CASE STUDY

IMPROVING QUALITY OF CARE WITH FASTER CRITICAL INSIGHTS USING NLP AND ML

BIOMARKER EXTRACTION FROM UNSTRUCTURED DATA

AXTRIA
INGENIOUS INSIGHTS

> "Over **80% of enterprise data today is unstructured**, including text, image, and voice, and the volume continues to expand."

PREFACE

Over 80% of enterprise data today is unstructured[1], including text, image, and voice and the volume continues to expand.

Traditional analytics use relatively small amounts of structured data like anonymized insurance-claims data and digitalized clinical trials data to gain insights into the real-world. This is due to the legacy data infrastructure that captures data in the form of relational databases and lack of experience and knowledge in transforming unfamiliar datatypes into usable information.

In the past, Electronic Health Record (EHR) and Electronic Medical Records (EMR) data were often limited to a single provider network. The process of obtaining data for specific business and research questions was time-consuming and inefficient due to the varied file types, unstandardized sharing, and the inability to analyze the unstructured data.

The rise of big data technologies and advanced analytics provide us the opportunity to better utilize these un-tapped assets.

In this case study, we dive into mining scanned files including EHR and EMR, specific to breast cancer lab test results. Axtria developed a process that automates the handling of scanned records and analyzes the unstructured data.

The case study will proceed in three parts –

Part 01
Covers the background, objectives and key challenges

Part 02
Explains the technologies evaluated, used and the approach taken to overcome the challenges

Part 03
Describes key takeaways beyond the scope of this use case, Axtria's differentiator and conclusion

PART 1

BUSINESS SCENARIO

A **diversified healthcare company** that provides **specialty practices support** were facing challenges with an inefficient data collection and information delivery processes. The stakeholders including specialty healthcare providers, operations, and senior management were viewing the data in raw image format.

Additionally, there was a critical need to capture biomarker data and other biographic information for breast and ovarian cancer patients, which was going uncovered.

> "To achieve this, we needed to accurately extract and analyze information from **~1,000,000 EHR/EMR** files in the form of PDF, image, and XML files, totaling **~7 terabytes of data.**"

BUSINESS OBJECTIVE

To provide a comprehensive view of patient reports and deliver critical insights to both healthcare providers, and the company, by:

Aiding **physician decision making** and **improving quality of care**

Identifying a new cancer patient population, better understand the disease area and treatment patterns

Improving patient adherence through **tracking** patients and their disease progression

PROJECT CHALLENGE

The work required **accurate extraction and analysis** of information from **~1,000,000 EHR/EMR files** in the form of PDF, image, and XML files, totaling **~7 terabytes of data.**

The key challenges of this process were three-fold:

- The data infrastructure was not set up for storage, rapid access and analysis of unstructured data
- The inherent complexity and variety of the files impeded the analytics process
- Lastly, the quality of the scanned records was low, leading to low-quality image-to-text conversions

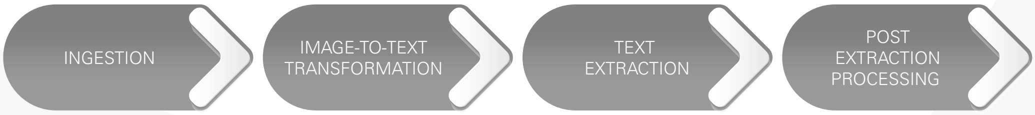
PART 2

AXTRIA'S METHODOLOGY

Big Data Powered by Machine Learning

The focus was on the process of using **cloud and distributed computing** to **upgrade data environments,** experimenting with different tools to **convert images to text,** and applying **Natural Language Processing (NLP)** and **Machine Learning (ML)** to accurately extract information from noisy text.

To address the objectives, a four-step approach was taken:



1. Ingest Big Data into the Cloud

Fast, efficient, and flexible ingestion, storage, and access to data is the pre-requisite of analytics, especially when dealing with unstructured and large volumes of data.

In this case, Spark and HBase were selected for their unique ability to meet project needs.

> “*The key consideration for tool selection: Spark and HBase are compatible with Hadoop interfaces, allowing the **future of the solution to be platform agnostic.** Once the tools were selected, the PDF, JPG and PNG files were organized into the Hadoop Distributed File System (HDFS) and an Oracle database for the next steps.*”

2. Data Transformations

The document resolution quality varied significantly. Some files were handwritten text over printed text, while others where tilted, which made the data transformation process and subsequent extraction challenging.

> “*For image-to-text conversion or OCR, our selection of Google Tesseract was driven by its open source nature, comprehensive settings, and integration options. However, even this best-in-class open source tool could not overcome all the issue of noisy input files, which called for creative text extraction.*”

Transforming images and scanned PDF files to text was done through image processing using OpenCV and Optical Character Recognition (OCR) using Google Tesseract.

The key image quality improvement techniques of filtering and de-noising were combined with subsequent text extraction to experiment with various methods and parameters to optimize data transformation.

Spark – processes large volumes of data quickly in memory

HBase – an agile NoSQL database with efficient storage

3. Text Extraction

In the text extraction stage, unstructured text data was transformed into structured data to provide a **summarized view of each document.**

Simple and complex text extraction was done in two phases with machine learning:

Phase I: Simple extraction employs Named Entity Recognition (NER), classifying words into predefined categories such as a person or organization. Regular expressions (RegEx) uses a computer syntax that finds strings using search patterns. The combination of these two methods was effective for fields with relatively predictable patterns, such as birthdate, names, and gender.

Phase II: For more complex fields such as the BRCA gene variants, which came in different patterns and lengths and added to the complexity, a machine learning-based method was used to improve extraction, discover patterns and rules to identify these fields as a business rules-based approach was not sufficient.

> *"To build machine learning models, we first compiled a training dataset of ~18,000 variants from online databases including the National Health Institute. (Unfortunately, these databases were not comprehensive, so they could not be used as a dictionary for variant detection). Next, features were created and adjusted based on characteristics of each token, which are semantic units in a string divided by blanks and punctuations. Then, we tested multiple algorithms with parameter tuning and feature engineering. Iterations, adjustments and cross-validation show that Gaussian Naïve Bayes algorithm was the most effective at identifying variants."*

The machine learning and testing were done in the **Spark environment** to take advantage of distributed computing, reducing run time of the modeling process.

"To build machine learning models, we first compiled a training **dataset of ~18,000 variants from online databases** including the National Health Institute."

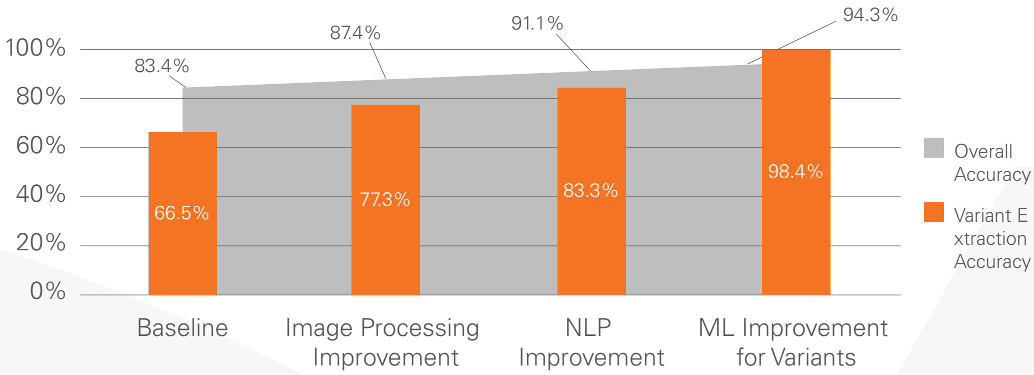
4. Post Extraction Processing

A confidence score was developed to enable efficient human intervention and aid future machine learning training on both field and document level, to determine whether the output value should be re-examined, saving time.

> *"We derived the confidence score from **regression models** created for each field, with metrics from the transformation and extraction stages using random samples of documents. Fields and documents that were lower than the confidence threshold was then re-examined by a human to monitor the output."*

RESULT

4 Stages of Improving Field Extraction



The gray area shows accuracy of all fields, with each stage showing accuracy gain of each improvement. The bar chart indicates the improvements of variant extraction accuracy at each stage.

94% accuracy across different fields, an 11% improvement

32% increase in accuracy, reducing variant error rate to only 1.6%

> "The most noteworthy improvement comes from machine learning, which alone contributed to a 16% improvement in accuracy."

PART 3

KEY TAKEAWAYS

Data availability and robust analytics environments are pre-requisites of utilizing big and unstructured data through image recognition, sentiment analysis, intent detection, and more. Efficient combination of tools and resources is crucial for data analytics that empowers companies to make timely and informed decisions.

Before using machine learning, deep learning, artificial intelligence, or other advanced methods, an organization should have a basic understanding and a realistic expectation of these methods, which are at the highest point of the hype cycle[2]. Meanwhile, understanding the strengths, limitations, and iterative nature of these approaches is critical to the proper usage of these tools.

Furthermore, it is important to note that there is often an accuracy – interpretation tradeoff in machine learning models. Therefore, we need to build in measures, allowing humans to monitor performance and intervene when needed.

AXTRIA DIFFERENTIATOR

Axtria's experience in big data and machine learning helps us identify different approaches that healthcare and life sciences/pharma industry can use to take advantage of the big and increasingly unstructured data. Some of the areas include:

- Auto-stewardship in Master Data Management (MDM)
- Marketing and salesforce empowerment with dynamic targeting
- Automated classification of adverse events
- Pre-screening of patients for clinical trials using NLP to enable adaptive clinical trials

CONCLUSION

Growing unstructured data presents challenges but, more importantly, also opportunities for pharma and healthcare to better understand patients, physicians, and other stakeholders.

Using computing software and hardware improvements, gaining real-world and patient insights is now a reality, and many companies are already heavily engaging in the space[3].

From clinical research to commercial analytics, the emergence of NLP and ML, together with big data enablement, will transform the industry, allowing us to obtain new value from new and existing drugs at a faster rate.

REFERENCES

1. https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/#198a8ffa493a
2. https://www.forbes.com/sites/louiscolumbus/2017/08/15/gartners-hype-cycle-for-emerging-technologies-2017-adds-5g-and-deep-learning-for-first-time/#550f2e885043
3. https://ils.unc.edu/MSpapers/3222.pdf

Founded in 2010, Axtria™ is a Big Data Analytics company which combines industry knowledge, analytics and technology to help clients make better data-driven decisions. Our data analytics and software platforms support sales, marketing, and risk management operations in the life sciences, finance, retail, and technology industries. We serve clients with a high-touch on-site and onshore presence, leveraged by a global delivery platform that focuses on reducing the total cost of ownership with efficient execution, innovation, and virtualization.

For more information, visit www.axtria.com

Follow Axtria on Twitter, Facebook and LinkedIn

facebook.com/Axtria www.axtria.com info@axtria.com Axtria – Ingenious Insights @Axtria +1-877-9AXTRIA