

Should We Change the Name of the Field of Statistics to “Data Science”?

Donald B. Macnaughton

Most fields of science have names that are neutral, being merely names. For example, the names “physics” and “biology” are neutral names for two large fields of science. In contrast, the name “statistics” is ambiguous. And, arguably, the ambiguity generates a negative perception of the field. So it is sensible to consider changing the name.

The name “statistics” is ambiguous because the word has four different meanings. It can mean:

1. the actual *values* of descriptive statistics, which are generally numbers, such as Babe Ruth’s major-league baseball batting average of 0.342
2. the raw data behind the numeric values of descriptive statistics such as 2873 “hit” indicators behind Babe Ruth’s 8399 at-bats in his major-league career
3. the abstract statistics that statisticians and data scientists have invented (e.g., the average, the *t*-statistic, the *p*-value), which are the *concepts* and *algorithms* that we use to compute (from data) the values of descriptive and test statistics, and
4. the entire field of “statistics” which, given the key role of data modelling in statistics, is much more than the sum of the preceding meanings 1 through 3.

The four different meanings of “statistics” give laypeople a sense that the field is vague.

In addition, the name “statistics” doesn’t transparently convey the high-level function of the field. And most laypeople don’t know the function, with some thinking that the function is merely to collect data. Thus the name would be better if it said what the field does.

“Statistics” is certainly a somewhat correct name for the field because a substantial percentage of work that statisticians and data scientists do involves descriptive statistics and test statistics. But the various statistics that we study are always computed from actual or theoretical *data*. So the data come first in the work and they, not the statistics, are (from a technical perspective) the collective main object of study in a typical scientific research project that is supported by statistics or data science. (Descriptive and test statistics are vital *measures* of data that help us to understand the data.) So it is quite reasonable for the word “data” to be in the name of the field.

John Tukey recommended that we call the field “data analysis” (1962). This was seconded by Frederick Mosteller in the title of a joint book chapter “Data Analysis, Including Statistics” (Mosteller and Tukey, 1968). Tukey wrote “*data analysis is intrinsically an empirical science*” (1962, p. 63, his italics), implying that he believed the field is a *science*.

Statisticians Jeff Wu (1997), Chikio Hayashi (1998), and William Cleveland (2001) were the perceptive first proponents of the name “data science” for the field (Wikipedia contributors, 2017). Their name is arguably slightly more effective than “data analysis” because “science” sounds more interesting and more general to many laypeople than “analysis”, which sounds complicated and dry. Also, the name “data science” gives a clear sense of the function of the field—it is the scientific study and interpretation of data.

For the layperson, if we apply data science to data, this suggests that we will obtain *meaning* from the data. But if we apply statistics to data, this suggests that we will obtain *statistics*, and the puzzled layperson may then wonder “What good is that—it’s just some numbers?”

Similarly, nowadays the idea of “big data” is often in the news. For the layperson, which name sounds better to study big data—data science or statistics? Arguably, “data science” sounds better, even though the field of statistics has produced most of the main tools for studying big data.

It is important to help laypeople to better understand the field of statistics because the field is widely misunderstood.

Perhaps tellingly, statisticians are using the name “data science” more often. For example, on September 12, 2017 an Abstract Keyword search for exact matches of “data science” (without the quotation marks) in the online program for the 2017 Joint Statistical Meetings found 34 (of more than 600) activities that used the name. For example, Jon Wellner’s talk is titled “Teaching Statistics in the Age of Data Science”. Searches of the program for earlier years reveal that the name was used with roughly decreasing frequency in the four years before 2017, and the name doesn’t appear in the program in the four years before 2013.

The name “statistics” has served the field since the late 1700’s (David 1995). Thus it would be a significant break with tradition to retire this respected name. And it would be regrettable to change the name in light of the ongoing impressive “This is Statistics” campaign developed by the American Statistical Association. (But we might quite reasonably change the name of the campaign to “This is Data Science”.) So some inward-looking statisticians will say that we should keep the traditional name.

But outward-looking data scientists say that the name “data science” better conveys the function of the field.

And the name isn't ambiguous. And the name “data science” (correctly) sounds like a path to meaning. Arguably, these various advantages greatly outweigh tradition.

Some people may think that the name “data science” is already taken. But, to use Robert Rodriguez' insightful metaphor (2013), it is sensible to view the name as a “big-tent” name that encompasses what we traditionally call “statistics”, but which also encompasses some areas of computer science. So if we change the name of the field to “data science”, we wouldn't be appropriating the name. But we would be acknowledging the correctness of the name for what data scientists and statisticians do—the name is sensibly applied to *all* activities aimed at systematically understanding data.

Some statisticians will take it for granted that we couldn't possibly change the name. But why not? What are the disadvantages? Of course, if we change the name, there will be conversion costs, and there may be some old-guard angst. But these are transient and are arguably outweighed by the substantial public relations benefits.

It is important to emphasize that the present discussion is only about changing the *name* of the field, with no intent to change the goals or activities of the field. (Changing the name might have the indirect effect of leading us to focus more on data but, arguably, that wouldn't be a bad thing.)

If we consider changing the name of the field, then it is also sensible to consider changing the names of some of the field's organizations. Therefore, would it be sensible to consider renaming the American Statistical Association and the Royal Statistical Society as the American Data Science Association and the Royal Data Science Society? Would it be sensible to consider renaming similar organizations?

Similarly, would it be sensible to consider renaming some journals (e.g., *The American Data Scientist*)? Would it be sensible to consider renaming each college or university Department of Statistics as the Department of Data Science?

Of course, any change of a primary name should be done with care to ensure that the public will properly understand the intent. Thus any change might be best done with the help of branding experts. They have experience in communicating a sponsor's vision to the public with appropriate positive public-relations impact. They would help to ensure that our intent—i.e., easy understanding of the function of the field—is properly communicated in a message that is simple, informative, and friendly. Ideally, a single public-relations firm or advertising agency would work with statistical organizations to provide a unified approach for the entire field.

If we were to decide to change the name of the field, then it would be important to strive for cooperation and unity among the various relevant organizations. This would show the strength of the field.

Is it sensible bold public relations—a leadership step—for statistical organizations (including journals and academic departments) to work with other organizations in the field to change the name? Would such changes help to propel the science of statistics to the forefront of the data-science revolution, where arguably it belongs? Would the changes generate substantial public interest and respect that such a long-established discipline has the flexibility to change its name to better represent its function?

The field of statistics has produced a large set of valid, versatile, and powerful methods to unlock data. Do we and our methods belong in our own small tent at the fair? Or are we a key part—sometimes the center ring—of the big-tent data-science culture? What is the best name for our beautiful field?

References

- Cleveland, W. S. (2001), “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics,” *International Statistical Review*, 69, 21–26. <http://doi.org/10.1111/j.1751-5823.2001.tb0047.Hx>
- David, H. A. (1995), “First (?) Occurrence of Common Terms in Mathematical Statistics,” *The American Statistician*, 49, 121–133. <http://doi.org/10.1080/00031305.1995.10476129>
- Hayashi, C. (1998) “What is Data Science? Fundamental Concepts and a Heuristic Example.” In *Data Science, Classification, and Related Methods, Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96)*, (C. Hayashi, K. Yajima, H. H. Bock, N. Ohsumi, Y. Tanka, and Y. Baba, eds) Tokyo: Springer. <http://doi.org/10.1007/978-4-431-65950-1>
- Mosteller, F. and Tukey, J. W. (1968), “Data analysis, including statistics.” In *Handbook of Social Psychology*, 2nd ed. (G. Lindzey and E. Aronson, eds.) 2, 80–203, Reading MA: Addison-Wesley.
- Rodriguez, R. N. (2013), “Building the Big Tent for Statistics [American Statistical Association 2012 presidential address],” *Journal of the American Statistical Association*, 108, 1-6. <http://doi.org/10.1080/01621459.2013.771010>
- Tukey, J. W. (1962), “The Future of Data Analysis,” *The Annals of Mathematical Statistics*, 33, 1–67. <http://www.jstor.org/stable/2237638>
- Wikipedia contributors (2017), “Data science”. In Wikipedia. Retrieved July 13, 2017 at 13:29 EST. https://en.wikipedia.org/wiki/Data_science
- Wu, C. F. J. (1997), “Statistics = Data Science?” Retrieved July 11, 2017 from <http://www2.isye.gatech.edu/~jef-fwu/presentations/datascience.pdf>

This essay was sent or linked on September 12, 2017 to the following communities: AllStat, ANZStat, ASA Connect, Statistical Society of Canada.