Data Science Intelligent Capture

LEVERAGING SMART LEARNING

PARASCRIPT®

PARASCRIPT

Page of Contents

PARASCRIPT

OVERVIEW

WHAT IT IS ALL ABOUT

Expectations for today's digital workforce automation are centered around higher speed and efficiency. As an enabling component for complex document-oriented robotic processes, intelligent capture must process as much document-based data as possible in a 100% *unattended* automation state. The return on investment lives or dies on this ability.

Yet most organization's use of intelligent capture still involves a significant amount of data verification by human staff. With all the automation available to organizations either on premise or via a cloud service, *why does intelligent capture still have a problem living up to its promise?*

When it comes to intelligent capture, organizations are not interested in implementing workflows that require staff to manual sort documents and enter data. They are interested in removing as much of the manual labor as possible. Explore here how to attain true *unattended document automation with high accuracy*.

Intelligent Capture

Living Up to the Promise

In a recent AIIM survey of professionals that manage intelligent capture systems for their organizations, almost 40% of respondents selected as either their first or second choice that accuracy of their system was not good enough. This is unsurprising given that over 45% selected as their first or second choice that complexity of configuration was a significant problem. If systems are overly complex, then it is reasonable that these systems will not deliver as expected.

With all of the automation available to organizations either on premise or via a cloud service, why does intelligent capture still have a problem living up to its promise? When it comes to intelligent capture, organizations are not interested in implementing workflows that still require staff to manual sort documents and enter data. They are interested in removing as much of the manual labor as possible. This is different from, for instance, a CRM system where there is little to no expectation of removing staff. The focus is on making human-centric workflows efficient and controlled, not to eliminate the human element.

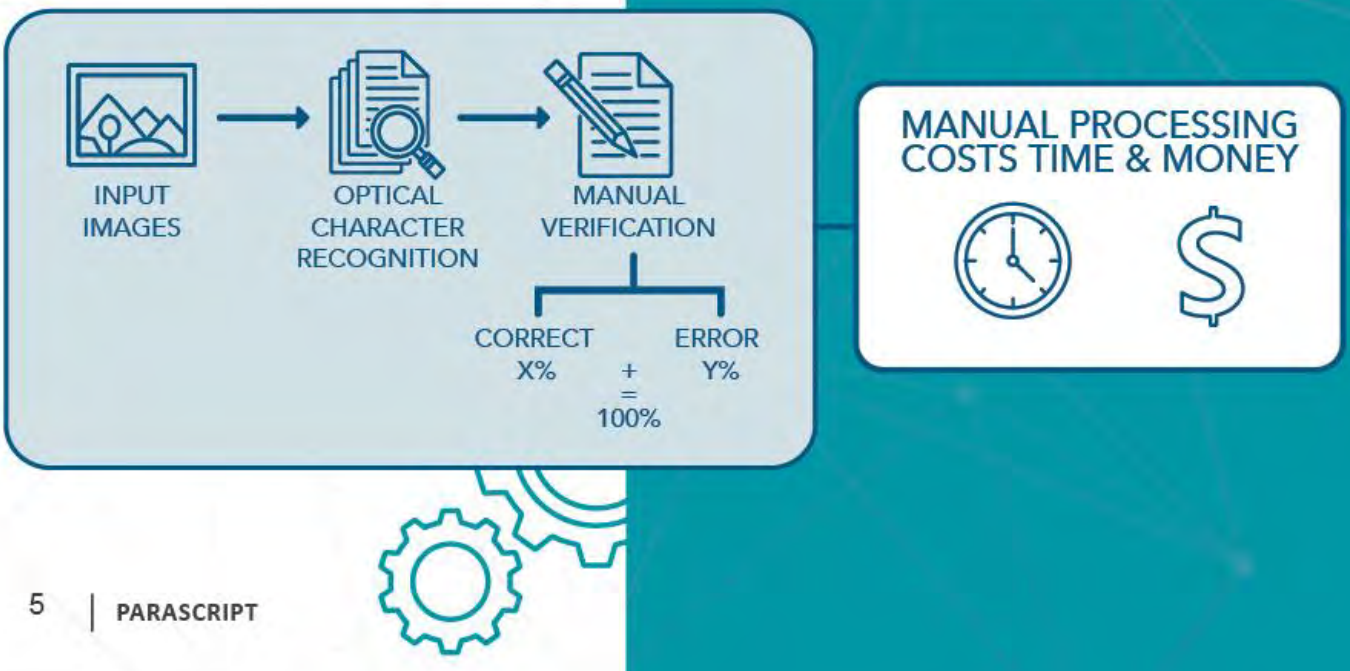
With intelligent capture, the single biggest factor of success is the ability for the system to automate as much document-based data as accurately as possible without involving staff. This is called ***unattended automation***. The return on investment lives or dies on this ability. This means that if you have 10 million invoices per year, understanding how much of this data can be accurately extracted is your objective. The ability to process document-oriented data with no human intervention is known as straight through processing (STP). If you use intelligent capture without the data science approach, it is likely that 5% of your volume could be consuming 50% of your workflow resources.

Precision is the Key to Success

In many respects, intelligent capture software resembles a data analytics platform. Both are measured on the quality of their results. Both require significant attention to sample data. Both require specialists who understand how to configure and measure the system.

However, unlike data analytics, many organizations have approached configuration of an intelligent capture system using a few sample documents with only minimal analysis of the output. The reality is that those organizations end up with almost zero automation due to the lack of reliability of the output. This is because using a few samples only allows a *functional configuration*—one that employs rules, but that has not been analyzed and optimized.

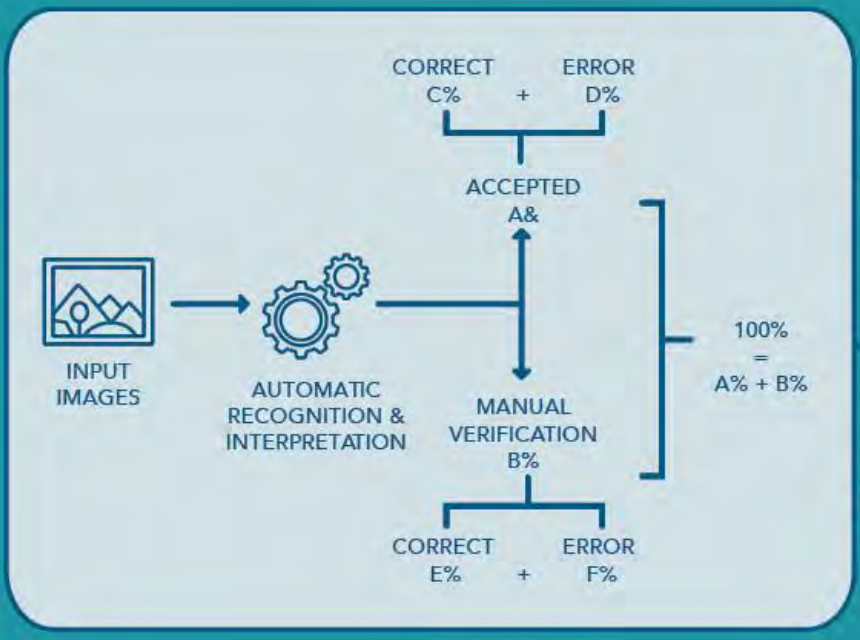
The result is that these organizations essentially review 100% of the data. Here we see a system that runs OCR on everything, but the output is so unreliable that all data is manually verified before it gets output. Even though 100% review is required, it still ends up with some level of error due to problems with manual verification. Humans aren't perfect, even when reviewing OCR output.



Precision is the Key to Success

The ultimate goal for intelligent capture is to have as much data flow through the system as possible, leaving only a small amount for the staff to handle. This maximizes the amount of data that goes straight through. It requires organizations to take a data analytics approach, spending time on curating adequate sample sets and evaluating the results of the system.

Below we see a system properly configured and optimized to reliably classify and extract data using statistical measurement, delivering a high percentage of straight through processing. Some data still requires manual verification, but since the system has been configured and optimized using proper statistical methodologies, a large amount of data can go straight through.



Benefits:

1. Fully automatic processing of a part of data - no manual intervention needed

2. Higher accuracy

3. Faster processing

In order to get to this level of reliability and efficiency, we need to use sample sets that accurately reflect production data. So unlike most, if not all other systems organizations employ, intelligent capture is about precision of document automation which requires a significant amount of attention on data analysis and system optimization.

PARASCRIPT

Intelligent Capture is Data Science

If you are only using intelligent capture to convert images into text, then you don't have to spend time on the precision of your data. However, you are also not getting the full benefit of intelligent capture. intelligent capture is not just *OCR*. It is the domain of technology focused on automating specific document processes including identifying and sorting documents as well as locating specific data elements within documents and reliably extracting them. intelligent capture can be applied to both scanned documents as well as born-digital documents such as Word files or emails, which don't require OCR at all.

Since intelligent capture is all about significantly removing manual work, the focus for understanding the value of intelligent capture is on answering one single question: ***how much work can flow straight through without any manual intervention?*** This question can often be answered with a single number, e.g., "85% flows straight through." In reality, to arrive at this single number or percentage involves performing many more calculations.

PARASCRIPT

What Accuracy Numbers Really Mean

1 **READ RATE**

This is the data field-level percent of ability to locate a particular data field on a page.

2 **PAGE-LEVEL ACCURACY**

Once the capture system locates a data field, it successfully transcribes the information.

3 **FIELD-LEVEL ACCURACY**

Some data is valued at a higher level than other data so system accuracy is measured at the data field level.

4 **CONFIDENCE THRESHOLD**

Calculate the percentage of fields located for each field and multiply that by the percentage of transcription accuracy.

READ RATE

This is the data field-level percent of ability to accurately extract data whether using OCR or other means.

For structured forms where there is good image quality, the percentage can be quite high, as high as a 95% to 99% read rate.

For variably-structured documents such as invoices, read rates are typically lower, depending upon the system's ability to apply the appropriate algorithm to locate any single data field.

FIELD-LEVEL ACCURACY

This is the measurement of read rate for each field. Since some data is more important than others, organizations often will prioritize performance for specific fields.

PAGE-LEVEL ACCURACY

This is the measurement of how many fields are read correctly at a page level. For instance, if a system reads 8 out of 10 fields on average for a single-page invoice, it has a page-level accuracy rate of 80%.

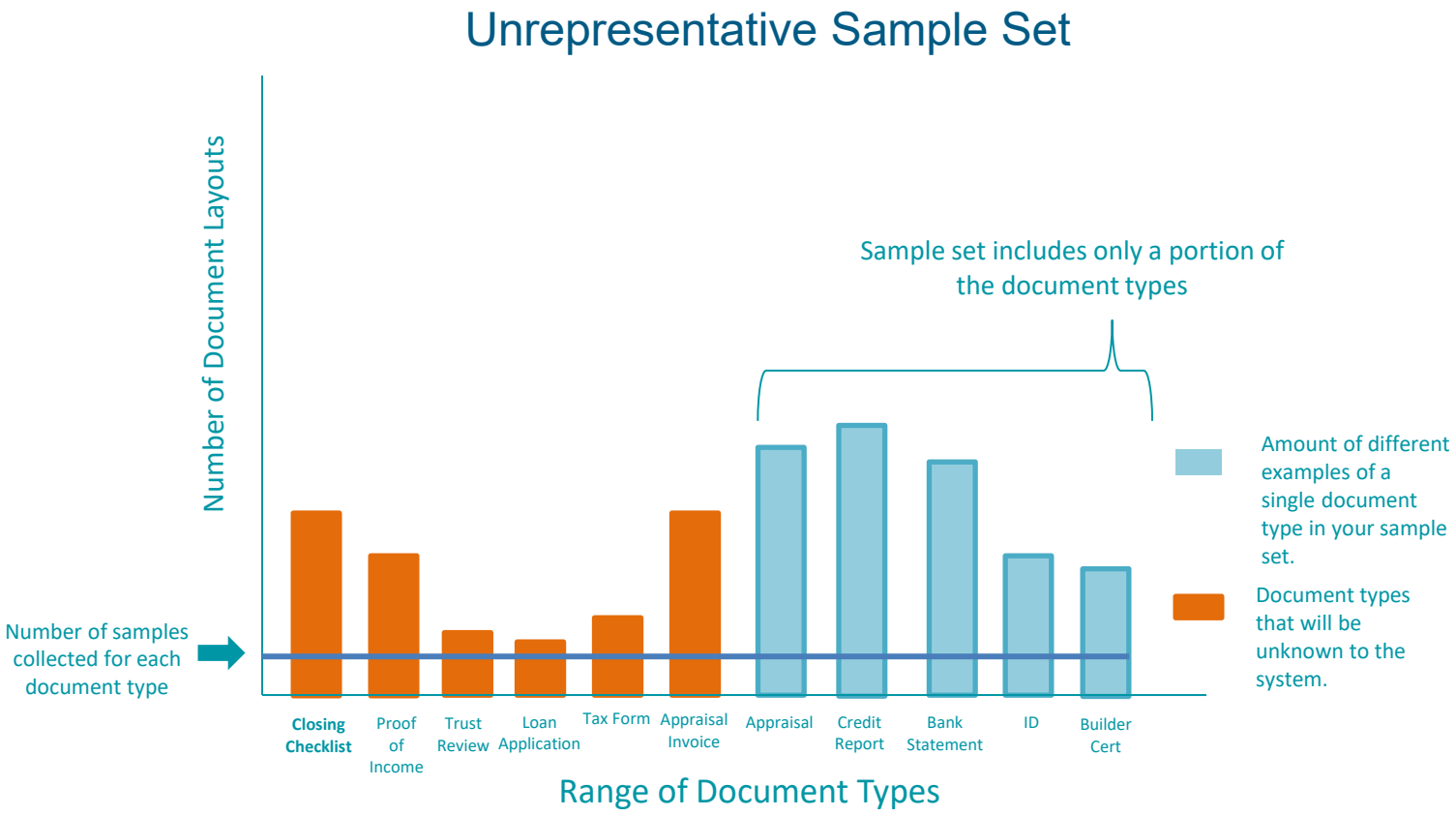
CONFIDENCE THRESHOLD

This is the field-level setting that governs whether data is accepted as accurate and sent straight through or sent for review / manual handling. The ability to reliably set thresholds to achieve specific accuracy rates is the single largest factor in achieving any level of unattended automation.

IN DATA SCIENCE, IT STARTS WITH DATA

When the primary focus for a system is precision, finding the right data to use for both configuration and measurement of the system is essential. This is where data science comes in. You have probably heard of concepts such as ***statistically significant*** or ***margin of error***. These are often used with polling and other survey-based research. For intelligent capture, we use similar measurements for a similar reasons: **to achieve precision**.

There are two issues that must be addressed when evaluating the appropriate sample set with which to use. The first is the coverage of the range of document types and the second is the coverage of variance of documents within each type. These are represented in the graph below.

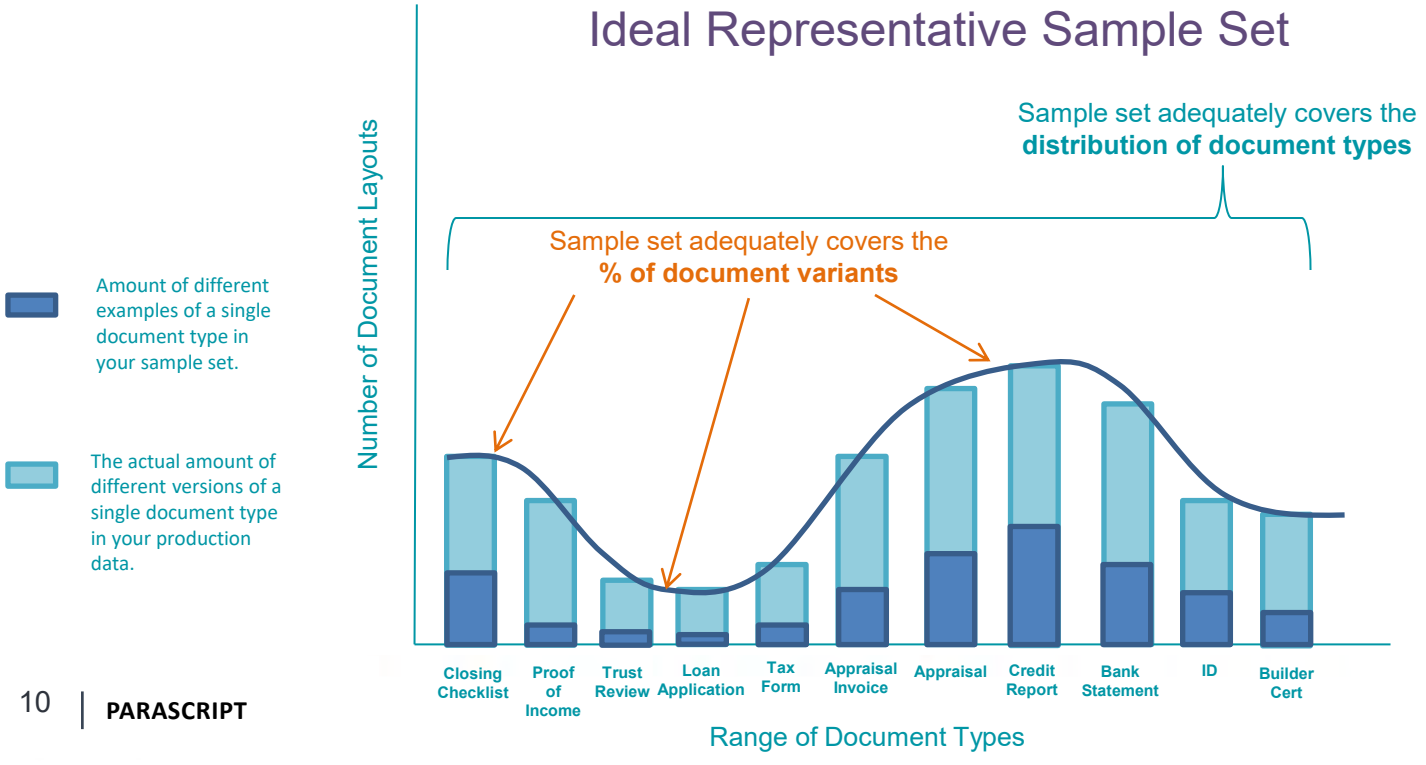


IN DATA SCIENCE, IT STARTS WITH DATA

In trying to estimate the savings rate of the US population, it would be irresponsible for a research group to restrict data collection to only one part of the nation. The results would not represent the true savings rate if focus was on a single slice of the country. It is the same for intelligent capture.

Let's say an organization wishes to automate classification of mortgage documents that includes an appraisal, tax form, credit report, application and a good faith estimate. If the sample set used for configuration does not reflect the range of document types to be automated in production (as represented in the graph below), then the resulting configuration will have a low rate of document classification. In this scenario, if the sample set does not include examples of appraisals, then these documents will be incorrectly classified. The result is the organization will have a large percentage, if not all, of these documents go to exception requiring staff to evaluate and organize these documents. If the sample set does not include examples of the variation within any given document type, then a large percentage of data fields will go to exception—requiring staff to perform data entry.

For example, let's say the organization receives over 1000 different variants (or layouts) of an appraisal yet only uses 20-30 examples in its sample set to configure the system (represented by the blue line). It is a high probability that the configured system will not locate the needed data on a large amount of these documents. The result is that, while perhaps they are properly classified, each appraisal document will require manual data entry. The key to a reliable configuration is to use reliable sample data. The graph below indicates that all document types are represented in the samples and the number of samples within each document type reflects their variance. Ultimately, the task of collecting and curating adequate sample sets is part art and part science. The only way to be 100% certain that your sample set properly represents your production data is to analyze everything. This is impractical and prohibitively expensive. Instead, we use statistical methods to get as close as possible without incurring a significant amount of cost in terms of both time and money.



IDENTIFYING GOOD DATA FROM BAD DATA

Once we have gathered a good sample set and configured the software, it is time to evaluate the results in order to optimize the system. Our sample data along with the *"answer key"* allow us to compare the results of the system to the correct answer. This allows us to calculate the read rate for each field. We also spend a lot of time analyzing another number called the ***confidence score***.

If you are a technical person who has worked with OCR software, then you probably have heard and even made use of a confidence score. All OCR software provides character-level and word-level confidence scores. These scores provide the developer an indication of whether the OCR software finds the answer to be correct. The scores are not representative of probabilities so a score of 80 **does not mean** *an 80% probability of being correct*.

These character/word scores can be useful. However, when it comes to actual data extraction—not simply converting an image to text—another confidence score comes into play, the ***data field confidence score***. Just like page-level OCR, software focused on data extraction produces the field confidence score.

FormXtra.AI is focused on field-level data location and extraction which differs from more generic full-page OCR software such as ABBYY Finereader, Nuance OmniPage SDK or OCR available through Google, Amazon and Microsoft. The field-level confidence score uses the raw OCR character and word-level scores and synthesizes them with other available information to arrive at a final score produced by the software.

This other information can be a data type (e.g., numeric, letters), format (e.g., phone number vs. credit card number), etc. When it comes to achieving true automation, these confidence scores are critical. Unfortunately, most solutions cannot support true automation. To understand why and the potential significant negative impact on your project, read on.

WHY CONFIDENCE SCORES ARE IMPORTANT

In using a field-level confidence score, the main objective is to identify a ***threshold*** that separates good data from bad data. Good data is a ***correct answer***, meaning an accurate, literal transcription of the field as represented on the page. If the input document has a date of birth as 1/1/1970, the field into which the data is transcribed should contain 1/1/1970 as well.

A confidence score is assigned and output by the OCR engine for each field answer. The field-level confidence score uses the *raw* OCR character- and word-level scores and synthesizes them with other available information to arrive at a final score. This other information can be, for example, the expected data type (such as numerals, letters) and format (such as phone number versus credit card number). For instance, if the answer to a phone number field provides confidence scores for each number, a field-level confidence score assembles all of the individual data for each number and combines it with other information about the field such as the expected length of the number (in this case 10 digits), as well as potentially the formatting resulting in a confidence score for the phone number.

When evaluating a field-level confidence score for instance, the OCR engine might output the *date of birth (DOB)* value as *12/5/2008* along with a confidence score of *60*. The field confidence scoring for each data element should output a consistent range of scores for correct answers. These scores should be higher than the scores for incorrect answers so that if you evaluated the results for 100 DOB fields and sorted them according to the confidence score of each, the correct answers should, on average, have confidence scores that are higher than incorrect answers. Although confidence scores are used to distinguish likely correct answers from likely incorrect answers, confidence scores are not probabilistic -- a score of 60 does not mean that there is a 60 percent likelihood that the answer is correct.

In reality, no OCR engine can produce a perfect correlation between a confidence score and whether or not the answer is correct. There will be instances where a correct answer has a low confidence score. Regardless, with tuned systems, the results should indicate an obvious score threshold where the majority of answers above it are correct and the majority of answers below it are incorrect.

USING CONFIDENCE SCORES

Once we understand field-level confidence scores, we can measure and tune field-level accuracy for higher quality data results. To be effective and reliable, this confidence score analysis should be based on several hundred to several thousand samples, ensuring the analysis includes the broadest array of variances in document quality and layout.

0114.tif	HCFA	HCFA:PATIENT_DOB	6/9/1922	92	Correct
0049.tif	HCFA	HCFA:PATIENT_DOB	1/30/1929	91	Correct
0126.tif	HCFA	HCFA:PATIENT_DOB	1/18/1933	91	Correct
0128.tif	HCFA	HCFA:PATIENT_DOB	10/18/1939	90	Correct
0636.tif	HCFA	HCFA:PATIENT_DOB	10/9/1921	89	Correct
0006.tif	HCFA	HCFA:PATIENT_DOB	11/15/1925	89	Correct
0130.tif	HCFA	HCFA:PATIENT_DOB	10/15/1950	89	Correct
0007.tif	HCFA	HCFA:PATIENT_DOB	11/15/1925	88	Incorrect
0005.tif	HCFA	HCFA:PATIENT_DOB	11/15/1925	87	Correct
0008.tif	HCFA	HCFA:PATIENT_DOB	11/15/1925	87	Correct
0002.tif	HCFA	HCFA:PATIENT_DOB	11/15/1925	86	Correct
0123.tif	HCFA	HCFA:PATIENT_DOB	10/31/1924	86	Correct
0078.tif	HCFA	HCFA:PATIENT_DOB	11/25/1932	84	Correct
0073.tif	HCFA	HCFA:PATIENT_DOB	3/29/1968	80	Correct
0649.tif	HCFA	HCFA:PATIENT_DOB	4/2/1978	77	Correct
0129.tif	HCFA	HCFA:PATIENT_DOB	8/21/1938	77	Correct
0090.tif	HCFA	HCFA:PATIENT_DOB	7/12/1920	74	Correct
0074.tif	HCFA	HCFA:PATIENT_DOB	7/6/1964	73	Incorrect
0103.tif	HCFA	HCFA:PATIENT_DOB	1/12/1949	62	Incorrect
0648.tif	HCFA	HCFA:PATIENT_DOB	4/7/1981	61	Incorrect
0079.tif	HCFA	HCFA:PATIENT_DOB	5/19/1932	60	Incorrect
0089.tif	HCFA	HCFA:PATIENT_DOB	3/29/1930	57	Incorrect
0644.tif	HCFA	HCFA:PATIENT_DOB	5/25/1938	55	Incorrect
0104.tif	HCFA	HCFA:PATIENT_DOB	1/12/1949	49	Correct
0001.tif	HCFA	HCFA:PATIENT_DOB	9/10/1918	40	Incorrect
0125.tif	HCFA	HCFA:PATIENT_DOB	3/17/1921	36	Incorrect
0120.tif	HCFA	HCFA:PATIENT_DOB	5/27/1959	24	Incorrect
0023.tif	HCFA	HCFA:PATIENT_DOB	9/2/1987	15	Incorrect
0660.tif	HCFA	HCFA:PATIENT_DOB	10/7/2091	6	Incorrect
0105.tif	HCFA	HCFA:PATIENT_DOB	9/30/2091	4	Incorrect
0628.tif	HCFA	HCFA:PATIENT_DOB	9/17/1949	1	Incorrect

In the above example, the information we are concerned with are the last three columns: the first two are the transcription of the field (date of birth) from the OCR engine and the confidence score for that field (also generated by the OCR engine). The final column shows the result from analysis as to whether the OCR answer matches exactly what is on the document image. Once this is done, answers can be ordered by confidence score from high to low in order to identify the optimal threshold. To find the optimal threshold, you must calculate accuracy provided at a specified threshold. We measure actual accuracy of a specified threshold by dividing the number of OCR answers above the threshold that are accurate by all answers provided above the threshold. In this scenario, all but one answer with a score equal to or above 74 are correct. There is also an answer below the threshold of 74 that is correct. Therefore, the majority of data can be segmented into two groups: one with a field-level confidence score of 74 or above and one group with scores of less than 74. Separation of data into these two groups is the goal of using confidence scores.

SEPARATING GOOD DATA FROM BAD

This ability for OCR to consistently output reliable confidence scores (i.e., erroneous data consistently has lower confidence scores than accurate data) to determine breakpoints is called ***establishing confidence thresholds*** and allows for **true unattended automation** of document processing. This ensures high accuracy and completely removing the need for manual verification of the majority of your data. Only data that falls below the identified confidence threshold (74 in this case) is probably inaccurate and must be manually reviewed. Due to differences in data fields, it is possible and realistic that some fields can use a low confidence threshold while others require a higher threshold; it all depends upon the analysis. Perhaps, the *date of birth* field has a threshold of *74*, but the *social security* field needs a threshold of *88*.

HCFA:PATIENT_DOB	10/9/2009	98	Incorrect
HCFA:PATIENT_DOB	1/19/2000	98	Correct
HCFA:PATIENT_DOB	9/17/1983	98	Incorrect
HCFA:PATIENT_DOB	4/5/1948	98	Correct
HCFA:PATIENT_DOB	4/5/1948	98	Correct
HCFA:PATIENT_DOB	8/17/1937	96	Incorrect
HCFA:PATIENT_DOB	10/15/1931	95	Correct
HCFA:PATIENT_DOB	11/29/1965	90	Correct
HCFA:PATIENT_DOB	11/29/1965	89	Correct
HCFA:PATIENT_DOB	2/26/1985	88	Incorrect
HCFA:PATIENT_DOB	11/25/1959	88	Correct
HCFA:PATIENT_DOB	2/11/1944	88	Correct
HCFA:PATIENT_DOB	12/22/1980	86	Incorrect
HCFA:PATIENT_DOB	10/20/2009	85	Correct
HCFA:PATIENT_DOB	1/16/1944	85	Incorrect
HCFA:PATIENT_DOB	11/16/1942	83	Correct
HCFA:PATIENT_DOB	9/17/2009	83	Incorrect
HCFA:PATIENT_DOB	3/15/1947	82	Correct

In some cases, OCR software cannot produce sufficiently consistent field confidence scores to establish an ordered list of answers that allow selection of a *single confidence score threshold* (where answers above the threshold are mostly accurate). The picture above shows a case where there are too many incorrect scores with relatively higher confidence scores and vice versa for correct scores. When confidence scores are unreliable, an ordered list of answers based upon confidence scores produces many incorrect answers above and correct answers below any threshold. When this is the case, rather than having accurate data to flow through from OCR to, say, the data warehouse without the need for verification, all data is forced to go through manual review. Even if most of the data is correct, the extra review is *costly*, and there is a higher probability that manual review will not identify all incorrect data due to *human error*.

INTELLIGENT CAPTURE REALIZED

Achieving true automation with intelligent capture involves a lot more than just evaluating features and configuring the system.

To create a reliable configuration, you must employ data science to gather an appropriate sample set with which to configure, measure and optimize for straight through processing of document-based tasks.

The good news is that Parascript has a reliable process to shepherd your organization through this journey and much of it is automated using machine learning. This is based upon our decades of experience using these advanced technologies. The result is the highest levels of unattended automaton with the lowest upfront investment.

CONTACT US TODAY

www.parascript.com

888.225.0169

info@parascript.com

PARASCRIPT®