

Understanding the Whole Picture

What techniques are available for analyzing information sources containing both structured and unstructured data?

eastport**analytics**



Restriction on Disclosure and Use of Data

The data in this document shall not be duplicated, used, or disclosed in whole or in part for any purpose other than to evaluate the reference concepts and techniques. The recipient shall have the right to duplicate, use, or disclose the data to the extent provided in any contract in-place between Eastport Analytics and the recipient. This restriction does not limit the recipient's right to use information contained within this document if it is obtainable from another source without restriction. The data subject to this restriction are contained on all sheets.

Today's information landscape is a complicated mix of structured data and text. Quantitative data, free text documents, forms, reports, web pages and a wide range of semi-structured internet sources, such as tweets, blogs, and discussion threads, can all hide valuable insights. Meanwhile, for a variety of business and government organizations, the need to assimilate and understand this information is greater than ever.

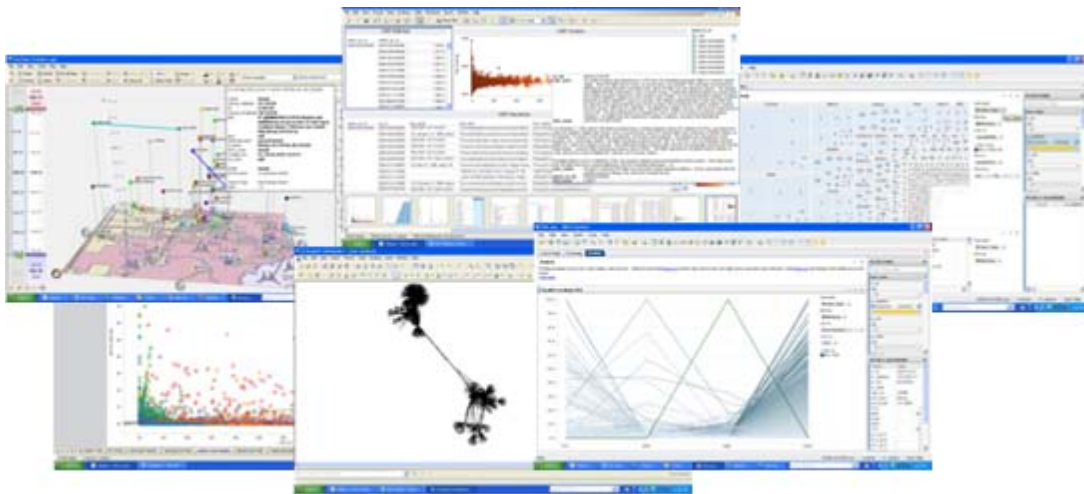
While there are individual analytical tools designed for either structured or unstructured data, using them in isolation against mixed sources can only provide half of the picture. Eastport Analytics solves this problem by selecting, combining and applying a variety of analytic capabilities to mixed data sources, and focusing on value added insights. Data visualization, search, statistical methods, and text mining are just some of the useful techniques that, when used in conjunction, can help surface answers to critical questions.

When facing a new, mixed datasets such as these, Eastport can call on a range of analytic methods. A typical approach would be to:

- gain an initial understanding of the structured data through basic data analyses
- generate value added metadata on the unstructured data (text) using Natural Language Processing (e.g., entity extraction) and statistical approaches (e.g. auto-tagging)
- explore the anomalies, patterns, trends and correlations across the full set of processed data using of variety of visualizations (e.g., charts and graphs, link diagrams, profiles (parallel coordinate plots), treemaps, geo-temporal visualizations, etc.) and other supporting techniques (e.g., search,

clustering, similarity detection, case based reasoning, etc.)

- communicate findings, along with supporting methods and reasoning, by capturing and conveying results in sequential, annotated, and repeatable analytic “stories”



To illustrate these techniques, Eastport recently analyzed FAA accident reports in an attempt to discern the factors that influence aircraft accident occurrences and severity. The FAA’s Accident Database is typically complex in that it contains both highly structured fields (dates, locations, inspection hours, injury counts, etc.) and textual data (pilot narratives, investigative findings, etc.) In the course of this analysis, Eastport applied several commercial tools and custom built processes.

Our findings included:

- the importance of pilot experience in both avoiding accidents and minimizing accident severity, especially in the specific model of aircraft being flown;
- some models of aircraft typically sustain more damage, others more injuries, while a few suffer both;
- surprisingly, most aircraft involved in accidents had recent inspections;
- notable patterns of similarity exist between clusters of accidents as described by their narratives, such as the involvement of student pilots, power lines, or oil starvation;
- ...and many more

These kinds of insights are only possible through the analysis the full data landscape, given that it is the full data that completely describes any given accident. Eastport Analytics believes that by understanding the required insights, then agnostically applying the right tools and techniques given the specific nature of the underlying information, can help organizations discover and communicate the most complicated facets and hidden factors in their data.