Dragos Condrea/iStockPhoto.com

What is data science? A closer look at science's latest priority dispute

Jonathan Auerbach, **David Kepplinger** and **Nicholas Rios** use two popular data science algorithms – naïve Bayes and eigencentrality – to examine the difference between data scientists, statisticians, and other occupations

What is data science, and where did it come from? Is data science a new and exciting set of skills, necessary for analysing twenty-first-century data? Or is it (as some have claimed) a rebranding of statistics, which has carefully developed time-honoured methods for data analysis over the past century?

Priority disputes – disagreements over who deserves credit for a new scientific theory or method – date back to the beginning of science. Famous examples include the invention of calculus and ordinary least squares. But this latest dispute calls into question the novelty of an entire discipline.

In this article, we use two popular data science algorithms to examine the difference ▶

Jonathan Auerbach is an assistant professor in the Department of Statistics at George Mason University.

David Kepplinger is an assistant professor in the Department of Statistics at George Mason University.

Nicholas Rios is an assistant professor in the Department of Statistics at George Mason University.

between data science, statistics, and other occupations. We find that in terms of the preparation required to become a data scientist, data science reflects both the work of natural sciences managers – individuals who oversee research operations in the natural sciences – and statisticians and mathematicians. This suggests that data science is a shared enterprise among science and mathematics, and thus those trained in the natural sciences have as much claim to data science as those trained in mathematics and statistics.

In terms of the role a data scientist serves relative to other occupations, however, we find that data science is closest to statistics by far. Both occupations are fast growing and central among the occupations that work with data, suggesting that a data scientist serves the same function as a statistician. But this function may be changing. While the centrality of statistics has declined over the past decade relative to other occupations, the centrality of data science has grown. In fact, data science has now surpassed statistics as the most central fast-growing occupation.

We examine the role of data science using data science

Everyone seems to agree that data science requires skills traditionally associated with a variety of different occupations. Drew Conway, for example, describes data science as a combination of mathematics and statistics, substantive (domain) expertise, and "hacking" skills (see Figure 1). In dispute is the relative importance of those skills. Some have argued that data science is basically statistics – and that twentieth-century statisticians such as John Tukey have long possessed the data science skills traditionally associated with computer science and the natural sciences. Others have argued that data science is truly interdisciplinary, and statistical thinking only plays a small role. But while opinions on data science abound, few appear to be based on data or science.

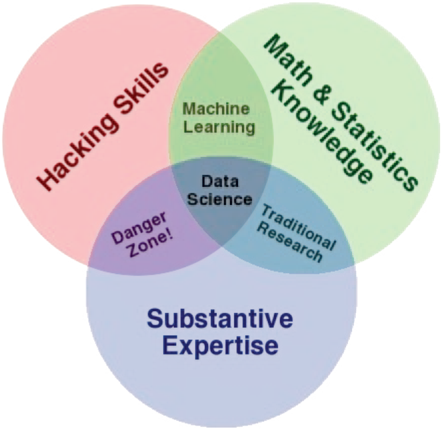

Figure 1: Drew Conway describes data science as a combination of mathematics and statistics, substantive (domain) expertise, and "hacking" skills. Conway's data science Venn diagram, reproduced here, is Creative Commons licensed as Attribution-NonCommercial.

Moreover, descriptions of occupations by government agencies are not particularly helpful in differentiating between data science, statistics, and related occupations. For example, according to the Bureau of Labor Statistics, data scientists use "analytical tools and techniques to extract meaningful insights from data". This description is similar to mathematicians/statisticians, who "analyse data and apply computational techniques to solve problems", and operations research analysts who use "mathematics and logic to help solve complex issues".

For these reasons, we use two popular data science algorithms, naïve Bayes and eigencentrality (eigendecomposition), to investigate the question: What is data science? Both algorithms use data listing the training a worker must generally complete to work in an occupation, such as data science. Specifically, we use the CIP SOC crosswalk provided by the US Bureau of Labor Statistics and US National Center for Education Statistics, which links the Classification of Instructional Programs (the standard classification of educational fields of study into roughly 2,000 instructional programmes) with the Standard Occupational Classification (the standard classification of professions into roughly 700 occupations).

Our main assumption is that the skills required to work in an occupation can be represented by the instructional programmes that prepare students to work in that occupation. For example, the occupation "data scientists" is associated with 35 instructional programmes, such as data science, statistics, artificial intelligence, computational science, mathematical biology, and econometrics. The occupation "statisticians" is associated with 26 instructional programmes, including data science, statistics, and econometrics, but not artificial intelligence, computational science, or mathematical biology.

> Everyone seems to agree that data science requires skills traditionally associated with a variety of different occupations

The algorithms we employ consider occupations to be similar if they have many instructional programmes in common. Data scientists and statisticians share 14 degrees, suggesting they are similar: half the programmes that prepare students to be a statistician also prepare students to be a data scientist. In contrast, data scientists and computer programmers share six degrees in common, suggesting they are less similar: computer programmers have 17 degrees overall, so only a third of the programmes that prepare students to be a computer programmer also prepare students to be a data scientist.

Our analysis treats all instructional programmes as equal and independent. We do not consider, for example, the number of workers who hold a degree from an instructional programme or whether two instructional programmes are similar or offered by similar academic departments. Our analysis could be adjusted to account for this or related information, although it is unclear to the authors whether such an adjustment would make the results more accurate.

Data science is a shared enterprise among science and mathematics

We use naïve Bayes to measure the similarity between each occupation and data science in terms of the preparation required to work in that occupation. Specifically, we first pretend that the occupation "data scientist" did not exist and then use Bayes' rule to calculate the probability that a hypothetical group of workers with the 35 degrees associated with data science could have come from one of the roughly 700 other occupations. The higher the measure, the more consistent that occupation is with data science.

The use of Bayes' rule is appealing because the similarity between a given occupation and data science takes into account the similarities between every other occupation and data science. Our use of Bayes' rule is naïve in the sense that – before collecting the data – we assume these workers are equally likely to have come from any occupation.

The occupations with the largest probabilities, and thus most related to data science, are summarised in Figure 2. We find that the hypothetical workers have a 50% chance of being natural sciences managers and a 50% chance of being statisticians or mathematicians. (Note that natural sciences managers share 18 instructional programmes with data scientists, while statisticians share 14.) We conclude that data science is a shared enterprise among science and mathematics, and thus those trained in natural sciences have as much claim to data science as those trained in mathematics and statistics.

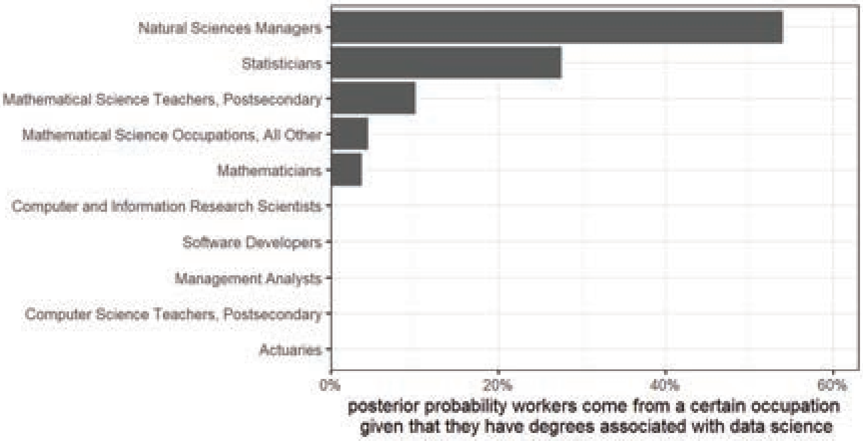

Figure 2: We use naïve Bayes to measure the similarity between each occupation and data science in terms of the preparation required to work in that occupation. We find that in terms of the preparation required to become a data scientist, data science is a shared enterprise among science and mathematics.

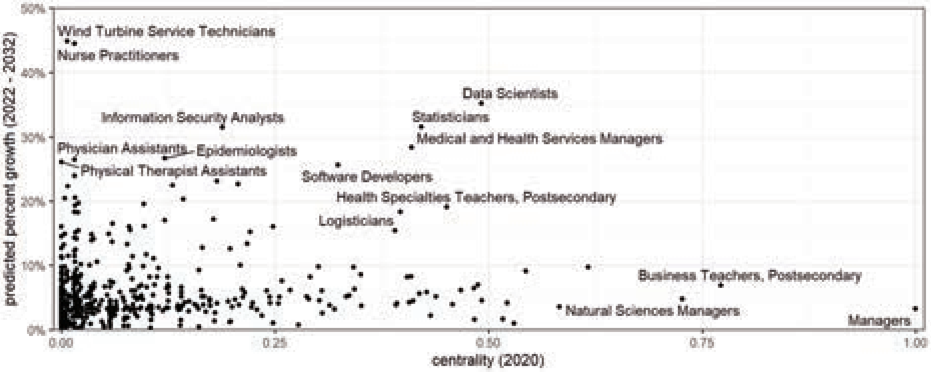

Figure 3: We use eigencentrality (eigendecomposition) to measure the similarity of each occupation in terms of its role relative to other occupations. We find that in terms of the role a data scientist serves relative to other occupations, a data scientist functions like a statistician.

We find that the centrality of statisticians has declined over the past decade relative to other occupations, while the centrality of data scientists has grown

Data science is closest to statistics in its role among other occupations

We use eigencentrality (eigendecomposition) to measure the similarity of each occupation in terms of its role relative to other occupations. Specifically, we calculate the principal right singular vector of the adjacency matrix denoting whether an instructional programme (row) is associated with an occupation (column). Or alternatively, the principal eigenvector of the adjacency matrix denoting the number of instructional programmes each occupation (row) has in common with each other occupation (column). An occupation has high eigencentrality when the instructional programmes that prepare a worker for that occupation also prepare that worker for many other occupations as well. This suggests that the higher the measure, the more central the role of the occupation relative to other occupations.

The eigencentrality of each occupation is displayed in Figure 3. Each point represents an occupation, the *x*-axis denotes the centrality of the occupation, and the *y*-axis denotes the percentage growth of the occupation as predicted by the US Bureau of Labor Statistics over the next decade. The figure demonstrates that data scientists and statisticians occupy nearly identical positions: both are fast growing and central to the other occupations that work with data. In contrast, natural sciences managers are central but growing much more slowly, suggesting a role closer to managers. We conclude that although data scientists are prepared similarly to natural sciences managers, a data scientist serves the same function as a statistician.

But this function may be changing. Figure 4 shows the centrality (*x*-axis) of each occupation (*y*-axis) in 2010 and 2020. Green bars denote increases from 2010 to 2020, while yellow bars denote decreases. We find that the centrality of statisticians has declined over the past decade relative to other occupations, while the centrality of data scientists has grown. In fact, data science has now surpassed statistics as the most central fast-growing occupation. We conclude that although a data scientist and a statistician serve similar roles today, those roles may change as the workforce changes. ▶

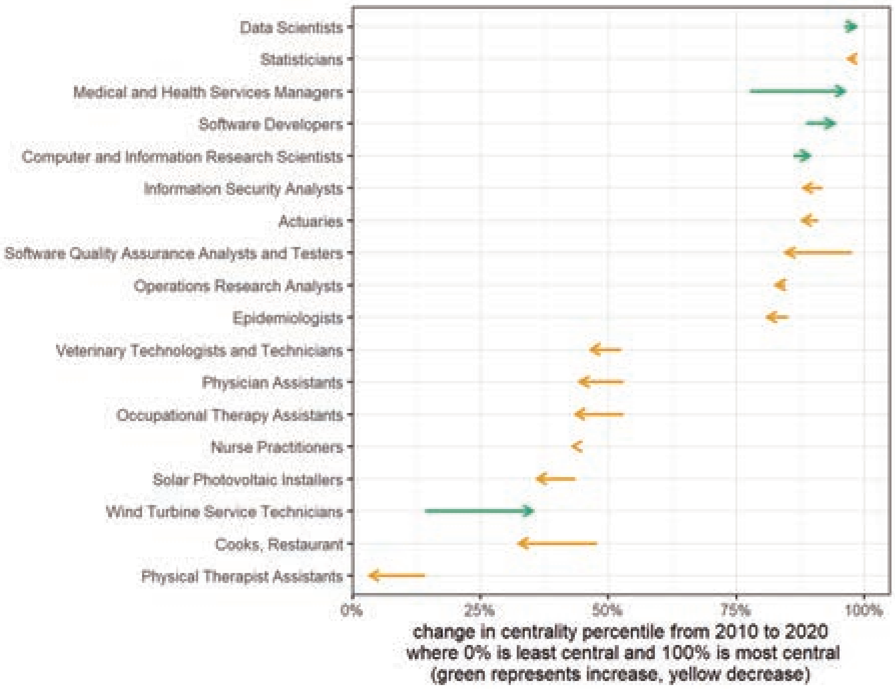

Figure 4: We use eigencentrality (eigendecomposition) to measure the similarity of each occupation in terms of its role relative to other occupations. We find that the centrality of statisticians has declined over the past decade relative to other occupations, while the centrality of data scientists has grown. Data science has now surpassed statistics as the most central fast-growing occupation. (Occupations predicted to grow more than 20% over the next decade shown.)

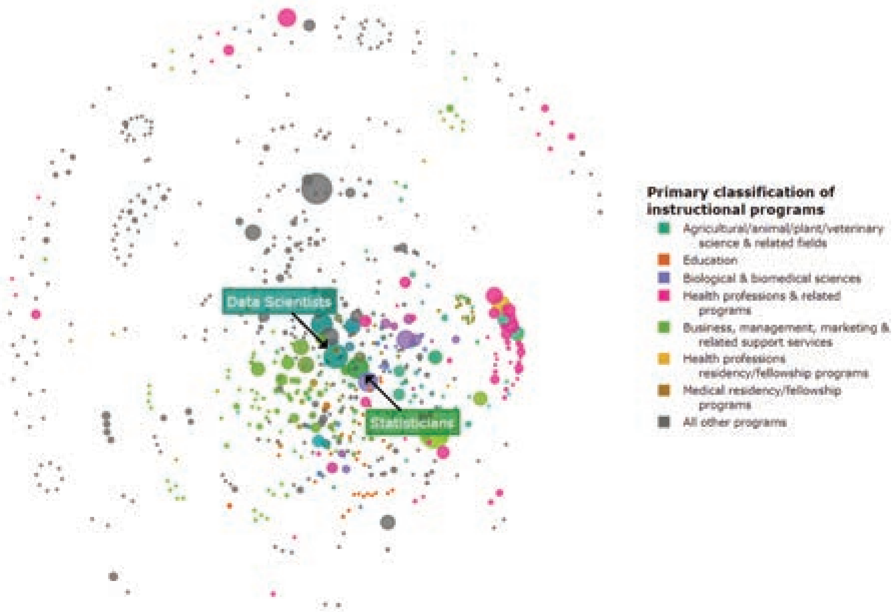
Data Scientists
Statisticians
Primary classification of instructional programs
🟩 Agricultural/animal/plant/veterinary science & related fields
🟧 Education
🟪 Biological & biomedical sciences
🟥 Health professions & related programs
🟩 Business, management, marketing & related support services
🟨 Health professions residency/fellowship programs
🟫 Medical residency/fellowship programs
⬛ All other programs

Figure 5: A visualisation of occupations as a network. Occupations are placed according to the instructional programmes that train students for that occupation, with occupations closer together sharing more instructional programmes in common. We find that data scientists and statisticians occupy nearly identical positions at the centre of the network. Occupations are coloured according to the primary classification of instructional programmes that train students for that occupation. Larger nodes represent occupations that are growing faster.

(Note that the occupation classifications changed in 2018, and we used the crosswalk provided by the US Bureau of Labor Statistics to make these comparisons).

The findings in this section are based on the adjacency matrix that encodes whether an instructional programme (row) is associated with an occupation (column). A more detailed summary of the matrix is provided in Figure 5, which depicts the matrix as a network graph. Larger nodes represent occupations that are growing faster, while nodes closer to the centre of the network represent more central occupations.

Is data science statistics?

We conclude that individuals trained in managing natural sciences research – a slow-growing occupation – are turning to data science – a much faster-growing occupation, and one which currently serves a role like that of a statistician. But if present trends continue, data science is poised to eclipse the historic role of the statistician as central to the occupations that work with data.

This suggests that while data science may be new and exciting, the role served by the data scientist is not particularly new. This does not mean that data scientists necessarily use the same time-honoured methods for data analysis as statisticians. It is the authors' experience, however, that many data science tools are in fact statistical. Indeed, the two data science algorithms we used in this article are both taught to students as new and exciting, but in reality are centuries-old methods steeped in statistical history.

Regardless of whether data science is or is not statistics, the occupation "data scientist" has proven immensely popular, capturing a zeitgeist that has eluded statistics. This is best evidenced by the fact that data science – and not statistics – has been crowned the sexiest job of the twenty-first century (tinyurl.com/2tumnd58). But if statistics has not enjoyed the popularity of data science, perhaps the real question in need of answering is: What is statistics? ■

Note

Data and code to reproduce the analysis and figures are available through GitHub (tinyurl.com/58pyndx4).

This article originally appeared on realworlddatascience.net.