**Project 1: Evolution of databases creations and discontinuations**

A database is an organized collection of structured information, or data, typically stored electronically in a computer system. Data within the most common types of databases in operation today is typically modeled in rows and columns in a series of tables to make processing and data querying efficient. The data can then be easily accessed, managed, modified, updated, controlled, and organized. In this CURE project, we investigated the long-term availability of biological databases to better understand how long databases are available after their creation and what factors influence their lifespan.

In this short project, I explore the evolution from 1995 to 2022 of database creations and discontinuations.

**Methods and dataset**

**Dataset:**

The dataset used for this analysis was collected during the 2022 BAT 102 class. The global dataset is available at <https://hurwitzlab.shinyapps.io/DS_Hero>

2282 entries (1 entry per database). Excluded the Databases never online and the Databases never archived in Wayback Machine.

The variables included in this analysis are defined below:

* db\_id : Unique identifier for the database in JL\_DB dataset
* resource\_name : Name of the database
* first\_publication : Date of the first article publication of the database
* available\_2022 : TRUE if the database is available online in 2022, FALSE if not
* last\_accessible\_year : If the database is not accessible in 2022, date of the last accessible year archived in the wayback machine

**Notebook and data availability:**

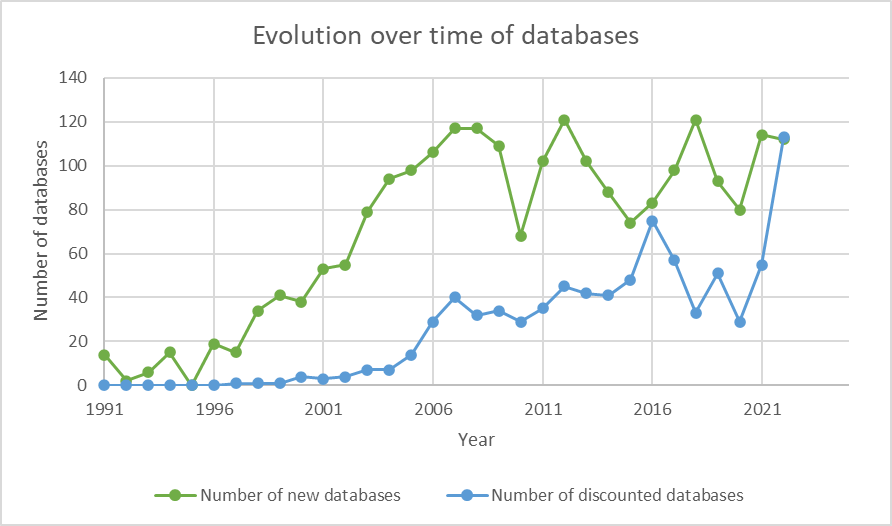
An electronic lab notebook for this analysis is available on Github and contains all collected and

analyzed data for this project: [**https://github.com/SamandarJakbarov/CURE2022\_SJakbarov/wiki/10-15-2022-Data-analysis-overview-and-hypothesis**](https://github.com/SamandarJakbarov/CURE2022_SJakbarov/wiki/10-15-2022-Data-analysis-overview-and-hypothesis)

**Results:**

**1:What is the evolution over time of the number of new databases published each year?**

**I first explored what is the evolution over time of the number of newly published databases. I computed how many databases were published each year as I learned from the video, which toughs how I can compute in the EXCEL program. I found so much helpful information about databases, for example in 1991 only 14 new databases were published and the most mew published databases were in 2012 and 2018. Also, the knowledge I got from the video was helpful not only for this assignment but in next future, because the formulas which I learned not only do faster but also correctly.**

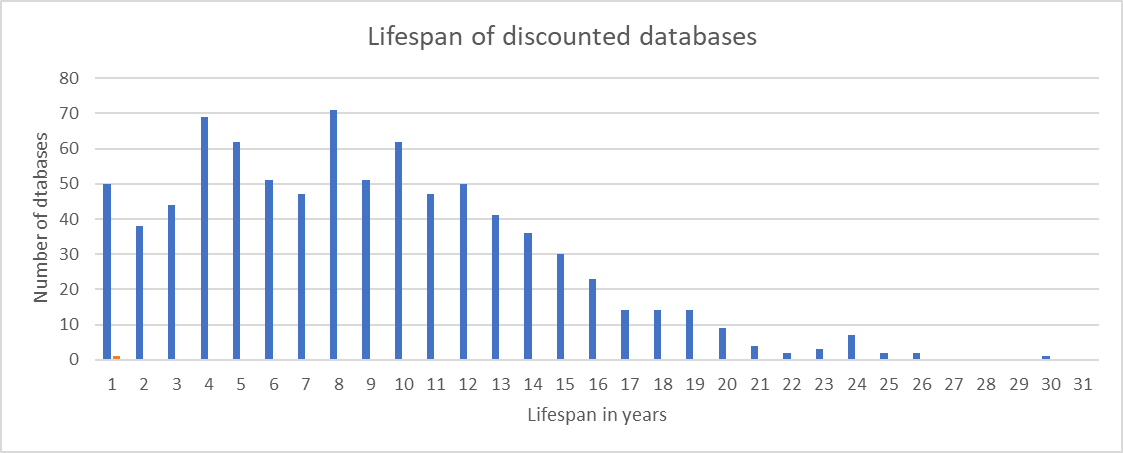
****

**2:What is the evolution over time of the number of databases discontinued each year?**

**I did also the same thing as I mentioned in the first question I just found the evolution over time of the number of databases which was discounted each year. As mentioned as an example that in 1991 was published 14 new databases, however, that year was not discounted database. The method I used, was the formula from the Excel program which helped me find all information correctly and save time. The most year when databases were discounted was 2022, with 113 databases.**

**3:NAR requires published databases to be maintained for at least 5 years after publication. For databases that are currently discontinued, how long did these databases were available?**

**The method I used to find information was the same as the previous question, I also, used an Excel program just to find the number of years of discounted databases in one word and their lifespan. It was clear, that around 70 databases, the lifespan was 8 years. But an interesting moment was that only one database's lifespan was 30 years.**

****

**Conclusion and future directions**

This analysis aimed to describe the evolution of new and discounted databases from 1991 to 2022, and their lifespan. In general, as we can see on the graph new databases were much more published rather than discounted, whereas last year new and discounted databases numbers were the same. Regarding their lifespan the much more databases which are 70 their lifespan was around 8 years but other we can see from the second graph that the more they live their number is decreased.

With this method would be better to find more information from different areas which can have an impact on scientific research.