Afraid of the Dark (Data)

I have a cabinet at home that is stuffed full of…well, I am not quite sure what. I am pretty sure there are chargers, although I am not as certain that the devices they once charged still exist. There are cables, connectors and whatnot—a lot of whatnot. If I were a betting man, I would wager that most readers have a cabinet, closet, box or drawer just like mine.

The Worst Reason

I bring this up because many, perhaps most, enterprise data centers have the digital equivalent of my cabinet. They have a lot of what is termed "dark data," most widely defined as any information that businesses collect, process, and store, but do not use for other purposes such as analytics and insight.[1]

This is accurate, but I feel that it does not go far enough. Organizations have data—lots of data—that they do not know about, do not use, cannot find and would have gotten rid of long ago if they knew they had them. Much like my cabinet. But if you do not know that you have them and you could not find them if you did know, this is functionally equivalent to not having them. So why do organizations keep them? For the worst possible reason: You never know.[2] This way lies an awful lot of data retention.

Using Dark Data

Most of the literature I have read on the subject of dark data bemoans the wasted opportunities presented by all these bits and bytes. A widely quoted figure is that 55 percent of all enterprise data are, in fact, dark.[3] Whatever organizations are paying for storage, it would seem that the majority of this investment is wasted. Even more, dark data represent wasted opportunities. There is value in these data, if only organizations would develop and use the tools that might bring them into the light. As it is, ignoring the challenge of dark data is leaving money on the table.[4]

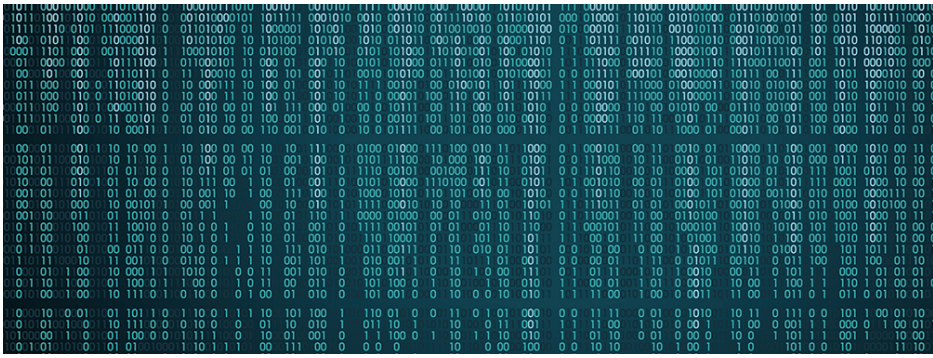
Artificial intelligence (AI), in particular machine learning, offers a means of identifying, researching, parsing and, ultimately, extracting value from dark data.[5] Machine learning offers the promise of putting this "junk" information to use.

I do not know what is in my cabinet, but I know how it all got there. Somebody (that would be me) put it there. Similarly, it is worth knowing where the dark data came from and where they are located. There are many sources of these data, including, but hardly limited to, logging, collection of employee-candidate data, geolocation, surveys, surveillance and email.[6] Other important sources are all the sensors used in industrial processes and, increasingly, in Internet of Things (IoT) devices.

The data are generally stored in network-attached storage (NAS) devices, the usual repository for unstructured data.[7] In particular, sensor data are by definition unstructured and easy to collect; it is easy to capture them and forget them. As a result, an industrial company might have piles of outdated sensor data, compiled for months and years, that are of no use to anyone.

Abusing Dark Data

Or almost anyone. There are a lot of bad guys who would love to know just how that chemical process works, how that pipeline runs, how that factory is operated. They can derive much of that from dark



Steven J. Ross, CISA, CDPSE, AFBCI, MBCP
Is executive principal of Risk Masters International LLC. Ross has been writing one of the *Journal's* most popular columns since 1998. He can be reached at stross@riskmastersintl.com.

data if they can get their hands on them. It occurred to me while reading about the recent cyberattack on a pipeline company that exactly this kind of detailed data may have been stolen. True, ransomware was the reported incident. But did thieves steal information as well, encrypting it? I have no inside insight and do not know if this happened, but the threat is real.

> "A MORE VALUABLE APPROACH IS TO ASSUME AT THE OUTSET THAT ALL UNSTRUCTURED INFORMATION, DARK OR OTHERWISE, IS VALUABLE, AT LEAST TO SOMEONE."

Most large companies and government agencies have information security functions that oversee the protection of their sensitive and critical data. In my experience, the dark, unstructured data do not figure as either sensitive or critical. To my knowledge, there is no published research on what is actually being done to protect dark data, though there is no shortage of advice as to what to be done.[8] Most of the suggestions I have read start (and sometimes end) with auditing unstructured data to find out what is actually there. Of course, once the data are located, identified and categorized, they are no longer dark. Hey, presto! Problem solved.

An Approach to Securing Dark Data

I think that a more valuable approach is to assume at the outset that all unstructured information, dark or otherwise, is valuable, at least to someone. Based on that assumption, the broad user population should be educated as to the potential value and risk of dark data and surveyed to identify those who use some portion of them. From there, it follows that basic security measures need to be in place (e.g., access control, encryption, logging, monitoring) for those data that are used, if they are not already. Those who perform data mining, analytics and other applications with some of the unstructured data should have no issue with legitimate controls.

Access to all the rest of the unstructured data should be prohibited. This step may aggravate a few people, but once they are identified as the users of these data, they may be granted access to what they need. After an acceptable period of time, perhaps one year, any unclaimed data may be discarded.

Some may consider a blanket removal of all these data to be too risky, because, well, you never know. There is the option of dumping them all onto some portable medium (magnetic tape comes to mind) and storing it somewhere secure. This would be the electronic equivalent of my cabinet, so who am I to object?

Once the unnecessary dark data are cleared away, the next requirement is to monitor the inflow of new unstructured data, to avoid refilling the dark hole. Any unexpected use of the data should be considered suspect. Remember that the legitimate uses would have been previously identified. It may make sense to intentionally create some tempting dark data, just to see who might be poking around in it.

In effect, dark data may be an unexpected back door to the kinds of business secrets I have addressed in several previous articles.[9, 10, 11] In my next column, I will give some consideration to private information that is kept in the dark.

Endnotes

1 Carter, R.; "What Is Dark Data and How Can You Use It?" UC Today, 23 October 2019, *https://www.uctoday.com/united-communications/what-is-dark-data-and-how-can-you-use-it/*

2 I recognize that there is a Law of Divine Retribution that states that the day after you throw out anything that has hung around unused forever, you will need it. This law was discovered by Newton. Or Murphy. Or Sod. Or someone.

3 There are a number of sources that present this figure. I believe they are all quoting an excellent study by the company Splunk, *The State of Dark Data*, USA, 2019, p. 3, *https://www.splunk.com/en_us/form/the-state-of-dark-data.html*. I consider the 55 percent estimate to be the most

accurate. There are others I have seen, ranging up to 90 percent (Johnson, H.; "Digging Up Dark Data: What Puts IBM at the Forefront of Insight Economy," SiliconANGLE, 30 October 2015, *https://siliconangle.com/2015/10/30/ibm-is-at-the-forefront-of-insight-economy-ibminsight/),* but only the Splunk study seems to be based on research rather than anecdotes.

4 Tully, T.; "Dark Data Has Huge Potential, But Not If We Keep Ignoring It," *Splunk*, 30 April 2019, *https://www.splunk.com/en_us/blog/leadership/dark-data-has-huge-potential-but-not-if-we-keep-ignoring-it.html*

5 Louvrier, L.; "The Rise of Machine Learning to Manage Dark Data," *Technative*, 2 August 2019, *https://technative.io/the-rise-of-machine-learning-to-manage-dark-data/*

6 Marsh, S.; "Dark Data—The Blind Spots in Your Analytics," iDashboards, 30 January 2019, *https://www.idashboards.com/blog/2019/01/30/dark-data-the-blind-spots-in-your-analytics/*

7 Unstructured and dark data are sometimes conflated. See, Alton, L.; "Is There Value in Unstructured Data?" ISACA Now, 27 November 2018, *https://www.isaca.org/resources/news-and-trends/isaca-now-blog*. Unfortunately, there are also many cases of structured data that are collected, filed and forgotten. In some cases, they are kept for reasons of regulatory compliance. In others, they are archive files that are gathering dust somewhere, never again to see the light of day.

8 A few examples will do: Kangaraj, A.; "Defend Yourself Against Dark Data," *InfoSecurity*, 13 November 2020, *https://www.infosecurity-magazine.com/opinions/defend-dark-data/;* Peters, M.; "Dark Data Is Hurting Your Cyber Security," Security Boulevard, 16 October 2019, *https://securityboulevard.com/2019/10/dark-data-is-hurting-your-cyber-security/*.

9 Ross, S.; "Secrecy and Privacy," *ISACA® Journal*, vol. 1, 2021, *https://www.isaca.org/archives*

10 Ross, S.; "Keeping Secrets," *ISACA Journal*, vol. 2, 2021, *https://www.isaca.org/archives*

11 Ross, S.; "Advanced Security for Secret Information," *ISACA Journal*, vol. 3, 2021, *https://www.isaca.org/archives*

Enjoying this article?

- Read *Achieving Data Security and Compliance. www.isaca.org/data-security-and-compliance-2020*
- Learn more about, discuss and collaborate on information and cybersecurity in ISACA's Online Forums. *https://engage.isaca.org/onlineforums*