory can accomplish all of these things because:

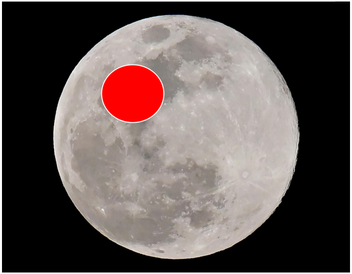
Installation of fiber optic cables on the telescope mount.
credit: www.rubinobservatory.org

- Rubin Observatory's telescope can move much faster than other telescopes its size - it can take pictures faster and create a more detailed map of the night sky
- Rubin Observatory also has a big field of view - one picture covers the same area as 40 full moons
- Rubin Observatory's camera is the highest resolution camera ever created for astronomy and astrophysics
- Rubin Observatory has a big, 8.4 m main mirror that lets it collect a lot of light and see faint objects

# LSST/ Rubin Observatory

big telescopes have (usually) small Fields of View

most telescopes in the 8 - 10 m class image the red circle:



sky: 40,000 deg$^2$
moon: 0.2 deg$^2$
large telescopes: $\sim$0.01 deg$^2$

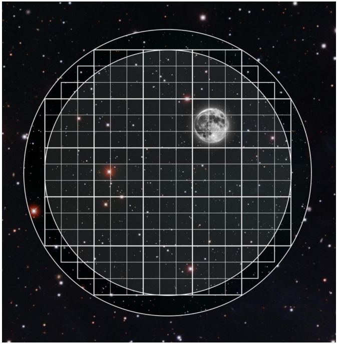
$\Rightarrow$ $\sim$4 million images to cover the sky

# LSST/ Rubin Observatory



big telescopes have (usually) small Fields of View

Rubin Observatory (VRO) FOV:



VRO: 9.6 deg$^2$

$\Rightarrow$ $\sim$4,300 images to cover the sky

$\Rightarrow$ image the whole sky once every 5 nights

# LSST/ Rubin Observatory

## Rubin Observatory Key Numbers[2] (excerpt)

**Telescope System:**

- FOV: 3.5 deg (9.6 $\text{deg}^2$)
- Primary mirror diameter: 8.4 m
- Mean effective aperture: 6.423 m
- Final f-ratio: f/1.234
- Etendue ($A\Omega$): 319 $\text{m}^2\text{deg}^2$
- Camera weight: 3060 kg

Etendue is a measure of the flux gathering capability of an optical system.

etendue = aperture [$\text{m}^2$] $\times$ FOV [$\text{deg}^2$]

[2] https://www.lsst.org/scientists/keynumbers

## Rubin Observatory Key Numbers[2] (excerpt)

**Telescope System:**

- FOV: 3.5 deg (9.6 deg$^2$)
- Primary mirror diameter: 8.4 m
- Mean effective aperture: 6.423 m
- Final f-ratio: f/1.234
- Etendue ($A\Omega$): 319 m$^2$deg$^2$
- Camera weight: 3060 kg

Etendue is a measure of the flux gathering capability of an optical system.
etendue = aperture [m$^2$] $\times$ FOV [deg$^2$]

**Dataset:**

- Nightly data size: 20TB/night
- Final database size (DR11): 15 PB
- Real-time alert latency: 60 seconds

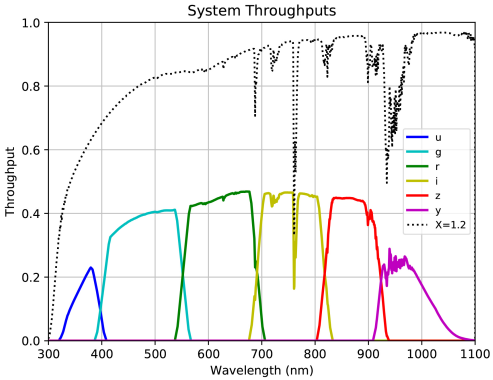
[2] https://www.lsst.org/scientists/keynumbers

# LSST/ Rubin Observatory



## Rubin Observatory Key Numbers[2] (excerpt)

**Spectral response/throughputs:**




[2] https://www.lsst.org/scientists/keynumbers

Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 $3\pi$

Big Data

Fitting a Model

Machine Learning

The LSST Survey

survey design

science cases lead to **competing constraints** on the LSST Survey Strategy:

e.g.:

**Cosmological parameter estimation** requires uniform coverage of 18,000 deg$^2$. Obtaining accurate photometric redshifts requires a specified number of visits in each filter.

**Weak lensing shear measurements** benefit from allocating times of best seeing to observations in the $r$ and $i$ bands. Maximizing S/N requires choosing the next filter based upon the current sky background.

**Supernova** cosmology requires frequent, deep photometry in all bands.

Detecting the motion of **solar system objects** and transients, characterizing **stellar variability** on various timescales, and acquiring the best proper motions and parallaxes place further demands upon the distribution of revisit intervals and observation geometries to each point on the sky.

# LSST/ Rubin Observatory



## survey design

science cases lead to **competing constraints** on the LSST Survey Strategy:

Synthesizing the requirements to accomplish the four primary science objectives of Rubin Observatory,

- Probing dark energy and dark matter
- Taking an inventory of the Solar System
- Exploring the transient optical sky
- Mapping the Milky Way

Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 $3\pi$

Big Data

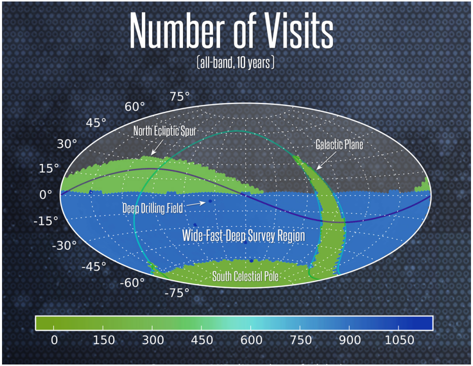
Fitting a Model

Machine Learning

The LSST Survey

survey design

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

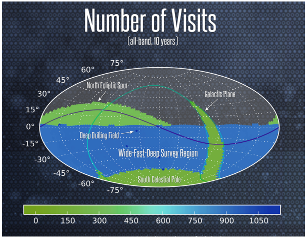



LSST survey strategy, number of visits incl. sub-surveys. (credit: www.lsst.org)

# LSST/ Rubin Observatory

## survey design

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




The survey area and cadence have been (and will be) fine-tuned to support all four science themes and enable the discover and characterization of transient objects.

*A small (<10% of time) set of "special survey programs" is designed to explore extreme corners of discovery space: deep drilling fields, mini-surveys

# LSST/ Rubin Observatory



## survey design

These **fiducial survey plans** can be optimized for science output:

- Operations Simulations (OpSim)
- Image Simulations (ImSim)
- Base catalogs of stars and galaxies in LSST filters (CatSim)
- Key Project Documents (Science Requirements Document, Data Products Definition Document)

The **Operations Simulator (OpSim)** is an application that simulates the field selection and image acquisition process of the LSST over the 10-year life of the planned survey.

It has a sophisticated model of the telescope and dome to properly constrain potential observing cadences.

LSST operations can be simulated using realistic seeing distributions, historical weather data, scheduled engineering downtime and current telescope and camera parameters.

# LSST/ Rubin Observatory



## survey design

The community is participating in this process by evaluating the simulated surveys:

**Cadence Note: Cadence impacts on reliable classification of standard-candle variable stars, including detection of amplitude period, phase modulation effects (e.g., Blazhko effect)**

Nina Hernitschek[1,*] and Keivan G. Stassun[1]

[1] *Vanderbilt University*

1. INTRODUCTION

The Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST) will carry out its science goal of "Mapping the Milky Way" through both astrometry and photometry, with a single-exposure depth of $r{\sim}24.7$ and an anticipated baseline of 10 years. This will enable LSST to access the Milky Way's old halo not only deeper, but also with a longer baseline and better cadence than e.g. PS1 $3\pi$ (Chambers et al. 2016), making this survey ideal to study populations of variable stars such as RR Lyrae (Hernitschek et al. 2016; Sesar et al. 2017a).

As members of the Transients and Variable Stars (TVS) Classification group, we focus on the specific science case of detecting period/ phase shift effects, so-called Blazhko effect (Blazhko 1907), of RR Lyrae stars. So far, due to depth and cadence of typical all-sky surveys, it was nearly impossible to study this effect on a larger sample. Surveys such as PS1 $3\pi$ with relatively few observations over a moderately long baseline allowed only for fitting the period and phase of RR Lyrae stars while integrating over the complete survey length, thus not giving any information regarding whether the period and/or phase of the light curve might have changed during the survey. On the other hand, surveys specialized for detecting slightly changing light curves due to very finely sampled cadence (such as TESS, see Ricker et al. 2015) usually have a relatively small footprint. LSST's cadence and depth, however, will allow for studying variable stars in the Milky Way's old halo in a way that makes population studies possible.

2. SCIENCE CASES

# LSST Data Products



| | |
|---|---|
| readout | **Raw Data**<br>30 s image, 20 TB/night |
| 37 s | **Prompt Data Products**<br>difference image analysis (DIA)<br>alerts: up to 10 million/night |
| 24 h | **Prompt Products Database**<br>images, object & source catalog<br>solar system bodies orbit catalog |
| yearly | **Annual Data Release**<br>via LSST Science Platform<br>& LSST Data Access Centers |
| 10 years | **Final 10 Year Data Release**<br>images: 5.5 million $\times$ 3.2 Gpx<br>catalog: 15 PB, 37 billion objects |

**Alert Brokers**

ALeRCE

FINK

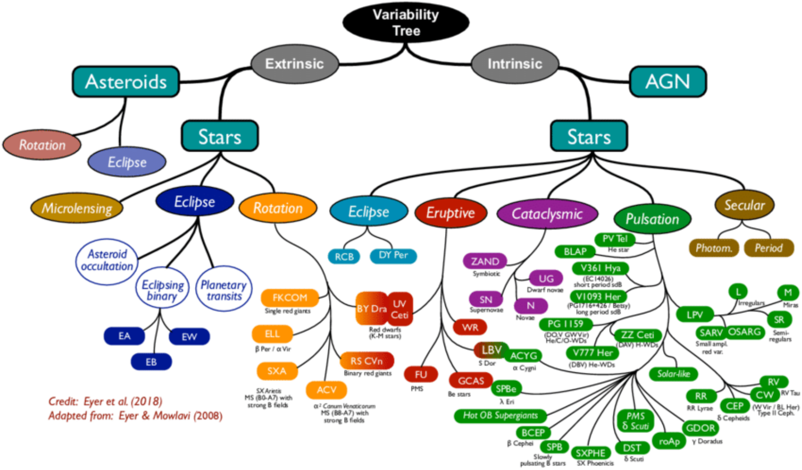
ANTARES

LASAIR

$\Rightarrow$ science as soon as LSST begins

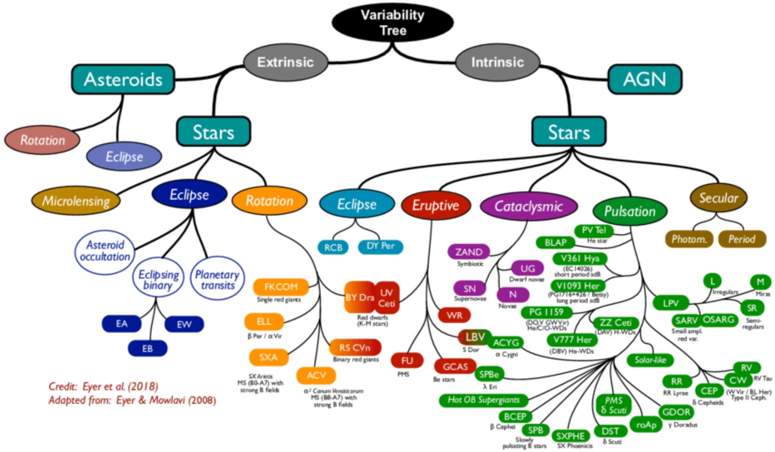
# Variable & Transient Sources with LSST






Credit: Eyer et al. (2018)
Adapted from: Eyer & Mowlavi (2008)

# Variable & Transient Sources with LSST



Credit: Eyer et al. (2018)
Adapted from: Eyer & Mowlavi (2008)

many astronomical sources vary - describe and classify astronomical sources by their variability

# Variable & Transient Sources with LSST



## Rubin LSST Transients and Variable Stars Science Collaboration

TVSSC

`https://lsst-tvssc.github.io/`

one of the LSST Science Collaborations

- Dark Energy
- Solar System
- Transients and Variable Stars
- Stars, Milky Way, and the Local Volume
- Galaxies
- Active Galactic Nuclei
- Strong lensing
- Large-scale Structure
- Informatics and Statistics

# Community



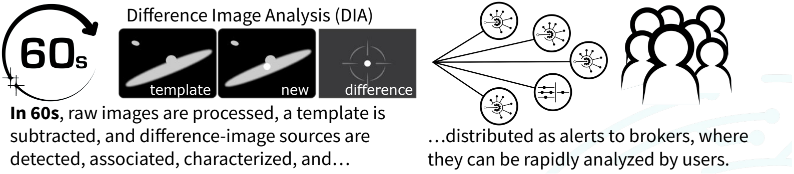
why **citizen science**?

- citizen science is vital for astronomy
- industry drives rapid advances in machine learning
- LSST data rate demands machine learning for identifying time-domain events
- citizen scientists now include thousands of machine learning experts

# Community



within 60 s after readout:
Stream of Alerts is released to Alert Brokers and to the LSST Alert Filtering Service




Alerts: packets of LSST data for a difference image
Brokers: receive & process Alerts (external to LSST)

Automatic Classification of Variable Stars (II)

Recap: Large Astronomical Surveys

Overview

Classifying Pan-STARRS1 $3\pi$

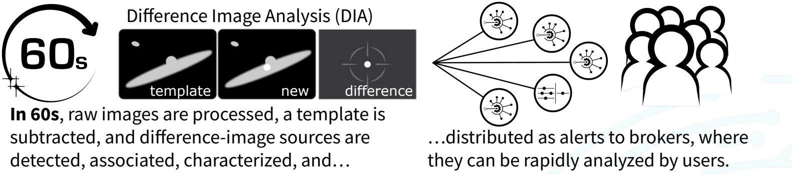
Big Data

Fitting a Model

Machine Learning

The LSST Survey

within 60 s after readout:
Stream of Alerts is released to Alert Brokers and to the LSST Alert Filtering Service




brokers will deliver scientific classification & interpretation to filter sources
Example uses include:

- collections of transient discoveries
- (pre-)classification using features & machine learning
- forwarding to *downstream brokers*
- alerting users
- alert distributions as ways to learn more about object types

# Community



Recap: Large Astronomical Surveys
Overview
Classifying Pan-STARRS1 $3\pi$
Big Data
Fitting a Model
Machine Learning
The LSST Survey

within 60 s after readout:
Stream of Alerts is released to Alert Brokers and to the LSST Alert Filtering Service

brokers currently process a stream from Zwicky Transient Facility (ZTF)

# Community

Cross-Collaboration Work: LSST Data Challenge

Data: Simulated LSST light curves of $\sim$3.5 million objects, including full range of astronomical phenomena

Challenge: Accurately classify the objects based on the available photometry

for the simulation, TVS members contributed models of galactic variability

# Community

## Data challenges: PLAsTiCC /ELAsTiCC

*Photometric LSST Astronomical Time-series Classification Challenge*
and its extension
*The Extended LSST Astronomical Time-Series Classification Challenge*

ELAsTiCC uses simulated alerts, delivered to the alert brokers, to mimic the future rate, volume, and complexity of the LSST prompt data products. Realistic contextual information is incorporated into synthetic alerts.

The Challenge:

- Types are unbalanced
- Small number in the training set
- The training set is not representative of the test data
- Seasonal gaps
- Non-uniform cadence

# Final Remarks

In modern astronomy, big data is crucial as a) challenge and b) chance to enable science.

**theory + data + astroinformatics = great discoveries**