SAS users come from many different business and government organizations. Students in this class are sometimes involved with various security and intelligence problems. The forensic linguistics demonstration is intended to show how SAS Text Miner can be used for these types of problems. Consider some background information for this example. Between 1978 and 1995, a person called the “Unabomber” (who is known to be Theodore Kaczynski) mailed bombs to selected individuals associated with technology research. His bombs killed three people and injured 23. In 1995, he sent a long article entitled *Industrial Society and Its Future* to the FBI and demanded that it be published in a major newspaper or he would strike again. This long article was eventually published in the *New York Times* and *The Washington Post*. The style and content of the writing was recognized by Theodore Kaczynski’s brother, and this ultimately led to Kaczynski’s arrest.

This demonstration uses 232 paragraphs extracted from Kaczynski’s long article, and 1726 paragraphs extracted from the writing of five other authors. The latter are used as “comparison” documents. There is a total of 1958 documents (paragraphs). You run both the Text Cluster and Text Topic nodes on this training data and then create a decision tree model in order to attempt to accurately classify the documents by their authors. Classification such as this is really a form of prediction modeling. In addition, 11 documents are used as unknowns. You use the two models to classify these 11 unknown paragraphs with regard to their likely authors. (Spoiler alert: In this setup, all 11 of the unknown cases were written by Kaczynski.)
Forensic linguistics typically uses predictive modeling to score a document of unknown, but suspected, authorship. The score represents an estimate of the probability that the document was written by a suspect. The value of text mining applied to forensic linguistics is that suspects can be identified for investigation. The text mining results are rarely if ever used as evidence in prosecuting a suspect, although testimony might include a discussion of techniques in describing how the suspect was identified.
The data for this study is real, but the situation is hypothetical. Separation of documents was enhanced for educational purposes. In actual forensic linguistic studies, there are rarely such pure results as those achieved here.

The six authors in the training data are coded as AM, CD, DM, DO, FE, and TK. The initials were changed for the first five authors. TK is Theodore Kaczynski, the so-called Unabomber. The TK documents are paragraphs from the manifesto written by Kaczynski and published in The New York Times and Washington Post. Obviously, when the manifesto was published, the author was not known to be TK. The 11 unknown documents are excerpts from interviews with Kaczynski after he was convicted of murder. Thus, although based on real data, this example is artificial.

Score Data Set: Eleven documents from the same unknown author

Problem: Build classification models on the known documents with six different authors. Apply these models to the unknown 11 documents to determine the likely author of each one.
Stylometry for Forensic Linguistics

This demonstration illustrates how to use text mining nodes and other Enterprise Miner nodes to build classification (prediction) models in a forensic setting. You analyze writing samples from six authors. For five authors, the writing samples in the training data have to do with technical material about statistics and SAS courses. For one of the authors (TK), the writing samples are paragraphs from his published manifesto.

1. Create a diagram named Forensic Linguistics.

2. The training data that is used is contained in one SAS data set called DOCEXTRACTS. It resides on the SAS server.
   Go to Data Sources for the project. Right-click to open Create Data Source. Select Browse to find the SAS data set DOCEXTRACTS in the DMTX51 library.

   ![Select a SAS Table](image)

   In step 5 of the Data Source Wizard when the variables are shown, change the variable roles to correspond to the following:

<table>
<thead>
<tr>
<th>Name</th>
<th>Role</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>Target</td>
<td>Nominal</td>
</tr>
<tr>
<td>DocID</td>
<td>ID</td>
<td>Nominal</td>
</tr>
<tr>
<td>Text</td>
<td>Text</td>
<td>Nominal</td>
</tr>
</tbody>
</table>

   Bring the data set into the diagram.

3. Connect a Data Partition node to the Document Extracts node and retain the default settings (40%/30%/30%).

4. Connect a default Text Parsing node to a Data Partition node.

5. Connect a default Text Filter node. Change the Weightings to explicitly show Log and Mutual Information. Remember, these are actually the defaults that are used here because a nominal target variable is present. It is always a good idea to make this clear in the Property Panel.
6. Attach a **Text Cluster** node to the **Text Filter** node and run it with the default settings. Open the results. Cluster 3 has descriptive terms, such as *power*, *society*, and so on, that are clearly associated with the Unabomber’s long published manifesto. Close the window.

7. Attach a default **Text Topic** node to the **Text Cluster** node and run it. Open the **Interactive Topic Viewer** and look for topics that are likely to be from the Unabomber’s writing. For example, select the third topic shown below. Look at the Terms and Documents windows associated with this topic.

8. Connect a **Decision Tree** node to the **Text Topics** node. Change the default leaf size from 5 to 25 and change the assessment measure to **Average Square Error**. Run the tree and look at the tree in the Results window. Notice how well the leaves of the tree separate the six authors. In particular, the author TK is accurately classified. The overall misclassification rate for all the document extracts can be seen in the Fit Statistics window. These rates are approximately .066, .108, and .073 for the **Train**, **Validation**, and **Test** data sets, respectively. Clearly, good accuracy was achieved.
9. Open the **TKSCORE** data set and designate it as a **Score** data set. It is in the same library on the SAS server, so you need to make another New Data Source for it.

This data set contains the 11 paragraphs that were drawn from TK’s interview after he was captured. You want to see how accurately the tree model classifies these paragraphs. To do this, open a **Score** node and attach it to both the **TKSCORE** data set and the **Decision Tree** node. Your complete diagram should look as shown below.

Run the **Score** node and open the **Exported Data** from the Property Panel. Select the **Score** data and click **Browse**.
Scroll to the far right in the browsing window and look at the last two columns.

The last column gives the model’s predicted author category (Prediction for Author) and the second to last column gives the model probability for this category (Probability of Classification). All 11 paragraph extracts are correctly classified as written by TK. (Remember, the data for this demonstration was enhanced to ensure such a clear-cut result!)