Io-Tahoe®

ASSESSING DATA SENSITIVITY:

Evolving Modes and The Need for Advanced Techniques

By David Loshin,
President, Knowledge Integrity, Inc.

Introduction

Within the past few years, there is a growing awareness of the degree of inadequacy of corporate protection of personal and private information. There are increasing risks of data exposure, including cybersecurity failures leading to massive data breaches, unauthorized sale and transfer of personal data among business partners, or even what might appear to be corporate misuse of consumer data. The recognition of risks of exposing individuals' personal and private information has raised awareness among many governments, with over 80 countries enacting laws mandating protection of private data[1] and correspondingly, there is a growing inventory of global regulations designed to address the need to secure and protect individuals' personal and private data.

Despite some familiarity with general concepts of protection of "sensitive" data, many of these different laws have slightly, and in some cases, even widely different definitions of what is meant by personal data. To complicate matters more, different types of laws mandate protection of non-personal data (such as "chain of command" laws for managing scientific study data) are joined by corporate directives for protecting confidential data from industrial spies. In essence there are four key challenges that can stymie programs for both regulatory compliance and implementation of good data handling procedures:

- Gaps in knowing what types of information need to be protected, i.e. translating between the specification in any specific regulation and the practical implications for implementation.
- Recognizing that the scope of data protection can easily expand beyond the realm of private or personal data.
- Gaps in the ability to determine which data assets contain information that requires protection.
- Acknowledging that the scale of data assets within the enterprise far outstrips the ability for manually assessing data sensitivity.

In essence, understanding and managing data sensitivity is a required component of an overall program for data protection. Established technologies like identity access management, access control, data encryption, and data masking are useful for preventing access to protected data, while the techniques for determining what data assets need to be protected continue to mature.

This paper examines the challenges of assessing data sensitivity by first discussing the complexity of the concept of "personal" data, especially in the contexts of the multiple global data privacy laws. The paper then suggests that the realm of data sensitivity is not limited to an individual's personal data, but encompasses a much wider set of classifications for both individual and corporate data objects. The paper then explores how automation techniques can simplify the organization's ability to contend with this emerging challenge, and finally reflects on the next steps for acquiring the right types of solutions to support data sensitivity analysis as part of an overall data protection strategy.

[1] What's Data Privacy Law In Your Country?" accessed via https://www.privacypolicies.com/blog/privacy-law-by-country April 4, 2019

The Complexity of “Personal Data”

The definitions of “personal” and “private” data” are fluid. We can start with generic definitions gleaned from the Web:

- Personal information is “information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual in context.”[2]
- Private information is “Information that a user wishes to keep from public viewing. Credit card, social security and financial account numbers, along with passwords to websites and other venues, are commonly kept private.”[3]
- Sensitive data is “critical, safeguarded information.”[4]

A deeper dive into the concepts of personal and private data reveals much more complexity. For example, the European Union’s General Data Privacy Regulation (GDPR) defines personal data as

> ***any information relating to an identified or identifiable natural person*** *(‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;*[5]

The California Consumer Privacy Act (CCPA) definition is even more comprehensive. That definition effectively encompasses the GDPR definition and augments it with biometric information, network activity information, geolocation data, professional or employment-based information, education information, as well as inferences drawn from any identified type of personal information used to create a profile reflecting preferences, characteristics, and psychological trends, among other inferred attributes. That encompasses both well-defined identifying information as well as the results of analytical efforts for summarization, creation of customer profiles, and application of more advanced analytical algorithms.

While these regulations provide a starting point for defining the types of data that need to be protected, you still need to survey the enterprise information landscape to identify the data assets that contain personal or private data, what types of data they contain, where the data assets reside, as well as additional computational dependencies that might be covered under the laws. And if that were not enough, realize that assessment becomes more complex when one needs to fold in continuous analysis of data in motion, as opposed to just scanning data at rest.

[2] Wikipedia, accessed via https://en.wikipedia.org/wiki/Personally_identifiable_information

[3] Yourdictionary, see http://www.yourdictionary.com/private-information

[4] “5 Examples Of Sensitive Data Flowing Through Your Network (& How To Protect It),” https://www.bitsighttech.com/blog/sensitive-data-examples-how-to-protect-it

[5] https://www.gdpreu.org/the-regulation/key-concepts/personal-data/

Evolving Modes of Data Sensitivity and the Value of Assessment

With the news about data breaches and awareness of data privacy regulations focusing attention on the protection of private and personal information, there are many other classifications of data sensitivity that require some type of protection, even if corresponding data items are not subject to regulatory compliance. Some examples include:

- **Confidential information**, such as internal corporate sales and production statistics that should not be made available to external parties;
- **Licensed information** that is subject to negotiated constraints, such as data purchased from a data aggregator that can only be used for specific purposes;
- **Privileged attorney-client data**, or confidential communications and legal advice;
- **Export-controlled research**, including reports that include national security information that should not be transferred out of the country;
- **Details of implemented security controls**, whose exposure could lead to more serious data breaches;
- **Credit card** and other payment information;
- **Public safety information**;
- **User name/password** combinations;
- **Calendars and individual schedules**;
- **Emails**;
- **Intellectual property and trade secrets**; or even
- **Corporate operations information**, describing the processes executed within an organization.

Sensitive data is not limited to just personally identifiable information – it encompasses a variety of different modes such as protecting organization interests, complying with contractual obligations, and guarding against actions of bad actors. Therefore, a strategy for data protection must take into account these different modes and embrace a means for assessing, tagging, and managing controlled access to any data asset that might contain any type of sensitive information.

The Need for Automation

Key aspects of implementing a data protection strategy, whether motivated by regulatory compliance or inspired by applying good data management practices, involve operational tasks for determining and mitigating data asset risk factors for exposure, including:

- Clarifying and formalizing the definition of "data sensitivity."
- Discovering data assets that might be subject to protection.
- Profiling data asset metadata and assessing content for sensitivity.
- Classifying each data asset according to its levels of sensitivity.
- Capturing data asset metadata, constraints, intent, and obligations in a shared data catalog.

The scale associated with growing data volumes and the expanding classes of sensitivity make all of these tasks unwieldy, if not impossible to do manually. Instead, look for ways for adopting technologies to simplify these processes by:

- Automatically crawling through the enterprise to surface the inventory of organizational data assets.
- Identifying incoming data feeds/streams that might contain sensitive data.
- Automatically assessing each data asset for sensitive data, both at rest and in motion.
- Presenting the inferred classifications to humans for validation.
- Taking advantage of machine learning (ML) and artificial intelligence (AI) algorithms to continuously enhance the precision and accuracy of the sensitivity analysis.
- Adding the discovered information to a shared data catalog that can be searched and browsed by a community of data consumers.

These techniques work together to reduce manual involvement. Increasing the level of automation while maintaining accuracy and precision allow you to scale the implementation of your data protection strategy by freeing the data stewards to focus on the compliance aspects of protection instead of the staggering manual effort of assessment and classification.

Summary and Considerations

Raised awareness of information threats and vulnerabilities, potential for exposure, coupled with an ever-growing list of global data privacy protection laws means that the need for a robust strategy for data protection will only continue to become more acute. A data protection strategy relies on the implementation of security protocols, including system perimeter protections, data encryption, and data masking. Yet these techniques are minimally effective if you cannot determine what data assets require which levels of protection in relation to specific assessments of data sensitivity.

An initial task is to devise a formal definition of what "sensitive data" means within the organization. Examine externally-defined regulations that mandate any type of data protection, whether that involves individuals' data privacy or other types of controls imposed on data workflows. Engage the business leaders to understand corporate protection demands for confidential information and intellectual property. Understand the different categories and validate your taxonomy by reviewing and classifying a selected set of data assets. In addition, document the obligations associated with each class of sensitive data (such as regulatory compliance directives or reporting requirements in the event of unauthorized exposure).

Next, solidify your requirements for technology support. Our evaluation of the challenges imply that the fundamental technical capabilities required for a data protection strategy are data discovery, assessment, classification, and cataloging. Identify the requirements for these capabilities and how tools like a data catalog can be used to support data obligation management. Technologies that integrate machine learning and artificial intelligence with human interactions leverage synergy that allows the systems to learn and get smarter.

Look for products that simplify the process through the incorporation of advanced analytics to provide high fidelity of the results. Importantly, do not limit your capabilities to data at rest – make sure that the selected platforms incorporate the ability to assess sensitivity with both data at rest *and* data in motion. Augmenting data sensitivity analysis, classification, and cataloging with machine learning algorithms allows the process to have AI and human intelligence work together to enable automation while establishing a high level of trust.

This file is private and may contain confidential and proprietary information. If you have received this file in error, please notify us and remove it from your system, and note that you must not copy, distribute or take any action in reliance on it. Any unauthorized use or disclosure of the contents of this file is not permitted and may be unlawful. Io-Tahoe LLC, 111 Broadway, Suite 601, New York, NY 10006. www.io-tahoe.com

Date of publication: May 2019.