GLYNT.AI

WHITE PAPER

Data Extraction from Documents

The Total Cost of Ownership

GLYNT's Total Cost of Ownership averages 56% below alternate solutions.

GLYNT's high accuracy rate leads to substantial savings for every level of operations from small document sets to high-volume extractions.

EXECUTIVE SUMMARY

For decades, data extraction from documents has lacked highly automated solutions. Business managers have used a suite of technologies, hand-coded software and teams of workers to support data entry and review. Machine learning (ML) solutions have enormous promise to change all of this, promising highly accurate data from a single system with minimum human intervention.

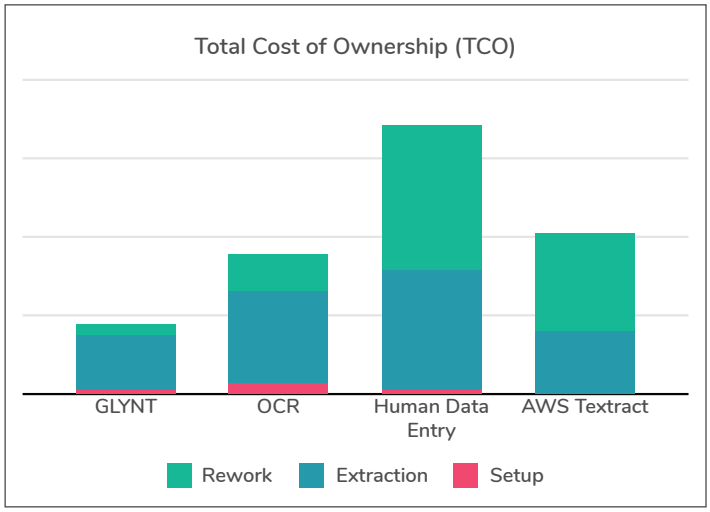
This White Paper, and the accompanying spreadsheet model, provide a detailed analysis of the components of cost for data extraction from documents, including machine learning solutions from GLYNT and AWS. The analysis also includes scenarios across document volume and variety.

The results show:

- GLYNT has the lowest Total Cost of Ownership (TCO), with average savings are 56% across a variety of scenarios.
- The accuracy of extraction is the largest driver of TCO expense.
- GLYNT has the highest accuracy rate, averaging 98%, and unlike most machine learning solutions, this high level of performance is available for smaller data sets. GLYNT has the TCO advantage at all volume levels.
- Straight-through processing (STP) is the automated extraction of data without any human intervention, and excludes any human review and repair of inaccurate data. Because of its high accuracy rate, GLYNT has the lowest TCO for STP data extraction.

The low TCO of the GLYNT machine learning system is shown in the graph below. The White Paper provides further scenario analysis. Additional inputs and assumptions are detailed in the accompanying spreadsheet model.

Figure 1
Data Extraction from Documents
GLYNT has the Lowest TCO



BACKGROUND

For decades, data extraction from documents has lacked highly automated solutions. Business managers have used a suite of technologies, one for each file type, with a host of hand-coded software and teams of workers to support data entry and review. Machine learning (ML) solutions have enormous promise to change all of this, promising highly accurate data from a single system for all file types, with minimum human intervention.

Market demand for new solutions comes from the desire to replace aging bespoke technology stacks with highly automated solutions, the desire to eliminate complex workflows for human data entry and review, and the

desire to capture documents and extract data at the point of use, such as data entry by office workers scattered throughout corporate settings.

In addition, the increased demand for business intelligence and analytics has led to a new market demand to liberate data from documents. Large industries, such as accounting, payments, medical data, and supply chains, require highly accurate data extraction.

With so much opportunity for automated data extraction from documents, it is an opportune time to carefully review current practice and new alternatives in light of real-world requirements and use cases. This makes for a longer TCO white paper, but one that forms a shared understanding of market needs and a clear basis of comparison across technology options.

Table 1 lists the market requirements for data extraction from documents that are addressed in this TCO analysis. Additional discussion of these items follows the table.

Throughout this paper the source of a document the "publisher." This could be a utility issuing utility bills, a vendor issuing invoices, a medical records system issuing a patient record, or a pharmacy issuing a prescription report. And each source that uses a different document layout is a different publisher. This leads to the challenge of document variety.

Table 1
Product Requirements for Data Extraction from Documents

FEATURE	NOTES
Produce clean, labeled data via API	Data flows into structured databases. API needed for speed and scalability
Produce highly accurate data	Errors are costly. Higher error rates lead to higher TCO
Single system to handle multiple file types (PDFs, scans, faxes, document images)	Multiple technologies and workflows leads to higher costs and more errors
Handle a huge variety of documents	Even small volumes of documents come from multiple publishers
Handle changes in document layouts	An estimated 20% of document publishers change the layout of their document each year
Highly transparent, with ability to trace data back to source	Demanding applications, such as accounting and healthcare, require complete transparency
Ability to avoid offshore labor	Often data governance clauses require all data and documents remain within the U.S.

Other market requirements include product features for data security, data retention and the usability of the product UI. For the purposes of this study, it is assumed that a technology solution is first selected based on how well it performs for the challenges listed in the table. If it passes that test, the customer team will then proceed with additional technology due diligence on other features.

Data Extraction Solutions Reviewed

Six data extraction technologies are included in the TCO analysis. Descriptions of each are in Table 2.

Table 2
Data Extraction Solutions Included in the TCO Analysis

DATA EXTRACTION SOLUTION	DESCRIPTION
GLYNT	A single machine learning system that handles PDFs, scans and faxes. User identifies data items to be extracted by document publishers. Point and click UI. Ground Truth created via same UI for 20 documents. Model training in minutes. Ready for production extraction. Verification Engine includes automated data transformations and validation, and a UI for human review.
ZONAL OCR	User creates a template document with zones, eg small regions, around the desired fields. Items outside the zones are not captured. User maps data items to be extracted to the zones and a field list. User then sets up a small test set of documents, adjusts as needed. Often used on large volume, highly regularized non-changing forms such as W-2s. It is not used on PDFs.
FULL-TEXT OCR	Software engineer sets up a script by publisher that searches a document after it has been through an OCR engine. A search is completed for each item on the field list, and then a mapping made to the appropriate data item. The search relies on the relative locations of the search item and the data value. Search terms are document specific, so each publisher requires a mapping from desired field list to the search term. Full-text OCR is more resilient to breakage than Zonal OCR. It is not used on PDFs.
PDF SCRIPTS	Software engineer sets up a script to read a PDF files by publisher. Typically a team of engineers maintains a library of scripts for the company's high-volume publishers. There is some code reuse across publishers, but in general the number of software engineers needed to setup and maintain these scripts increases linearly with the number of publishers.
HUMAN DATA ENTRY	Large-scale human data entry shops deal with a huge amount of document variety and c onstant document change, so they continue to rely on humans. For best results, two or three data entry teams are used on the same documents. Small-scale data entry occurs at sites such as medical offices and corporate settings. Human teams don't have the speed or scalability required by the market.
AWS TEXTRACT	In December 2018 Amazon announce Textract, its machine learning system for data extraction from documents. It is currently in preview mode and not available for general use. However, reading the Textract and Ground Truth documentation allows some estimation of its capabilities. Textract does not need a setup by publisher, but in this study it is assumed that a small validation data set by publisher will be prepared by the user.

ADDITIONAL DISCUSSION OF MARKET REQUIREMENTS

The Need for Accuracy

There are two general use cases for data extracted from documents, and both require highly accurate data. The first use case requires the data to be extracted from a constant stream of incoming documents that arrive each day such as invoices, shipping manifests and so on. The need for accurate extracted data is obvious, as the basic accounting and payment tasks cannot be done with inaccurate data.

In the second use case, data in documents is liberated for use in business analytics or data-hungry AI algorithms. Whether in highly transparent visual displays, software applications or when the liberated data feeds downstream AI applications, inaccurate data causes user confusion, decreases application value and leads to weaker AI models.

The need for accuracy is not surprising. Keynote addresses at AI World in 2017 and 2018 speak to the corporate need for accuracy, typically above 95%, before user acceptance. Without high accuracy rates, users of AI model results are confused and frustrated by errors and corporations don't adopt.

A Single System for All Document Types

Current data extraction solutions are are specialized by file type, with separate technology stacks and workflows for pdfs, scans, faxes and document images. This is enormously frustrating to software engineers and managers who want a clean, easy solution.

PDFs are automatically read via libraries of customized scripts. Document scans, faxes and images either go software applications that leverage OCR technology, known as Zonal OCR and Full-Text OCR. To handle document flows, larger data extraction shops use a mix of PDF and OCR technologies, supported by human data entry as needed. With multiple vendors, and multiple workflows data extraction needs continual management attention.

Machine learning is a powerful alternative, providing the simplicity of a single system. All document types are ingested through the same API. One Data Extraction as a Service provider, one workflow.

The Data Extraction Workflow

There are four components in a typical data extraction workflow.

- **Setup:** The effort of setting up a document publisher. Typically this requires some element of manual intervention and work. There is significant variation in setup costs across the technologies reviewed.

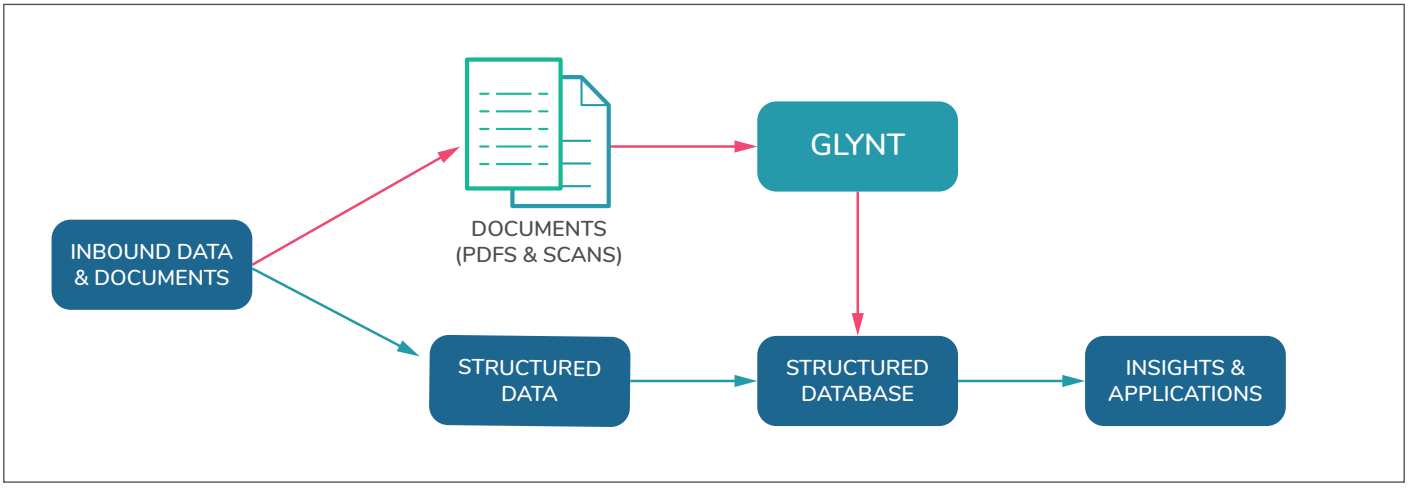
- **Maintenance:** As documents change in layout, the setup by publisher must be redone. This is typically a manual fix.

- **Extraction:** This is the cost of lifting the data out of documents once the system has been setup. Extraction cost is largely license fees to third parties, with the exception of human data entry.

- **Verification:** Every data extraction system will have errors. Machine learning promises to reduce these to negligible levels as algorithms improve and more data is processed. But errors will persist. A Verification Engine collects together automated scripts to clean extracted data. In addition, many tools have human review interfaces to speed manual corrections.

Straight-Through Processing (STP) does not include the human review. When data is extracted from documents and slotted into a larger workflow, STP may not be sufficient. Downstream error correction is even more costly than immediate Verification as the data is extracted.

Figure 2
Enterprise Workflows Demand Accurate Data Extraction from Documents



About Verification

The TCO analysis shows that the Verification expense per field is 10.5X greater than Extraction expense, so the best way to reduce data extraction costs is accurate extraction from the start. But errors will persist, even as they grow smaller over time, and meanwhile the market's demand for accuracy is immediate and key to customer adoption. So the ability to flag and correct inaccurate data is a key product feature for data extraction systems.

Legacy products, such as Zonal OCR, have software triggers that send problematic data items to humans for review, and a user interface that enables low-skilled labor to modify flagged values. Full-text OCR and human data entry shops do the same, often build out in-house code to do the same. It is not clear at this time how verification is done in AWS Textract.

To avoid human review, or at least to reduce the effort, automated software scripts can be used to check business logic or expected data values. One can also use external data sets to reduce errors by running sub-sets of extracted data against the external source, flagging discrepancies.

Because of the market requirement for highly accurate and transparent extracted data, Verification is must-have, not a nice-to-have in the world of data extraction. A combination of highly accurate data with automated Verification is the best strategy for closing any performance gap while keeping TCO low. But one must still review the results, so always, some sort of human interface to the data extraction system is needed.

Batch vs Continuous Processing

Any mention of AI, analytics and business intelligence solutions includes the requirement of "real-time processing." There are several aspects to real-time data extraction from documents.

First consider the Extraction phase in the workflow. Whether a document is extracted immediately upon ingestion or queued up for extraction is largely dependent the customer's willingness to pay for rapid data processing. With the exception of human data entry, all the technologies reviewed here can scale to real-time processing if needed. And this can be done on a continuous basis, as documents arrive.

Next consider the Verification step in the workflow. Automated scripts can strip away predictable errors in the data extraction. But to the extent that there are unpredictable errors, never-before seen data items, and so on, an alert is issued and a human review needed. Human review breaks the continuous real-time processing flow.

So real-time data extraction requires a low error rate in Extraction and reliance on automated scripts in the Verification. Without these, human review becomes the chokepoint and reduces the solution to batch processing.

Scalability

It is often stated that document management is about the three Vs: Volume, Variety and Velocity. These features of document management determine scalability.

Automated technologies can be configured to handle high volumes, and modern solutions will elastically expand as needed. The three Vs would not exist if this was the end of the story. But often Velocity is held back by document Variety, and the lack of throughput speed in manual interventions, such as human review in the Verification.

Document variety often presents challenges. expense. One online lender for example, requires all loan applicants to submit three recent paystubs. The lender has millions of paystubs in its document library, but they come from multiple publishers. And document layout is not constant over time.

Customers report to us that about 20% of documents they process have a change in layout each year. The rate at which a data extraction technology fails to automatically navigate the change is "breakage". For example, a technology that has a 75% breakage rate will need to rework 15% of its document publisher setups. (20% x 0.75)

There are some documents, such as W-2 forms, that have a highly structured layout that seldom changes. Current technologies, such as Zonal OCR, work well. For this small segment of the data extraction market, legacy technologies might be cost-effective. But for the vast majority of the market, document variety and breakage are a constant challenge.

Finally, note how Verification can prevent scalability. A human review to correct errors takes time, management attention and money. And when a first of the month surge in document inflow can create a significant human review bottleneck. Mitigation strategies such as more staffing and overtime only drive expense up further.

A scalable data extraction solution must solve for document variety and human verification. These are the two key frictions to scalability.

BASE CASE RESULTS

Document Scenarios Analyzed

Table 3 summarizes the document variety and scale scenarios analyzed in this study.

Table 3
TCO SCENARIOS

ITEM	SMB	COMMERCIAL	ENTERPRISE
Documents per month	750	75,000	750,000
Number of document publishers	50	500	10,000

ITEM	BASE CASE	HIGH VARIETY	HIGH VOLUME
Documents per month	75,000	75,000	750,000
Number of document publishers	500	1000	500

It is assumed that each document has 3 pages and 12 fields of interest.

Table 4 reports the TCO results for the Base Case. The TCO reflects all costs over a three-year period. Setup expense, for example, is amortized across all documents processed in a three-year period. Similarly, annual Maintenance Expense amortized over all documents processed each year. As is clearly seen from the results, these two costs are negligible over a three-year period. TCO results are largely comprised of Extraction and Verification expenses. Extraction expense is based on license costs, with fees taken from the websites of leading vendors. Verification expense depends on the error rate of the data extraction solution, and these rates are shown in Table 4.

AWS Textract is announced, but not yet available for use. The Base Case assumes an 8% error rate, and a sensitivity analysis was done across a range of error rates, 5% to 12%. The direction of conclusions below are unchanged. See the accompanying spreadsheet for details.

Table 4
TCO Result Detail, Base Case

Three-Year TCO Results ($ Per Document Processed)						
SOLUTION	**GLYNT**	**ZONAL OCR**	**FULL-TEXT OCR**	**PDF SCRIPTS**	**HUMAN DATA ENTRY (1 TEAM)**	**AWS TEXTRACT**
Error Rate	2%	6%	6%	2%	12%	8%
Setup Expense	$0.00265	$0.01061	$0.02953	$0.01969	$0.00354	$0.00149
Extraction	$0.16	$0.31	$0.30	$0.03	$0.38	$0.20
Maintenance	$0.0001194	$0.0015917	$0.0026575	$0.0011811	$0.0001592	$0.0000669
Verification	$0.04	$0.11	$0.11	$0.31	$0.46	$0.31
Total Three-Year TCO	**$0.20**	**$0.44**	**$0.45**	**$0.36**	**$0.84**	**$0.50**
GLYNT Savings		**54%**	**54%**	**44%**	**76%**	**60%**

For all the Base Case results, GLYNT is the lowest cost technology, with average TCO savings of 56%.

Note that the technology solution PDF Scripts has a relatively low TCO. Unfortunately, this solution is only applicable for high-volume PDF file flows. A typical document flow has a variety of small volume publishers and a sizable number of documents that come in as scans. So the PDF script solution must be used in conjunction with other technologies; it is not a complete solution. Thus other technologies and additional workflows are required.

Verification expense is a significant portion of the TCO. Analyses in the accompanying spreadsheet show that Verification expense is 10.5X higher than the original extraction expense per field. For some use cases, there is an eagerness to consider Straight-Through Processing (STP), a fully-automated solution with no human intervention. Results are shown in Table 5.

Comparing STP results across technologies is a bit of comparing apples to oranges, because each technology has a different accuracy rate. So the final two lines of Table 5 show a normalized comparison, extraction costs per accurate field.

Table 5
TCO Result Detail, Straight-Through Processing

Straight-Through Processing Expense Per Document						
SOLUTION	GLYNT	ZONAL OCR	FULL-TEXT OCR	PDF SCRIPTS	HUMAN DATA ENTRY (1 TEAM)	AWS TEXTRACT
Error Rate	2%	6%	6%	2%	12%	8%
Fields Per Document	12	12	12	12	12	12
Fields Extracted Accurately	11.76	11,28	11.28	11.76	10.56	11.04
Setup Expense	$0.00265	$0.01061	$0.02953	$0.01969	$0.00354	$0.00149
Extraction	$0.16	$0.31	$0.30	$0.03	$0.38	$0.20
Maintenance	$0.0001194	$0.0015917	$0.0026575	$0.0011811	$0.0001592	$0.0000669
Total Straight-Through Processing (STP)	**$0.16**	**$0.32**	**$0.33**	**$0.33**	**$0.05**	**$0.39**
GLYNT Savings		**49%**	**50%**	**-201%**	**57%**	**16%**
STP Per Accurate Field	**$0.01**	**$0.03**	**$0.03**	**$0.00**	**$0.04**	**$0.2**
GLYNT Savings		**51%**	**52%**	**-201%**	**62%**	**21%**

TCO RESULTS: ADDITIONAL SCENARIOS

To better understand the TCO results a number of scenarios were analyzed. The results are in Table 6. As is expected, the relative cost ranking of each technology solution remains the same because of the predominance of Extraction and Verification Expense in the TCO costs. For simplicity in presentation, Zonal OCR results are used to represent both Zonal and Full-Text OCR solutions. Human Data Entry and PDF Scripts are dropped from the scenario analysis.

The results of Table 6 shows GLYNT provides savings at levels of document volume and variety. GLYNT performs as well for SMB customers as it does for Enterprise customers. With the importance of "Try before You Buy" sales strategies, only GLYNT provides the on-ramp that allows users to get started, get great results, and grow their usage organically, via a self-serve sales model.

Table 6
TCO Results by Scenario

	SMB	COMMERCIAL	ENTERPRISE	BASE CASE	HIGH VARIETY	HIGH VOLUME
No of Publishers	50	500	10,000	500	1000	500
Documents per Month	750	75,000	750,000	75,000	75,000	750,000
OCR Application	$0.44	$0.44	$0.45	$0.44	$0.48	$0.43
AWS Textract	$0.50	$0.50	$0.50	$0.50	$0.50	$0.50
GLYNT	$0.22	$0.22	$0.22	$0.22	$0.21	$0.20
Avg GLYNT Savings	**56%**	**56%**	**56%**	**56%**	**58%**	**60%**

For all the additional scenario results, GLYNT is the lowest cost technology, with TCO savings of 56% or more.

CONCLUSIONS

This detailed study of the TCO of data extraction has shown that GLYNT has the lowest TCO under a range of real-world conditions. GLYNT's unique ability to deliver highly accurate results on small data sets keeps its TCO low at every operational level, and makes it the dominant technology on the market today.

The study has shown that the expense of Extraction and Verification comprises all of the TCO. Reducing extraction errors is the single best way to lower the TCO.

Setup costs per publisher are negligible when amortized over a three-year period, and while Setup represents a large legacy investment for user of legacy solutions, with GLYNT providing 56% average savings, the payback to change is quick.