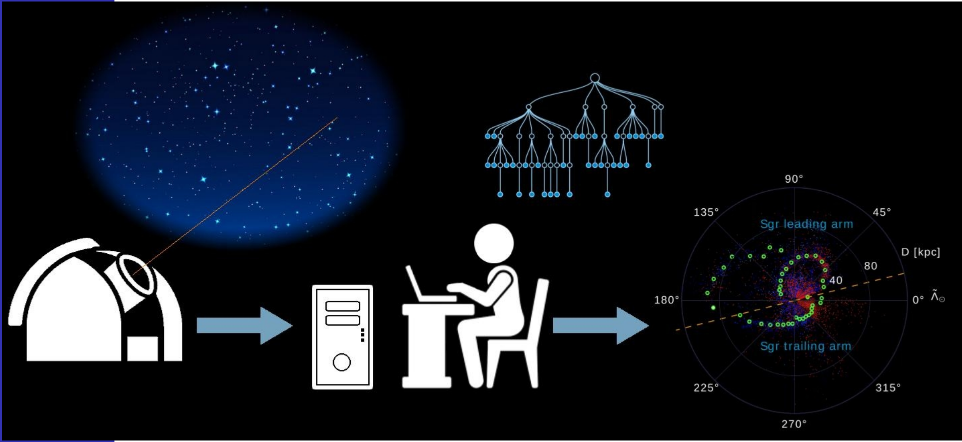
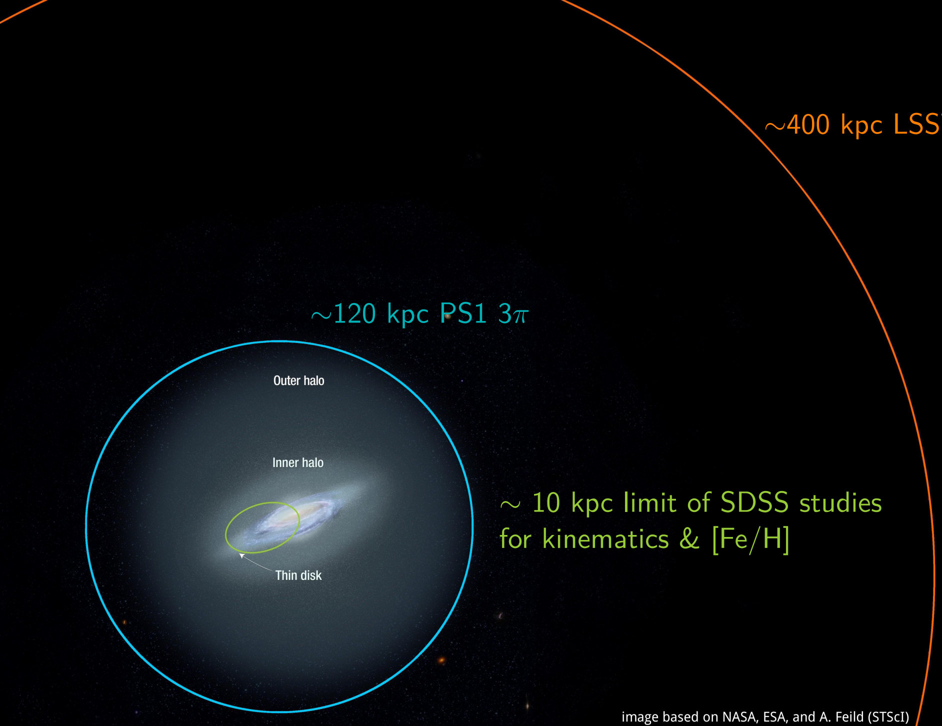
ASTR 3890 - Selected Topics: Data Science for Large Astronomical Surveys (Spring 2022)

Introduction (course, Python, github)

Dr. Nina Hernitschek
January 24, 2022

Motivation

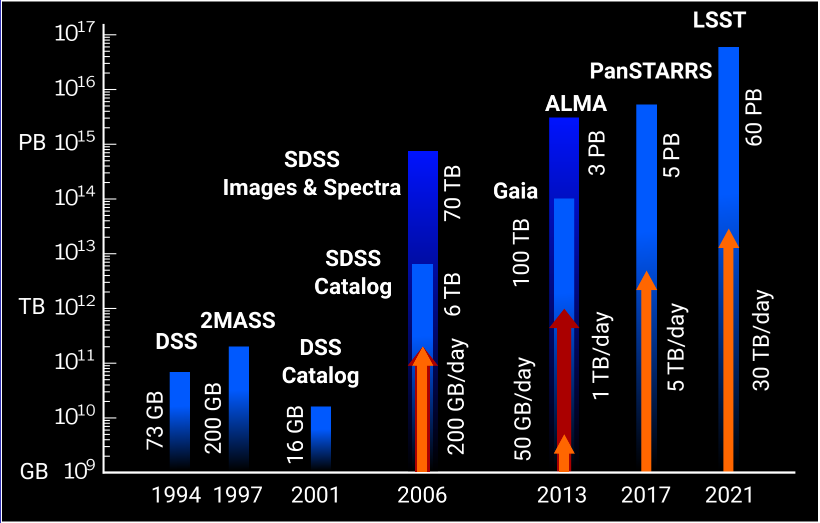

90°
45°
135°
Sgr leading arm
D [kpc]
80
40
180°
0°
$\tilde{\Lambda}_{\odot}$
Sgr trailing arm
225°
315°
270°



~400 kpc LSST
~120 kpc PS1 3π
Outer halo
Inner halo
~ 10 kpc limit of SDSS studies for kinematics & [Fe/H]
Thin disk
image based on NASA, ESA, and A. Feild (STScI)

Challenges in Data Handling

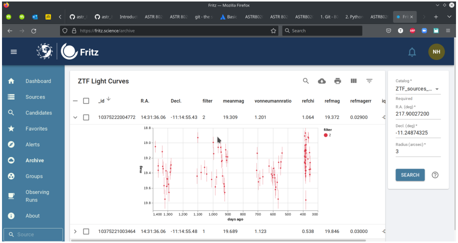
increasing data volume in astronomical surveys



what you will learn in this class

this course will prepare you for "doing science" with current and upcoming large astronomical surveys:

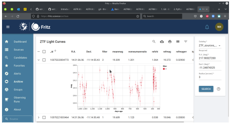
accessing astronomical survey data



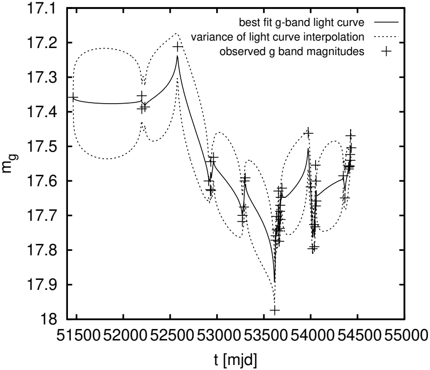
what you will learn in this class

this course will prepare you for “doing science” with current and upcoming large astronomical surveys:

accessing astronomical survey data



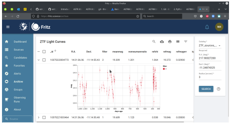
time series analysis



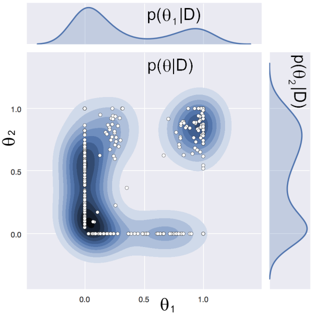
what you will learn in this class

this course will prepare you for “doing science” with current and upcoming large astronomical surveys:

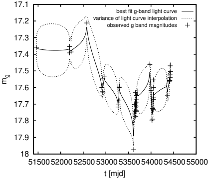
accessing astronomical survey data



statistical methods



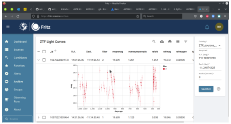
time series analysis



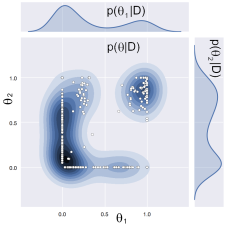
what you will learn in this class

this course will prepare you for “doing science” with current and upcoming large astronomical surveys:

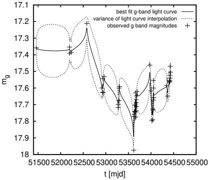
accessing astronomical survey data



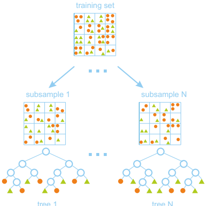
statistical methods



time series analysis



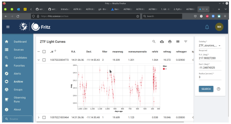
machine learning



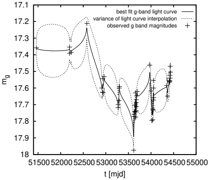
what you will learn in this class

this course will prepare you for “doing science” with current and upcoming large astronomical surveys:

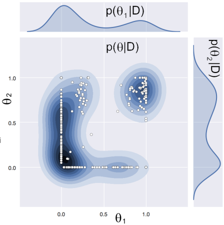
accessing astronomical survey data



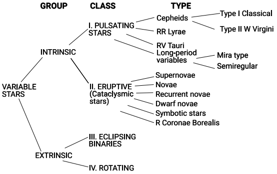
time series analysis



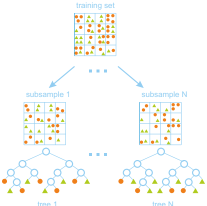
statistical methods



classification



machine learning



Lecture Schedule

Jan 24 Lecture 1: Introduction & Introduction to Python

Jan 31 Lecture 2: Astronomical Survey Data

Feb 7 Lecture 3: Introduction To Probability & Statistics: I

Feb 14 Lecture 4: Introduction To Probability & Statistics: II

Feb 21 Lecture 5: Classical/Frequentist Statistical Inference

Feb 28 Lecture 6: Bayesian Statistical Inference: I

Mar 5 - Mar 13 *Spring Break*

Mar 14 Lecture 7: Bayesian Statistical Inference: II

Mar 21 Lecture 8: Time Series Analysis: I

Mar 28 Lecture 9: Time Series Analysis: II

Apr 4 Lecture 10: Data Mining & Machine Learning: Intro to Scikit-Learn

Apr 11 Lecture 11: Dimensionality Reduction, Density Estimation & Clustering

Apr 18 Lecture 12: Classification: Introduction, Supervised Classification

Apr 25 Lecture 13: Unsupervised Classification

Grading

Participation credit will be assigned by submitting your completed copy of the lecture Jupyter notebook. The completed lecture notebook must be submitted by 11:00am Central Time the following Monday. Credit is given for making a reasonable attempt at all tasks in the Jupyter notebook.

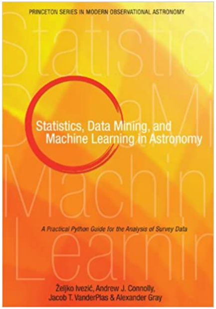
Final Take-Home Exam:
After the last session of the class I will assign a take-home exam that includes a series of statistics, coding and data analysis tasks that assess the course material.

Grading metric:
class participation and collaboration: 25%
homework assignments: 45%
final take-home exam: 30%
total: 100%

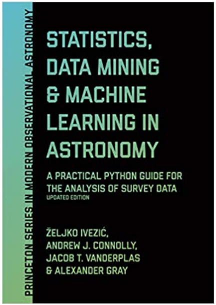
Textbook

Statistics, Data Mining and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data - Ž. Ivezić, A. J. Connolly, J. T. VanderPlas, A. Gray

The textbook is available from the Princeton University Press website as well as from Amazon. Used books can be found on Amazon from about $20. An older version (1st edition) will do.



or



Plan for Today

- introduction to the idea "big data", data science in astronomy
- quick intro to `github`
- a `Python3.x` mini tutorial

Challenges in Data Handling

decades ago: small astronomical surveys, catalogs in form of books

today: astronomy is largely determined by the available 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

Big Data is transforming how and which discoveries are made

Big Data

Laney et al. 2001: Big Data means that the data growth challenge is **three-dimensional**

Big Data is data with at least one big dimension:

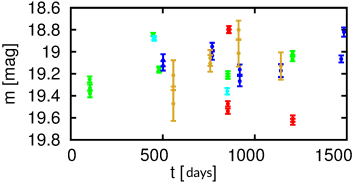
- volume: total amount of data
- velocity: bandwidth, response speed
- variety: number and size of individual assets

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

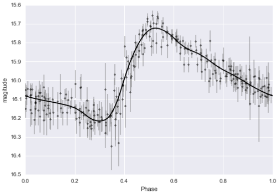
Big Data

one **example** for finding the *right* data :

Pan-STARRS1 3π survey with about 10^9 light curves like that:



goal: finding RR Lyrae* stars whose light curves look like that (if better sampled):



*less than 1 % of the light curves are expected to be from that type

Big Data: Typical Methods

Feature Extraction

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable derived values, the *features*.

typical features for astronomical time series data:

- statistical features: minimum, maximum, median...
- amplitude
- period
- astronomical: colors...
- features from fitting a function: template fitting, structure function...

typical features in image processing:

- edges
- shapes
- motion

Big Data: Typical Methods

Feature Extraction

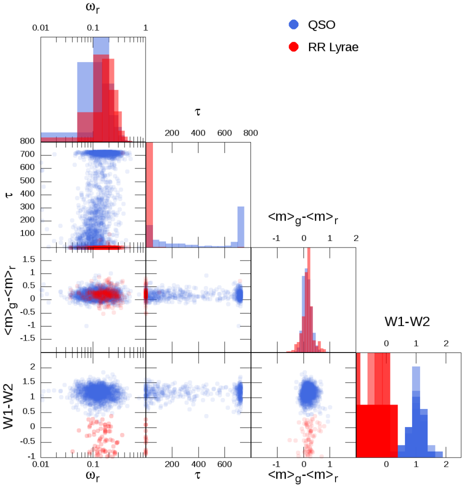
Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable derived values, the *features*.

Why is this useful?
reduce the number of data points for processing without losing important or relevant information – extracting *relevant* information

makes it easier for algorithms (machine learning) and humans (scientists) using and interpreting the data

Big Data: Typical Methods

one **example** for feature extraction:
features extracted from the Pan-STARRS1 3π dataset:



Statistical Data Analysis

Data-driven methods like statistical methods can reliable **quantify information** embedded in scientific data **without the biases of physical 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!

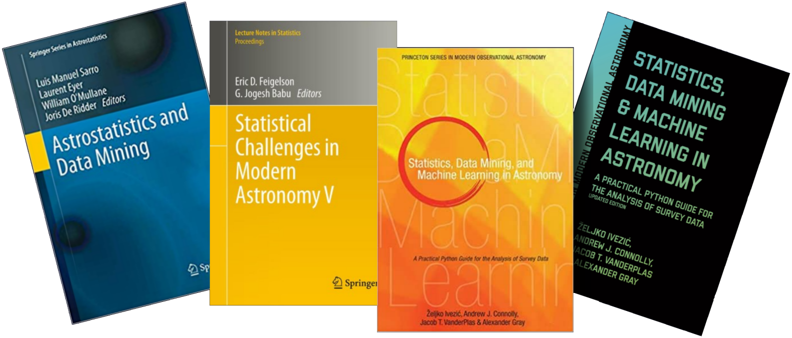
(Astro-)Statistics

19th and 20th centuries: statistics moved towards human sciences (demography, economics, psychology, medicine)

"traditional" methods, literature, software often not applicable to astronomy/astrophysics (take a look at a typical bookstore... or mention you are doing astrostatistics - surprise!)

(Astro-)Statistics

beginning of 21st century:
increasing size of data sets forces scientists to think about statistical methods



(Astro-)Statistics

a lot is possible:

galaxy clustering	spatial point processes, clustering
galaxy morphology	regression, mixture models
weak lensing morphology	geostatistics, density estimation
strong lensing morphology	shape statistics
faint source detection	false discovery rate
variable source preclassification	structure functions + classifier

(Astro-)Statistics

a lot is possible:

galaxy clustering		spatial point processes, clustering
galaxy morphology		regression, mixture models
weak lensing morphology	⇔	geostatistics, density estimation
strong lensing morphology		shape statistics
faint source detection		false discovery rate
variable source preclassification		structure functions + classifier

$\Rightarrow$ **fitting models**

Break & Questions

git

`git` is a **version control system**. Specifically, `git` is a distributed version control system: each user actually clones the entire **repository** locally.

Think of a repository as a collaborative directory to which multiple people have access, and that it tracks how changes are made (who, when, what).

`git` vs. `github`: `git` is a version control system. `github` is a cloud-based hosting service that lets you manage Git repositories.

git

In this class, all work will be submited through `git`:

- completed copy of the lecture Jupyter notebook
- homework

caution: everything submitted will be available for everyone in the class!
but: `diff` command makes it obvious to me if solutions were copied

git

let's get started:
[here insert link to first lecture notebook]

Python

Python is an open-source, object-orientated high-level scripting language that is useful for manipulating data. It is widely used in science, especially physics and astronomy.

As with any programming language, Python has some undesirable features, such as some relatively slow processes. But Python can be used to wrap faster code such as C. (I use this often.)

advantages: Python it has a large number of existing packages for manipulating large datasets including data access, machine learning and plotting.

Python

we continue with `lecture_1.ipynb` from the `github` repository