XSEDE Scientific Impact Metrics for the Yearly Report between May 1, 2020 to Apr 30, 2021 and the Period between Feb 1, 2021 - Apr 30, 2021

Fugang Wang, Gregor von Laszewski
Apr 30, 2021
Technical Report, Indiana University, Bloomington, IN

1.1. Scientific Impact Metrics (SIM)

This appendix presents the current Scientific Impact Metrics data as of April 30 of year 2021 as well as updated analysis based on a new data set. This is part of the *XD Metrics Service (XMS) (*formerly *NSF Technology Audit Service (TAS))* effort.

1.1.1.Summary Impact Metrics for XSEDE

Table SIM-1 shows the essential scientific summary impact metrics as of April 30 of year 2021. The increasing values for each metric are listed in the table indicating the changes during the last quarter. By calculating such metrics periodically we can show the trends, as depicted in Figure SIM-2 and Figure SIM-3. Both show steadily increasing trends.

Table SIM-1: Overall Scientific Impact Metrics Data

	Number of externally verified unique publications*	i10-index (Number of publications cited at least 10 times)	Overall citation count*	h-index	g-index
Since 2005 (TG+XD)	19,939	11,756	745,107	281	508
Since 2011 (XD)	17,375	9,699	531,764	233	405
Change since last quarter (TG+XD)	+499	+422	+35,766	+6	+12
Change since last year (TG+XD)	+2,098	+1,703	+139,172	+26	+48
Change since last quarter (XD)	+497	+415	+32,312	+9	+14
Change from a year ago (XD)	+2,080	+1,676	+124,521	+33	+56
* Data updated as of April 30th, 2021 Note: The quarterly comparison was against Jan 31, 2021. The yearly comparison was against April 30, 2020.					

1.1.2. Historical Trend

Figure SIM-2 and SIM-3 show the increasing quarterly trend regarding publications, citations, and other impact metrics such as H-Index and G-Index. Both suggest the increasing impact of XSEDE during the past years, based on verified unique publication count; citation count; H-index and G-index.

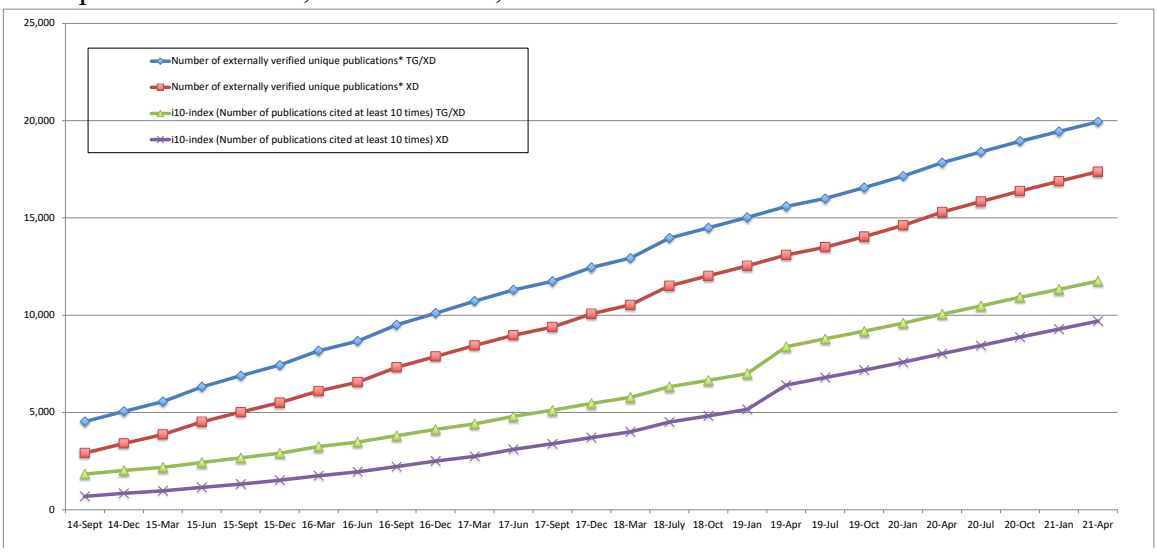

Figure SIM-2. Counts of all externally verified publications for TG/XD (since 2005) and XD (since 2011) and of those being cited at least 10 times (i10-index).

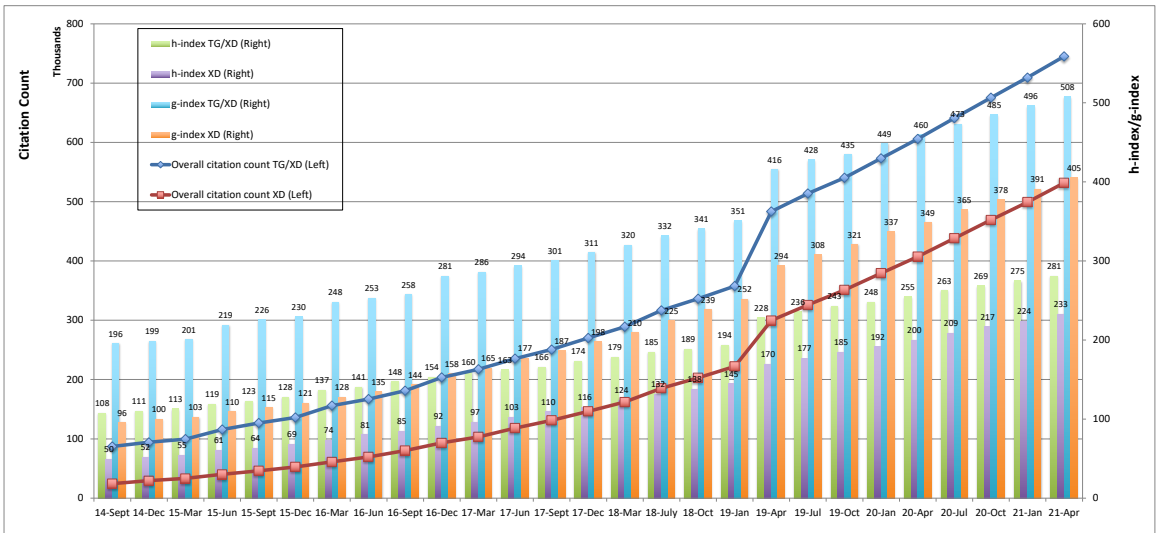

Figure SIM-3. Accumulated citation count (line, left axis) as well as h-index and g-index metrics (bar, right axis) for TG/XD (since 2005) and XD (since 2011).

1.1.3. Top FOS trend analysis based on historical data

We have conducted some new historical trend analysis based on the data we have been collecting. Figure *SIM-4* and *SIM-5* show the Top 20 Field of Study by number of publications and citation count for the past 5 years. The data point for each year was the snapshot at end of that year.

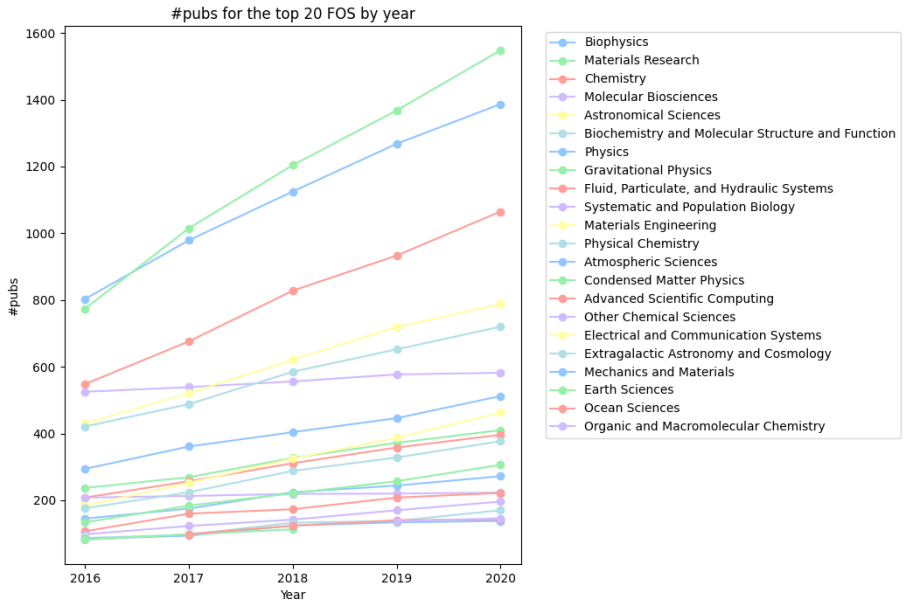

Figure SIM-4. Top 20 Field of Study with the greatest number of publications by year

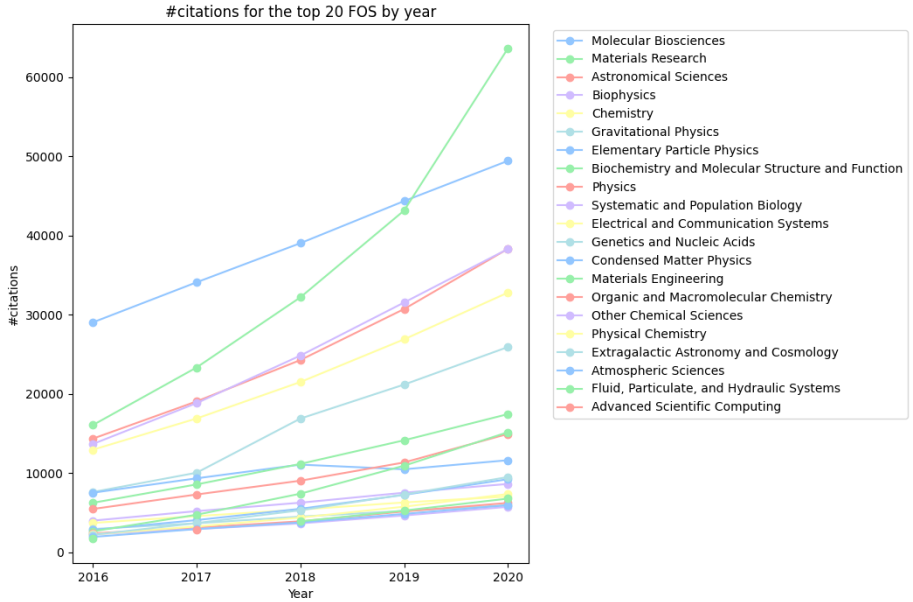

Figure SIM-5. Top 20 Field of Study with most citations received by year

Figure *SIM-4* shows the new publication increasing trend for different FOS. E.g., We see consistent increasing for the top 3 FOS, Biophysics, Material Science, and Chemistry, while the Molecular Bioscience almost stayed flat. In Figure *SIM-5* we see that the Material Science has gained significantly on the citations, which seems reflecting well that that has been a hot research field recently.

1.1.4. Updated analysis with the new Semantic Scholar dataset

We have been exploring other data sources to be integrated into our data integration and analytical framework. Recently we have assessed the use of Semantic Scholar, a public publication dataset, in the use of our analysis. This dataset has over 186 million publication records (as of end of February 2021) which was published as a series of compressed text files. We have developed the data injection and integration workflow (Figure *SIM-6*) and was able to use this dataset to carry on updated analysis of peers comparison and Field Weighted Citation Index (FWCI) analysis. We are showing the results in the following subsections.

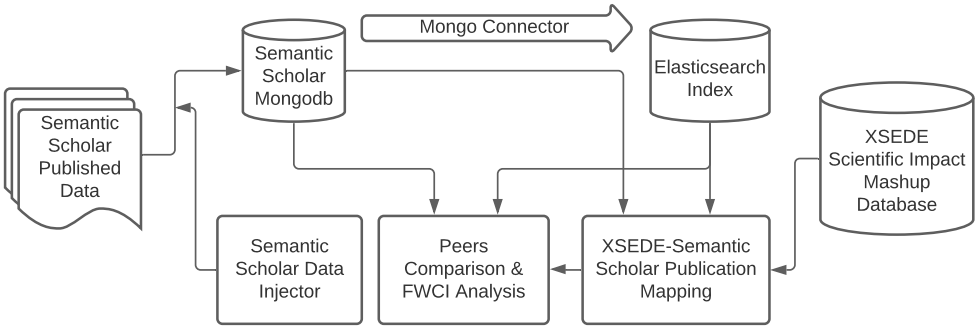

Figure SIM-6. Semantic Scholar dataset integration and analysis workflow

1.1.4.1. Peers comparison results

We had introduced the peers comparison as a way to evaluate the scientific impact of a group of publications for a virtual organization and have published some analysis results for XSEDE before. With the access to the new dataset we were able to conduct an updated peers comparison analysis. In this updated study we have compared 12461 XSEDE publications from 231 publication venues against their peers. Each publication venue has at least 10 XSEDE publications published on them to avoid the possible statistically insignificant results if too few publications appear in a publication. Figure *SIM-7* shows the average percentile ranking score for each publication venue, sorted in descending order. The horizontal red line is the baseline of score 50. It clearly shows that for the majority of the 231 publication venues the XSEDE publications had received more citations compared to an average peer.

Figure *SIM-8* shows the histogram of the averaged percentile ranking score for each publication venue. The distribution skews towards the right side suggesting that the XSEDE publications, in general, received more citations than the average peers.

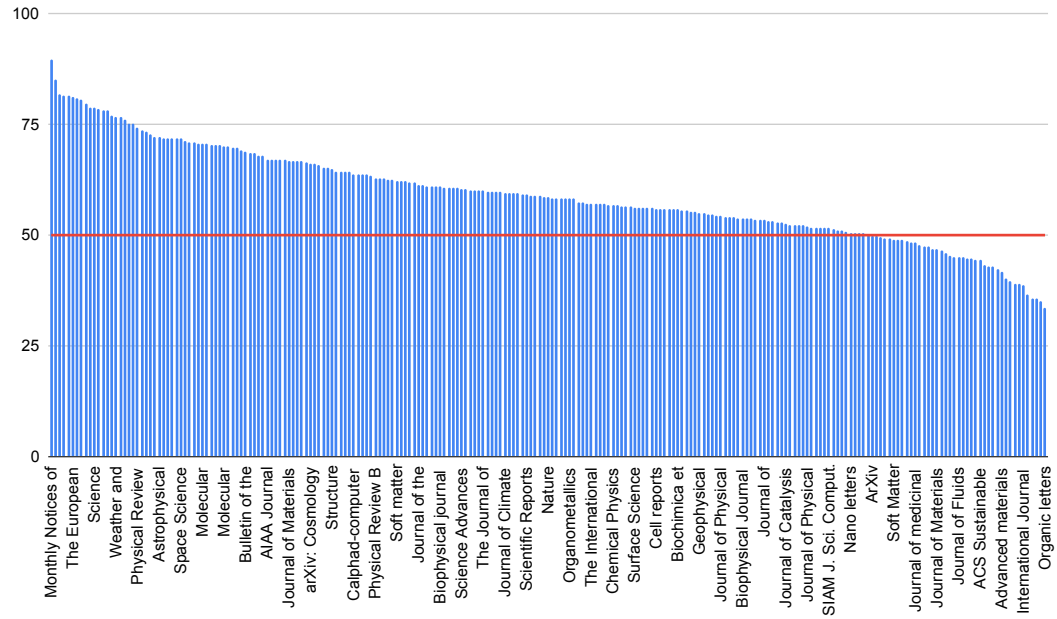

Figure SIM-7. Average percentile ranking score of XSEDE publications by journal

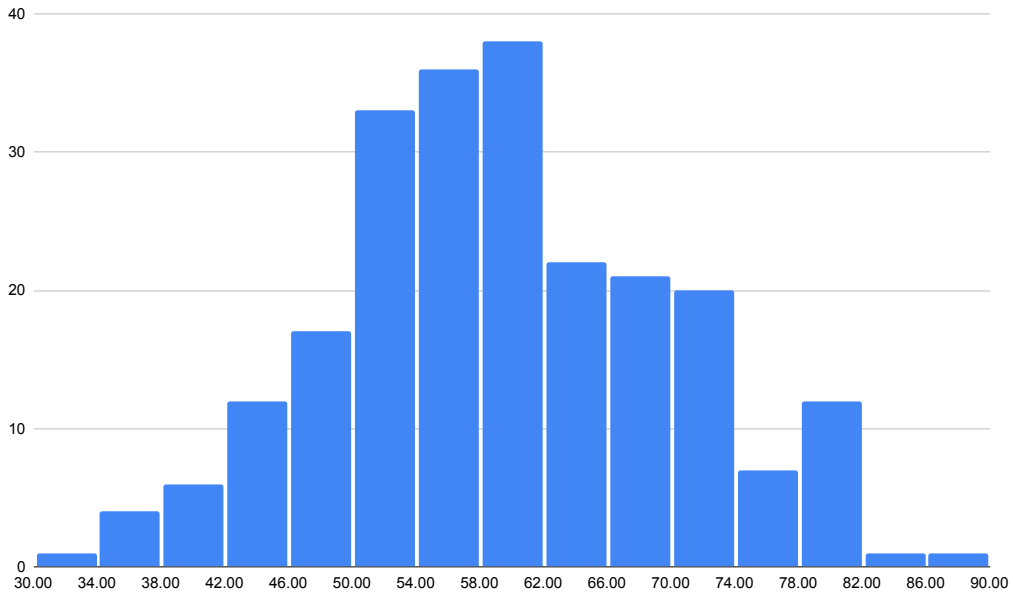

Figure SIM-8. Distribution of the percentile ranking scores of XSEDE publications

When grouping the percentile ranking scores by the publication's identified field of study, we obtain the result as shown in Figure *SIM-9*. Here we also discarded the data points for a few fields of study due to the limited number (less than 10) of XSEDE publications. The results show that the average percentile ranking score is above 50 for most FOS.

In one of our previous studies we have conducted the peers comparison analysis based on the data of more than 5000 publications from about 120 publication venues.

Now with more data available over a longer period of time we could verify the trend due to its similar results. Hence, we assess that the studies' results validate each other and show the consistent performance of XSEDE publications.

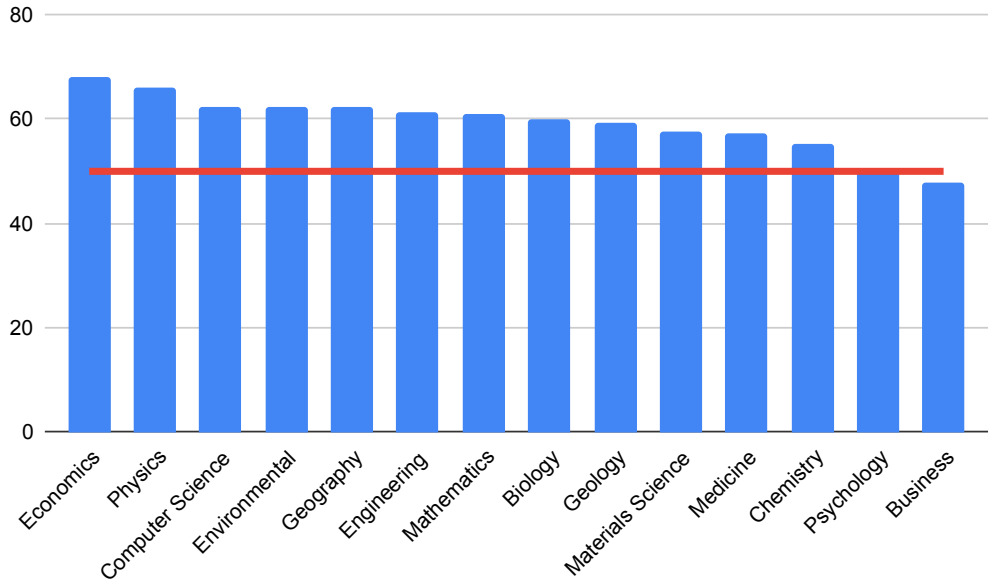

Figure SIM-9. Average percentile ranking score of XSEDE publications by Field of Study

1.1.4.2. Field Weighted Citation Impact analysis

We also repeated the FWCI analysis based on the new dataset to obtain updated results as depicted in Figure *SIM-10*. It lists the FWCI values for each field of study. The horizontal red line shows the baseline at 1. It shows that for all the FOS the FWCI values are at least around 2, which indicates that the XSEDE publications in each FOS received far more citations than expected. For the Geography, Computer Science, and Engineering the publications received about 8 times more citations than expected from the field average.

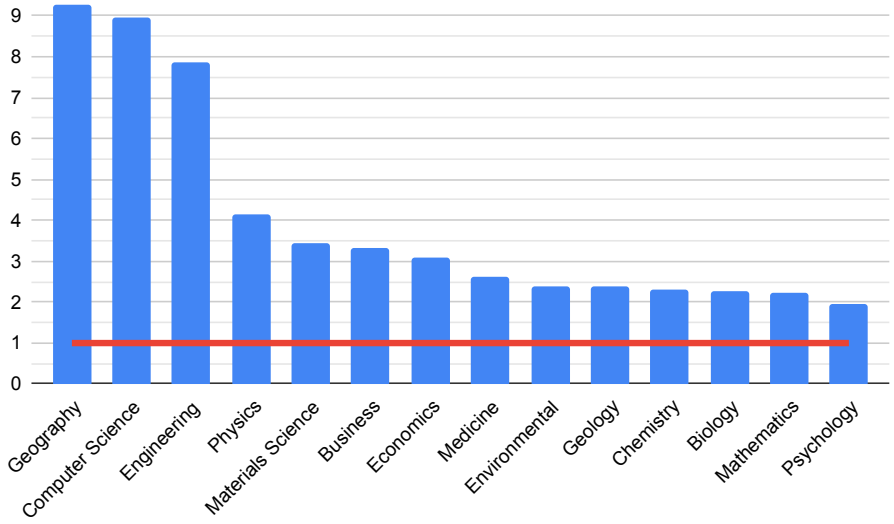

Figure SIM-10. Field Weighted Citation Index for XSEDE publications

1.1.4.3. Highly cited papers

With access to the much larger publication dataset, we can also identify for each field of study the highly cited papers. We consider the top 5% and top 1% most cited papers for the results reported here. Hence, we can find out what portion of XSEDE publications fall into those groups. Again, we excluded the few fields of study that had a very small number of publications identified as belonging to them (less than 10) due to the same reason as mentioned in the previous sections. Figure *SIM-11* shows the actual numbers, percentages for each field of study that fall into the top 5% and top 1% categories, as well as the number of XSEDE publications belonging to that field of study. The blue and yellow lines indicate the 5% and 1% baseline, respectively. All fields of study show a disproportionately higher percentage of XSEDE publications fall into the highly cited papers categories. As an example, we find specifically that 38.4% of geography related papers were in the top 5% highly cited papers, and 13.7% were in the top 1% most cited papers. Table SIM-12 summarizes the data for all the fields of study.

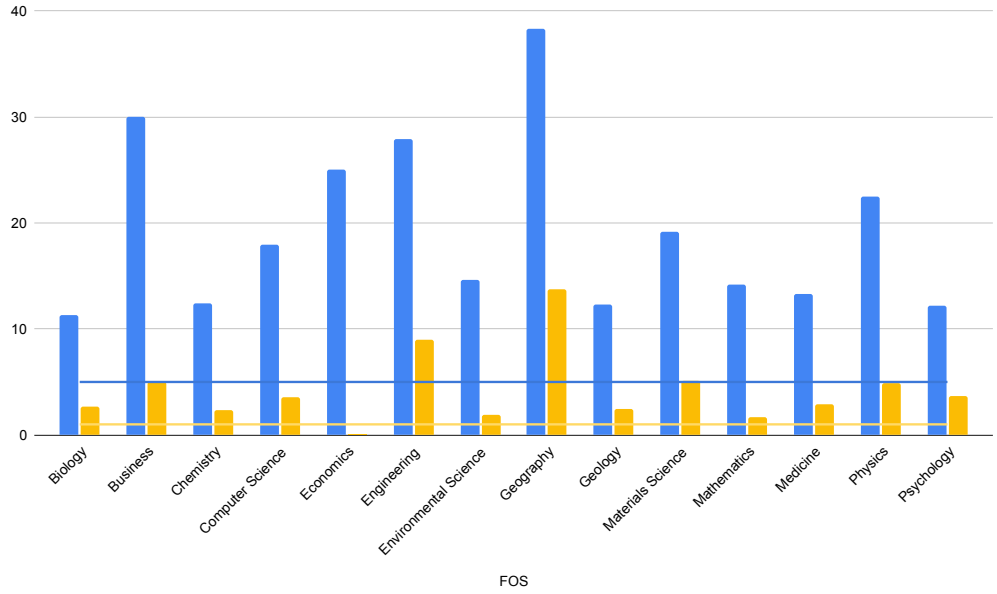

Figure SIM-11. Percentage of XSEDE publications falling into the top 5% and top 1% for each field of study

Table SIM-12. Highly Cited Papers Statistics (in top 5% and 1%)

Field	# in top 5%	% in top 5%	# in top 1%	% in top 1%	# XSEDE publications
Biology	213	11.3	50	2.7	1886
Business	6	30.0	1	5.0	20
Chemistry	609	12.4	114	2.3	4924
Computer Science	476	17.9	95	3.6	2658
Economics	7	25.0	0	0.0	28
Engineering	65	27.9	21	9.0	233
Environmental Science	76	14.6	10	1.9	521
Geography	28	38.4	10	13.7	73
Geology	70	12.3	14	2.5	571
Materials Science	615	19.2	165	5.2	3202
Mathematics	76	14.2	9	1.7	534
Medicine	1062	13.3	231	2.9	7978
Physics	1236	22.5	268	4.9	5504
Psychology	10	12.2	3	3.7	82