# Advanced Data Similarity Clustering

**What is the ?data similarity? clustering process?**

Using our automated ?data similarity analysis? all documents are organized algorithmically by actual content into proposed clusters, and reviewers assess the clusters for classification relevance, privilege or other purposes. This approach allows us to group similar content rapidly, associating documents in a common context. Initial decisions on inclusion and associating to various issues or a taxonomy can be made on a cluster by cluster basis rather than document by document. The non-textual files contained in document collections are not really susceptible to automated analysis with semantic text based tools, such as Boolean searching or predictive coding. When using that approach they need to be examined separately and, often, manually.

Data similarity analysis provides a holistically complete and accurate assessment of data not offered with other methodologies. All documents will be reviewed in this process. Keyword, Boolean, or predictive coding, really involve sophisticated guesses and return only documents that can associate to those guesses. In the foregoing scenario, there is no guarantee that all documents are reviewed consistently and objectively. Using this new process, all documents in a given data set will have been uniformly considered as part of the clustering process.

**How do we work with this process?**

- The files would be shipped to us and loaded on one of our servers.
- You would have remote access by browser during the processing time.
- If we required, we would do a standard dedupe and denist process first. Second, we would move to the clustering.
- The system presents suggested clusters of documents that you, in turn, associate with particular legal issues that you have identified or to a taxonomy.
- The issues, and accordingly the clusters, can be adjusted or changed as you learn more about the information during this process. During the clustering processing, we would work directly with you and your team.

**Understanding use of ?percentage of similarity?.**

- Both text and visual similarity clustering operate by clustering documents based upon their similarity within a given percentage threshold of similarity. This percentage can be adjusted based on the specific objectives or requirements of the process arbiters ? reviewers, lawyers or other decision makers. So we might start with a given percentage, review the results and adjust percentage up or down as we assess the results.
- We will assist you directly in doing a compare of results at varying percentages,

obtaining a consistent level of comfort in the result.

- The process enables the deployment to human judgment at the right point ? the point at which similar documents suggesting common attributes have been identified, with duplicates and system files culled.

  Note: It is rather easy to eliminate clusters on non-relevant documents. They stand out. Once they are eliminated, the process may include a second look or pass as well. With larger clusters, we review statistically significant samples of documents (based on the total document set) with you to assure that the cluster contains documents that are of sufficient similarity. The sample can be chosen randomly. Again, the goal is to get to clusters that are sufficiently discrete, and that is surprisingly easy to achieve.

**The process is speedy.**

- The speed of process is further enhanced by the preservation and continued application of the coding that supported the creation of any cluster.

  Note: Many people ask whether the machine learning can be applied to newly introduced documents. The answer is yes. So, what has been learned can be applied directly to newly added documents, such as those you might have received from your adversary. New documents are automatically aggregated to existing clusters where appropriate and/or new discrete clusters are created. This speeds, by a large factor, the management of large projects with continuing introduction of new documents to a dataset.

- The process is speeded by breaking up the datasets, even arbitrarility, into smaller blocks of data and processing them in succesStion.

  Note: With very large datasets, we have found it better to work iteratively with smaller chunks of data, even arbitrarily selected. The automated clustering process creates code with the early chunks that is passed on to subsequent chunks. After, say 50% of the data has been analyzed; often no new clusters need to be created. This result, in turn, tends to speed completion of the processing.

- There is usually a significant culling of irrelevant or unusable data

  Note: We find that there typically is a very significant cull of information made during the process. In some data classification projects, the cull rate has exceeded 70%. However, the actual cull rate will vary from project to project. The culling effect and the analysis by cluster results in substantially fewer documents requiring review on a document by document basis. The labor savings are significant.

**What are the deliverables?**

Once the clusters have been reviewed and assigned in support of particular issues, we make a delivery to a review tool. Usually this is the standard review tool that you have on site. If not, we can provide one. The deliverables will also typically include a data discovery and analysis work plan consisting of recommendations to triage and organize the data in a fashion consistent with e-Discovery best practices ? inclusive of hash based file de-duplication.

The deliverables include:

- All information clustered according to data similarity (both visual and textual);
- Mapping as to communication paths reflected in the emails;
- Complete listings of senders and receivers of emails;
- Accurate timelines reflected in the emails and selected other documents;
- Text files in non-searchable format having been rendered to text searchable format; and
- Standard metadata to documents having been identified, extracted and organized for later review.

Once the delivery has been made, we can continue collaborate with you on a mutually agreed upon basis, to complete a more subjective drill down into each of the clusters. This further analysis would be done with apply granular issue codes and search methods which can include the setup of a rules framework to associate particular documents, existing and newly introduced, with various issues and theories that you deem important. Again, we work directly with you at all times during this process to assure the best results.

**How are the results of clustering presented?**

The **presentations of results** using this technology are illustrated in the following pie chart and schedule:

*Chart, Figure 1 represents the ?force multiplier? effect of our automated document grouping methodology. Scenario under which the chart was created ?*

- *We were presented 110,000 documents in a data collection that classified into various ?top level? categories.*
- *Our automatic text clustering algorithms were applied to the data, and the chart represents a sample subset of ~27,000 documents that have been grouped into 30 clusters.*
- *Because the documents in the collection are now automatically grouped by their similarity to one another, the numbers of decisions that need to be made to group the documents are a function of the number of clusters that exist.*

*Chart, Figure 1- 25% of the documents in this collection are represented by 30 clusters.*

*Table 1? The table below shows the number of documents that happen to be in particular clusters. It also shows the ?document type? and the issues for which the documents will be used to support or refute.*

| Cluster ID | Count of Documents in Cluster | Document Type |
|---|---|---|
| CL_001 | 2547 | Purchase Order |
| CL_002 | 2428 | Contract |
| CL_003 | 2209 | Memorandum |

| Cluster ID | Count of Documents in Cluster | Document Type |
|---|---|---|
| CL_004 | 2156 | CAD Drawing |
| CL_005 | 1655 | Email |
| CL_006 | 1311 | Bid Solicitations |
| CL_007 | 1205 | Audit Documents |

*Table 1 above shows how documents in different clusters can be combined, based on the issues they relate to. In the above example the individual clusters of CL_001, CL_003 and CL_007 can become a ?super cluster? of documents that are germane to the issue of fraud. When new documents are introduced into the process, they would be analyzed by the code associated with all three of the clusters to be included or excluded in the ?fraud? super cluster*.

**What are the important characteristics of the process?**

There are **several important characteristics** of the results that demonstrate savings in time and cost and accuracy of results:

- The distribution of documents to associated clusters puts more documents in fewer clusters, i.e. 50% of the documents might appear, say, in the first 20% of the clusters speeding decisions on relevance of a group;
- The document categories (codes) created by this automated process are carried forward and applied to new information introduced to the system on a particular project, i.e. the coding basis for the classification is preserved and, used over and over again as appropriate;
- Human judgment in this organizing stage is leveraged in making decisions on blocks of documents rather than each document;
- The size of clusters and their focus are adjusted by adjustment of the relative percentage of similarity of their content; we work with you and from our experience to test the data on an iterative basis to come up with an appropriate percentage.
- All documents will be reviewed in this process. Keyword, Boolean, or predictive coding, really involve sophisticated guesses and return only documents that can associate to those guesses. There is no guarantee that all documents are reviewed. Using this new process, all documents will have been considered as part of the clustering process;
- Email and communication threads are identified and charted for reference; attachments are separately analyzed but parent child relationships are maintained. This same process is done with all documents associated with container files, such as PST, Zip etc. as well.
- Documents can be reviewed through a browser on our server or delivered to you fully classified in a load file suitable for your in house review tool;
- Clusters can be worked with dynamically ? several clusters can be combined to form a super-cluster and the aggregate coding that creates the super cluster is carried forward and applied to newly added information as if it were a single cluster.

**Our scope and pricing will typically include the following items:**

- Dedupe and denisting to identify and eliminated duplicates and system or application files that typically contain no substantive content;
- PDF normalization that is required for clustering ? all documents must be put in image formate before processing for analysis;

- Text extraction from content identified as important for numerous fiels including : dates, proper nouns, company names, monetary values or other ?entities?, (additional fields can be extracted at a different price point); and
- Clustering

.

---

**Xpriori, LLC ? 2864 South Circle Drive, Suite 401 ? Colorado Springs, CO 80906 ? 719-425-9840 ? Fax 719-203-6496?© 2014-2018**

---