MLD 2019- Problem set 1

Installing programs

The tasks below require internet connection and a web browser. For data analysis we will use python and the astroML framework as well as a few other python packages (`numpy`, `matplotlib`, `astrolib`, `scipy` etc) so make sure this is installed on your computer.

If you want to explore R instead of python, that is ok but there will be no support for this in the course. To start with this go first to the main web page: http://www.r-project.org/ and go to the 'CRAN' link in the left column (under `Downloads, Packages`). This will bring up the possible mirrors and led me to a page in Utrecht:

```
http://cran-mirror.cs.uu.nl/
```

and I downloaded my version from there (for Linux it might be preferable to use your packaging system to install R - as long as you have version > 2 it is fine).

For the database we will use `sqlite3` - this is distributed with the standard Python library so should always be available.

Finally we will need `git` - which is normally available on laptops running Linux or Mac OS and for Windows you can install Git For Windows (https://gitforwindows.org/). A general overview of the installation can be found on Atlassian's pages: https://www.atlassian.com/git/tutorials/install-git

General instructions

This problem set has a number of possible tasks. There is no expectation that you do all of them, nor that you do them sequentially. They are instead a list of possibilities to choose from depending on your interests/needs. A guideline to them is as follows:

If you are not very familiar with SQL: Do problem 3 & 4.

If you have never used git: Do problem 1.

If you would like to have some experience in creating tables using SQL: problem 2 & 5

If you are doing anything, do at least problem 6!

If you want to explore different simple regression methods, have a look at problem 7.

Problem 8-10 are good for getting some more experience with SQL.

Problem 11-12 introduce kernel density estimators and I recommend doing those although 11 can not be done in full before next lecture since I did not have time to cover cross-validation.

After these problems there are more things to try out - but those are problems that I expect most of you do not need to look at and I haven't checked that they are all doable!

1. Trying out git

1. Create a working directory for the course. Check out the MLD2019 repository, or update it if you have checked it out before. This is located at https://github.com/jbrinchmann/MLD2019.
2. Find out what the last commit to this repository was.
3. Create a GitHub account if you do not yet have one. Go to https://github.com/ to create an account.

That's it for now - but you will use it occasionally in the following.

2. Create the Stars and Observations tables

If you did not create the Stars and Observations tables yet, follow the lecture notes (p 163 onwards in the notes of Lecture 1), to create these. The scripts and relevant files are available on GitHub (ProblemSets/MakeTables) - including the database itself, but I would suggest creating a new one

After creating the database, try this out on the command line by typing:

```
> sqlite3 MLD2019.db
```

for me this gives:

```
> sqlite3 MLD2019.db
SQLite version 3.26.0 2018-12-01 12:34:55
Enter ".help" for usage hints.
sqlite> .tables
Observations  Stars
```

which shows us that the two tables have been created.

3. Querying using Python.

I assume that you have put the Stars and Observations tables in a file called MLD2019.db. Substitute this whenever relevant below:

1. Start up ipython: `> ipython`
2. Now, we need to import the sqlite3 package:

```
In [1]: import sqlite3 as lite
```

3. We can now use this to connect to the database - this is done as follows:

```
In [2]: con = lite.connect('MLD2019.db')
```

4. With a connection one can carry out operations on the database. In particular it is possible to execute SQL queries. This is done using con.execute and this returns an iterator that lets you iterate over the rows returned. So we have

```
In [3]: rows = con.execute("Select ra, decl FROM Stars")
In [4]: for row in rows:
   ...:     print "Ra={0}  Dec={1}".format(row[0], row[1])
   ...:
Ra=198.8475  Dec=10.5034722
Ra=198.5654167  Dec=11.0231944
Ra=198.9370833  Dec=9.9168889
Ra=199.2516667  Dec=10.3486944
```

Check that you get the same (or similar) results.

4. Completing the lecture

In the lecture notes there are two questions that did not have an SQL query specified:

4. Where is the FITS image stored for star S5?

5. Give me a list of all stars observed with the same FieldID

Construct the SQL queries necessary to answer those two questions. Check that it works - ideally using python.

5. Creating simple tables

Create the following two tables in sqlite and run the queries below. Check that your results make sense! Use python to populate the tables to get some experience with that.

Table name: **MagTable**

Name	Ra	Dec	B	R
VO-001	12:34:04.2	-00:00:23.4	15.4	13.5
VO-002	12:15:00.0	-14:23:15	15.9	13.6
VO-003	11:55:43.1	-02:34:17.2	17.2	16.8
VO-004	11:32:42.1	-00:01:17.3	16.5	14.3

Table name: **PhysTable**

Name	T_{eff}	FeH
VO-001	4501 K	0.13
VO-002	5321 K	-0.53
VO-003	6600 K	-0.32

---

a) Find the Ra & Dec of all objects with B > 16.

b) Output B, R, Teff and FeH for all stars (is this question well-defined?).

c) Output the same for all objects with FeH > 0

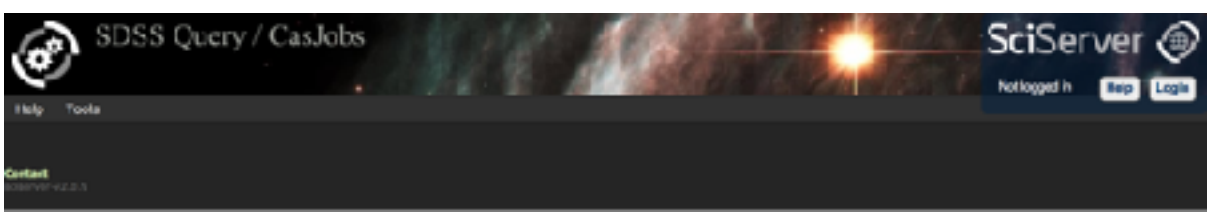
d) Create a table with the B-R colour

6. Running SQL queries on the SDSS databases - CasJobs

CasJobs is the standard interface to interact with the SDSS database and it is also the default database to use on the MAST archive (which contains HST, Kepler, TESS, GALEX and other mission). To use this you need to get a CasJobs account. I will start with using the SDSS version for concreteness but the skill you learn here can be used on the MAST archive as well and I will return to this later.

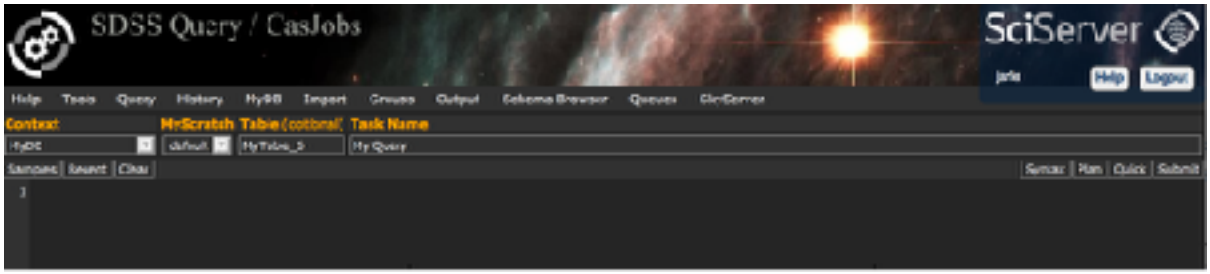
This is not a terribly challenging task but it is very useful to be familiar with CasJobs - and at first glance it can be a bit bewildering. Note that the schema (layout) for the SDSS database is available within casjobs if you use the skyserver one (see below), but it can also be seen at http://cas.sdss.org/dr7/en (for DR7) and http://skyserver-.sdss3.org/dr14/en/ (for DR14 - the latest). The two releases have slightly different sets of tables and column names.

To set up an account or later log in you should use the SDSS-III site: http://skyserver.sdss.org/casjobs/login.aspx
There is also an older and widely used site at http://casjobs.sdss.org/CasJobs but we will use this only as a



backup. You will actually discover that this is part of the SciServer programme but that is not important for us here

a) On the page - use the Login button on top to create your account. You can then create your first query - you should have a page that looks like:



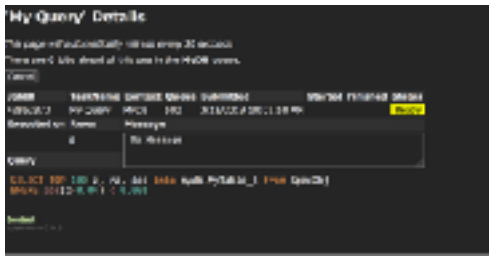
The best now is just to make a very simple one - choose the first 100 objects in SpecObj table (actually a view) that is within 0.001 of z=0.04:

```
SELECT TOP 100 z, ra, dec
FROM SpecObj
WHERE abs(z-0.04) < 0.001
```

Click on 'check syntax' to see that you did not make a typo.

(If you prefer you can of course run whatever query you want but make sure it does not take too much time and does not return too many data points).

Click 'submit' to submit the query. It should look like this:



This is a queue of queries and it does not end immediately You can reload the page to see what is happening. What you will find is that this will fail!

Why does this fail? It fails because on the search page there is a menu called 'Context' which says what Database you want to search within, and the default is your private database which is called MyDB. What you actually want is DR15 (data release 15) which is the latest

So change the context and run the query again (in this case you could also click 'Quick' to run it quickly but then the rest of the problem set is not possible to do).

b) After the query has finished - or while it is running, click on the History link to be informed on its progress. This is a very convenient functionality btw - when you come back to CasJobs after some time away!

c) Your search has ended up in MyDB - click the link for this and verify that there is a table there - click on the table name to get up an interface to look at it. Explore this and at the end click Download to download the finished table in your preferred format. If you download a CSV or FITS file you might want to use TOPCAT to explore the results - or Python/R/IDL etc if you prefer that.

---

7. Linear regression

In the git repository under Problemsets, you will find a Jupyter notebook named `'Exploring Regression for problemset.ipynb'`. There is also a PDF version in case you prefer not to use `jupyter`. Go through this and answer the three problems.

8. Returning to the SDSS - Finding stars

Use CasJobs and find the first 100 stars with psfMag_r in each of the ranges 10-11, 11-12, 12-13, 14-15, 15-16 and output psfMag_r and psfMag_g for these stars. Save the result in one or more tables for later use.

Useful: SQL has a BETWEEN option that can be used in WHERE statements: WHERE zBETWEEN 0.2 AND 0.5 works the way you would expect it to.

9. Aggregating functions in SQL

SQL has the concept of aggregating functions. These are functions like SUM or COUNT, that sum or count a particular column. See pages 147-160 in the lecture notes and try these out in the CasJobs interface.

10. And an extra challenge - the fraction of red galaxies with redshift.

Write and execute an SQL statement to get the fraction of galaxies that have u-g>2 as a function of redshift. You can use the SpecPhoto view in CasJobs and should create 50 bins between z=0.001 and 0.5

Useful to know: To convert an integer to a float, you can write `cast(number AS float)` anywhere where you would use `number` in the SQL statement.

11. Determine an optimal bandwidth

In this problem set you will use a sample of galactic black hole masses from the literature. You will use Kernel Density Estimation to determine the distribution of their masses and for that you need to determine the best band-width to use.

The file to use is available in the git repository as `Datafiles/joint-bh-mass-table.csv`. A csv file can be easily read in using e.g.

```
from astropy.table import Table
t = Table().read(fname)
```

but if you prefer other approaches, feel free to do so.

a) Select the best band-width for kernel estimation by linear search. Try band-widths between 1 and 7 solar masses. What is the best band-width? Is this a useful estimate?

b) **This you can not do until after Lecture 3 without some individual study:** Use cross-validation to find the optimal bandwidth. I recommend using the KFold routine in `sklearn.model_selection`.

12. The likelihood of gravitational wave neutron stars

The file `Datasets/pulsar_masses.vot` contains masses of pulsars from Özel & Freire 2016, Annual Reviews of Astronomy and Astrophysics, taken from the associated web page.

On October 15, 2017, the LIGO/Virgo consortium and host of ground- and space-based observatories announced the discovery of an optical counter-part to the gravitational wave source GW170817. The two neutron stars that are thought have merged to produce the gravitational wave signal, have estimated masses of $M_1 = 1.81\,\mathrm{M}_\odot$ and $M_2 = 1.11\,\mathrm{M}_\odot$.

a) Use the pulsar mass catalogue to estimate the likelihood of finding a neutron star with $M > 1.8\,\mathrm{M}_{\odot}$.

b) Actually the mass estimates are ranges, and the masses were $M_1 \in [1.36,2.26]$ and $M_2 \in [0.86,1.36]$. What are the likelihoods of those mass ranges and the likelihood of the binary?

c) Simulate the next 5 detections of merging pulsars with LIGO+Virgo. What is your prediction for the average mass neutron star they would detect? [hint: to draw N from a `KernelDensity` object, you can use the `.sample(N)` function]

d) There are many assumptions made that I did not spell out - make a list of some that you worry about.

If you have little experience with databases in astronomy, or with topcat, the following pages might be useful but these are old problems that are not guaranteed to work!

The following pages go through some simple tasks to test out some commonly used databases in astronomy as well as some `topcat` exercise. Since many of these databases and topcat will be familiar to you, this is not part of a standard problem set, but if you want a refresher this is definitely a good idea to go through:

Standard databases in astronomy

Aims: Get familiar with the a few important data bases in astronomy and explore their capabilities. This set of problems can be done at your leisure.

These are:

- **Vizier** [http://vizier.u-strasbg.fr/cgi-bin/VizieR], an interface to a large number of astronomical data bases.
- **Simbad** [http://simbad.u-strasbg.fr/] a system to find information about astronomical sources.
- **NED** [http://nedwww.ipac.caltech.edu/], a comprehensive collection of data known for a specified object.
- **NASA ADS** [http://cdsads.u-strasbg.fr/ads_abstracts.html], where you can find astronomical literature.
- **MAST** [http://archive.stsci.edu/], a multi-mission archive at Space Telescope Institute.
- **IPAC** [http://www.ipac.caltech.edu/], an online interface to a large number of surveys, including several well away from the infrared the centre is named after.
- **IRSA** [http://irsa.ipac.caltech.edu/] is the IR focused subset of IPAC but it does in fact contain a lot more.
- The **ESO archive** [http://archive.eso.org], is the foremost archive for European astronomy.
- **SDSS** [http://www.sdss.org/dr12/], the largest and most complex web site for an existing survey. (we will not go in depth here - that is for another lecture).
- **Atomic Line List** [http://www.pa.uky.edu/~peter/atomic/] is a useful resource to find information about lines in a particular region of a spectrum.
- **NIST Atomic Spectra Database** [http://physics.nist.gov/PhysRefData/ASD/index.html] is another great resource for atomic transitions.

Here you will find first a set of minor tasks, one or two for each of the databases. These are small - the aim here is to bring everyone up to speed - some of you have extensive experience with these databases, some do not. This first lecture is to ensure that everyone sing from the same sheet.

A couple of astronomical reminders

I will use a few standard astronomical terms in the following. It is therefore useful to refresh your memory already here:

Redshift - the distance to galaxies is often quoted by giving the observed redshift of their spectrum and this is normally denoted by z. This can then be related to the luminosity distance for instance. However it is not a very reliable indicator of distance for very nearby galaxies and for these it is often useful to construct other, redshift independent, distance measures. Direct observations of Cepheids and use of their period-luminosity relation is one of these, but other techniques also exist. The details are not important. It is also useful to note that redshift often is quoted in terms of velocity, then it is defined as $z = v/c$ - as long as the velocities are much smaller than c, or $z \ll 1$. In practice this is only used for nearby galaxies with $z<0.01$ or so.

Forbidden lines - In the formation of emission lines, the most physically natural distinction is between re-combination lines (such as the Balmer series of Hydrogen) and collisionally excited lines such as [O III]5007Å. The latter are usually also forbidden transitions and proceed as magnetic dipole (often denoted M2), electric quadrupole (often denoted E2) or other multipole transitions. Forbidden lines are usually denoted by square brackets, so that [O III] is a forbidden transition of doubly ionized oxygen whilst O III is a permitted transition of doubly ionized oxygen.

Tasks

To get up to speed, we will do a few simple tasks - these are basic and it is important that you are able to each one of these without too much fuss:

1. Go to SIMBAD and search for 47 Tuc. This is a well-known Globular Cluster - check that you understand how to find this information as well. What information is available for this object through SIMBAD? Is it in the northern or southern hemisphere?
2. Use Vizier to find catalogues of observations of 47 Tuc. There are subtly different ways to do this - one is to indicate it in the 'Target Name' box - and then use the 'Find Data' button. How does this contrast with using the 'Find Catalogues' button? What about entering 47 Tuc in the Direct access box at top?
3. Use NED to find information about NGC 300 - this is a nearby galaxy which has been repeatedly studied over the years. What is the distance to this galaxy (use redshift independent estimates if possible - see above for a refresher on redshifts if necessary)? What is the quoted uncertainty on this distance and how does the distance compare to that inferred from the redshift? Find an image of the galaxy from GALEX through NED - what other data sources can you get images from?
4. Go to the SDSS home page and check where SDSS has observed, compare the area covered by imaging and those for imaging ("Sky Coverage"). Is SDSS a good survey for those that are interested in sources in the Milky Way disk?
5. VV Cephei is one of the largest stars known - it would engulf Jupiter and almost Saturn if placed in our Solar System. Go to ADS and find articles about this object - after you find the full list, find out which one has the most citations (hint: it is very easy, do not click on individual articles.). How many citations does it have? Are all of these in refereed journals?
6. Staying in ADS - sometimes you want articles where you know the first author. This is actually easy: Type '^' before the name. Try this, how many first-author papers do you find for Andy Fabian? And for Jan Oort?
7. The Annual Reviews of Astronomy & Astrophysics can be a gold-mine for when you start research in astronomy. Use ADS to find the titles of the articles in ARA&A in 2004. Hint: Look down the page for where you can choose bibliographical sources, the shorthand for Annual Reviews is ARA&A.
8. Go to MAST and search for HST observations of HH 47. This Herbig Haro object is a high-ionisation region caused by a jet from a young stellar object.You can download a FITS file for the object if you wish, but for a more satisfactory view do a Google Image search for HH 47.
9. Go to IRSA. This contains a large amount of data, you will explore it a bit more later - note for now that a lot of the navigation choices are given on the left. But for now go to The SINGS public survey - this is a survey of nearby galaxies in the thermal IR using the Spitzer space telescope. Try to find a 24μm image of M 51 in JPEG format. [Tip: The easiest way is to go through the SINGS page from the left column menu to the 'Summary page' for the survey].
10. Go to the ESO archive. There is a z=5.4 quasar at Ra = 37.906881 Dec = -7.481803 (J2000 desimal degrees), are there any observations of this object in the ESO archive? [the link above is to the ESO Science Archive] When you are at the ESO Archive page - find out what the wind speed was at Paranal at UT=5 on the night starting 5th of November 2006.
11. Use the NIST database given above to find all [O III] lines between 4300Å and 5200Å - it is useful to know that forbidden lines will be indicated in the Type column. How many are there and what are their wavelengths?
12. A common situation is that you have a spectrum and an unidentified line at some wavelength. Let us assume that you measure it close to 4069Å - use the Atomic Line List database to find a likely identification. Only use the elements H, He, C, O, N, Ne, S for this check and assume that the line appears to be nebular in nature so is likely to be a forbidden transition. The night sky also has a very strong emission line at 5577Å - where does this originate? (you might want to choose to quote wavelengths in air when you search for this).

Using topcat

This project will take some time but this is “learning the tools” time.

The steps we will go through here is to download a file from an archive server. Inspect the data and extract a useful subset of the data using a graphical interface. We will then output these and analyse them in a separate program.

a) The first task is to start-up Topcat. It might be installed already but if you need it for your own computer you can find it at:

```
http://www.star.bris.ac.uk/~mbt/topcat/
```

Go there and download the topcat-full.jar file and start this. If this is presenting problems, try the WebStart version a bit further down on the page. If it is installed, just open a terminal window and type topcat. All ok?

b) Your first real task is to go to Vizier and download the appropriate data. We will look at the Tully-Fisher relation for spiral galaxies, but the actual science is not terribly important. This is a relation between the rotation velocity of a spiral and its absolute magnitude.

1. Go to Vizier and look for a table by Mathewson (from 1996). The title is “Parameters of 2447 southern spirals (Mathewson+ 1996)”. Can you find it?

2. Start by exploring Vizier a bit - can you find the abstract of the paper that presents these data? What publication does this table refer to - and do the authors refer to any other similar tables?

3. The Readme file often contains useful information - how do they calculate Aext?

4. What kind of output formats can you choose between - are there unfamiliar formats (I presume so)?

5. Download this table - make sure to choose ALL columns & unlimited number of rows. Choose VOTable as the format. Name the downloaded file “TF-Mathewson.xml” - well, name it what you want but that is what I will use to refer to it in the following!

You should also try to download other formats if you will - if you try this at home you might want to play with Google Sky.

If you look at the file now in an editor you can see the structure of VOTables - can you make sense of it? We will look at this in a later lecture as well.

c) Where we look at the data in a practical way.

The VOTable you downloaded might be hard on the eyes - let us look at it in Topcat instead. You should have started Topcat already - if not do it now.

1. Open the table in Topcat. You do this using the “Load Table” option in the File menu. After opening you should now have a few columns in the drop-down menu next to ‘Sort order’ - check that you do!

2. First task: Check the table metadata/Table parameters - you can either do this using the menus or clicking on one of the icons - what is the table title? If you downloaded different formats, do they contain different information here?

3. Plot Galactic longitude (GLon) versus Galactic latitude (GLat). What is the horzontal gap around Glat = 0? Why is there a gap in the northern Galactic cap do you think? To try to understand this, try to look at a histogram of the declinations of the galaxies - how do they distribute & and does this have any impact on the gap you saw? [you might need to make some more plots to figure this out].

4. The column names are a bit cryptic - one way to get more information is to use the Column Metadata/ Column Info option. What is aESO and what are the units it is measured in?

5. It is possible to create new columns in Topcat - this is quite useful at times. This is done in the Column Info window you just used. Click the ‘+’ button to open the relevant window. The $ID column shows the name you refer to this column by. Let us create an ellipticity column: e=b/a. If you didn’t click the’+’ button, do so now.

Fill in a descriptive name, ellipticity works fine and in the expression, we want b/a - in the window we see that b is $15 (at least in my file!) and a is $14. So the expression to fill in is:

```
$15/$14
```

6. The Tully-Fisher relation is a relation between circular velocity and absolute magnitude. Thus to examine this we need to calculate the absolute magnitude in the I-band (because that is what we have information to do). The absolute magnitude is given as:

$M_I = m_i + k_z(i) + 5 - 5 \log_{10} (d)$

where k_z is the k-correction in the i-band - this is ID $21 in my table. You get the i-band magnitude, m_i, from the table and you can get the distance d in Mpc from Hubble's law. You can assume that, H_0, Hubble's constant is 75.0 km/s/Mpc and you need to find a recession velocity for the galaxy.

Create now a new column M_I which contains the absolute magnitude. You might want to do this in two steps - first create a column with the distance and then use this column in the calculation of M_I. (log base 10 is written log10 and 10^6 is best written 1e6).

The first galaxy in the sample should have M_I=-21.628

7. Ok, you are now ready to look at your data. First brows the table data using View/Table Data or the relevant button. This shows a grid of your data and can be useful for browsing. Then choose Graphics/Plot and plot M_I versus Vopt - what is clearer, logarithmic axes, linear or a mixture?

8. Now identify a few points - what is the object at M_I just above -15? If you kept the data table display open you should now see this row selected when you click the point.

9. Selection of subsets is another important task. We now want to check whether the TF relation looks the same for heavily attenuated galaxies. Create a plot of inclination (i) and extinction (Aext). We are interested in the region of Aext > 0.25 and i > say, 40. Either zoom in by dragging a box around this region and choose Subsets/New Subset from visible, or drag a box around by Subset/Draw Subset Region (note that in the second case you need to choose it twice to finish the region). [check that the axes are not logarithmic when doing this because it might cause problems].

When this is done, go back to the M_i vs Vopt plot and see whether your new subset has a similar distribution - if you find the dots to be too small or the colour hard to read, click on the small symbol in the bottom right of the plot panel for your subset and change as you wish.

10. Finally save the result as a FITS and CSV file to be able to work with this in Python or other programs later.

To save a FITS file: Go to 'Save Table' in the File menu and in the window that pops up make sure that 'Row Subset' is "All". For output format choose 'fits'. Choose some sensible name - to make sure you know where it is saved, use the Filestore Browser to choose the name.

To save a CSV file: Do the same as above but choose 'CSV' as format (not CSV-No Header).