The data science life cycle:

National academies report: "Repeated exposure to the data science life cycle (i.e., posing a question, collecting cleaning and storing data, developing tools and algorithms, performing exploratory analysis and visualization; making inferences and predictions; and communicating results) is needed to help hone the skills required to assess the data at hand, extract meaning from them, and communicate those findings to non-experts."

We divided it up as follows:

- Data wrangling
- Visualization
- Statistical thinking
- Modeling
- Computational thinking
- Communication skills
- Attitudes toward data and research

The content, topic, assignments, and assessments need to be slaved toward gaining acumen in these categories.

Data Wrangling:

This category includes accessing and downloading data and preliminary preparations, but also exploratory data analysis.

In the context of this class, it will encompass skills and activities such as how to use online catalogs, how to use SQL, and basics of exploratory data analysis, probably in the context of comparing one data set with another.

Visualization:

Data visualization is the graphical representation of data. This also includes tables.

In the context of the class it will include proper preparation of graphs (figure check list), as well as different types of graphs that may be useful for exploratory data analysis.

https://en.wikipedia.org/wiki/Data_visualization - this has a lot of useful links

Statistical thinking:

Statistical thinking, roughly speaking, is the recognition that all data is influenced by statistics. This ranges from uncertainty in data (all data points have errors) to underlying assumptions about distributions (a normal distribution does not describe everything). In astronomy, statistical thinking includes features such as flux limits in surveys.

In the context of the class, this category will include understanding uncertainty primarily, but also what it means for data to be characterized by a distribution. E.g., the luminosity function of quasars is not a normal distribution. Also to recognize where the assumption of a normal distribution is inherent in a method.

https://www.fharrell.com/post/introduction/
https://amstat.tandfonline.com/doi/full/10.1080/10691898.2002.11910677#.XhTC1S2ZNBw

Modeling:

We will use this category to mean fitting physical or empirical models to data. "Data modeling" means something else from the data analytics point of view.

In the context of the class, this category includes exercises of fitting data, and understanding what constitutes a good fit. In a sense, it overlaps with statistical thinking, since by necessity it would include a discussion of chi2 etc.

https://courses.washington.edu/matlab1/ModelFitting.html
https://physiology.arizona.edu/sites/default/files/psio472_12_26awe.pdf

Computational Thinking:

Computational thinking involves expressing problems and their solutions in ways that a computer could execute. Apparently this term has a long history.

Features that are (from the Wikipedia page below) include:

- Using abstractions and pattern recognition to represent the problem in new and different ways
- Logically organizing and analyzing data
- Breaking the problem down into smaller parts
- Approaching the problem using programmatic thinking techniques such as iteration, symbolic representation, and logical operations
- Reformulating the problem into a series of ordered steps (algorithmic thinking)
- Identifying, analyzing, and implementing possible solutions with the goal of achieving the most efficient and effective combination of steps and resources
- Generalizing this problem-solving process to a wide variety of problems

https://en.wikipedia.org/wiki/Computational_thinking

In the context of the class this category includes basics of the python language, but also the ideas above, e.g., breaking the problem down into smaller parts, and reformatting the problem into a series of ordered steps. But also the idea of not doing tasks by hand repeatedly – write a program to do it for you. It also will encompass what people do when they do theoretical modeling.

Communication Skills:

An important part of research is sharing the results with your peers and the public. At the most fundamental, this includes writing and speaking. Scientific or technical writing differs from the ordinary writing or essay writing in the following:

- **clear** - it avoids unnecessary detail;
- **simple** - it uses direct language, avoiding vague or complicated sentences. Technical terms and jargon are used only when they are necessary for accuracy;
- **impartial** - it avoids making assumptions (Everyone knows that ...) and unproven statements (It can never be proved that ...). It presents how and where data were collected and supports its conclusions with evidence;
- **structured logically** - ideas and processes are expressed in a logical order. The text is divided into sections with clear headings;
- **accurate** - it avoids vague and ambiguous language such as about, approximately, almost;
- **objective** - statements and ideas are supported by appropriate evidence that demonstrates how conclusions have been drawn as well as acknowledging the work of others.

(taken from:
https://www.le.ac.uk/oerresources/ssds/writingskills/page_65.htm’
https://www.unl.edu/gradstudies/connections/scientific-writing

In the context of the class, we will focus on technical aspects (appropriate use of latex, probably can’t really do pptx also?) but also on the logical structure and the fact that the statements need to be based on evidence, not opinion.

Attitudes towards Research and Science:

We will assume that if you are in this class, you already have a positive attitude for research.

Properties of Stars – Stars are not all the same, and the color yields the temperature

Outline of activities: Students are tasked with writing a function to compute a blackbody spectrum. Then they are given stellar spectra and asked to match, as best they can, the stellar spectrum with the appropriate temperature and amplitude black body. Students are then given filter functions. They must write a program that can obtain the color of the star. A plot is made of color versus temperature, and we infer that color measures temperature.

Data analytics concepts exercised:

- Computation – first introduction to writing functions
- Computation – first exposure to convolution.
- Model fitting – chi-by-eye
- Visualization. Making appropriate plots of the data and models.

Duration: two class sessions

Distribution of Stars – Hot Stars indicate current star formation

Outline of possible activities: Look at properties (colors and magnitudes) of nearest and brightest stars. We'll want to get luminosities, so we'll have to get distances from Gaia.

Data analytics concepts exercised:

- Statistical thinking – stars don't have a normal distribution. The distribution influences profoundly what we see.
- Data wrangling (maybe) to get properties of stars.
- Data wrangling – use of catalogs
- Questions involving how many cool stars does it take to have the same luminosity of a hot star

Duration – two class sessions

References:
http://www.astro.wisc.edu/~dolan/constellations/constellations.html

The Colors of Galaxies indicate the rate of star formation:

Outline of possible activities. The simplest would be to provide stellar population synthesis spectra, and students could compute the colors by convolving the spectra with the filter functions. The example is the Bruzual & Charlot. Students could go farther than this and actually compute the result for a project. I seem to have spectral SED templates. And there might be others at STScI.

Another possibility is to do it from spectra of galaxies – except the colors are already available.

We'll need the fuel consumption theorem, but that can just be given.

Data analytics concepts exercised:

- Statistical thinking – the concept of an IMF.
- Computation – computing colors of galaxies
- This mostly emphasizes the previous.
- An alternative would be to use portions of this for a HW.

References:
http://stev.oapd.inaf.it

Duration:

Expansion of the Universe:

Outline of activity: use old data to explore the Hubble expansion, specifically fitting the data. Then, as an activity, analyze / fit new data. Discussion of how it means that we need a more complicated model.

Data analytics concepts exercised:

- Model fitting. What it means to have an appropriate model fit.
- Statistical thinking. The role of uncertainty in constraining the model.

Duration:

Galaxies evolve, becoming redder with time:

Outline of activity: Students extract colors from galaxies in clusters from SDSS using redshift and position cuts. They clean the data (probably stick with the most luminous?) and analyze the result.

Data analytics concepts exercised:

- Data wrangling: use of catalogs, use of SQL, data cleaning

- Visualization: plots that show the effect
- Statistical thinking: how to show the differences in cluster content, e.g., KS test.

References:

https://en.wikipedia.org/wiki/Butcher–Oemler_Effect

Duration:

The Masses of Galaxies I

Outline of activity: The science is to investigate the velocity dispersion in a spectrum. It doesn't seem like it will be very easy to do this in fact (i.e., fit a convolution model to it). However, it looks like it may be a measured product of some galaxy spectra in SDSS:

http://www.sdss3.org/dr8/algorithms/veldisp.php
https://www.sdss.org/dr12/algorithms/redshifts/

In this case, one could make it a data wrangling exercise, e.g., find galaxies in SDSS, learn how to read and use fits files.

At the same time, we could also make it a computational exercise, i.e., use of a convolution kernel, as the previous exercise was more of a weighting than a convolution, I would say.

Data Analytics concepts exercised:

- Data wrangling – structure of fits files.
- Computation – convolution and kernel.

Duration:

Masses of Galaxies II

Outline of activity: Once we have a velocity, we need a radius. A good radius is the galaxy half-light radius. So download the images of the galaxies with the dispersion measurements and extract their half-light radius. Those can be converted to distance using angular size.

The question is, how to analyze half-light radius? Well, again, SDSS does it for you – they fit a de Vaucouleurs profile to the images and those are available from the pipeline. Still, it would be great to do some data analysis here. Maybe it could be done by hand, i.e., given an array of data, find the brightest pixels, create a radial profile.

Data Analytics concepts exercised:

- Computation (Kepler’s laws, angular size also.)
- Depends on how we get half-light radius, but can perhaps be done from an image.

https://github.com/brittlundgren/SDSS-EPO

Duration:

Black Hole Masses I

Outline of activity. Measuring the radius of the broad line region. Here they can go further with convolution by thinking about a transfer function. Collin has already drafted a lag visualizer. But can explore other topics such as computation of lags, what can you do when the data is gappy, etc.

Data Analytics Concepts exercised:

- Computation – transfer function
- Statistical thinking ? Effect of gappy data?
- Visualization – constructing and animated notebook?

Duration:

Black Hole Masses II

Outline of activity: Measuring the velocity of broad line clouds. That will be obtained through spectral fitting using Sherpa, which will hopefully work on the sciserver.

Data Analytics Concepts exercised:

- Modeling data – choosing the appropriate model
- Statistical thinking – what constitutes a good fit.

Duration:

Galactic Dynamics Simplified

Outline of activity: N-body dynamics begins with programming the equations of motion for two bodies and evolving them using the Euler method. They should yield orbits around one another. Another point is that this system can be solved for analytically, but it is easy to imagine that more than two bodies can’t be solved for. Finally, there are choices to be made when performing this modeling, e.g., step size, and the integration method. Note that something similar to this may be available on the PICUP website. The restricted three-body problem might be suggested as a project?

Data Analytics Concepts exercised:

- Computation – equations of motion for two bodies

Duration:

Galaxy Dynamics and Star Formation

Outline of Activity: Not sure. Want to link galaxy collisions with star formation. Ok, apparently, there was a time when the Butcher Oemler effect was discounted, and it was thought that rather blue galaxies resided on the outskirts of clusters. So could try to find a data set that exihibits this.

http://www.cfht.hawaii.edu/~morrison/home/notes/BO.html

Data Analytics concepts exercised:

- Data wrangling – data mining
- Statistical thinking – may take some work to see the differences in the color distributions

Duration:

Cosmological N-body Simulations

Outline of activities: Investigating results from n-body simulations. I wonder if there are some already developed classroom exercises for this. I believe that for ASTR 1504 I found a movie that somebody had programmed. There are certainly youtube videos about how this is done. Otherwise, students might systematically investigate the "Galaxy Collider" app.

Data Analytics concepts exercised:

- Computational – parallel programming ?

http://www.cs.ucy.ac.cy/~ppapap01/nbody/

Duration:

The Subgrid

Outline of activities: Students will compute the scale of accretion onto a black hole in the center of a galaxy and compare with the typical computational size of the “universe in a box”. Order of magnitude computations will show that computing all scales from the black hole accretion size scale to the dynamical size scale would take a very long time.

Data Analytics concepts exercised:

- Modeling concepts – how theoretical models are set up and how choices are made.

Duration

So it ends up being 13 topics, which fits rather nicely into the 15 week semester, if the last two weeks will be for projects.