The Changing Landscape of Data Science

Thomas Speidel

November 19, 2024

Content

1 TL;DR

1. Data science has undergone significant transformations over the past decade, influenced by several factors.

2. **The Beginnings**: Initially rooted in the scientific community, data science emerged from the convergence of fields like statistics, computer science, and operations research. This evolution was catalyzed by reduced costs of storage and computing, along with increased bandwidth, which in turn amplified the scale, velocity, and variety of data.

3. **Interdisciplinary Nature**: Unlike most STEM disciplines, data science did not formally exist until recently and has been driven by commercial rather than intellectual developments, leading to its constant reinvention.

4. **Technology Firms' Influence**: By 2015, technology firms had recognized the potential of data science and made significant investments and acquisitions, often prioritizing revenue-generating aspects of the field. This focus frequently overshadowed important applications, scientific breakthroughs and new methods that may not require large-scale solutions or sophisticated technology stacks.

5. **Scientific Challenges**: The scientific aspect of data science involves experimentation, understanding, and failures which can be difficult to monetize and are often viewed with suspicion by organizations.

6. **Democratization**: As data science became more accessible, many practitioners entered the field from non-scientific backgrounds, often up-skilling through online courses or boot camps. As a result, the newest cohorts often lack experience in scientific inquiry, as well as the principles and procedures that ensure the validity and reliability of their work.

7. **Current Developments**: The field has faced four significant developments: (a) the influence of technology firms, (b) the complex relationship between business and science, (c) the democratization of data science, and (d) the advent of generative AI.

8. **Evolving Focus**: As a result, data science has shifted from being *insight-focused* to *data* and *product-focused*, resembling more the role of a developer in IT than that of a sophisticated analyst. Good data pipelines are now seen as more valuable than good models. The value of a data science initiative is often measured not by the quality of the insight and the decisions it helps support, but rather by the product's deployment, uptake and sustainment.

9. **Unmet needs**: As a result, the need for insight-focused approaches in many areas of the organization may no longer be fully met by the data science professional.

10. **Need for reform**: Data science jobs need clarity. Curricula need to include research and principled methods of analysis. Hiring managers need to more accurately match relevant skills to the needs of the role. Organizations need to understand that there is no one-size-fits-all approach. A data scientist who excels at building good data pipelines may be a good data professional but not necessarily a strong scientist. Conversely, a data scientist who excels at generating valid insights from a model may be a good scientist but not necessarily a strong data professional. Roles that are focused on *insight discovery* may not always require to build data pipelines or deploy models in production.

2 Background

The history of data science is rather convoluted as it is best understood as the convergence of multiple, yet historically isolated fields such as statistics, computer science and operations research to name a few. This convergence occurred more or less when key barriers of entry were lowered, specifically the costs of storage, computing, and the availability of large bandwidth.

One point I want to emphasize is that, unlike other formal STEM disciplines, data science did not formally exist until recently and has only now become an interdisciplinary field of study in post-secondary education. In other words, data science was "*motivated by commercial rather than intellectual developments*"[1]. However, those commercial developments were largely built upon existing areas and tools.

Why is this important? Because as data science evolved in industry, it lacked a formal, unified discipline, a unified set of nomenclature and norms, it lacked a body of knowledge to adhere to, and this explains in part its constant evolution. But there is more to it.

SAS started in 1976, R in the early 1970's, Python in 1991. For algorithms: logistic regression was developed in 1958, neural networks in the 1950's, cross-validation in 1974.

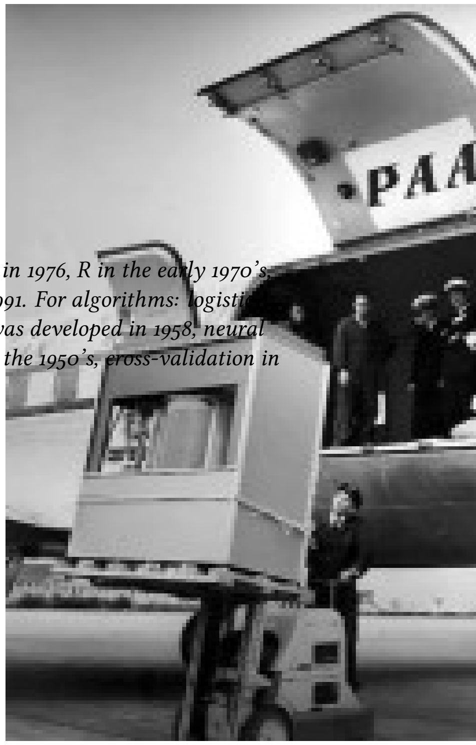

Figure 1: *A 5 MB hard drive from IBM is being loade in 1956.*

[1] Donoho, David. "50 Years of data science." Journal of Computational and Graphical Statistics 26, no. 4 (October 2, 2017): 745–66. https://doi.org/10.1080/10618600.2017.1384734

3 An Evolving and Ever-Changing Landscape

Data science originated in the scientific community. Researchers used statistical and mathematical tools together with other computing software and hardware in order to test hypotheses, collect data, simulate outcomes and predict responses in order to understand the world around us.

3.1 Humble Origins

Around 2012, the tools and skill-sets slowly begun making their presence in organizations, although, arguably, more mature organizations had started leveraging those skills much earlier. It was around this time or a few years later that we heard tongue in cheek definitions of data scientist as someone “*who is better at statistics than any software engineer and better at software engineering than any statistician*”.

Who invented the term? As far back as the early 1960’s, John Tukey called to reform statistical education into a science that learns from data. In 1997, Jeff Wu gave a speech at an inaugural lecture titled “Statistics = data science?”. In 2001, Bill Cleveland suggests the term “data science” in a paper.

3.2 The Domineering Role of Technology

By 2015, technology firms had already figured out that we were at the dawn of a new era and began making large bets in the data science and adjacent spaces. These investments have included acquisitions and sponsorship of open-source projects.

For example, in 2015, Microsoft acquired Revolution Analytics. A few years later, they acquired GitHub while continuing to integrate and expand these technologies in their Azure ecosystem. Additionally, Microsoft continued to make investments in various companies, including OpenAI, or supported and contributed towards open source projects such as Apache Spark (the underlying technology behind Databricks and Microsoft Azure Synapse). Microsoft was not the only player in this space, but it was arguably one of the pioneering large technology firms.

It would not be accurate to say that the data science tools and skills did not exist in government and corporations. Rather, they were not widespread. For instance, the SOUNDEX phonetic algorithm was utilized by the US Census as early as in the 1930s. In 1965, the Fellegi-Sunter theory developed at Statistics Canada, became of fundamental importance in record linkage as it enabled records from different data sources to be joined using non-unique identifiers. The AT&T Bell Labs made so many landmark contributions to scientific computing and statistics, it would impossible to list them all here.

In the current realm of data science, AI and machine learning, technology firms have dominated the discourse, often sidelining important applications and new methods that may not require large-scale solutions or sophisticated technology stacks.

> *…public dialog about these issues too often uses the term AI as an intellectual wildcard, one that makes it difficult to reason about the scope and consequences of emerging technology.*
>
> Michael Jordan

3.3 What Happened to the "Science" in data science?

The word "science", has always been problematic for technology firms and other organizations alike: science relies on experimentation to explore and understand the world, often leading to failures that either yield no new knowledge or challenge our existing hypotheses. Monetizing science is challenging to say the least (ask a pharmaceutical company!): developing cycles can be exhaustingly long, funding it is not for your average venture capitalist, and the outcome is anything but certain.

The principles of data science don't have to be lengthy or costly. They are embedded into the data science process as a way of thinking: we start with extensive conversations with the subject matter experts, conduct literature reviews to understand existing research and build on established theories, prove and disprove our own hunches and expectations in exploratory data analysis, identify sources of uncertainty and measurement error, and precisely define the response in line with the objectives. All of this is a highly iterative process.

In the first phase of data science, approximately from 2015 to 2020, practitioners were more focused on *insight generation* and attuned to the scientific method of inquiry. Indeed, many of the early practitioners were researchers because data science originated in academia and research institutes.

As data science evolved, it became more accessible to practitioners outside of academia, leading to its widespread adoption in industry and other sectors (this is sometimes referred to as *democratization*).

However, with that evolution the scientific foundations became more fragile: an increasing number of practitioners entered the field from backgrounds that did not emphasize research. Many have up-skilled themselves through online courses or boot camps, often focusing on specific technologies stacks.

From my experience in hiring numerous data scientists, and observing the composition of data science teams, the vast majority of applicants for data science roles have a computer science background, which often lacks research methods and a strong statistical foundation. Consequently, the curricula and experience of data scientists increasingly resemble those of developers rather than scientists.

This post on a reddit Data Science conversation illustrates my point:

> *ML in industry really doesn't care about the model chosen, it's more about building good data pipelines, getting your model callable in prod, and getting automated refresh processes.*

Notice the emphasis on 'good' data pipelines, yet, the same quality is not expected from the model. I believe this is not a coincidence: it too reflects my experience working alongside large ML teams. In other words (and at the cost of exaggerating a bit), the quality of a model is not important. What matters is that the model be in production and callable with an API.

But, if the model is of poor quality, aren't we doing a disservice to the industry? If we deploy a poor model in production, attempts at deriving value from it may be seriously compromised. In some cases, the repercussions may be amplified by virtue of having been productionalized.

A rebuttal I often hear is that the model is not used for mission-critical applications. While that may be true, I would argue that it feels more of an excuse to avoid deeper reasoning. Another rebuttal I sometimes hear is that the model will be regularly re-trained. That, again, is no excuse for poor modelling and re-training will do little to fix the underlying problems.

I once witnessed a team spending far more time and resources attempting to shave-off a few seconds of computational time in production than to ensuring a particular metric they were using was sensible. On another occasion, I had to advise a product manager against deploying a product in production because its predictions were severely misleading.

In 2020 the technology company Palantir went public. In its registration statement it stated: "If our customers are not able or willing to accept our product-based business model, instead of a labor-based business model, our business and results of operations could be negatively impacted". This product-based model contrasts with a customized analytics consulting model, which, like the original evolution of data science, can be difficult to monetize from.

In summary, I feel that the evolution towards technology at the expense of sensible and principled methods, towards *good data pipelines* at the expense of *good models* came at the cost of that very *insight* that was the premise of the field. This *product-centric view* of data science was arguably influenced by the software industry (see sidenote on Palantir). Nonetheless, in my view, the need for *insight-focused* approaches in many areas of the organization is no longer being fully met by data science.

3.4 Where Are We Now?

We now find ourselves at the intersection of four significant developments. **First**, the prominent influence of technology firms, promoting a culture of bigger-is-better[2], thus overshadowing important applications of the field. **Second**, the complex relationship between business and science, with its high failure rate and long development cycles, causes organizations to view science with suspicion. **Third**, the democratization of data science, which has introduced many practitioners with limited background in scientific inquiry and principles and procedures that ensure the validity and reliability of their work. **Fourth**, the advent of generative AI, where models are predominantly pre-trained, has led to the deployment of generative AI being more akin to data plumbing.

People coming from traditional software development and IT tend to struggle empathizing with a data scientist. Katie Malone put it eloquently: "[in IT/software development] there's the notion that if you keep working on something, you're going to get there. People who are used to thinking this way can sometimes get pretty confused when a data scientist comes back and says: I'm sorry, this is isn't working and there's nothing we can do about that. That sounds like a foreign idea".

As a result, data science has inevitably evolved to focus more on *data projects* where the challenges are quite different, where finding the *signal* in the data may not even be the point or perhaps it's straightforward (or deterministic), where generating insights may not even be necessary or where the focus is to automate a task at scale.

Data science is increasingly resembling IT, with a mindset focused on completing tasks on time and within budget. We can see tangible signs of this evolution whereby many data science teams have now been brought under the IT function.

[2] Varoquaux, Gaël, Alexandra Sasha Luccioni, and Meredith Whittaker. "Hype, Sustainability, and the Price of the Bigger-Is-Better Paradigm in AI." arXiv, September 21, 2024. https://doi.org/10.48550/arXiv.2409.14160.

3.5 Now What? Is This Bad?

In my experience, organizations have well-justified needs to handle pure data projects. These opportunities are often plentiful, and the main risks involve budget overruns, the establishment of new processes or managing a large team of contractors. Unlike projects with a strong scientific focus, the risks in these projects are manageable as long as sufficient resources are allocated.

On the other hand, there is less and less science in the way data science is leveraged in many organizations. And I wonder if we should stop calling it data science. It's data and it is important, but there is not much science in the way it has evolved.

The field has now become a confusing mixture of several somewhat disjointed fields with only data as a common denominator. Data science jobs often lack the specificity to understand what the job is about. It is not uncommon for many data science positions to be disguised business reporting job posts.

We need to also recognize that despite this evolution, the need for the full breadth of data science still exists in most organizations. There are still opportunities to understand things, solve problems and take decisions based on the evidence the data is providing.

Some organizations make the needs for this type of insight quite explicit. For instance, in A/B testing we aim to understand whether a change, like changing the background colour of a web site, is associated with better outcomes. This is science because it involves identifying a suitable response measure, forming a hypothesis, conducting controlled experiments, and training models to draw evidence-based conclusions. It's also data-centric because another important aspect is to systematically deploy these experiments at scale.

In HR, one might be interested in identifying drivers of turnover or whether systematic pay equity issues exist. This is also science because it requires one to identify and adjust for factors that would otherwise legitimately explain any observed difference. Such projects can be pretty complex because we try to make conclusions from observational data (as opposed to the randomized A/B experiment) and so proper methods of analyses can be difficult to implement. They are also more likely to result in single point decision making. What I mean by that is that the insight is often

One common myth that emerged as data science evolved is the belief that for data science to be valuable, it must result in productionalized machine learning models. However, that is not always where the true value lies. It is important to remember the core principles of data science and its broader potential to uncover insights and drive decision-making.

utilized to design new policies or programs aimed at addressing the problem: there is limited value in deploying a model in production.

4 Conclusions

In conclusion, there is a need for greater clarity in data science roles. First, we should consider reclassifying jobs labelled as data science that involve little to no scientific discovery. Second, we should properly recognize and equally value data science roles that are insight-focused to help businesses uncover insights. Third, data science jobs need clearer definitions, and curricula should include research skills. Hiring managers need to more accurately match relevant skills to the needs of the role.

But the central theme of this essay is that organizations need to understand that there is no one-size-fits-all approach. A data scientist who excels at building good data pipelines may be a good data professional but not necessarily a strong scientist. Conversely, a data scientist who excels at generating valid insights from a model may be a good scientist but not necessarily a strong data professional.

The main value of productionalizing a model is when frequent decisions need to be made. Looking back at the first examples of productionalized models, such as FICO scores, exemplifies this need.

However, not all applications require frequent decisions to be made. And that insight can be just as valuable. Therefore, roles that are focused on insight discovery may not always require the ability to build data pipelines because the value of the model is self-evident once interpreted and properly communicated.

Besides that of frequent decision making, another use for productionalized models is educational. For instance when we need to let the reader play with the output in order to understand a complex problem.