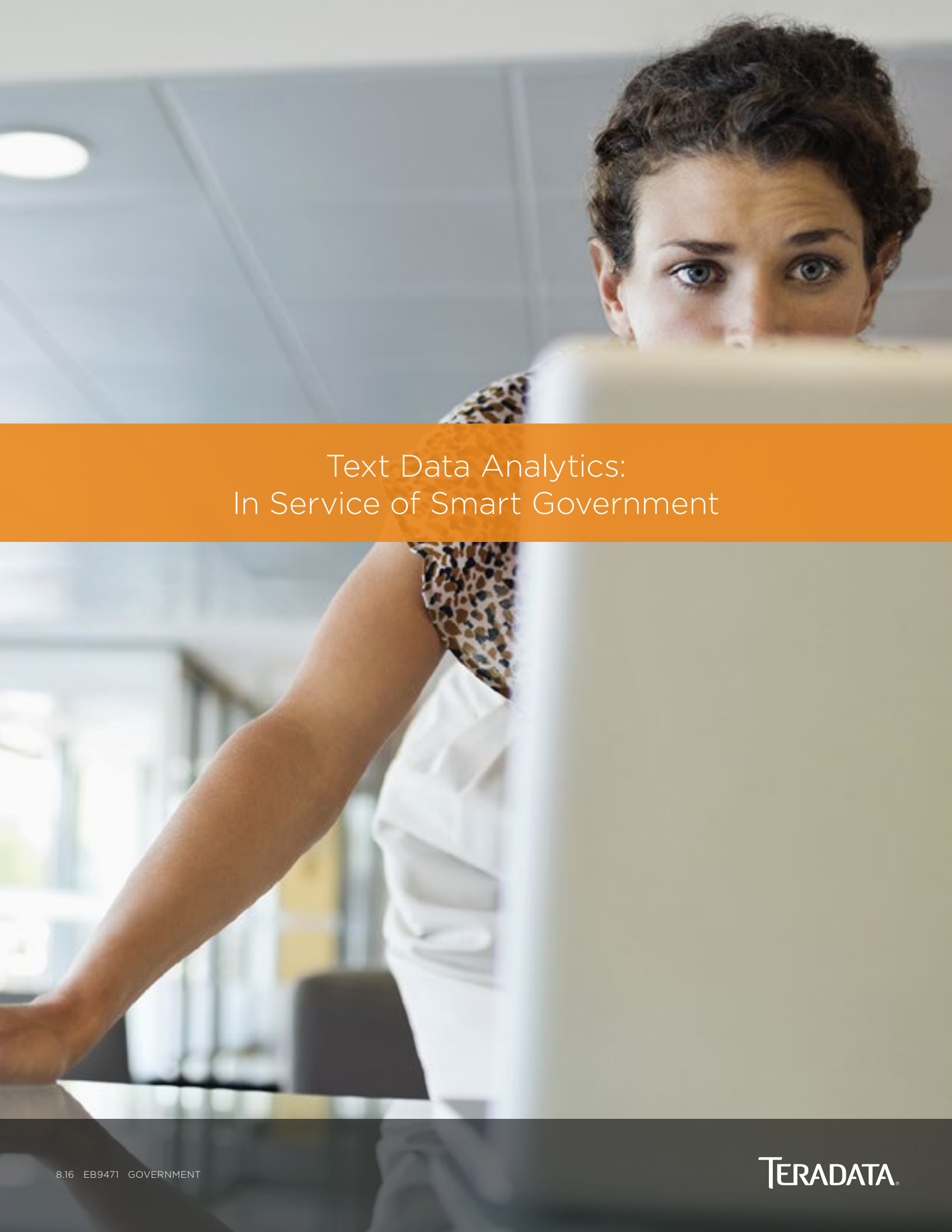

Text Data Analytics: In Service of Smart Government

8.16 EB9471 GOVERNMENT

TERADATA.

Table of Contents

Big data includes documents and text data of all sorts, and today's data science can help take you beyond reports and charts—allowing you to head off threats before they hatch.

Executive Summary

Of the many challenges facing government, the most compelling may be the protection of its constituents from continuously changing threats—from border control and war fighting to cyber security, and from an increasingly global, connected society that's steadily increasing the scope and impact of "bad actors". Smart government can use analytics to search across mounds of documents and deliver crisp new insights into predicting where investigators should look next.

Meeting Government Challenges in a Fast-changing World

Like most organizations, government often needs to do more to meet the rising challenge of security threats with limited or shrinking resources. And even before current missions are accomplished, new priorities, emerging threats, and a rapidly-changing landscape are demanding attention. What would you do if you could head off emerging threats—before they became fully developed problems?

In the not-so-distant past, the use of data was often limited to a report card of performance. Today, in the era of big data, private industry is catching on—leveraging data to channel its efforts and influence its customers, and signaling a crucial opportunity for government to step up its game. And one way government can achieve this is by starting to look deeper into its treasure trove of documents.

Big data includes documents and text data of all sorts, and today's data science can help take you beyond reports and charts—allowing you to head off threats before they hatch.

What Would You Do...

- if you could use psycholinguistics to understand the intent of bad actors, or those being influenced to become insider threats?
- if you could use text data to flag potential illegal activities and secure our borders?
- if you could focus law enforcement and intelligence resources on the highest value documents, sifted by predictive analysis of all types of documents, including text, criminal records, licenses, conversations, and emails?

Why Smart Government Should Mine Big Data

When security matters, failure is not an option. And when more resources are not an option, smarter approaches are needed. Government needs to anticipate and react quicker to emerging needs, and across a wider spectrum of conditions, more than ever.

"Most crime is not random, but rather part of a complicated pattern of events. For crime investigation, it has typically been about unraveling and connecting the dots after the crime is committed. Historically, investigators took notes, interviewed witnesses, and gathered forensic data at the scene in the hope of piecing together a picture of the perpetrators."

See Reference

Enter the world of predictive policing: an evolution from dependence on investigative hunches to a modern-day capability that brings years of crime fighting experience and information where it's needed most—at the front lines. The uses of big data in law enforcement comprise a rapidly evolving science that aims to anticipate areas of potential crime by leveraging historical law enforcement data.

To gain a greater advantage, it's not enough to connect historical dots—you need to be able to anticipate the perpetrator's next move. "Criminologists use the term 'near repeat phenomenon' to describe criminal patterns of behavior." These phenomena can provide police a powerful predictive technique, if law enforcement has the analytic capacity to identify patterns and capitalize on them.

See Reference

Big data analytics examines large data sets to unearth hidden patterns, trends, preferences, unanticipated correlations, and other useful actionable knowledge. It differs from traditional analytics because it taps into dramatically different data sets. Known as behavior data sets, they are useful for tapping into explanatory variables and critical linkages that are ultimately connected with illegal activities.

Patterns identified from the behavior data of bad actors or illegal activities in the past can be used to create predictive models, which are used to identify likely future actions. Combining history with current data helps repurpose forensic efforts of the past into a powerful resource for law enforcement professionals.

Detecting Behaviors from Documents and Text Data of all Types

One of the richest forms of behavior data is text data. Text is a special form of behavior data because, like most big data, it provides rich context around activities or transactions of interest, and is readily available in machine-readable forms such as emails, surveys, social media, and user-generated content.

However unlike most other big data, use of language has been around since the dawn of civilization—and the patterns in word choice have become rich and highly evolved. While this can make text data complex to analyze, it can vastly increase the potential value of delving into it. If you could detect changing behaviors (e.g., suspicious documentation of cross-border travel or shipping), language usage patterns (e.g., language usage connected with illegal activity), or conversational linkages (e.g., communications across/within suspicious social media networks), it would leverage effectiveness of government resources.

Government: Staying Ahead of the Game

What makes data science valuable to law enforcement efforts is predictability of human behavior, especially with the use of language. To illustrate, let's look at an example of how analytics can help government fight terrorism by countering use of social media by radicals. Use of social media by terrorist groups is very sophisticated and, when used for radical propaganda and terrorist recruitment, its broad reach can be difficult to combat without ways to isolate highest value targets.

The number of potential accounts to analyze can quickly become overwhelming. Just three layers into friends/followers of a single individual can result in over one million users—far too many to analyze manually. And the account turnover is significant. Even if you could identify bad actors today, the accounts could shortly be under different names.

By combining multiple types of analytics, we can identify content of interest, analyze the social networks of members and their supporters, and recognize them when they re-appear under different names. A five-step approach reduces a large social media dataset of interest to a

Use Cases

Understanding meaning, intent, and behavior from documents and text data may be helpful in anticipating and heading off illegal activity.

Content Analytics: Understand intent to collaborate on illegal activities

- Flag conversations that have intent to bypass regulations (e.g., illegal immigration) or content designed to escape notice (e.g., falsified cargo shipping)
- Spot increasing companionship and trust between individuals and groups

Recruiting and Trafficking: Understand intent to recruit for terrorist, gang, drug, or human trafficking activities

- Develop psychological "look-a-like" models of bad actors and recruiters
- Filter social media to identify previously unknown bad actors and recruiters
- Differentiate ringleaders from recruits

Financial Forensics: Identify patterns such as those associated with fraud, money-laundering, and bribery

- Develop psychological "look-a-like" models of people who commit fraud (e.g., Medicare) or participate in illegal activities (e.g., drug abuse)
- Identify deceptive language, claims, patterns

Aster Analytics Example

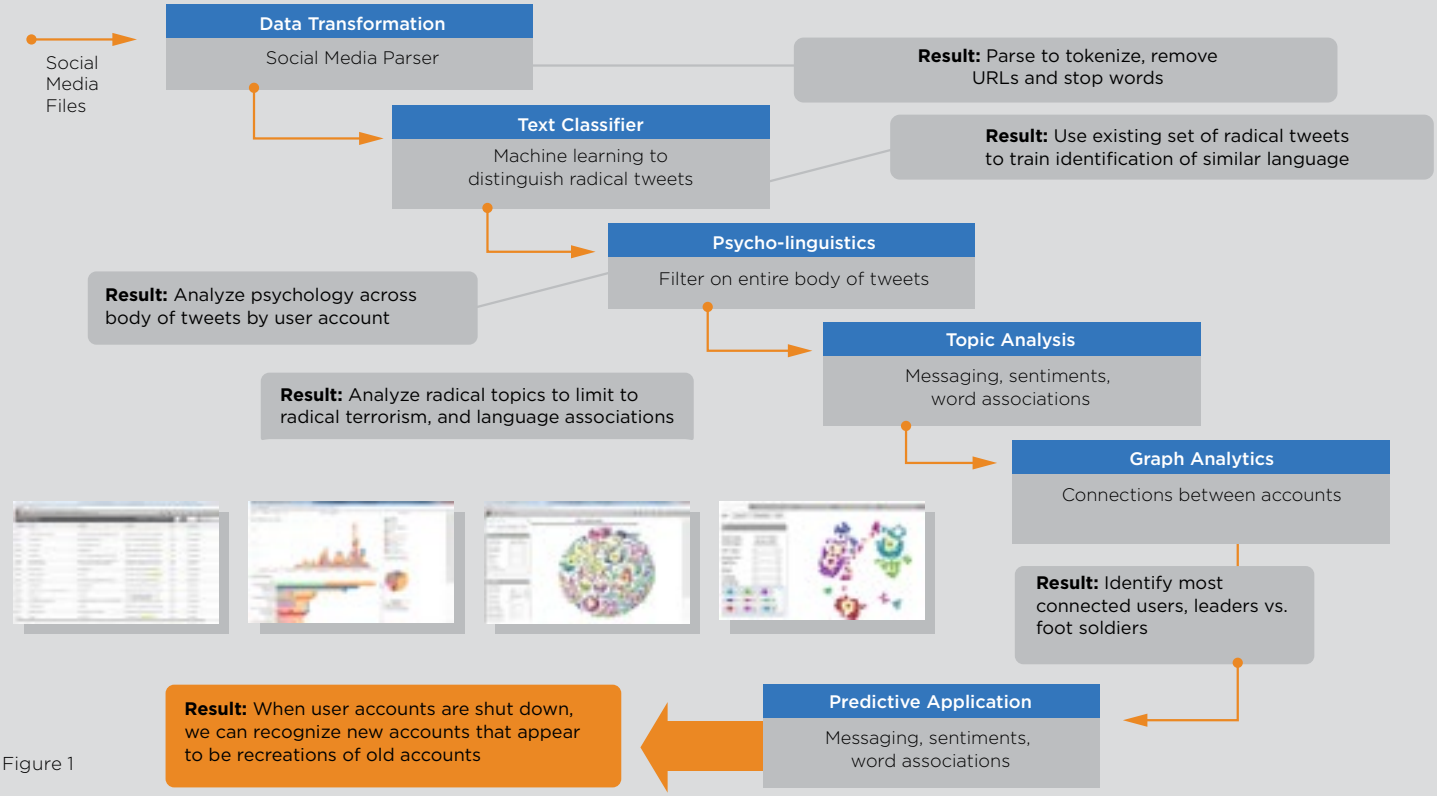

Figure 1

Communications Compliance: Activities non-compliant or deceptive in nature

- Develop insight on truthfulness of people or documents (e.g., testimonies, interrogations, and transcripts)
- Identify deceptive language

Cyber Security: Identify incidences of cyber security intrusions

- Spot spear phishing emails that are intended to get the user to disclose personal information, such as login information and SSNs
- Scan communications for sensitive information before sending out of the network

Information Risk Management: Identify problematic situations that create risk or exposure

- Spot individuals who exhibit markers of depression, anxiety, PTSD, and other psychological states
- Determine the best approach for responding to each individual

Insider Threats: Identify insider threats in your organization

- Spot anomalous user behavior by analyzing communication patterns

TERADATA.

manageable level, and with a sufficient degree of accuracy so that manual intervention can be effective—and from there the next step of predictive interventions can evolve (see Figure 1).

Text data first needs to be transformed to reduce the variability of written and verbal speech (e.g., different spellings, errors, and abbreviations)

Putting Text Data to Work

What makes it possible to draw compelling conclusions from examples such as the extremist social media recruiting, in spite of massive amounts of data? To grasp the power of big data analytics, it's important to understand more about the different kinds of analytics that apply in cases of big data that involve text (see Figure 2).

Big data analytics is about connecting the dots (datasets); and the challenge to connecting them comes from the differing nature of what we need to extract from each piece of data. For example, classifying emails as "radical" is statistical modeling, which is very different from knowing the patterns of who is central to communications among people with radical interests. And that is further differentiated from association analysis, which identifies topics shared by certain communicators.

Enabling this sequence of analytics across different data attributes is what Teradata calls "multi-genre analytics". Multi-genre analytics enable analysts to use a single tool with optimized techniques for each stage of the problem. It enables them to pivot from one analytic capability to another as the use case evolves. Multi-genre is extremely helpful with text data because of the four basic capabilities involved in deriving its value:

Data Ingestion

Text data first needs to be transformed to reduce the variability of written and verbal speech (e.g., different spellings, errors, and abbreviations). "Stemming" reduces variations of the same word, and "stop word elimination" drops words that convey little meaning. More complex transformations extract named entities (individuals or accounts) or facilitate identity matching where names are used differently in different source systems. This needs to be done at scale, and be adaptable to different formats of documents.

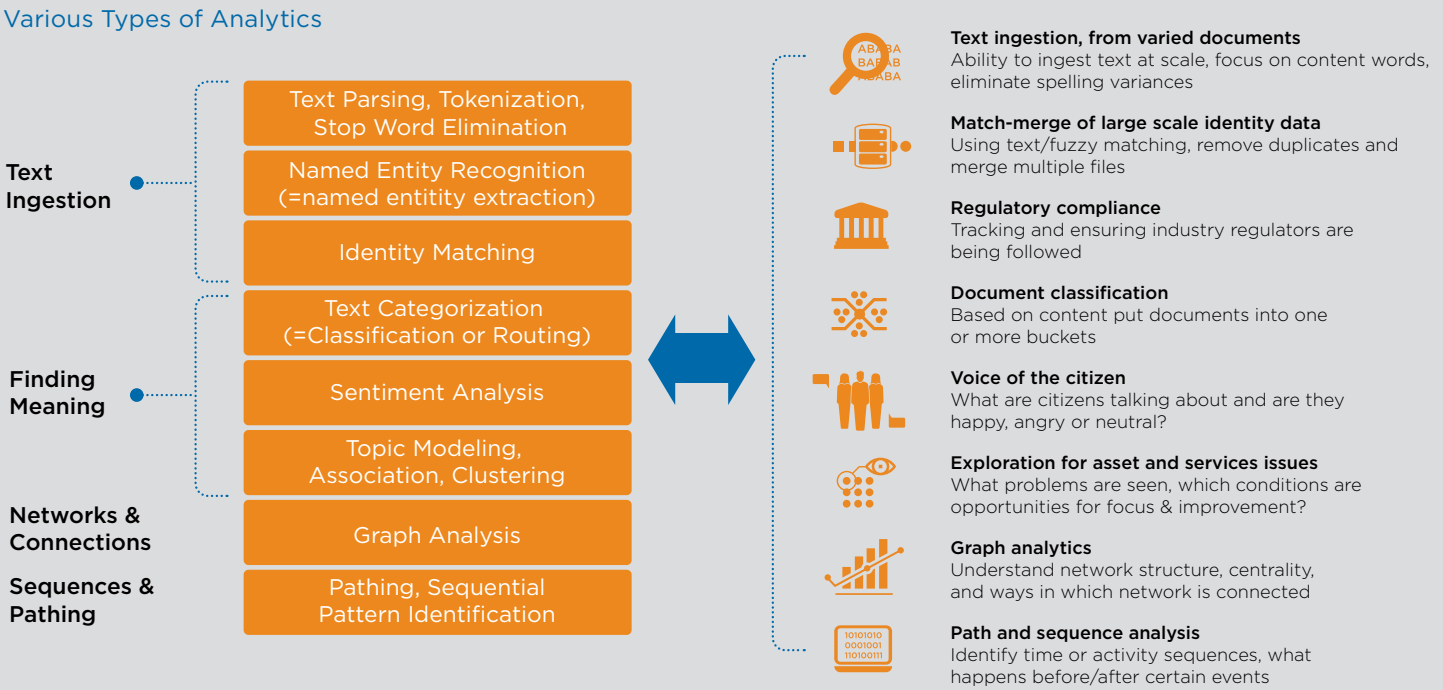

Figure 2: Different kinds of analytics apply in cases of big data that involve text

Finding Meaning

Analytics can establish meaning from the way the words are used. Sentiment extraction (positive, negative or neutral), classification of content, and topic analysis can automatically analyze every message, chat, email or post, and flag those that fit a pattern. Extraction of meaning is enabled by natural language processing (NLP), artificial intelligence, and computational linguistics. These techniques can cluster (or be trained to identify) suspicious, informative, specific topics, or potential violations of homeland security laws and other government regulations.

Establishing Networks and Connections

Today's citizens and accounts are connected and influenced in remarkable ways. Understanding these connections is vital to being able to cull out those of greatest interest and find ways to leverage the text data. Multiple types of graph analytics can identify the most central individual, the connectedness of the network, the relative importance of certain individuals, and extent to which accounts collaborate.

Recognizing Sequences and Pathing

Sometimes sequences or combinations of factors are of interest. Pathing analytics can identify time sequences, activity sequences, or the boundary conditions in certain kinds of events. Pathing analytics were particularly difficult queries to code before big data tools became available—you had to know what you were looking for before you could write code. Now the data can speak to the analyst with very little effort needed to predefine the query, allowing a broader net to be cast around potential illegal activities.

Take the Analytics Challenge

Analytics goes beyond reports and summaries, providing answers to the next question. While reports recap "what happened", analytics identify characteristics that are common among them—providing critical answers to the next question, "what are the common profiles?"

While investigators working after-the-fact are limited by resources, profiles built from past investigative work can be applied proactively to scan new data coming in every day. By proactively applying profiles to trap illegal activities that have not yet occurred, government has a powerful scalable weapon. Profiles tackle the "near repeat phenomena" that is inherent to criminal activity, and more.

While it may be impossible to stop illegal activity altogether, proactive profiling can get you closer. Illegal activity is not likely to stand still. And while illegal activity stays in motion as long as it's not stopped, as soon as one angle is shut down, variants emerge shortly thereafter. With proactive profiles, not only do you have a tool that keeps working forever, but one that evolves into a portfolio of profiles over time to keep up with the changing nature of criminal activity.

Because of the constantly changing nature of illegal activity, using data to stop and hinder it will always be an analytics problem as opposed to an application problem. While an application approach has a fixed set of data, and a specific outcome, analytics is always exploring the next set of questions to emerge. This is what Teradata calls a data-driven approach.

Ask us about the Teradata Architecture for Smart Government

A data-driven approach to smart government isn't about adding a little data—or even a lot—to your current activity. It's about opening a window into your current data and letting it speak, which requires a purpose-built approach. Teradata provides the flexibility and scalability of a data-driven approach through the following capabilities:

Architected for Analytics

With scalable parallel processing, Teradata enables the analysis of full sets of data, avoiding problems and limitations that come from data sampling or limiting the attributes. We enable you to ask multiple questions of the data, so you can narrow your search to questions you really want answered.

All Your Data

With unified data architecture, Teradata accesses all your data regardless of where it resides; in a data warehouse, data mart, or a big data platform. Our approach connects all your descriptive, structured data—the foundation of your enterprise—with the complete spectrum of behavior data that is the foundation of predictive analytics.

Discovery Analytics

With Teradata discovery tools, we can help you navigate your constantly changing data sources. When you're answering the next question, the nature of analytics tools required will always be changing. Our multi-genre

discovery analytics tools let you explore new data you've never seen before, and allow you to select any combination of approximately 200 techniques to open up your text data—as well as other unstructured data types, including web logs and voice recordings—to gain even more insights.

By leveraging discovery analytics, data science is embedded into your investigative approach—allowing you to reduce mountains of text into a few documents and individuals that matter most.

With Teradata big data and analytics solutions you can collect, unify, and analyze all your data—including text and documents—to identify potential or emerging threats to public safety. By leveraging discovery analytics, data science is embedded into your investigative approach—allowing you to reduce mountains of text into a few documents and individuals that matter most. The goal is to triage all available document data down to a workable, manageable size for analysts.

Conclusion

Big data helps you go beyond traditional reporting to provide strategic insights into what is happening—and what will happen. This approach has applications in tracking everything from extremism to any environment where a large number of individuals may be involved and/or collaborating on illegal activities (e.g., border patrol, drug traffickers, human trafficking, and pedophile rings).

To learn more about Teradata Government Systems and text data analytics solutions for smart government, visit teradata.com/government.

Reference: Predictive Analytics within Law Enforcement: Accelerating the Pace with Big Data

About the Author

Peeter Kivestu—Industry Consultant, Teradata Government Systems

Based on private industry experience spanning finance, marketing, operations, and technology, Peeter focuses on business transformation through data. Today, he's an avid proponent of the Teradata industry approach to analytics—supporting government clients by finding new ways to put data to work for solving the toughest business challenges, and mitigate future ones.

10000 Innovation Drive, Dayton, OH 45342 **Teradata.com**

8.16 EB9471