Moving From Dark Data to Clear Insights: A Framework

by Rajesh Narayan

Abstract

This article examines how data that is not easily accessed by traditional means, such as documents, notes, and other internal and external sources, can be leveraged to generate valuable insights. Such data, which flows naturally within an enterprise but is not fully accessible to systems because of the way the information is expressed, is referred to as "dark data." This article aims to help readers identify patterns of dark data sources. It also provides a framework for leveraging artificial intelligence and analytics to generate insights where such patterns are found.

In his landmark book *Blink*, Malcolm Gladwell highlights how thin-slicing data to make decisions can yield effective results. This approach tends to be more effective than poring over large quantities of analysis and facts, which can bias decisions and result in "overfitting" the data.

This strikes me as clearly the case with insurance, as the facts presented in an application are only one aspect of decision making needed to analyze the risk exposure. The more time underwriters invest in risk analysis, the more likely they are to adopt the insured as a client, even if doing so is not in the best interest of the organization. This is human nature—and the reason that decisions made in a "blink" can be better than those made after extensive analysis.

The insurance industry can apply this knowledge by thin slicing up front. This is now possible because data that was previously underused can be extracted with new technology. Such data,

Continued on page 12

Continued from page 11

typically called "dark data," is collected during normal business operations for specific business activity but not leveraged for additional insights using analytics. The sources of data in typical insurance operations include inspection reports, emails, and external data and reports.

Once the context within these sources is understood, we can retrieve signals that uncover additional exposure details, the intent and behavior of the broker or insured, and more.

Thin slicing can help industry professionals identify signals that indicate the brokers' intention to bind, the true scope of exposure, and the nature of business beyond what the application tells you. These signals lead to better risk selection and allow underwriters to focus their time on risks that warrant further assessment.

Think of how gmail is able to find behavior signals and correlations to sort inboxes into promotional and primary material. Imagine if underwriters could detect in advance which submissions are truly ready to bind and best-suited to their market appetite through an artificial intelligence (AI) intervention. The innocuous email with attachments that brokers send provides enough information to thin slice. By taking advantage of advances in technology, the industry can build smarter underwriting practices for commercial insurance.

The Data Challenge

Most organizations have excessive amounts of data—to the point that they are nearly drowning in it.

"Thin slicing can help industry professionals identify signals that indicate the brokers' intention to bind, the true scope of exposure, and the nature of business beyond what the application tells you"

Organizations and individuals generate and store data at phenomenal and ever-increasing rates. The volume and speed of data collection have been associated with new technologies and roles focused on managing, organizing, and storing data. The challenge, then, is how to thin slice from this abundance of information and glean the necessary insights?

The steps below show how an example of dark data (i.e., error codes generated by software) can be converted to insights:

1. Signals exist in system logs (serving as an example of dark data) — Collect all the error logs from the corresponding system.

2. Convert to structured data—Build an organized report about error codes that are interspersed in the logs with the comments that typically occur when those codes are found.

3. Build a knowledge base—Build a tool to document a set of codes and comments that can help identify the underlying issue. Broader patterns of problems that stem from hardware capacity, connectivity, user error, etc., are more easily identified when multiple issues are combined.

4. Generate insights—An understanding of how different and seemingly unconnected issues are related symptoms of a larger, underlying problem. Once these patterns emerge, we could reach a point of predicting system failures before they happen and proactively correct such issues.

From an underwriting perspective, the work process and necessary insights do not usually follow a similar, direct path; instead, the risk assessment process requires a holistic evaluation of the exposure across various perspectives. Some perspectives that underwriters need to keep in mind, in addition to the specific account or location characteristics of risk, are overall exposure aggregation and accumulation across the book, the nature of business and emerging challenges, and catastrophe-related modeling for large accounts. In fact, as knowledge about the risk increases, the type and nature of data collection become clearer.

Insight is gleaned from all directions. Knowledge inputs include inspection reports, loss runs, schedule of values, business credit reports, catastrophe models, and exposure aggregates, in addition to the initial submission.

Insights about the business value of an account could override other decisions, however. Here are some examples:

- The account is part of a desirable larger group or company.

- Upselling opportunities are present in the submission, making it desirable.
- The quote is a favor to a broker, to establish a long-term relationship.

The Evolution of Technology

Most relevant technologies used to solve the data challenge come from attempts at developing effective web search engines. Countless individual humans have used their own knowledge to put content (information) onto the web. Using graph databases to turn the web into data about connections and natural language processing to turn information into data about content made the web navigable. Humans are now able to find and use information that was previously inaccessible. And this core technology is widely available in applications for industries such as insurance.

To illustrate further, the basic unit of a knowledge graph is (the representation of) a singular entity, such as a submission that is under review, a policy that is coming up for renewal, or anything you would like to describe. Each entity has its own attributes. For example, the attributes of a submission include broker name, producer name, broker office, received date-time stamp, customer name, and customer's nature of business, among others.

Furthermore, entities are connected to each other by relations, so producers might specialize in one type of risk or line of business. Relations can be used to bridge two separate knowledge graphs: for example, a producer who specializes in a particular area can effectively distinguish him or herself. This is more easily done if

"Most relevant technologies used to solve the data challenge come from attempts at developing effective web search engines"

the producer, line of business, and submission have unique IDs that establish the relation.

Entities and relations are defined in special dictionaries called ontologies. The standard ontology language is called OWL (Web Ontology Language), which helps underwriters understand and give context to data.

Combining to Build Analytics From Dark Data

Much of the current excitement in AI and data science is about generating information and modeling knowledge. Classification tools—which really allow machine learning to excel—create new information. Natural language processing tools can summarize text and extract information expressed within a sentence. Multidimensional vector analysis can categorize inputs based on a near-limitless number of factors. Computer vision technology, combined with trained neural networks, can identify faces and street signs.

In a sense, knowledge of how a particular emotion manifests in text is contained in the model or neural network that is able to produce that information. More advanced AI uses data to simulate procedural knowledge. This could be relatively simple, like finding an efficient driving route. It could also be unimaginably complex, like assembling a self-correcting robot. In either case, knowledge is a process enabled by data.

The skills required by this process are typically dependent on learning from a large number of datasets. To identify the street sign, the machine learning needs to have seen tens of thousands, if not millions, of related pictures before it develops a level of accuracy. This amount of data may not be available through an insurer as it relates to a new peril or exposure (like flood or cyber), serving as an impediment to leveraging deep learning algorithms.

Another form of language-recognition AI mimics how children learn new words, by building on vocabulary through continuous feedback. This kind of technology as applied to reading the context and content within paragraphs and sentences is called computational linguistics. It takes the value of AI beyond the classification or procedural knowledge available with neural networks and deep learning, moving it into the realm of attainable experiential insight, where one not only knows what a thing is, but also why and in what context it should be used.

Linguistics allows a person to be reminded of things: so someone recognizes that "installed a firewall" is like "installed a sprinkler," which was previously learned, and can articulate why. This type of learning simulates how underwriting experts dive into their memories, retrieving past experiences that are similar to present matters, to match patterns and connect the dots.

Continued on page 14

Continued from page 13

Accordingly, some AI platforms and similar technologies extend, rather than replace, human capabilities. The machines are not wise and will not replace key resources, but they help make humans wiser and supplement humans' efforts. An underwriting case illustrates this point.

Leveraging Dark Data for Submission Selection

In a typical day in the life of a commercial lines underwriter, emails arrive with submission data embedded and attached. Typically, submissions have to be read by underwriters or their assistants to gauge broker of record (clearance), sufficiency of attributes for risk selection and further processing, and sanctions and other checks.

But consider this: If computational linguistics reads through the body of the email and the attachments to glean the necessary data for clearance, underwriters will be able to leverage a consistent process to digitize the data for automated sanctions checks and more, generating additional patterns of signals. For example:

- Does the submission from the broker provide data sufficiency, and if so, how complete is it?
  - –Can we supplement the missing data from external sources instead of requesting more information from the producer?
  - –Does the behavior of the producer who submitted such information indicate a higher likelihood to bind by way of comparing past behaviors related to similar submissions (evidence of producer specialization)?

> "The machines are not wise and will not replace key resources, but they help make humans wiser and supplement humans' efforts"

- If the producer sends out a submission with an indicated premium and it falls within a defined range that has yielded a better bind result based on past results, does this mean better chances to bind for this particular submission? (Could we find out the exposure aggregation for this risk and consider prioritizing this over other submissions if the chances to bind are higher?)

By combining the ability to read dark data with analytics, underwriters can gain insight and increase their hit ratios. By converting signals to insights, underwriters also gain a true competitive differentiator for the insurer, which can save on operational costs, while being able to respond to high-value submissions much more rapidly.

Humans will continue to build, operate, and understand the tools that turn dark data into information, information into knowledge, and knowledge into insights. However, the intersection of computer insights and human wisdom can enhance a variety of roles in the insurance industry, from executives to data analysts, based on an understanding of the fundamental tools of data science and the value that data-driven wisdom can create. ■

Many thanks to the Information Technology Interest Group for its contributions to this article.

Resources

1. Jeff Z. Pan, Guido Vetere, Jose Manuel Gomez-Perez, and Honghan Wu, editors, *Exploiting Linked Data and Knowledge Graphs in Large Organisations* (Switzerland: Springer International Publishing, 2017).
2. Mike Barlow, *Learning to Love Data Science* (Sebastopol, California: O'Reilly Media, 2015).

Rajesh Narayan, CPCU, LSSBB, serves as managed services business lead for Risk Management Solutions, where he explores the use of emerging science and technology to increase the reach of insurance. Specifically, he leads services for carriers to leverage the new science and methods of analyzing exposure at the micro level (for underwriting) and the macro level (for