HOW MACHINES LEARN IN HEALTHCARE

Machine learning is transforming every facet of healthcare, as computer systems are being taught how to use Big Data to derive insights and support decision making. In this respect, teaching a computer, no less than teaching a child, is to "shape the future." Educating a computer is a surprisingly labor-intensive process, requiring massive amounts of data, a nuanced understanding of every data element from every data source, years of trial and error, and extensive domain expertise. The key differentiator in machine learning is not the specific technology and science applied; it is in the volume and quality of the instructional material and the knowledge of the instructor.

ANDREI STOICA, PHD
Vice President
Systems Development
IQVIA
astoica@iqvia.com

Machine learning requires human, healthcare knowledge

Today, we're surrounded by computing systems that can learn from experience and handle new situations. Behind our internet searches, spam filters, online music curation, and virtual assistants, computers are studying away, becoming "smarter" with each interaction we have with them.

Machine learning is destined to accelerate the pace of healthcare transformation, as it allows us to extract meaning from otherwise insurmountable volumes of data. It is proving valuable in supporting research and development, identifying populations at risk, improving diagnostics, providing clinical decision support and optimizing sales and marketing.

A little understood fact is that a machine learns in much the same way as humans. The ingredients are a scientific model (from simple rules to complex algorithms), information, and a knowledgeable teacher (the domain expert). When these elements come together in the right way, machines are able to perform high-volume automation, recognize patterns, spot anomalies, provide linkages, offer recommendations, run simulations and make predictions about future outcomes with great reliability.

IQVIA was a Big Data company long before the term "Big Data" was coined. Thirty years ago, we had more data than we could effectively move over the internet and had to use elaborate "sneakernets" to transport data. Today, we are in a similar position with machine learning. The term has been over-hyped by vendors that have limited experience with healthcare data. This article is the first in a series examining what it takes to do machine learning in healthcare based on the knowledge of experts that have been applying these models and algorithms for decades. We define the entire process from data processing to analytics and the intrinsic interdependencies between the various stages underpinning the quality of results. This article provides additional details about data processing as the foundation for analytics.

GOOD DATA HYGIENE

Sometimes, answering healthcare business questions calls for data of great breadth. Other times, for data of great depth. But in most cases, and especially for business critical decisions, the data must be clean. That's why most data mining systems that claim they work on "dirty data" have in fact an intensive data-cleansing step prior to data processing. It's worth reviewing the three basic steps involved in data cleansing and processing: bridging, coding and linking. These steps not only prepare the data, but they are the foundation for quality machine learning in processing and analytics stages.

All healthcare records contain multiple references to entities (such as diagnoses, products, physicians, procedures, outlets and companies, etc.) And there can be thousands of attributes linked to each entity. For example, a medical encounter can have hundreds of attributes, including details on the procedures, imaging, notes, etc.

In some cases, there are standard codes by which these entities can be referenced, such as the National Drug Code (NDC), a universal product identifier for human drugs in the U.S. Where standard codes such as this exist, they must be assigned to the entity in the data record and subsequently validated. This assignment is called bridging. If the entity does not have a standard code, a unique one must be created as a reference in a process called coding.

To prepare data for bridging and coding, simple rules are first introduced to the computer. For example, one basic rule might be to remove all extra blank spaces in names.

continued on next page

Machine learning is destined to accelerate the pace of healthcare transformation, as it allows us to extract meaning from otherwise insurmountable volumes of data

Then, with greater exposure to more data and situations, complex machine-learning algorithms (such as neural networks, score engines, Ngrams, random forests, Bayesian networks, genetic algorithms and many others) begin reasoning about data attributes in every individual data stream.

At this stage, machines can differentiate, for example, between a doctor who has just changed her name after marriage vs. another doctor with the same last name that just graduated medical school and started practicing in the same city.

Experience shows that complex machine-learning inferences used to bridge and code data on pharmaceutical packaging and dosage, doctor addresses, distribution outlets and volumes, patients and medical procedures and hundreds of other attributes must be highly specific – even down to the individual data stream or data supplier. Achieving this level of specificity requires deep, detailed knowledge of the field.

Only someone with a history of working with a given supplier would know, for instance, that promotions for an individual pharmacy store are coded as one record while promotions at the chain-store level are coded as multiple transactions. This example is just one simple rule that IQVIA's massive machine-learning infrastructure has learned over time. Currently there are hundreds to, in some cases, thousands of rules for each of our 800,000 data suppliers around the world.

Once an entity in a record is properly bridged or coded, it can be linked to records in other data sets to allow for cross-referencing. Billions of healthcare records are generated and exchanged between different parties in the healthcare industry each year, and accurate bridging and coding must take place with every data exchange to maintain the quality and usefulness of the data.

"Dirty" data can have significant implications for the decisions that pharmaceutical companies make – from research and development to commercial planning.

The privacy laws of different countries add complexity to the challenge of linking records. Where laws allow de-identified data to be collected, the de-identification algorithms have to be both secure (irreversible) as well as versatile to allow linkage for longitudinal patient analytics for example.

For de-identification, homogenous systems are key to data security and linkage. It is less likely that data de-identified with different standards can be linked, or if it is, that it will not affect the security of the data.

A CASE IN POINT

Machine learning is changing healthcare in real time. IQVIA built a decision-support system using machine learning to help sponsors manage physician selection in clinical trials – a task fundamental to trial success.

Experts in the therapeutic area defined multi-dimensional models to express all of the study protocol details. Data scientists with complementary expertise worked as part of the same team to define matching multi-dimensional models to express all physician prescribing/treatment patterns and history. Through deep learning, the system was trained on eight petabytes of clean (bridged, coded and linked) claims and electronic medical records (EMR) data. The result was a prioritized list of investigators with the highest probability of success. This resulted in a double-digit percentage decrease in non-enrolling investigators and, at the same time, a double-digit percentage increase in patient enrollment for rheumatoid arthritis studies.

Since both physician behavior and study protocols can be very complex, it is incumbent upon domain experts to understand how to model this complexity. The accuracy of the results depends on training, and that is directly influenced by the quality of the data. Machine learning depends on the availability of clean data for training and constant supervision from domain expert operators who tune the results.

For this reason, IQVIA has global standards for de-identification that are constantly reviewed by security experts and designed to support linkage where permitted by law. Machine learning plays a crucial role in identifying patterns that weaken the security and in cleaning the data prior to de-identification.

THE ONGOING HUMAN ROLE

Even after a machine-learning system is mature and has been successfully processing billions or trillions of transactions every year – as our systems have been doing for the past 60 years – day-to-day human tasks are still critical to data operations and data mining. Domain experts must work alongside the machine, overseeing, and sometimes correcting, its work.

Healthcare is continuously evolving, and data structures are volatile; every day there are variances with unaccounted scenarios that the machine has not yet learned. Imagine that the machine learning does 99 percent of the work but misses identifying one new pharmaceutical product in the market. In that case, domain experts must teach the machine about that product so that it now can be automatically processed. In the absence of domain experts who sample data and direct the machine when confidence levels are low, we find that machine learning quality degrades anywhere from 0.1 to 10 percent every month, depending on the data source. Even at 0.1 percent level, a machine operating for a year without a domain expert will eventually produce outputs at unacceptable quality levels.

For instance, IQVIA operates a machine-learning algorithm in one set of product data that processes seven million records and learns 300,000 new product reference keys each month. About half of the new records undergo some human-assisted quality control operations.

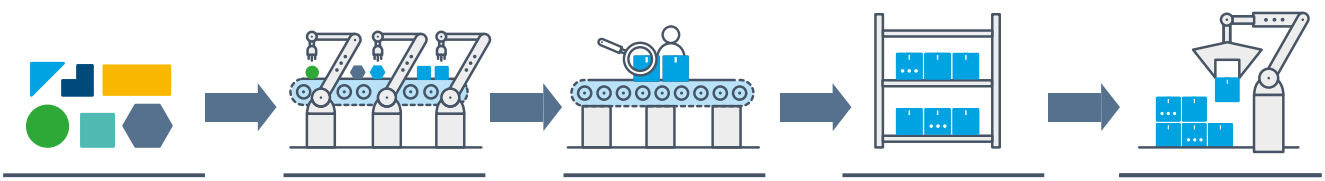
MACHINE LEARNING + DATA + DOMAIN EXPERTISE = IMPACTFUL ANALYTICS

Clean, linked and attributes-rich Big Data represent the foundation for quality and impactful analytics. Every stakeholder within healthcare (pharma, payers and providers, etc.) along every dimension (countries, languages, suppliers, data type), and for every specific use case (therapy, clinical, commercial, R&D, etc.) requires its own machine-learning algorithms and specific configurations. These can only be developed over time by constantly building upon an ever-growing knowledge base.

Analytics are divided into three categories descriptive, predictive and prescriptive. Descriptive analytics summarize, describe, quantify and analyze the past trends of an activity. Predictive analytics use past behavior to train machine-learning algorithms to predict possible future trends. Prescriptive analytics look at different possible outcomes in the future, why they can happen, how to navigate them and the impact of possible decisions. Prescriptive analytics are used in intelligent and decision-support systems.

continued on next page

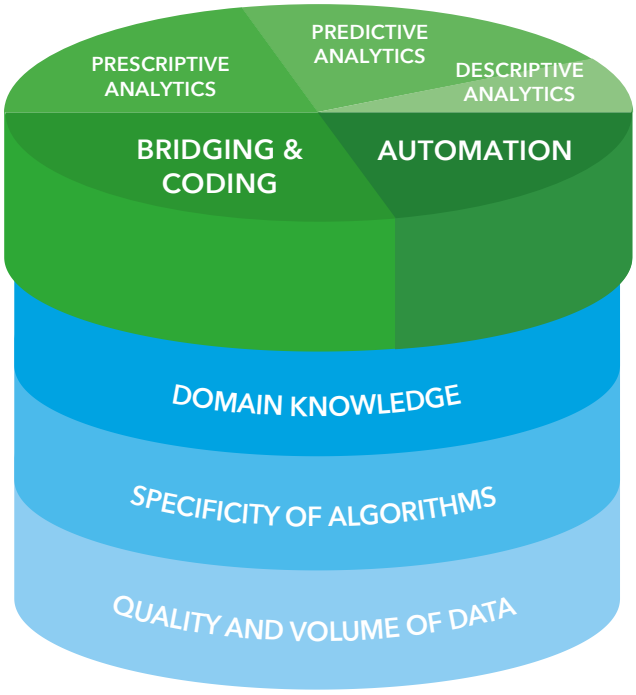
Figure 1: Humans and Machines Work in Tandem from Data Processing to Advanced Analytics



In both predictive and prescriptive analytics, the first phase is modeling. A team of domain experts including clinicians and data scientists analyze the problem and the available data and select the machine-learning algorithm that will have the highest success rate. This is a "human only" intensive process that requires great expertise in clinical practices, data, computer science and machine learning algorithms, and region-specific healthcare delivery and application processes.

At this point we have a capable computer system (a machine) with the right "instructions" to learn. If we go back to our initial analogy, we have a capable "genius baby" with no knowledge – just capacity to learn.

Figure 2: Building Blocks for Machine Learning in Healthcare



There are supervised or unsupervised learning methods, but they all depend on large amounts of data already categorized by human operators; "teachers," or domain experts, that must tell a computer when it is right or wrong. With every choice that the human makes, the computer "learns" and will be able to apply that decision process in the next similar instance. This iterative process can be set so that experts review only the machine's choices that fall outside of certain confidence limits. Or, it can be set so that the machine recommends a choice for the expert's approval.

There is a basic data volume threshold that must be met for the process to work at all. If too little data are used, the system will simply not have enough information to train the model. After that threshold is met, the more data available to represent every situation with enough examples, the better the machine can be taught. Just how much data are required depends on the situation.

When patterns repeat infrequently, it may take several years' worth of data to train a computer. When data are clean and bear standard reference codes, smaller amounts are needed than when the data are unstructured.

Because data and the practice of healthcare are constantly changing, machine learning demands constant review of core algorithms and settings. Only domain expert teams with detailed knowledge of the data, science and healthcare can determine when and how to refine the model. Such permanent day-to-day interaction with the machine and regular supervision from domain experts is resource intensive, but absolutely essential for the performance of the computer systems.

Large scale data processing using machine learning is the strong foundation required to build a meaningful data repository, train machines, and constantly refine domain knowledge expertise.

Only domain expert teams with detailed knowledge of the data, science and healthcare can determine when and how to refine the model

The various steps of teaching a computer system demands extensive knowledge of the subject area under study

From this perspective, machine learning is a very human matter

With this foundation, now it's possible to perform the next generation of clinical decision support analytics including use cases such as non-adherence, disease progression and care management, safety signal detection and evaluation, therapy dosing and response, or using the digital footprint of diagnosed patients to find new, undiagnosed patients. Machine learning models such as vectorization, natural language processing term extraction, ensemble suites of algorithms using deep-learning and many others can return high-quality results for these and many other use cases in real world evidence.

CONCLUSION

Every year, the healthcare industry generates billions of records. The more data we have access to, the more accurate healthcare becomes. There are thousands of analytics problems and machine learning can have a major impact providing all foundational building blocks are in place: high-quality big data, advance science, and the "secret sauce" - deep domain knowledge.

The science involved in machine learning has been around for several decades, and software developers all draw upon the same statistical, mathematical or computer science algorithms and techniques. But, the difference is in the knowledge base that determines how those techniques are applied. The various steps of teaching a computer system demands extensive knowledge of the subject area under study. From this perspective, machine learning is a very human matter.

HARD LESSONS OF MACHINE LEARNING

As machine learning has moved from theory to applications, the space has been filled with unverified claims and marketing hype. Here's what we know to be true

1. **The more data, the better.** Having a large, pre-existing repository of data governed by well-established rules is an absolute pre-requisite for teaching a computer. A minimum volume of data are needed for learning to occur, and after that, the more data, the better.
2. **Garbage in, garbage out.** It's trite, but true. Data must be clean and linked to be useful.
3. **Machines can't do it alone.** The only way for machines to learn is for a human teacher to review their work and to educate them on "right" from "wrong." This iterative process improves the computer's output. And the collaboration between man and machine must be sustained. Experts must be on hand at all times to validate what machines do, especially as data and data models are constantly changing.
4. **Domain knowledge rules.** It is impossible to select the scientific algorithms, configure the systems, and evaluate the computer's output without being intimately familiar with the subject matter. This requires years of experience in the specific domain. There is no crash course on this.
5. **There are no shortcuts.** Training computers takes time and trial and error. Data investigations of possible errors on the machine's part translate, over time, into cleaner data and better rules and algorithms. But the demands keep escalating over time. The level of training that was good enough today will be insufficient tomorrow.
6. **The only good algorithm is a specific algorithm.** A machine trained on financial systems will produce poor results in healthcare. Even in healthcare, a machine trained on a specific data set will not perform on another data set. Algorithms must be specific to the use case, country, language, data type, and even regional healthcare delivery procedures.

485 Lexington Ave, 26th Floor
New York, NY 10017, USA
Tel: +1 646 596 6053

CONTACT US

accesspoint@iqvia.com
www.iqvia.com
www.linkedin.com/iqvia/real-world-insights
www.twitter.com/iqviaRWI

IMS Health & Quintiles are now IQVIA™