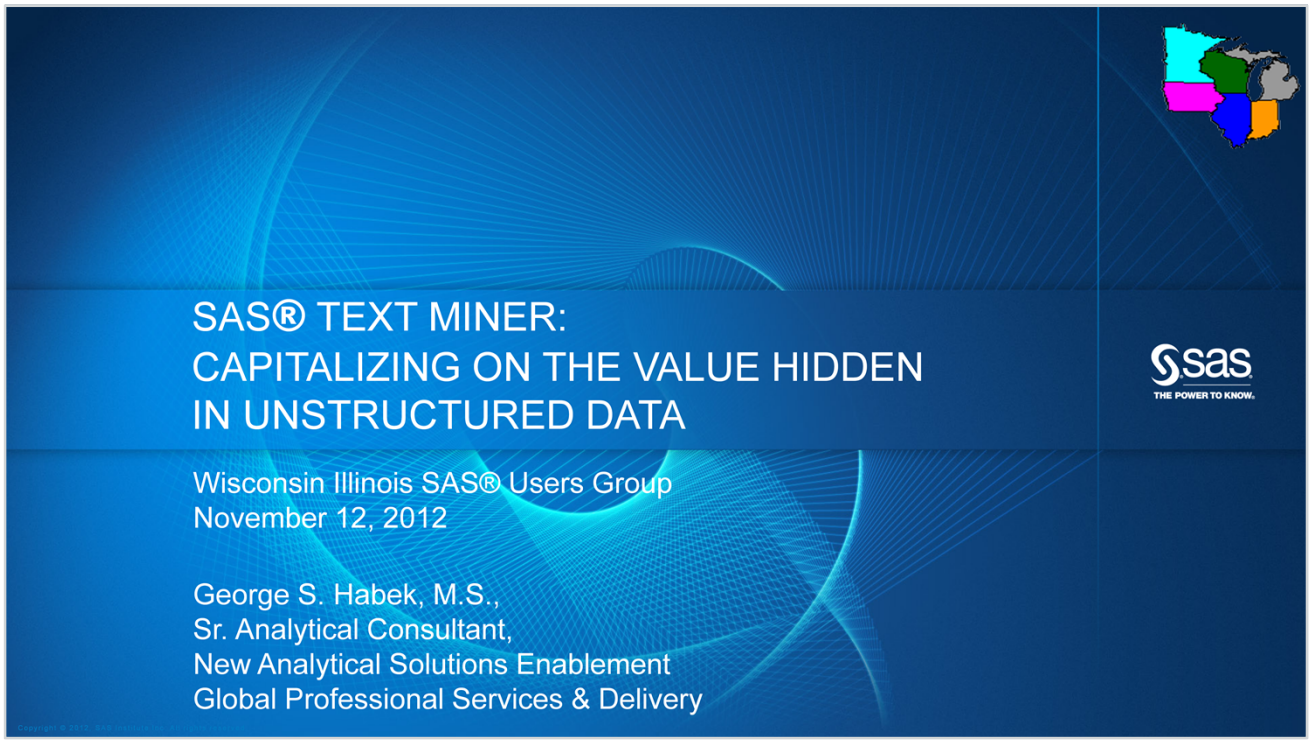
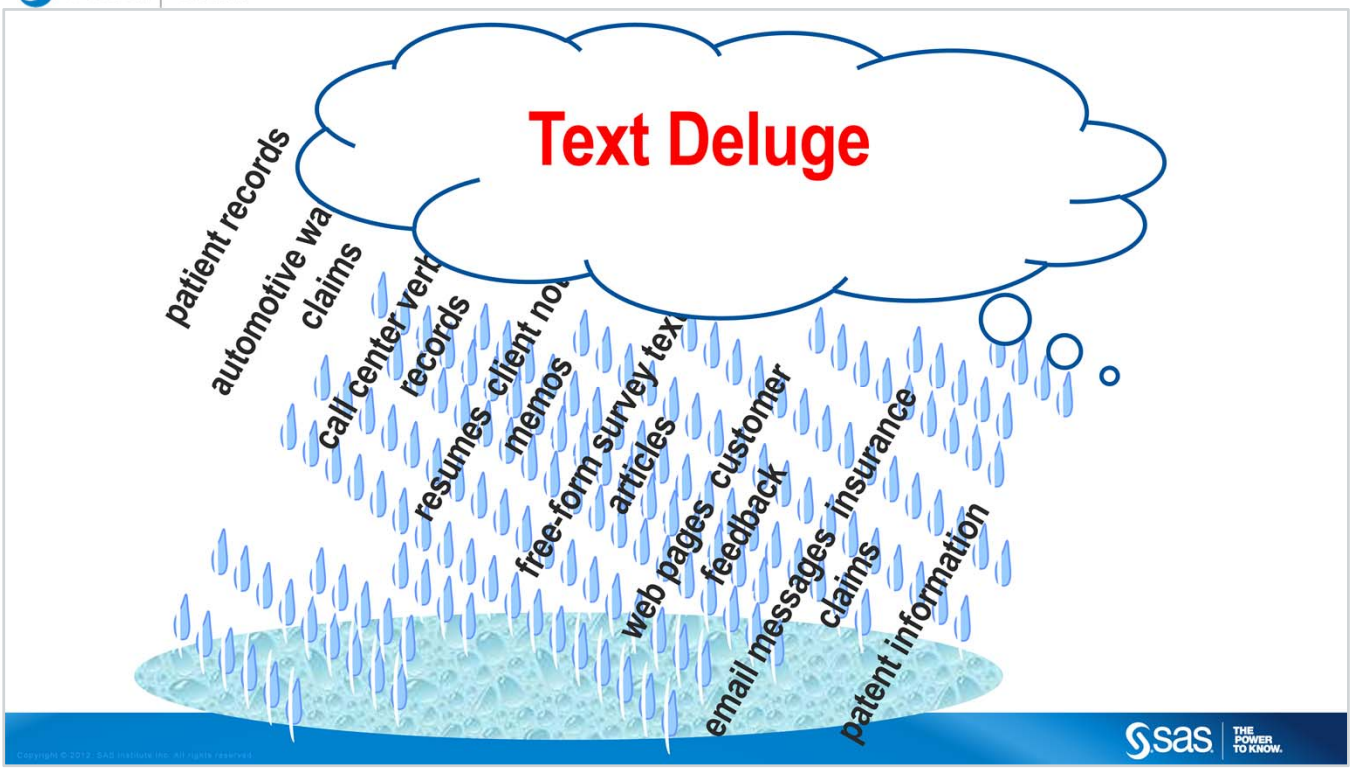




Everyone is aware of the huge quantities of text stored by their organization. Analysts are reporting that up to 85% of customer databases are made up of text. In fact, all data can be stored as text and text mining can help analyze it. SAS customers are using these sorts of texts (e.g. warranty claims, customer feedback, patent information) for a wide array of applications.

Text Mining Definition

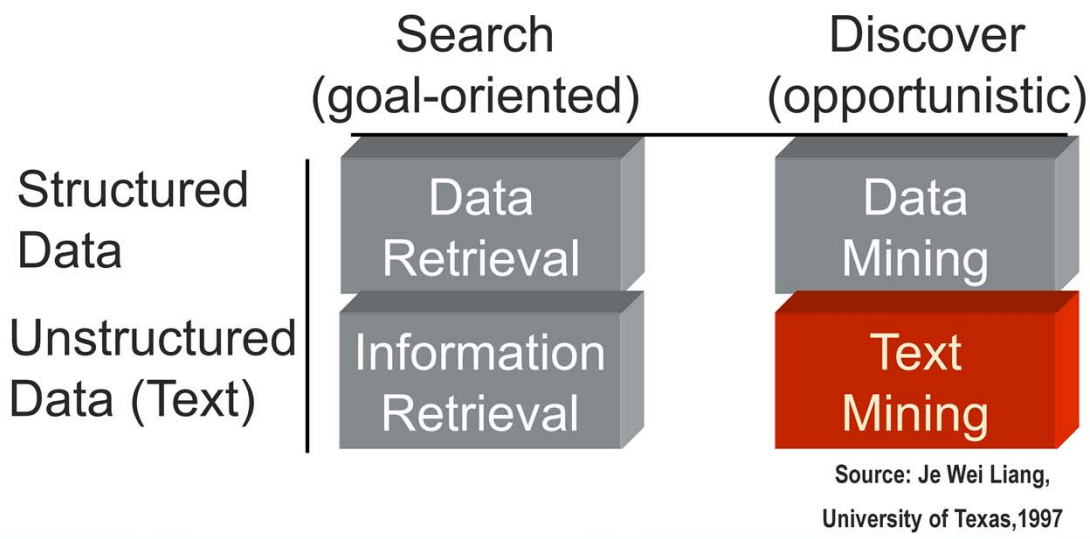
- The process of discovering and extracting meaningful patterns and relationships from text collections.
- Converting textual data into its **numerical representation** for modeling through dimension reduction.

sas THE POWER TO KNOW.

Text mining means a lot of different things to different people so I would like to give you the SAS definition to start us off with a level understanding.
Text mining is data mining (applied to text) combined with natural language processing. The emphasis is on analysis of large document collections.
And the biggest differentiator for SAS is the fact that TM is fully integrated with EM allowing the combined analysis of related structured data.

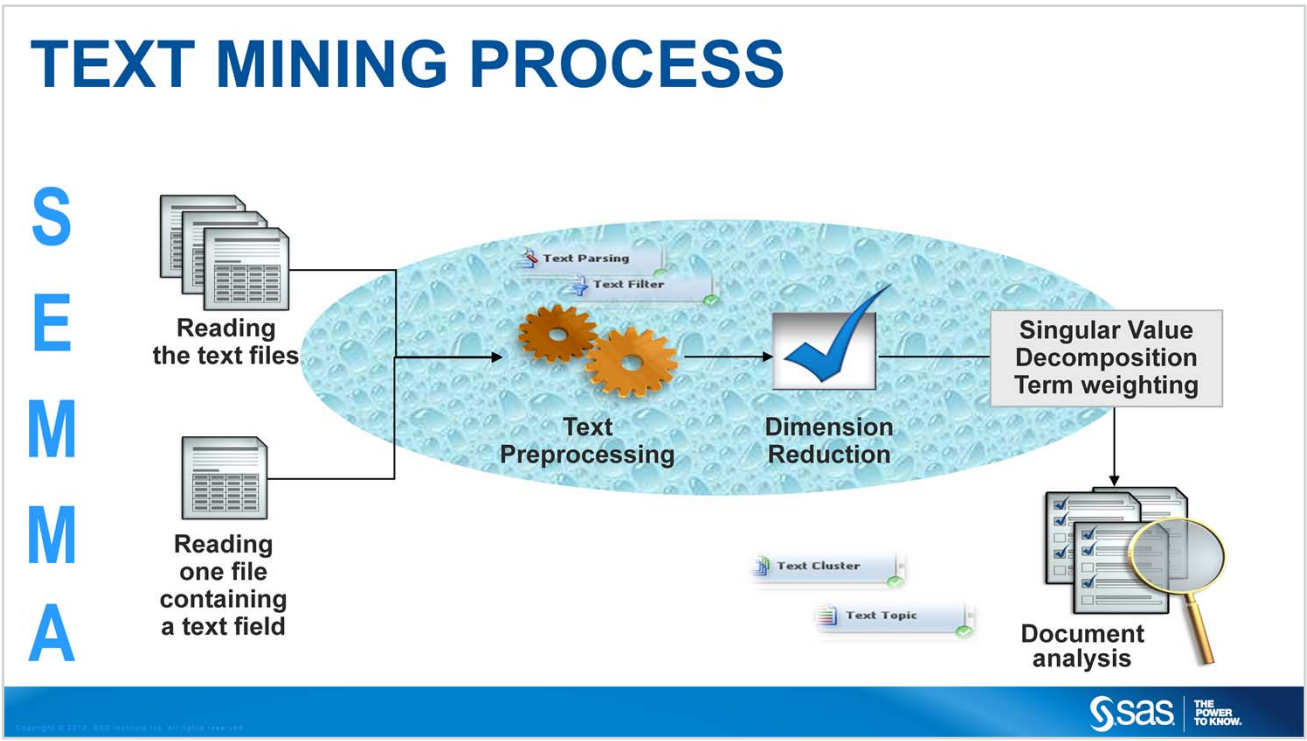
Positioning: Text Mining is *not* Information Retrieval. We focus on analyzing large document corpuses rather than individual documents.

"SEARCH" VERSUS "DISCOVER"



Search is goal oriented meaning you have to know what it is that you are looking for. Discovery focuses on capturing unknown information
The point is, a search can bring back volumes of information. Text mining helps discover unknown issues and patterns in that retrieved data.

TEXT MINING PROCESS



The process is one of reading the text, preprocessing to identify terms etc, transforming and reducing dimensionality so that the document is in an 'analysis ready' format. As with data mining the text can be sampled, explored, modified, modeled and assessed.

SAS® TEXT MINER CAPABILITIES

- Exploratory Analysis & Visualization
  - View term statistics, identify similar documents
  - Graphically display term relationships

- Automatic Classification
  - Taxonomies – e.g., Identify Medline abstracts by hierarchies
  - Discrete groupings – e.g., Identify clusters of call center comments
- Predictive Modeling

  - Text alone – e.g., predict patients conditions based on physicians comments
  - Text and structured data – e.g., predict future purchasing based product comments combined with demographics

Demo Time