

Machine learning will continue to break boundaries



By [Christiaan Viljoen](#)

14 Jan 2022

A lot of marketing and hype in the artificial intelligence space has resulted in data scientists being touted as modern magicians who can solve any company's toughest problems with those magic two letters: "AI". As a result, recruiters voraciously hunt down mathematically- and statistically gifted individuals - who understand the fine details of how ML algorithms work - but without the hiring company addressing the essential technical prerequisites to enable AI in the first place.



Christiaan Viljoen, ML-Ops engineer, Teraflow

That is a mature data engineering stack, model CI-CD, model monitoring and all the other necessary ML-Ops elements which ensure that this new, slightly unpredictable component of your software stack behaves as it should.

This mismatch between expectations and reality often ends in mutual disillusionment: 1) for the companies hiring the data scientists, which don't see the massive ROI they were expecting; and 2) for the data scientists themselves, who find that their shiny models might never make it into production.

This disillusionment is, however, entirely preventable, since it usually arises simply because the company's tech stack isn't yet mature enough to allow for such complexity and because the company has not taken the time to generate a decent ML roadmap or an overarching data strategy (ie. instead of saying AI is cool, let's find a business use-case for it, one should instead employ this perspective: we have a specific business use-case, which could benefit from having some AI elements baked into it, let's find the right AI solution to integrate with it).

The sad fact however remains, that many brilliant data scientists end up doing data engineering and business intelligence work instead of ML... and let's face it, a couple of broken BI dashboards are a lot less sexy than a CEO's grandiose vision

of a company run entirely by robots.

Auto-ML and transfer learning beats your in-house human data science team

While “pure-bred PhD data scientists” might look down on Auto-ML, it is not uncommon for Auto-ML to beat a custom-built model that a data scientist spent weeks perfecting. Auto-ML conducts hundreds of automated experiments with multiple ML algorithms or even ensembles of algorithms, as well as automating various tedious processes such as feature engineering, hyperparameter tuning and cross-validation.

This study in 2020 concluded that Auto-ML was able to achieve better or the same results as human data scientists on seven out of 12 OpenML tasks. One can only wonder what a similar study will find at the end of 2022, but with so many large research teams pouring money into this research area, it's not looking great for humans.

Today, state of the art neural network architectures and weights can also be saved and made publicly available by experts through online model catalogues such as the TensorFlow Model Garden. This means one can quite easily download and “freeze” the lower layers of a highly performant model built by an expert and only train the last layer or two for your specific dataset and use-case, allowing for much faster time-to-market than hiring a bespoke ML research team for the intensive R&D work required to build a model from scratch.

Given these trends, more and more companies should shift focus towards the integration of machine learning models into their existing software stack, instead of necessarily spending money on hiring bespoke ML researchers. ML in Industry is often more of a software problem than a math or stats problem.

Academics vs industry: ML-Ops tools knowledge is more important than actual ML knowledge

There is a widening gap between what is taught in academic programmes and what is actually relevant in the industry today: dusty old academics teaching “the data scientists of the future” aren't able to keep up with the pace of change in the ML space in Industry.

This does not mean that advances made in academia don't eventually make their way into the latest tech (think: “they make the honey and we make the money”), but there is a huge difference between being an ML researcher (an academic in the space of ML) and an ML practitioner (an ML engineer, data scientist or ML-ops engineer in industry).

Having in-depth knowledge of managed ML-ops platforms, such as Amazon SageMaker, Google Cloud's Vertex AI and Azure Machine Learning, as well as managed data engineering platforms such as AWS Glue, DataFlow and Data Factory and the attendant Big Data tools, such as Apache Spark, Apache Beam, Apache Kafka, TensorFlow, PyTorch and many more, should make one better equipped for ML in Industry than an advanced degree in mathematics, statistics or even data science likely ever will.

The abovementioned ML-Ops and data engineering platforms will play an increasingly central role because they provide a managed environment in which all the aspects of modern ML have already been thought of; i.e. all major concerns (scalability, monitoring, fault-tolerance, security, etc.) are already being addressed.

As an added benefit, Cloud ML-Ops platforms are mostly “serverless” technologies, providing a layer of abstraction away from managing the underlying hardware or installing the software dependencies required in the ML lifecycle.

ABOUT CHRISTIAAN VILJOEN

ML-Ops Engineer at Teraflow

■ #BizTrends2022: Machine learning will continue to break boundaries - 14 Jan 2022

[View my profile and articles...](#)

For more, visit: <https://www.bizcommunity.com>