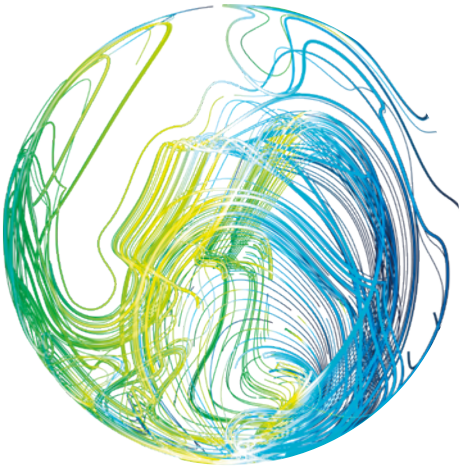
Deloitte.



Machine learning: things are getting intense

Deloitte Global predicts that in 2018, large and medium-sized enterprises will intensify their use of machine learning. The number of implementations and pilot projects using the technology will double compared with 2017, and they will have doubled again by 2020. Further, with enabling technologies such as ML application program interfaces (APIs) and specialized hardware available in the cloud, these advances will be generally available to small as well as large companies.

ML is an artificial intelligence (AI), or cognitive, technology that enables systems to learn and improve from experience – by exposure to data – without being programmed explicitly.

Despite the excitement over ML and cognitive technologies, and the aggressive forecasts for investment in these technologies, most enterprises using ML have only a handful of deployments and pilots under way. According to a 2017 Deloitte Consulting LLP survey of executives in the US who said their companies were actively using cognitive technologies and were familiar with those activities, 62 percent had five or fewer implementations and the same number of pilots under way.[87]

But progress in five key areas should make it easier and faster to develop ML solutions while also removing some of the barriers that have restricted adoption of this powerful technology. Progress along these vectors should lead to greater investment in ML and more intensive use within enterprises. This in turn should cause enterprises to double the number of ML pilots and deployments by the end of 2018. By then, over two-thirds of large companies working with ML may have 10 or more implementations and a similar number of pilots.

Analysts are predicting strong growth in investment and adoption of ML globally. International Data Corporation (IDC) forecasts that spending on AI and ML will grow from $12 billion in 2017 to $57.6 billion by 2021.[88] But adoption of ML is still in its early phases.

Deloitte Consulting LLP recently surveyed "cognitive-aware" executives in the US at companies that are active in cognitive computing with at least 500 employees. Half of the respondents worked for companies with 5,000 or more employees. Qualifying respondents had a moderate or better understanding of the technology and were familiar with their company's use of it.

While respondents were highly enthusiastic about the potential of cognitive technologies, the majority (60 percent) had just a handful of implementations and pilots per company under way.[89]

What has held back the adoption of ML? Qualified practitioners are in short supply.[90] Tools and frameworks for ML work are immature and still evolving.[91] It can be difficult, time-consuming and costly to obtain the large data sets required by some ML model-development techniques.[92] Even when they work well, some ML models are not deployed in production, as their inner workings are inscrutable and some executives will not run their business on systems they do not understand. Others may be constrained by regulations that require businesses to provide explanations for their decisions or to prove that decisions do not discriminate against protected classes of people.[93] Black-box models, no matter how accurate or useful their outputs, cannot be deployed in such situations.

However, Deloitte Global has identified five key vectors of progress in ML that should unlock more intensive use of the technology in the enterprise.

Three of these five advancements – automation, data reduction and training acceleration – make ML easier, cheaper or faster (or some combination thereof). They will have the effect of expanding the market for ML. The others – model interpretability and local ML – enable applications in new areas, which should also expand the market.

[87] 2017 Deloitte State of Cognitive Survey, November 2017: https://www2.deloitte.com/us/en/pages/deloitte-analytics/articles/cognitive-technology-adoption-survey.html#.

[88] IDC Spending Guide Forecasts Worldwide Spending on Cognitive and Artificial Intelligence Systems to Reach $57.6 Billion in 2021, IDC, 25 September 2017: https://www.idc.com/getdoc.jsp?containerId=prUS43095417.

[89] Deloitte study op. cit. 2017 Deloitte State of Cognitive Survey.

[90] For a discussion of supply and demand of data science skills, see The quant crunch: How the demand for data science skills is disrupting the job market, IBM, accessed 6 November 2017: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IML14576USEN&.

[91] This article provides a perspective on the early stage of development of ML tools: The evolution of machine learning, TechCrunch, 8 August 2017: https://techcrunch.com/2017/08/08/the-evolution-of-machine-learning/.

[92] For a discussion of the challenge and some ways around it, see Weak supervision: The new programming paradigm for machine learning, Stanford University, 16 July 2017: http://dawn.cs.stanford.edu/2017/07/16/weak-supervision/.

ML continues to improve in other ways as well and is evolving so rapidly that another key improvement is likely to arise over the course of the year.

Our top five vectors of progress – ordered by breadth of application, with the widest first – are detailed below.

1. Automating data science. Time-consuming ML tasks, such as data exploration and feature engineering, which typically take up as much as 80 percent of a data scientist's time, can increasingly be automated.[94]

Data science, an often misunderstood, specialist discipline, is in reality a blend of art and science. Much of what data scientists spend time on – from data wrangling to exploratory data analysis, feature engineering, feature selection, predictive modeling, model selection and so on – can be wholly or partially automated. For instance, while building customer lifetime value models for guests and hosts, data scientists at Airbnb used an automation platform to test multiple algorithms and feature engineering steps – which they would not otherwise have had the time to do. Automation enabled them to discover changes they could apply to their algorithm that increased accuracy by more than five percent – a significant impact.[95]

A growing number of tools and techniques for data science automation, offered by established companies as well as venture-backed start-ups, should help shrink the time required to execute an ML proof of concept from months to days.[96] Automating data science means data scientists can be far more productive.
It thereby helps overcome the acute shortage of data scientists, enabling enterprises to double their ML activities.

2. Reducing the need for training data. Training an ML model can require up to millions of data elements. This can be a major barrier. Acquiring and labeling data to be used for training can be highly time-consuming and costly. Consider, as an example, a project that requires MRI images to be labeled with a diagnosis. It might cost over $30,000 to hire a radiologist to review and label 1,000 images at a rate of six images an hour. Privacy and confidentiality concerns can also make it difficult to obtain the data in the first place.

But a number of promising techniques are emerging that aim to reduce the amount of training data required for ML. One involves the use of synthetic data, generated algorithmically to mimic the characteristics of the real data.[97] A team at Deloitte Consulting LLP tested a tool that was able to build an accurate model with only a fifth of the training data previously required; it synthesized the remaining 80 percent of data.

Synthetic training data can also open the door to the crowdsourcing of data science solutions. A number of organizations have engaged third parties to devise ML problem-solving models and are posting data sets appropriate for sharing that outside data scientists can work with.[98] Researchers at MIT used a real data set to create synthetic alternatives that could be used to crowdsource the development of predictive models without needing to disclose the original data set. In 11 out of 15 tests, the models developed from the synthetic data vault performed as well as those trained on real data.[99]

Another technique that could reduce the need for training data is transfer learning. With this approach, an ML model is pre-trained on one data set as a shortcut to learning a new data set in a similar domain, such as language translation or image recognition. Some ML tool vendors claim their use of transfer learning can cut the number of training examples customers need to provide by several orders of magnitude.[100]

3. Accelerating training. As detailed in the prediction *Hitting the accelerator: the next generation of machine-learning chips*, established and start-up hardware manufacturers are developing specialized hardware (such as GPUs, FPGAs and ASICs) to slash the time required to train ML models, by accelerating the calculations required and the transfer of data within the chip. These dedicated processors can help companies speed up ML training and execution manyfold, which in turn brings down the associated costs.

For instance, a Microsoft research team using GPUs completed
a system in one year to recognize certain conversational speech
as capably as humans could. With CPUs, it would have taken
five years.[101]

Google stated that designing its own AI chip, a TPU, for neural networks execution and adding TPUs to CPU and GPU architecture helped the company save the cost of building a dozen extra data centers.[102]

[93] For a high-level discussion of this challenge, see The business case for machine learning interpretability, Fast Forward Labs, 2 August 2017: http://blog.fastforwardlabs.com/2017/08/02/business-interpretability.html; for a discussion of how the European Union's new General Data Protection Regulation effectively creates a "right to explanation" that will increase demand for interpretable algorithms and models, see European Union regulations on algorithmic decision-making and a "right to explanation" report, Bryce Goodman and Seth Flaxman, Cornell University, 31 August 2016: https://arxiv.org/pdf/1606.08813.pdf.

[94] For a discussion of ML automation, see Driverless AI blog, H2O.ai, 13 July 2017: https://blog.h2o.ai/category/automl/.

[95] Automated machine learning – paradigm shift that accelerates data scientist productivity @ Airbnb, Medium, 10 May 2017: https://medium.com/airbnb-engineering/automated-machine-learning-a-paradigm-shift-that-accelerates-data-scientist-productivity-airbnb-f1f8a10d61f8.

[96] For a partial list, see Automated data science and data mining, KDnuggets, 4 March 2016: https://www.kdnuggets.com/2016/03/automated-data-science.html; as of October 2017, one start-up in this area, DataRobot, had raised $125 million from venture investors. Google has introduced machine learning modeling techniques called AutoML. Using machine learning to explore neural network architecture, Google, 17 May 2017: https://research.googleblog.com/2017/05/using-machine-learning-to-explore.html.

[97] New resources for deep learning with the Neuromation platform, Medium, 9 October 2017: https://medium.com/neuromation-io-blog/new-resources-for-deep-learning-with-the-neuromation-platform-55fd411cb440.

[98] 9 Reasons to crowdsource data science projects, InformationWeek, 19 February 2016: https://www.informationweek.com/big-data/big-data-analytics/9-reasons-to-crowdsource-data-science-projects/d/d-id/1324377.

Early adopters of these specialized AI chips include major technology vendors and research institutions in data science and ML, but adoption is spreading to sectors such as retail, financial services and telecom. With GPU cloud computing offered by all the major cloud providers (IBM, Microsoft, Google, AWS), accelerated training should become mainstream, increasing the productivity of teams working on ML and multiplying the number of applications enterprises choose to undertake.

4. Explaining results. ML achievements get more impressive by the day. But ML models often suffer from a critical flaw: many are black boxes, meaning it is not possible to explain with confidence how they make their decisions. This makes them unsuitable or unpalatable for many applications, for reasons ranging from trust in the answers generated by a model – as when customers are offered incentives – to regulatory compliance. For example, the US financial services industry adheres to the Fed's Supervisory Letter, SR 11-7, Guidance on Model Risk Management, which among other things requires that model behavior be explained.[103]

A number of techniques have been created that help shine light into the black box of certain ML models, making them more interpretable and accurate. MIT researchers have demonstrated a method of training a neural network that delivered accurate predictions and the rationales for those predictions.[104]

Some techniques are finding their way into commercial data science products, such as H2O Driverless AI, a data science automation platform;[105] DataScience.com's new Python library, Skater;[106] and DataRobot's ML-powered predictive modeling for insurance pricing.[107] As it becomes possible to build interpretable ML models, companies in highly regulated industries such as financial services, life sciences and health care can be expected to intensify their use of ML and significantly expand the number of pilots and deployments over coming years.

Some of the potential applications include credit scoring, recommendation engines, customer churn, fraud detection, and disease diagnosis and treatment.[108]

5. Deploying locally. ML use will grow along with the ability to deploy it where it is needed. As we predicted last year, ML is increasingly coming to mobile devices and smart sensors, expanding the technology's applications to smart homes and cities, autonomous vehicles, wearable technology, and the industrial Internet of Things.[109]

Technology vendors including Google, Microsoft, Facebook and Apple are creating compact ML software models to undertake tasks such as image recognition and language translation on portable devices. Google is using TensorFlow Lite, Microsoft has an embedded learning library, Facebook has Caffe2Go and Apple Inc. is using Core ML for on-device processing.[110] Microsoft Research Lab's compression efforts resulted in ML models that were 10 to 100 times smaller.[111]

Semiconductor vendors including Intel, Qualcomm and Nvidia, as well as Google and Microsoft, are developing their own power-efficient AI chips to bring ML to mobile devices.[112] With smartphones an increasingly viable deployment option for ML, the number of potential applications is growing, and the number of enterprise ML pilots and deployments will rise too.

99. Artificial data give the same results as real data – without compromising privacy, Massachusetts Institute of Technology, 3 March 2017: http://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303.

100. For instance, see Extract insight from data with indico's API, Indico, as accessed on 7 November 2017: https://indico.io/product?api=custom.

101. How AI is shaking up the chip market, Wired, 28 October 2016: https://www.wired.com/2016/10/ai-changing-market-computer-chips/.

102. Building an AI chip saved Google from building a dozen new data centers, Wired, 4 May 2017: https://www.wired.com/2017/04/building-ai-chip-saved-google-building-dozen-new-data-centers/.

103. Supervisory guidance on model risk management, The Federal Reserve System, 4 April 2011: https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf.

104. Making computers explain themselves, Massachusetts Institute of Technology, 27 October 2016: http://news.mit.edu/2016/making-computers-explain-themselves-machine-learning-1028.

105. Machine-learning interpretability with H2O driverless AI, H2O.ai, September 2017: https://www.h2o.ai/wp-content/uploads/2017/09/MLI.pdf.

106. DataScience.com releases python package for interpreting the decision-making processes of predictive models, DataScience.com, 23 May 2017: https://www.datascience.com/newsroom/datascience-releases-skater-python-package-for-predictive-model-interpretation.

107. New DataRobot release extends enterprise readiness capabilities and automates machine learning in insurance industry pricing models, DataRobot, 24 July 2017: https://www.datarobot.com/news/new-datarobot-release-extends-enterprise-readiness-capabilities-and-automates-machine-learning-in-insurance-industry-pricing-models/.

108. For more information, see Research FF06 – Interpretability, Fast Forward Labs, July 2017: https://www.fastforwardlabs.com/research/FF06.

Definitions and explanations – a layperson's guide

Data science: An interdisciplinary field that generally employs data management, analytics modeling and business analysis to gain insight from complex data sets that are often very large or unstructured.

Training data: Used to discover and model a relationship between a set of data inputs and a corresponding set of data outputs, or labels. For example, records of home sales might include three attributes, such as square footage, year of construction and school district, as the inputs, and the sale price as the output. An algorithm would be used to discover a relationship between those three attributes and the sale price. Capturing that relationship in a model might make it possible to predict the sale price for other homes when only those three input attributes are known. The use of training data to create or learn such a model from training, or labeled, data is known as supervised machine learning.

Black box: Anything with inner workings that are not apparent. A black-box ML model produces answers – such as medical diagnoses or credit underwriting decisions – without explaining the rationale. A white-box model, by contrast, would reveal its inner workings, making it possible to understand how it arrives at its results.

Interpretability: In this context, the ability to explain why and how a system makes a decision.[113]

Data wrangling: The process of cleaning and sorting complex, unstructured data sets for ease of use and analysis.

Data exploration: The first step in data analysis to understand the data set and to summarize key characteristics of the data.

Feature engineering: The process of using domain knowledge to create relevant features of the data in a tabular format, from the existing raw features, for an ML model.

Neural networks: Includes layers of interconnected nodes, inspired by neurons in the human brain, to perform a form of ML in which the system learns to perform a task by analyzing training data on its own.

[109] See Deloitte Global's TMT Predictions 2017 – Brains at the edge: machine learning goes mobile, 14 January 2017: https://www2.deloitte.com/us/en/pages/technology-media-and-telecommunications/articles/tmt-predictions.html.

[110] Google's TensorFlow Lite brings machine learning to Android devices, TechCrunch, 17 May 2017: https://techcrunch.com/2017/05/17/googles-tensorflow-lite-brings-machine-learning-to-android-devices/; Microsoft wants to bring AI to Raspberry Pi and other tiny devices, ZDNet, 30 June 2017: http://www.zdnet.com/article/microsoft-wants-to-bring-ai-to-raspberry-pi-and-other-tiny-devices/; Facebook open sources Caffe2, its flexible deep learning framework of choice, TechCrunch, 18 April 2017: https://techcrunch.com/2017/04/18/facebook-open-sources-caffe2-its-flexible-deep-learning-framework-of-choice/; Apple announces new machine-learning API to make mobile AI faster, The Verge, 5 June 2017: https://www.theverge.com/2017/6/5/15725994/apple-mobile-ai-chip-announced-wwdc-2017.

[111] Microsoft wants to bring AI to Raspberry Pi and other tiny devices, ZDNet, 30 June 2017: http://www.zdnet.com/article/microsoft-wants-to-bring-ai-to-raspberry-pi-and-other-tiny-devices/.

[112] The rise of AI is forcing Google and Microsoft to become chipmakers, Wired, 25 July 2017: https://www.wired.com/story/the-rise-of-ai-is-forcing-google-and-microsoft-to-become-chipmakers/; Apple is working on a dedicated chip to power AI on devices, Bloomberg, 27 May 2017: https://www.bloomberg.com/news/articles/2017-05-26/apple-said-to-plan-dedicated-chip-to-power-ai-on-devices.

[113] For more information, see Research FF06 – Interpretability, Fast Forward Labs, July 2017: https://www.fastforwardlabs.com/research/FF06.

The bottom line

Collectively, the five vectors of ML progress should double the intensity with which enterprises are using this technology by the end of 2018. In the long term, these vectors should help make ML a mainstream technology. Advances will enable new applications across industries where companies have limited talent, infrastructure or data to train the models.

Companies should:

- Look for opportunities to automate some of the work of their oversubscribed data scientists, and ask consultants how they can use data science automation.

- Keep an eye on emerging techniques, such as data synthesis and transfer learning, that could ease the bottleneck often created by the challenge of acquiring training data.

- Find out what computing resources optimized for ML are offered by their cloud providers. If they are running workloads in their own data centers, they may want to investigate adding specialized hardware to the mix.

- Explore state-of-the-art techniques for improving interpretability that may not yet be in the commercial mainstream, as interpretability of ML is still in its early days.

- Track the performance benchmarks being reported by makers of next-generation chips, to help predict when on-device deployment is likely to become feasible.

Deloitte.

This communication contains general information only, and none of Deloitte Touche Tohmatsu Limited, its member firms, or their related entities (collectively, the "Deloitte Network") is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that might affect your finances or your business, you should consult a qualified professional adviser. No entity in the Deloitte Network shall be responsible for any loss whatsoever sustained by any person who relies on this communication.

Designed and produced by The Creative Studio at Deloitte, London. J14285