# Introduction

The purpose of this research paper is to track the rate of published papers for Artificial Intelligence-related topics of research. The dataset was pulled from the arXiv.org search webpage using their API, following the documentation to implement the API for various search terms.

## Objectives

The objectives I would like to achieve for this paper are:

* Has the speed of research increased over the past 23 years for data-related subjects?
* Does it look like research is speeding up or slowing down?
* Are there any under-researched topics?

## Data Collection

To collect the research papers, I had to find an appropriate website to scrape from that was free to use, which is why arXiv.org was chosen. It should be noted, this overreliance on arXiv due to being free may have effects on the findings, since it's entirely possible there are a number of research papers missing from the arXiv.org database on the search term subjects. After figuring out which site I was going to use, I found their API and followed the documentation to implement it into my Python project. After getting the API to pull the data for one search term, I implemented the export to CSV to save the information to a data document. Once I verified that everything was working with one search term, I updated the file to include all the search terms related to Artificial Intelligence. After implementing the search terms, I pulled a total of 4515 scientific papers and their information to use in the graphs.

To make future updates of the dataset, I created a program to run. It has an option to update the core dataset by running the scrape again with the current search terms. There are also options to show three different graphs to view the data: a bar graph, a heat map, and a stacked area graph, with an option to save the graph to your computer.

## Search Terms Used:

[

"Artificial Intelligence",

"Machine Learning",

"Deep Learning",

"Neural Networks",

"Natural Language Processing",

"Computer Vision",

"Robotics",

"Data Mining",

"Data Science",

"Big Data",

"Data Analytics",

"Data Visualization",

"Data Engineering",

"Data Warehousing",

"Data Modeling",

"Data Architecture",

"Data Governance",

"Data Quality",

"Data Security",

"Data Storage",

"Data Integration",

"Data Migration",

"Data Lake",

"Data Lakehouse",

"Data Catalog",

"Data Dictionary",

"Data Lineage",

"Data Profiling",

"Data Wrangling",

"Data Cleansing",

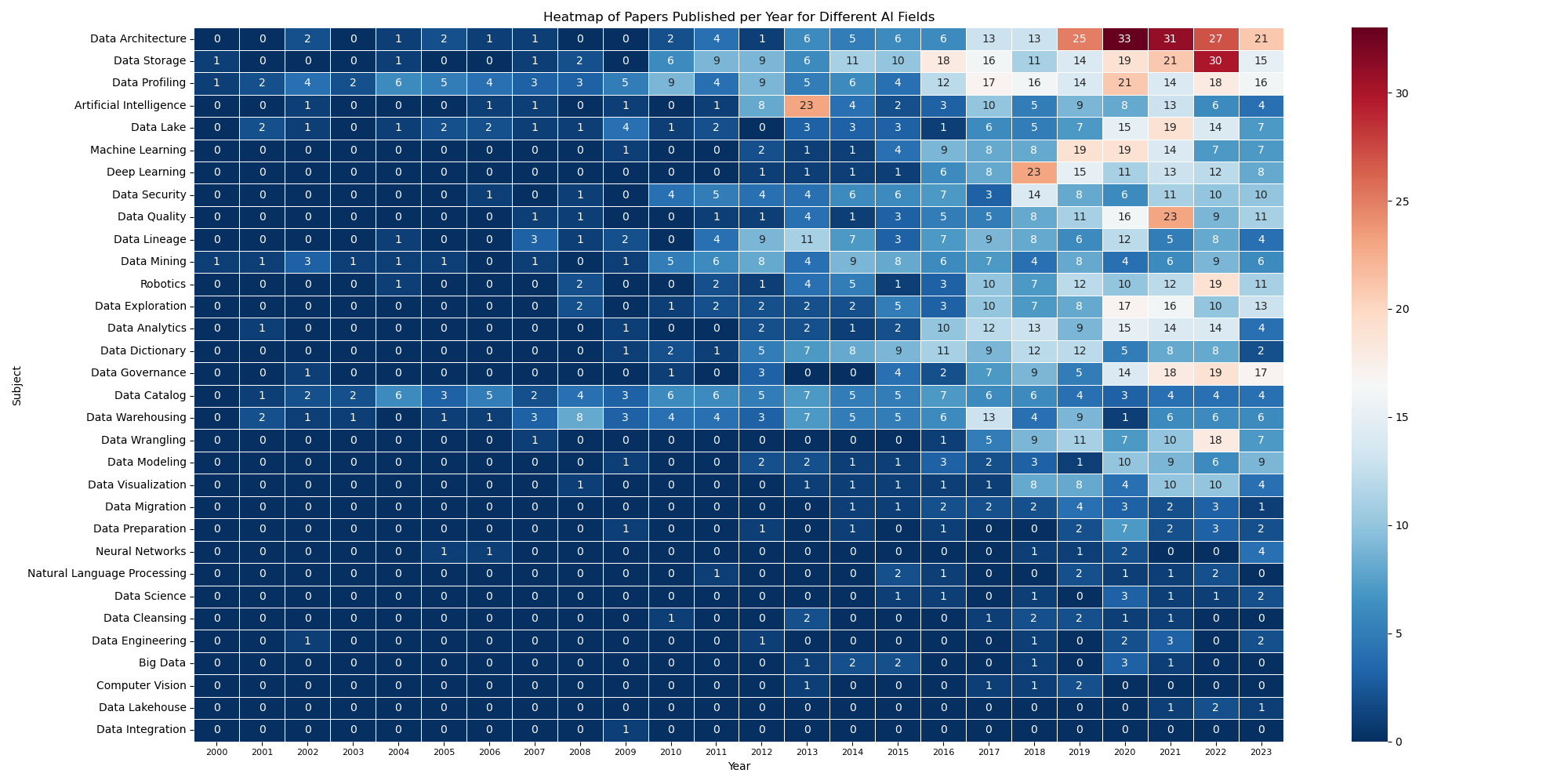
"Data Preparation",

"Data Exploration",

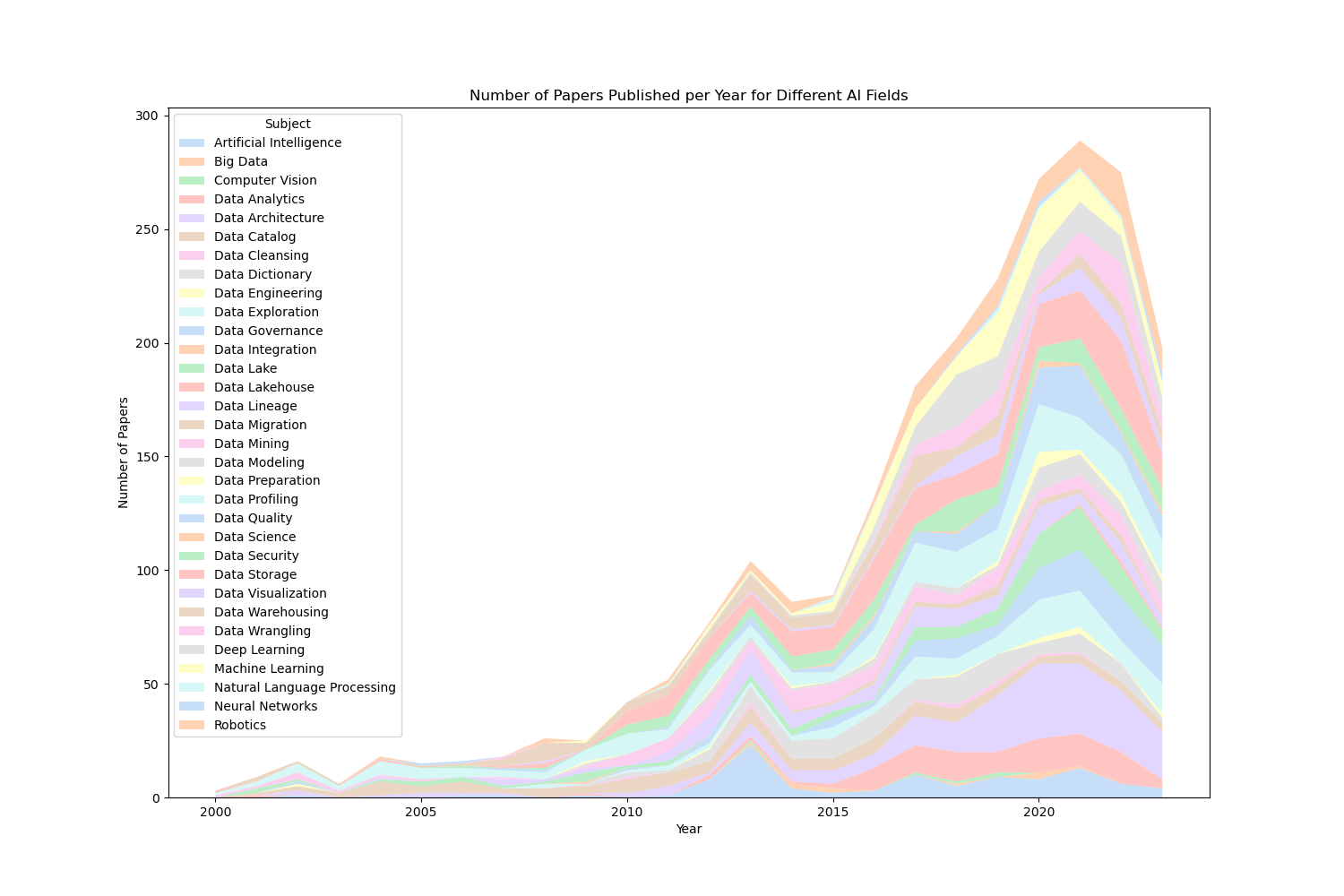
]

## Graphs

The first graph that I wanted to see was a basic bar graph. I wanted to have a clear visual representation of the number of papers researched over time, and this sort of gave me that. It's fairly easy to tell that interest in all data-related subjects has increased over the past 20 years, with AI-related products being researched in the past 10 years. There has not been a real decrease in the number of data-related papers being published even with the increase of AI-related papers being published, which indicates that the two subjects complement each other scientifically, and one's research leads into the others.



I thought that bringing in a heat map and showing the least-studied to most-studied subjects would more clearly display the more popular subjects. The heat map reinforced my hypothesis of the subjects working together. Once data scientists started publishing papers on innovative storage solutions, the increased interest in AI became apparent.



The last graph I created was a stacked area graph because I thought it might show the interest in the various subjects more clearly, and it looks nice.

## Interesting Findings

The heat map revealed a strong correlation between the development of data storage solutions and subsequent growth in AI research. This relationship may indicate that advancements in data handling paved the way for more complex AI algorithms. An unexpected finding was the steady interest in Data Wrangling despite the rise of more advanced topics, reflecting the ongoing importance of data preparation in various research fields. Certain areas, such as Data Lineage and Data Lakehouse, showed less research activity, suggesting potential areas for exploration and innovation.

## Conclusions

The analysis of the research papers over the past 23 years reveals consistent growth in AI and data-related fields. The complementary nature of data science and AI research reflects a symbiotic relationship that has fueled innovations in both domains. While there has been a surge in certain areas, some under-researched topics offer untapped potential for future exploration. It's recommended that academic and industry focus be directed towards these lesser-explored subjects to foster balanced growth in the field.