DATA SCIENCE AS AN ACADEMIC DISCIPLINE

F. Jack Smith

Queen's University Belfast N. Ireland
E-mail: fj.smith@qub.ac.uk

I recall being a proud young academic about 1970; I had just received a research grant to build and study a scientific database, and I had joined CODATA. I was looking forward to the future in this new exciting discipline when the head of my department, an internationally known professor, advised me that data was "a low level activity" not suitable for an academic. I recall my dismay. What can we do to ensure that this does not happen again and that data science is universally recognized as a worthwhile academic activity? Incidentally, I did not take that advice, or I would not be writing this essay, but moved into computer science. I will use my experience to draw comparisons between the problems computer science had to become academically recognized and those faced by data science.

The first computers were invented in the 1940's, but the debate about computer science continued until the 1970's. Was it an academic discipline or was it not? There were many who took the view that computers were only tools. Chemists were the best people to develop computer systems in chemistry, engineers for engineering, and so on. Only a few technicians were needed to build and maintain the computers and write the software. This attitude was slow to change. Exactly the same arguments have been used and occasionally are still being used about data science. Is it an academic discipline? Is it not the realm of technicians rather than academics?

Surprisingly, our discipline never had a title until the 1990's when the term 'Data Science' began to be used. It was the Data Science Journal that confirmed the name. The term 'Computer Science' also evolved after a long time and only dominated in the early 1970's.

To be taken seriously, any discipline needs to have endured over time. Unlike computers, scientific data has a long history. Without astronomic data, Newton would not have discovered gravitation. Without data on materials, the Titanic would not have been built, and with good data on the location of icebergs, it might not have sunk! Data then consisted of tables of facts and quantities found in textbooks and journals, but data science did not yet exist. Then computers and mass storage devices became available, and the first databases were designed holding scientific data. Data science was born soon afterwards, about 1966, when a few far seeing pioneers formed CODATA.

Data science has developed since to include the study of the capture of data, their analysis, metadata, fast retrieval, archiving, exchange, mining to find unexpected knowledge and data relationships, visualization in two and three dimensions including movement, and management. Also included are intellectual property rights and other legal issues.

Data science, however, has become more than this, something that the pioneers who started CODATA could not have foreseen; data has ceased being exclusively held in large databases on centrally located main frames but has become scattered across an internet, instantly accessible by personal computers that can themselves store gigabytes of data. Therefore, the nature and scope of much scientific and engineering data and, in consequence, of much scientific research has changed. Measurement technologies have also improved in quality and quantity with measurement times reduced by orders of magnitude. Virtually every area of science, astronomy, chemistry, geoscience, physics, biology, and engineering is also becoming based on models dependent on large bodies of data, often terabytes, held in large scientific data systems.

If all of this activity is communicated effectively, who can doubt that data science is a serious academic subject? However, has this been communicated effectively? It is true that data science has begun to spawn a lively body of literature, growing in detail and breadth, but it is still in its infancy and scattered throughout a multitude of journals primarily devoted to something else. The one exception is CODATA,

which has been publishing a long informative series of bulletins, books, reports, and most significantly, almost 3000 papers published as proceedings of biannual, international conferences. This invaluable source is the largest collection of excellent papers in data science [which CODATA should publish in its entirety on the Internet]. Unfortunately, this collection's great academic value is not widely recognized because it was not rigorously refereed and is not readily available. A young scientist wanting promotion based on CODATA conference publications would be in difficulty.

Now let us look at what computer science did about publications. First of all, computer science held international conferences (e.g. IFIP conferences) and published proceedings as well as several bulletins and other publications. CODATA has also been doing this. But in addition, computer science refereed journals at an early date: the IEEE Computer Society formed in 1946 published its first journal 6 years later in 1952; the Association of Computing Machinery founded in 1947 published its first journal in 1954, and The Computer Journal in the UK formed later in 1956 published its first journal in 1958. These papers were all refereed although refereeing was less rigorous than today. Now 50 years later, these three societies together publish 20 journals in computer science. No one questions the academic standing of computer science any longer.

In comparison, it took 38 years for CODATA to publish its first journal in 2002. This new electronic journal not only freely disseminates knowledge on data science to a world wide readership but also gives data science professionals a refereed quality journal where they can publish and get deserved recognition for the quality of their papers. The journal was the most important step taken by CODATA since it was formed, and it deserves our support.

Once a body of literature is in place, academic courses can begin at universities. In computer science, single modules began to be taught about 1960, text books followed about 1970, and full degrees before 1980. Although data science does not have the breadth of computer science, we should now be starting single modules, perhaps initially to graduate students. Textbooks need to follow (perhaps sponsored by CODATA).