Embedded Analyst: AI Without Borders

How one healthcare system is breaking through barriers that shackle AI's potential in medicine.
And in doing so, it may not only be saving patients.
It may even be saving itself.

June 2018

Analyst: Mike Feibus

in association with
Intel Corp.

Artificial intelligence has the potential to make huge improvements in just about every aspect of healthcare, from helping doctors make people healthier to improving operations. But too often, the loftiest of ambitions get tripped up out of the gate by some decidedly low-level hurdles: incompatibilities that prevent researchers from combining different sets of records to build their models.

One of those ground-level obstacles is nomenclature. Different systems often label things differently. In fact, there are at least 47 different ways a single substance – potassium – is identified in electronic health records. Sometimes, even different versions of the same system use different terms. To compound matters, metadata – companion information vital to building accurate models, like which analyzer at which lab tested the sample – may be lacking or absent.

This devil-in-the-details conundrum is the prime reason why so many data silos still exist anachronistically in the connected age – now isolated not by communications capability, but rather by structural incompatibility. Researchers who do manage to stitch divergent systems together for their models commonly fix the inconsistencies by hand, which results in intelligent systems that are costlier, take longer to build and are narrower in scope than would otherwise be possible.

One healthcare system has been pioneering efforts to free researchers from these constraints. With origins in post-9/11 municipal emergency projects, Montefiore Health System's platform – called PALM, short for Patient-centered Analytical Learning Machine - is beginning to prove itself out in the Intensive Care Unit, helping doctors save lives by flagging patients headed toward respiratory failure.

And that's just the first step. PALM is enabling Montefiore to develop a wide variety of models to do everything from predicting and preventing costly, life-threatening conditions like sepsis and heart failure to improving operations by predicting appointment no-shows and guiding walk-in traffic to urgent care clinics and emergency rooms with better availability.

As with the lion's share of AI projects in play today, PALM is powered by Intel Xeon processor-based systems, including Intel's newest Xeon Scalable processors. The underpinnings of the PALM platform demand processors that can address very large amounts of system memory, which Xeon is far better suited to do than GPUs. And because of their flexibility, Xeon's can actually be used for inferencing once the heavy lifting of AI training is complete. More than inferencing, in fact, the Xeon AI systems also make great servers for email, security, file sharing and other traditional enterprise activities. GPUs flat-out aren't built for that.

With cooperation from Montefiore and Intel, FeibusTech Principal Analyst Mike Feibus
was embedded in the Bronx-based healthcare system in the spring of 2018
to produce this in-depth, first-hand look at PALM, and its impact on care
in one of the nation's most challenging environments.
His account follows.

Parsa Mirhaji was wrong. And that pleased him.

Mirhaji is Director at Montefiore and Albert Einstein College of Medicine's Center for Health Data Innovations, and PALM is his brainchild. When we'd met earlier that week in the Bronx, he told me he'd been preoccupied with how to expand PALM's capabilities. Powerful though it is, PALM's very untraditional approach had been designed to tame very traditional datasets: electronic medical records, insurance billing codes, drug databases and clinical-trial results – even online catalogues of genes and genetic disorders.

But Mirhaji didn't think PALM was prepared to perform its magic on different types of information, like voice, images and sensor inputs from IoT devices. He was worried it would need to be rearchitected. Or worse – that he and his team would need to build a second, rich-data version of PALM.

He needn't have been concerned, as it turned out. Stuffed into a conference room with most of the team at their offices in Yonkers, what we heard was music to Mirhaji's ears. PALM didn't need to be rearchitected. They just needed to inject some new metadata to give the platform the ability to cope with the new data types.

"It's really just a matter of adding modular pieces to the existing data structure to deal with the new data types," Boudewijn Aasman, who manages the PALM framework, told us.

"That actually makes it much simpler than I was thinking," Mirhaji smiled. "This is really too good to be true."

In theory, yes. But as with all things PALM, of course, the devil is in the details.

Nuts and Bolts

Although PALM is a foundation for building machine learning algorithms, what sets it apart is its ability to enable people who write those algorithms to tap myriad data stores, regardless of where the information is physically located, or how it is structured. The potential impact of that on artificial intelligence cannot be overstated.

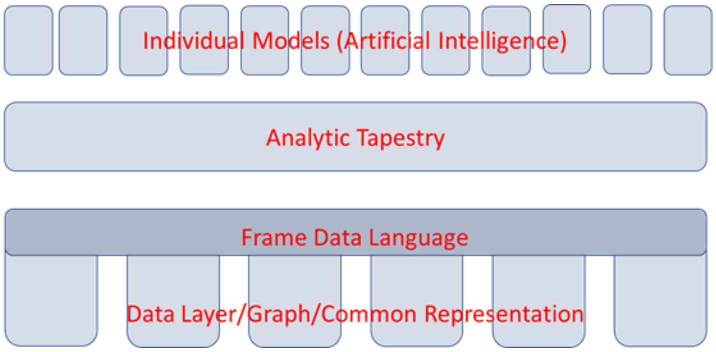
> *AI needs large numbers of cases to reliably draw conclusions. Which means if you work at a small healthcare system – or if you're part of a large system and want to build algorithms focused on rare conditions – then you need to combine your health records with others if you're going to get anywhere.*

PALM is what's come to be known as a *semantic data lake* architecture. In addition to data stores, PALM also taps into *ontological* databases, which together define more than 2.5 million terms and their relationships to one another. There are ontological databases of ailments, medications and genetic disorders. An important part of PALM's magic is its ability to quickly bolt on those outside definitions and integrate them into the platform.

Just atop all the databases is what's known as a *triplestore*, or *triple*, construct. That's a key piece of any semantic data architecture. A triple is a three-part data series with a common grammar structure: that is, subject-predicate-object. Like, for example, John Smith has hives. Or Jill Martin takes ibuprofen.

Triples are the heart and soul of *graph databases*, or graphs, a powerful, labor-saving approach that associates John and Jill to records of humans, hives to definitions of maladies and Ibuprofen to catalogues of drugs. And then it builds databases on the fly for the task at hand based on those associations.

But what if one health system refers to hives as rashes, and logs ibuprofen as Advil or Motrin, two brand names for the drug. That's what PALM's *Frame Data Language*, or FDL, is designed to tackle.


Individual Models (Artificial Intelligence)
Analytic Tapestry
Frame Data Language
Data Layer/Graph/Common Representation

That's only one piece of FDL's power. It can also leverage broader relationships, which gives researchers tons of latitude to target precisely what they're investigating without concern for what's actually recorded. With FDL's help, for example, John Smith will show up in a search for patients with allergic reactions. And Jill Martin will appear in a query for patients taking NSAIDS.

A word about John and Jill. PALM deidentifies personal data, so it's not possible to drill down and uncover their identities. The way PALM taps data sources, in fact, researchers at Montefiore can't get their hands on any raw data from, say, New York Presbyterian or Northwell Health, even though the data potentially can be used to train their AI models.

PALM's final layer is called the Analytic Tapestry, which is like the control center between the models and the data. Mirhaji likes to call the tapestry a "social network of algorithms," because it keeps an eye on what each model is doing, and networks algorithms that are doing similar work. Which means that the tapestry will see your new model to improve hospital readmissions for patients with chronic conditions, for example, and link you with an existing model that's lowering readmissions for patients with diabetes.

PALM is built on AllegroGraph, a semantic data lake platform from Oakland-based Franz Inc., and run locally on Xeon servers in Montefiore's datacenter in Yonkers. Semantic data lake architectures require lots and lots of system memory to address the tailor-made databases they build on command out of other, more structured databases. And Xeon processors are better able to address large banks of system memory than graphics processors, or GPUs. In addition, semantic data lakes also demand a stable of processor cores and super-fast storage.

The team has three very different server configurations, a sign that Montefiore and Intel are still zeroing in on the best combination. The oldest server in the rack has dual octa-core Xeon processors and 768 GB of system memory. It has 13TB of fast solid-state drives, or SSDs, as well as 22TB of traditional hard-disk storage.

There's also a cluster of eight servers bound together by a super-fast optical network. Each system is powered by has dual 8-core Xeon E5 V4s, with 256GB of memory and 12TB of traditional disk drives. The team has been experimenting with how many shards, or database partitions, to spread across the eight systems, Manuel Wahle told me. Wahle is the team's technology lead and a former student of Mirhaji's at the University of Texas. They've found that if each system is responsible for four shards – for a total of 32 –there's a sizeable jump in performance, even though they're configured with traditional disks, he said.

At the other end of the spectrum is the team's third system, a developmental platform from Intel configured with Optane drives. Built on Intel's groundbreaking 3D XPoint memory, Optane drives are the fastest SSDs on the planet, by far. With two 20-core Xeon Gold processors, an 800GB SSD and two 375GB Intel Optane SSDs, the server has the makings of a semantic data lake powerhouse. Except that it only has 32GB of memory, not nearly enough for these workloads.

Fortunately, Optane SSDs can be configured as either system memory or storage. So the team was able to get the system chugging on PALM workloads, and give them a glimpse of what Optane can do for performance, Wahle said. Likewise, Jans Aasman told me he's really excited about Optane and how much it can augment Xeon system capability. Aasman should know. He literally has PALM in his DNA – and not just because he's the CEO of AllegroGraph maker Franz. His son Boudewijn manages the framework for Mirhaji.

The Inaugural Model: Predicting Respiratory Failure

The first PALM graduate to make it onto the hospital floor is a model that predicts respiratory failure, which the team defines by the need to put the patient on a ventilator. Artificial respiration is a pretty grave step, and one that many patients don't recover from. Indeed, their risk of dying within the next six months is nearly 50 percent. As many as a third don't even leave the hospital alive. And only a third function as well as they did before they were admitted to the hospital.

So there's clearly a strong incentive to keep as many patients off the ventilator as possible. If Montefiore had an alert that could give doctors and other clinicians enough of a warning, they could save lives and dollars by keeping more patients off ventilation.

That's exactly what Dr. Michelle Gong, Director of Critical Care Research, Division of Critical Care Medicine at Montefiore and Einstein, set out to achieve when she teamed up with Mirhaji and the PALM team to build a model to predict – and, hopefully, prevent – respiratory failure.

As a critical care physician, Dr. Gong takes care of the sickest patients in the hospital who are experiencing life-threatening organ failure such as respiratory failure. And as a clinical researcher in the area of respiratory failure, she has spent most of her research career trying to predict and prevent severe respiratory failure and the severe consequences in patients in the hospital.

Respiratory failure occurs in only a small percentage of patients in the hospital but when it does occur, it can be devastating to the patient's life, function and well-being. Finding these patients early may give clinicians the opportunity to prevent respiratory failure or death resulting from respiratory failure.

“The longer you’re on a vent, the worse it is,” Dr. Michelle Gong told me. “We do it to keep patients alive. But with it comes consequences.”

With funding from the NIH, Dr. Gong worked with Dr. Mirhaji to develop and validate a machine learning model to identify patients at high risk for respiratory failure or death in the hospital. It takes about 18 months from the time PALM researchers begin training a model until the resulting algorithm can move into the clinical workflow. Most of that time is taken up not by building and training the model, but by validating its clinical reliability and safety.

The respiratory failure model was no exception. The team started by randomly dividing hospital records into two parts. They used one part to train the model. And then, they tested the model with the remaining data. For good measure, they let the model practice for months by predicting respiratory failure on live hospital data flowing in real-time.

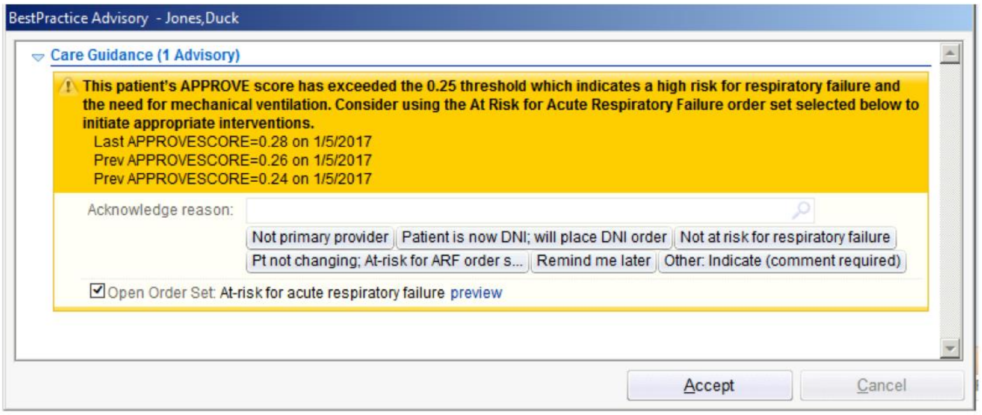
That’s a critical piece. And there are no shortcuts there.

“You have to get the algorithms right in order to get the predictive analytics correct. And that takes some validation, it takes a lot of work behind the scenes,” said Dr. Andrew Racine, System Senior Vice President and Chief Medical Officer at Montefiore, and Professor of Clinical Pediatrics at Einstein. “Before you put it in front of clinicians you have to have confidence that it’s going to do what you say it’s going to do, and it won’t surprise you with unintended consequences.”

Finally, once the model had proved itself, it was time to roll it out.

That happened in January 2017. The algorithm was added to the electronic medical record, or EMR, in the form of an alert at the smallest and – comparably speaking – quietest of Montefiore’s three Bronx hospitals, the 369-bed Wakefield Hospital in the north Bronx. The following year, they expanded the program to the other two hospitals, the 396-bed Jack D. Weiler Hospital and 726-bed Montefiore Hospital at the Moses Campus.

The alert comes in the form of a “Best-Practice Advisory,” or BPA, in the Epic EMR system deployed by Montefiore. The BPA, dubbed APPROVE, or Accurate Prediction of Prolonged VEntilation, appears on clinicians’ screens like the example below:


BestPractice Advisory - Jones,Duck

Care Guidance (1 Advisory)

This patient’s APPROVE score has exceeded the 0.25 threshold which indicates a high risk for respiratory failure and the need for mechanical ventilation. Consider using the At Risk for Acute Respiratory Failure order set selected below to initiate appropriate interventions.

Last APPROVESCORE=0.28 on 1/5/2017
Prev APPROVESCORE=0.26 on 1/5/2017
Prev APPROVESCORE=0.24 on 1/5/2017

Acknowledge reason:

☑ Open Order Set: At-risk for acute respiratory failure preview

Not primary provider
Patient is now DNI; will place DNI order
Not at risk for respiratory failure
Pt not changing; At-risk for ARF order s...
Remind me later
Other: Indicate (comment required)
Accept
Cancel

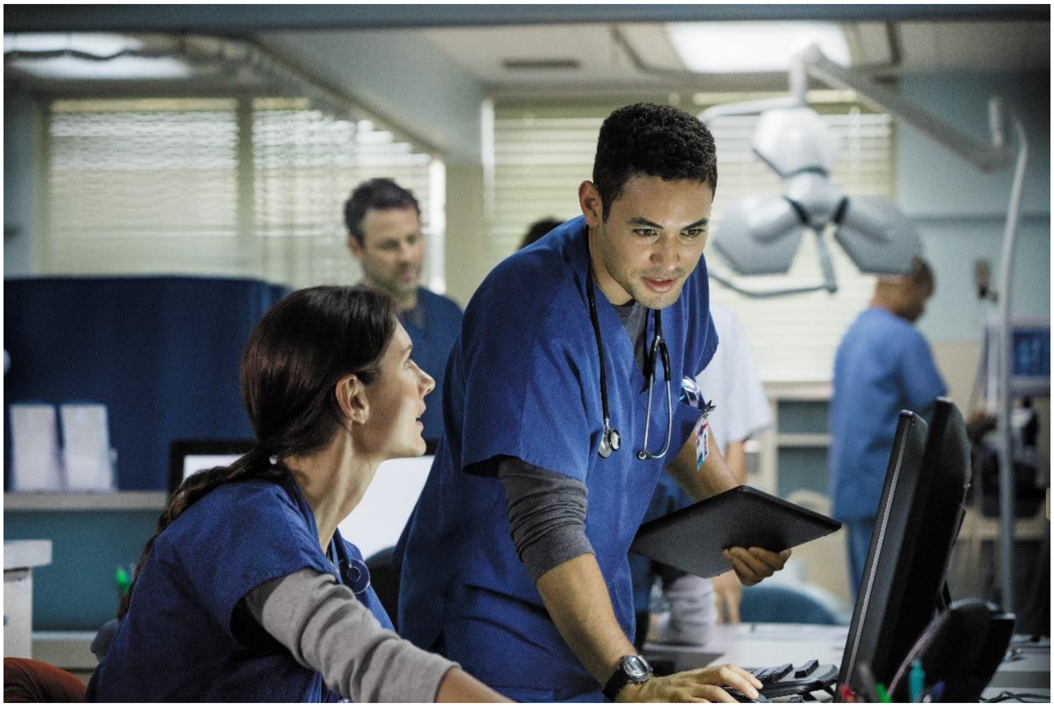
Once providers accept the advisory, a to-do list pops up, called PROOFCheck, or PRevention Of Organ Failure Checklist. Chief among the actions, or orders, on the checklist is a consult with critical care physicians like Dr. Gong.

The AI delivers a score that's an assessment of risk on a scale of zero to one. The BPA gets tripped when a patient's risk score crosses the 0.25 threshold. It's important to note that while the model generates risk scores, the AI did not draw the line in the sand for generating a BPA. Team members took that decision upon themselves.

Dr. Gong acknowledged that the threshold of 0.25 is on the "high side." ~~W~~ith any prediction rule, there has to be a balance between setting a threshold that is low enough to capture as many patients as possible vs a threshold that is high enough to minimize any false positives which can contribute to alert fatigue and lack of response from clinicians.

Certainly, there's no disputing Dr. Gong's high-side assessment. The average score of patients in the hospital is 0.06. Less than three percent of scores top 0.25. The percent of patients in the range is even lower, because most high-risk patients log multiple scores greater than 0.25.

But the best validation of Dr. Gong's high-side assessment may be this: a score of 0.25 captures about two thirds of the patients who end up needing to be put on a ventilator or die in the hospital. In addition, one out of every 4 or 5 patients with the high score will end up with the event.



Dr. Gong and I discussed the tradeoffs of choosing the right threshold as we toured the Montefiore Hospital at the Moses Campus, the health system's largest. As we walked through the emergency room, one of the country's busiest, on what she observed was a "slow day," I was struck by the incessant cacophony of buzzes and beeps from the collection of monitoring devices. The systems all seemed to clamor for attention. But the clinicians seemed oblivious.

At that moment, I could see the dilemma the PALM team faced: how do you balance the algorithm's ability to predict respiratory failure while still command the clinicians' attention? They didn't have the luxury of training the model to choose the right threshold. Because in doing so they risked simultaneously training doctors and nurses to turn a deaf ear.

I heard that message. Loud and clear.

Not Another Alert!

The PALM team knew that the last thing clinicians wanted to see was another BPA. In April 2017, just before the team introduced APPROVE, the number of unique BPAs in use at Montefiore stood at 289, up 9.5 percent from 264 in January.

From January through August last year, Epic issued 247,326 unique BPAs a month – each one delivered to an average of more than eight Montefiore clinicians. It'd be one thing if most of those more than 2 million interruptions a month carried information clinicians urgently needed to better care for patients. Suffice it to say, that's not the case. Many have to do with procedural protocols. Or drug interaction advisories issued regardless of whether the patient is prescribed any of the troublesome combinations.

If you thought those statistics were mind-blowing, then try this one on for size: in total, clinicians take no action on – that's health record-ese for flat-out ignore – more than three-fourths of all BPAs, according to data gathered by the PALM team.

"You cry wolf so many times that when you cry fox, they still think you're crying wolf," Dr. Gong told me.

Too true. You should hope that "Base Epilepsy and Seizure" never applies to your case. In their research, the team found that this particular BPA was not acted upon by clinicians even once in the 12,037 times it was issued during the first eight months of 2017.

By now, it should be clear that earning clinicians' trust with a new BPA is no small feat – and how important it is that any prediction tool balance the sensitivity of the tool to detect the event with the alert burden on the healthcare providers and their ability to respond to the alert.

Approving of APPROVE

With that as a backdrop, we're better equipped to assess APPROVE's adoption. And it's doing pretty well. Relatively speaking.

On average, at least one clinician interacts with the APPROVE alert about 93 percent of the time it is issued.

It's not the best-performing BPA. There is a smattering of BPAs that engenders a positive response from at least one clinician each and every time they're issued. But that number pales in comparison to the number of BPAs that, like "Base Epilepsy and Seizure," did not get a single positive response during the eight-month study period.

I found that oddly comforting, because it suggests that, as with the beeps and buzzes of the ER, clinicians do find a way to hone in on truly urgent alerts while tuning out the others. If you're a clinician, of course, that's no consolation.

"If you walk on the floor, there's tons of sound – and no one seems to care," Dr. Michoel Snow, a PALM team member who's not just a data scientist. He's also a physician. "You tune it out within a day or two. You learn which alarms matter, and which ones don't. Which BPAs matter, and which ones don't. And they shouldn't have to do that."

The team found that residents are the most likely clinicians to respond to the APPROVE BPA. Dr. Gong said that made sense, as residents typically log into the EMR more often.

In the meantime, that may help explain why I got an unequivocal two thumbs up for APPROVE from the residents I spoke with. They told me that when they're evaluating whether to involve the critical care department, the alert can give them a little boost of confidence to act.

"Sometimes there are small things you can overlook, and the BPA helps you pick up on those things," Dr. Sahil Virdi, a second-year resident at the Montefiore's Wakefield Hospital, told me. "And when someone is critically ill and you are doing your best, and still the patient is not moving in the right direction, these BPAs definitely help."

FT

What's Next for PALM: the Sky's the Limit

With BPAs as the only sure way to pass in front of clinicians eyes' through Epic, it stands to reason that finding a way to reduce or eliminate irrelevant alerts would go a long way toward ensuring that APPROVE – and future AI-generated algorithms – get the attention they need. Not surprisingly, that's one of the upcoming projects in the PALM hopper.

There are lots of models underway, in fact. Which may be the best validation yet of PALM's architecture.

One new model, for example, is being developed in concert with Dr. Deb White, Director of Emergency Medicine at Montefiore's Jack D. Weiler Hospital, the network's busiest and most overstressed ER. The PALM team and Dr. White are collaborating on a model and a companion app to help triage and route people to the most appropriate care. For some, that might mean pointing them to a doctor's office or urgent care clinic, avoiding the ER altogether.

The team also is developing a model to predict appointment no-shows, a problem that needlessly squanders precious care resources. Another model in the works will help manage patient logistics as well as forecast and allocate beds for those admitted to the hospital. Among other things, the team hopes to help the hospitals keep patients only as long as they need to be there. And in doing so, help ensure that beds are available that require hospitalization, as opposed to those who can be cared for in an outpatient setting.

Another model in the works to detect the first signs of sepsis began as a compliance reporting project. The New York State Department of Health changed reporting requirements for sepsis last year. And a group worked to rearchitect the report for the new data. After nine months, it became clear they weren't going to be able to produce the new report in time. So they called on the PALM team, which tapped their platform to produce the new report in just six weeks.

"With the sepsis study, we were able to cut down the amount of manual research from a couple of years into a couple of weeks," said Jack Wolf, Montefiore's Chief Information Officer. "It's phenomenal."

The team is now turning the reporting into real-time. From there, researchers will build a model to alert clinicians about patients who are starting to show signs of sepsis. That's critical, because the longer it takes to spot and treat sepsis, the deadlier it gets.

There are other models in development to enhance detection and care of specific conditions, following in the footsteps of the APPROVE model. There's one being built to help physicians spot heart failure earlier. There’s another one to predict spinal cord compression, an effectively irreversible paralysis that can hit cancer patients. Spinal cord compression can be prevented if it’s spotted early enough. But that’s tough to do, because the symptoms sound an awful lot like common complaints from cancer patients.

The spinal cord compression model will be the first to leverage images and other rich-data types. Assuming they are as straightforward to integrate into PALM as the team believes. Remember, the devil is in the details.

And for the future, the sky’s the limit. For example, the team is looking ahead to tying in data from the New York Genome Center, of which Montefiore’s Albert Einstein College of Medicine is a founding member.

CIO Wolf is certain that the future is in semantic data lake and platforms like PALM. But with so many demands for here-and-now resources, how that transition happens keeps him awake at night.

“Where do you put your dollars? That is one of the things that weighs heavily on my mind,” Wolf said. “It’s a difficult to balance that. Today, I have a lot more people on the day-to-day than I do on the semantic data lake. But I see the semantic data lake slowly overtaking that. When can I totally rely on (PALM) and the AI that’s associated with that? When does AI start to play a role for patient-driven healthcare?”

How close are we to that?” he said. “That’s exciting stuff. I hope I’m still working when we get there.”

FT

`

Certainly, no one will dispute that big data is an enabler of artificial intelligence. Indeed, the bigger the data, the easier it is to draw conclusions – and the better they are.

That’s been the stumbling block for so many AI projects. And it’s made others far costlier and more limited in scope than architects of these projects had envisioned. Indeed, for them the AI is the easy part. The hard part is putting together the database to fuel their models.

That’s what makes PALM so special. The Montefiore Health System’s AI platform possesses a unique ability to quickly and easily concoct purpose-built big-data stores from smaller, seemingly incompatible and unconnected datasets.

And now that they’ve had early success with APPROVE, its first algorithm to be released into the wild, the PALM team is leveraging the platform to attack AI with a swiftness and simplicity that few people would have believed was possible even a couple of years ago.

In the meantime, the team is working with Intel to identify the best combination of Xeon processor cores, system memory and storage for its semantic data lake architecture. Given that graphs require large amounts of system memory and ultra-fast storage as well as an ample supply of compute cores, it seems like a foregone conclusion that Optane SSDs will emerge with a role that’s as central to the PALM platform as even the Xeon processors.

FEIBUSTECH

clear ▪ critical ▪ independent

FeibusTech
P.O. Box 25685
Scottsdale, AZ 85255
www.feibustech.com
+1-480-922-3244