The Average Data Scientist is an Outlier Achiever in an Information Science Department

Ofer Bergman[a], Noa Gradovitch[a] and Tamar Israeli[b]

[a]Bar-Ilan University, IL

[b] Western Galilee College, IL

OferBergman.biu.ac.il, noahgrad@gmail.com, tamarisraeli@gmail.com

ABSTRACT

The aim of this study was to test whether data science and information science have similar or different publication and citation standards. In order to test this, we applied serpAPI to compare one hundred random Google Scholar data science profiles to their equivalents – one hundred information science profiles. The results indicate that: (a) The yearly average number of data scientist publications (15.70) was over three times higher than for information scientist (4.26). (b) The average citation per paper for data scientists (23.06) was over 4 times higher than the average for information scientists (5.70). (c) The total number of citations of papers published in 2021 for data scientists (334.54) was over 14 times higher than for in-formation scientists (23.38). These results clearly indicate that when making academic career decisions, data scientists should be evaluated according to data science standards which are very different from information science standards.

ALISE RESEARCH TAXONOMY TOPICS

Education; Bibliometrics; Standards.

AUTHOR KEYWORDS

Data science; Information science; Publications; Citations.

DOI: https://doi.org/10.21900/j.alise.2024.1689

INTRODUCTION

Academic career decisions related to choosing between candidates for a tenure-track faculty position, whether and when a faculty member should be tenured or receive an academic promotion are heavily based on the candidate's number of publications and the citations of these publications. Typically, each academic department has its own standards based on past experience, because the expected yearly publications and citations differ from one academic field to the other (Lillquist, & Green, 2010 ; Podlubny, 2005). Data scientists rarely work in data science departments and more often work in other departments such a computer science, physics and biology. Recently, data scientists are also applying to information science departments and are being promoted by committees in academic institutions as information science faculty members. Naturally, these committees tend to judge the data scientist file by information science standards. After all, 'information science' and 'data science' may sound very similar to an outside observer. The aim of this study is to test whether data science and information science has similar or different publications and citations standards.

LITERATURE REVIEW

Information science

Researchers generally agree that the roots of information sciences are in the 19th century, when the documentation movement in Europe shifted the focus of the field of librarianship from the collection of historical documents and certificates to the systematic management of these sources (Buckland & Liu, 1995). The emergence of information science as a distinct field of study began after World War II, with the computer revolution and the introduction of advanced automatic systems for information retrieval and search (Stock & Stock, 2013).

Information science is interdisciplinary in nature and constantly evolving. Some associate it to librarianship and assert that the use of new technologies is employed for the conventional tasks of the library, like information storage, retrieval, presentation, and distribution. Others view information science as a subfield of computer science or engineering (Rayward, 1996). Information sciences are crucial to the evolution of the information society and have a strong social and human dimension that goes beyond technology. That's why some researchers consider them an integral part of the social sciences (Stock & Stock, 2013).

Defining information science is challenging. It combines aspects of an applied science and a pure science and is a relatively new field in comparison to more established ones like mathematics and physics (Stock & Stock, 2013). According to Webster's dictionary concise definition, information science is a science that deals with the effective storing and retrieval of information (merriam-webster.com). However, information science continues to expand over the years, and new specializations are constantly emerging. Broader definitions take into consideration the development of the internet, social networks and the relations of information science with other fields.

The evolution of data science

Data science is a new field that evolved as a result of the advancement of digital technology in the twenty-first century. The term "Big data" refers to data amounts that cannot be handled by ordinary data application software and require technological and analytical procedures. The necessity to manage massive volumes of information that flow quickly from several sources and in various forms gave birth to the new profession.

Computer scientist Peter Naur first used the phrase "data science" in 1960. At that time, methods from mathematics, statistics, and computer science helped the discipline of data manipulation flourish. Leading statisticians called to change the term statistics to data science and to refer statisticians as data scientists (Priestley & McGrath, 2019). However, the occupational term "data scientist" didn't exist until 2008 when two team leaders DJ Patil and Jeff Hammerbacher coined the phrase during a meeting of analytics groups at LinkedIn and Facebook and launched a distinct professional specialization (Patil, 2011).

Data science was initially focused solely on little amounts of data and used to advance the social sciences. The technology that enabled the collection of massive volumes of data through internet-based apps took it to a whole new level. Now, data science is defined as a field where statistical and computational methods are used to extract insights and knowledge from data to improve diagnostic and decision-making processes (Waller & Fawcett, 2013), and transform data into real value (Van der Aalst, 2016). Converting data into value requires more than analytical and technological methods. The development of the algorithms also includes the collection, cleaning, and preparation of the data, as well as the integration of prior knowledge in the data processing processes (Avnoon, 2021).

Many sectors—including the economy, healthcare, manufacturing, and banking— have undergone a "datafiction" process that has pushed computer science and statistics out from the center and toward the periphery. The traditional statistics practice which was based on a small, static, organized database is no longer relevant for problems that are determined by a huge, dynamic, unstructured database. Computer science, which enables the collection and storage of vast volumes of organized and unstructured data also falls short in addressing the requirement to convert the data into information through modeling, sorting, and analysis (Priestley & McGrath, 2019).

The differences between information science and data science

While there is some overlap between data science and information science, they differ in their primary focus, methods, and applications. Data science revolves around the analysis of data such as raw facts and figures to derive insights and handle problems in various fields (Fred & Ma, 2023 ; Marchionini, 2016), whereas information science encompasses a broader study of information and focuses on the creation, use, and preservation of knowledge and its management (Rayward, 1996). Large-scale data analysis produces new, frequently unexpected relationships, patterns, and laws. It appears that no theory is needed because the latter are produced using a bottom-up method based on inductive procedures and statistical manipulation. Researchers in information science typically employ the conventional hypothesis-driven research design, which enables the development of particular research questions and the testing of particular hypotheses. It offers a methodical approach to scientific investigation and permits the assessment of cause-and-effect connections (Mazzocchi, 2015).

RESEARCH QUESTIONS

1. Do data scientists have different *publication* standards than information scientists?

2. Do data scientists have different *citations per publication* standards than information scientists?

3. Do data scientists have different *total citation* standards than information scientists?

RESEARCH METHOD

In order to test these research questions, we used serpAPI[1] to do the following:
(a) Find the thousand most highly cited Google Scholar scientist profiles with a 'Data science' label in their research interests field; and the thousand most highly cited Google Scholar scientist profiles with an 'Information science' label.
(b) For each category (i.e. 'data science' and 'information science') we chose at random a hundred profiles. Of the thousand 'information science' profiles 30 also contained a 'data science' label. We excluded these profiles before the random selection in order to differentiate between the two categories.
(c) For each profile we extracted the number of yearly publications the researcher had published during 2020, 2021, 2022 and the first half of 2023 (we gathered the information in June 2023). We also extracted the number of citations of the papers published in 2021 (about two years after they were published).
(d) We omitted from our data profiles of researchers which were not active during the examined time period (i.e., that did not publish any articles for two years or more of the 3.5 years examined). For the 'data science' category 18 profiles were omitted and 38 profiles were omitted for the 'information science' category.
(e) We compared the remaining data scientists and information scientists for yearly number of publications, average number of citations per paper and total number of citations of their 2021 published papers.[2]

RESULTS

Publication Standards

For each researcher sampled, we computed the yearly number of publications by dividing the total number of papers published from Jan 2020 until Jun 2023 by 3.5 years. Table 1 indicates the yearly number of publications for the sampled researchers in the different fields.

Table 1: Publications per year by research fields

Research field	Yearly publication *M* (*SD*)
Data scientist	15.70 (15.86)
Information scientists	4.26 (3.19)

[1] https://serpapi.com/google-local-api

[2] The reason we chose 2021 was so that there will be enough citation years to analyze.

Table 1 shows that on average, the sampled data scientists published 15.70 papers each year compared to 4.26 a year for the sampled information scientists. An independent samples t-tests indicate that they have published significantly more than information scientists $t(141)=5.55$, $p<0.001$. Note the large difference between information scientists and data scientists. The average data scientist published over three times more than the average information scientist. The average data scientist would be considered an outlier in an information science department because he/she publishes well over two standard deviations beyond the mean (4.26 + 2 * 3.19 = 10.64).

As the data science distribution has high standard deviation (15.86) and high skewness (2.05) perhaps the median would represent it better. The median data scientists publish 10.14 papers a year. This median achievement would put him/her in the top 10% of the information science distribution.

Citation Standards

We looked at the number of citations of the papers published in 2021 by the sampled scientists (i.e. citations of papers published about two years ago). Table 2 presents the average number of citations per paper published and the total number of citations for the researchers in each field.

Table 2: Citations per paper and total number of citations of papers published in 2021 by research field.

Research field	**Citations per paper *M* (SD)**	**Total number of citations *M* (SD)**
Data scientist	23.06 (31.63)	334.54 (678.72)
Information scientists	5.70 (8.07)	23.38 (58.44)

Citations per paper: Table 2 shows that the average number of citations per publications for the sampled data scientists was 23.06 compared to 5.70 for information scientists. An independent samples t-tests indicate that each of their publication was citied significantly more than the sampled information scientists $t(141)=4.18$, $p<0.001$. Note that the large difference, the average citation per paper for data scientists was over 4 times more that for information scientists. Again, the average data scientists would be considered an outlier for citations per paper in an information science department. The median citations per paper for in the data science distribution is 12.79. This median achievement makes the top 10% of the information science distribution.

Total number of citations: Table 2 indicates that the total number of citations for data scientists was 334.54 compared to 23.38 for information scientists. An independent sample t-test indicate that this is significantly more than the total number of citations for information scientists $t(141)=3.57$, $p<0.001$. Note the extreme difference: the total number of citations for data scientists is over 14 times larger than for information scientists. Needless to say that average data scientists would be considered extreme outliers for total number of citations in an information science department. The median total citations for the data science distribution is 132. This median achievement makes the top 5% of the information science distribution.

DISCUSSION

When deciding which candidate to choose for an academic job and when to promote a candidate, information science departments and higher level university committees tend to treat data scientists in the same standards as information scientists. However, our results indicate that there are big differences between the two disciplines: On average data scientists publish over three time more than information scientists, their per publication citation rate is over 4 times, and their total citation rate is over 14 times higher. For all three parameters an average data science would be consider an outlier achiever in an information science department, and the median data scientists would be considered at the top 10% of the department. Our results strongly suggest that data scientists should be considered by using different standards than information scientists.

One possible reason for this difference is that information science research typically take a research design approach derived by theoretical hypotheses, while data science takes a data driven approach where no theory is needed (Mazzocchi, 2015. Information scientists research begins with finding a theoretical gap (i.e. something theoretical which is not known) in order to form theoretical research questions. Then they need to choose a research method in order to test these research questions in a concrete operational manner. Then they need to get approval from ethical committees, approach potential participants and attempt to recruited them to the study. Then they typically conduct test their research hypotheses using inferential statistics, and if results are not significant there is little chance for publication. In contrast, data science takes a bottom-up (data driven) approach: they don't need to read the literature in order to find theoretical gaps as they begin with the data. Often the data is given to them or they program simple crawlers to gather their data. Because they are not committed to research hypotheses and analyze (using mostly automatic tools) a large set of data with many variables, they are bound to some statistical connections between them, so the chance for unpublished results is small. This is not to say anything about the value of the studies in these two disciplines. Only to that it is easier, faster and less risky to conduct data science research than to conduct an information science research.

CONCLUSIONS

Academic career decisions related to choosing between candidates for a tenure-track faculty position, whether and when a faculty member should be tenured or receive an academic promotion in information science departments should take into consideration that data scientists should be judged according to very different publication and citation rates than information scientists in the department. This would help avoid unfair advantages of data scientists over information scientists when competing for the same job offer, and unrealistic expectations from information scientists from promotion committees when compared to data scientists' achievement in their department.

REFERENCES

Avnoon, N. (2021). Data scientists' identity work: Omnivorous symbolic boundaries in skills acquisition. *Work, Employment and Society*, *35*(2), 332-349.

Buckland, M. K., & Liu, Z. (1995). History of information science. *Annual review of information science and technology*, *30*, 385-416.

Fred, Y. Y., & Ma, F. C. (2023). An essay on the differences and linkages between data science and information science. Data and Information Management, 7(1), 100032.
Lillquist, E., & Green, S. (2010). The discipline dependence of citation statistics. *Scientometrics, 84*(3), 749-762.
Marchionini, G. (2016). Information science roles in the emerging field of data science. *Journal of Data and Information Science*, *1*(2), 1-6.
Mazzocchi, F. (2015). Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. EMBO reports, 16(10), 1250-1255.
Patil, D. J. (2011) Building Data Science Teams. Beijing; Cambridge; Farnham; Koln; Sebastopol, CA; Tokyo: O'Reilly.
Podlubny, I. (2005). Comparison of scientific impact expressed by the number of citations in different fields of science. *Scientometrics, 64*, 95-99.
Priestley, J. L., & McGrath, R. J. (2019). The evolution of data science: A new mode of knowledge production. *International Journal of Knowledge Management (IJKM)*, *15*(2), 97-109.
Rayward, W. B. (1996). The history and historiography of information science: some reflections. *Information processing & management, 32*(1), 3-17.
Stock, W. G., & Stock, M. (2013). *Handbook of information science*. Walter de Gruyter.
Van Der Aalst, W., & van der Aalst, W. (2016). *Process Mining - Data science in action*. Springer.
Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics, 34*(2), 77-84.