Decadal Plot Redesign

Jay Frothingham

1. Background

The plot in Figure 1 below was included in section 4.5.1 of the 2020 Decadal Survey (*Pathways to Discovery in Astronomy and Astrophysics for the 2020s*). This particular section of the Decadal Survey is focused on data in astronomy, and the importance of curating archival data. The Decadal Survey explains the current shift in astronomy towards large survey datasets and increased use of archival data. However, the Decadal Survey states that "...while some facilities place their data into public archives, these resources are often difficult to tap. The net result is an opportunity lost, for the scientists who could be exploring data immediately rather than spending months reducing it or making new observations, for the observatories that invested in instruments whose data are underused, and for the science that could be done if that data could be easily accessed" (*Pathways* 4-17). In contrast to these inaccessible data, the plot is meant to show an example of a well-curated data archive and its benefits. The original figure description praises the Chandra X-ray Observatory archive and "[demonstrates] the impact of a well-organized archive" (*Pathways* 4-17). This figure is intended for readers of the Decadal Survey with little background in astronomy research who need to be convinced of the importance of archival data curation in order to prioritize and provide funding and resources for the purpose of data infrastructure. It is meant to show that organizing and curating archival data will produce valuable results in the form of publications.

Figure 1. Original figure from the Decadal Survey.

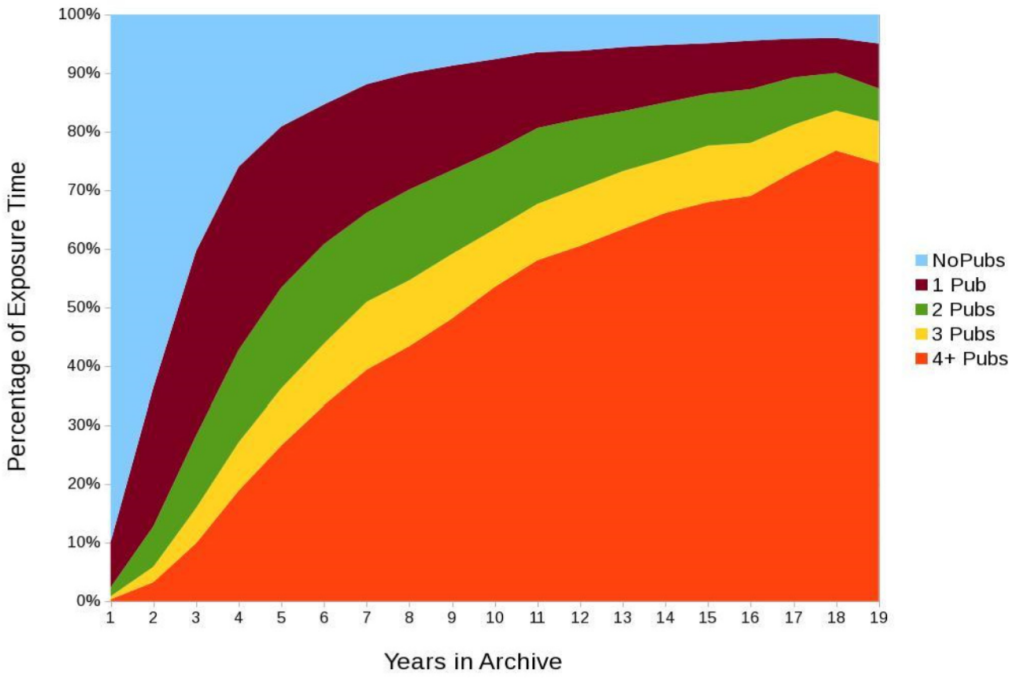


Figure 1 shows a sand plot chart. The vertical axis, labeled as "Percentage of Exposure Time," represents a percentage of the total data collected by the Chandra X-ray observatory. The horizontal axis, labeled as "Years in Archive," represents how many years have passed since data was collected by the Chandra X-ray observatory. The rightmost end of the plot would include older data, while the leftmost end of the plot would include more recent data. Most prominently, the plot features brightly shaded regions representing what percentage of the total data has been included in publications against the age of the data archive. The width of the blue shaded region represents the percentage of data not included in any publications, the width of the maroon shaded region represents the percentage of data included in one publication, the width of the green shaded region represents the percentage of data included in two publications, and so on.

The plot shows that for older data, a higher percentage of it is included in four or more publications and only a very low percentage of it is completely unused in publications. The percentage of data included in publications increases rapidly with time in the archive within the first four to six years in the archive, then increases more slowly.

The original source of the data is simply listed as "Courtesy of the Chandra Data Archive operations team" (*Pathways* 4-17). Rather than being collected in a formal study and published, it was likely drawn from the Chandra X-ray Observatory archival records keeping track of when data were collected, archived, and used in publications.

Bias in the data could be introduced in the motivations of the makers of the graph. For instance, the horizontal axis reverses the chronological flow of time by including older data further from the vertical axis. This creates the appearance of an upward trend in the data, which may give viewers of the figure a slightly more positive interpretation of the data. The figure may also be somewhat biased in that it doesn't specify how quickly data are reduced, or at what point data are introduced to the archive. If data are collected and immediately archived, there will be a delay of reducing and processing the data before they can be used in publications, so there will naturally be a smaller percentage of the archived data in publications. The Decadal Survey did not specify whether this is the case, or whether the figure accounts for that difference between older and newer data.

2. Analysis

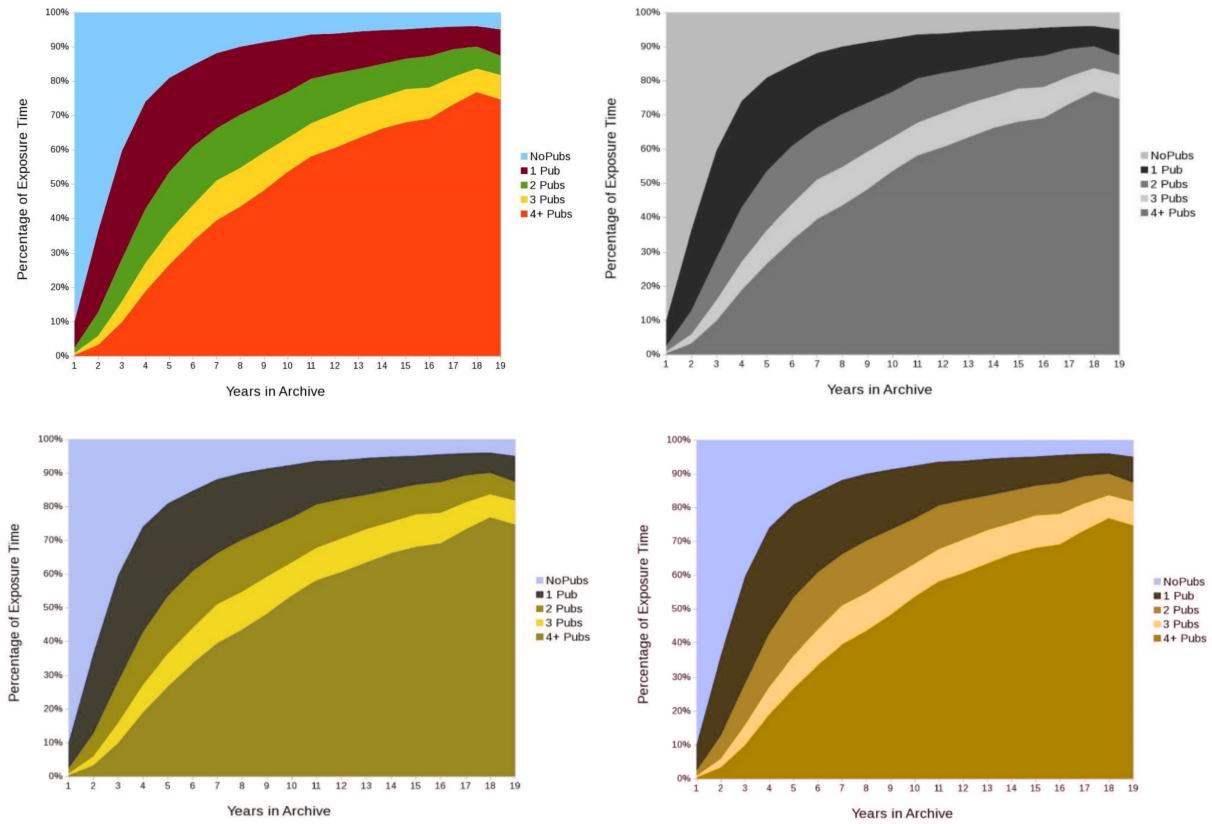
The original plot does succeed in some areas of graphical excellence. It certainly follows Tufte's primary principle of statistical graphics: "Above all else show the data" (Tufte 92). The eye-catchingly shaded regions of the sandplot, and the trend lines separating each region, are the main features of the graphic. The graphic also follows Tufte's principle "Erase non-data-ink, within reason" (Tufte 96). There is no grid, and the figure does not include excessive tick marks on the axes. As a whole, the figure has a fairly high data-ink ratio, following Tufte's principle "Maximize the data-ink ratio, within reason" (Tufte 96). The figure does so through its

compliance with the two previously mentioned principles. The original plot does a good job of emphasizing the data without distraction from unnecessary non-data ink.

However, the original plot violates a number of other principles for statistical graphics. According to Tufte, each of his principles should be followed *within reason*. The exact criteria for what is and is not within reason is somewhat subjective. In class, we decided on a new principle for statistical graphic design: identify your audience, and let their ability to clearly understand the graphic drive decisions about reasonable plot features. With that principle in mind, the original figure is not very well-designed.

While the data-ink ratio is quite high, not all of the data-ink is necessary. There is a lot of empty space on the graph filled only by shading. This shading could be considered redundant data-ink or even chartjunk, as the width of each shaded region could be more efficiently presented with just lines. Additionally, the reduction of data-ink has been taken too far. Without some type of grid, it is very difficult to match data points to numbers. Labeled tick marks on an axis are meaningless if they can't be used to quantify the data. Erasing non-data-ink and increasing the amount of data-ink have affected the graph's understandability and are no longer following Tufte's principles within reason. The original graphic also fails from an accessibility standpoint.

Figure 2. The original graphic with simulated effects of color blindness.



Clockwise from top left: original graphic with no added effects; original graphic with simulated effect monochromacy/achromatopsia; original graphic with simulated effect green-blind/deuteranopia; original graphic with simulated effect red-blind/protanopia. Figures generated using Colblindor's Color Blindness Simulator.

The colors are clear and distinct for those without color blindness or with specific forms of color blindness. For others with achromatopsia, deuteranopia, or protanopia, the colors used would present a significant barrier to understanding the legend and distinguishing between plotted populations, as shown in Figure 2. In particular, the red and green regions are difficult to distinguish.

Figure 3. Annotated version of original graphic.

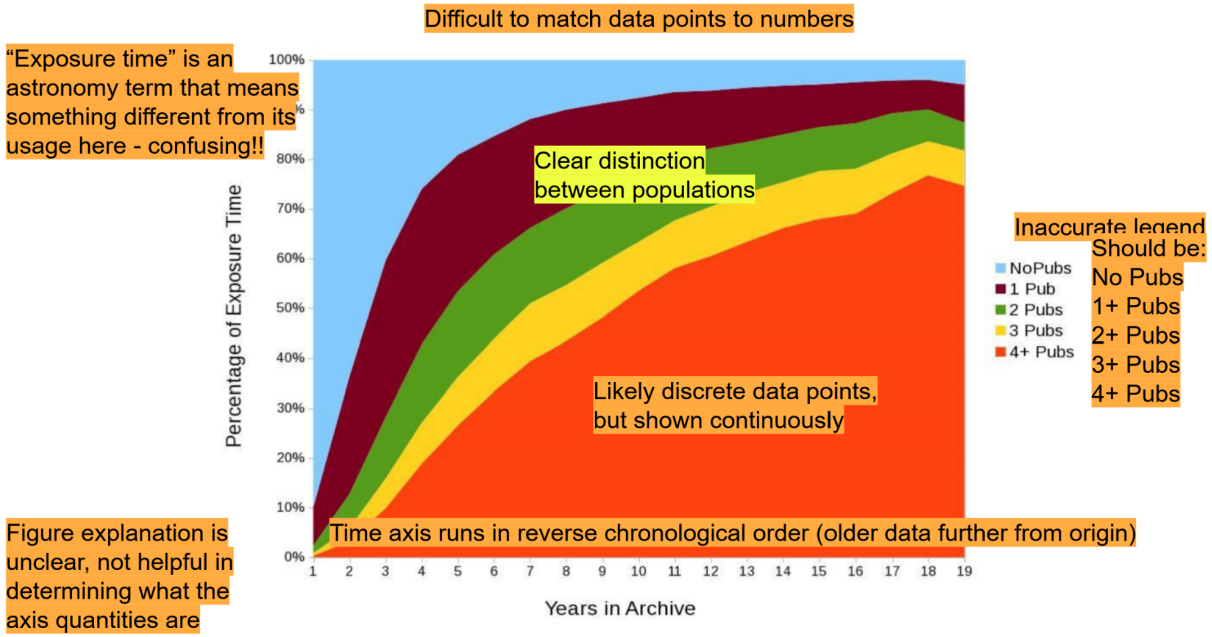
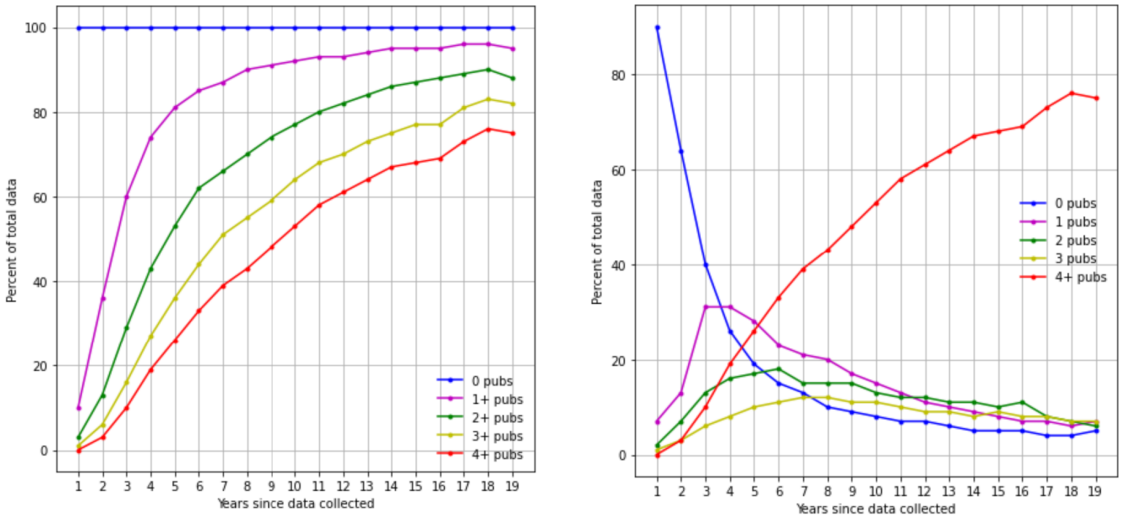


FIGURE 4.8 The percentage of data published as a function of time for data taken from the Chandra X-ray Observatory archive, demonstrating the impact of a well-organized archive. Data here is quantified as exposure time. Here, 70 percent of the oldest data sets have four or more publications using the data. SOURCE: Courtesy of the Chandra Data Archive operations team.

In redesigning the plot, my main aim was to try and improve understanding of what exactly the quantities plotted were. The original graphic had unclear axis labels, an inaccurate legend, and a figure description that was not helpful in alleviating confusion at a glance.

I tried redesigning the original graphic as a scatter plot and a bar chart, and I also experimented with plotting different quantities on the axes. Ultimately, I decided to go forward with a scatterplot redesign, as it seemed to be the simplest way to present the data and clearly show important trends.

Figure 4. Scatter plot redesigns



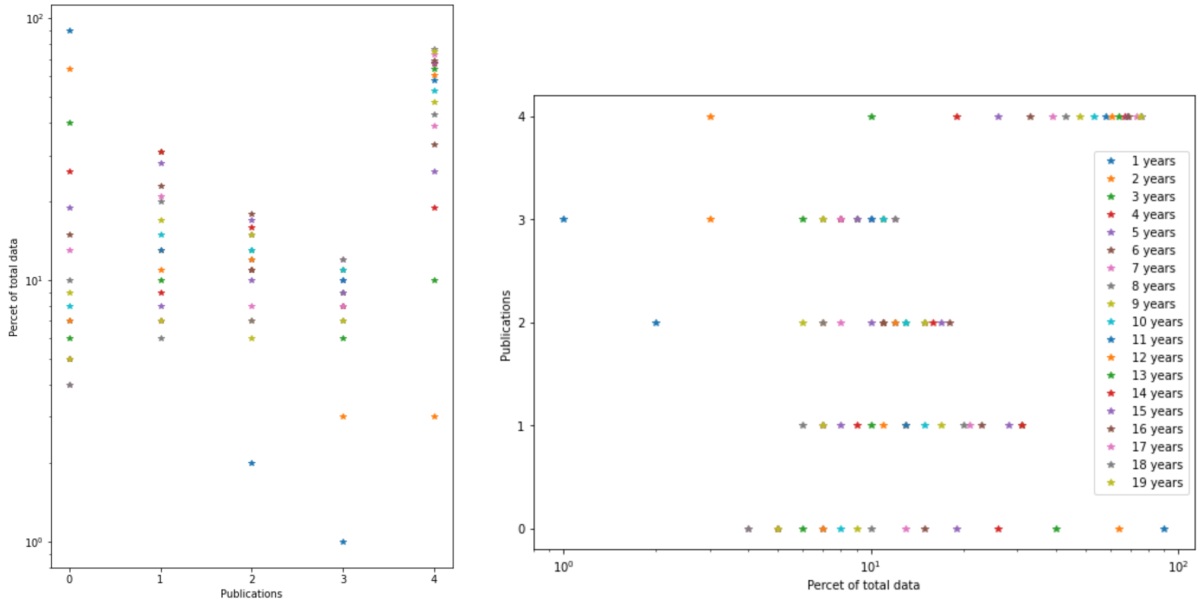
Left: original data categorizations. *Right: new data categorizations.*

For both redesigned scatter plots shown in Figure 4, the improved axis labels, clearer legends, and added gridlines improved readability of the plot. I also indicated discrete data points and connected them with trendlines rather than plotting a continuous line. This makes it clearer how the data were acquired, and that the information on the plot is not continuous, but rather there is only known information at particular points.

The first plot I made transformed the original plot from a sandplot to a scatter plot. It was essentially the same plot, just with shading removed, gridlines added, and clearer axis labels and legend. This figure was useful for checking that I had transcribed the data correctly from the original figure, but like the original figure, was not the clearest way to present the data. In particular, the line labeled “0 pubs” is unclear. When examined on its own, it appears to show that all of the data is unpublished, no matter how old it is. This is definitely false. What it actually means is that all of the data is published in *at least* zero papers. That’s a confusing way to categorize data.

The next plot I made plotted slightly different data. There was confusing overlap in the original figure’s data categorizations. I recategorized the data so that the plotted populations no longer had overlap (i.e., populations were exactly zero publications, exactly one publication, exactly two publications, etc. instead of zero or more publications, one or more publications, two or more publications, and so on). The non-cumulative nature of the populations shows the data trends in a different way. Looking at the line labeled “0 pubs” again, in this plot it shows that a very high percentage of data is unpublished soon after it is collected, but that the percentage of unpublished data drops dramatically as it gets older. This trend is much more clearly visualized in this scatterplot than it is in the previous scatterplot or in the original figure.

Figure 5. Scatter plots with data categorized by age instead of number of publications.



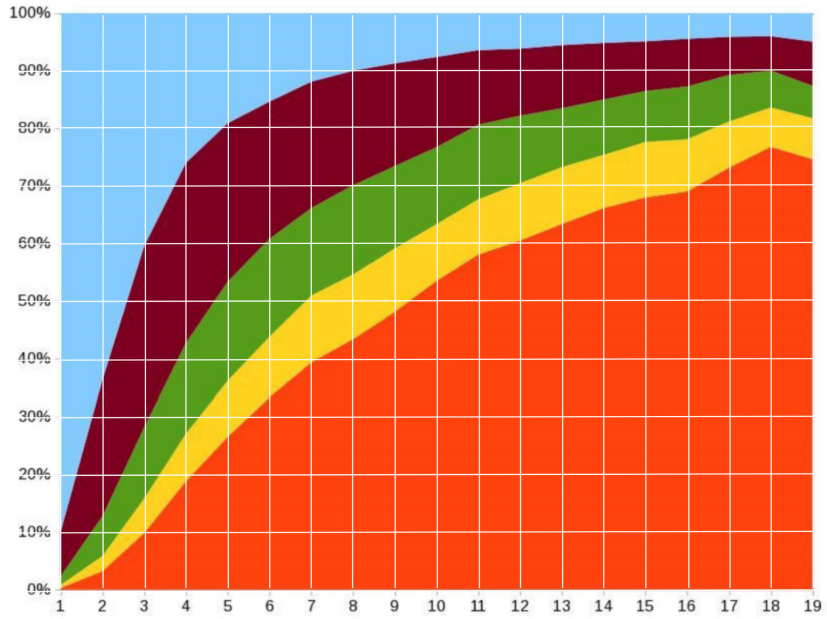
Left: Percent of total data plotted against number of publications.
Right: Number of publications plotted against percent of total data.

I was interested to see how plotting the data with different axes would turn out. I took each population to be an age, or a particular amount of time since data was collected, and plotted the number of publications and percent of total data against each other as scatter plots, shown in Figure 5. Since each year had a data point for each number of publications, this formed distinct columns (or rows, depending on which quantity was plotted on the x-axis). I envisioned it sort of like a bar chart without filled-in bars. The goal was to make it easy to compare how different ages had different percentages of data in different numbers of publications. I think this was an interesting way to consider the data, but not necessarily a good way to see trends. Particularly because there were so many possible ages, there were too many populations to really distinguish between them. This is especially noticeable in the plot with a legend included, which shows that there were not enough default colors for each population to be plotted with a unique color. That could be overcome by including custom colors, but would not solve the problem.

I think it could be interesting to bin the years, making populations of 0-4 years, 5-9 years, etc. Or plotting each population in the same color but with different shades: darker colors indicating older years, like an intensity plot. However, simply being interesting is not a good motivation to make a plot. I chose not to pursue this idea for my final redesign because it would not show trends as clearly, and I thought the relationship between the quantities on the axes would be less intuitive.

3. Procedure

Figure 6. Original figure with manually-added gridlines.

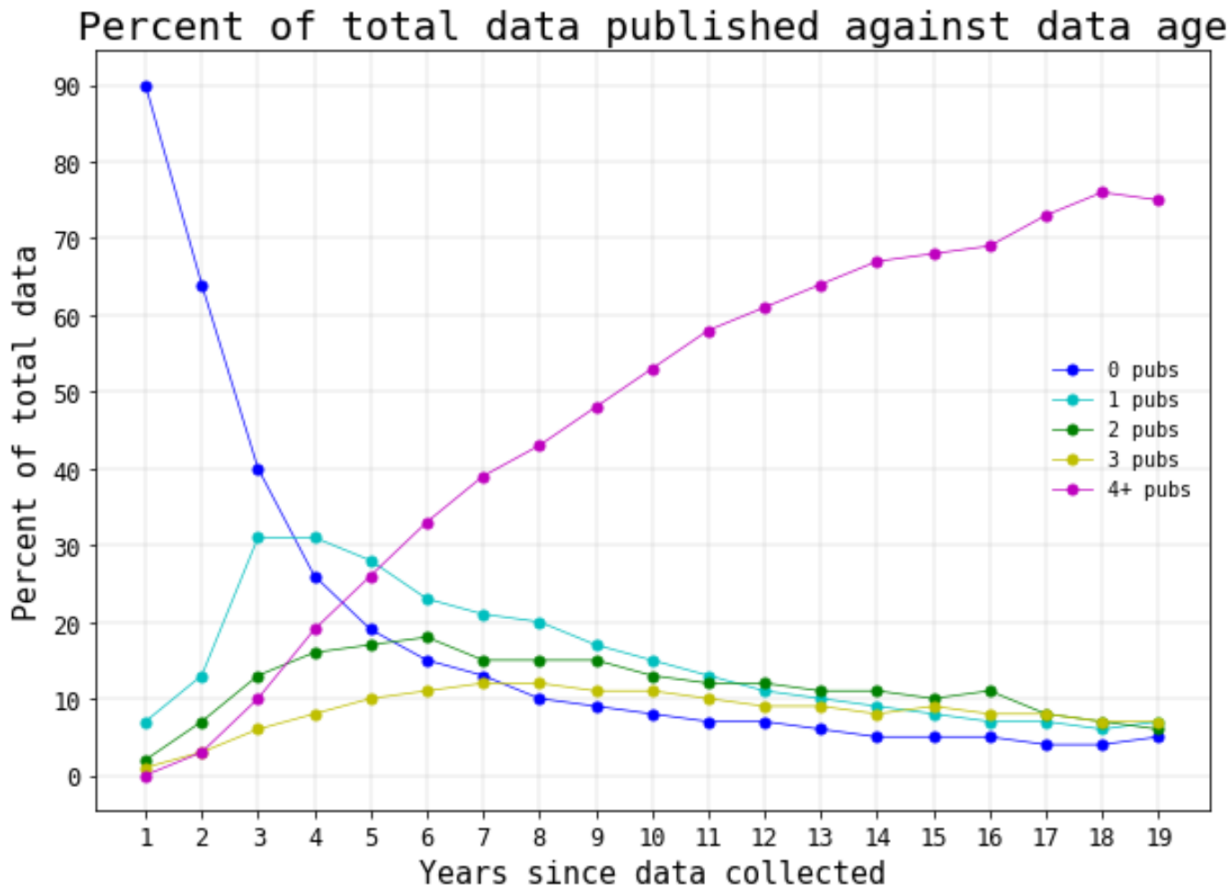


I used Google Slides' image editing tools to manually insert parallel gridlines that matched up with the tick marks already present on the figure's labeled axes. Figure 6 shows the results of this work. It is likely that each gridline is offset from its intended tick mark by up to a few pixels, but they are close enough to help approximate the data points.

I lifted the data from the original figure with careful by-eye observations. I added gridlines to the original image and created a data table with rows for year and columns for cumulative number of publications. In each column, I wrote the estimated percentage of total data for the corresponding year and cumulative number of publications. I also included four columns to record the estimated percentage of total data for each year and exact numbers of publications rather than cumulative numbers of publications. The data for each exact number of publications were determined by finding the width of each shaded region. For instance, in the year labeled 5 on the x-axis, 36% of the total data was published in three or more publications, cumulatively. The yellow-shaded region corresponding to three or more publications intersects the 5-year gridline for approximately 10% of the total data as marked on the y-axis. This means that 10% percent of the total data was published in exactly three papers.

The Python code to manipulate the data and generate the redesigned plot is fairly simple. The extracted data is hardcoded into arrays. The plotting library Matplotlib is used to create the final redesigned plot with a title, axis labels, a legend, data markers, gridlines, and more accessible colors and fonts.

Figure 7. Final redesigned graphic.



4. Discussion

The main differences between the original graphic and the redesigned graphic are the chart type and the data categorizations. The redesigned graphic is a scatter plot with connected individual data points, while the original graphic was a sandplot with shaded regions. The redesign's format is an improvement because it more clearly and simply shows trends in the data. This is partially due to the second difference. The original graphic categorized data cumulatively and had overlap in each category. This allowed the makers of the figure to shade in the entire plot as data-ink. The redesign categorized data without overlap in categories. Rather than shaded regions, each population is represented by a single line connecting data points. The same information is conveyed with less ink and less distraction, and the plotted quantities are easier to understand at a glance.

The more minor differences between the two graphics also constitute improvements. The redesign includes axis labels that more clearly convey the plotted quantities, a descriptive title, and a correctly labeled legend. It also includes a faint grid. This improves readability while still taking into account Tufte's recommendations about clutter and non-data-ink. The colors chosen

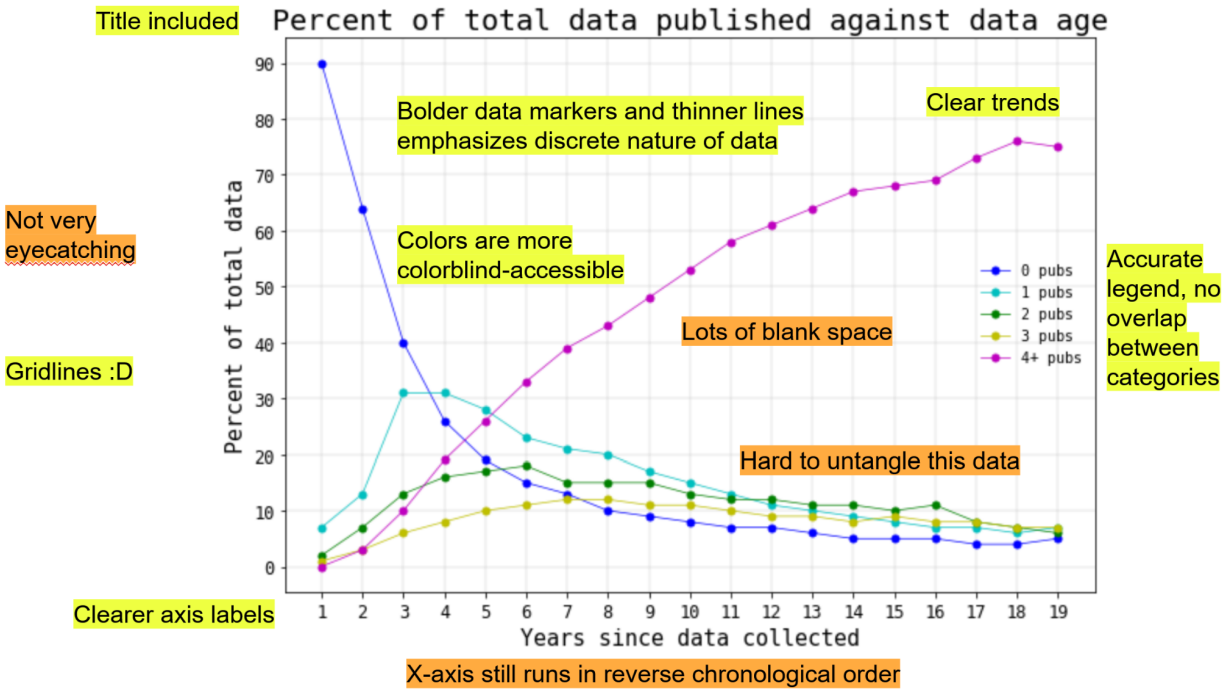
for the redesigned plot are more accessible to color blind readers. They are still not well-chosen for those with achromatopsia, but they are more understandable for other forms of color blindness that the original figure was not designed for. The redesigned graphic clearly shows discrete rather than continuous data. This makes the limitations of the data clearer. There are only known percentages at particular years, but these data suggest a trend.

The redesigned graphic preserves the intended purpose of the original graphic. The important trend conveyed by the original graphic, that more data is published in multiple papers the older it is, is more clearly conveyed by the redesigned graphic. The redesigned graphic also more clearly shows that the highest percentage of unpublished data occurs close to when it is first collected, another trend that supports the idea that older data accessible through an archive is valuable scientifically.

One area in which the redesign may have altered the original graphic's intended purpose is in its simplicity. The plain lines, more muted colors, and blank space rather than shading make the graphic less eye-catching. As a graphic included in the Decadal Survey, the original figure would be intended for a non-scientific audience with significant budgetary power. A more interesting and memorable graphic would invite the reader to spend more time looking at it, and make decisions based on the information conveyed.

Parts of the graphic are difficult to read. In particular, for older years, the percentages of data published in one, two, or three papers are similar to one another, so it is difficult to distinguish individual data points. I have to wonder how many of the populations are really necessary to plot. It's possible that a figure with basically the same trends and takeaway could be produced by only including percentages of data published in four or more publications and percentages that remain unpublished.

Figure 8. Annotated version of final redesigned graphic.



References

National Academies of Sciences, Engineering, and Medicine. 2021. *Pathways to Discovery in Astronomy and Astrophysics for the 2020s*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26141.

Tufte, Edmund R. *The Visual Display of Graphical Information.* 2001. Cheshire, Connecticut: Graphics Press.