Corrected: Author Correction

Machine learning, practically speaking

To apply machine learning, labs needn't have years of computational expertise, but they do need a cautious mind-set.

Vivien Marx

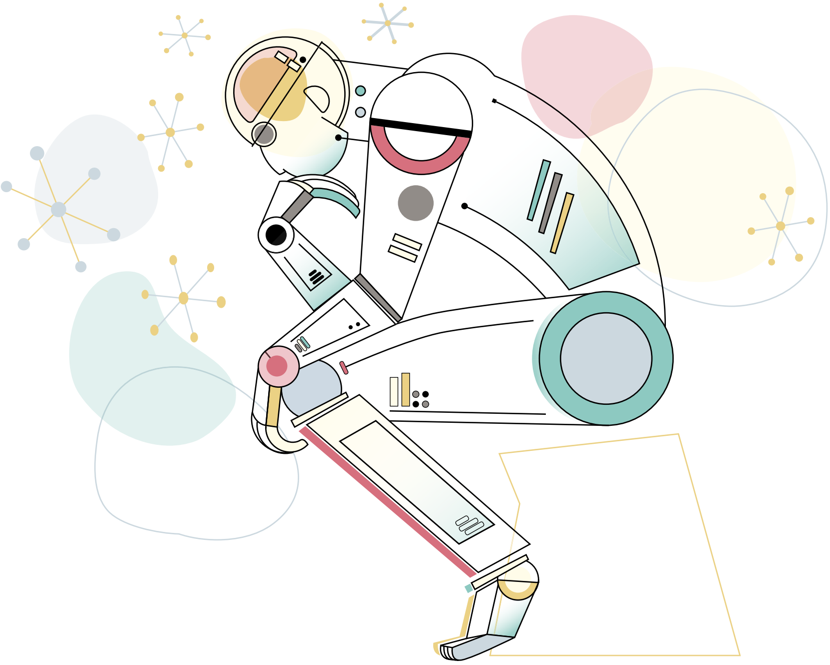
'Artificial intelligence' (AI) is hard to beat as an enigmatic term. Within the AI field, many projects involve machine learning (ML), in which a computer can learn iteratively from data and make predictions[1]. Basic ideas of ML architectures have been around for decades and have fallen in and out of favor, says Don Geman of Johns Hopkins University. What has changed of late is the availability of massive amounts of labeled data, such as faces, which can be used to train ML systems. Computing power has grown because of graphics-processing units. "Those two things allowed basically the same methodology to suddenly work far better than it had before and far better than anything else was working," he says.

This year's Association for Computing Machinery's Turing Prize, nicknamed the 'Math Nobel', went to three scientists who have been called "fathers of deep learning" and the "godfathers of AI": Geoffrey Hinton of the University of Toronto, who is also part of the Google Brain team; Yann LeCun of New York University, who is chief artificial intelligence scientist at Facebook; and Yoshua Bengio of the University of Montreal.

ML-based systems can win games, even beat the world Go champion; they power many aspects of Facebook and Google's operations. ML can discern faces in photos, translate text, target ads on the web, power autonomous driving on Earth and Mars.

These systems have various architectures and algorithms that all need training in which they are presented with the 'correct' answer, such as slides of breast cancer tissue of the luminal A molecular subtype. Algorithms extract features and tweak inner parameters to model the data and learn to generalize from them so they can, for example, assess whether tissue slides the system has not 'seen' before show features of luminal A breast cancer.

Deep learning can be used when, for example, linear regression algorithms do not suffice to model complex data. Algorithmic processing takes place in many nodes, and learning tunes each node's parameters or 'weights'. The nodes are organized in layers: calculations from one layer become input to the next layer.



Machine learning (ML) architectures have been around for decades. They've fallen in and out of favor. Now, there is heightened interest in applying ML in biomedicine. Credit: E. Dewalt/Springer Nature

As learning progresses, a deep learning system adapts its internal parameters, sometimes millions of them, to accurately map input data to output predictions, says Stanford University computer scientist Jure Leskovec. Well-trained machines can generalize beyond training data to new data and find patterns humans miss or simply can't see.

ML is arriving in biomedical research labs, and toolkits abound[2–5]. Excitement and enthusiasm about ML have drawn researchers in and are leading many beyond the piling of ever-higher data mountains to creative approaches that reliably "put those mountains to work," says Casey Greene, a computational biologist at the University of Pennsylvania Perelman School of Medicine. In the past, says Leskovec, labs could 'see' all data, but now that instruments deliver such data piles, ML and data science "are the only ways we will be able to 'see' and 'understand' the data to reveal new discoveries."

Excitement and caution

ML projects might involve training a system to find and classify patterns indicative or predictive of disease in images or gene expression data, to predict protein structures from genetic sequence or to design chemical scaffolds in drug discovery. MIT computer scientist Regina Barzilay likes seeing how popular and modular deep learning frameworks for building ML systems, such as PyTorch or Google's TensorFlow, have become. "Now you have the big Lego blocks and you can put it together," she says. Collaborating with computer scientists is still advisable to better understand what the system does, "but you can start using some of these methods even though you are not expert in them," says Christos Davatzikos of the University of Pennsylvania Perelman School of Medicine.

But Barzilay sees some biomedical researchers try AI, make big claims that don't materialize and then turn their backs on these methods. Perhaps, she says, they

To see how "data-hungry" an ML method is, labs can feed a network iteratively, says Regina Barzilay.

forget that ML is unlike electricity, which "I can just plug it in and it works." Geman and Barzilay say they see too many published papers in which labs describe training on small datasets and claim high prediction accuracies.

Libraries such as Selene from the Troyanskaya lab at Princeton University[6] help to overcome some technical hurdles of deploying deep learning models with massive sequence datasets, says Marinka Zitnik, postdoctoral fellow in the Leskovec lab. The library offers "a unified interface" to a number of sequence-based deep models, she says, which lets labs compare models for the same prediction task, standardizes their use to interpret functions of genetic variation and can speed up the development of new models.

Greene initiated a crowdsourced Deep Review[7] as his lab tinkered with a type of neural network architecture that had not yet been applied to biology but seemed to offer useful properties for the team's transcriptomic data. "I wanted some colleagues to help read the literature with me because there were a bunch of examples of what felt like hype and fewer examples that felt truly transformative," he says. The project has made him feel more positive about ML methods. Generally, he would prefer to see fewer published benchmarking tests that characterize many ML methods across a "modest" number of datasets, and more emphasis on methods that learn different types of signals but still provide "reasonable performance," which is problem-specific.

When constructed, trained and used properly, ML is not a "black box," says Christos Davatzikos.

The optimization of complex models and their many parameters requires training with large datasets, says Barzilay. Labs should know where they are driving their ML system for a given question. The tools will always tell you something, she says, but it can be a meaningless something. For some questions, classic statistical methods might prove more efficient. She advises a cautious mind-set when applying ML.

When ML sees

ML is good at processing images, says Geman, and such systems can learn to distinguish malignant and nonmalignant skin spots, for example. "There are many problems where deep learning could have a nice impact," he says. Combining the genomic correlates of an image is also "really exciting," he says.

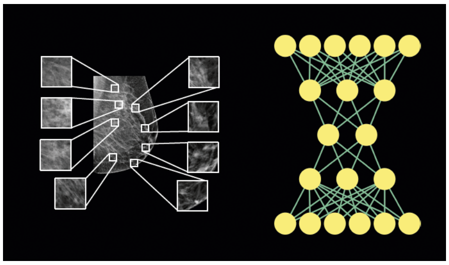
Imaging reveals morphology and physiology, which are used in basic research and patient treatment, says Davatzikos, who organized an ML session at this year's annual meeting of the American Association for Cancer Research (AACR). Radiology meets genomics in the emerging field of radiogenomics, in which labs explore how ML techniques help with finding subtle signatures such as important phenotypic indications of a tumor's genome. In glioblastoma, that can be a mutation in the *IDH1* gene or methylation of the *MGMT* promoter. "It's a bit of a surprise how much information is hiding in those images," he says. Such "subvisual" molecular composition patterns still must be confirmed in assays, but they indicate what a machine can 'see' that scientists and clinicians cannot. Ideally, one would want to tag every image voxel, but "we're not there yet."

To help scientists extract potentially thousands of features from images of brain, breast and lung cancer samples, Davatzikos and colleagues developed the Cancer Imaging Phenomics Toolkit (CaPTk), through which users can tap into some commonly used ML-tool libraries. It's one of several toolkits the National Cancer Institute (NCI) has funded.

To find training datasets, labs can mine The Cancer Genome Atlas or The Cancer Imaging Archive, but they might not always find what they need, says Davatzikos. Along with 12 institutions around the world, he is building a glioblastoma dataset. They have 500 glioblastoma cases; "hopefully we'll get around 3,000 datasets or so."

Large-scale datasets such as ImageNet help those hunting for training data, but in biomedicine, such datasets are just emerging. "Big data—the power and potential it has to change practice has been a focus of mine at the NCI," said Ned Sharpless in his keynote at the annual AACR meeting. He is leaving his post as NCI director to lead the Food and Drug Administration. The Cancer Genome Atlas, an NCI 'big data' foray, has become the NCI Genomic Data Commons, which by the end of 2019 will hold data from over 70,000 patients. Other data types to be added include proteomics data and radiology, pathology and clinical annotations so that it is a "multimodal dataset," Sharpless said.

Barzilay looks forward to computer scientists, biologists and clinicians collaborating to build biomedical datasets at scale to enable better ML models. In her presentations, she speaks about how her breast cancer was missed in her mammogram. (She has been successfully treated.) Along with the radiology department at Massachusetts General Hospital, she has developed, validated and deployed an ML system to help radiologists detect patterns indicative of cancer. She engineered existing ML approaches to account for common variance between images and draw conclusions about malignancy. Next, the team wants to integrate genomic with radiology data.



With the radiology department at Massachusetts General Hospital, MIT computer scientist Regina Barzilay has set up an ML system to help detect patterns indicative of cancer. Credit: Barzilay lab, MIT & MGH

Separately with colleagues in several MIT departments, she co-leads Machine Learning for Pharmaceutical Discovery and Synthesis, which involves over a dozen biopharmaceutical companies and is about using ML to automate chemical syntheses. ML tools work well when a lab knows there is a strong pattern in the data and when training data are plentiful, says Barzilay. She has met frustrated researchers who design ML systems trained on one set of structures but then do not generalize to a different "chemical space." It's possible, she says, but one needs diversity in the training dataset.

ML needs

When choosing training data, "you need to make sure that the distribution of your

To make predictions more robust, says Luigi Marchionni, scientists can use biological information to select features for training data.

Credit: AACR

training data is similar to the scenario on which you are going to be testing," says Barzilay. Labs can assess their data with a "learning curve." To see how "data-hungry" a method is, they can divide training data in half and 'feed' the network iteratively. "Look at how the increase in data increases your accuracy," she says.

Labs need to be sure about data quality. Some datasets might have been collected with exclusions of sorts. A dataset can be biased in many ways, and that can make the ML system biased.

When using ML with small datasets or when a lab wants to apply ML on quite different data than the training data, "you really need to tread carefully," says Barzilay, and work on the "mathematical machinery." Labs can try transfer learning, says Greene, in which training data from one domain are used for predictions in another. But the path between the two domains needs to be solid.

In some situations, it can be advisable to steer clear of ML, says Greene, such as when a lab has insufficient data or not enough similar data for training, or if there is no way to augment data. Labs can use tools such as PyTorch and others when they have the right type of data on hand, the amount of which will be problem-dependent, and when it's obvious what a suitable neural network architecture might be. Even with enough data, it can be unclear what ML approach is suitable, which is when it's prudent to reach out to a computer scientist.

When constructed, trained and used properly, ML is not a "black box," says Davatzikos. To train and debug, says Leskovec, labs using TensorFlow can apply TensorBoard to visualize their network's architecture, plot quantitative metrics and see the data passing through the network. Parameters can be observed in real time as they change, and one can see the predictions getting better. Barzilay also sees no 'black box' risk, given that labs can query a deep learning system so that it reveals the data applied to a prediction.

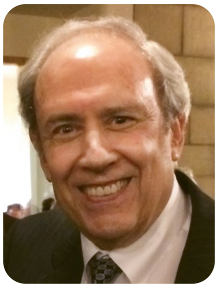
Overfitting happens when an ML system cannot generalize beyond a training dataset, says Geman. It's a common issue when ML is applied to molecular data, such as exome sequencing, RNA sequencing or transcriptional profiles, says Luigi Marchionni, who is also on the Hopkins faculty and who collaborates with Geman. The data have many features and, he says, "what's going to happen with these algorithms, they're going to overfit." Noise will look like signal, the model won't generalize and predictions won't hold up to scrutiny.

Researchers should maintain a special level of suspicion about ML when a system performs extremely well, says Greene. "It can be helpful to analyze real data side-by-side with permuted data," he says. "If the model is starting to look 'good' on the permuted data, something is wrong." He and his team explored whether there is an ideal algorithm or ideal dimensionality to solve specific problems when feature construction is decoupled from the supervised learning, which might be when a lab seeks "universal" features that could be reused across many tasks. But, in his view, "there is no method that completely dominates other methods," he says. "There's also no single dimensionality that was the optimal for all problems."

Watch out for the curse

To make predictions more robust, says Marchionni, one can use biological information to select features for the training data. In cancer, that might be the topology of gene regulatory networks or metabolic data. He and Geman are working on algorithms that can capture general biological mechanisms, which they call "mechanism-driven classifiers." Biologists would predetermine where, in the dark room of their data, they want the ML system to "shine a light," says Geman. This is, he says, a more promising way to use ML with molecular data and with an eye toward uses such as predicting drug response on the basis of molecular signatures and other data.

The high-dimensional data in biomedicine—genomics, therapeutics, environmental data and others—can lead to more complex diagnostic and prognostic categories than currently in use, says Leskovec. New tools are needed to analyze such complex and diverse data. He, Zitnik and colleagues have developed SNAP, an algorithm toolbox for handling complex and multimodal network data. The tools can scale to networks with billions of interactions and thousands of modalities. Their BioSNAP is a public repository of high-quality, rich biomedical interaction datasets. The resources can be used for algorithm development and benchmarking, he says.



Overfitting happens when an ML system cannot generalize beyond a training dataset, says Don Geman.

One haunting issue with ML and molecular data is the so-called curse of dimensionality. Geman says he is often approached by biologists and clinician-scientists. Can the proficiency of ML for discerning cats from dogs in photos transfer to predicting cancer prognosis on the basis of gene expression data? "Shouldn't the same methods work?" they ask. No, says Geman, the challenges are "radically different." The cat-versus-dog scenario requires 1 million images for training. A cancer research lab might have data from only 50 patients, not 1 million patients. And they're high-dimensional, heterogeneous data.

Omics data in basic research and medicine can contain 1 million dimensions, such as gene species, splice variants, different RNAs. But perhaps only around 100 of them might have bearing on a phenotype of interest that a lab is making predictions about. "We don't know which ones are informative for the task at hand," says Geman, and the signal is likely to be weak.

Classic statistics is powerful, he says, when n, the sample size, is much larger than the number of dimensions, d. When n is closer to d, perhaps within an order of magnitude, standard ML can come into play. But with cancer prediction and omics, n is much smaller than d: n is in the tens or hundreds, and d can be 10,000 or 1 million. It's kind of a "worst-case scenario for machine learning," says Geman. "It's so difficult to find the signal," he says, and there are no off-the-shelf solutions.

Dogs and wolves

Hinton, LeCun and Bengio have pointed out that when an ML system processes images or language, input is often complex[1]. Photos of the same object shot at different angles or against different backgrounds need to not throw off a system. And subtle differences need to be detected, such as those between a type of dog called a Samoyed and a white wolf, which, in photos, can look similar. When they apply ML to molecular data and 'personalized medicine', says Marchionni, labs are trying to tell individual dogs and wolves apart and make decisions about how to act with each. Instead of 1 million images, they have a handful of blurry images, including ones of unknown animals, and they need to discern a particular, well-behaved Samoyed from a rowdy one that runs off for days at a time to raid a barn, and from a specific white wolf near a barn in Yakutsk, Siberia.

With small sample size and large dimensionality, says Geman, complexity reduction can help, but it must be done with care, given that this step risks the loss of important information. Because of these fundamental issues, he worries that toolkits used by those less experienced in ML might not yield valid results. He is concerned that the field "may be trending in the wrong direction." He advocates working in cross-disciplinary groups. "I think of each discipline as exposing the fantasies of the other," he says. At Hopkins, his collaborators include cancer researchers such as the Vogelstein lab, computational biologists including Marchionni, applied mathematicians and others.

With ML, labs can find that their "real problem" is the weak signal in their data, says Geman. To avoid swamping their millions of measurements with false positives, "you're going to have to bring mechanism into the story, biological knowledge." To apply ML in biomedicine, one needs a problem-driven methodology. "There's no off-the-shelf solution," he says. Packages certainly help, but he sees too many papers reporting high prediction accuracies, and few stand up to rigorous follow-up, which is needed for reproducibility and for these tools to eventually affect clinical practice.

Machines hunt drugs

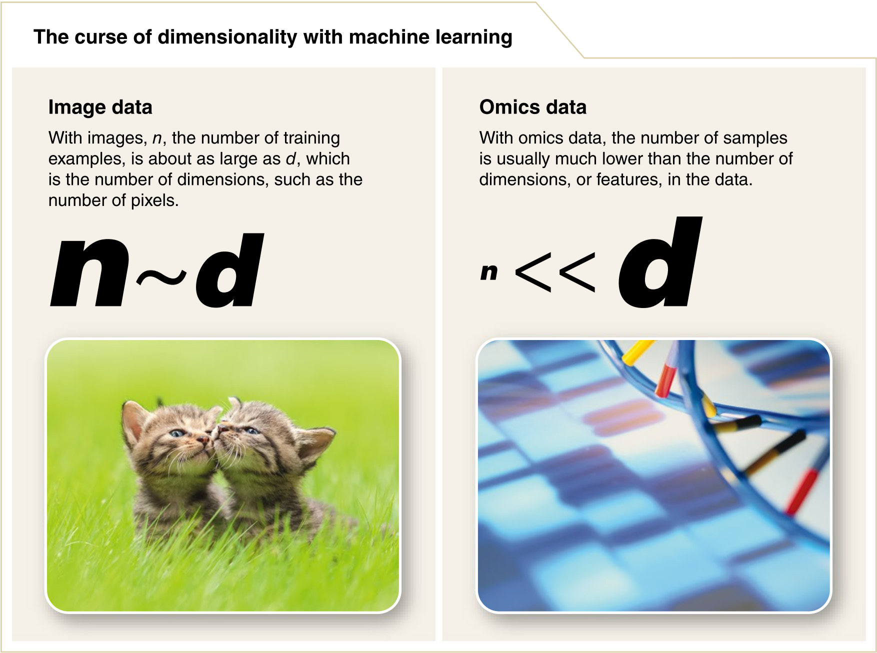
ML techniques are being put to work across the biopharmaceutical industry[8]. "AI can be incredibly useful to us, but there are certain caveats," says Saurabh Saha, who directs global strategy in translational medicine as a senior vice president at Bristol-Myers Squibb (BMS). But applying ML involves focused, well-defined questions, large labeled datasets and cross-disciplinary collaborators. "There's a lot of hope, there's a lot of hype," says Joe Szustakowski, who oversees translational bioinformatics as an executive director at BMS and was interviewed jointly with Saha. The Turing Prize is "very well deserved" and he is excited about ML prospects. But no matter how solid the algorithm, if the data input into an ML system are not high quality and structured, its output "is not going to matter," says Szustakowski. "We're not believers that you can just take a whole bunch of data, pour it into a black box and a target or a drug is going to fall out on the other side."

In recent months, the company has been applying ML to integrate digital pathology and genomics information with clinical trial data. "We're trying to extract more signal through the integration of those datasets than we get from either one of them on their own," says Szustakowski. Although many platforms measure hundreds or thousands of features, with sequencing data, there can be millions of data points per patient, but the data are from hundreds, perhaps thousands, of patients, which is not large. To build a deep learning model that predicts response or non-response to a drug, one would need many thousands, possibly millions, of patients who have responded one way or the other. "We may have millions of variables, but we don't necessarily have millions of observations," he says.

ML is helping the team extract information from pathology slides so they can, for example, train a system to detect the gene expression signatures indicative of a tumor's inflammation status. The BMS scientists are integrating these data with a range of other information about each patient, which is done in-house and with external companies such as PathAI, among others, says Szustakowski.

In cancer and immunotherapy, human pathologists do well at quantifying the expression of programmed death ligand 1 (PD-L1), which matters in immunotherapy-based cancer treatment and research, says Szustakowski. To apply deep learning for identifying tumor cells and immune cells in slides, they are manually cataloging, for example, where an immune cell is—in the tumor, at the boundary to healthy tissue or in healthy tissue, they look at the proximity of PD-L1-expressing cells. Such annotation of the tumor cells and immune cells, even when collected on a small number of patients, leads to a large library, he says, that might include a catalog of tumor cells and immune cells in melanoma that helps to train deep learning approaches for discerning cell types that will then allow the researchers to query the data in more complex ways, he says.

Drug developers are now collecting much more data about patients than previously, says Arshad Ahmed, who directs a new digital initiative at Avantor, a supplier to the pharma industry. That might be imaging data, single-cell RNA-sequencing data, tumor-microenvironment-based data or proteomics data. Companies want to use ML to learn from hundreds or even hundreds of thousands of patients. AI is one of several methods being applied that include statistical methods, "but machine learning, certainly, this is where things are going to go, to find the next level of biomarkers." He expects new, multimodal biomarkers to emerge.


The curse of dimensionality with machine learning
Image data
With images, n, the number of training examples, is about as large as d, which is the number of dimensions, such as the number of pixels.
n~d
Omics data
With omics data, the number of samples is usually much lower than the number of dimensions, or features, in the data.
n << d

Source: D. Geman, Johns Hopkins Univ. Credit: Right: PHOTODISC. Left: zhao hui/500px Prime/Getty; E. Dewalt/Springer Nature

Previously, Ahmed co-developed data-integration platforms at Novartis Oncology and at Philips. Finding patterns with “predictive power” helps to identify the patients most likely to benefit from a given treatment. “Right now the field is mostly focused on gene expression, the DNA- and RNA-level markers,” he says, but it’s moving to more protein-level and cellular data such as tumor-infiltrating lymphocytes. Adding imaging data is “still very new,” he says, and much validation needs to happen, but there’s promise.

The industry’s “data silo problem” is a hindrance for implementing ML, says Ahmed. Data batches have to be harmonized, normalized, QC’d. Companies also need algorithms to work across all data types such as genomic and proteomic data, T-cell activation data, flow cytometry data and imaging data, so they can build a “single data lake” to which ML applications can be applied.

Given the needs for large amounts of training data, many want to merge datasets, but without normalization, it’s “comparing apples and oranges,” says Fiona Nielsen, founder and CEO of Repositive of Cambridge, UK. The company helps contract research organizations and pharma companies as a data scout to find cancer models and, for ML-based companies, find datasets in order to train their model, which can be data on a specific type of metastatic cancer. Nielsen trained in computer science, and launched the UK charity DNAdigest a few years after her mother was diagnosed with cancer (she is now well). Repositive is a social enterprise spun out of DNAdigest.

As the team discovers datasets, they assess the metadata characteristics and highlight missing information that would preclude certain ML uses. “You need to make sure that the data represents those particular aspects of evidence that you need to test that particular hypothesis,” says Nielsen. Data are often captured in a standardized way, but combining data from different sources can be challenging. “Yes, they’re systematic about it, but they’re each systematic in their own way,” she says. They might use different types of annotation, for example. “Real-life data is so messy,” she says, which adds to biology’s general messiness.

To prepare ML training data, many want to merge datasets. But without normalization, it’s “comparing apples and oranges,” says Fiona Nielsen.

If an ML model looks ‘good’ with permuted data, something is wrong, says Casey Greene.

Credit: A. Greene

There are many celebrated technological successes with ML, but he would like the balance to tip toward assuring eventual clinical benefit through robust validation of these methods, says Gerrit Meijer, who heads the pathology department at the Netherlands Cancer Institute. When he reads slides, he says, “I’m using a biomarker that is based on formaldehyde, candle wax, water color stain and an instrument that was invented in the 17th century, the microscope.” Having more data and computing power helps, but what’s also needed is to annotate clinical phenotypes to the same depths as genomic data. In the end, “value comes from the integration of both,” he says.

Meijer looks forward to datasets helpful for standardized training of ML-based systems. “Wouldn’t it be great to have some kind of Spotify for research data?” If it exists for music and MP3s, “why wouldn’t we have that for cancer research?” To enable sharing of large standardized datasets, the Netherlands is building Health-RI, a nationwide research infrastructure for personalized medicine and health.

Engineers or scientists?

Applications of ML in biomedicine lead to a fundamental question: “are we engineers or scientists?” says Greene.

“If we’re scientists, I want to understand the mechanisms,” he says. “Why does a method work?” As a scientist, he wants to know why a system makes certain predictions, not why it works or doesn’t in a specific case. He wants it to be able to make predictions about living systems in ways that reach beyond the model. It might lead to, for example, the discovery of a new motif or co-regulated group of genes. “As scientists, we want that motif or co-regulated group of genes to be related to a process that matters elsewhere, too.”

“If we’re engineers, we want things to work,” says Greene, which highlights the importance of prediction accuracy. “We care a lot less about causality as long as something is predictive,” he says. This approach faces issues, especially around biased models that perpetuate unfairness, “but perhaps we don’t need to understand causality if we can build in adjustments that address them.” Imaginably, the optimal endpoint for an engineer, as one of many possible endpoints, he says, is a clinical trial that uses an ML model and shows it improves outcomes for those “treated” by the deep learning method.

Greene feels he is more of a scientist than an engineer. “I can see an argument for both, but I really worry about the hazards of an entirely engineering-based approach in this domain,” he says. He is concerned that the field is “tilting” more toward that side of late, “so perhaps my choice is also a bit of a counterbalance to that perspective.”

Vivien Marx
Technology editor for Nature Methods.
e-mail: *v.marx@us.nature.com*

Published online: 9 May 2019
https://doi.org/10.1038/s41592-019-0432-9

References

1. LeCun, Y., Bengio, Y. & Hinton, G. *Nature* **521**, 436–444 (2015).
2. Wainberg, M., Merico, D., Delong, A. & Frey, B. J. *Nat. Biotechnol.* **36**, 829–838 (2018).
3. Zitnik, M. et al. *arXiv* Preprint at https://arxiv.org/abs/1807.00123 (2018).
4. Ma, J. et al. *Nat. Methods* **15**, 290–298 (2018).
5. Eraslan, G., Avsec, Z., Gagneur, J. & Theis, F. J. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-019-0122-6 (2019).
6. Chen, K. M., Cofer, E. M., Zhou, J. & Troyanskaya, O. G. *Nat. Methods* **16**, 315–318 (2019).
7. Greene, C. et al. *GitHub* https://github.com/greenelab/deep-review (2019).
8. Vamathevan, J. et al. *Nat. Rev. Drug Discov.* https://doi.org/10.1038/s41573-019-0024-5 (2019).