# Predicting Research Trends From Arxiv

No Author Given

No Institute Given

## 1 Introduction

Knowing trends in research has been a long-standing dream of scientists. Projects on popular research topics often lead to higher acceptance rates at conferences and journals, as well as funding application approvals. Further, knowing future research trends immediately also has implications for society as a whole, because these trends will most likely directly affect the labor market, technological orientation and bias, consumer end products, as well as cultural metaphors and definitions of the human identity (this is even more true for fields such as artificial intelligence, as we focus on here). However, with the accelerating number of papers made available each year, it becomes ever more difficult to digest the incoming information and thereby identify topics that will have long-term scientific impact. We have developed an automatic system whose goal is to uncover important research trends, and, therefore, aims at helping researchers better plan their academic endeavors.

To illustrate our system, we crawl papers published in the Machine Learning (cs.LG) and natural language processing (cs.CL) categories of Arxiv, with information about how often they were cited. In this dataset, we identify influential papers and categorize them by hand and automatically. Using Arxiv papers for our exploration appears promising, because Arxiv is a very popular pre-print (and post-print) server for scientific publications, whose impact has, moreover, considerably increased over the last few years.

## 2 Data

We created two datsets: One with papers from the machine learning (cs.LG) category on Arxiv and one with papers from computation and language (cs.CL). We focus on these two prominent subfields of artificial intelligence because they appear concurrently particularly dynamic, with drastic changes and performance improvements witnessed each year, mainly due to the impact of artificial neural network (aka deep learning) models. The data includes papers with their title, abstract and authors. We also harvested citation information for the papers from semanticscholar. Our crawled papers date between mid-2017 and mid-2018. We crawled roughly 2k papers from cs.CL and 5k papers from cs.LG.

## 3 Identifying Influential Papers

The simplest measure for a paper's influence is the number of citations it has received. But plain citation counts may be misleading. They may vary depending on the research field and the date of the publication. Instead, citation counts can be normalized by comparing only papers in the same research fields and adjusting citation count scores by the paper's age. This is the idea of the *z-score* approach suggested by Newman [2, 3]. This is calculated by subtracting the mean citation count of papers within a time window from the citation count of a paper and dividing it by the standard deviation. With this method, Newman [3] does not only find papers that have a high number of citations because they were published earlier than other papers on the same topic, but also papers with only a few citations. To illustrate, after 5 years, the 50 papers with the initially highest z-scores received 15 times as many citations than the average paper in a randomly drawn control group *that started with the same number of citations*. Thus, Newman argues (and empirically demonstrates for his sample) that the z-score can indeed identify short- and mid-term research trends.

To find influential papers, we calculated the z-score for the papers in our datasets, using a time-window of $\pm 10$ days. We ignored papers with less than 4 citations, because we deemed such small numbers as unreliable.

## 4 Evaluation

For both cs.CL and cs.LG, we identify the top-100 papers according to their z-score. We then (manually, initially) cluster these papers into groups in order to detect more general patterns: which areas will dominate in the upcoming years? In the following analysis, we focus on cs.LG, for space reasons.

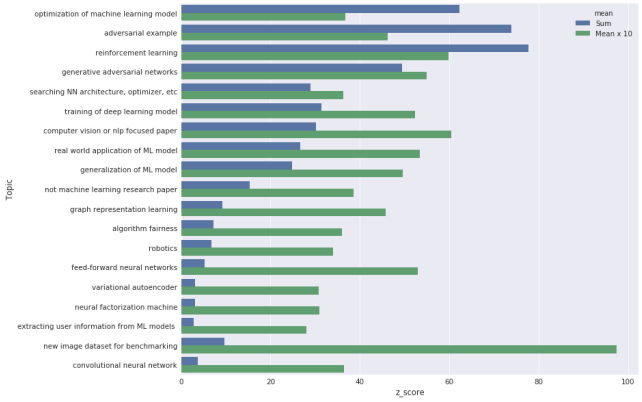


**Fig. 1.** Topics and corresponding z-score for cs.LG. Topics ranked by $n_C$ score. The score $s_C$ is on its original scale, and $m_C$ is multiplied with 10.

We assign cs.LG papers into 19 categories/topics. We then introduce three statistical quantities to measure the importance of each category:

$$n_C = \sum_{p \in C} 1, \quad s_C = \sum_{p \in C} \text{z-score}(p), \quad m_C = \frac{1}{|C|} \sum_{p \in C} \text{z-score}(p).$$

for each category $C = \{p_1, p_2, \ldots\}$ consisting of papers $p_1, p_2, \ldots$. The value $n_C$ gives the number of papers in each category. In contrast to $n_C$, the value $s_C$ does not only count the number of papers per category, but also weights them by their z-score. The value $s_C$ thus favors categories that have more top papers. The value $s_C$ favors categories whose average z-score is high; therefore, categories with very few papers may also rank highly.

*Results* Given the current hype and success of deep learning, it is not surprising that most categories deal with different aspects of deep learning, see Fig. 1.

The categories "optimization [...]", "adversarial examples" and "reinforcement learning" rank highest according to $n_C$ but the ranking is reversed according to $s_C$. As the biggest category, the majority of papers in "optimization [...]" focuses on the optimization process of deep learning models, like their saddle points, local minima, and properties of gradient descent. While optimization is a well-established field in machine learning, the fact that adversarial examples (how to attack deep learning models) and reinforcement learning rank higher according to $s_C$ may indicate that the field will be dominated more heavily by these two categories in the future.

"generative adversarial networks" has rank 4 both according to $n_C$ and $s_C$. This area of machine learning deals both with *generating data*, rather than classifying it as in traditional machine learning, and with fooling a machine learning system, as does "adversarial examples", but this time in order to render the system more robust.

## 5 Concluding remarks

We have developed a system to rank Arxiv papers according to their z-score, in order to detect short- and mid-term research trends. Our manual evaluation and clustering of ranked papers indicated some potentially interesting paradigm shifts in machine learning in the upcoming years. It appears that the field will be heavily concerned with fooling and attacking deep learning systems, as well as defending against such attacks, with data generation (instead of classification, as in traditional machine learning), and with reinforcement learning, which can be seen as an endeavor towards proper artificial intelligence where artificial agents learn in a way similar to humans.

Further automatization of our methodology is necessary and has partly already been realized: we want to assign topics to ranked Arxiv papers automatically rather than manually. We will use neural text embeddings [1] and machine learning for this. In the same vein or alternatively, clustering of Arxiv papers

must be done automatically. As a secondary signal next to the z-score ranks, we want to predict citation counts from a model that incorporates the text and the meta-data (such as authors and publication venue) of a paper. Finally, we want to conduct a diachronic study: how does the ranking of high z-score papers/topics change over time? This may be considered as an additional layer of estimating and quantifying trends and their shifts.

## References

1. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). pp. 681–691. Association for Computational Linguistics (2017), http://www.aclweb.org/anthology/D17-1071
2. Newman, M.E.: The first-mover advantage in scientific publication. EPL (Europhysics Letters) **86**(6), 68001 (2009)
3. Newman, M.: Prediction of highly cited papers. EPL (Europhysics Letters) **105**(2), 28002 (2014)