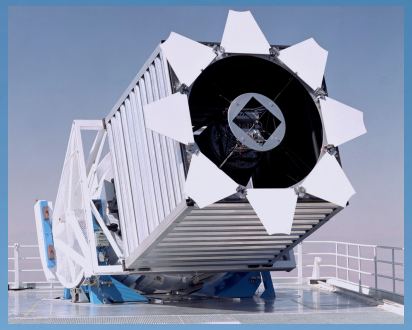
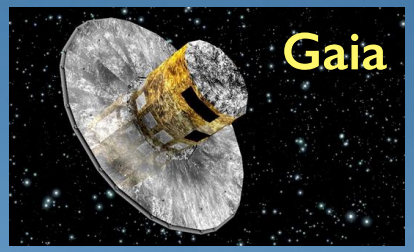
Week 0: Introduction to class

Andy Connolly & Željko Ivezić, Department of Astronomy, UW



SDSS



LSST

Outline

- Syllabus and class organization
- Class Project
- Motivation for astrostatistics and this class
  - ever increasing data volume and complexity
  - sophisticated analysis, need for reproducability
  - open-source approach
  - generally useful tools

- Syllabus

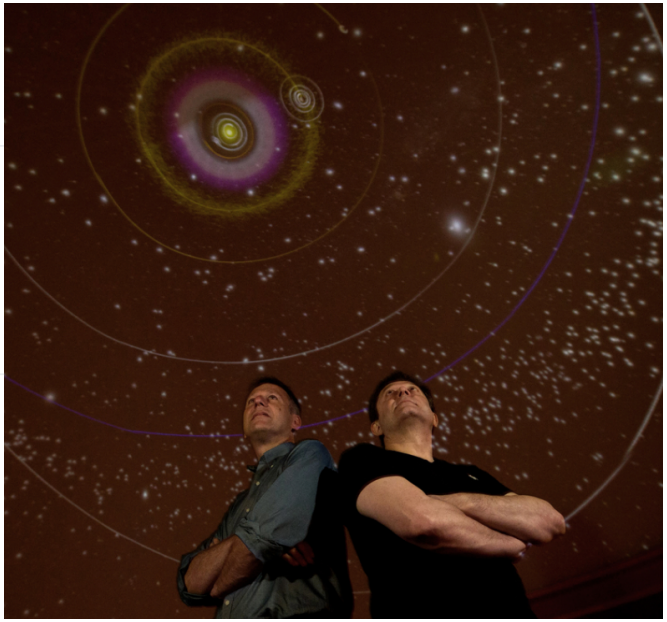
everything is on GitHub:

https://github.com/dirac-institute/uw-astr598-w18

ASTR 598, Winter 2018, University of Washington:

Astro-statistics and Machine Learning

Andy Connolly and Željko Ivezić



Location

- When: 11:00-12:20, Tuesday & Thursday, Winter quarter 2018
- Where: PAB B305 (close to the end of the grad student hallway)

Class Materials

- Syllabus and course description
- Lectures
- Class Project

Reference textbook

Ivezić, Connolly, VanderPlas & Gray: Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data (Princeton University Press, 2014)

Class Description

Class Schedule (tentative!):

- WEEK 0 (first class on Jan 4, Željko): Introduction to class (syllabus, literature, astroML, class project, DataLab)
- WEEK 1 (starting Jan 8, Željko): Introduction to statistics (probability, distributions, robust statistics, Central Limit Theorem, hypothesis testing).
- WEEK 2 (starting Jan 15, Željko): Work on the Class Project (step 1).
- WEEK 3 (starting Jan 22, Andy): Maximum likelihood and applications in astronomy.
- WEEK 4 (starting Jan 29, Andy): Bayesian statistics and introduction to MCMC; Class Project step 2 report.
- WEEK 5 (starting Feb 5, Željko): Model parameter estimation and model selection.
- WEEK 6 (starting Feb 12, Željko): Regression and Time series analysis; Class Project step 3 report.
- WEEK 7 (starting Feb 19, Andy): Dimensionality reduction; Class Project step 4 report.
- WEEK 8 (starting Feb 26, Andy): Density estimation and clustering; Class Project step 5 report.
- WEEK 9 (starting Mar 5, Željko): Supervised Classification; writing assignments for the Class Project paper
- FINAL EXAM: Mar 13 (Tue, 11:00-12:20, B305): cake and closed book final exam.

Class Project

A quarter-long survey data analysis project is described separately (see the class website). We will use GitHub and Jupyter notebooks for progress tracking.

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

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

Class Project: Deep Proper Motion Catalog for SDSS Stripe 82 Regi

Scientific Goals: The main scientific aim of this quarter-long class project is to produce a catalog with improved proper motion measurements for faint stars ($14 < r < 22$) in the SDSS Stripe 82 region (250 sq. deg.). The SDSS-based catalog[1] for 3.7 million objects constructed by Bramich et al. will be augmented with newer additional data obtained with the Dark Energy Camera[2], and possibly from other surveys, and the proper motions will be refit. If all goes well, the results should be publishable in a top astronomical journal.

Learning Goals: While working on this project, students will develop a working knowledge of the NOAO Datalab interface, astropy, pandas, astroML and other astronomical python tools, and of selected methods from astro-statistics (e.g., robust regression, Bayesian statistics, clustering, visualization).

Prerequisites: The students taking this class are required to open an account at the NOAO Data Lab site[3]. The Data Lab allows users to

- access, search, and filter databases containing large catalogs;
- create custom databases and analyses from large catalogs using familiar tools;
- combine catalog databases with data from NOAO telescopes, analysis results, and data from external archives in one place;
- share custom results easily with collaborators and create and publish catalogs derived from large data sets through a central workspace;
- experiment with tools being developed for LSST using existing large data sets.

We will use GitHub and Jupyter notebooks for progress tracking.

Brief Project Outline

We will discuss detailed work plan in class (the 3rd week) and here we list only an overview of the main steps:

1. Download Bramich et al. catalog (HLC files) from the SDSS website[4] and perform preliminary analysis (e.g., the variation of proper motion errors with position and magnitude; also, see examples in the last bullet below and, in particular, reproduce Figs. 22 and 23 from Sesar et al.); Using Data Lab, perform similar analysis for DECaLS and other available data from the Stripe 82 region.

2. Cross-match to DECaLS data using Data Lab (technical details TBD).

3. Preliminary analysis and quality assurance of the assembled data set (e.g., coordinates vs. time plots for stars with large SDSS proper motions); decision whether astrometric recalibration is required (perhaps using galaxies).

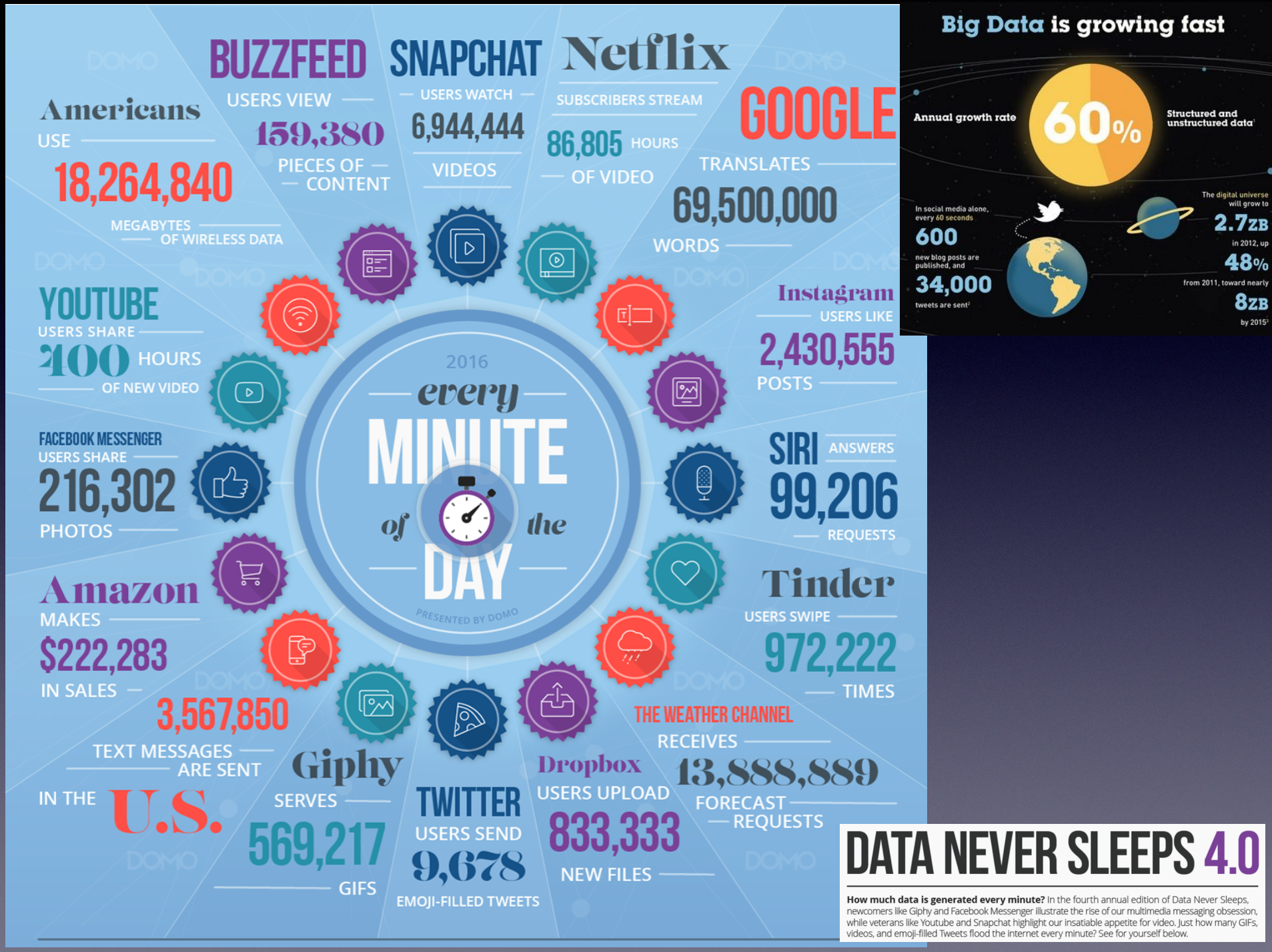
4. Fit proper motions (see eqs. 3-6 in Bramich et al.) using the Bramich et al. SDSS values as priors. For coordinate sytstems, please see Section 2.6 in Bond et al. (2010, ApJ 716, 1).

5. Analyze updated proper motions (e.g., Fig. 2 in Vidrih et al. 2007, MNRAS 382, 515; Figs. 22 and 23 from Sesar et al. 2010, AJ 708, 717, color-coded with the mean proper motions), including using quasars for the assessment of systematic errors.

6. Write paper(s) and become rich and famous!

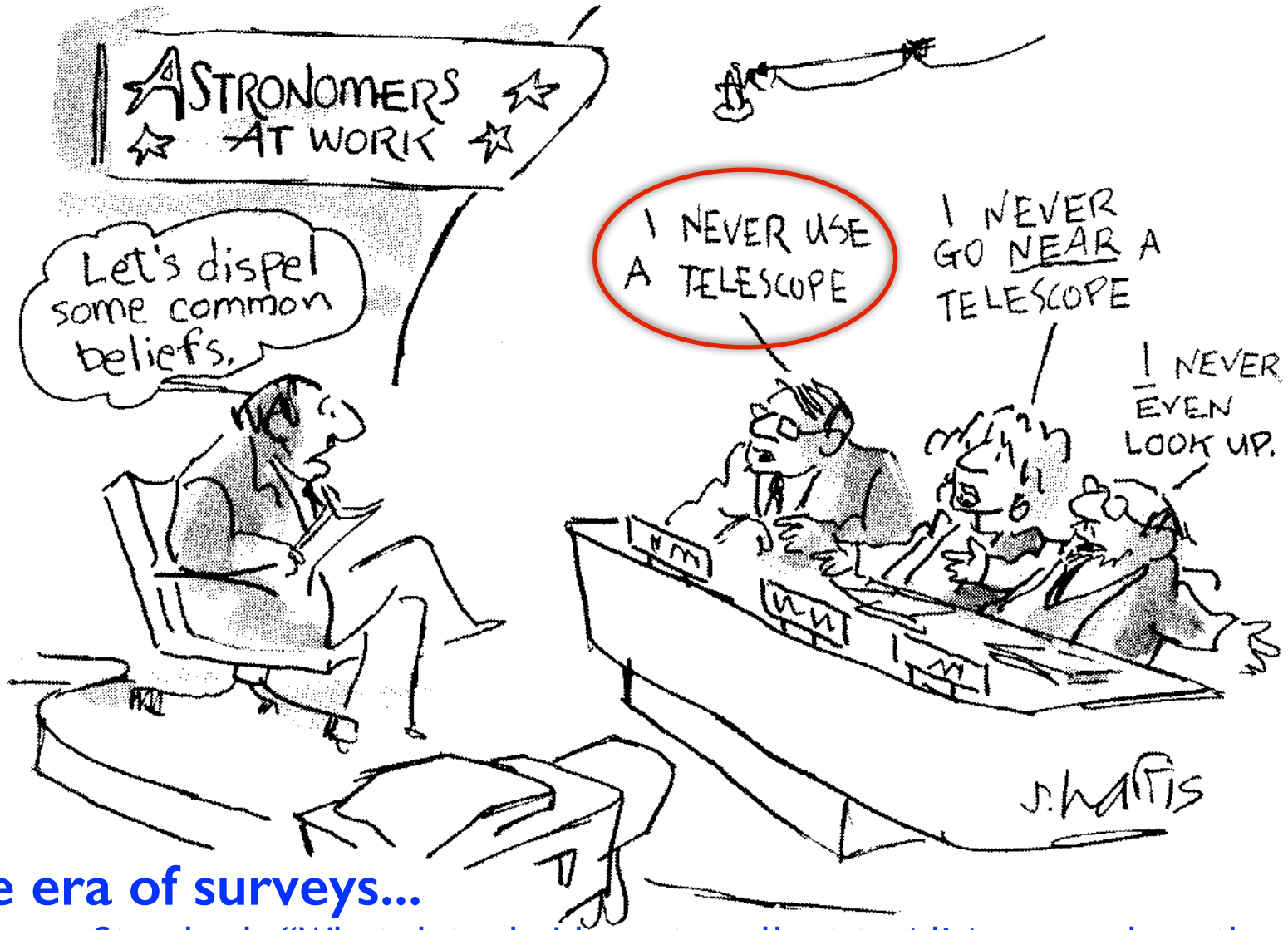
Outline

- Syllabus and class organization
- Class Project
- Motivation for astrostatistics and this class
  - ever increasing data volume and complexity
  - sophisticated analysis, need for reproducability
  - open-source approach
  - generally useful tools



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.



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?"

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

Zhang & Zhao (2015): http://datascience.codata.org/articles/10.5334/dsj-2015-011/

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.

https://astrocompute.wordpress.com/2016/01/29/how-does-an-astronomer-become-a-data-scientist/

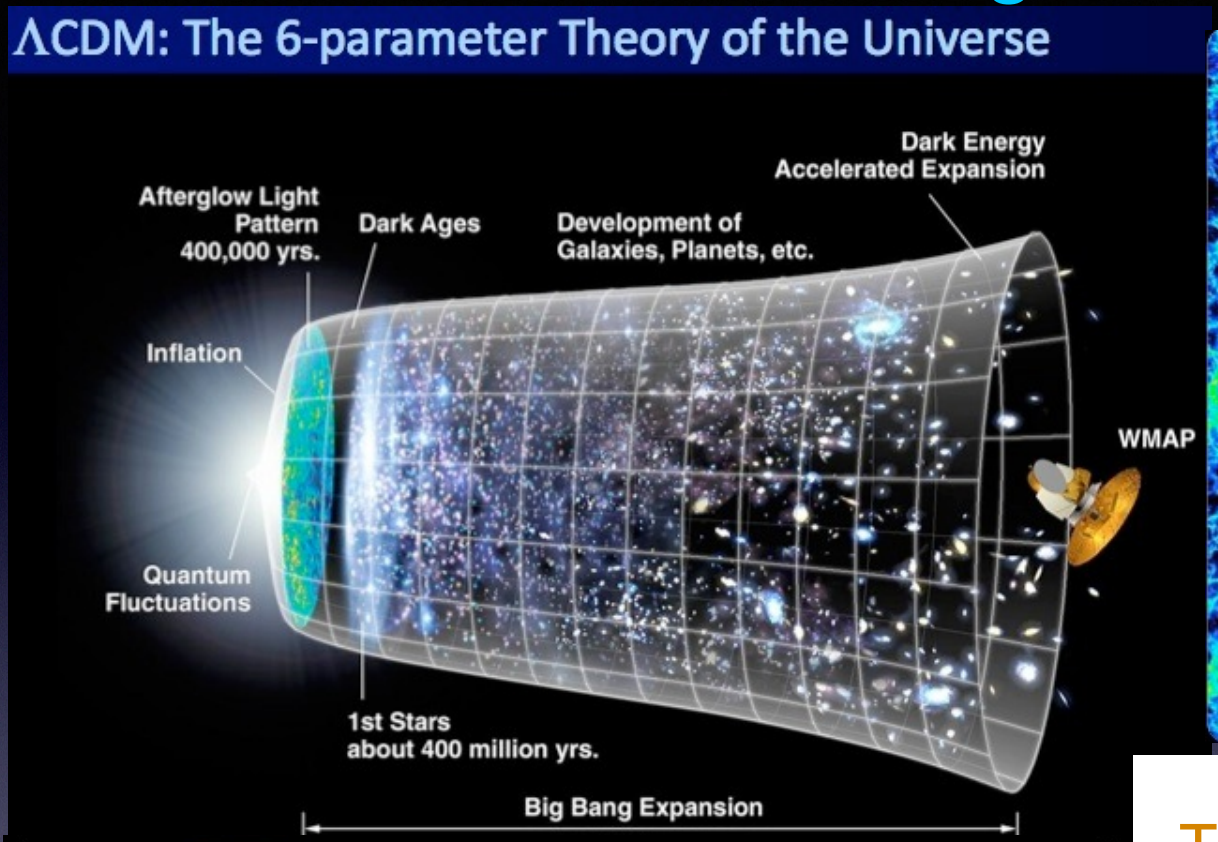
LSST in one sentence:

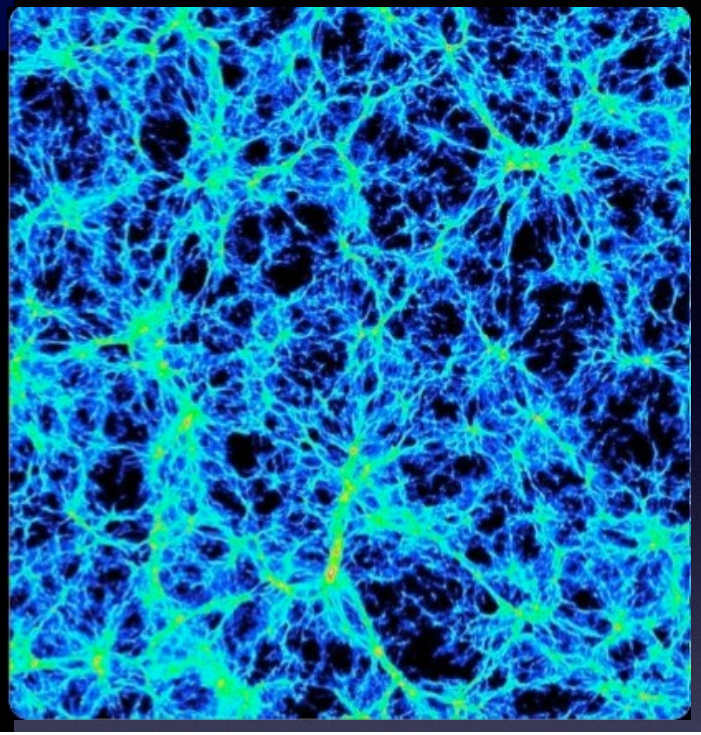
An optical/near-IR survey of half the sky in ugrizy bands to r~27.5 based on ~800 visits over a 10-year period:

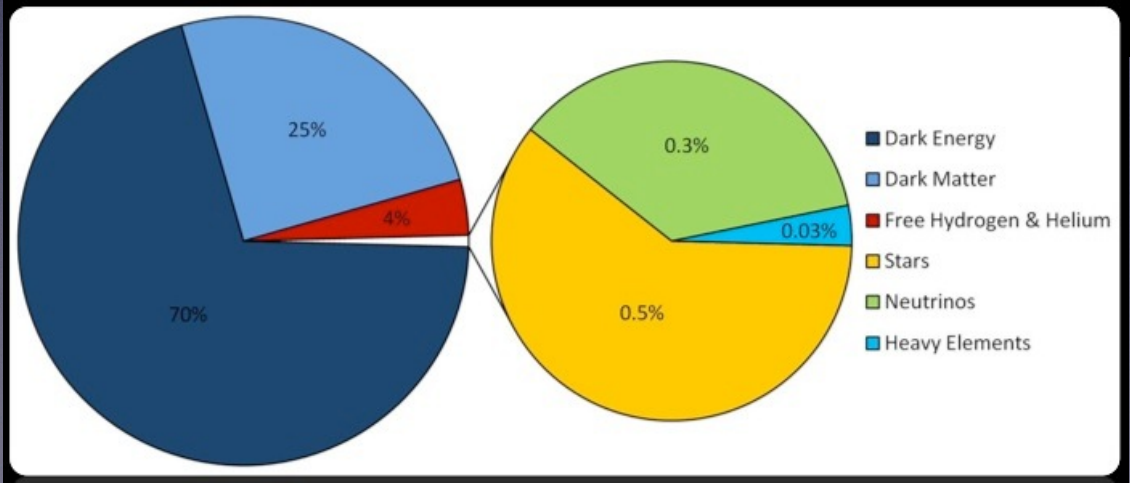
More information at www.lsst.org and arXiv:0805.2366

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

New Cosmological Puzzles







The modern cosmological models can explain all observations, but need to **postulate** dark matter and dark energy (though gravity model could be wrong, too)

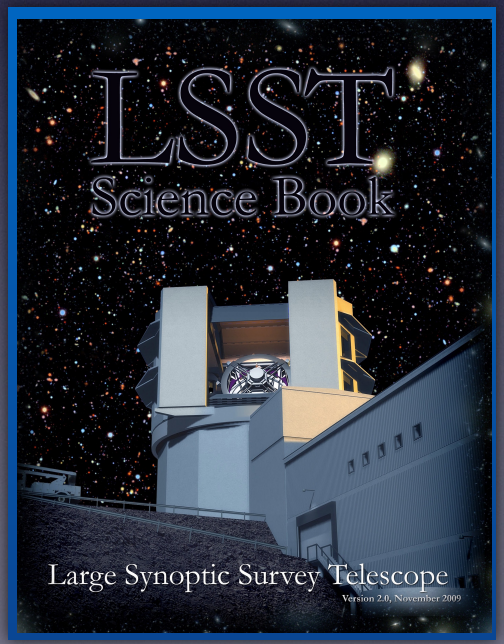
LSST Science Themes

- 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

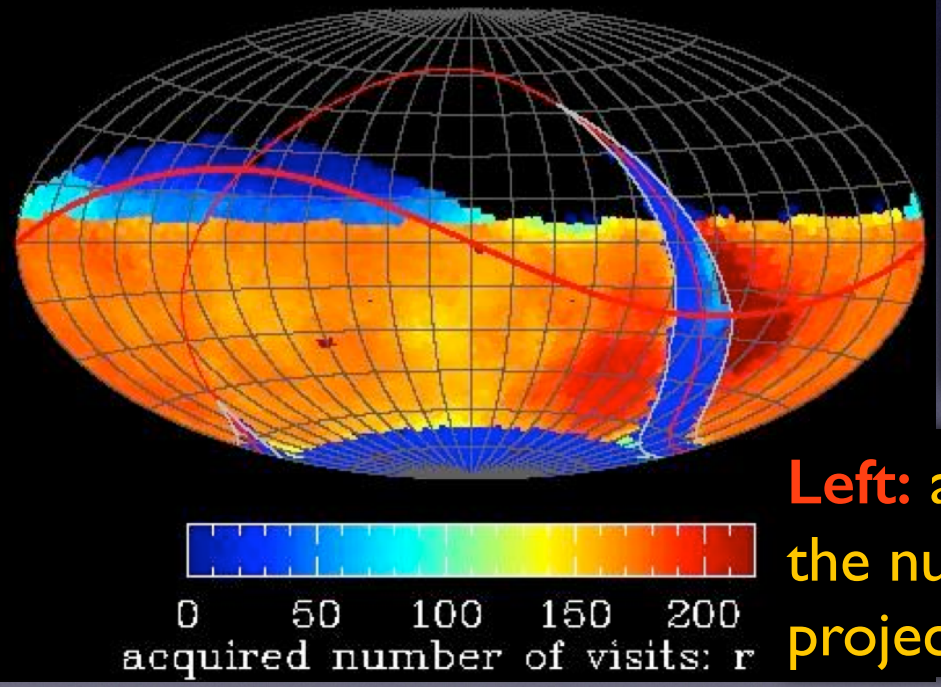
Summarizes LSST hardware, software, and observing plans, science enabled by LSST, and educational and outreach opportunities

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!**



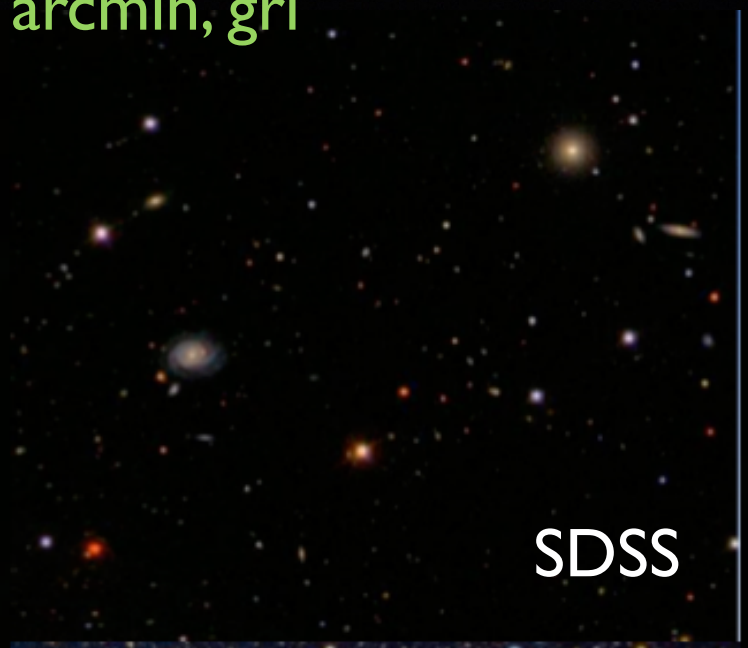
LSST in one sentence:
An optical/near-IR survey of half the sky in ugrizy bands to r~27.5 (36 nJy) based on 825 visits over a 10-year period: deep wide fast.

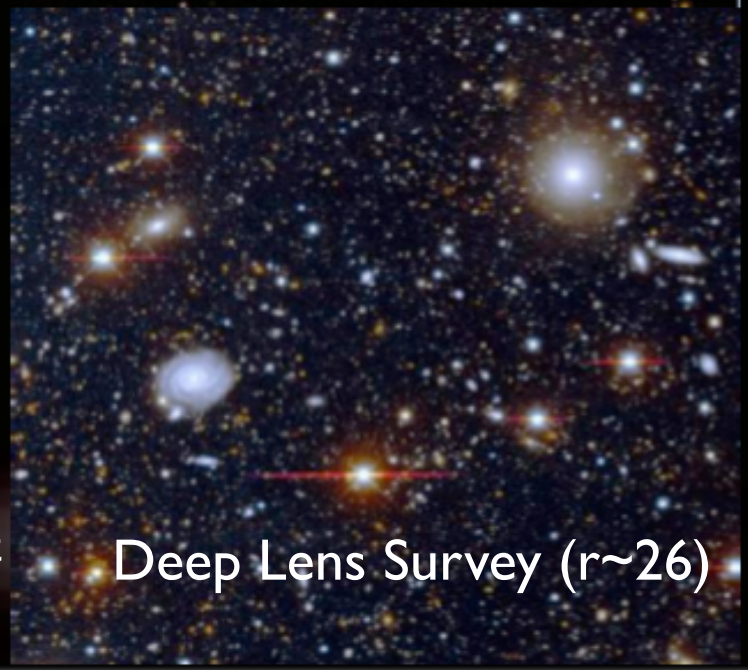
Left: a 10-year simulation of LSST survey: the number of visits in the r band (Aitoff projection of eq. coordinates)

SDSS vs. LSST comparison: LSST=d(SDSS)/dt, LSST=SuperSDSS

3x3 arcmin, gri

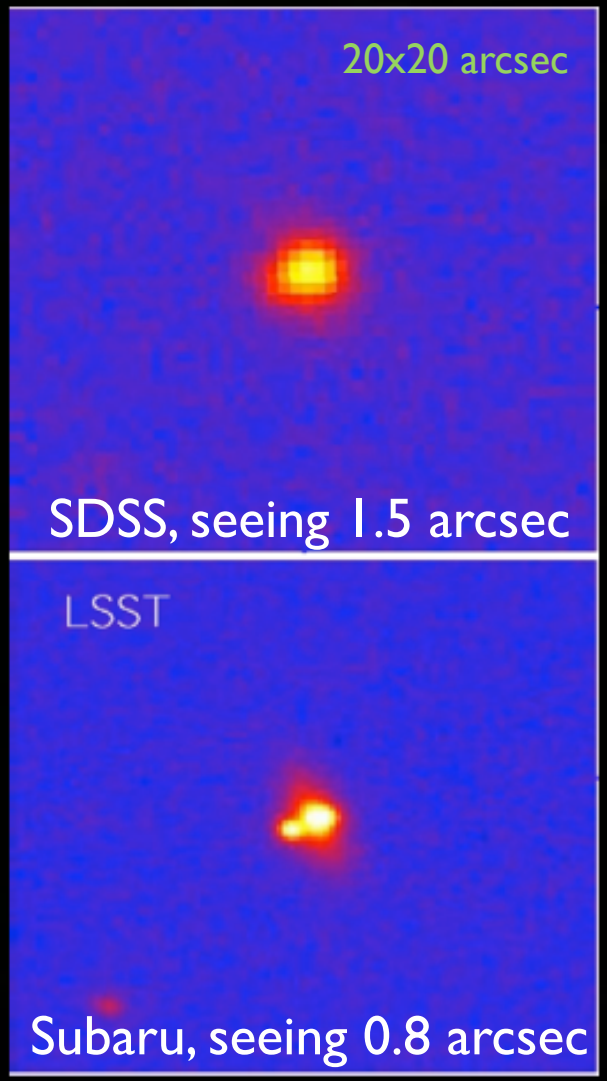
3 arcmin is 1/10 of the full Moon's diameter





(almost) like LSST depth (but tiny area)

20x20 arcsec; lensed SDSS quasar (SDSS J1332+0347, Morokuma et al. 2007)



Motivation for studying astro-statistics

- **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

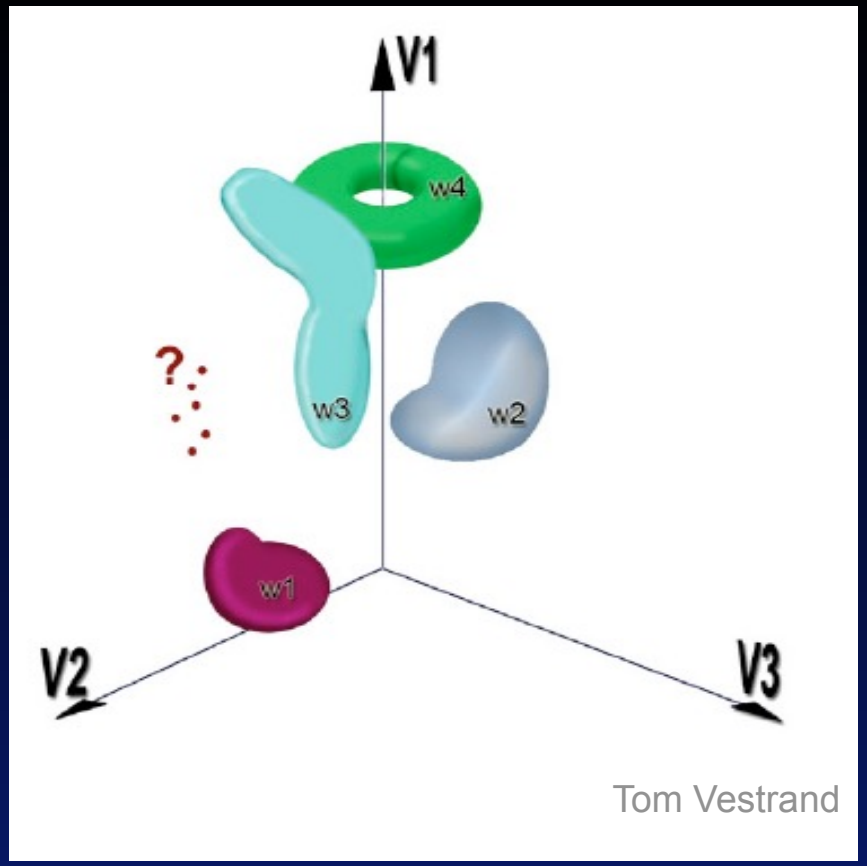
5) Time-series data (irregular sampling)

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.

- Characterize the known (clustering)
- Assign the new (classification)
- Discover the unknown (outlier detection)

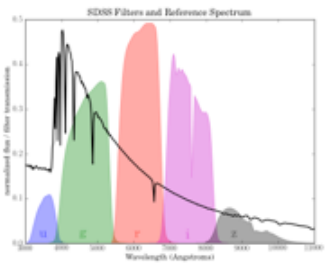


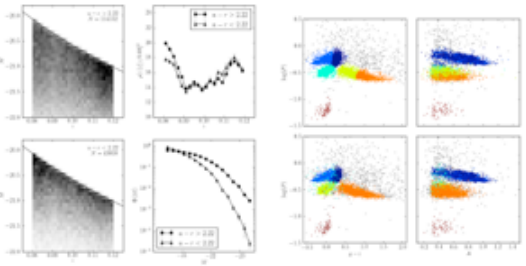
Benefits of very large data sets:

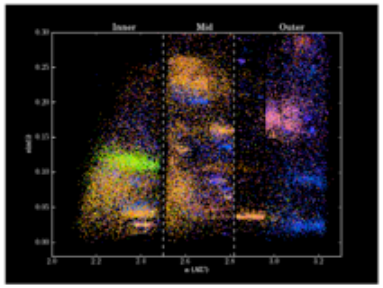
- best statistical analysis of "typical" events
- automated search for "rare" events

In this class, you will learn how to do all that.
Btw, it need not be an astronomical application!

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.

Downloads

- Released Versions: Python Package Index
- Bleeding-edge Source: github

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(pdf)`**.

User Guide

1. Introduction

- 1.1. Philosophy

Open source!
www.astroML.org

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

Scipy 2012 (15 minute talk)

Citing

If you use the software, please consider citing astroML.

Textbook Figures

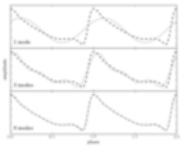
This section makes available the source code used to generate every figure in the book *Statistics, Data Mining, and Machine Learning in Astronomy*. Many of the figures are fairly self-explanatory, though some will be less so without the book as a reference. The table of contents of the book can be seen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

Figure Contents

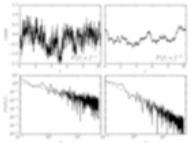
Each chapter links to a page with thumbnails of the figures from the chapter.

Chapter 10: Time Series Analysis

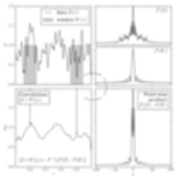
This chapter covers the analysis of both periodic and non-periodic time series, for both regularly and irregularly spaced data.



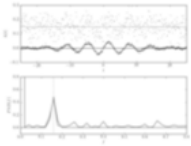
Fourier Reconstruction of RR-Lyrae Templates



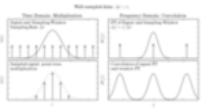
Generating Power-law Light Curves



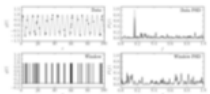
Plot a Diagram explaining a Convolution



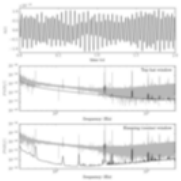
Fast Fourier Transform Example



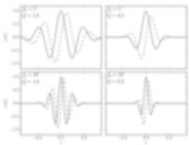
The effect of Sampling



The effect of Sampling



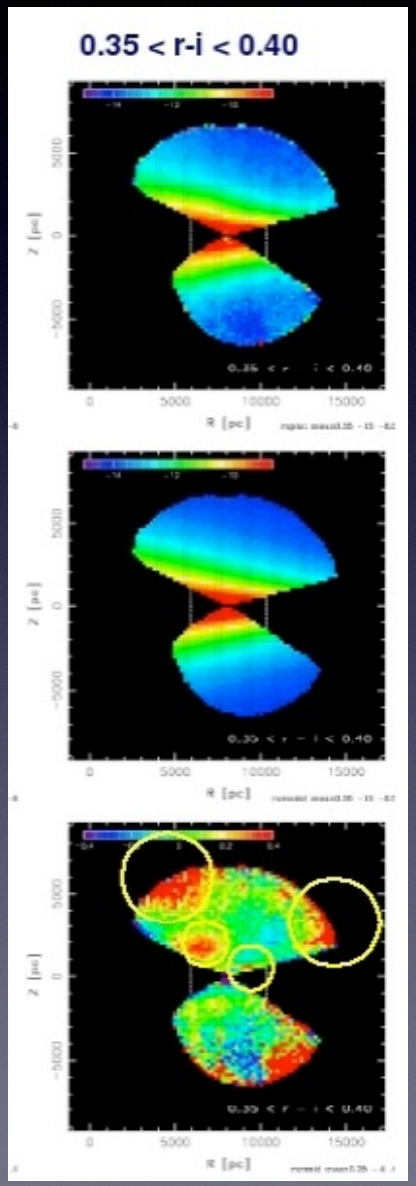
Plot the power spectrum of the LIGO big dog event



Examples of Wavelets

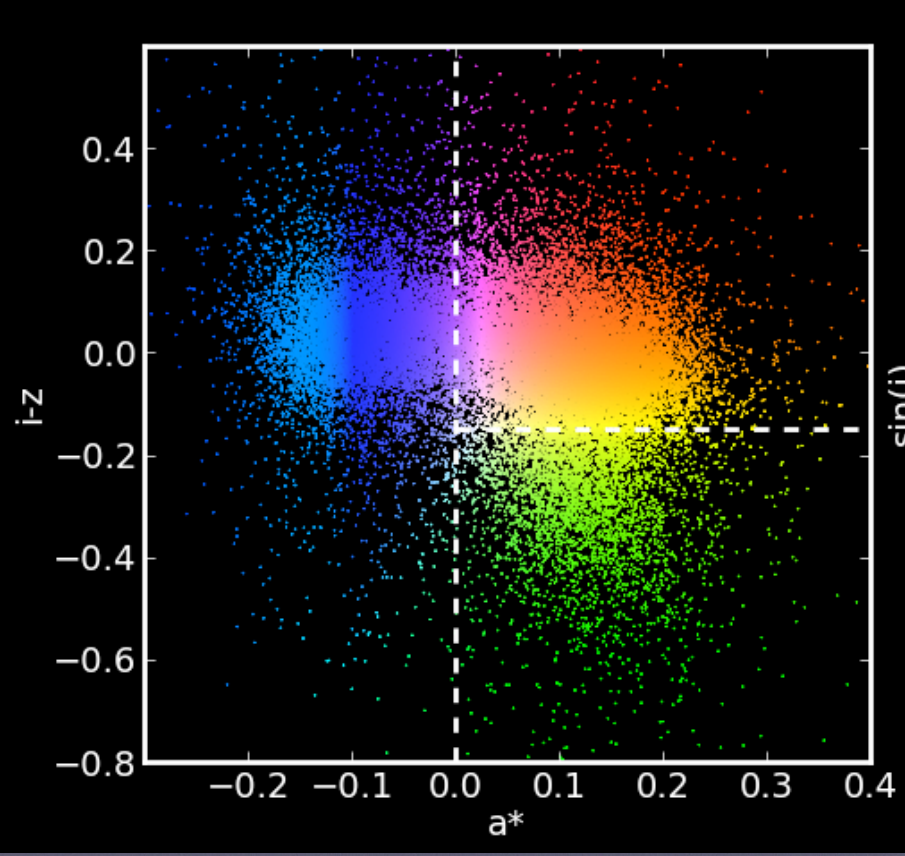
Visualization of multi-dimensional correlations and search for patterns

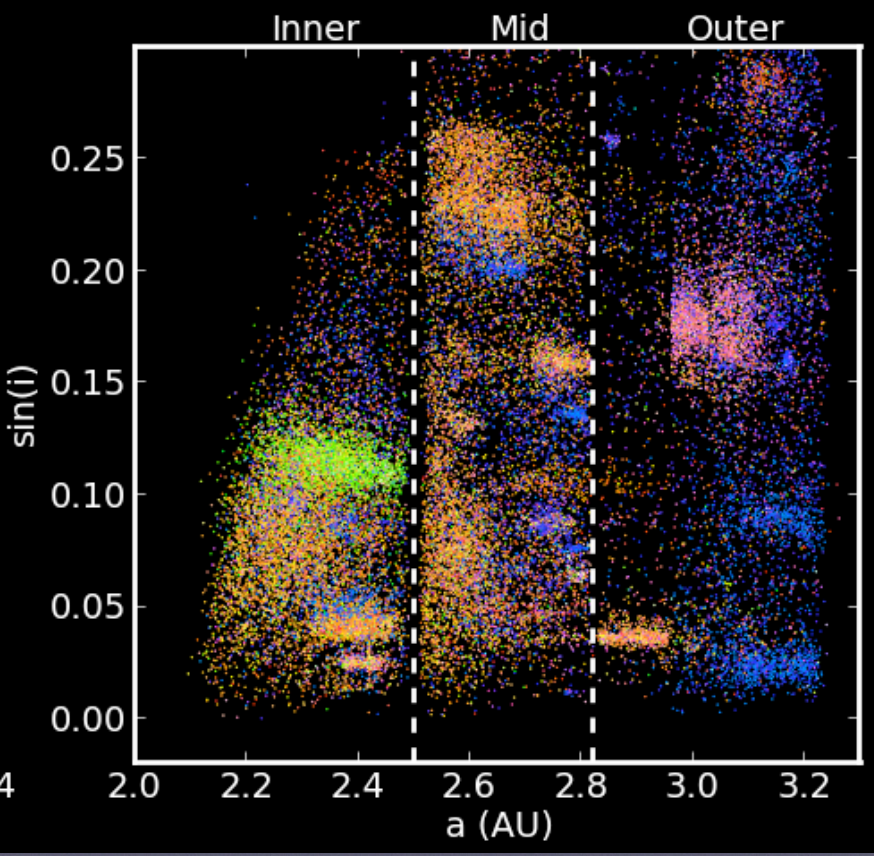
"The greatest value of a picture is when it forces us to notice what we never expected to see." (John Tukey, 1977)



- An example from Jurić et al. (2008): our Galaxy grew by cannibalizing nearby smaller neighbors
- Four-dimensional data: stellar color, apparent magnitude, position on the sky (RA, Dec)
- Step 1: derive three dimensional positions and bin the data in cylindrical coordinates
- Step 2: fit a (relatively complicated) model
- Step 3: subtract the best-fit model from data
- We will learn about all these steps in this class; today, let's focus on visualization.

Visualization of 4-dimensional correlations





- 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”)
- we will use this method extensively for data analysis and visualization, as part of the Class Project

AstroML

1) First install it:

Go to:
http://www.astroml.org/user_guide/installation.html

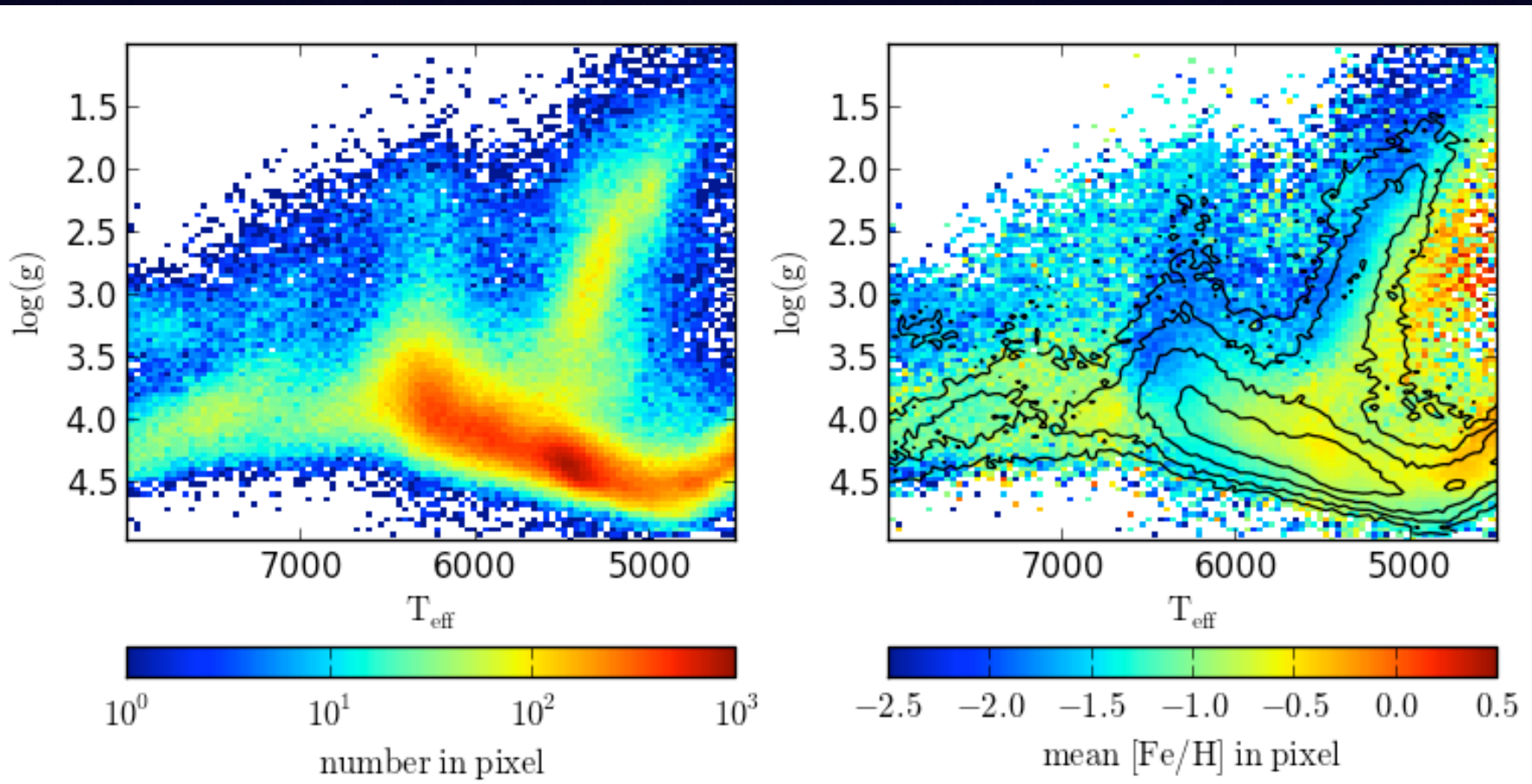
2) And then you need to test your installation:

Run:
http://www.astroml.org/book_figures/chapter1/
fig_SSPP_metallicity.html

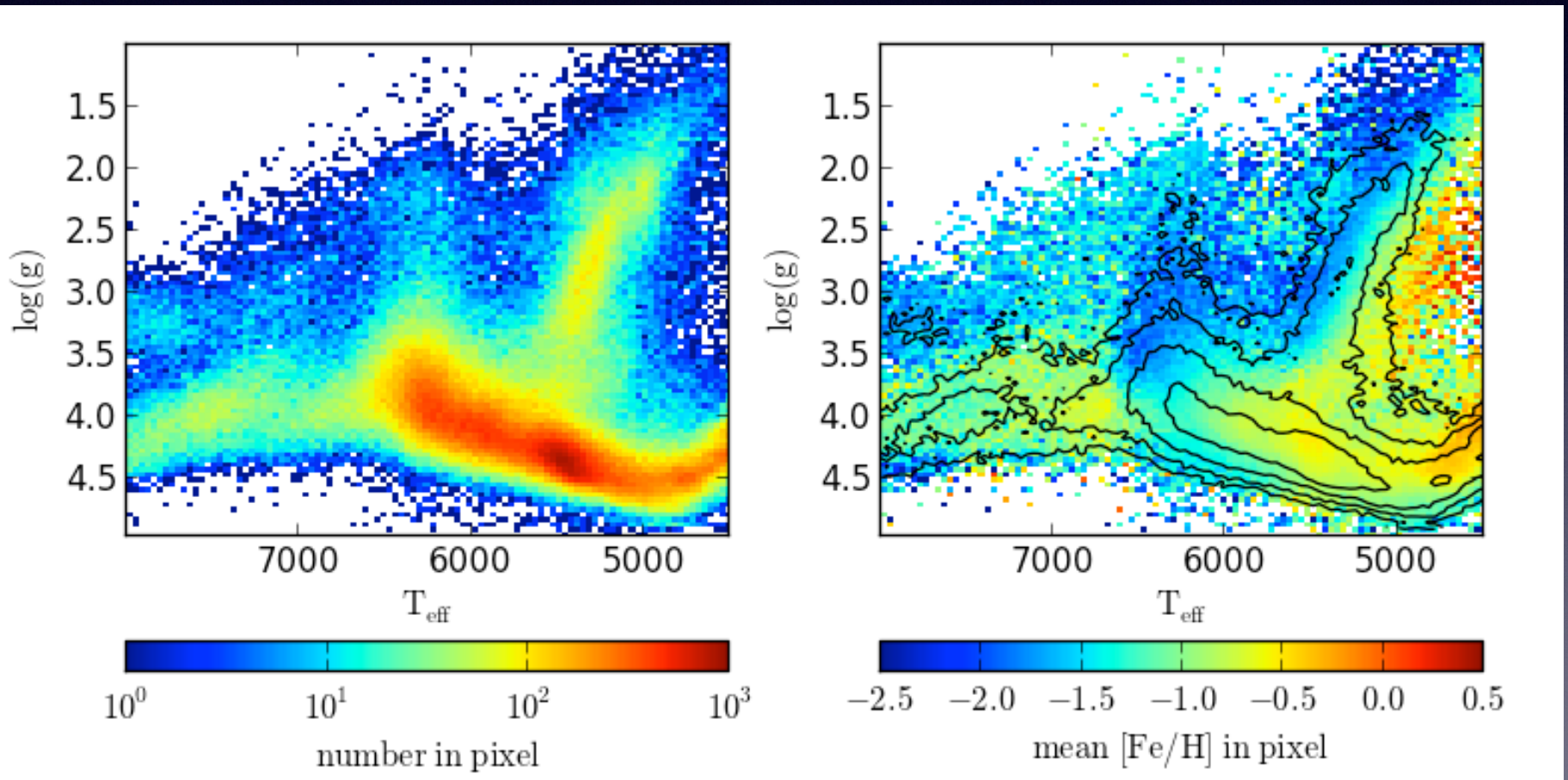
By the way, there is a very nice video about astroML by Jake at:

http://pyvideo.org/scipy-2013/opening-up-astronomy-with-python-and-astroml-sci.html

- **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



- Of course, the pixels don't have to be coded by the number of objects in it - eventually, we (as in you) will redo these plots from Sesar et al. and use proper motion for coding

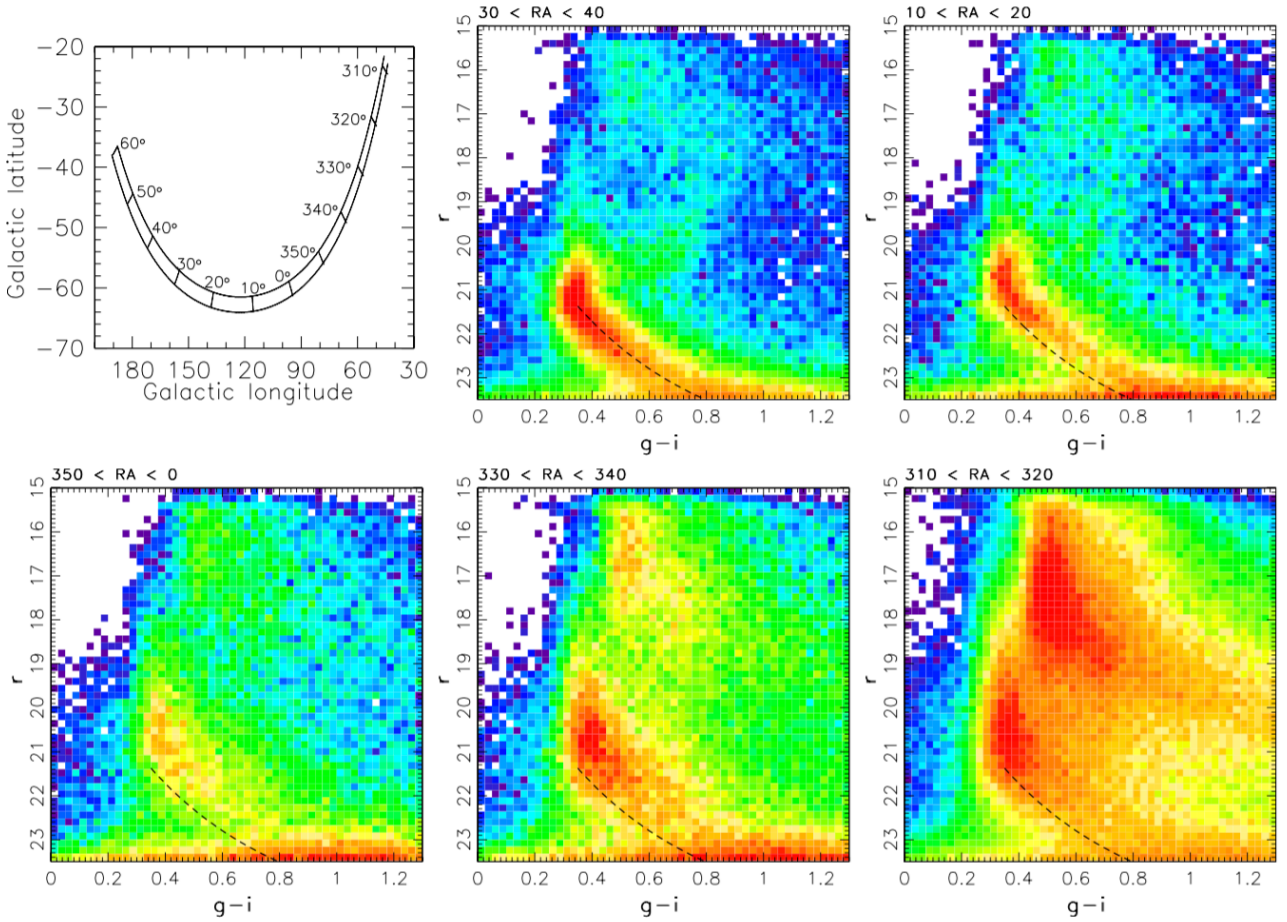


Figure 23. Top left panel shows the stripe 82 footprint in the galactic coordinates. The corresponding R.A. along the stripe is marked (decl. ∼ 0). The other five panels show the r vs. $g - i$ color–magnitude (Hess) diagrams for five 25 deg^2 large regions from stripe 82 selected by R.A. (the range is listed on the top of each panel). The top middle panel corresponds to a region intersecting the Sgr dSph tidal stream; note the MSTO at $g - i \sim 0.3$ (F-type dwarfs) and a well-defined subgiant branch. The MSTO at $g - i \sim 0.3$ indicates that the main-sequence stars in the Sgr trailing arm are at least 8 Gyr old. The counts are shown on a logarithmic scale, increasing from blue to yellow and red (with varying normalization to emphasize features). The dashed lines are added to guide the eye and to show the position of main-sequence stars with [Fe/H] $= -1.2$ at a distance of 30 kpc. Note the lack of stars with $g - i \sim 0.4$ and $r > 22$ in most panels. For the median halo metallicity of [Fe/H] $= -1.5$, $r = 22$ corresponds to 28.7 kpc. The feature at $g - i \sim 0.5$ and $r \sim 15$–18 is due to thick disk stars.

Homework (due Jan 12)

- From the website http://das.sdss.org/va/stripe_82_variability/SDSS_82_public/ download eight HLC*fits.gz files
- For each HLC file, and separately for stars and galaxies (use MEAN_OBJECT_TYPE > 5 to classify stars), make a 3-panel plot with one panel showing the Hess (counts) r vs. g-i diagram and another two using isoplets for counts and color scheme to show the median RA and Dec proper motions.
- Make a few other interesting and judiciously chosen plots!
- You need to submit your work as a jupyter notebook to directory …/uw-astr598-w18/homeworks using pull request
- You are encouraged to work as team(s), but the submitted notebook should briefly describe who did what