“Out-of-the-Box” Scans for Sensitive Data: Easy Solution to a Difficult Problem?

Maureen O’Neill:
SVP, Discovery Strategy and Data Privacy & Security, DiscoverReady
maureen.oneill@discoverready.com
Phil Richards:
Chief Technology Officer, DiscoverReady
phil.richards@discoverready.com
Eric Willis:
Director, Product Development, DiscoverReady
eric.willis@discoverready.com

The Sensitive Data Problem

In the context of handling litigation, regulatory inquiries, internal investigations, compliance programs, and other legal matters, organizations routinely collect vast stores of electronic information. A critical challenge faced by companies is the identification of “sensitive data” found within these collections.

Sensitive data takes many forms. It includes personally identifying information (“PII”), such as social security number, birthdate, and address; financial and payment information, such as credit card numbers and account numbers; protected health information (“PHI”); intellectual property and trade secrets; source code; attorney-client privileged content; and a wide range of other proprietary and confidential business information. Once identified, sensitive data can be protected as appropriate for the matter at hand—by masking/redacting, tokenizing, defensibly deleting, withholding from production in legal discovery, or producing subject to certain controls and restrictions.

Why must organizations identify and protect sensitive data? First, various laws and regulations mandate that companies take reasonable steps to secure and prevent unauthorized disclosure of PII, PHI, and other specified types of sensitive data. These laws include U.S. federal and state statutes and regulations, laws of international jurisdictions, court rules and procedures, and ethics rules applicable to attorneys. Second, companies possess sensitive information that, although not legally protected, provides enough value to the organization that it should be tightly controlled. Examples include trade secrets, product formulas, source code, and proprietary processes. In this context, “sensitive” is in the eye of the beholder—what’s important for one company to protect may be wholly uninteresting for another.

Companies must protect sensitive information in the ordinary course of business, while it resides within the corporate environment. Whether resident in systems of record, or found outside those systems as data “exhaust,” sensitive data behind the corporate firewall must be identified and properly managed.[1] The enterprise also must take steps to secure sensitive information when it leaves the organization in connection with legal matters, and into the possession of outside counsel, consultants and experts, opposing parties, government agencies, and courts. When sensitive information is compromised, the consequences can be dire. The company may face legal liability, to government regulators and to aggrieved victims. And as breaches now routinely land in the media headlines—think Yahoo!, LinkedIn, Sony, Wendy's—the organization's reputation with customers and business partners may suffer. Fines, money damages, and legal fees can mount into the millions. And if a company's valuable trade secret gets turned over to an adversary, there may be no way to repair that damage.

So, why do we characterize sensitive data as a "problem" for organizations? Because finding sensitive data—at least with a reasonable degree of accuracy and completeness—can be quite difficult.

Current Approach to the Sensitive Data Problem

In our experience, the most effective solution for finding sensitive data relies on a combination of technology, analytics, and consulting services. This strategy leverages a multi-pronged approach that includes custom-built scripts for processing, indexing, and organizing large data collections; Boolean searches; data scan tools; human review of data samples; iterative rounds of statistical sampling, measurement, and validation; and different types of concept analytics.

This approach is a methodical, sometimes resource-intensive process. It requires collaborative consultation with each organization, and often for each new matter. Some data sets require multiple rounds of iteration to achieve satisfactory results. No question, this type of technology-enabled service is not a Staples® "easy button" for finding sensitive data.

Do Commercially Available Software Scan Tools Offer an Easier Solution?

Over the last several years, various commercially available "scans" for sensitive data have come to market. These scans—essentially search software that runs against stores of data—purport to find sensitive data in both structured and unstructured formats. They claim to identify common types of sensitive data such as PII, credit card numbers, and source code. To date, however, none of these tools has emerged as an industry standard, and there are no commonly accepted best practices for using the tools.

We examined a number of these scan tools, and conducted some preliminary testing of their effectiveness at finding sensitive data. Our goal was to assess: Do these tools offer a better way

[1] Sensitive data "exhaust" is the (undesirable) by-product of routine business practices, in which small amounts of sensitive data leak into streams of communication, user files, spreadsheets, reports, and other data stores where it doesn't belong. The by-product then spreads, finding its way into yet other data sources.

to identify sensitive data? And by “better,” are they faster and more efficient than our current recommended approach? Are they reasonably effective at finding the various types of sensitive data in a collection, without bringing back an unreasonable number of false positives?

Based on our investigation, the answer to each of these questions is “no.” While scan tools can be an important component of an overall assessment process, when used “out-of-the-box”—without any modification, fine-tuning, or supplemental searches—none of the tools we examined provide a satisfactory replacement for a technology-enabled, human-driven process for finding sensitive data. There is still no easy solution for this difficult problem.

A Quick Statistical Primer

For those readers familiar with basic statistical sampling terminology, please skip ahead to the next section. But for those who need some background, we define a few terms that will be used in the discussion:

> ***Richness (or Prevalence)***: This refers to the percentage of documents or files in the collection being searched that contain sensitive data. For example, if statistical sampling shows that approximately 30% of documents in a collection contain sensitive data, we would say that the collection has an estimated richness (or prevalence) of 30%.
>
> ***Recall and Precision:*** When testing the efficacy of a search and retrieval process such as a sensitive data scan tool, we use statistical sampling to generate two important measurements—recall and precision.

- *Recall* is the fraction of sensitive data in the collection identified by the search; recall measures the completeness of the search. For example, if a scan hits on 80% of the documents containing sensitive data in the collection (missing 20%), we say that the search has 80% recall. When recall is low, the results of a scan may give the user a false sense of security—believing that most sensitive data has been found, when in fact it has not. Recall correlates inversely with risk—the lower the recall, the higher the risk created by the scan’s failure to find sensitive data.

- *Precision* is the fraction of files identified by the search as containing sensitive data that are in fact sensitive; precision measures the accuracy of the search. For example, if a scan tool identifies 10,000 documents as potentially containing sensitive data, but only 9,000 of the documents actually do contain sensitive data (1,000 of the documents are false positives), we say that the tool’s results have 90% precision. Precision correlates inversely with inefficiency and cost—the lower the precision, the more time and resources are wasted on “wild goose chases” caused by the scan’s false positive hits.

Commercially Available Sensitive Data Scans: Some Preliminary Findings

Based on our initial efforts to examine some widely used scan tools, we offer the following preliminary findings.

1. *Sensitive Data Exists Almost Everywhere*.

In our experience, sensitive information exists in virtually every collection of data. It's found in expected locations, like organized, well-managed databases; but it's also found in many unexpected places, like individual e-mail accounts, personal folders on employees' computers, and freely shared network folders. For example, even when a company states with confidence that "you won't find source code" in a particular custodian's e-mail collection, we often find source code there (most likely a result of data exhaust).

2. *The Prevalence of Sensitive Data Typically is Very Low.*

Even though it's found almost everywhere, sensitive data in most collections typically exists in very small quantities. In other words, the richness/prevalence of sensitive data tends to be low—generally less than 5%, depending on the type of sensitive data element.[2] The lower the richness of the information being searched for, the harder it becomes to find—low richness presents the classic "needle in a haystack" problem.

A high-quality scan for sensitive data should maximize both precision and recall–which often is challenging, as the concepts intrinsically are in tension with each other. Typically, the higher the recall, the lower the precision, and vice versa. As a practical matter, the low richness of sensitive data means that scans for these data must trade off low precision for high recall; to find all the needles in a haystack, the scan must gather a lot of hay. However, an organization's risk profile can determine how precisely a scan must target its efforts. In a more risk-tolerant scenario, a scan could leave behind some sensitive data (allowing for lower recall) and potentially improve the precision of the results. Conversely, the more critical it becomes to find all sensitive data and boost recall as high as possible, the more likely the scan will achieve low precision.

3. *Out-of-the-Box Scans Suffer from Both Poor Recall and Precision.*

Unfortunately, the tools we examined performed poorly on both precision and recall. Even where large volumes of false positives were dragged in by the scan, the scan still left behind significant quantities of sensitive data. So, if the scans were designed to cast a broad net and

[2] We attribute the circumstance of sensitive data existing everywhere, but in small quantities, to the problem of sensitive data exhaust discussed above. However, once an organization finds rogue sensitive data, it can trace the data lineage back to the source of the exhaust, and then work to improve information governance practices to eliminate or reduce the exhaust.

bring back most sensitive data at the expense of low precision, their nets are poorly aimed—the recall of the scans typically was unacceptably low.

4. *Pay Attention to the Index.*

No matter how well a scan might potentially perform, if the collection of data being scanned has not been indexed appropriately, the scan will fail to find certain sensitive data. For example, some types of sensitive data contain punctuation and special characters (SSN and source code, for example). If the index settings for the data collection treat these characters as "noise" and prevent them from being searched, the scan can't find them.

5. *Out-of-the-Box Scans, Standing Alone, Will Never Find All the Sensitive Data.*

Even the best commercially available scans will not find all the different types of sensitive data existing in an organization. Scans can target common, standard formats of sensitive data (such as SSN), but a commercial scan can never find the idiosyncratic forms of sensitive data virtually every company generates. When an organization uses unique language to express sensitive concepts, a search for those concepts must be similarly unique.

Some scans offer to identify "junk" or non-essential business documents in collections, to help cull down those collections for more efficient handling and disposition. But in our experience, those tools can interfere with a search for sensitive (and often highly important) information, and therefore require individualized consideration by each organization using them. For example, a junk scan might look for terms associated with fantasy baseball teams, and cull those documents from a collection. But what if the company using the scan produces fantasy baseball game software? Using that tool out-of-the-box, without modification of its terms, would prove disastrous.

Finally, even fairly good "regular expression" based scans—those aimed at finding sensitive data entities that typically appear in a consistent, regularized format, such as SSN—benefit from testing and assessment on each company's unique data. We found that even minor fine-tuning of these scans (by adjusting or adding terms) can improve their effectiveness substantially.

What's Next?

In the course of our investigation of scan tools, and in working with the providers of several tools, we discovered an interesting explanation for the poor performance of these tools out-of-the-box—the software developers simply do not have access to sufficiently large and varied sets of data on which to test and refine the scans. So while the scans may seem logical and well-crafted in theory, until they are tested against "live" data generated by actual companies, they will not perform well in practice.

As we continue to explore improved methodologies to identify and protect sensitive data, we have embarked on a project that we hope will mitigate that limitation. We are harnessing the power of the many billions of files we store on behalf of our clients to conduct a methodical, statistically validated test of sensitive data scans. We have secured permission from a number of clients, across a diverse group of industries, to gather data from a variety of different types of matters into a large collection for testing. Using that collection, we will perform a thorough examination of the capabilities and limitations of several of the most widely used commercial sensitive data scan tools.

Later this year we plan to publish a research paper reporting the results of our tests (with all client data fully anonymized). We also intend to include some recommendations for how best to optimize the scans, and suggestions for best practices in using the scans. Please reach out to us if you'd like to receive a copy of our paper.