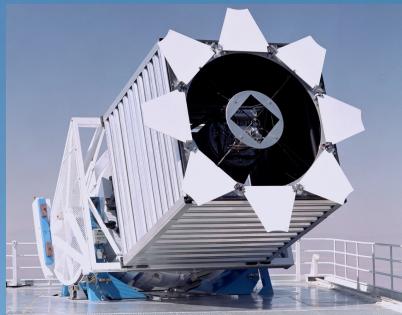


Lecture 0: Introduction

Željko Ivezić, Department of Astronomy, University of Washington

LSST



SDSS



Outline

- Big Data in astronomy
- What is LSST and why should you care?
- Motivation for astroML
 - ever increasing data volume and complexity
 - sophisticated analysis, need for reproducability
 - open-source approach
 - generally useful tools

Big Data is growing fast



Structured and unstructured data

The digital universe will grow to
2.7ZB in 2012, up

48%
from 2011, toward nearly
8ZB by 2015

Americans
USE

18,264,840

MEGABYTES
OF WIRELESS DATA

YOUTUBE

USERS SHARE
400 HOURS
OF NEW VIDEO

FACEBOOK MESSENGER
USERS SHARE
216,302
PHOTOS

Amazon
MAKES
\$222,283
IN SALES

3,567,850

TEXT MESSAGES
ARE SENT

U.S.

DOMO

BUZZFEED

USERS VIEW
159,380
PIECES OF
CONTENT

SNAPCHAT

USERS WATCH
6,944,444
VIDEOS

Netflix

SUBSCRIBERS STREAM
86,805 HOURS
OF VIDEO

GOOGLE

TRANSLATES
69,500,000
WORDS

Instagram

USERS LIKE
2,430,555
POSTS

SIRI ANSWERS
99,206
REQUESTS

Tinder

USERS SWIPE
972,222
TIMES

THE WEATHER CHANNEL

RECEIVES

13,888,889

FORECAST
REQUESTS

DOMO

2016
every
MINUTE
of
the
DAY

PRESENTED BY DOMO

Giphy

SERVES
569,217
GIFS

TWITTER
USERS SEND
9,678
EMOJI-FILLED TWEETS

DOMO

DATA NEVER SLEEPS 4.0

How much data is generated every minute? In the fourth annual edition of Data Never Sleeps, newcomers like Giphy and Facebook Messenger illustrate the rise of our multimedia messaging obsession, while veterans like YouTube and Snapchat highlight our insatiable appetite for video. Just how many GIFs, videos, and emoji-filled Tweets flood the internet every minute? See for yourself below.

Sky Survey Projects	Data Volume
DPOSS (The Palomar Digital Sky Survey)	3 TB
2MASS (The Two Micron All-Sky Survey)	10 TB
GBT (Green Bank Telescope)	20 PB
GALEX (The Galaxy Evolution Explorer)	30 TB
SDSS (The Sloan Digital Sky Survey)	40 TB
SkyMapper Southern Sky Survey	500 TB
PanSTARRS (The Panoramic Survey Telescope and Rapid Response System)	~ 40 PB expected
LSST (The Large Synoptic Survey Telescope)	~ 200 PB expected
SKA (The Square Kilometer Array)	~ 4.6 EB expected

**"Ask Not What Data You Need To Do Your Science,
Ask What Science You Can Do With Your Data."**



The era of surveys...

- Standard: “What data do I have to collect to (dis)prove a hypothesis”?
- Data-driven: “What theories can I test given the data I already have?”



Alternative Careers: Leveraging your Astronomy Degree for Data Science

by Ben Cook | Jun 1, 2016 | Career Navigation, Personal Experiences | 0 comments

Big Data in Astronomy

Alongside the recent explosion of “[Big Data](#)” into the public consciousness, there has been a similar transition into the age of “[Big Astronomy](#)”. Astronomers have always been adept at drawing conclusions using [advanced statistics](#) and [data analysis](#). Now, with the advent of extremely large simulations like [Illustris](#) and surveys like the upcoming LSST, astronomers are increasingly gaining experience in dealing with [datasets vastly larger](#) than could ever hope to fit on a single computer.

For early career astronomers looking for advice, I think you can do no better than look at the posts made by Jessica Kirkpatrick, who obtained a PhD in Astronomy and then became a data scientist at Microsoft/Yammer, and I understand she has since taken a position as Director of Data Science at the education start-up [InstaEDU](#).

The term “Data Scientist” is extraordinarily broad. For example, the post “[What is a Data Scientist?](#)” describes some of the Data Analyst roles a Data Scientists may play:

- Derive business insight from data.
- Work across all teams within an organization.
- Answer questions using analysis of data.
- Design and perform experiments and tests.
- Create forecasts and models.
- Prioritize which questions and analyses are actionable and valuable.
- Help teams/executives make data-driven decisions.
- Communicate results across the company to technical and non-technical people.



LSST in one sentence:

An optical/near-IR survey of half the sky
in ugrizy bands to $r \sim 27.5$ based on
 ~ 800 visits over a 10-year period:

$3.6 \times 10^{-31} \text{ erg/s/cm}^2/\text{Hz}$

More information at
www.lsst.org
and arXiv:0805.2366

A catalog of 20 billion stars and 20 billion galaxies with
exquisite photometry, astrometry and image quality!

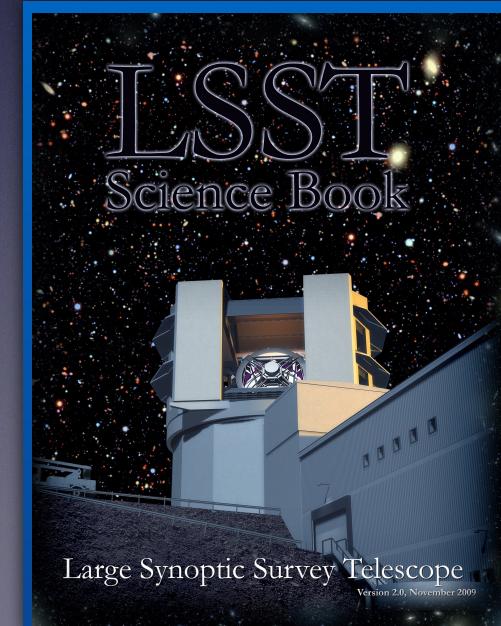
LSST Science Themes

- 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

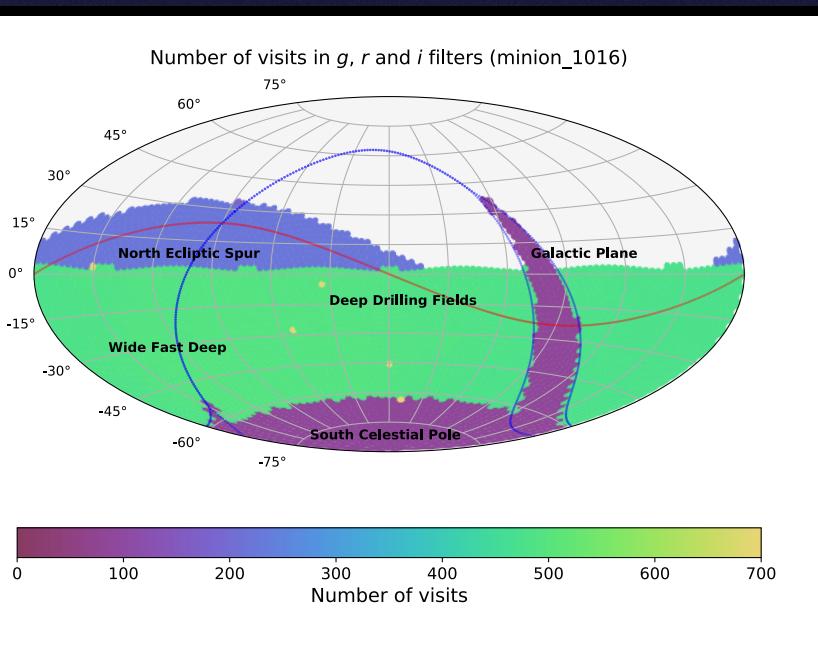


Large Synoptic Survey Telescope

Version 2.0, November 2009

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 $r \sim 27.5$ (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

gri

3.5'x3.5'

r~22.5



HSC

gri

3.5'x3.5'

r~27

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

like LSST
depth (but
tiny area)

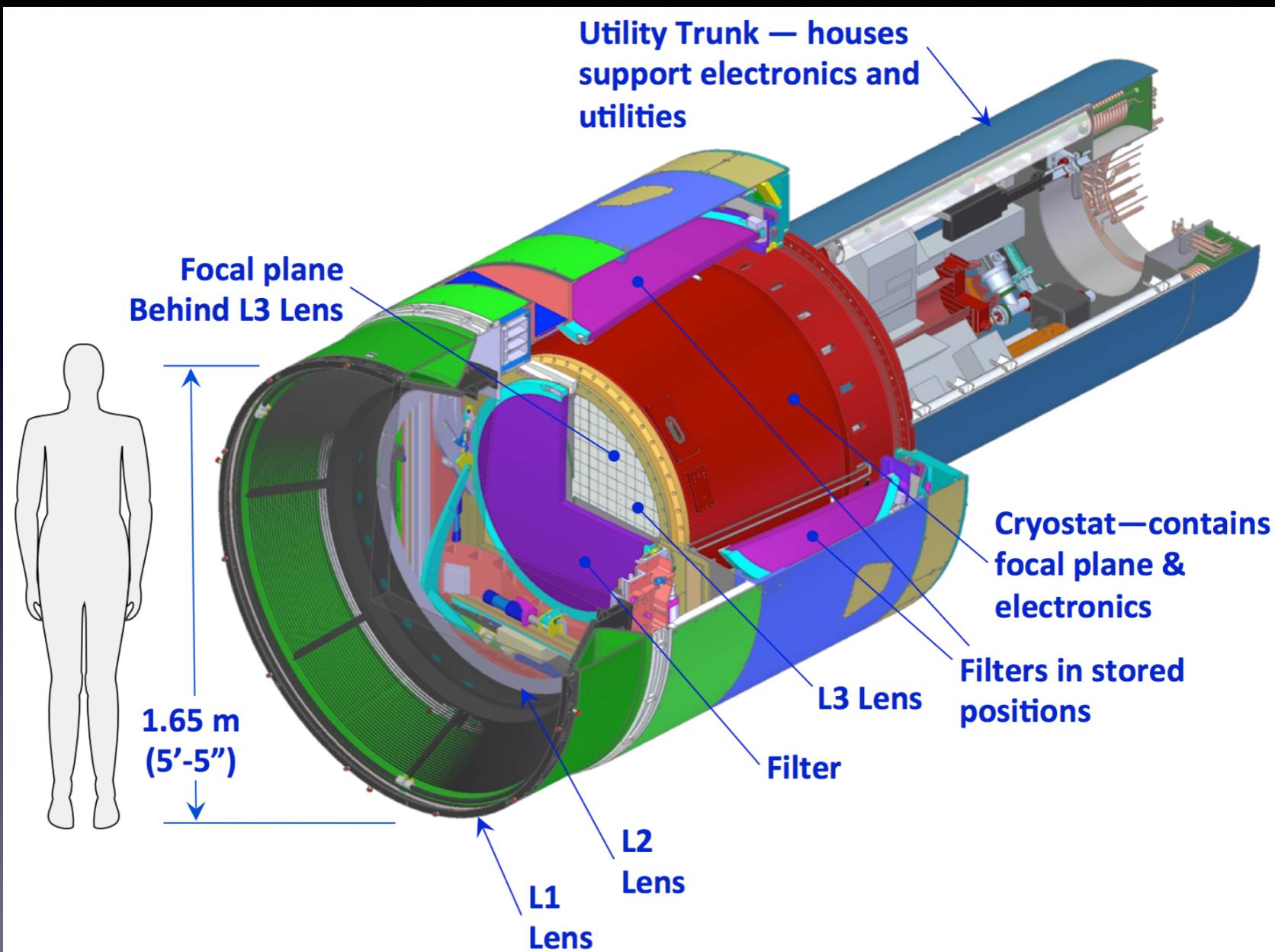
LSST will
deliver 5
million such
images



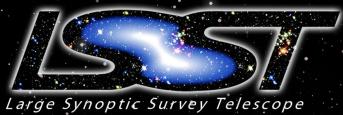


Telescope Mount Assembly before going from Spain to Chile

LSST camera



The largest astronomical camera: 2800 kg, 3200 Megapix



LSST Operations: Sites & Data Flows

HQ Site

Science Operations
Observatory Management
Education & Public Outreach

Base Site

Base Center
Long-term storage (copy 1)
Data Access Center
Data Access & User Services



French Site

Satellite Processing Center
Data Release Production
Long-term Storage (copy 3)

Archive Site

Archive Center
Alert Production
Data Release Production
Calibration Products Production
EPO Infrastructure
Long-term Storage (copy 2)

Data Access Center

Data Access and User Services

Summit Site

Telescope & Camera
Data Acquisition
Crosstalk Correction

Mar 10, 2019



First light: 2021
Operations: 2022



Some obvious points...

- **Ever increasing data volume and complexity**
 - SDSS is ~30 TB; LSST will be one SDSS per night, or a total of >100 PB of data (40 billion objects); of course, also Gaia and many other surveys
 - who and how will do the required data analysis?
- **Sophisticated analysis, need for reproducability**
 - with the increasing data complexity, analysis becomes more complex, too; what do we do in case of disagreement?
- **Open-source approach improves efficiency**
 - we are not data starved any more!
 - the bottleneck for new results is in human resources (as in “grad students and postdocs”) and analysis tools
 - nobody has an unlimited budget; collaborate and share!

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 (irregular sampling)
- 6) Heteroscedastic errors, truncated, censored and missing data
- 7) Unreliable quantities (e.g. unknown systematics and random errors)

The bottleneck is not any more data availability but instead our 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)

● 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)

These topics are not typically taught to astronomers in graduate school.
Mario Jurić: “Astrostatistics and astroinformatics are the new calculus!”

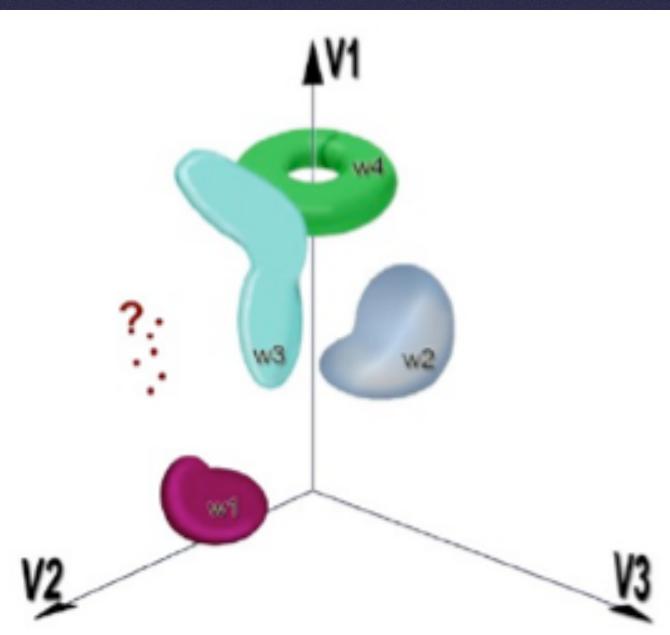
● Some tools and methods...

- 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)

Statistical analysis of a massive LSST dataset

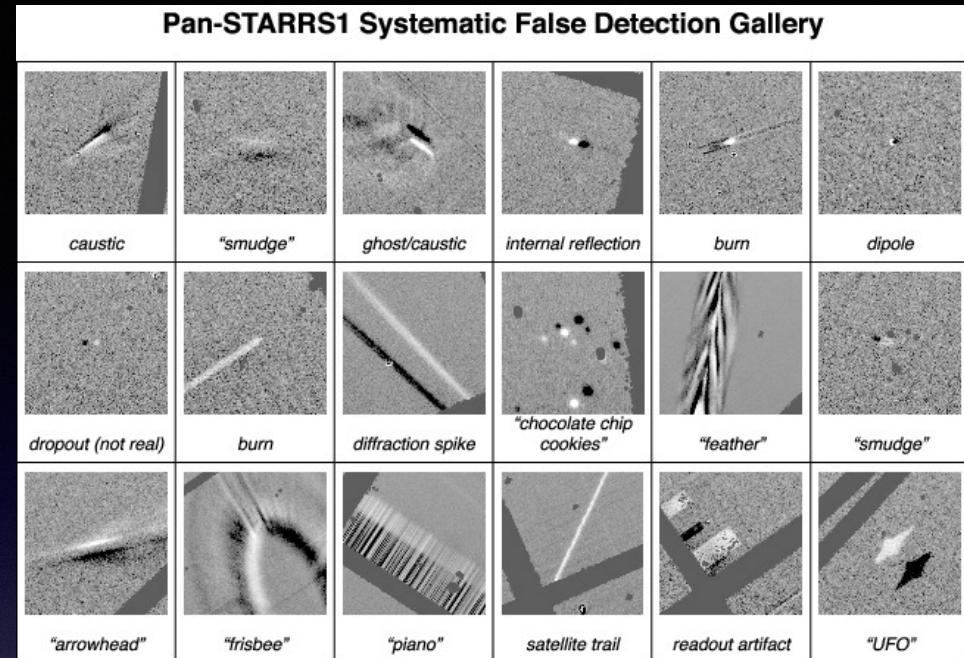
- A large (100 PB) database and sophisticated analysis tools: for each of 40 billion objects there will be about 1000 measurements (each with a few dozen measured parameters)

Data mining and machine learning:

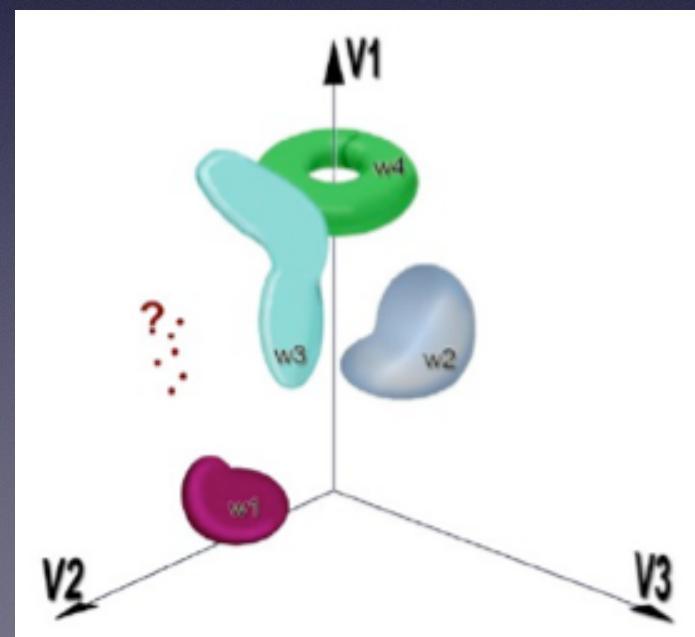


- 10,000-D space with 40 billion points
 - Characterization of known objects
 - Classification of new populations
 - Discoveries of unusual objects
- Clustering, classification, outliers**

It is often not straightforward to identify alerts worth following: need to first reject instrumental artefacts, and then **classify** astrophysical sources to select those that are “most interesting”



Clustering, classification, outliers...



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
- Everything I'd like to do with LSST data, but don't know (yet) how**
- For more details: Ivezić, Connolly & Jurić 2016,
arXiv:1612.04772
- Main work areas:**
- astronomical digital image processing
 - statistical modeling and analysis
 - data mining and machine learning
 - high performance computing

News

October 2012: astroML 0.1 has been released! Get the source on [Github](#)

Our Introduction to astroML paper received the CIDU 2012 best paper award.

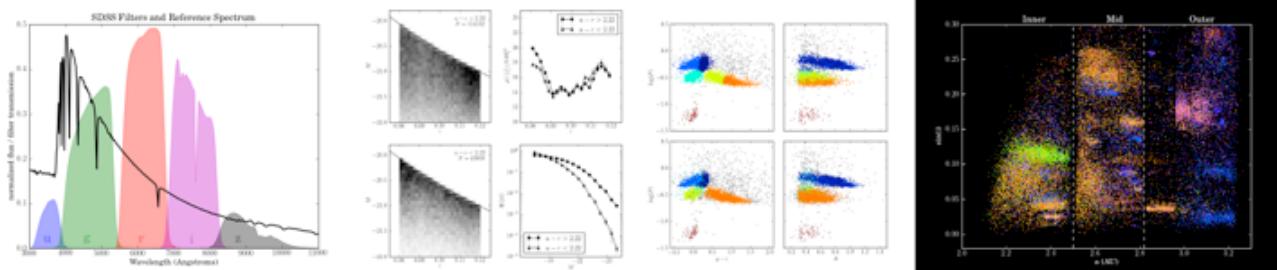
Links

[astroML Mailing List](#)

[GitHub Issue Tracker](#)

Videos

AstroML: Machine Learning and Data Mining for Astronomy

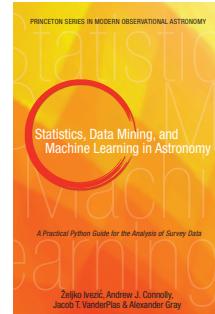


AstroML is a Python module for machine learning and data mining built on [numpy](#), [scipy](#), [scikit-learn](#), and [matplotlib](#), and distributed under the 3-clause BSD license. It contains a growing library of statistical and machine learning routines for analyzing astronomical data in python,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

How do we efficiently train our astronomy students to use sophisticated methods from statistics, data mining and machine learning?

Downloads

- Released Versions: [Python Package Index](#)
- Bleeding-edge Source: [github](#)



User Guide

1. Introduction

- 1.1. Philosophy

Open source!
www.astroml.org

Textbook Figures

This section makes available the source code used to generate every figure in the book *Statistics, Data Mining, and Machine Learning in Astronomy*. Many of the figures are fairly self-explanatory, though some will be less so without the book as a reference. The table of contents of the book can be seen [here \(pdf\)](#).

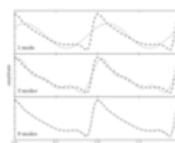
Figure Contents

Each chapter links to a page with thumbnails of the figures from the chapter.

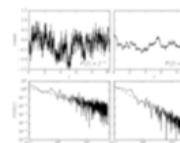
- Chapter 1: Introduction
- Chapter 2: Fast Computation and Massive Datasets
- Chapter 3: Probability and Statistical Distributions
- Chapter 4: Classical Statistical Inference
- Chapter 5: Bayesian Statistical Inference
- Chapter 6: Searching for Structure in Point Data
- Chapter 7: Dimensionality and its Reduction
- Chapter 8: Regression and Model Fitting
- Chapter 9: Classification
- Chapter 10: Time Series Analysis
- Appendix

Chapter 10: Time Series Analysis

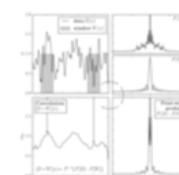
This chapter covers the analysis of both periodic and non-periodic time series, for both regularly and irregularly spaced data.



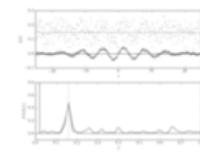
Fourier Reconstruction of
RR-Lyrae Templates



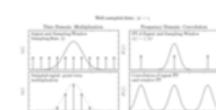
Generating Power-law
Light Curves



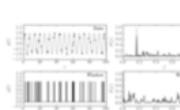
Plot a Diagram explaining
a Convolution



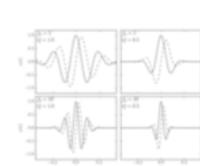
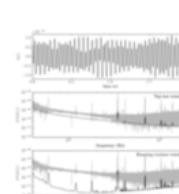
Fast Fourier Transform
Example



The effect of Sampling



The effect of Sampling



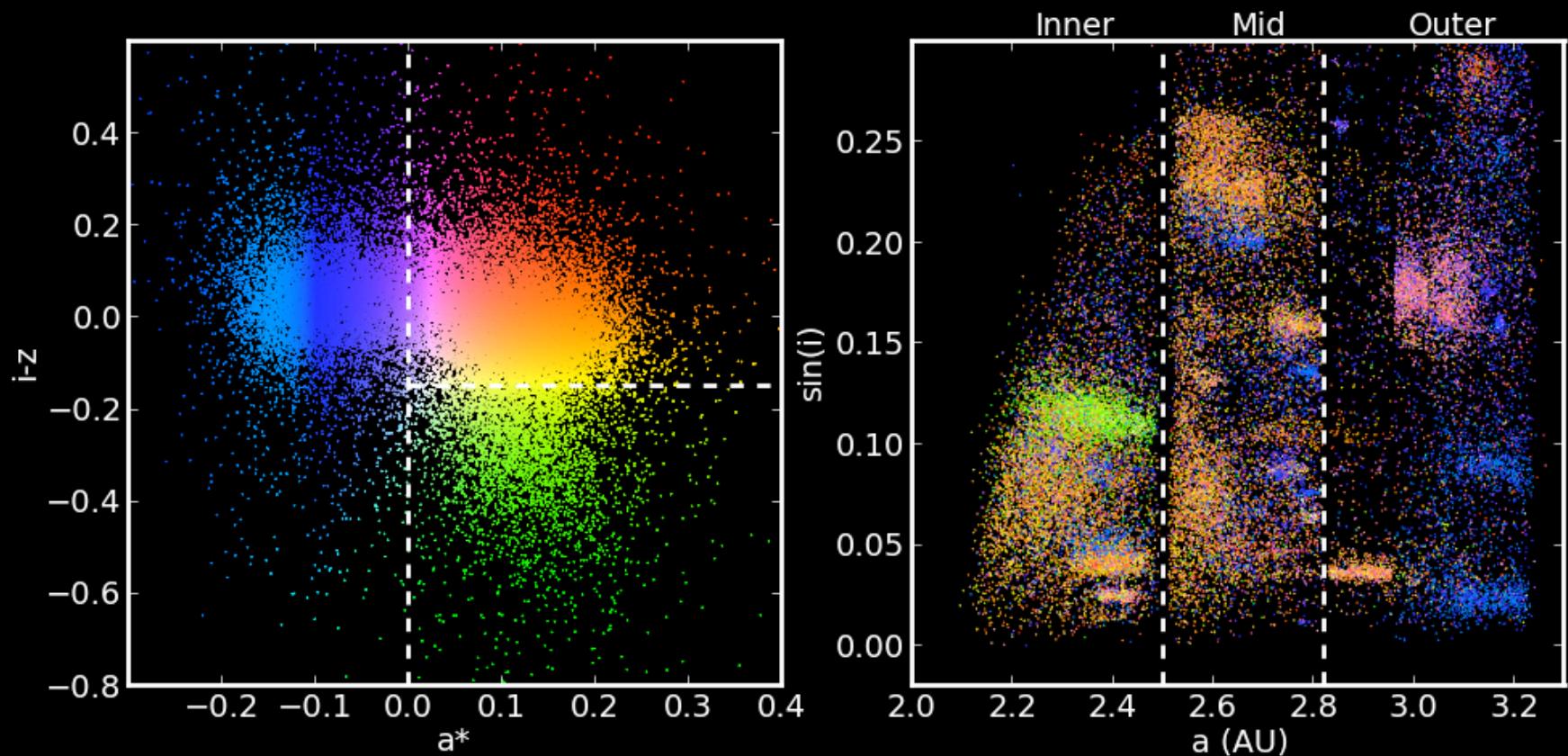
Examples of Wavelets

Plot the power spectrum of
the LIGO big dog event

Example:
SDSS
asteroids

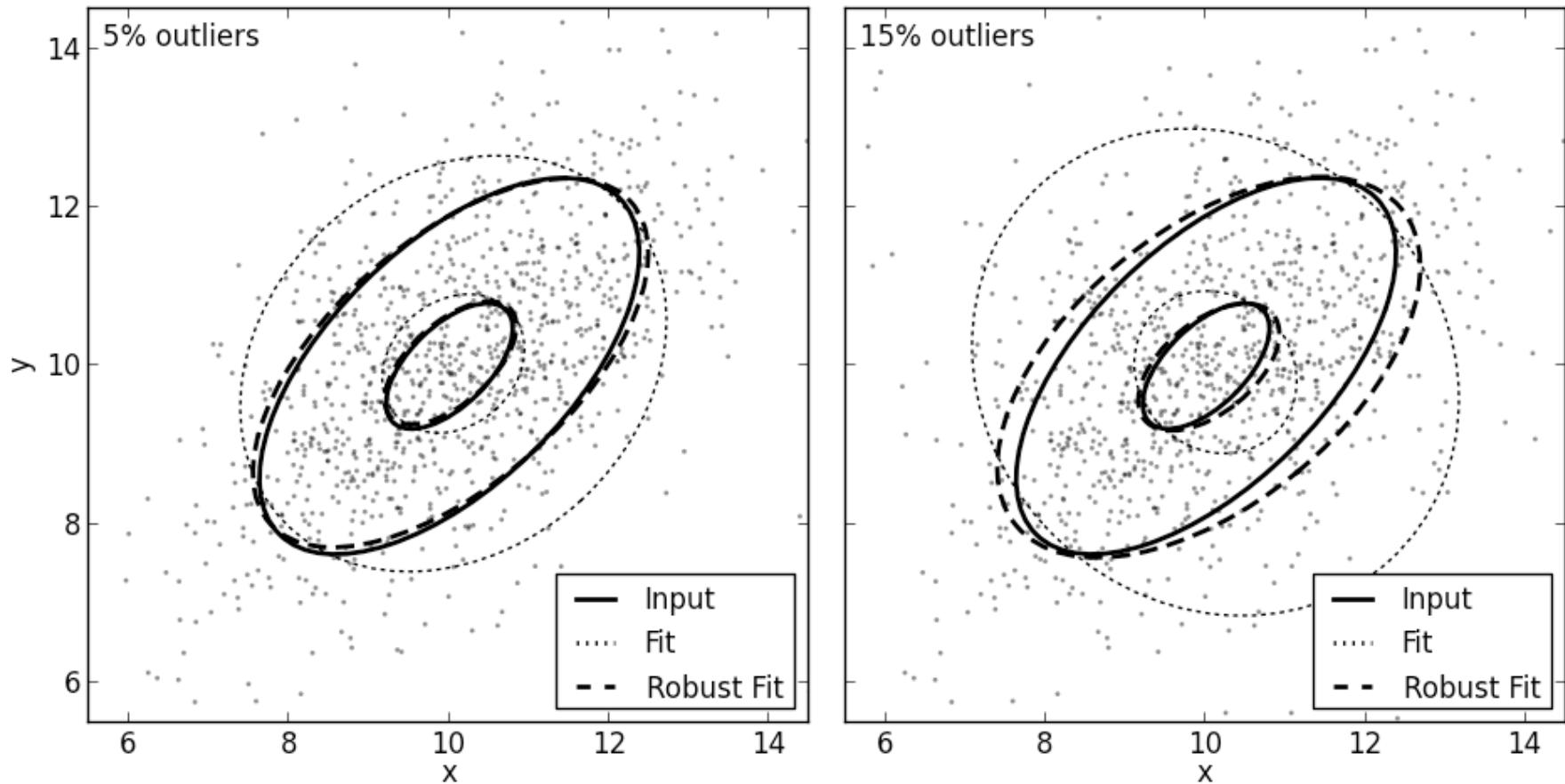
You can make this plot from scratch in <1 hour!

Visualization of 4-dimensional correlations



- Robust parameters for a 2-dimensional gaussian

standard methods are not robust to outliers



If you feel uneasy with git and/or python/jupyter,
please peruse

[https://github.com/uw-astr-324-s17/astr-324-s17/blob/master/notebooks/
Week-1-Thu.ipynb](https://github.com/uw-astr-324-s17/astr-324-s17/blob/master/notebooks/Week-1-Thu.ipynb)

Additional Resources:

- E-Science 2015 seminar on python:
<https://github.com/uwescience/python-seminar-2015>
Includes introduction to python, git, matplotlib and pandas.
- Big Data in Astronomy: Hands-on with Large Surveys (Astr 597,
by Mario Juric) https://github.com/mjuric/astr597b_wi16
excellent lectures on python, numpy, github, matplotlib, databases
- Scikit-learn Tutorial by Jake VanderPlas
https://github.com/jakevdp/sklearn_tutorial

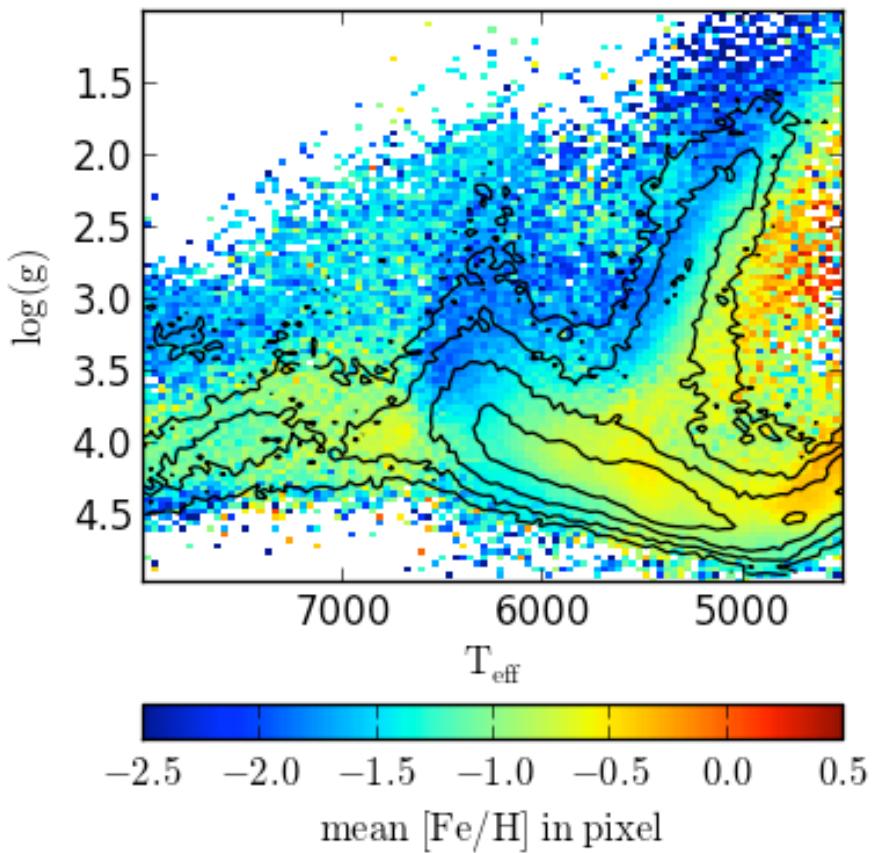
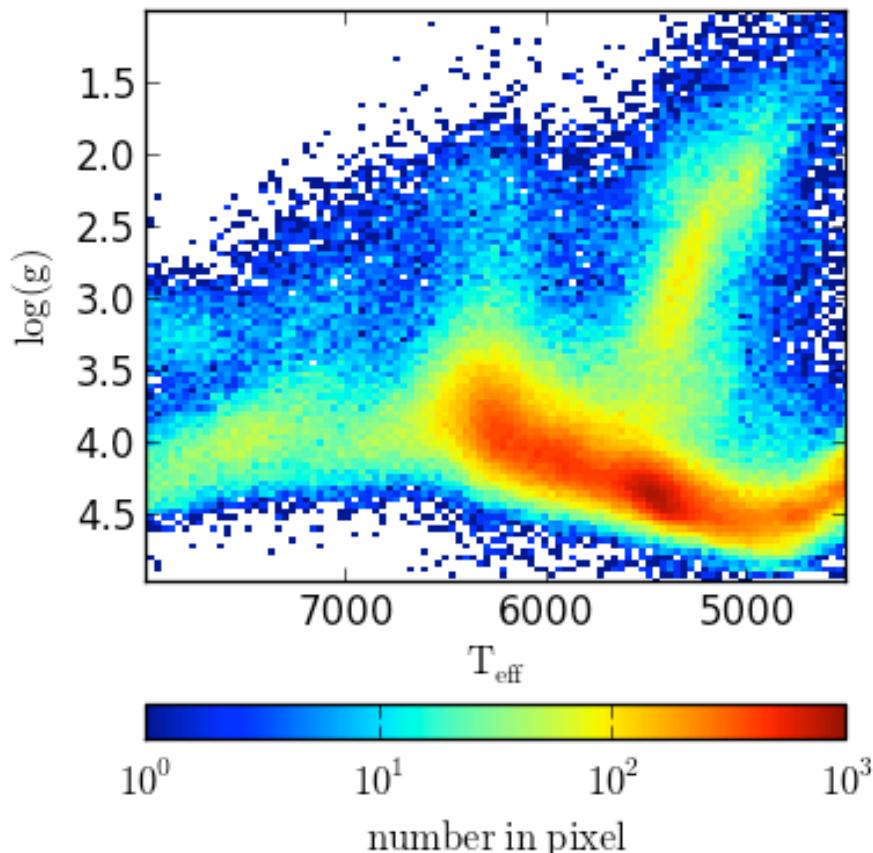
Even more resources:

- Concise “handbook”: *Notes on statistics for physicists* by Orear,
<http://www.astro.washington.edu/users/ivezic/Teaching/Astr507/orear.pdf>
- A great book: *Probability Theory: The Logic of Science* by Jaynes,
<http://bayes.wustl.edu/etj/prob/book.pdf>
- A book about python and data science by Jake VanderPlas:
<https://github.com/jakevdp/PythonDataScienceHandbook>
- An intro class by Gordon Richards (Drexel University):
https://github.com/gtrichards/PHYS_T480
- An advanced class by Phil Marshall (Stanford University):
<https://github.com/KIPAC/StatisticalMethods>
- LSST Data Science Fellowship Program:
<https://github.com/LSSTC-DSFP/LSSTC-DSFP-Sessions>
- TED talk “The best stats you’ve ever seen” by Hans Rosling:
<http://ls.st/0dt>

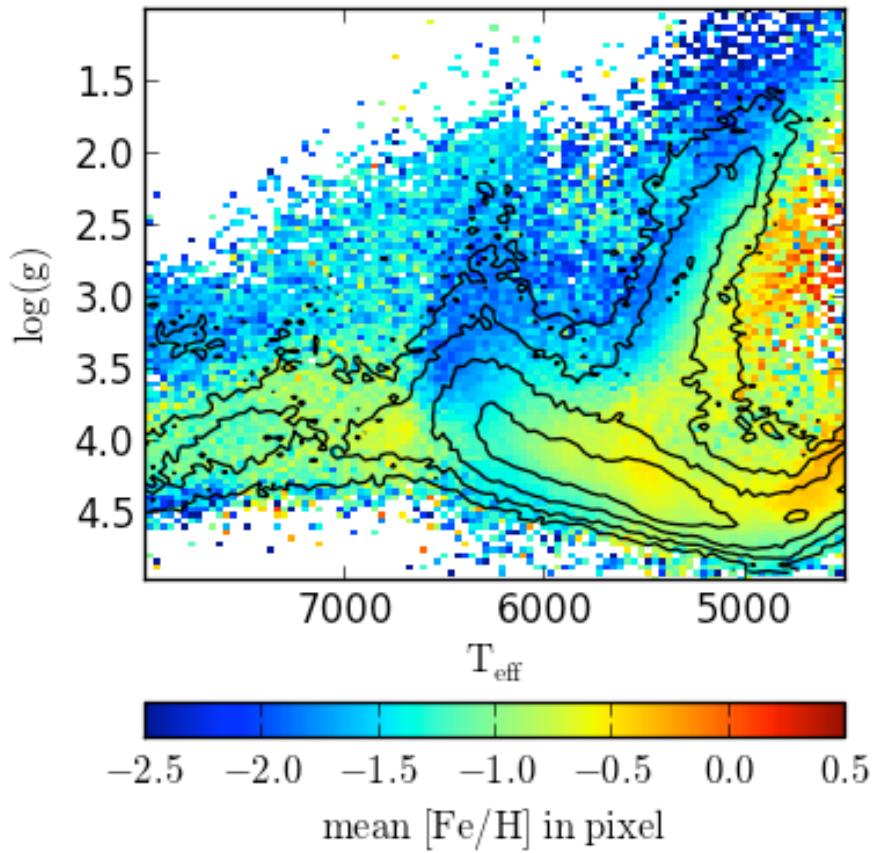
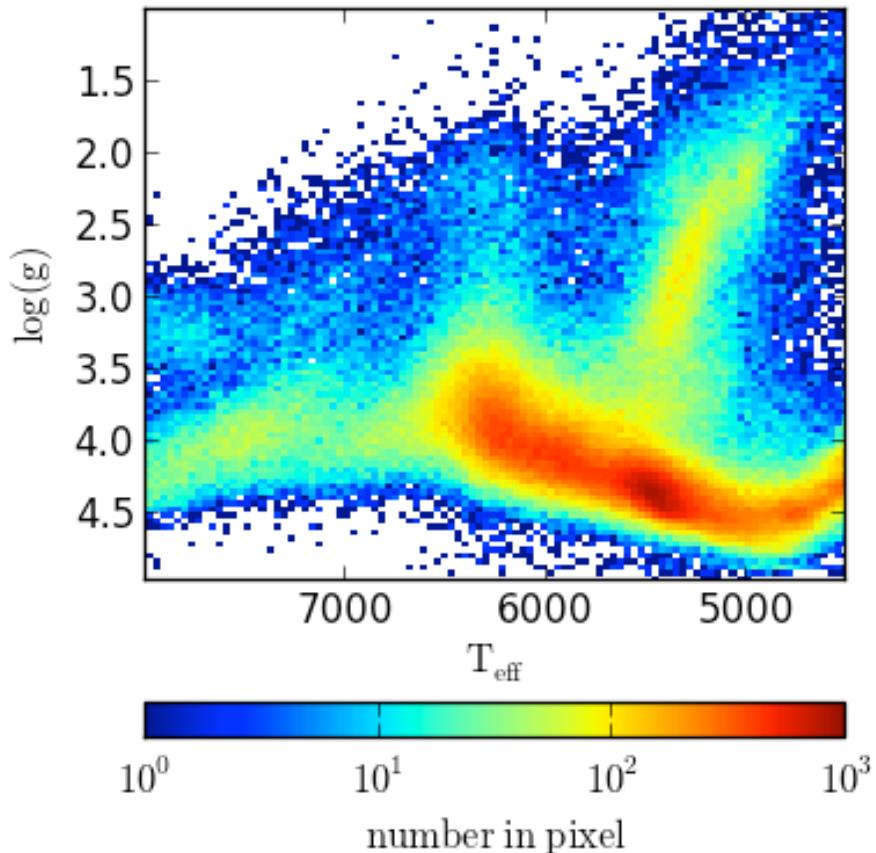
- A simple example of astroML use:
a Hess diagram coded by a third quantity
- Hess diagram is astronomical term for pixelated
color-magnitude diagram, where each pixel is coded
to display the number of objects in it (also known as
“two-dimensional histogram”)

- A Hess diagram coded by a third quantity

- Here we plot a measure of the star's surface gravity strength vs. effective temperature (both estimated from a spectrum obtained by SDSS); the left panel shows the count of stars in each pixel



- Of course, the pixels don't have to be coded by the number of objects in it - we can use instead any other statistic (the mean, median, scatter, etc): the right panel shows the mean metallicity [Fe/H] by color and the same counts as in the left panel, but now using contours



AstroML testing

[https://github.com/carmensg/IAA_School2019
/blob/master/lectures/Day3-Zeljkolvezic/notebooks/astroMLtesting.ipynb](https://github.com/carmensg/IAA_School2019/blob/master/lectures/Day3-Zeljkolvezic/notebooks/astroMLtesting.ipynb)

Slide Type

```
1 ! curl -O https://raw.githubusercontent.com/astroML/astroML/master/examples/datasets/plot_SDSS_SSPP.py
```

```
% Total     % Received % Xferd  Average Speed   Time      Time      Time  Current
                                         Dload  Upload   Total    Spent    Left  Speed
100  3049  100  3049      0       0  9292      0 --::--- --::-- --::-- 12147
```

Slide Type

```
1 %matplotlib inline
2 %run ./plot_SDSS_SSPP.py
```

