Marcos Barreto June 13th, 2023

# A fundamental problem at the heart of data science teaching

1 comment | 1 shares

Estimated reading time: 10 minutes

*Data science techniques are held up as a beacon of objectivity and interdisciplinarity – but are universities instead training future data scientists in a form of data colonialism? Marcos Barreto investigates.*

Data science applies a mixture of mathematical, statistical, and computing-related concepts to domain-specific problems, such as recommendation systems for online purchases built on consumer behaviour, identifying trends and forecasts in financial markets, and classifying and predicting health outcomes based on patient demographics and medical history.

Based on a full understanding of a domain-specific problem, data scientists need to know which statistical tools should be used to analyse the problem's input data and generate accurate outputs, and which technological resources, including data science software, are most appropriate for that.

## Training data scientists

The growing demand for data scientists across different sectors to make use of vast amounts of data has led to an increasing, although still insufficient, number of data science courses and programmes. Universities and other educational settings have no standard curriculum for data science and can struggle to find a perfect balance between statistical/computing and domain-specific concepts.

Most data science programmes and courses use a mixture of statistical and programming concepts and activities based on standard examples from literature (books, blogs, and technical documentation) and research. Domain-specific concepts are less common in data science curricula, as they are harder to generalise. In this sense, educators tend to stick to industry-led standard examples to provide training on the basic building blocks of modelling and analysis.

But are educators simply reproducing these standard examples and thus asking students to learn how to do the same? Or are they truly promoting the hard and soft skills required in a growing knowledge-based society, including thinking skills, ethics and responsibility, digital skills, and knowledge management?

### Tech giants dominate

One problem is that data science is based on the assumption that data scientists are given different tools – such as programming languages, specialised libraries for data pre-processing, analysis, and visualisation – as well as computing infrastructures, and can freely choose the best combination for a particular problem. This is not entirely wrong but, in practice, this freedom is restricted.

Data scientists tend to follow one or more standardised approaches. These are a combination of tools and analytical models that big tech companies promote as leading to the best results, such as convolutional neural networks pre-built over ImageNet training data. This scenario reflects technological monopoly (or a form of colonialism), where a small number of big tech companies impose their big computing infrastructures and associated case studies as references for deploying different types of applications (classification, prediction, anomaly detection, among others).

### AI challenges

Besides the issue of technological dominance, there is increasing discussion of data colonialism (or data capitalism), in which data is exploited for economic expansion. Data is continuously extracted and exploited based on systems, institutions, and values established in the past and which persist or remain unquestioned in the present.

One of the practical effects is the proliferation of toy datasets. These are samples extracted from a given data source and used as a reference for explaining a particular data science approach or application. This adoption of standard use cases and associated data can perpetuate bias (including historical and social) and misinformation across data science solutions and practitioners.

The increasing adoption of generative artificial intelligence (AI) tools across different domains brings additional challenges to educators, including its ethical use as supportive technology for teaching, learning, and assessment. We are witnessing a plethora of “best prompts for a particular problem” targeting foundational and large language models trained over a variety of (not necessarily unbiased) data and backed by big tech companies. Will “prompt collections” (or any other similar names) and prompt engineering guides be another type of data colonialism?

### The problem of objectivity

If, as David Narter argues, the goal of objectivity is to make all things equal in all places, when applied to data science, objectivity means that data scientists are expected to make reliable, informed decisions based on objective data representative of a particular domain; no subjectivity is expected. The more data the better, if computerised algorithms are to provide reproducible results as a sign of objectivity. This idea seems to match perfectly with teaching quantitative courses, as solving mathematical and statistical problems is an irrefutable way to remove subjectivity and ensure that all students have the same chance to succeed.

This rationale can help explain why most data science teaching and assessment resources are strongly based on standard examples and data, as they allow the students to learn a safe and well-trodden path through the building blocks with minimal chances of error. In addition, they allow for easy marking criteria, such as whether a particular software library was used or whether the correct labels were chosen to tell the story in visualisation graphs. They also enable fast (sometimes impersonal) feedback, as the results are also predictable and comparable.

> *Are we looking at our students as objective data points or repeatable algorithms assumed to reproduce standard outputs?*

But, by adopting marking criteria that are heavily embedded in technical aspects and well-known problems, are we (educators) looking at our students as objective data points or repeatable algorithms assumed to reproduce standard outputs for the proposed problems, without any margin for subjectivity?

The way data science courses are assessed risks worsening the problem. Assessment is mostly based on in-class and homework assessments as well as practical coursework and written exams. The students work individually or in groups to apply data science skills to different datasets and problems, in much the same way as expected in the industry and other labour markets. But to what extent does this effectively assess soft skills and allow for experimentation and innovation, or simply assess whether students will reach similar results to those found in standard examples and thus be graded based on that?

In addition, we assume that data science is intrinsically interdisciplinary, as even the simplest data science course will ask students to perform statistical calculations and code writing over domain-specific data. By adopting standard use cases from a diversity of domains, we educators assume that interdisciplinary awareness will be developed and enhanced among our students, and this will allow them to generalise to less known or new scenarios while taking social context and data ethics into account. Does this really work?

### Towards a more diverse practice

In reality, these teaching and assessment practices can distort the idea of objectivity and interdisciplinarity, while helping to perpetuate technological and data colonialisms. This can only be fought by critically assessing the basis and contexts where data and reference use cases are produced and promoted, to identify bias and colonial traces and then work towards a more diverse and interdisciplinary practice.

The data-centric AI movement is a good sign of the industry's awareness of the need for better-quality data. It argues for a systematic improvement of the data based on the observed results from a particular model, so as to keep the model barely changed and invest in different techniques for iterative data acquisition, labelling, cleansing, and preparation. How can we incorporate such practices around data quality in our teaching and learning paths? Besides standard datasets and use cases, which other data sources would help us to ensure our models are able to generalise to more complex and diverse settings while minimising (or even eliminating) different sources of bias? Where can we find them among the hundreds of data repositories?

Data literacy must be promoted as an approach to identify data relevant to a particular field of study but also linked to the larger world, as David Schuff proposes in *Data Science for All*. There is also a case to be made for designing a "competence profile and framework", as emphasised by Sampson et al in *Educational Data Literacy*. Some technical steps can be also useful when accounting for data quality and representativeness, such as designing comprehensive documentation spanning the entire dataset lifecycle (as proposed by datasheets for datasets) and applying a framework to check different aspects of analytical solutions, including training and testing datasets.

> *Promoting group interactions and prioritising critical analysis over technical outputs can be a valuable approach…*

Objectivity must be revised and somewhat relaxed to give room for scientific thinking and modelling, even when exploring well-known problems, and reflecting on new teaching and assessment strategies. Promoting group interactions and prioritising critical analysis over technical outputs can be a valuable approach, especially in an era when "prompt engineering is a fine craft", as Janna Lipenkova suggests when discussing the hallucinations (semantically correct but factually incorrect texts) and silent failures (dubious outputs presented with great confidence) affecting today's large language models.

However, even with the constant evolution of technologies surrounding data science, the data itself is the main artefact. Embracing all these new concepts and visions related to "better data practices" into our teaching and assessment will be an ongoing challenge – but one that data science educators must embrace.

- This post was awarded runner-up in the *LSE HE Essays in Education Blog Challenge* in June 2022 in the open category.

This post is opinion-based and does not reflect the views of the London School of Economics and Political Science or any of its constituent departments and divisions.

*Top image credit:* photo by Oriol Portell on Unsplash

## About the author

**Marcos Barreto**

Marcos Barreto is an Assistant Professorial Lecturer in Data Science at the London School of Economics, UK

**Posted In:** Essays in Education