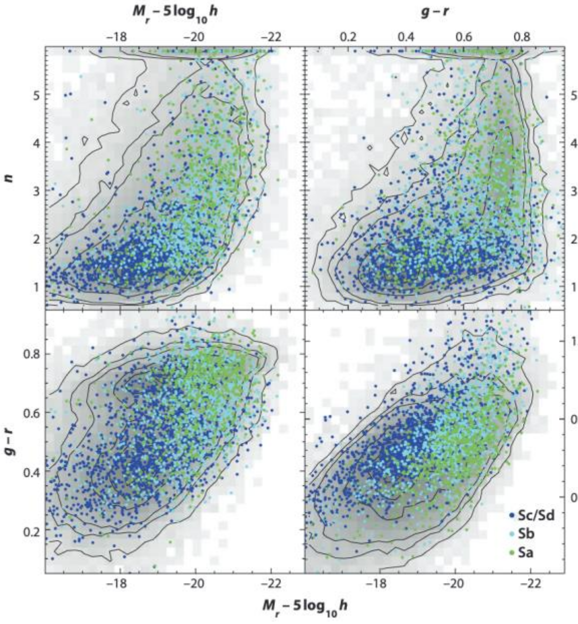
Katya Gozman
SI 649

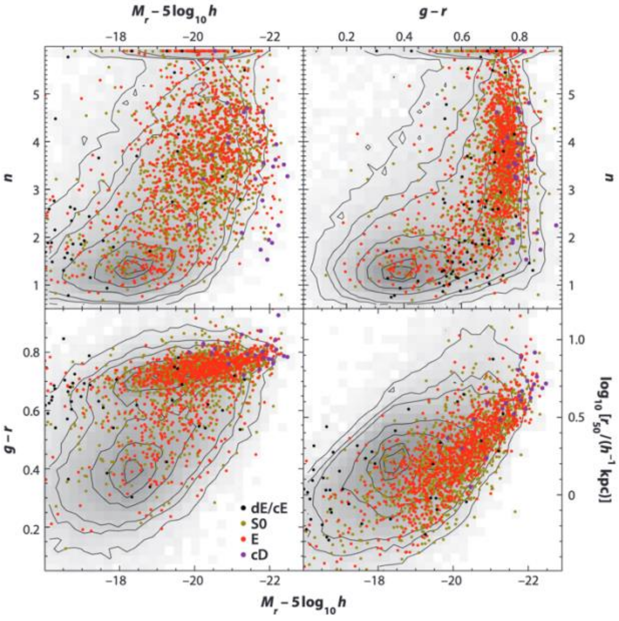
Scientific Viz Project

My project can be found here:
https://huggingface.co/spaces/kgozman/galaxy_properties_scivizproj

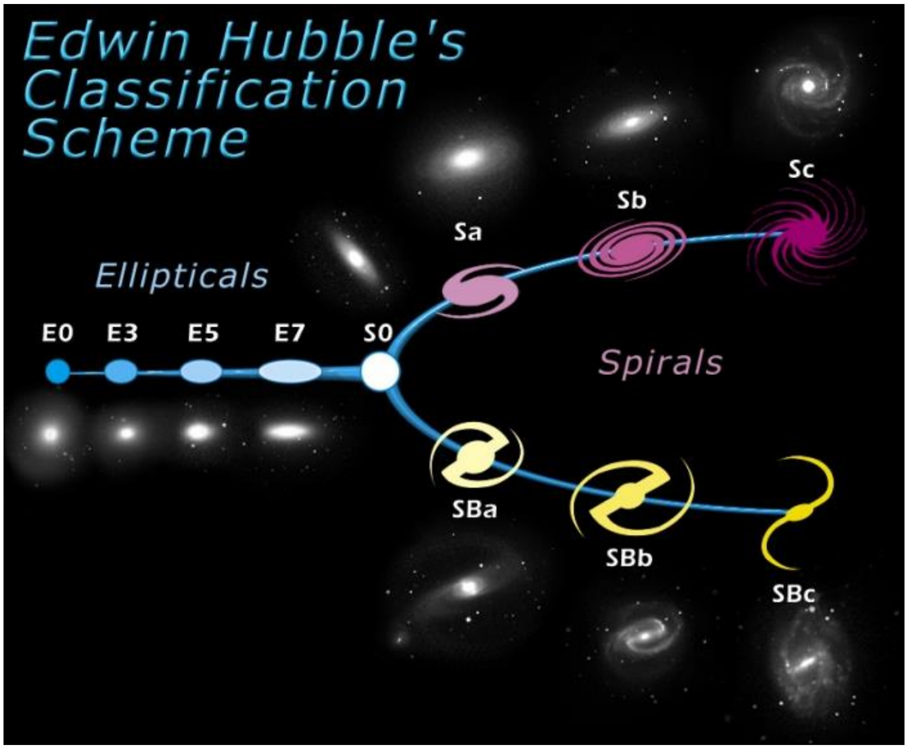
Initial Idea and Background

Since this project focuses on scientific visualizations, I wanted to try and recreate a figure from a paper in my own field, astronomy. I thought about different plots in subfields I'd see that might be interesting to make interactive, things like data on stellar streams or exoplanet populations, but after throwing around some ideas with people in my department and looking at the feasibility of obtaining data used to create those visualizations in the first place, I decided to go with a plot from a well-known review paper in the galaxy subfield, Physical Properties and Environments of Nearby Galaxies by Michael Blanton and John Moustakas. This paper gives a general overview of different properties of nearby galaxies and how they change with different galaxy types. I chose to recreate/combine figures 8 and 12, which both show the same type of plot, just for different galaxy types. They are representatively shown below.





This figure shows the relationship between four different galaxy properties: M_r-5log$_{10}$h is the r-band absolute magnitude, which is basically how bright the galaxy is in a certain filter. g-r is a metric measuring the color of a galaxy, basically the difference in how bright it is in two different filters. log$_{10}$[r$_{50}$/(h^{-1} kpc)] is the log of the galaxy's half-light radius, i.e. the radius that contains 50% of the total light from the galaxy, and n is the galaxy's Sérsic index, which describes how concentrated the galaxy's surface brightness profile is. These are all fundamental and important measured quantities that describe a galaxy, and different kinds of galaxies display different distributions in these parameters. The different colored points are individual galaxies, color-coded by their morphological type. This code was introduced by Edwin Hubble in the 1920s to classify galaxies based on their visual appearance. At its core, this splits galaxies into ellipticals (uniform and smooth looking), spirals (with spiral arms, like our Milky Way!), lenticulars, and other less-common types like compact dwarfs or irregular galaxies. A simplified graphic of this is shown below (from Wikipedia/Cosmogoblin - Own work).



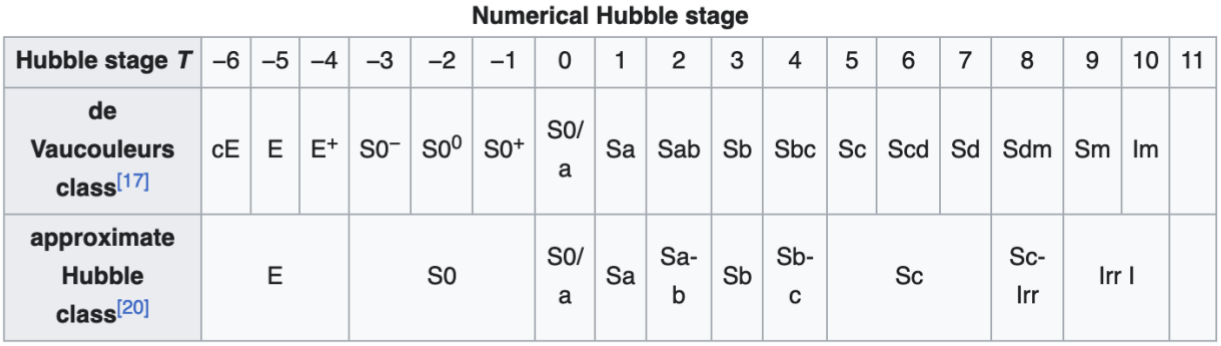
We actually read and discussed this paper in my graduate class on galaxies that I took here and we focused a lot on these figures in our thinking about how different galaxies evolve and what their properties look like, so I thought it would be a relevant plot to recreate. Readers of this paper are likely in the field of astronomy but might not be experts in galaxies specifically, might be students, or might be doing some general background reading, so I thought of this plot as a pedagogical tool for anyone that wanted to get intuition about how different galaxy properties are related and how those numbers translate to what a galaxy actually looks like. It used data from the NASA/IPAC Extragalactic Database (NED) which has information about each galaxy including its morphological class and the Sloan Digital Sky Survey (SDSS), which is publicly available and a famous survey in astronomy that imaged and got spectra of millions of galaxies in the northern hemisphere that began almost 25 years ago.

Recreating the static figure

Though the plot used data from NED and SDSS, the actual final/clean tables used to create the plot are not publicly available. The paper stated that they used “77,153 galaxies with $z < 0.05$ in the SDSS Data Release 6 (DR6; Adelman-McCarthy et al. 2008), an update of the low-redshift sample of Blanton et al. (2005c)” (Blanton & Moustakas 2009). The variable z is what astronomers call redshift, which you can think of as a proxy for distance. $z<0.05$ means all the galaxies are “nearby” on galactic scales. The sample the text refers to I believe is the NYU

Value-Added Galaxy Catalog (NYU-VAGC), so I looked there for the data tables and galaxy cutout images, but unfortunately the website did not have the low-redshift catalog with DR6 data, only DR4 which had a smaller galaxy sample than the original dataset (only ~40000 galaxies). After a lot of digging and URL-breaking I managed to find the hidden DR6 low-z sample. A challenge is that neither of these catalogs had the morphological class for each galaxy listed; I assumed the authors of the paper had cross-listed each of their galaxies with NED somehow. To my great dismay, the DR6 low-z catalog did not have the name of each galaxy that is listed in NED, unlike the DR4 catalog. Both of these catalog columns were also formatted in bytes (they were FITS files, a commonly used file format in astronomy) and were semi-challenging to work with, and I didn't know how to cross-list them with the NED service.

I decided to look for other catalogs that might have all the data I want listed, which was another challenge–many catalogs I found either had morphological data but not other structural parameters or vice versa. I got a tip from a professor in my department to look into the NASA-Sloan Atlas (NSA), which is actually a descendant of the data used in the original paper. This atlas had the same range of redshifts as the original low-z DR6 catalog, so it had the same type of galaxies, and contained all the structural parameters I needed. The table itself didn't have morphologies, but the website also includes links to ancillary catalogs, including one with NED data, meaning I could cross-match between the NSA and the NED catalog to get morphologies. Unfortunately, the morphologies listed in the NED catalog were complex, with over 6000 unique categorizations ranging from "(L)SB(rs)0^+" to just "AGN". Galaxy classification can be very complex and nuanced, with each part of the code standing for a different feature in the galaxy like whether it includes a bar, has an inner ring, or what stage it's in; but for the purposes of this project I needed to simplify to 9-10 categories. Viewers reading the paper didn't need to know whether a galaxy was a (R')SA(rs)ab LINER or a (R')SA(rs)b;Sy2HII? galaxy. The paper states that most of the NED classifications come from the The Third Reference Catalog of Bright Galaxies (RC3), so I hunted down this catalog and found that besides the complicated morphological code, it also listed every galaxy's numerical index T of the stage along the Hubble Sequence – a number that corresponds to a general Hubble class, which I show below for reference (from Wikipedia).



Numerical Hubble stage

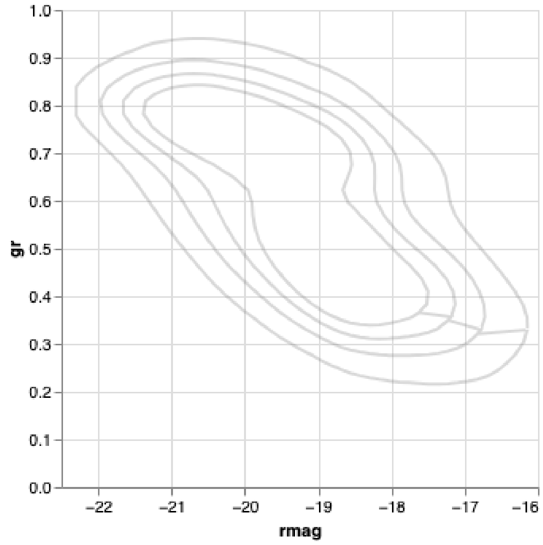
Hubble stage *T*	–6	–5	–4	–3	–2	–1	0	1	2	3	4	5	6	7	8	9	10	11
de Vaucouleurs class[17]	cE	E	E$^+$	S0$^-$	S0^0	S0$^+$	S0/a	Sa	Sab	Sb	Sbc	Sc	Scd	Sd	Sdm	Sm	Im	
approximate Hubble class[20]	E			S0			S0/a	Sa	Sa-b	Sb	Sb-c	Sc			Sc-Irr	Irr I		

Therefore I made the decision to use the RC3 catalog and cross-match with my NSA catalog using the astronomical coordinates of each galaxy, manually checking a few to make sure the matches were accurate. Not all the galaxies in the NSA catalog (which had over 100,000 galaxies!) were in the RC3, but after cross matching I still had over 6000 galaxies with morphological classes, which I decided was plenty and even more rows than Altair can display by default. Though the dataset was different from the one used in the original figure, the overall sample is still volume-limited (all nearby galaxies) and therefore the overall distribution of galaxy properties should be fairly similar between one and the other, which wouldn't change the effectiveness or scientific accuracy. Afterward I did a lot of filtering, cross-matching, and cleaning each of the catalogs to get rid of poorly formatted columns I didn't need, converting the Hubble T type to a string with the corresponding simple class of "E", "S0", etc., and creating a few new columns with some calculated fields like getting the color and calculating the log of the radius in kiloparsecs instead of arcmin.

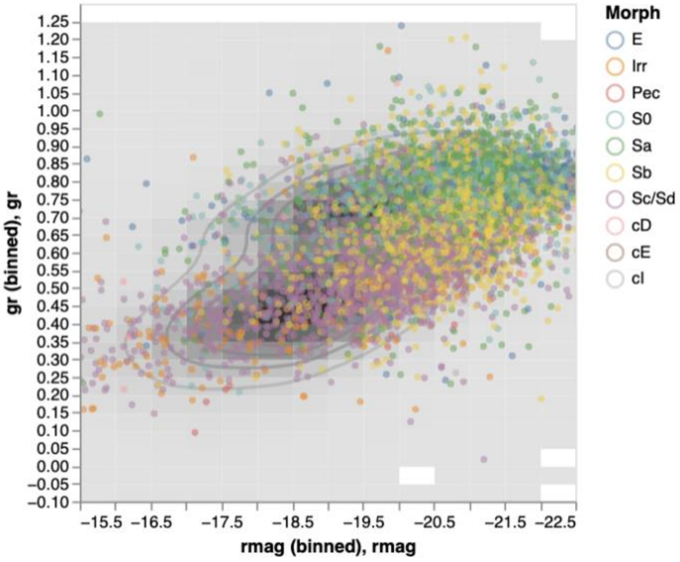
I finally had two final catalogs, one with all the galaxies from the NSA, and one with only the NSA galaxies that had known morphologies. Using these I was able to start recreating the static figure. Since my figure is four plots arranged in a box, I decided to first try making one, and then writing a function to make it create the other boxes where I could easily replace the variables being plotted. I used Altair as my choice of plotting tool since we had been learning it in class, it seemed less limited than Tableau, and is python-based and the only interactive library I now know how to use since I've never learned D3/JS or other frameworks.

I started layer-by-layer: the first step was the background histogram/KDE, which was fairly simple to implement with mark_rect() in Altair. I immediately realized that since my dataset was ginormous, Altair ran into the max row limit of 5000 rows, so I started by running the alt.data_transformers.enable("vegafusion") line, recommended in the Altair documentation, to deal with large datasets (later I would abandon this idea, but more on that later). I also had to implement some sort of dictionary to keep chart specifications for different axes like their limits and titles because different variables needed different axis limits based on the data – the Sersic index n only goes from 0-6, while the magnitudes not only go from about -22 to -16ish, but also need to be flipped. In astronomy, magnitudes use a backward logarithmic scale, so a star with a magnitude of -22 is actually brighter than one that's magnitude -16 (and a magnitude 6 star is 100 times fainter than a magnitude 1 star!), so we usually invert the magnitude axis on a plot to match this. This meant that for the plots with magnitude in them I also had to have a Boolean to take into account whether or not the axis needed to be flipped.

Next came the contours, which were much more of a pain to implement than I thought. After many hours of digging and looking online, I was very surprised to learn that Altair/Vega has never implemented the ability to plot 2D contours, so there was no native way to make these for my binned data. I could make them in other libraries like seaborn or matplotlib, but I couldn't overlay them in the same Altair plot. After chatting with Prof. Card, he showed me resources that let me generate contours with matplotlib, save them in an array, and then plot them in Altair using mark_line(order=None). This worked, except because I was plotting the contours as one array, I was getting a line going between each contour as it plotted from one contour to the next (see figure below). I didn't like how this looked, so in my contour plotting function I saved the number of points in each contour level (hence the indices in the array that corresponded to the start and end of each contour) and then plotted each contour in a loop on top of a "base contour". Though it took a bit of messing around to get working, in the end it worked wonderfully, though did slow down the chart-making function a bit.



Finally, I could move on to implementing the scatter plot, which was relatively straightforward for the static version, just mark_point and encoding the morphology in the color to get the legend to split different galaxies by shape. I could then layer all these charts together and ended up with a nice panel that very much resembled the one in the paper that I was happy with (see below).



Adding interactivity

I decided to get interactivity working with this one panel of my chart, reasoning that I could just as easily implement it in my other three panels through a function. I started first with the part of interactivity that I thought would be the most interesting and give the most improvement to the static figure, and also the one I thought might take a while to get working: having an image of the galaxy pop up when a user clicks on any point in the four-panel plot. I thought it was important for the tool that users would be able to connect the seemingly arbitrary numbers that describe all a galaxy's properties to what the galaxy actually looks like and thereby make comparisons and draw connections between different types of galaxies, both numerically and visually. This would be a big part of intuition-building for the reader –the original plot gives readers a few select examples of galaxies with different morphological classes but doesn't connect these images to the plot.

Looking at the Altair documentation, I saw there was an easy way to pop up an image as a tooltip on hover if you give your plot specification a URL to the corresponding image. I found after some URL/directory digging that all the NSA galaxies had jpeg images associated with them at a certain URL that could be accessed through their subdirectory URL, which was part of the NSA dataset. I added an "image" column to my dataset by generating the URL for each galaxy and then put them in my function to display the tooltip. I found that there was a lag in their loading time and that the images were much too big, which I thought would hinder user interactivity since they blocked a large portion of the plot. Since the NSA data is from SDSS, I knew from previous experience that you could access image cutouts of SDSS data by modifying a certain URL with the corresponding object coordinates, and that you could modify this link to adjust how zoomed in your image was. I switched to encoding this URL in the image column and this worked much better in terms of lessening the lag time.

Though the tooltip worked just fine, I had reservations about this: the image was still big and blocked portions of the plot and sometimes you'd have to scroll to see it if you were hovering over a point in the lower part of the image, and I did also want the numerical properties for each galaxy displayed somewhere. Since I had many data points, this also meant an image would pop up almost anywhere you hovered on the plot and might lag if the user made quick motions around the plot. The Altair documentation showed that there was a way to instead make images appear on the side by using a selection interval and a faceted chart, so I tried that instead. Though I couldn't get the facet part of the chart working and since I had so many points that doing a selection interval would mean potentially displaying images of hundreds of galaxies which wasn't feasible, I figured out how to do a selection such that clicking on a point in the plot would display that one galaxy as an image to the side, which worked great. There was very little lag in displaying the image and also let the user see both the image and the entire scatter plot at the same time so that they could link a picture of a galaxy to its physical properties.

Now that I had an image for each galaxy, I wanted to deal with the fact that this viz had so many scatter points and 10 morphological classes. The original viz dealt with this by breaking it up into two plots, one for spirals and one for ellipticals and other minority classes. My viz had all these categories on one plot, which could be overwhelming, so I wanted users to be able to select one or multiple morphological categories and only see those. I followed the example on the Altair site to create a clickable legend where you can select one or more legend categories and lower the opacity of other points. I originally made the non-selected points completely opaque, but I then switched to giving them an alpha of 0.1 so you could just barely see them. My reasoning for this was that I still wanted users to situate different galaxy classes in the context of the overall distribution of all galaxies in these properties. I also figured out how to make it so that when a user selected only a certain number of morphological classes, clicking on a point not in the class would not display any galaxy, because it would be unclear whether the user is clicking on a galaxy of the selected class or not.

I put my plotting code in a function that took in the x and y variables to be plotted and their axis limits and managed to create the four panel plot by concatenating four plot function call results side by side and also concatenated the galaxy image region next to the four panel plot successfully.

After this, I wanted to implement a hover selection. There were a few problems with the current viz that I wanted to solve with this: 1) it wasn't clear exactly which galaxy a user was clicking on because of the density of points on the plot unless they zoomed in, which I also hadn't implemented yet, 2) since there were four panels, when selecting a galaxy in one panel, I wanted users to be able to see where that galaxy fell on the other three panels, and 3) the numerical values of all the relevant galaxy properties weren't being displayed anywhere. I added a hover which made the hover-over point larger and also would pop up a tooltip with the name of the hovered galaxy and its other structural parameters. Due to the way I concatenated my plots, hovering over a point in one plot automatically did the hover selection in the other three panels, which is exactly what I wanted – I wanted users to be able to see where a galaxy fell in the context of all four plots. Since there were many points and some points were hidden partially behind others, simply making a hovered point larger was not always effective and was sometimes hard to see, especially in the other panels. Unfortunately, in Altair/Vega there is no way to dynamically change the z-order of a point (i.e. where it appears in the layers of points) such that a hovered point could always appear on top of its neighbors. With this software limitation, I decided to increase the size of a hovered point by a lot and also make it black, contrasting with the other colors in the plot. This made the point easily stand out in all four panels, even if it was partially obscured by other points. A problem I noticed after this was that when only one or multiple morphological classes were selected in the legend, hovering over non-selected points still triggered the tooltip. I didn't want this confusing or distracting the user and making it more difficult to interact with the viz, so I split my scatter point plotting into two charts that I could layer together so that the tooltip only appeared on galaxies part of the currently selected morphological class(es). I also added .interactive() to the fully concatenated four-panel plot, so that users could zoom in and pan around and zooming in on one plot would also zoom into other panels if they were plotting the same variables.

My next step was addressing the colors of the viz – users might not like the default colors or they might not be accessible to them if they are colorblind, etc. The default colors I used were the ones used in the original paper for continuity, except two morphological classes that I had in my data that weren't present in the paper which I chose other colors for. I wanted them to have control over all the different colors for each morphological class, so I figured out how to add color parameters and bind them to the color of the points. Now below the four panel chart I had boxes that a user could click and change the color of each class to anything they wanted using their browsers' built-in color picker.

Now that I had an interactive plot, I needed to make it servable so I could host it online somewhere. I decided to use Panel since that's what we had learned in lab and the only framework I knew how to do this in. This is where I ran into a major hurdle: the Vega Fusion data transformer (and as it turns out, Vega Fusion in general) I was using for displaying my large dataset and getting around the 5000 row limit is not compatible with Panel. It's on their roadmap, but I spent probably more time than I should have failing at finding any code or workarounds to get them working together. I initially decided to just limit my data by cutting my datasets at the 5000th row so I could get the Panel implementation working before I dealt with this issue. After doing this and also chatting with Prof. Card, I decided to implement a slider that would let the user choose the number of galaxies to randomly sample from the dataset. I initialized it with 4000 galaxies and let the user go between 0 and the max number of galaxies with morphological properties, having the code recalculate the contours and histogram as well so users could see how this changes depending on how you sample. I also used the alt.data_transformers.disable_max_rows() to display more than 5000 rows, thinking that the difference between 5000 and 6000 galaxies wasn't enough to deeply lag out the viz. Randomly sampling doesn't change the general distribution and is interesting for users to play around with sample size. I did think about implementing something where users could cut galaxies based on distance or magnitude but didn't have time to do so.

One thing that was still bothering me was the placement of the color boxes so users could select the color of each morphological category. They were all in the line at the bottom of the viz, making it hard to see/find and required the user to scroll to see the result of the color change. Unfortunately, Altair is pretty limited when it comes to the placement of parameters and other customization like this, so with the help of another student, stack overflow, and Chat GPT, we figured out the css and code to let me move those color boxes from the bottom to the right side of the entire view, below the image of the selected galaxy. Now everything was more or less contained in one view a user could see without scrolling. The boxes still aren't lined up, but I couldn't figure out the code to do that (and might very well be beyond Altair's capabilities).

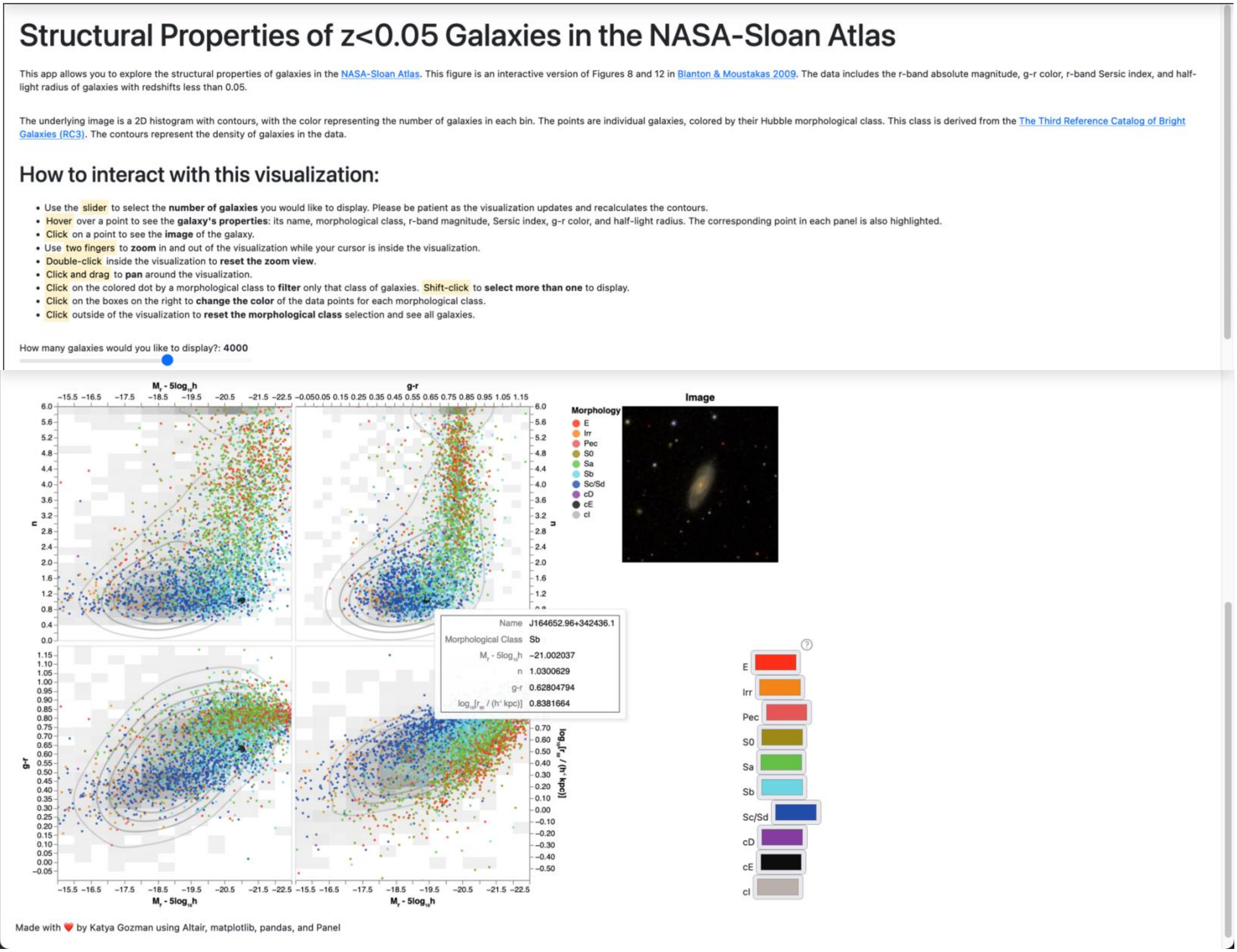
My last steps were to add explanatory text to tell users about the viz and how to interact with it, a footnote with authorship, and I added a little tooltip icon from Panel by the color boxes that lets the user hover over it to pop up instructions on how to change the colors.

Overall, I think I picked a plot that has many different domain tasks it could answer and is rich in information. The plot is situated in the learning/exploratory context of a review paper, so I wanted the plot to allow a reader to use it as a guide to build intuition about different types of galaxies and how their properties are similar/different, as well as related numerical properties to its visual appearance. Users can select data points of interest and explore the viz, filtering by categories of interest and get more details with pop-ups of the galaxy name and numerical values of its properties, change the encoding (color channel) of points, and reconfigure the viz with a smaller or larger dataset – many modes of interaction to explore the data in different ways.

There are a lot of features I tried or would love to implement to improve the viz but either found too difficult/impossible in Altair and Panel. I tried for quite a while to make it so that users could choose what variables to plot on different axes, or to at least give them a choice of what filters to focus on. Right now the plot is showing r-band magnitude and g-r color but in astronomy other magnitudes and colors are also possible and relevant – this dataset listed u,g,r,i,z,fd, and nd, which are just filters that let in different wavelengths of light. The example code on the Altair page to do this wasn't even running on my notebook, as it turns out because of the Vega Fusion transformer I had on, and even with limiting the amount of rows this was still really difficult to implement especially with all the different layers in my plot. I got it working semi-successfully for one panel when I was playing around with this idea, but it also took quite a while to recalculate and reload the data with new variables. You also cannot dynamically change axes labels in Altair/Vega, so this was another challenge with changing what data is plotted. I also tried this with Panel instead of Altair's default, but couldn't do the binding while concatenating the four plots together. I might have been able to implement this if I had more time, but I had already spent way too much time implementing what I already had. I also tried using Panel to implement the color selection instead of the Altair params since this might give me more freedom over their placement, but once again ran into binding and displaying issues. My legend selection also has a bug that might very well be a bug in Vega itself – when you multi select morphological classes and then click one of the selected points, sometimes it deselects all but the last category instead of showing the galaxy and you'd instead have to shift-click to select that galaxy (while deselecting the last class selection). I'm not sure why my selections are interacting in this manner (and neither did Prof. Card after I asked him about this). I would also love to figure out how to make changing the number of samples take less time as well – you must be patient while it recalculates and redraws the contours and points whenever you change the number. The zooming and panning is also a tad laggy, but we are working with thousands of data points. Lastly, if I had more time and if it was less challenging I would have wanted to change the layout/design of the application to make it look nicer, add ticks on all sides of the panels and also add a toggle for users to choose between light and dark mode. It would be cool to have an overlay pop up on the viz the first time it's opened that is semi-opaque with little annotations showing the usage instructions with arrows (almost like its hand drawn) that you could close and open at will instead of the wall of bullet points which would have made the viz feel more friendly. I think a lot of these features could be implemented with a lower-level more customizable framework like D3 which I'd love to learn in the future.

Screenshots of final viz is below.

Katya Gozman
SI 649



Structural Properties of z<0.05 Galaxies in the NASA-Sloan Atlas

This app allows you to explore the structural properties of galaxies in the NASA-Sloan Atlas. This figure is an interactive version of Figures 8 and 12 in Blanton & Moustakas 2009. The data includes the r-band absolute magnitude, g-r color, r-band Sersic index, and half-light radius of galaxies with redshifts less than 0.05.

The underlying image is a 2D histogram with contours, with the color representing the number of galaxies in each bin. The points are individual galaxies, colored by their Hubble morphological class. This class is derived from the The Third Reference Catalog of Bright Galaxies (RC3). The contours represent the density of galaxies in the data.

How to interact with this visualization:

- Use the slider to select the **number of galaxies** you would like to display. Please be patient as the visualization updates and recalculates the contours.
- Hover over a point to see the **galaxy's properties**: its name, morphological class, r-band magnitude, Sersic index, g-r color, and half-light radius. The corresponding point in each panel is also highlighted.
- Click on a point to see the **image** of the galaxy.
- Use two fingers to **zoom** in and out of the visualization while your cursor is inside the visualization.
- Double-click inside the visualization to **reset the zoom view**.
- Click and drag to **pan** around the visualization.
- Click on the colored dot by a morphological class to **filter** only that class of galaxies. Shift-click to **select more than one** to display.
- Click on the boxes on the right to **change the color** of the data points for each morphological class.
- Click outside of the visualization to **reset the morphological class** selection and see all galaxies.

How many galaxies would you like to display?: **4000**

Name	J164652.96+342436.1
Morphological Class	Sb
M_r - $5\log_{10}h$	−21.002037
n	1.0300629
g-r	0.62804794
$\log_{10}[r_{50}$ / $(h^{-1}$ kpc)]	0.8381664

Made with ❤️ by Katya Gozman using Altair, matplotlib, pandas, and Panel