PARIVEDA

UNCOVERING DARK DATA

How natural language processing can uncover value from the immense amount of unreachable data in the modern medical ecosystem

THE DARK DATA PROBLEM

Up to 80% of your organization's medical patient data is untapped, undervalued or unused

Imagine walking into a physician's office for a checkup. As you describe some pain you've been having in your left knee, the doctor barely looks up from the screen as she clatters away on the keyboard, taking notes. After she resolves your concern and you leave, she will spend another 11 minutes[1], on average, documenting your visit in the electronic health record (EHR) system. While driving home, you wonder how, or even if, those notes are really used to take care of you.

DARK DATA IS VASTLY UNDERUTILIZED

The truth is that those notes, along with other free-form or unstructured pieces of medical data, are not fully utilized. Care providers spend hours each day creating notes, but many of the potential benefits and uses of these notes are untapped. They belong to a collection of data that researchers and industry experts label "dark data": data whose value is unknown and unrealized. A 2016 study estimated that 52%[2] of all data stored and processed globally falls into this category, meaning that the value of the data is unknown and unrealized. For the medical field specifically, an oft-cited statistic pegs the amount of dark data as high as 80%[3]. As data volumes and types continue to increase, this number is expected to rise.

Forward-thinking medical executives look for effective ways to extract value from medical dark data. Unstructured data such as that from progress notes, referral notes and operation notes are a gold mine of rich patient data. They can provide a clinician with more complete context for a patient's care, aid in decreasing research time or cost, ensure accurate billing or save time for busy clinicians.

CURRENT SOLUTIONS ARE INSUFFICIENT

Those who have recognized the value of the data contained in unstructured notes have tried various methods of extraction, including manual extraction, pattern matching and natural language processing (NLP) algorithms. While each has varying degrees of effectiveness, each method is insufficient for meeting the goals of today's healthcare executives.

MANUAL EXTRACTION

The costliest approach to extract information from unstructured fields has traditionally been to ask medical professionals to read through the notes manually. This is often done in research scenarios where the team may not have the technical capabilities to build a more automated approach. Often the type of information the research team is looking for is not easily extracted even with traditional technical aptitude, so researchers must rely on human expertise.

PATTERN MATCHING

Pattern matching is a rudimentary way to look for specific phrases or textual patterns in freeform, unstructured text. Building the appropriate queries and data processing infrastructure requires technical aptitude, yet the end result is too rigid. It can only look at very specific pieces of data in the set. Usually, for pattern matching to work, unstructured notes must have a common template. Free-flowing natural text won't work.

NLP WITH SELF-TRAINED OR SELF-HOSTED MODELS

Some researchers use this advanced technique, which requires extensive machine learning expertise as well as access to a large set of training data and large compute expenditures to train the model. It is the most flexible approach, able to perform best over a wide array of varied input text. However, since patient data is highly regulated, it is very difficult to find a set of training data large enough to build a general-purpose medical NLP model.

Data typically not collected or under-collected in structured fields include

- Patient-reported symptoms
- Provider-observed signs
- Tumor size
- Cancer staging
- Detailed information about a diagnosis or procedure
- Medical/family history
- Social determinants
- Details of current or past medical conditions

UNTAPPED UNDERVALUED | **UNUSED UNTOUCHED**

DARK DATA

Additionally, in late 2018, Amazon released a new managed service called Comprehend Medical, which improves on the current solutions for extracting unstructured data. Building on top of the existing NLP foundation that Amazon's Comprehend provides, Comprehend Medical uses a pre-trained model to extract medical information with high accuracy[4] with a small amount of investment, healthcare organizations can realize large gains in improved patient care. Comprehend Medical provides the broad flexibility, speed and lower cost of earlier NLP solutions but without the requirement of deep machine learning expertise. Additionally, this service provides a general-purpose platform that allows institutions to keep and reuse the structured results, unlike some vendors who focus in one NLP area or whose product does not return reusable information for further study and analysis.

SHARE THE VISION

Driving adoption by verifying operational value through meaningful use cases

Modern advances in machine learning and natural language processing (NLP) are inspiring institutions around the country to look for ways to improve patient care, increase revenue, reduce costs and make research efforts more effective and efficient. These institutions are already seeing huge benefits from incorporating this technology:

- Fred Hutchinson Cancer Research Center[5] uses NLP technology to lower the time needed to match patients with clinical trials from hours to seconds.
- A large children's hospital[6] uses machine learning to identify infectious disease up to three days earlier than normal detection with 70% accuracy.
- Drexel University[7] uses NLP technology to screen for missed billing opportunities for comorbidities treated adjunctly.
- Mercy Health System[7] uses NLP technology to show the life cycle of a heart failure patient and evaluate associated risk factors.

This list will continue to grow as more and more institutions start reaping the benefit.

PARIVEDA CONDUCTED PILOT AT ACADEMIC RESEARCH HOSPITAL

Hearing about the benefits of this technology prompted a large academic research institution to partner with Pariveda to pilot a data pipeline in just five weeks based on the cutting-edge Amazon Comprehend Medical service. Pariveda took a sample of 1,200 patient encounter notes with the goal of processing them with Comprehend Medical and then evaluating the results with the providers to determine valuable areas in which this technology could make a meaningful impact.

EARLY ADOPTERS EXPLORE VISION WITH PILOT PROJECT

The pilot had three main goals:

- Prove that the technology could be put in place by a small, nimble team to start producing results quickly.
- Garner excitement across the organization by showcasing the preliminary results and gaining broad stakeholder buy-in.
- Explore additional use cases with potential project stakeholders to determine the highest value areas to focus on first.

After building the initial data pipeline, Pariveda identified early adopters across the organization to help them understand how extracting unstructured data could aid their goals. Pariveda initially focused on three groups of people:

- Researchers and Data Owners: Researchers are continually looking for more efficient and effective ways to identify potential study participants. NLP of unstructured progress notes uncovers information such as the reason for patient's visit, past medical history, previously reported symptoms and family history information. Accessing this data allows researchers to more quickly determine prospective participants.
- Physicians: Improving quality of care is top of mind for physicians. Pariveda created visuals to emphasize both population-level and individual-patient-level data extracted from medical notes. This dual-pronged illustration was instrumental in sparking creative ideas for how to use the results for quality measures and patient safety indicators.
- Billing Staff: Medical coding continues to grow in complexity every year. The documentation required for each code is often sprinkled through structured and unstructured data. During the pilot, Pariveda showed the various diagnoses, signs and symptoms pulled from a medical note to share how this data might be used to automate or simplify the time-consuming coding process, reducing the likelihood of inaccurate codes and providing additional revenue opportunities.

PRIORITIZING USE CASES

Stakeholder meetings resulted in a plethora of potential use cases. To help prioritize use cases, Pariveda used the goal of maximizing usage and value of NLP as a guiding principle across the organization. Pariveda considered the following three factors:

1. Solution with the biggest impact
2. Stakeholder with the most influence
3. Stakeholder with the appropriate budget

With the top use cases in hand, we began building out extensions of the existing pipeline and the extracted data.

BUILD THE TECH

New machine learning tools allow innovative organizations to extract value from previously untapped dark data

The pilot project was built in five weeks and took advantage of a hybrid on-premise and cloud-based architecture. Amazon Comprehend Medical provided the natural language processing (NLP) engine, and insights were evaluated using an analytics and reporting tool.

EFFECTIVELY STRUCTURE DATA WITH HYBRID ARCHITECTURE

The straightforward data pipeline architecture that was proven with the pilot consists of the following pieces:

1. Data is de-identified first on-premise. A new identifier is assigned to the record to provide continuity and traceability, and all protected health information (PHI) data is maintained within the institution's secure boundary.
2. The records are sent to the cloud where they are queued, processed by Comprehend Medical and stored. The service is accessible through a simple API call. No machine learning expertise is required, no complicated rules need to be written and no models need to be trained. The flexible and scalable nature allowed Pariveda to design for the multiple use cases that needed addressing.
3. An automated process brings the structured data back on-premise. The structured data can be linked back to the original data set with the unique identifier.

A sample architecture for the cloud-structuring step is shown in Figure 1.

AMAZON COMPREHEND MEDICAL EXTRACTS DIVERSE DATA

The managed service extracts the following types of data from an unstructured note:

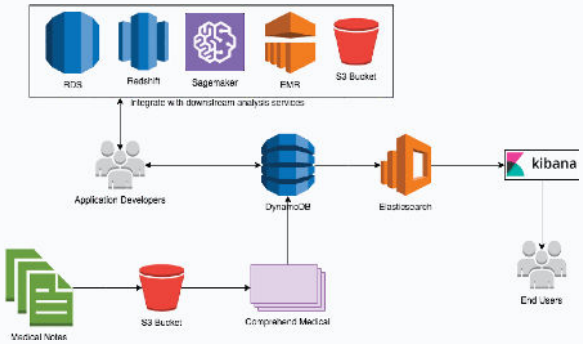
- Anatomy
- Medical Condition
- Medication
- PHI
- Test Treatment Procedure

A sample note (Figure 2) illustrates how Comprehend Medical can tag words or phrases in each of these categories.

NEW DATA MEANS NEW INSIGHTS

After running unstructured data through this pipeline, the output will be leveraged differently for each use case. For the pilot, Pariveda highlighted this new domain of data with an analytics and reporting tool. This allowed the team to showcase the capabilities of this technology and allowed providers and stakeholders to discover new insights from previously inaccessible data.

Figure 1

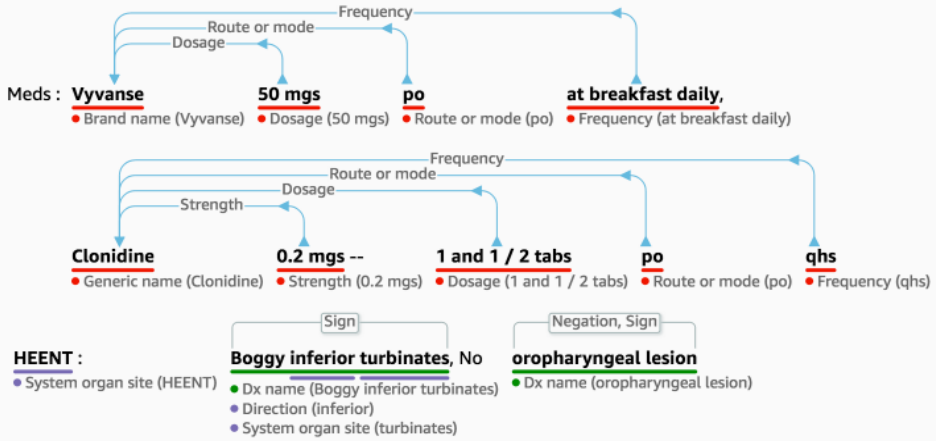


SAMPLE CLOUD ARCHITECTURE

This architecture shows medical notes being stored, structured and then used in a variety of ways, including data warehousing, machine learning and search.

Figure 2

SAMPLE INPUT/OUTPUT



LEARN THE LESSONS

The unique healthcare environment adds additional considerations to any successful healthcare-data project

Healthcare is a unique environment that adds additional considerations to any successful healthcare-data project. Below are some of the dimensions that Pariveda considered as part of the pilot project, which will need to be considered for any similar effort.

ENSURE PATIENT PRIVACY IS TOP PRIORITY

The Health Insurance Portability and Accountability Act (HIPAA) was passed in 1996 to safeguard patient privacy. Organizations are rightly concerned with the benefits of HIPAA to patients as well as the reputational and financial risk of a potential privacy violation. For this reason, the organization Pariveda worked with required complete de-identification of all protected health information (PHI) before allowing the data outside the on-premise environment and into the cloud.

HIPAA provides for two methods to de-identify PHI in order to render it safe for use outside of a highly restricted environment.[8] expert determination and safe harbor. The method most suitable for use in an automated fashion is the safe harbor method.

The safe harbor method of de-identification enumerates 18 pieces of sensitive information that must be removed before the data is marked as not containing PHI. An obstacle that often derails the use of unstructured data in non-secure environments is the challenge of providing automated removal of these 18 pieces of sensitive information. Redaction is simple in structured data, but complex in unstructured fields.

Since our focus was on unstructured data, Pariveda made sure to pick a de-identification solution that would work with it. This involved a combination of open source and proprietary tools to de-identify records as part of the data pipeline on-premise before leaving the protected environment. This solution involved pattern-matching, natural language processing (NLP) and a backup layer of cloud-based de-identification based on Amazon Comprehend Medical.

Ensuring appropriate data privacy and controls is an essential first step to any medical data project.

START SMALL, BUILD EXCITEMENT

Management journals are littered with case studies of projects that began too ambitiously and whose value wasn't proven before scaling. Pariveda's strategy began with a pilot project, then took advantage of periodic small successes to spark excitement across the organization. Through a five-week pilot project, Pariveda showcased the potential value to stakeholders across the administration and business, clinic, research and surgical departments.

As excitement grew, stakeholders shared with other potentially interested parties, and soon additional department heads reached out to discuss potential use cases. Starting small allows value-driven, organic expansion while limiting the complexity, expectations and overhead of large efforts.

IMAGINE AND MOVE FAST WITH INITIAL DATA LOAD

The center of patient data will always rightly be the Electronic Health Record (EHR) system. While the EHR does have integration points, we found that using a data warehouse attached to the EHR provided just as much, if not more, benefit than a direct EHR integration, even though the data was not as fresh. Because the strategy was to start small and build excitement, Pariveda primarily focused on historical data and smaller data sets. This information is readily accessible from the data warehousing team and allows the project to move quickly without slowing for focus on integration concerns. Once value has been shown, integration with the EHR can be accomplished.

CONSIDER SUITABILITY OF THE NLP MODEL

As previously mentioned, earlier efforts in using NLP to extract information have been riddled with challenges. It is important to select the appropriate NLP model for the benefits it provides. We selected the Comprehend Medical model for this use case because it overcomes these challenges in the following ways:

- Comprehend Medical is built on a proven general-purpose model that has been tuned for years across Amazon's wide customer base.
- Comprehend Medical was further trained on the highly domain-specific language of the medical field through a wide training set.
- Comprehend Medical is a hosted solution with a team of experts dedicated to continuous improvements.
- Comprehend Medical has a wider domain of suitability than a model built for one particular use case or discipline. Once the data pipeline is in place it can be used without any adjustments to process data from multiple domains, encouraging quick experimentation and driving down cost.
- Comprehend Medical does not simply provide the output of the use case (e.g., the patients to be contacted as a result of a cohort identification project), but rather the structured underlying data that drives this decision. This means that the same data can be processed once and leveraged for multiple use cases, including new use cases not determined before processing, and the data is fully controlled by the organization.

DISCOVER THE USE CASES

Utilize dark data in revenue cycle management, research, clinic and quality assurance

Pariveda encountered dozens of potential uses while workshopping with stakeholders across the organization and noticed four themes of value that fit into two primary categories: operational excellence and patient care.

OPERATIONAL EXCELLENCE

INCREASE REVENUE AND REDUCE RISK IN REVENUE CYCLE MANAGEMENT

When uncovered, the information contained in unstructured dark data fields can help refine revenue cycle management by exposing over-coding or under-coding in billing claims. Resolving these over- and under-coding errors leads to a greater first-pass acceptance rate and a lowered risk of fraudulent billing. Eventually, this analysis could be applied to claims in near real-time.

Currently, medical coding is done using costly manual review and aided, in some institutions, by computer-assisted-coding (CAC) software. A dark data pipeline, like the one Pariveda implemented, helps to expand computerized coding assistance to more institutions at a lower cost. It also improves quality. Current toolsets find it challenging to extract diagnoses that can be represented in different ways. For example, "atrial fibrillation" is sometimes written as "AF." Amazon Comprehend Medical can accurately identify abbreviations, misspellings and typos in medical text. This intelligence reduces the time a medical coder must spend analyzing unstructured notes, decreases the time burden on clinical staff and improves efficiency.

IMPROVE CLINICAL OPERATIONS

Clinical operations can also take advantage of dark data. Current clinical decision-making systems use only structured data. With dark data, clinical decision-making systems can be augmented and enhanced with a richer data set, including symptoms, signs and diagnoses found only in unstructured notes. Provider time can be streamlined by highlighting the most important information before a patient visit or in a hospitalization summary. Unstructured data can provide a needed boost in the continual search for quality and operational improvements.

PATIENT CARE

ACCELERATE AND ENHANCE RESEARCH EFFORTS

Researchers stand to benefit quickly from a dark data pipeline. One of the most salient uses is to dramatically reduce the cost of using symptoms, signs, medication and medical history in cohort identification for clinical trials or studies. One Comprehend pilot organization reduced the amount of time it took to find suitable participants from hours to seconds[9] – and that's recognizing that "hours" is already on the quick side for how long most cohort identification phases currently take. Because some study requirements, such as family history or social indicators, do not have a dedicated field or are recorded in a form that is not easily accessible, natural language processing (NLP) is a perfect fit to actualize these types of data. And by using this data well, researchers are finding more eligible participants in days rather than months, accelerating study timelines, freeing time and funds available for analysis, ultimately leading to higher-quality studies. These benefits translate into improved healthcare for patients.

ENRICH QUALITY ASSURANCE MEASURES

Unlocked, unstructured data increases quality of patient care by facilitating iterative reduction and avoidance of side effects through data-driven quality efforts. In addition, data warehouse owners can augment existing data stores and cross-reference existing data with the additional depth from unstructured fields. This allows analytics teams to have the most up-to-date and highest quality data for running reports and driving operational decision-making.

CONCLUSION

As machine learning and natural language processing (NLP) continue to mature and new use cases are refined and introduced, forward-thinking healthcare executives need to understand the value of this technology. They must also look for ways to use these tools to drive better outcomes for their organization and for their patients. Pariveda's experience building and validating a dark data extraction pipeline shows that with a small amount of investment, healthcare organizations can realize large gains in improved patient care, speedier research efforts, increased revenue and lower cost.

REFERENCES

1. https://med.stanford.edu/content/dam/sm/ehr/documents/EHR-Poll-Presentation.pdf, retrieved 5 Nov 2019
2. https://www.veritas.com/news-releases/2016-03-15-veritas-global-databerg-report-finds-85-percent-of-stored-data, retrieved 5 Nov 2019
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6372467/, retrieved 5 Nov 2019
4. https://aws.amazon.com/blogs/machine-learning/extract-and-visualize-clinical-entities-using-amazon-comprehend-medical/, retrieved 5 Nov 201
5. https://aws.amazon.com/comprehend/medical/, retrieved Nov 8, 2019 -
6. Large Children's Hospital – Work done by Pariveda Solutions
7. https://www.healthleadersmedia.com/clinical-care/using-natural-language-processing-crunch-your-clinical-data, retrieved Nov 8, 2019
8. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html, retrieved 5 Nov 2019
9. https://aws.amazon.com/comprehend/medical/, retrieved 5 Nov 2019

APPLYING A DARK DATA EXTRACTION PIPELINE CAN PROVIDE IMPROVEMENTS SUCH AS THESE:

IMPROVED PATIENT CARE

MORE EFFECTIVE RESEARCH EFFORTS

LOWER COSTS

INCREASED REVENUE

PREPARED BY

SEAN FISHER

Manager
Pariveda Houston
sean.fisher@parivedasolutions.com

Sean Fisher has over 10 years of experience using technology to solve business problems across healthcare, energy, financial services and education industries. With a background in cloud, emerging technologies, mobile, DevOps and the unique power of software, he focuses on using his deep technical expertise and analytical capabilities while leading teams to unlock unrealized value for organizations.

PARIVEDA

parivedasolutions.com

ABOUT PARIVEDA

Pariveda is a consulting firm dedicated to solving complex technology and business problems by aligning our people-development focus with the mission of our clients. We desire to help our clients achieve lasting success now and into the future. As an employee-owned company, our naturally curious people have strong technical, business and strategic skills to help our clients identify, architect and develop custom solutions. We believe in challenging our clients' thinking, and we are comfortable solving problems without a clear solution. Our solutions create new opportunities for our clients, and we manage the change each solution brings to their company.

Visit our website to explore our perspective on the healthcare industry and read the stories on how we help organizations just like yours align to their full potential and achieve lasting success.