Research on Research on Research: Analyzing historical trends in statistical and computational research from the 1990s to modern day

Naren Prakash*
Department of Statistics and Data Science, University of California, Los Angeles

March 20, 2025

Abstract

This paper aims to analyze the changes in research paper output for different statistical and computational fields over the time period from the 1990s to modern day (2025). The paper also projects short term growth for recently emerging fields in an effort to predict the fields that will receive further resource funding and attention in the near future. The research papers used for this analysis are sourced from a dataset of papers from the pre-print journal arXiv.

Keywords: retrospective analysis, arXiv, publication analysis, forecasting, time series

*The author gratefully acknowledges Professor Lew for their encouragement and advice regarding the production of this paper.

1 Introduction

In recent years, with the fields of artifical intelligence and machine learning becoming important parts of the public lexicon and increasingly becoming involved in our day to day lives, we've seen firsthand large changes in statistical and computational research. With statistical methods increasingly becoming intertwined with computational principles, such as its integration with aspects of computer science, the future of statistics and computation appear to be one and the same. How does this current research landscape compare with that of the landscape a mere 30 years ago? This paper aims to analyze historical trends in statistical and computational research, as tracked by papers submitted to the online pre-print journal arXiv, in order to visualize the dramatic changes we've seen over the years and find any subfields growing in the present that could yet transform the landscape of the future. This analysis of historical trends will be conducted using a specific dataset available on Kaggle Mishra (2025).

2 Literature Review

Looking at themes in statistical and computational research is nothing new. For instance, Gelman & Vehtari (2021) analyzed the dominant statistical ideas of the past 50 years, suggesting inferential methods, computational algorithms, and data analysis have been the most impactful in the shifting of the research landscape. Smaller subsets of time have also been analyzed, with Jun et al. (2018) using Google Trends to track the growth of different subfields of research (with an emphasis on big data and application). This paper aims to look at a similar problem with a different lens, using publication outputs themselves as a way of analyzing changes in research focus and interest. In doing this, the paper aims to also obtain an indication of the subfields with increasing research interest in the short term that may lend itself to future publications. Evaluating the research trends of the future has often involved modeling itself, such as the hype cycle model Dedehayir & Steinert (2016). This model aims to track the life cycle of technological innovations. In a similar vein, this paper aims to use current and recent paper production output to indicate trends of the near future. As the methodology involves using the research paper pre-prints themselves, it may provide a clearer picture of specific publication interests and trends rather than topics and concepts in general.

3 Research Questions

- What statistical and computational fields have seen the largest increase in publications?
- How have the most published statistical and computational fields changed over time?
- What statistical and computational fields are projected to grow the most in the coming years?

4 Data

4.1 Data Description

The arXiv paper dataset consists of 136,238 observations and 10 columns. The 10 columns present in the data are: id, title, category, category code, published date, updated date, authors, first author, summary, and summary word count. Only the summary word count is a numeric variable. This data is scraped directly from arXiv, aiming to provide a representative sample of research published on the platform.

4.2 Data Pre-processing

The original dataset will have its variables converted to categorical variables for grouping and analytical purposes, with the exception of the summary word count due to its numeric nature. Following this, two selected lists of subtopics will be created for the purpose of data partitioning and separate analysis.

4.3 Data Partitioning

While acknowledging the connected nature of statistical and computational research in the present and future, this paper will partition the data into two halves. One half will be comprised of research deemed statistical in nature, and the other half will be comprised of research deemed as computational. This split in the data is done to narrow down the problem and allow for ease of analysis and interpretation of the results. Along with this, the data will also be subset in terms of time. In order to keep each yearly subset of papers as a representative sample of all research output for that year, the time range will be limited to exclude publication in 2025. This is done so that the resulting analysis will focus on comparison with full yearly samples of research data, rather than extrapolating from the research output in the year 2025 as of now.

5 Methods

5.1 Quantifying growth

For the purposes of this paper, growth will be represented by two metrics. Firstly, the simple percentage change from year to year for each subfield will be considered. In addition to this, the proportion of overall research represented by each subfield over time will also be used to evaluate growth in research interest and output.

5.2 Trend analysis and short-term prediction

Lastly, the metrics of growth as well as the partitioned data will be used to create a prediction model for the short term growth of research subfields. In correspondence with the data partitioning, separate models will be constructed for the statistical research and the computational research. The objective here is to produce a time series model for short term projections of growth in research interests and outputs.

5.3 Limitations

Focusing on primarily numerical data as a sign of growth indicates a relatively simple way of quantifying growth. In reality, growth is a more complex idea and could benefit from the use of paper content for text data processing to supplement the numerical figures of growth. This is a potential avenue of further exploration and work.

6 Results

6.1 Exploratory Data Analysis

6.1.1 Original Data

Before examining each partition of the original data for the purpose of directly answering the research questions, it is important to understand the context behind and the general appearance of the original dataset itself. For this reason, many plots were created to visualize parts of the data during the data pre-processing and data partitioning stages.

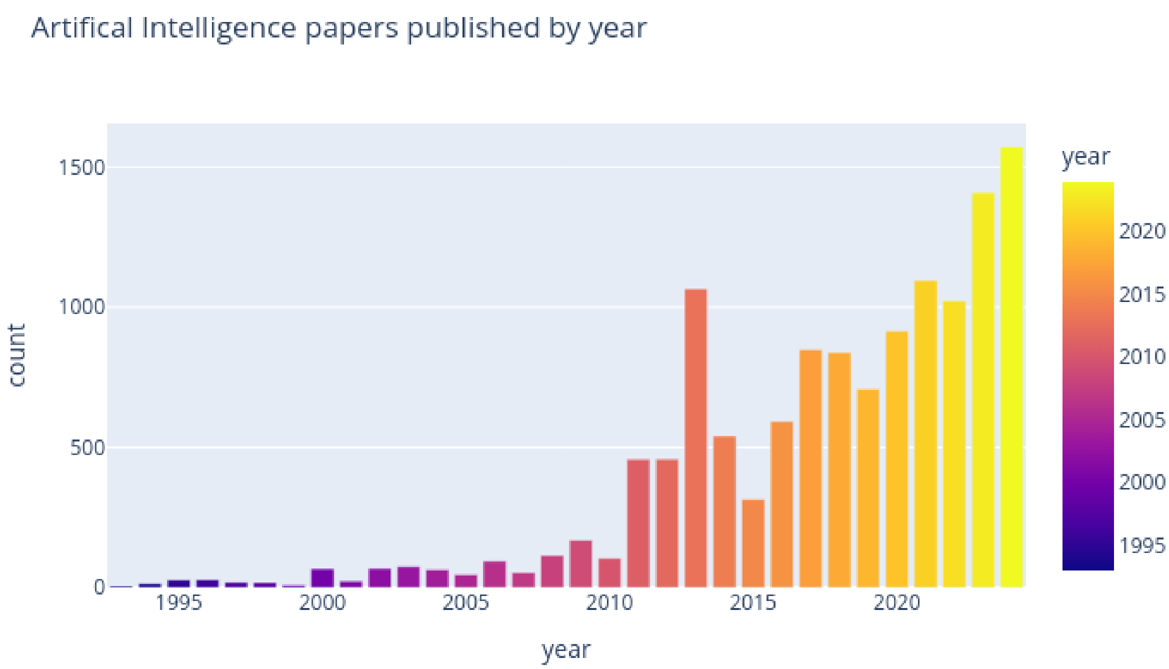

Figure 1: Artifical Intelligence Paper Output Plot

This initial plot of the research paper outputs for the category of Artifical Intelligence indicates several patterns we will continue to see in this data. Though Artifical Intelligence is often seen as a recent breakthrough that has only gotten larger year by year, we can see here that the yearly trend is far from consistent.

From here, we then further explore each of the partitioned datasets rather than focusing on the original alone, as these will be the foundation of our future analysis and modeling.

6.1.2 Statistical Data

The statistical data subset is comprised of the following topics: Data Analysis, Statistics and Probability, Machine Learning (Statistics), Methodology (Statistics), Computation (Statistics), Other Statistics, Applications (Statistics), and Statistics Theory. In order to explore the relative frequencies of paper output by these topics, we present an initial plot of this data.

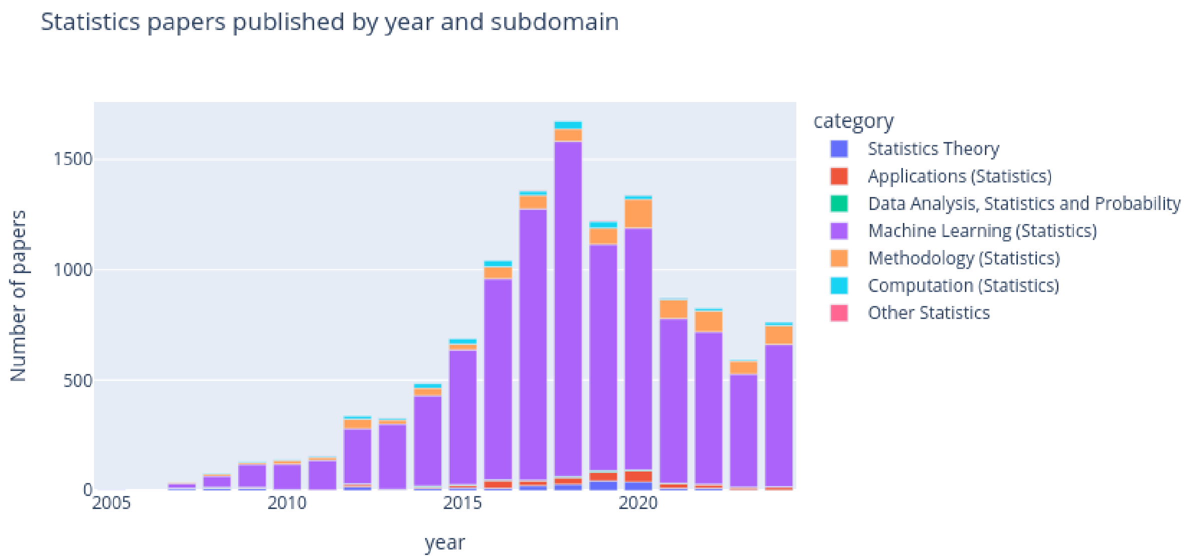

Figure 2: Statistical Data Paper Output Plot

From this initial plot it is immediately apparent that Machine Learning (Statistics) and Methodology appear to be the most popular subdomains, with other subdomains varying and not having a clear edge over each other. This provides us with a general impression of the data prior to delving into the specific yearly figures and relative frequencies.

6.1.3 Computational Data

The computational data subset is made up from the following topics: Artifical Intelligence; Computation and Language (Legacy category); Computation and Language (Natural Language Processing); Computer Vision and Pattern Recognition; Distributed, Parallel, and Cluster Computing; Neural and Evolutionary Computing; Computer Science and Game Theory; and Computational Physics. Once more, we create an initial plot to explore the relative frequencies of paper output by subfield.

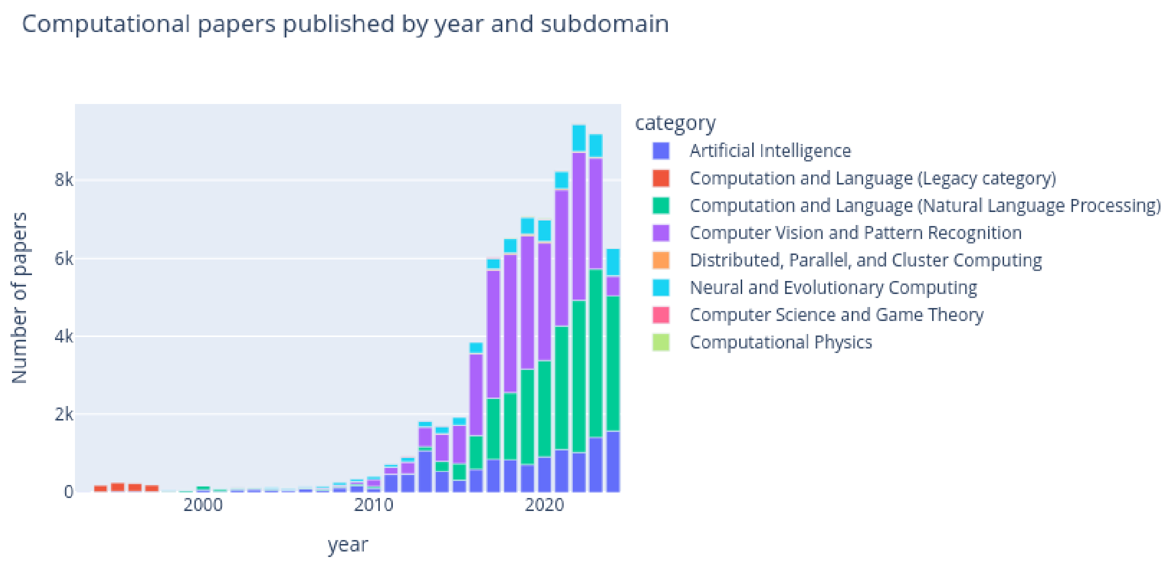

Figure 3: Computational Data Paper Output Plot

Once more, we can see a few subtopics are the most popular visually. In this case, the Computer Vision and Pattern Recognition, Computation and Language (Natural Language Processing), and the Artificial Intelligence subtopics have the highest proportions of paper outputs, especially in more recent years. This is not unexpected, considering the rise of LLMs and the popularity of Artifical Intelligence in modern society.

6.2 Statistical Data Analysis

6.2.1 Fields with the largest publication increase

In order to quantify the largest publication increase for the statistical fields, two methods were employed. One method involved looking at the average growth (in percentages) year over year for each subtopic. The other involved comparing the raw paper output from the first year of comparison for each subtopic to its more recent paper output total. The results of both forms of analysis are shown in the following plots.

Figure 4: Statistical Data Average Growth and Ratio Plots

Based on the two plots above, we can see some patterns as many of the subfields retain the same order in both the paper output ratio and the average yearly percentage growth. The apparent exception here is Statistics Theory, which despite having an extremely high percentage growth year by year, is among the lowest for growth in terms of raw totals. The reason for this, after further investigation, is due to the volatile nature of Statistics Theory publications on a yearly basis, as many years would have little published but be followed by a year with a larger output. This inconsistency produced a very large average percentage growth yet an overall small increase in raw output. From considering these graphs, we can see that the fields with the most growth in this time period are Machine Learning, Methodology, and Applications.

6.2.2 Changes in the top fields over the years

In order to analyze publication output on a yearly basis, the top 4 subfields in paper output for a given year were calculated and plotted. The visualization produced from this is below.

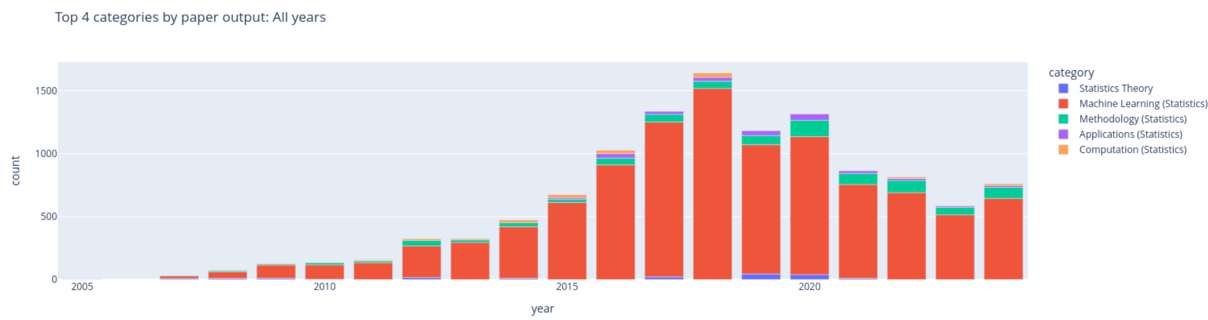

Figure 5: Statistical Data Top 4 Subfields Yearly

From this analysis, we can see that the most published subfields remained consistent over the years. Machine Learning, Methodology, Applications, and Computation often comprised the top 4 with Statistics Theory also in the mix. Overall, the top 3 fields remained Machine Learning, Methodology, and Application / Computation with some variance between the presence of Application and Computation in any individual year.

6.2.3 Projected growth of statistical fields

In order to tackle the short term prediction of paper outputs, we want to look at recent trends in publications. An initial look at this is through limiting the time frame. Here we restrict the data from 2015-2024 and look at the average percentage growth in publication output.

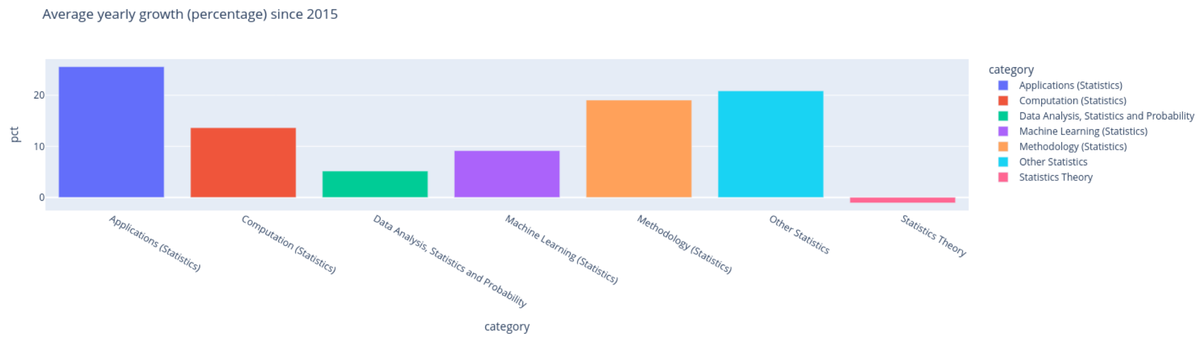

Figure 6: Statistical Data Recent Publication Growth

From this, we can see that publication growth is actually highest in Other Statistics, Methodology, and Applications. Somewhat surprisingly, the publication average growth of the Machine Learning subfield is relatively low. In order to fully capture the historical trends and use them for future predictions, we want to utilize a time series approach.

6.2.4 Creating the time series model

In order to create a suitable model for the future time series prediction of paper outputs by subfield, we first need to understand the relationships between subfields. We first look at the correlation between each subfield to understand if the patterns in paper output are dependent on other subfields.

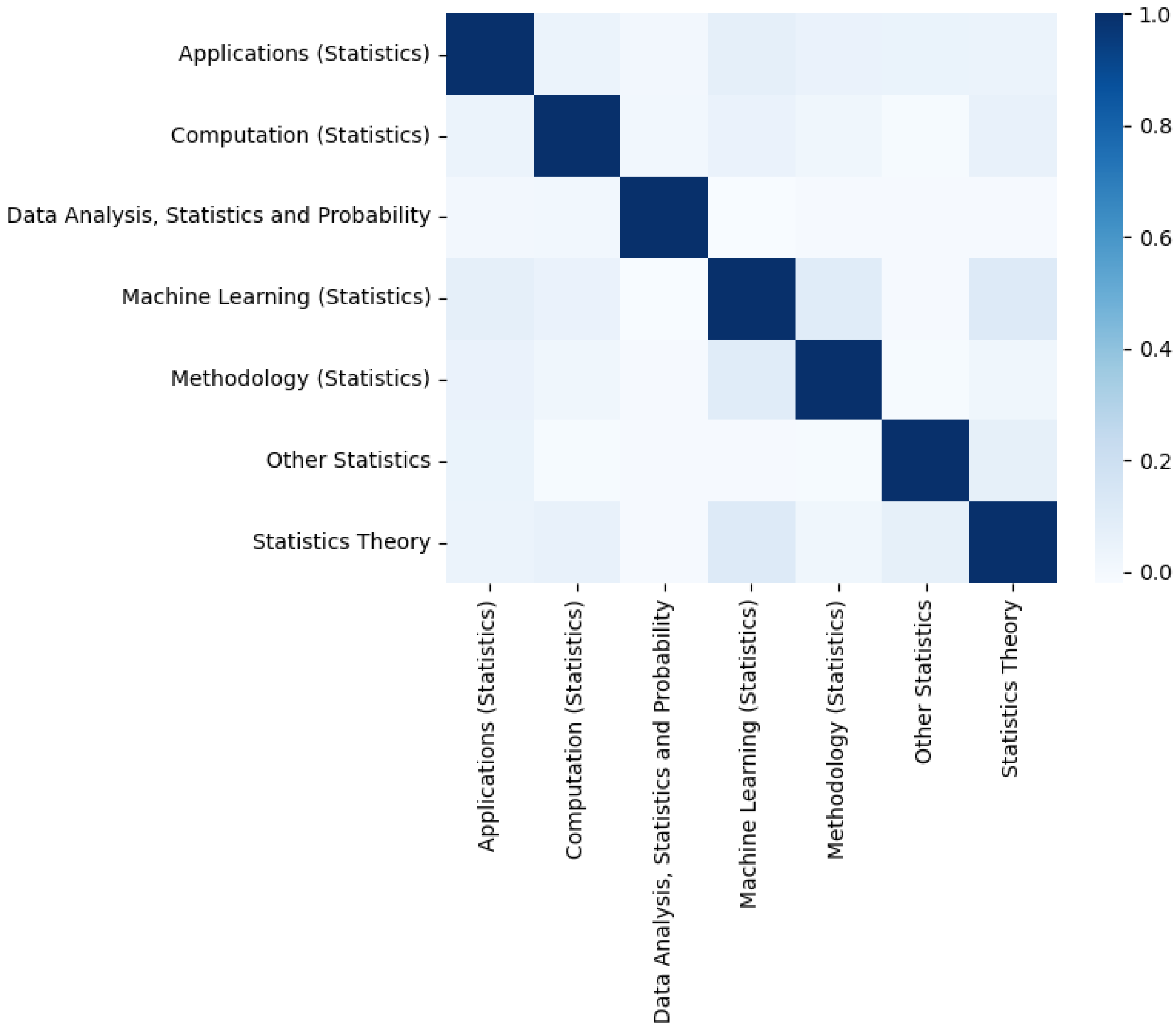

Figure 7: Statistical Data Correlation Heatmap

From this heatmap, we can conclude that each subfield is largely independent of other subfields, with the correlation values across subfields being fairly low. As a result, we choose a model that predicts future values based on the previous values of each individual subfield. Essentially, we treat each subfield as an individual time series.

The model itself is constructed using the skforecast package from Amat Rodrigo & Escobar Ortiz (2025). We use a LightGBM regressor to create a ForecasterAutoregMultiSeries object from the aforementioned skforecast library. From previous fitting of a Vector Autoregression Model, we specify 9 lag periods for the object.

We create a training set from the statistical data of the first 70% of observations, with the remaining 30% serving as the test set. After training and predicting on our test set, we see a mean absolute error of about 0.5913. Finally, we use all the original data as the training data for a model object identical to that of our original and supply the future time frame of the first 6 months of 2025. From those predictions, the following trends are seen.

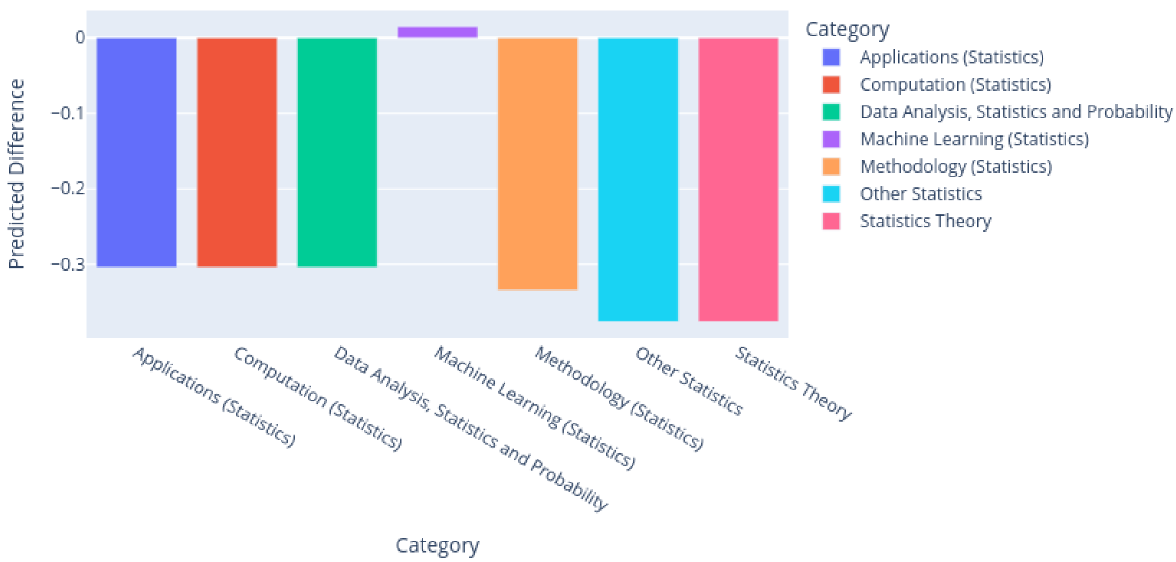

Figure 8: Statistical Data Predicted Differences

As seen here, the time series model predicts a decline in paper output for each subfield over the first 6 months of 2025, with the exception being in Machine Learning. Further inspection into the changes of the percentages over time show an inconsistent rate of change, with many ups and downs.

This further analysis shows a trend commonly seen in the predictions of each of the individual subfields. Considering the nature of the statistical data from our observations, it seems apt that the predictions themselves also reflect the inconsistency seen in paper output. Ultimately, from these predictions, in the short term Machine Learning is expected to see the most paper output.

6.3 Computational Data Analysis

6.3.1 Fields with the largest publication increase

Once more, we use the same two methods to quantify this. We look at average growth (in percentages) year over year for each subtopic before then comparing the raw paper output from the first year of comparison for each subtopic to its more recent paper output total. The two plots below show us the results.

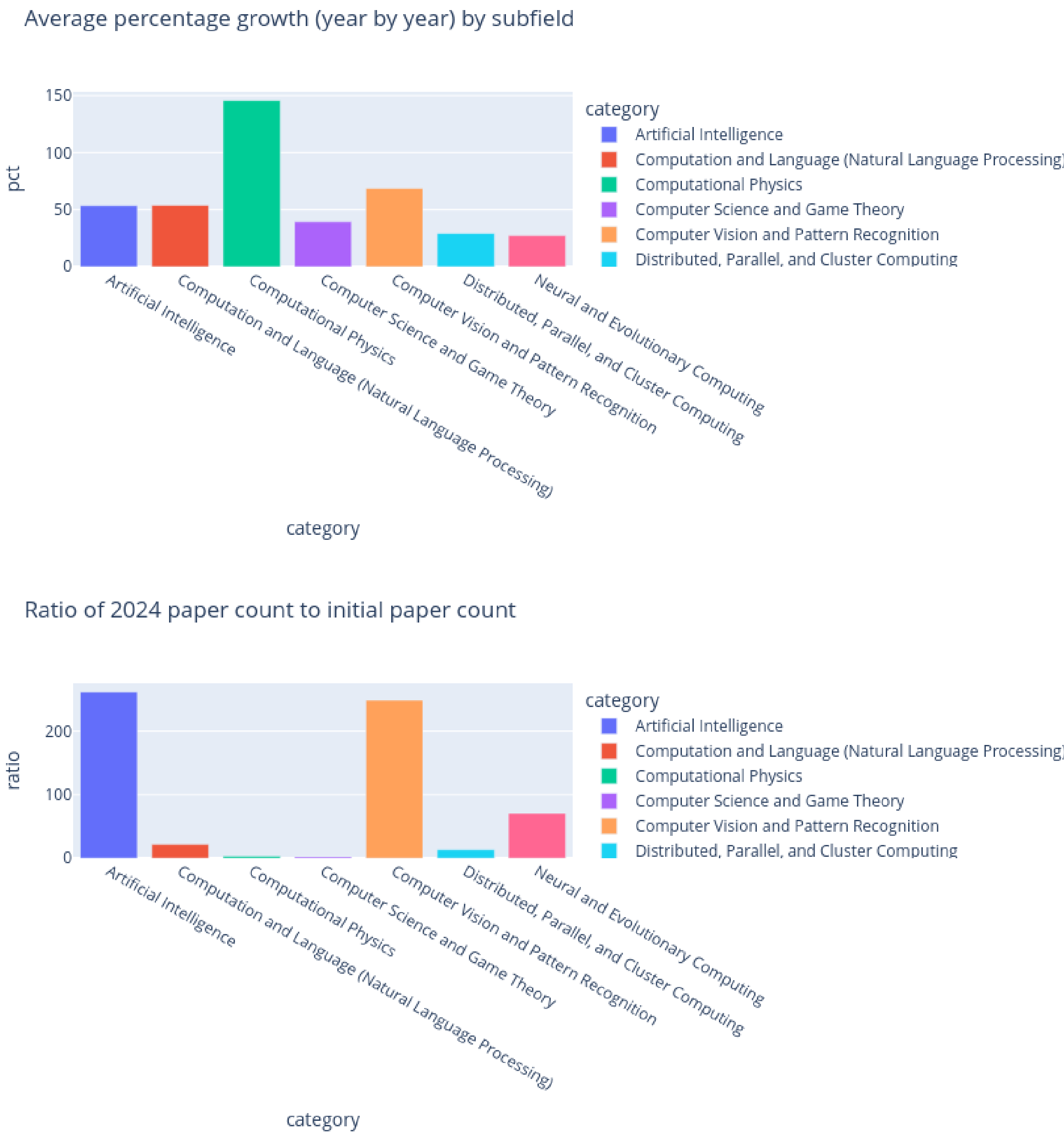

Figure 9: Computational Data Average Growth and Ratio Plots

Similar to Statistics Theory in our statistical data yearly growth plot, Computational Physics has the highest average growth but a small ratio, indicating that it is the most volatile subfield with publication outputs varying between relatively small and relatively large. From these plots, we can see that Artifical Intelligence, Computer Vision, and Neural and Evolutionary Computing seem to be the largest subfields in terms of growth.

6.3.2 Changes in the top fields over the years

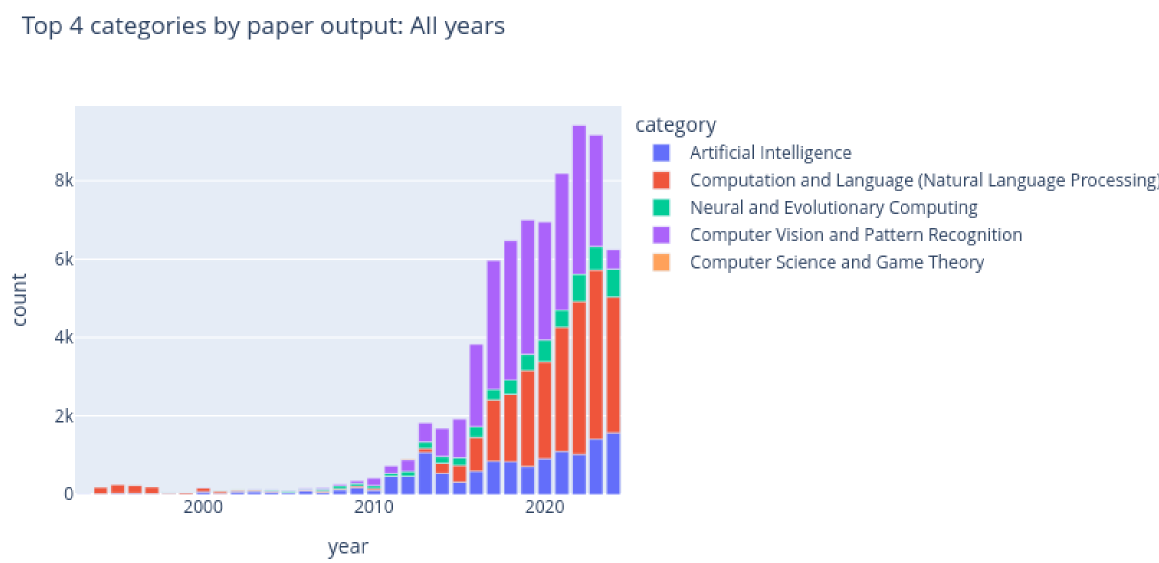

Figure 10: Computational Data Top 4 Subfields Yearly

From this, we can see that the subfields comprising the majority of computational papers have been Artifical Intelligence, Computation and Language (Natural Language Processing), and Computer Vision and Pattern Recognition. Perhaps surprisingly, the Artifical Intelligence subfield has been one of the larger subfields for much longer than its stay in the public lexicon. Despite it only recently becoming known to the general public, the research output indicates it has been an emerging and popular publication subfield for quite a while.

6.3.3 Projected growth of computational fields

Using a subset from 2015 to the present, the average growth of each subfield is extremely similar to that displayed in Figure 9. Thus, to gain additional insight into the projected growth of the computational fields in question, we turn toward the time series model once again.

6.3.4 Creating the time series model

We again choose a model that predicts future values based on the previous values of each individual subfield due to a weak correlation between the subfields.

The model itself is constructed using the skforecast package from Amat Rodrigo & Escobar Ortiz (2025). We use a LightGBM regressor to create a ForecasterAutoregMultiSeries object from the aforementioned skforecast library. From previous fitting of a Vector Autoregression Model, we specify 13 lag periods for this object.

We create a training set from the statistical data of the first 70% of observations, with the remaining 30% serving as the test set. After training and predicting on our test set, we see a mean absolute error of about 2.0295. Notably, we can see that our model is less accurate in predicting the differences in computational fields as compared to the similar model used for the statistical data. We then use all the original data as the training data for a model object identical to that of our original and supply the future time frame of the first 6 months of 2025. The predicted trends are seen below.

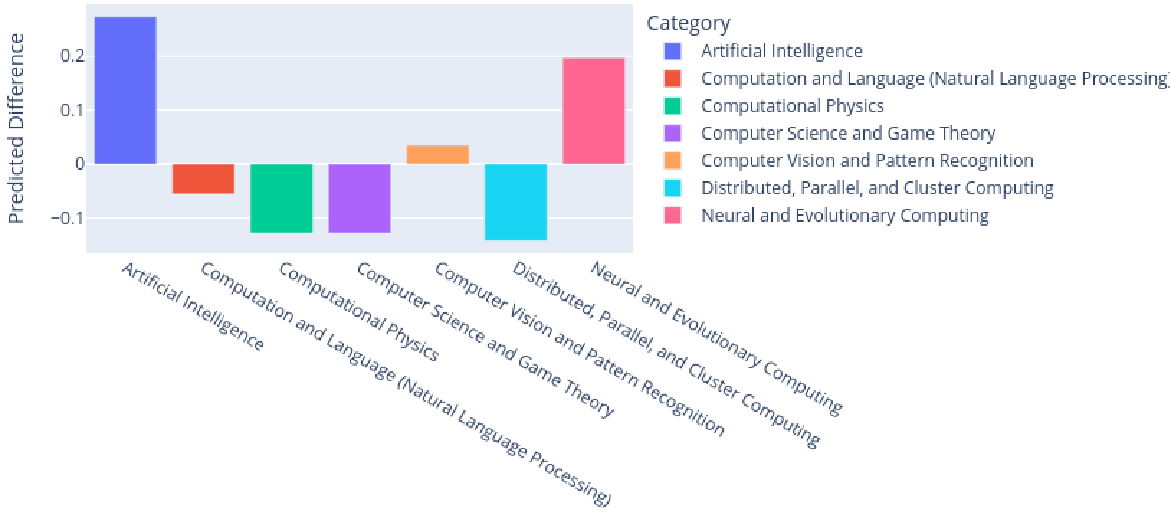

Figure 11: Computational Data Predicted Differences

From our time series model, we see a projected growth in publication output for Artifical Intelligence, Neural and Evolutionary Computing, and a smaller growth for Computer Vision and Pattern Recognition. This model projects more publication growth than the comparable statistical data model, which only projected growth for one subfield. However, it is important to note that the mean absolute error for this model is larger, and as such the model predictions may be less accurate.

7 Conclusion

From looking at the various statistical and computational subfields from the 1990s to the present day, we can see that the popularity of research and paper output for subfields remains fairly consistent. While the actual outputs, as in the raw paper totals, vary a lot year by year, the top subfields tend to keep their places. For statistical data that was

Machine Learning, Methodology, Applications, and Computation. For computational data, that was Computation and Language (Natural Language Processing), Artificial Intelligence, and Computer Vision and Pattern Recognition. With that being said, there are several subfields our models predicted to grow in the near future. The computational time series model projected a growth in the Neural and Evolutionary Computing subfield, a subfield not often in the top subfields in terms of paper output. It also projected continued growth for the field of Artifical Intelligence and to a lesser extent Computer Vision and Pattern Recongition. On the other hand, the statistical time series model projected continued growth for the Machine Learning subfield, while projecting short term decreases in the output of other subfields. It is important to note that the differences in predicted growth between the statistical and computational models may be affected by the source of publications used in this analysis, as arXiv hosts more computational papers than statistical papers. As a result, future analysis may want to incorporate several different sources and/or using oversampling techniques to ensure comparable analysis for each subtopic. These insights aim to provide a framework for future retrospective analysis, and to capture the trends and expectations of research from a specific point in time.

8 Source Code and Data

All code used in the production of this analysis as well as the original dataset from arXiV can be found at this repository.

References

Amat Rodrigo, J. & Escobar Ortiz, J. (2025), 'skforecast'.

URL: *https://skforecast.org/*

Dedehayir, O. & Steinert, M. (2016), 'The hype cycle model: A review and future directions', *Technological Forecasting and Social Change* **108**, 28–41.

URL: *https://www.sciencedirect.com/science/article/pii/S0040162516300270*

Gelman, A. & Vehtari, A. (2021), 'What are the Most Important Statistical Ideas of the Past 50 Years?', *Journal of the American Statistical Association* **116**(536), 2087–2097.

URL: *https://doi.org/10.1080/01621459.2021.1938081*

Jun, S.-P., Yoo, H. S. & Choi, S. (2018), 'Ten years of research change using google trends: From the perspective of big data utilizations and applications', *Technological Forecasting and Social Change* **130**, 69–87.

URL: *https://www.sciencedirect.com/science/article/pii/S0040162517315536*

Mishra, S. (2025), 'arxiv scientific research papers dataset', Kaggle.

URL: *https://www.kaggle.com/datasets/sumitm004/arxiv-scientific-research-papers-dataset*