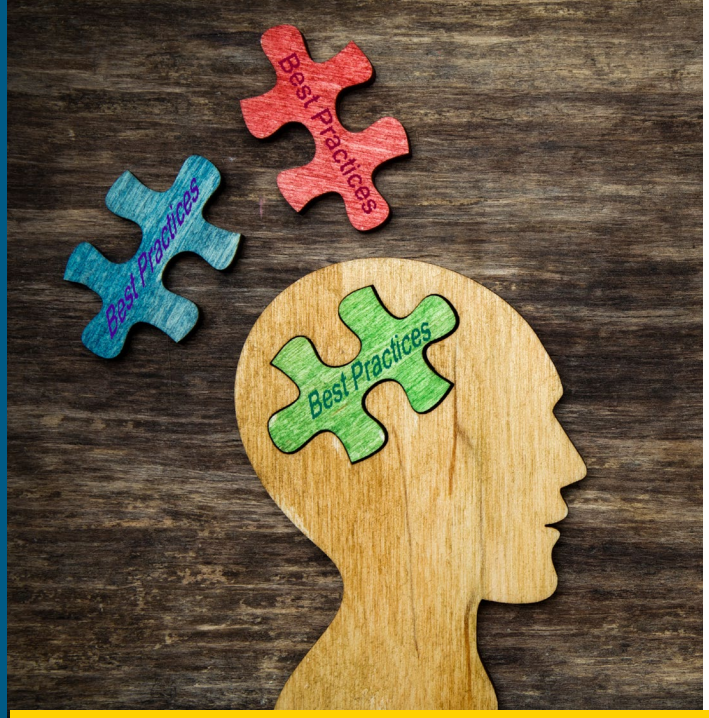


PARASCRIPT®

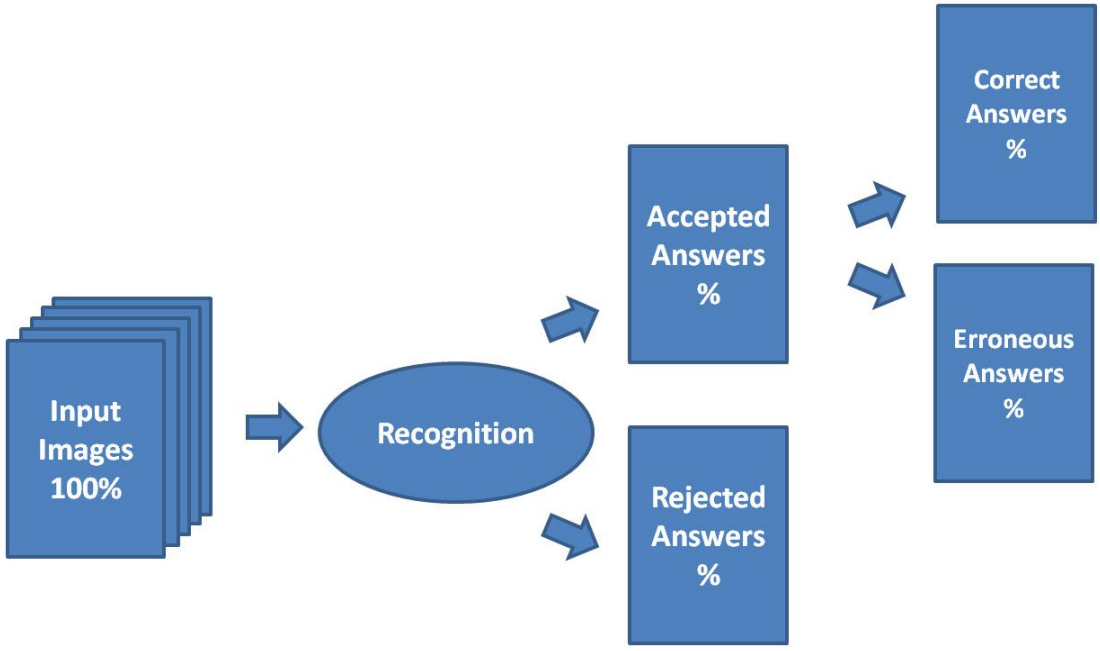
Best Practices:

Leveraging Truth Data to Improve Recognition

Importance of Data Quality

Parascript technology solutions reduce manual data entry while simultaneously increasing accuracy and boosting productivity. Reducing manual intervention through automation reduces the costs associated with forms processing as long as it's possible to achieve an acceptable level of accuracy.

Different tasks require different levels of accuracy to achieve *accepted* results, which have a high probability of being correct. Some applications are tolerant up to a 2 percent error rate, while others require less than a 0.5 percent error rate. The higher the required accuracy, the fewer the results that will be accepted during recognition. Here is a high-level diagram depicting how productivity is calculated.



6273 Monarch Park Place, Longmont, CO 80503 USA
T: 303.381.3100 | toll free: 888.772.7478
F: 303.381.3101 | info@parascript.com | **parascript.com**

PARASCRIPT®

Understanding Confidence Values & Thresholds

Every recognition result has a confidence value associated with it. A confidence value is a number ranging from 0 to 100. However, if a confidence value of an answer is 50 it does not mean that the answer has a 50 percent probability of being correct. A confidence value is an aggregated parameter that is calculated using a complex algorithm and considering many aspects that indicate how confident the recognition engine is about a particular recognition result. Confidence values vary depending on the specific data quality, type and complexity, and on how the data is being used, as well as on many other factors. For example, the confidence value for poor quality data is rarely, if ever, in the 90s so the confidence value scale shifts based on the specifics of the data being extracted. The greater the confidence value, the more confident Parascript is about the result. Confidence values are highly useful when you follow these best practices:

- **Select a certain confidence value** as an acceptable **threshold** for what you determine to be reliable results.
- **Unreliable Results** have confidence values *below—or equal to*—the threshold value. In this case, the field is rejected and must be manually processed.
- **Reliable Results** have confidence values *above* your chosen threshold that are accepted.
  - **Read Rate/Accept Rate** is the percent of the accepted answers to total number.
  - **Accuracy** is the number of correct results among the accepted answers expressed in percent
  - **Error Rate** is the number of erroneous results among the accepted answers expressed in percent.
- **Determining Thresholds**—the higher the chosen threshold, the *lower* the number of the accepted answers, and the *higher the accuracy* of the accepted answers. The challenge is knowing where to set the threshold in order to achieve the required accuracy.

To establish a confidence value as the threshold for each document type requires **ground truth data**.

Ground Truth Data

When testing the system, it is necessary to have a strong and representative sample set of documents. In addition, the actual data must also be recorded that describes what should be extracted from the sample sets—the data that represents nearly 100 percent accuracy.

In machine learning, this sample set along with the true data is called, "**Ground Truth Data.**" Typically, with existing data entry operations, the output from the data entry process can be used as the ground truth once it undergoes quality control in order to understand the error rate. Another common way to get this truth data is to review each sample document and manually record the actual data. A few staff members must dedicate the time to review each sample and enter the correct data. Having this information is crucial to tuning the software to meet specific accuracy requirements. Without it, there is no way to provide any real assurance about recognition accuracy.

In projects and even production environments, truth data allows for objectively measuring the system and understanding how well it performs. Continuously adding to the samples and truth data allows the system to be accurately measured over time. Measuring the error rates is part of this crucial process. The costs of sending bad data into other processes may be *small* or *enormous*. It's impossible to know the truth until you have "the truth."

6273 Monarch Park Place, Longmont, CO 80503 USA
T: 303.381.3100 | toll free: 888.772.7478
F: 303.381.3101 | info@parascript.com | **parascript.com**

PARASCRIPT®

Establishing Threshold Values

To establish a threshold, it is necessary to prepare a representative set of documents and process or run recognition on these documents. Best practices for preparing a set of representative documents to set the appropriate threshold value are as follows:

- **Ensure the number of documents in the set is sufficient.** If the set has too few documents, the results of the investigation will not be representative. Some random variations in the set of documents may occasionally alter the results of the analysis.
- **Determine the number of documents based on the level of accuracy required for the project.** Parascript recommends evaluating the error rate for at least 10 errors that occur among the accepted results. For example, if an acceptable error rate for recognition of a field is 1 percent and at least 10 documents make 1 percent, then the set of documents represents statistically significant results (in this case, a single non-typical error changes accuracy by only 0.1 percent). Accordingly, the set should contain at least 1000 documents. If the required error rate is 0.5 percent, then the minimal size of the set should be 2000 documents.
- **Ensure the *quality* of the documents is taken into account.** The set of documents should be representative, not just with regard to their number, but in the quality of the documents of this particular field, the variety of character styles, and the variety of words that may be encountered in the field.

N	Answer	Confidence Value	Expected Answer	Result Is Wrong	Number Of Images	Number Of Wrong Answers

Each table row contains the following data for each recognized image:

N	The table row number
Answer	The result of recognition
Confidence Value	Its confidence value for the answer
Expected Answer	Correct ASCII value, determined by human processing of the image (truth information)
Result Is Wrong	Enter 0 or 1 for this column according to the following: 0 – The fields "Answer" and "Expected answer" are the same 1 – The fields are not the same
Number Of Images	The number of rows up to and including the current one
Number Of Wrong Answers	The sum of "Result Is Wrong" column

- **Process/run recognition on the documents.** In order to select a threshold value, prepare a table and sort the results in decreasing order of the confidence value.
- **Determine a confidence value for the threshold** by building a new table extracting the following three columns from the first table. *(Note: If there are several rows with the same confidence value, take the lowest one for this step in the process.)*

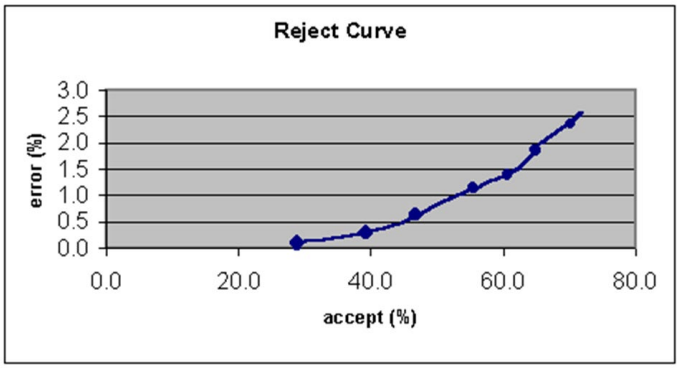
Confidence Value	Number of Images	Number of Wrong Answers

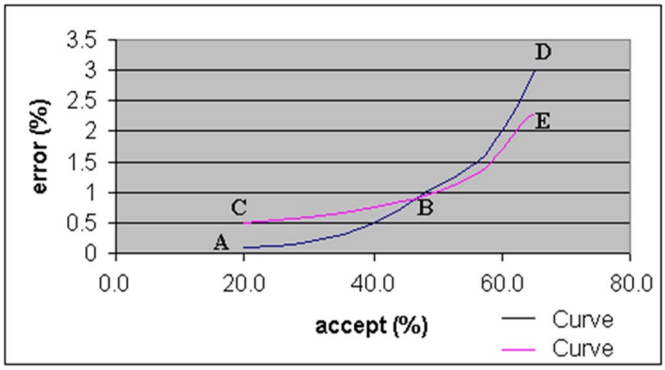
- Once a confidence value is selected, then the number of documents (images) represents the *number of incorrect* results (wrong answers). The number of accepted results for the chosen threshold represents the *number of recognition errors that occur among the accepted results*.

6273 Monarch Park Place, Longmont, CO 80503 USA
T: 303.381.3100 | toll free: 888.772.7478
F: 303.381.3101 | info@parascript.com | **parascript.com**

PARASCRIPT®

- **Relationship displayed as a graph.** The confidence value associated with the required error rate determines the threshold level.





The reject curve allows us to choose a threshold value more accurately than using a table because of the curve shape in the required error value. The reject curve not only describes the effect of moving a threshold, it also allows an effective comparison of the performance of different recognition engines or the performance of the same recognition engine working with differently-tuned parameters. The better the recognition engine and the more tuned the parameters, the closer the reject curve is to the X-axis.

However, it is not always possible to compare two curves, making it worth comparing different parts of them. In the chart to the right, two reject curves are shown of the recognition engine working with differently tuned parameters (Curve 1 and Curve 2). It is clear that as requirements change, one curve reflects better results.

For example, if low error rates (for example 0.5-1 percent) are required, reject curve 1 (section AB) shows better results than reject curve 2 (section CB). This means that the first option of tuning parameters of the recognition engines is preferable. For higher error rates, curve 2 (section BE) shows better results than reject curve 1 (section BD).

Different tasks require different levels of accuracy to achieve acceptable results. Setting the threshold to reflect the task and project accuracy requirements requires some fine-tuning.

Ensuring High Quality Data

Together, ground truth data and machine learning algorithms—that automate the entire development of inferences and rule sets—ensure that the software recognizes and extracts the required data. The net result is higher quality data and performance that improves as more data is fed into the system.

About Parascript

Parascript automates the extraction of meaningful, contextual data from image and document-based information to support transactions, information governance, fraud prevention, and business processes. Parascript software processes any document with any data from any source with its easy-to-use, image-based analysis, classification, data location, and extraction technology. More than 100 billion documents for financial services, government organizations, and the healthcare and life sciences industry are analyzed annually by Parascript software. Parascript offers its technology both as software products and as software-enabled services to our partners. Our BPO, service provider, OEM and value-added reseller network partners leverage, integrate and distribute Parascript software in the U.S. and across the world.