data iku

ADVANCED ANALYTICS IN HEALTHCARE

TOWARD EFFICIENT, PRECISE, AND PERSONAL MEDICINE

A WHITE PAPER BY DATAIKU

www.dataiku.com

INTRODUCTION

No domain better demonstrates the potential for data science to improve the world than healthcare. The past century has seen major advances in medicine and health, leading to dramatic increases in life expectancy. But data science applications may be able to help healthcare professionals combat some of the world's most devastating diseases in addition to helping cope with logistical and economic issues that influence patient care and the emotional effects of an aging Baby Boomer population.

Deloitte highlighted in their 2019 Healthcare Report that the top industry trends in America include ways technology can put patients first and increased adoption of virtual care options. In their Health Systems report from 2018, the EU Commission stated that in addition to patient-centric care becoming a focus, the access to and adaptability of health systems is a top priority.

The Organization for Economic Co-operation and Development (OECD) analyzed health spending across member countries and found that the average spend per capita was $3,453, with the United States spending more than double that, coming out to 16.4% of the American GDP; that's over $1 trillion of public funds as of 2015.

Even if tax dollars are being spent on medical inefficiencies, the true cost of the lack of healthcare access and inefficiencies in the current medical system are fundamental issues. After heart disease and cancer, medical errors are the third leading cause of death in America. While some of these deaths are the result of human errors, many are the result of systematic issues that AI could help address: system mistakes, organizational inefficiencies, or overworked healthcare workers on long shifts. And even in cases of human error, an AI-enabled assistant could "sanity check" processes to prevent catastrophes.

Data science offers the healthcare sector the opportunity to decrease costs and improve patient outcomes across all work streams by:

- Reducing errors and inefficiencies
- Improving administration and logistics of health systems
- Reducing fraud
- Alleviating over-burdened specialists and imaging technicians
- Helping diagnose disease

Data science offers the healthcare sector the opportunity to decrease costs and improve patient outcome

And if that's not enough, data scientists are leveraging their skills to explore wider population health in order to help target spending to improve the health of entire communities.

We interviewed several expert doctors and researchers to find out about the cutting-edge work being done regarding data science, ML, and AI in the health sector. A lot of the problems they are facing are no different from other industries—data access and wrangling, industry and organizational buy-in, productionalization (or operationalization)—however, there are challenges that make the healthcare industry unique. Not the least of which is that lives are at stake.

This white paper will delve into three main topics, with insights from these expert interviews dispersed throughout each:

I. Cutting edge techniques and endpoints being explored in the industry.

II. Challenges to applying data science, ML, and AI in healthcare.

III. Top use cases (with an in-depth look into each use case).

I. DATA SCIENCE TECHNIQUES AND ENDPOINTS IN HEALTHCARE

NATURAL LANGUAGE PROCESSING (NLP)

NLP is the process through which machine learning (ML) models extract information from unstructured human speech or written language. It has nearly endless applications, especially in the world of health, but the lowest-hanging fruit includes:

- Analyzing forms and social media to help hospitals and clinicians better understand their performance.
- Extracting useful information from doctors' notes and from patients' prior histories (including applications to improve the clinical trial process by streamlining patient recruitment).
- Helping doctors parse research to diagnose disease.
- Aiding patients in exploring doctor reviews to find the best clinician for their needs.
- Allowing providers to more quickly identify patients who should be screened for certain conditions for which they are at risk, enabling early detection that is easier and cheaper to treat.

But how does NLP work in practice? Taking the example of clinical trials, NLP in combination with deep learning models can aggregate massive amounts of data - including unstructured sources like patients' medical histories - to create a comprehensive clinical profile for recruitment to trials. From there, ML can be useful in pinpointing as diverse a patient pool as possible (or patients with a desired biometric profile for more personalized medicines) and predict any existing drug interactions that may compromise the trial.

It's easy to see why NLP has so much potential in healthcare - currently, extracting and interpreting information from the myriad documents that make up a patient's record involves an immense time and technical spend. Yet this type of analysis often yields insights into a patient's medical situation that might have eluded a medical professional.

This extraction technology will also make it easier for patients to provide all of the information about their symptoms in writing, or even as an audio recording. This way the patient or medical professional doesn't need to attempt to simplify; instead, the NLP model can analyze all of the information and factor all relevant parts into a diagnosis model. This is especially useful for accurate analysis in the event of language barriers or slang that the doctor may not understand.

As an example of NLP already springing into action in the healthcare space, in November 2018, Amazon launched Amazon Comprehend Medical, a healthcare-focused version of its natural language processing service. While electronic health records made the storage of sharing of records more convenient, Amazon Comprehend can extract and organize pieces of unstructured medical information that is found in doctors' notes, discharge summaries, case notes, lab results, and other important medical documents.

BETTER NLP FOR HOSPITALS

While NLP offers huge opportunities to healthcare providers, taking the easy road can be damaging and produce models that don't serve patient needs or improve workflows. For example, oftentimes, sentiment analysis based on social media is used to evaluate hospital or doctor performance. Twitter, Reddit, and other platforms are scraped to see what people are saying and whether the sentiment is generally good or bad. This technology has been around for decades and is relatively easy to do. It's also easy for patients and the end-user to understand. However, sentiment analysis is a crude measure of success or failure.

In order to perform accurate sentiment analysis, you need a solid baseline. Social media is self-selecting because users don't post about an average experience; they usually only p ost about the very good or very bad ones. This makes social media sentiment analysis not particularly credible.

Instead, in order to develop better NLP for hospitals, it's more productive to evaluate specific metrics, not just "good" or "bad." Checking in on keywords for when users are tweeting about wait times or cleanliness can be a useful tool. Large scale text summarization is harder but can be much more powerful for hospitals large enough to have a significantly big social data set.

But even with robust analysis, language represents a moving target. Linguistic drift occurs when words shift in their colloquial or local meaning. A phrase like, "that's sick!" can indicate illness or that something is really cool. And this is a difficult challenge for any NLP model. If a new subpopulation joins a health network or a news event shifts language usage, the only way to keep models accurate is to retrain them.

HEALTHCARE ON THE EDGE: THE ROLE OF IOT

Wearable technology and the Internet of Things (IoT) offer incredible opportunities for data collection and patient notification. Dr. Anthony Chang, the founder of AIMed, thinks that wearables are key to an AI-driven health future.

Just one example of how the healthcare industry is leveraging IoT is Sensoria Fitness, which creates AI-driven wearables. The gear not only reads in basic biometrics but also captures precise information on the way people move through a combination of sensors and conductive fibers. They've developed an app to help runners improve performance, but they also work to aid in patient recovery for those with gait impairments with their Walk app. Their socks can also analyze and help prevent diabetic ulcers and falls.

The FDA recognizes the potential for augmented personal health analysis and has approved AliveCor's Kardia system, which gives patients the ability to take a medical-grade heart anomaly check (EKG) in 30 seconds. The ML system determines whether the user's heart rate is normal or if there is an issue and can send abnormal scans to the user's doctor. With over 30 million EKGs recorded using the system, it has reached relatively widespread adoption.

When we seed funded AliveCor many years ago, we could see that the future of healthcare was going to exist at the Edge," said Dr. William Paiva, Executive Director of Oklahoma State University's Center for Health Systems Innovation. "More and more consumers—from millennials to seniors—are demanding solutions that allow them to manage their healthcare themselves.

Dr. Chang describes an event last year where similar wearables were the key to peace of mind:

> *I received a text from a patient's mother who is concerned about her daughter's heart rate being low in the middle of the night and queried if we can set up a monitor "real-time" so she can be reassured. This level of data transparency with accompanying anticipatory guidance from clinicians or surrogates will be the rising expectation of our patients and families as more accurate and intelligent wearable devices become available.*

"I think it's an amazing future considering that wearables have not even really hit the mainstream yet," Dr. Chang said. "That's all an additional tsunami of data that we're not well prepared for yet."

While wearables may not be applicable for every population, as they become more reliable and less expensive, they will likely achieve more widespread integration. But even as IoT devices become more prevalent in healthcare, it's not as easy as collecting sensor data and calling it a day. Long term success – that is to say, innovation - resides in the transformation of putting data and analytics (or, going a step further, data science, machine learning, and AI) at the very core of the healthcare experience for larger institutions. And this is no easy task.

Processing this tsunami of data is where ML comes in. Once the data is pulled into an integrated environment (like an end-to-end data science, ML, or AI platform), models can process vast datasets to pinpoint deviations (otherwise known as anomaly detection).

IoT data presents plenty of unique challenges that most healthcare players (that is, those not born of the online era) are not fully prepared to take on. Therefore, turning an investment in IoT into real business value vis-à-vis ML and AI requires a strategic approach and a fair amount of organizational transformation. To succeed and remain a step ahead of traditional competitors as well as IoT pure-players seeking entry into the market (especially from information systems pure players), healthcare institutions will need to cultivate data as a core competency.

INTEGRATED SENSORS FOR OLDER ADULTS

While wearables seem to offer solutions to problems of data quality and over-notification, they are not applicable or useful in every situation. Each population must have its local needs determined to find the best solution for them.

The Health Collaborative for Research & Education in Aging and Technology (C.R.E.A.T.E. Health) is part of the College of Engineering at the University of Southern Florida. Their goal is to design or re-engineer integrated systems that can help achieve positive health outcomes for older adults and improved peace of mind for their caregivers.

They created the Homesense project, which utilizes passive sensors placed throughout the home to monitor older adults' health and activity as they continue to live independently in their own residence. When we asked why the Homesense system doesn't take advantage of wearables, Dr. Carla VandeWeerd explained that:

> *As part of our development process for HomeSense, we interviewed older adults about their homebased technology preferences. Through this process, the older adults we spoke with clearly expressed the sentiment that they do not want to be RFID-ed – it felt too invasive and in some cases almost disrespectful or demeaning. In one focus group we had an older adult stated, "I microchipped my dog, I do not want to be microchipped or tagged like this and have people basically following me all around all the time." By contrast, Homesense can actually give us the same sort of information that you can get if you were RFID-ed, but it feels different because the sensors are permanently fixed so it feels more like a person is just moving through their environment rather than giving them the feeling that they have been tagged so that someone can track them 24/7.*

> *The second reason really in terms of wearables is that there's a lot of evidence-based research that supports the idea that, though this is likely to change in the future, for the current cohort/generation of older adults the idea of having to remember to wear something everyday, to plug it in, to constantly put it back on --that's not really as popular as it was originally hypothesized it was going to be. And even when it comes to alarm devices for example, it turns out that people often don't have them on at night when they get up to go to the bathroom, which is one of the most likely times when they are at risk for falling. Further, often, even when they do have it on, they can forget to push the button when an event occurs, or they may be unable to push the button of an alert style wearable, iso even some of the wearables that have been put to market so far have not had this widespread utilization that people were ultimately hoping that they would.*

The notification process of the Homesense system is hyper customizable to the needs of each older adult. Dr. Ali Yalcin describes the notification process:

> *A lot of it is related to unique health indicators and observed patterns. A notification in the form of an email or text message goes out if you have done things out of the norm or you've broken a hard-set rule. I'll give you two examples. If you're using the bathroom more or less than you normally do. Or if there's a concern with leaving the house at night time between let's say 10PM and 6AM, that would trigger a notification. Essentially you can program it for anything you need. We have a list of notifications that we've created that's useful for us, but each one is just processing the data a different way. And they're very flexible, in that you can say if you deviate more than 20% from the normal, I could send an alert to the doctor or caregiver. So you can have a fixed rule or a deviation by a certain percentage from what's normal for you. In designing this system we tried to go with a platform that's as configurable as possible, because you quickly notice that once you get into it, what you think should work works in one house doesn't work in the other house.*

Dr. Carla VandeWeerd and Dr. Ali Yalcin are Associate Professors at the College of Engineering at USF.

II. CHALLENGES TO APPLYING ML AND AI TO HEALTHCARE

"I think doctors inherently will trust other doctors"

Dr. Anthony Chang, founder of AIMed

TRUST, RESPONSIBLE AI, AND EDUCATION

AI in healthcare is not new; in fact, it's been trying to gain footing for years with one major hurdle: doctors don't trust a black box. Many systems that healthcare providers have attempted to adopt have a fundamental issue, which is that the system cannot accurately explain how it arrived at its conclusions (and there are too many factors for humans to understand, so hospitals have been dropping these systems).

In order for AI to help deliver on its promise in the healthcare sector, trust is required to encourage buy-in from every technician, nurse, insurance provider, and clinician. This part, at least, is the same as any business, except the density of "experts" is (luckily) higher. This means that more demonstrated value is required to achieve buy-in to the system.

Taking a step back, one of the main concerns with AI (and building responsible AI) is that there's undetected bias in algorithms. But bias comes from the data that the machine is given, and trained data experts can identify biased data. In fact, it's much easier to minimize bias in data than to unravel all the past experiences and potential biases of a human. So that means before there can be trust, there must be education.

For example, Dr. Chang is working on sharing his specialized knowledge of AI with other clinicians:

> *I think we need to gather the clinicians that have a strong interest in this area as champions and then also hope to inspire other clinicians to get this education and training. I'm in the process of forming a doctor's group that will gather clinicians with a strong interest in AI and data science. I think doctors inherently will trust other doctors when they're given special domain expertise. I think we need to gather the clinicians that are passionate about this and believe in it. I think that's going to go much faster than just the computer scientists working by themselves, thinking that they can replace doctors and sort of starting off with perhaps not as strong a position for a collaboration.*

Keeping doctors and other industry experts in the data workflow is critical to maintain responsible AI. When it comes to ML interpretability and putting an end to the black box, people and processes are critical, but tooling matters too.

Many data science, ML, and AI platforms nowadays (including Dataiku) have built-in checkpoints that keep a human in the loop by allowing people to assess the validity of any step in model construction. Plus, they can provide clear visual analyses and metrics to help people understand the algorithm's reasoning.

On top of the lofty goals of building responsible AI, and doing so through widespread education and democratization, there is

also one other facet to building trust that often gets overlooked - in all industries, but especially in healthcare. And that is the idea that AI systems must be complicated, taking on the largest and most complex problems.

In fact, healthcare institutions would be wise to start small - there are lots of little wins to be had in operational efficiencies that can help build trust in AI systems while also reducing costs and increasing employee (not to mention patient) happiness. Once these systems solving smaller inefficiencies are accepted, introducing other larger systems solving those more complex issues will be much easier. This white paper will talk more about organizational change management in this respect in the section ***Top Use Cases for Data Science, ML, and AI in Healthcare.***

DATA ACCESS AND REGULATIONS

An interesting dilemma facing health researchers is where to look for data to feed their ML models. Data regulations mean that handling even anonymized patient data is difficult and often expensive. HIPAA in the United States and GDPR in the European Union (and beyond) are legislation governing the storage and use of personal (health) data. GDPR classifies all healthcare data as sensitive data, a hyper-protected sub-category of personal data.

Under GDPR, those authorized can still process sensitive data provided they meet certain stipulations, which can be reviewed in Article 9. Under HIPAA, there are no restrictions on the use or disclosure of de-identified health information. However, all identifiable data on the past, present, or future health or payment for healthcare is under rigorous protections.

The most important step in working with sensitive data for an organization is ensuring that the way consent is obtained for collecting that personal data is in line with regulations. From there, the challenge is clearly restricting and controlling access - that is, being able to separate by teams (and realistically probably more granularly, down to individuals), by topics, and by purposes to ensure proper use and access. Keeping track of:

- **Who** the data is about,
- **Where** the data is stored,
- **Why** it is being kept,
- **When** can it be kept until,
- **What** protections are in place, and
- **How** is the data processed.

Outside of a clinical trial in which patients sign specific waivers, there are ways to access patient data in order to train accurate models. For example, the Oklahoma State University's Health Access Network (HAN) is a program funded through Oklahoma's state Medicaid payer, the Oklahoma Health Authority. The HAN provides care and case management services to over 30,000 Medicaid patients in Oklahoma using traditional Registered Nurse (RN) and Licensed Clinical Social Worker (LCSQ) case managers.

Within that organization, there is an active data analytics team that uses advanced analytic tools to identify, stage, and monitor these patients by mining the patient's clinical and claims data. However, this information is rarely a complete picture and often lacks information that could benefit researchers. HAN records rarely include genetic information or sequencing, for example, and only include the minimal demographic data common to health records (age, sex, race, etc.)

Outside of the public sector, pharmaceutical and academic institutions often have ways of accessing data sets to aid their endeavors. Academic institutions can partner with teaching hospitals to request patient data. However, this often provides a limited local picture of health risks and complications. When research teams collaborate cross-institutional boundaries, it improves the quality of their research by expanding the patients they are exposed to. Pharmaceutical companies are better situated with cross-community data through their network of healthcare providers.

Even when data requires investment, this isn't necessarily a bad thing. Dr. Mark Sendak, the Population Health & Data Science Lead at the Duke Institute for Health Innovation, says that *"the first signal of institutional buy-in is that they've invested to actually curate the data themselves to solve their own problems."*

Since organizational buy-in is one of the biggest challenges facing data innovators, this sort of assurance is no small thing. *"I recently listened to a podcast with Alberto Savoia. One of his major points is the value of Your Own DAta. I see the investment that an organization puts into curating it's own data on a problem as a big step towards building a successful solution to that problem."*

But buy-in and access to data is only the beginning of the problems facing those who research health systems; healthcare data is notoriously messy and difficult to deal with. Dr. William Paiva, Executive Director of Oklahoma State University's Center for Health Systems Innovation, affirms that, *"healthcare data is dirty data. Maybe the most difficult data you will find to deal with. It's got duplicate values, it's got missing values. It's just not a very analytic-friendly data."*

Once you have access to the data, cleaning it to the point where the data is in a workable format is critical to extracting value. And even when large datasets are available, you often need to put in the work to build your own datasets.

Dr. Sendak explains that:

> *If a model is developed in external data, and you're trying to get it adopted in your local setting, often times it needs to be completely rebuilt, as local data is critical to benefit the community your building health solutions for. My **first publication** was a case study of the cost required to take an externally developed and validated model and implement that model in **our local settings**. The total cost was upwards of $200,000.*

Yes, the cost in developing usable datasets is high; but it's also a critical component (if not the critical component) to the data-driven future of healthcare institutions. Without good, reliable, data, moving forward with responsible AI based on trust is outside of the realm of possibility.

III. TOP USE CASES FOR DATA SCIENCE, ML, AND AI IN HEALTHCARE

FRAUD

Insurance fraud is a major issue that drives up healthcare costs for insurers, providers, taxpayers, and patients. The risk of unnecessary or nonexistent medical services due to misrepresentations by patients or providers is a costly one; in the U.S. alone, the National Healthcare Anti-Fraud Association estimates that payers spend up to $68 billion a year due to fraud.

Patients are often billed for complicated bundled services they have not received. Alternatively, doctors are deceived into prescribing services or prescriptions that aren't necessary. The threat of fraud comes from a plethora of sources.

In 2016, roughly 46 percent of fraud was a result of billing for unnecessary or unperformed services. Another 25 percent of fraud is attributed to providers who bill under false names, often in an attempt to bill for services they are not licensed to perform. About 10 percent of fraud occurred on unnecessary prescriptions.

While fraud represents a huge cost to every participant in the healthcare system, (so, everyone) it is critical to note that current anti-fraud measures often have detrimental consequences. In rural healthcare environments, anti-fraud legislation preventing doctor kickbacks is slowing down the progress of digital integration and hiring at a time when rural hospitals are running with an impossibly small number of doctors and clinicians. We'll dive into the intricacies of rural healthcare in more depth later, but this demonstrates that a new anti-fraud solution is sorely needed. The health insurance provider Aetna already uses 350 machine learning models to combat fraud, and new models are coming out of research centers regularly.

Machine learning - specifically anomaly detection - presents a solid opportunity for the industry and regulators to fight back against fraudsters.

But again, it's not easy, and will involve fundamental organizational change. It also comes back to the data - having as much data for anomaly detection as possible allows for more accurate models because one never knows which features might be indicative of an anomaly.

Using multiple types and sources of data is what will allow the healthcare industry to move beyond point anomalies into identifying more sophisticated contextual or collective anomalies. In other words, variety is key.

It's also important to have real impact with an anomaly detection system, which means models should be scoring data real time in production. Anomaly detection is generally time sensitive, so going to production to make predictions on live data rather than retroactively on test or stale data is more important than ever.

And putting a model in production isn't the end. Iteration and monitoring of anomaly detection systems is critical to ensuring that the model continues to learn and be agile enough to continue detecting anomalies even as user behaviors change.

All this to say that fraud detection in healthcare is one of the most important use cases, but that doesn't mean it's easy. Because of the breadth of its utility, especially as fraudulent activity become more pervasive and complicated, anomaly detection will continue to become more sophisticated, and implementing good fraud detection systems will require an orchestration of not only technology, but people and processes in healthcare institutions as well.

DIAGNOSIS SUPPORT

Increasingly, medical providers are turning to clinical decision support systems (CDSS) to keep better tabs on patients, reduce errors, and lower costs. Algorithms that analyze a patient's electronic health record to flag information that a clinician may not have considered, such as an allergy or a past negative reaction to a drug, can help medical providers avoid costly hospital readmissions.

A notable example of algorithmic implementation comes from the Huntsville Hospital in Alabama, where researchers implemented a sepsis surveillance algorithm that sent nurses alerts "for all positive sepsis screenings as well as severe sepsis and shock alerts." Over a period of 10 months, the hospital saw a 53 percent decrease in sepsis deaths and the 30-day readmission rate declined from 19 percent to 13 percent.

In the coming years, artificial intelligence will become a mainstay of CDSS, helping providers to make better-informed decisions that will lead to healthier patients and less waste.

A recent Nature paper explored the success of this sort of a model for diagnosing common childhood illnesses. The paper found a great deal of promise in the system and anticipates incorporating a similar model to help diagnose more uncommon diseases in the future.

But we may not be quite at sweeping AI diagnostic support yet. Arijit Sengupta published a response to the article in Wired, saying that the AI system was actually worse than doctors at diagnosing more serious illnesses and that this error would be significant for health outcomes. Arijit rightly states that, "when it comes to evaluating AI, we should recognize that model accuracy is not the only measurement to consider. We have more questions to ask—and answer—about how AI can best operate both accurately and ethically so it properly augments medical professionals without adversely impacting overall public health."

"We have more questions to ask—and answer—about how AI can best operate both accurately and ethically so it properly augments medical professionals without adversely impacting overall public health."

-Arijit Sengupta

PATIENT CARE

Hospitals and other providers produce an immense amount of data that until recently has been largely unharnessed because it has been unstructured and isolated from other relevant data points. However, the ongoing transition to electronic health records and the development of innovative data-mining technology opens up vast opportunities for healthcare providers to pursue revolutionary prevention and treatment strategies.

Hospital-acquired conditions (HAC) represent a major strain on hospitals. Infections, surgical errors, and falls in medical facilities lead to further medical treatment that payers (insurers, public health programs) will often refuse to cover. While HACs are inevitable at even the best-run facilities, minimizing HACs can go a long way in improving a hospital's financial position and patient care.

Hospitals seeking to reduce the rates of HACs struggle to identify underlying causes due to the high volume of reported incidents across multiple departments. A thorough analysis previously required significant staff time and cost.

However, some providers have made significant progress in identifying sources of HACs by leveraging advanced data analytics and ML. In an effort to understand the spread of multi-drug resistant organisms (MDRO) among patients, Augusta Health[34] in Fishersville, Va., pioneered an innovative approach that combines electronic health records with geospatial mapping tools to visualize the incidence of MDROs in the facility.

Penny Cooper, Augusta Decision Support Manager, to HealthITAnalytics. explained that:

> *We used images of the hospital floor plan imported into Tableau, then we geocoded the locations of the patient rooms on top of that using XY coordinates on those images, so it looks like the rooms correspond to the floor plan. Then we incorporate the positive organisms, the date, and all of the rooms the individual was in – they might move from the ICU to a step-down unit before they might be discharged from our 3 East medical unit – so you would see the time period that they spent in each one of those locations.*

As a result, the hospital was able to track the spread of infection and respond more quickly to stop it from spreading further, all while gaining insights into process failures to avoid in the future.

Dr. Richard Embrey , Augusta Health Chief Medical Officer explained that:

> *Once we understand those patterns, we can work with our infection prevention experts to say, for example, that we need to be sure to use ultraviolet cleaning techniques in these rooms, or spend extra time making sure that room to the right of the entrance is receiving a little extra monitoring,*

A related problem to HACs are Adverse Drug Effects (ADE), which represent a major threat to patient outcomes, provider liability, and the financial position of healthcare providers. While all drugs that are on the market are subject to robust clinical trials, a robust understanding of their ADEs progresses for years after they are approved for treatment. The more data that can be quickly analyzed, the easier it will be for individual providers as well as the medical establishment as a whole to prevent harm from medical prescriptions.

The wealth of data from medical journals presents an enormous opportunity to gain better insight into how different drugs interact with different patients. The challenge, of course, is that the data is too vast for even an army of experts to analyze. Between 2007 and 2016 there were more than 342,000 medical journal articles relating to adverse drug effects, amounting to roughly 1 terabyte of data. Additionally, patients contribute useful information about ADEs on social media every day just by describing their symptoms or the medication they're using.

Machine learning can help medical authorities, medical providers, insurers, and other industry stakeholders delve into this ocean of information to understand whether drugs are producing negative reactions. Why? Because ML is specifically valuable where there is an intersection of massive amounts of data in different formats that needs to be combined and then value extracted not just once, but continually as data morphs and updates.

In 2017, a group of researchers at the Marshfield Clinic designed a neural network that analyzed 1.45 million medical journal articles and 420,000 social media posts in an effort to discover information about ADEs. The results yielded more than 12,000 sentences related to ADEs in the journal articles and 181 ADE-related sentences in the social media posts. They used that data to create a visualization that linked various medications to different symptoms, from weight gain to vertigo.

Thus, in a matter of minutes, the researchers had scoured the world of medical information and discovered important information about the side effects of a number of different drugs. Just as important, the information they found was able to be presented in a way that is easy to understand and therefore be communicated rapidly within the medical field.

Similar to side effects, drug interactions are difficult for physicians to determine until it's too late. Yet advanced analytics can simulate a drug's protein structure and predict its atomic interactions with other drugs. This analysis will enable a better understanding of drugs that interact strangely with each other or toxicity risks.

COMMUNICATION IS CRITICAL

The Duke Institute for Healthcare Innovation focuses on promoting and implementing transformative innovations in healthcare through high-impact pilot programs, incubators, leadership development, and encouraging entrepreneurship from healthcare practitioners who best understand the problems they face.

As part of this work, they are very thoughtful about the language they use to describe innovations. Dr. Sendak explained that language can be critical to clear communications between operations systems and technical tools:

> *I don't think people realize how important change management is when you're trying to change an organization. For example, in the last six months we stopped using the word deploy when we talk about technologies. We use the word "integrate." Because "deploy" has military associations and makes it sound like you just drop something into an environment and it works. Versus, you're always integrating into workflows. You're integrating into existing processes and roles.*

> *One of the other things we really value is the need for external perspectives when you're doing this work. Emerging technologies can cause so many potential unintended consequences and in health care, it's especially important to be prepared to detect and address problems. So we've had a long-standing collaboration with Data & Society, a social science research organization that came out of Microsoft Research. So we've been working with Madeleine Elish, one of their anthropologist research leads, to evaluate our SEPSIS3 Program.*

> *She's done interviews with us all. She's done fieldwork observations with our clinicians. And it was actually out of a conversation with her that we discussed the important differences between using the word "deployment" versus "integration." I think it's valuable to get outside perspective, outside of healthcare, because we're late to the game. Other industries have been embedding machine learning and artificial intelligence into operations for years. And we need to be benefiting from all the learning and all the questions that have been asked from other settings.*

Dr. Mark Sendak is the Population Health & Data Science Lead at the Duke Institute for Healthcare Innovation at Duke University. He holds an MD from Duke in addition to a Masters in Public Policy from Duke and a bachelor of science in Mathematics from UCLA.

OPERATIONS

AI applications that reduce human labor have already become mainstays of the medical world. Virtual nursing assistants, automatic patient booking and other AI-enabled tools have helped providers, pharmacies and insurers reduce costs.

Dr. Chang says that operations are a great stepping stone: “medical administration is just another type of AI. Process automation is something that is a little bit easier to adopt than let's say an ICU readmission criteria model. I think you can start with the easier project and just get used to what it's like to have that kind of capability.” Since these are not medical models, but operational ones, they can sometimes be easier for healthcare providers who are unfamiliar with ML technology to trust.

A major use case for ML in operations is predicting hospital bed or emergency room availability in hospitals. Emergency room overcrowding is an expensive and dangerous problem that predictive modeling can help address. If a surge in need is predicted, hospitals can prioritize discharging patients to free up space or redirect incoming cases to nearby hospitals. ML powered models for bed availability prediction are already in use in Vanderbilt, Boston, and Australia.

In 2018, Amazon, J.P. Morgan, and Berkshire Hathaway announced a venture aimed at developing new technology and data-driven strategies to reduce the cost of healthcare. So far the venture has kept its work a closely-guarded secret, only making public vague goals about reducing healthcare costs. But operations are rife with inefficiencies the task force could be addressing.

RURAL HEALTHCARE

Rural populations present unique healthcare challenges. Even as health organization modernizes, managed care often forgets about or ignores rural markets because it's not where the money is. According to researchers[43] at UNC Chapel Hill, "rural America has 20 percent of the nation's population but less than 11 percent of its physicians. This imbalance has become worse over time."

Additionally, rural populations face higher levels of poverty and barriers to healthcare access than urban populations, making their need for technologically enabled access greater. Rural patients often have to travel farther to reach providers, and more than half lack reliable high-speed internet, another barrier to information.

Patients are less frequently insured and more likely to forego care to save limited funds. Rural life expectancies are anywhere from eight to twenty-two years below the national average, according to Dr. William Paiva, who runs the Center for Health Systems Innovation at Oklahoma State, "Rural America is our largest health disparity zone with 70 million citizens dealing with inadequate healthcare."

The Center for Health Systems Innovation has been trying to address the unique healthcare problems facing rural markets for going on five years now. The Cerner Corporation endowed the Center with data from 63 billion patients over the span of 20 years, so data access is not an issue for them. Everything the Center does has been directly informed by a review process in which they surveyed rural providers to evaluate their most pressing pain points and challenges.

After analyzing this information, they came up with two main categories of work:

WORKFLOW CHALLENGES

According to Dr. Paiva, several organizations work on incentives to get young doctors practicing in rural markets, so the Center sought to fill a different issue: making these existing rural doctors and health centers as efficient as possible.

In an interview with Dataiku, Dr. Paiva explained that "we try to knock workflow impediments out, like transportation issues, scheduling issues, pharmacy, pharmacy pre-authorization issues, all these things that suck operational efficiencies out of rural clinics and therefore take away from available time to provide clinical capacity." Since clinics and rural doctors are overworked and understaffed as is, they need a sustainable path towards survival.

One example the Center has been focusing on is combating no-show rates with automated reminders and robust prediction systems. In one clinic, fixing the patient no-show problem resulted in an ROI of about $95,000 annually, which made the difference between cash negative and cash positive. Dr. Paiva explains that until doctors feel that they are a bit more stable, you cannot address more advanced improvements to patient care: "It's Maslow's hierarchy of needs [...] only once they're viable do they see improving patient care and these types of things as welcome."

LACK OF SPECIALISTS

While there is a dearth of primary care physicians, Dr. Paiva explains that the lack of specialists is even more pronounced: "I call it the lack-of-ologist problem. You know, cardiologists, neurologists, nephrologists. Pick your favorite ologist, they're just not there."

Clinical decision support systems (CDSS) solutions are critical to assisting patients in the rural markets. The Center is developing a suite of tools to assist rural Primary Care physicians in managing diabetic complications based on data collected in routine primary care visits.

Since Opthamologists in rural America are scarce, their first product, which is undergoing clinical validation, is a tool that allows primary care physicians to identify patients who have diabetic retinopathy, the leading cause of vision loss in diabetic patients. While diabetes is more common in rural markets than urban ones, this system is applicable in any health market.

> *Rural markets are ripe for AI and machine learning CDSS tools to assist primary care doctors in the management of their patients because the lack of subspecialists there. Think about it. 20 percent of America, or 70 million citizens, live in rural markets dealing with inadequate healthcare. If you have developed tools to help Primary Care physicians in those health disparity markets manage subspecialty issues you have a real winner.*

Thomas Ricketts at the Cecil G Sheps Center for Health Services Research states that,[45] "only a comprehensive policy, one that links healthcare resource distribution with underlying economic forces and overall economic planning, can deflect or reverse the factors that cause the imbalance."

Yet if legislatures are unwilling or unable to pick up this baton, health innovators can help fill the gaps. Dr. Paiva believes that rural markets are the ideal testbed for emerging health technologies and AI solutions, as they will not be displacing anything:

> *The world and rural markets, in my opinion, are ideal spaces for testing these new alternative AI tools in the marketplace. Most of these new tools are going to disenfranchise part of the healthcare delivery value chain somewhere, somehow. Some position group is going to have their market share gored by an AI platform. If you've got an AI platform automating the reading of radiographs or radio films, well, the radiologists are going to go nuts because you're disrupting their business, and they believe that they can do it better. But if I show that same technology in a rural setting, their response is always, 'Can we get the tool tomorrow? Because we haven't seen an ophthalmologist out here ever.'*

Dr. William Paiva, Center for Health Systems Innovation Executive Director, is focused on transforming rural and Native American healthcare. Within CHSI, the Institute for Predictive Medicine is applying advanced analytics to the largest healthcare database, with 63 million patient's clinical information covering 16 years. Dr. Paiva has 20+ years of venture capital, management consulting, and investment banking experience.

IMAGING

Clinical imaging has skyrocketed in the last ten years, both in number of scans and the size of imaging files. There are simply not enough radiologists to support this need. An additional pre-screening step—between the scan and the radiologist—would free up enormous amounts of time that radiologists currently spend declaring scans normal.

This relatively simple operational process would greatly reduce strain on health imaging systems, granting them more time and energy to focus on the complicated cases. Yet we can afford to dream a little bigger than that too, and dial in on how ML can impact our understanding of tumor growth. Clinical imaging has skyrocketed in the last ten years, both in number of scans and the size of imaging files. There are simply not enough radiologists to support this need. An additional pre-screening step—between the scan and the radiologist—would free up enormous amounts of time that radiologists currently spend declaring scans normal.

ONCOLOGY

The traditional approach to diagnosing brain cancer is histology, where a tumor sample is examined under a microscope. The problem, of course, is that the observation is inevitably subject to differences of interpretation among pathologists, leading to missed diagnoses and traumatizing false positives.

In recent decades, the medical community has increasingly turned to technology that can analyze large amounts of molecular data, reducing the risk of human error. And yet rare types of tumors still demand the histological approach due to the shortage of data points and molecular identifiers.

In 2016, the World Health Organization recommended a combination of histological and molecular analysis for certain tumors. Nevertheless, the process remains fraught with uncertainty; who wins in a knife fight? Researchers explained that, “histological diagnoses face many challenges, including [….] the fact that similar histological features can be shared by many different types of brain tumor.”

In such cases, developments in ML have provided major improvements in terms of accuracy of diagnosis and prognosis. A clustering model can assess a large number of tumors histologically and quickly learn to identify patterns that an individual might not recognize.

Researchers have further enhanced the potential of ML to accurately identify brain cancer by supplying additional factors to inform models. DNA methylation, or epigenetic change, is one notable technique that researchers have long believed holds significant potential but have historically struggled to factor into tumor assessment because it is costly and requires staff with high-level data analysis skills (another argument for education and democratization).

In a 2018 study, however, researchers used ML to analyze "genome-wide methylation data for samples of almost every CNS tumor type classified by the WHO" and may have encountered a major breakthrough for cancer diagnosis and prognosis. The models recognized distinctions between histologically similar tumors and grouped them into sub-categories with more precision than traditional classifications as recognized by the medical establishment.

When asked to analyze over 1,100 tumor samples that had been analyzed by pathologists, the ML model came to the same diagnosis 60 percent of the time, while in 15 percent of the time the computer reached the same conclusion but provided additional nuance by assigning the tumor a subclass. In 12.6 percent of cases, the computer came to a different conclusion, and in many of those cases it assigned the tumor a different grade, "a recategorization that might have implications for prognosis or treatment." This is just the beginning; even the authors noted that the technology was not yet practical on a wide scale.

Here we come back to the problem of access to health data; the more researchers working on larger data sets, the faster data analytics can reach saturation. Another example of the potential of data-sharing relates to the treatment of lung cancer, one of the deadliest cancers.

In 2010, researchers showed that subjecting patients to yearly screenings with low-dose computed tomography, a technique that generates 3D images by taking multiple X-rays from different angles, could significantly enhance the likelihood of early detection of lung cancer and reduce the mortality rate by up to 20 percent.

But the success came with a price: a large number of false positives. False positives put patients at risk of serious psychological and physical harm, both from unnecessary medical interventions and the tremendous stress associated with such a morbid diagnosis. They also represent a substantial financial cost and operation strain on oncology resources.

The more researchers working on larger data sets, the faster data analytics can reach saturation

In an effort to reduce these false positives, the Data Science Bowl, an annual competition led by Booz Allen Hamilton, an IT consulting firm, and Kaggle, an online community of data scientists, offered a $1 million prize (split between the top ten contestants) for algorithms that could more accurately detect a cancerous tumor. The participants were provided with 1,500 scans (and diagnoses) to craft a better algorithm. They then used their algorithms to analyze 500 additional scans (or a test set) where the diagnosis was not disclosed.

The Data Science Bowl showcased the power of human ingenuity coupled with terrific technological horsepower. The NIH and the FDA began work with the winning teams to further study their algorithms; the submissions were creative and took multiple different tacks. While advancements in cancer diagnosis will continue to depend heavily on dedicated researchers, Data for Good efforts like this may continue to increase the probability of a breakthrough discovery. In the context of cancer research, the $1 million prize provided by the Laura and John Arnold Foundation was extremely inexpensive.

CONCLUSION

Machine learning presents enormous opportunity within the healthcare industry to reduce inefficiency and costs while increasing the quality and accuracy of patient care. But before healthcare institutions can realize this potential, they all need to be on the same page.

Every expert interviewed agreed that education, communication, and change management were critical to improving the success of integrating AI into health systems. Dr. Chang stated that from the clinician's side, "we need the education and training to couple with the technology. We need to be in step with the technology, not playing catch-up."

Dr. Paiva agreed that the challenge of communication and understanding diverging industries can be prohibitive to the success of data initiatives:

> *The issue is more with the business people and their understanding of technical clinical stuff, than it is the technical clinical people with the business stuff. In my experience, most researchers and medical people throw all things business under the broad heading of 'that's easy,' because most clinicians and researchers, they think they're the smartest people in the room, and if they can cure cancer and do a [coronary artery bypass grafting (CABG)] procedure, then anything in business can't be that hard. So invariably, one of the biggest hurdles is sometimes to communicate the value of the business principles to the clinician and technical people because they sometimes pooh-pooh its worth. And then the business people, it's often a challenge just to communicate the clinical side in a way that doesn't overwhelm them. But it is a little bit of an art.*

Yet the responsibility is not on doctors alone. It requires enabling every nurse, technician, insurance provider, health start-up, and legislature to understand a bit more about machine learning so that they can begin to trust it and imagine how it can make their lives—and our health—a little bit better.

NEXT STEPS

So what are the next steps for healthcare institutions looking to increase their data literacy overall and start on the path to more widespread AI systems?

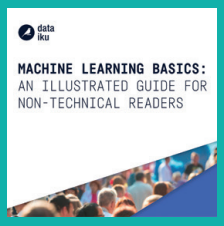
1. CONDUCT AN ASSESSMENT:

Take a critical look at your organization and conduct a full assessment of the current state of data affairs:

- Run a **data audit** to determine first and foremost what data already exists, where it's stored, what format it's in, what it's already being used for, and who has access. The data audit can also help identify potential new sources of data to be collected to aid machine learning efforts.
- Identify key **use cases** for the specific organization (both low-hanging fruit as well as stretch goals) for applying machine learning to data. Ideally, this happens after the data audit so that it's simple to determine whether data needed for each use case is already being collected. Prioritize use cases that address obvious pain points and have operational champions.
- Assess the **data competency** of your teams - who has what skills, what departments do they work in, and where are their skills gaps? What educational resources can you give them access to in order to enable them to improve their familiarity with ML, and eventually grow in their trust of augmented systems.

2. GET EDUCATED:

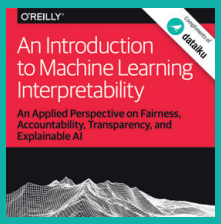
From nurses and technicians to seasoned surgeons, education is key to acceptance and progress. Here are some recommended follow-up readings on a range of topics and technical fluencies that will help you understand machine learning at a fundamental level, learn how it can be applied practically, and grasp the keys to pertinent data privacy and interpretability challenges:



[Machine Learning Basics - An Illustrated Guide]

[How To: Do Anomaly Detection at Scale]

[Executing Data Privacy-Compliant Data Projects]



[O'Reilly - An Introduction to Machine Learning Interpretability]

3. ITERATE ON SUCCESS (AND LEARN FROM FAILURES):

Just like any initiative, not every attempt in the path to AI will be successful. But don't let that dissuade you from evolving and adapting the aspects that do go well to improve the quality of healthcare you offer and the ease at which you can provide it. Again, start small in the beginning, choose several use cases, and see what works (or what doesn't).

data iku

Your Path to Enterprise AI

Dataiku is the centralized data platform that moves businesses along their data journey from analytics at scale to enterprise AI. Data-powered businesses use Dataiku to power self-service analytics while also ensuring the operationalization of machine learning models in production.

20,000+

ACTIVE-USERS

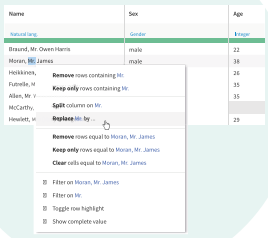
*data scientists, analysts, engineers, & more

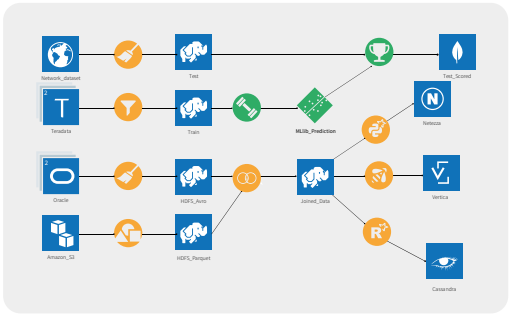
200+

CUSTOMERS

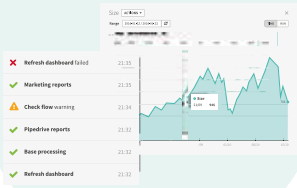
GE · Santander · paloalto NETWORKS · PREMERA BLUE CROSS · FOX NETWORKS GROUP · Unilever · NXP · SEPHORA · KUKA · BNP PARIBAS

1. Clean & Wrangle



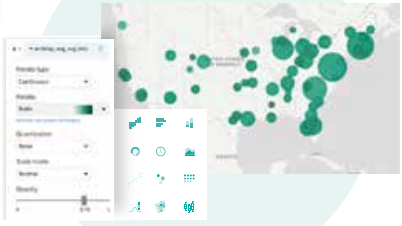


5. Monitor & Adjust

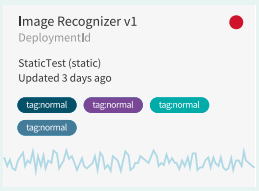


2. Build + Apply Machine Learning

3. Mining & Visualization



4. Deploy to production



data
iku

WHITE PAPER

www.dataiku.com