netwrix

# 10 Ways You Can Improve Your Automated Data Classification with Netwrix Auditor

www.netwrix.com

# Table of Contents

# Introduction

Organizations have limited resources to invest in safeguarding their data. Knowing exactly what data needs protection the most will help you set priorities so you can develop a sound plan that allocates your budget and other resources wisely — in other words, a plan that maximizes security and compliance while minimizing costs. But where's the best place to start? Data classification.

For many years, data classification was purely a user-driven process; users were tasked with classifying the documents they created, sent, modified or otherwise touched. This approach can be very precise, but it's not very efficient. Moreover, by the time an organization starts thinking about data classification, it usually already has terabytes of data — who's going to classify all those documents?

Automated data discovery and classification addresses both these issues. It is much more efficient than a manual approach, and it can tackle the backlog of all your existing data without overwhelming your staff resources.

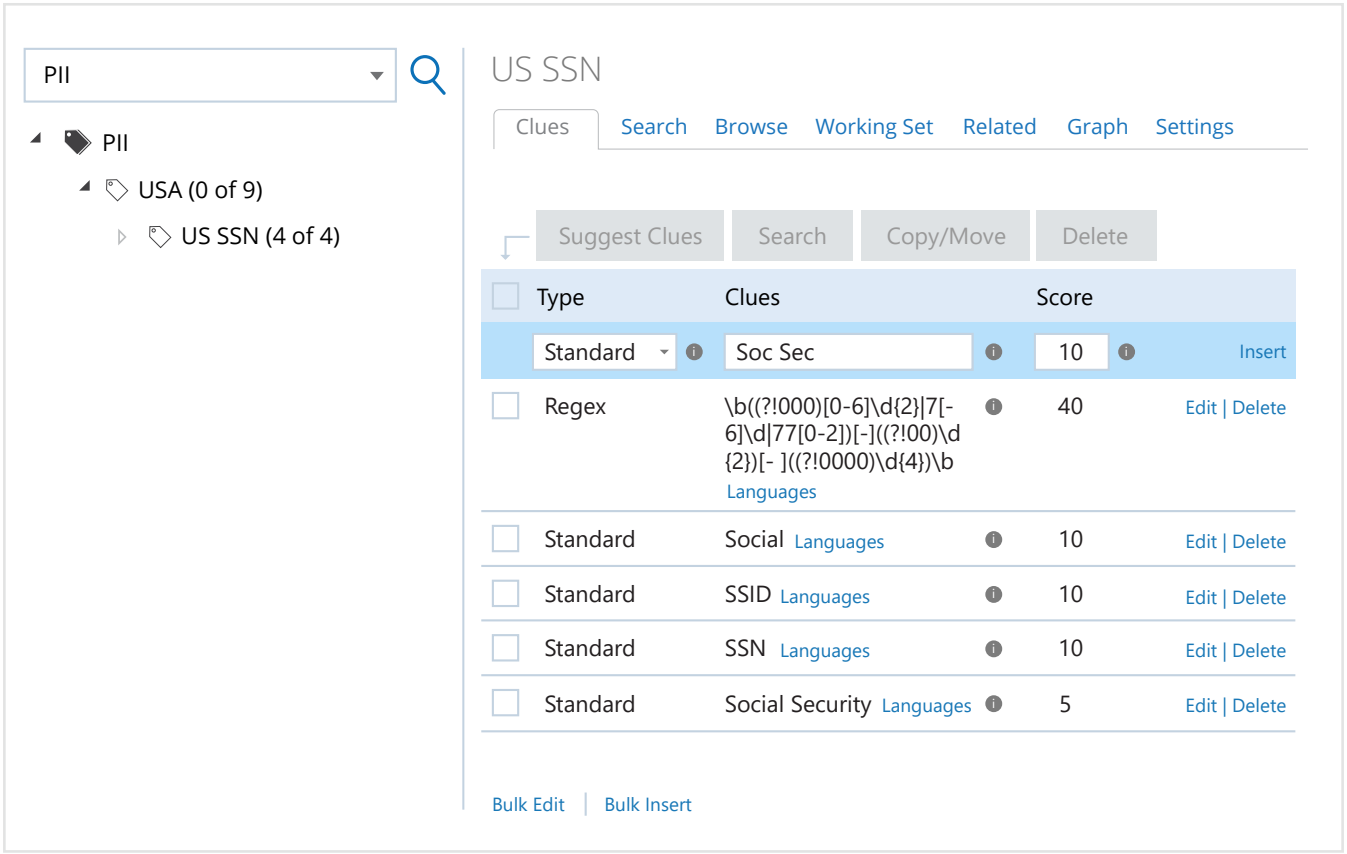
However, to reach a high level of precision, you have to find a way to eliminate all the false positive and negatives. The more sensitive data that the solution doesn't recognize (false negatives) and the more documents it classifies incorrectly (false positives), the more you're at risk of not protecting your sensitive data properly. As a result, you'll have to spend more time on fine-tuning the solution and less time actually managing and securing data. Improving the precision of data discovery classification is no easy task, but it's one you can't afford to disregard. What good is an automated solution if it gives you results that you can't trust and use?

This eBook can help. We'll discuss 10 ways in which Netwrix Auditor empowers you to increase the precision of automated data discovery and classification, so you can derive more value from this technology.

# 1. Complement your RegExes with keywords and keyword phrases

Keyword matching is nothing new in data discovery and classification. However, most solutions use it as a substitute for RegExes, instead of using it to complement them to increase precision. That is, they simply look for specific words or phrases that characterize sensitive data, such as “bank card”, and then assume that any document with the keyword actually contains credit card numbers. This assumption can easily be false, which increases the number of false positives you have to filter out.

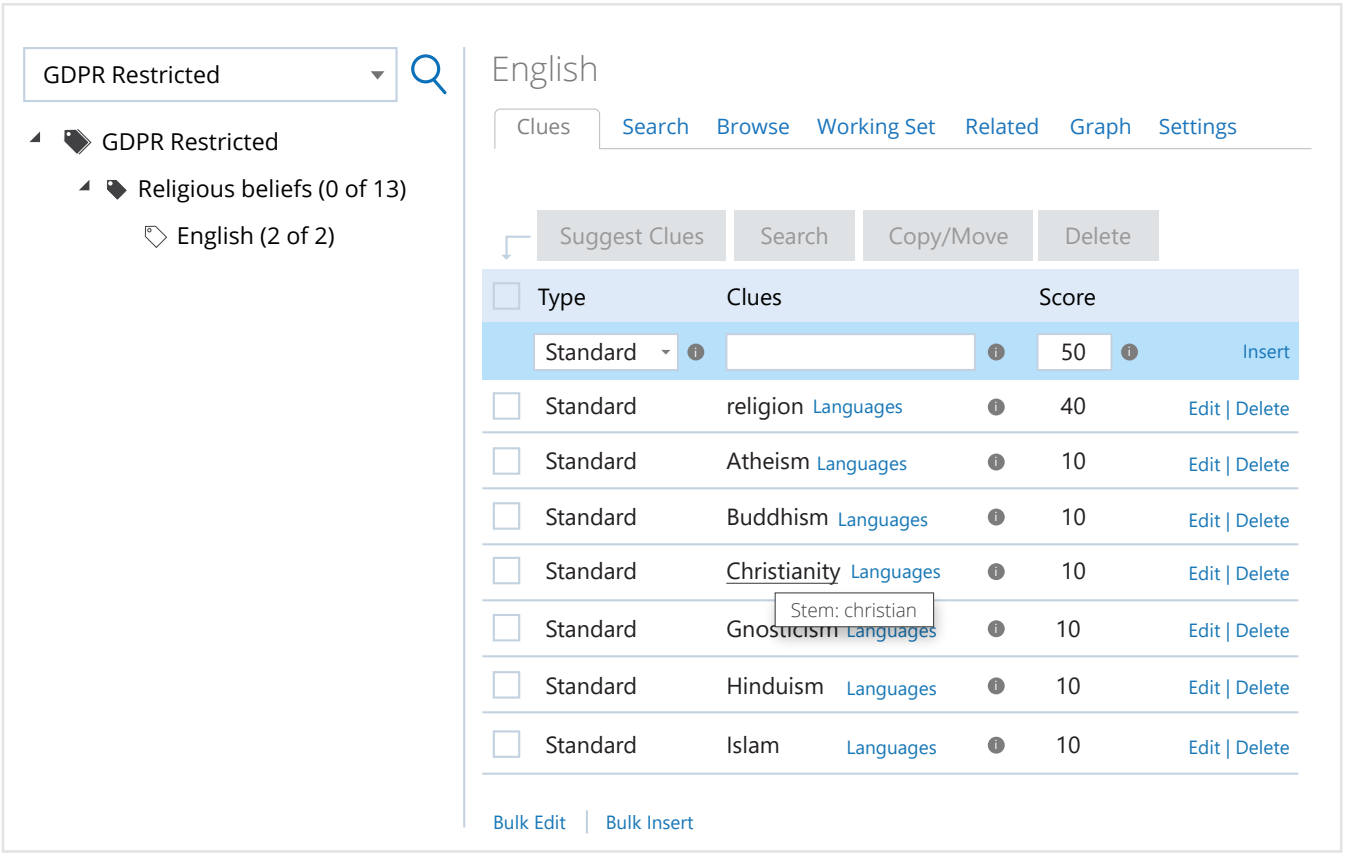
Netwrix Auditor’s Data Discovery and Classification, on the other hand, enables you to use keywords to “validate” your regular expressions. Requiring files to match against both a RegEx and a keyword before they are classified can dramatically reduce the number of false positives, saving you all the time you would have spent dealing with them.



# 2. Take into account keyword variations

There are lots of variations on most keywords: “license,” “licenses,” “licensing” and “licensed” all reflect the same concept. But most legacy data classification solutions can handle only exact keyword matches, so users have to list every variation they want to catch. That’s a time-consuming and error-prone task, so they are often left with a lot of false negatives. And if your tool fails to spot sensitive data, you don’t know to protect that data, which puts you at risk of security breaches and compliance failures.

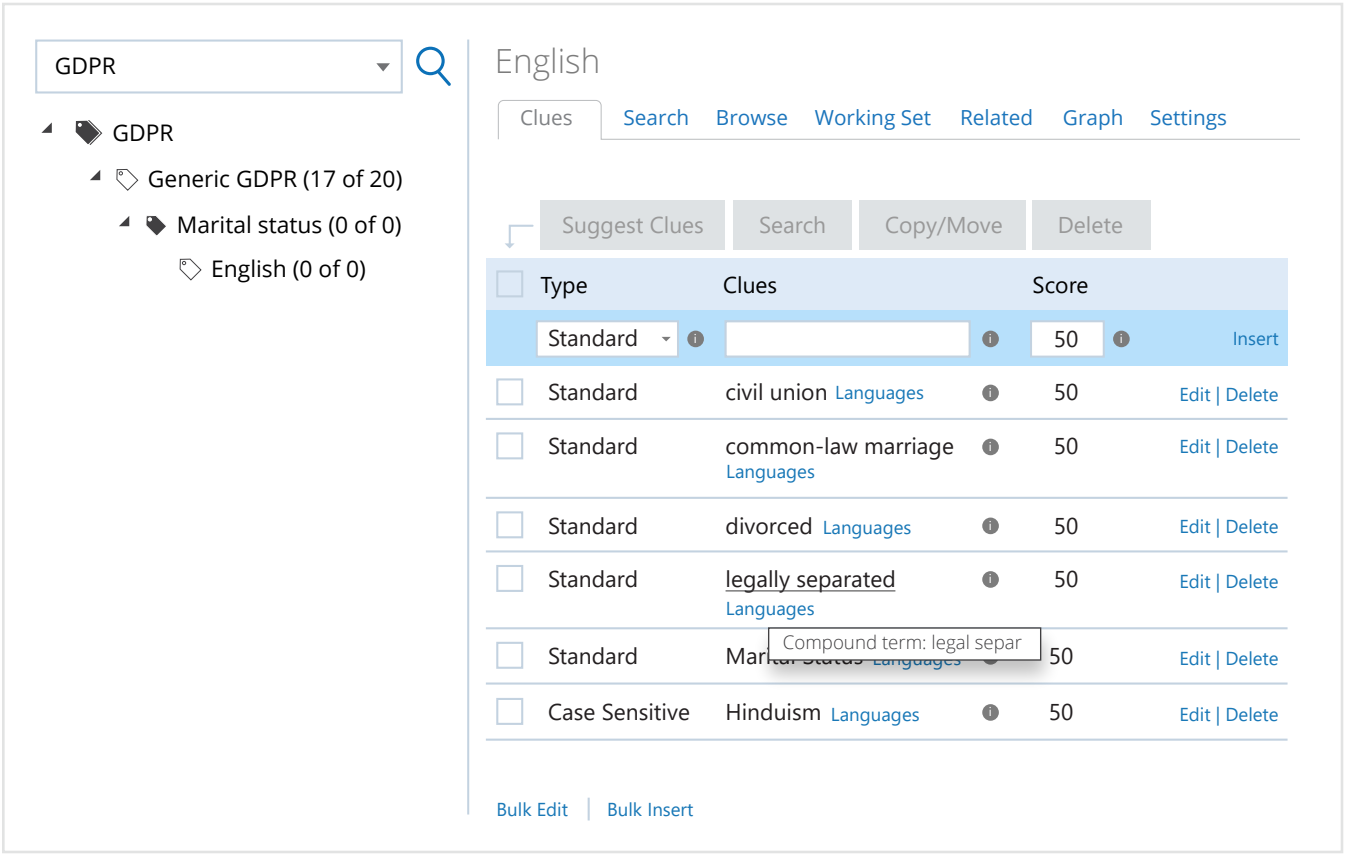
But it doesn’t have to be this way. Netwrix Auditor’s Data Discovery and Classification uses stemming to save you the time and effort of having to deal with word inflections. It automatically reduces any keyword to a stem that contains the meaningful part of the keyword, so you don’t have to spell out countless variations of the same term; you can simply put in one keyword and Netwrix Auditor will match its variations automatically.



# 3. Allow for fuzzy or partial matching of key phrases

What if you want to use a key phrase instead of a simple keyword? The more words in a phrase, the less the chances for an exact match. Handling inflections of the constituent words is just the start; parts of the phrase might be split up, with other words in between. Most solutions are unable to recognize that “passport UK” and “passports of UK citizens” reflect the same concept, and the only option you have to circumnavigate that is to split each key phrase into its individual keywords. But individual keywords simply don’t have the same meaning as the larger key phrase. If you make “UK” a keyword by itself, for instance, you’re bound to get lots of false positives.

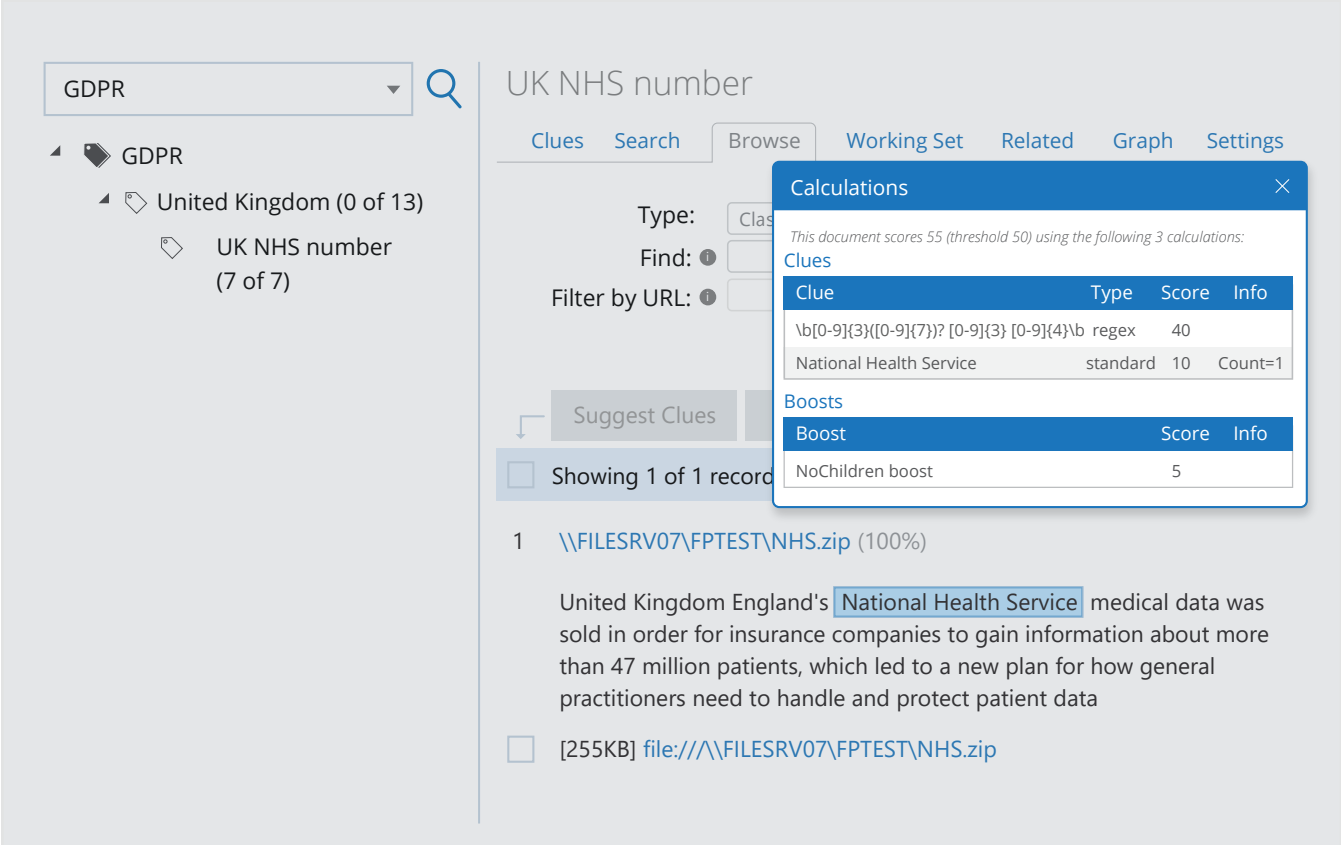
Fortunately, Netwrix Auditor’s Data Discovery and Classification can work with key phrases much the same way it works with individual keywords. It automatically identifies the compound term behind each keyword phrase that covers most of its variations, and supports partial or fuzzy matching. As a result, it will spot the key phrase even if the constituent words are separated by prepositions, their order is shuffled or some words are missing altogether.



# 4. Examine the results using classification scores

Most data classification tools don't give you much visibility into their behind the scenes. You simply know that a given file was classified into a particular category (or not classified at all). A few tools go a bit further and show you which piece of content matched against a RegEx or keyword. But even that still doesn't give you enough data to analyze and improve the precision of automated data classification in your organization.

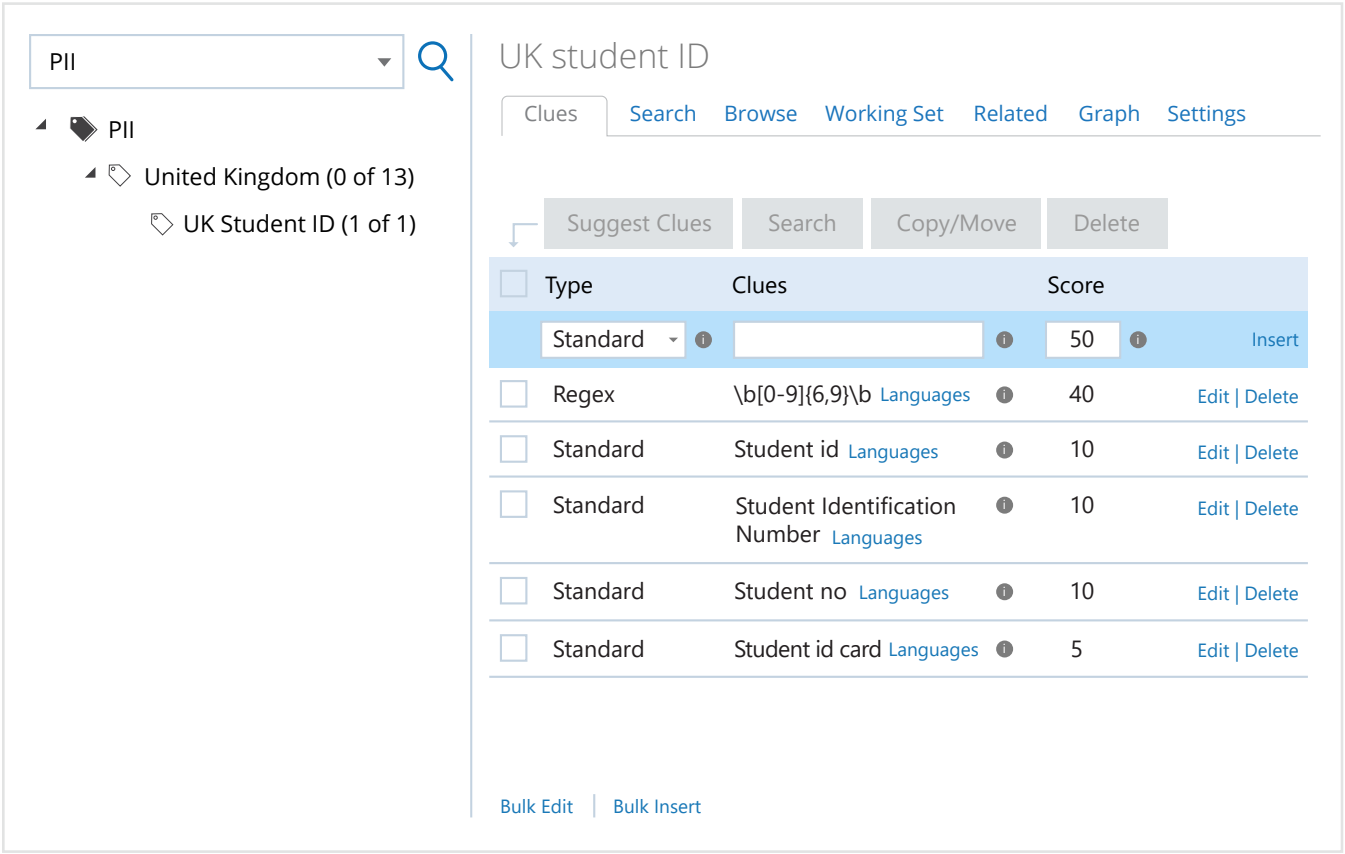
Netwrix Auditor's Data Discovery and Classification shows you precisely how and why each document was classified the way it was. Each rule that matches against a document adds to the score for that document; once the total score hits a certain threshold, the file is classified. You can review exactly which rules were involved and how much they contributed to the classification score. With this deep insight into why files were classified the way they were, you can fine-tune your rules to achieve much greater precision.



# 5. Assign different weight to different RegExes and keywords

Chances are, the keywords and RegExes you're using are not equally important. For instance, a match on a well-designed RegEx for a credit card number makes it much more likely that the document contains sensitive data than a match to the simple key phrase "credit card" — it's simply a better clue. Even two RegExes can have different relevance. Consider this: Current VISA cards have 16 digits, while older ones have only 13. Therefore, a match on the 16-digit RegEx is a better clue that you've found a VISA card number than a match on the 13-digit RegEx. Unfortunately, most tools don't have the flexibility to weight various clues differently, which limits your ability to reduce the number of false positives and negatives.

With Netwrix Auditor, you can decide how much weight to assign to each individual RegEx, keyword or key phrase, so that only the right combinations of these clues pushes the document over the classification threshold.

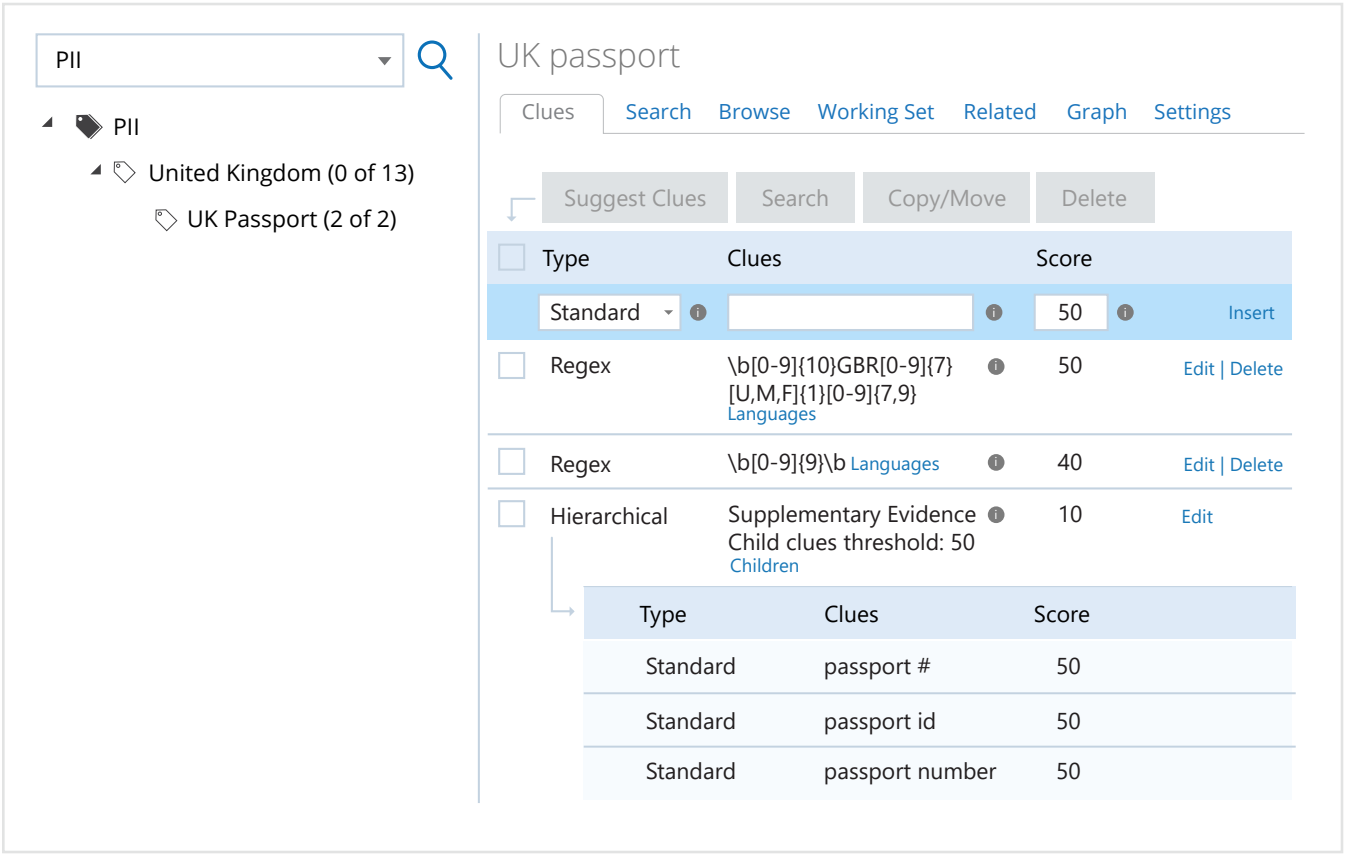


# 6. Prevent multiple keyword matches from skewing the results

Creating a large enough set of keywords and complementing them with RegExes is critical — but it's just the foundation for increasing the precision of data classification. The next step is interconnecting them.

For instance, suppose you created a very precise VISA card number RegEx and came up with tens of relevant keywords for it. Since each keyword contributes a certain value to the classification score, a given file might be classified simply because of the sheer volume of keywords inside it. This of course, increases the rate of false positives dramatically.

Netwrix Auditor allows you to create a single "master" keyword, which is the set of all keywords that you listed, and assign it a score. The master keyword will match against a document only if a sufficient number of the individual keywords are found. In that case, the single score associated with the master keyword will contribute to the document's total score, rather than the scores associated with each individual keyword, which might artificially inflate the total. That way, the total score will push the file over the classification threshold only if it matches both the master keyword and a RegEx.

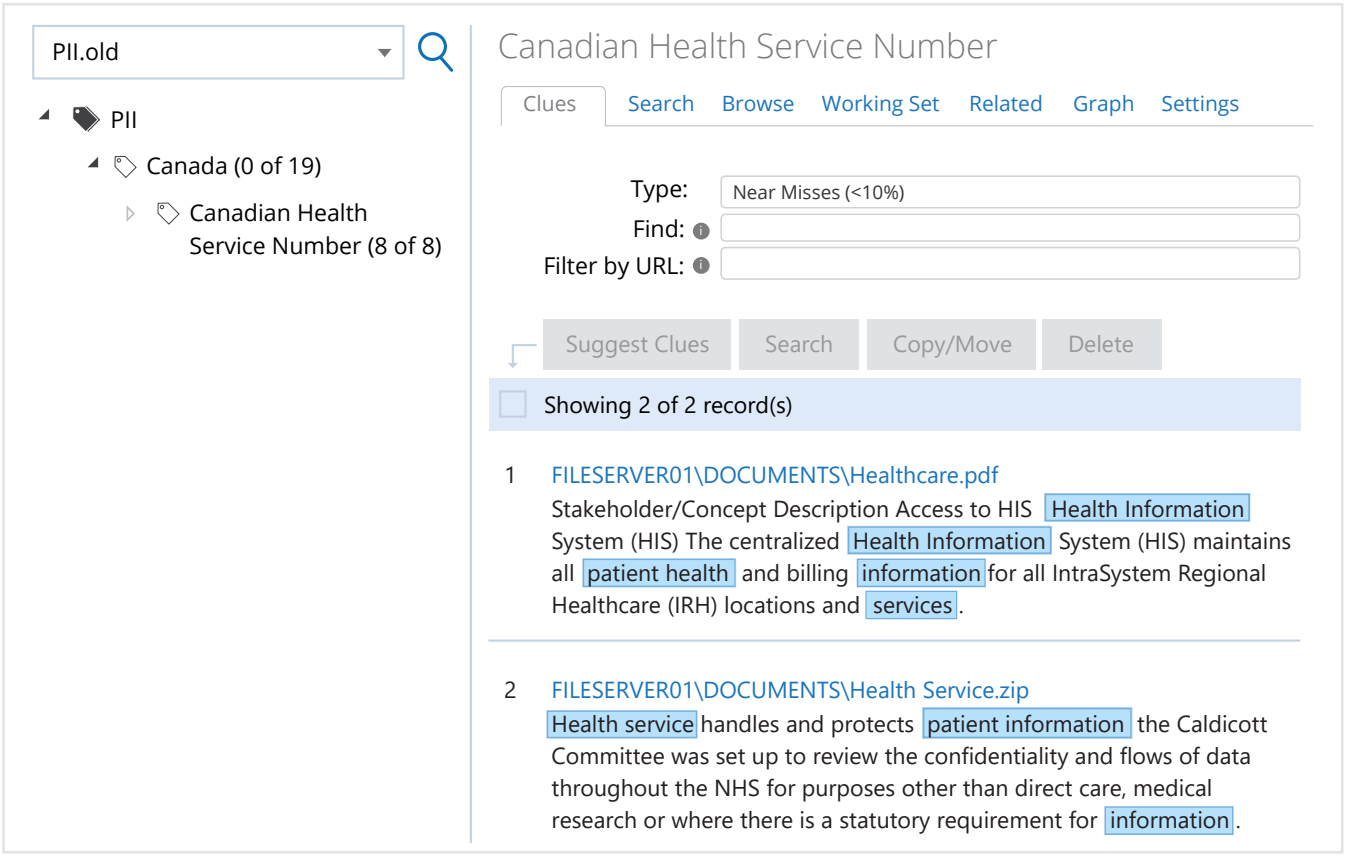


# 7. Analyze near misses

We described how Netwrix Auditor enables you to you fine-tune your rules to achieve much greater precision by giving you deep insight into why files were classified the way they were. Let's look at another thing you can analyze to improve the precision of data classification.

Netwrix Auditor's Data Discovery and Classification assigns a score to each document it scans — including those that scored just below a threshold. These documents warrant special attention. You'll probably spot some that should have been classified, and a little digging will enable you to extract new keywords to include in your classification rules. These new keywords will reduce false negatives by pushing the near misses over the classification threshold.

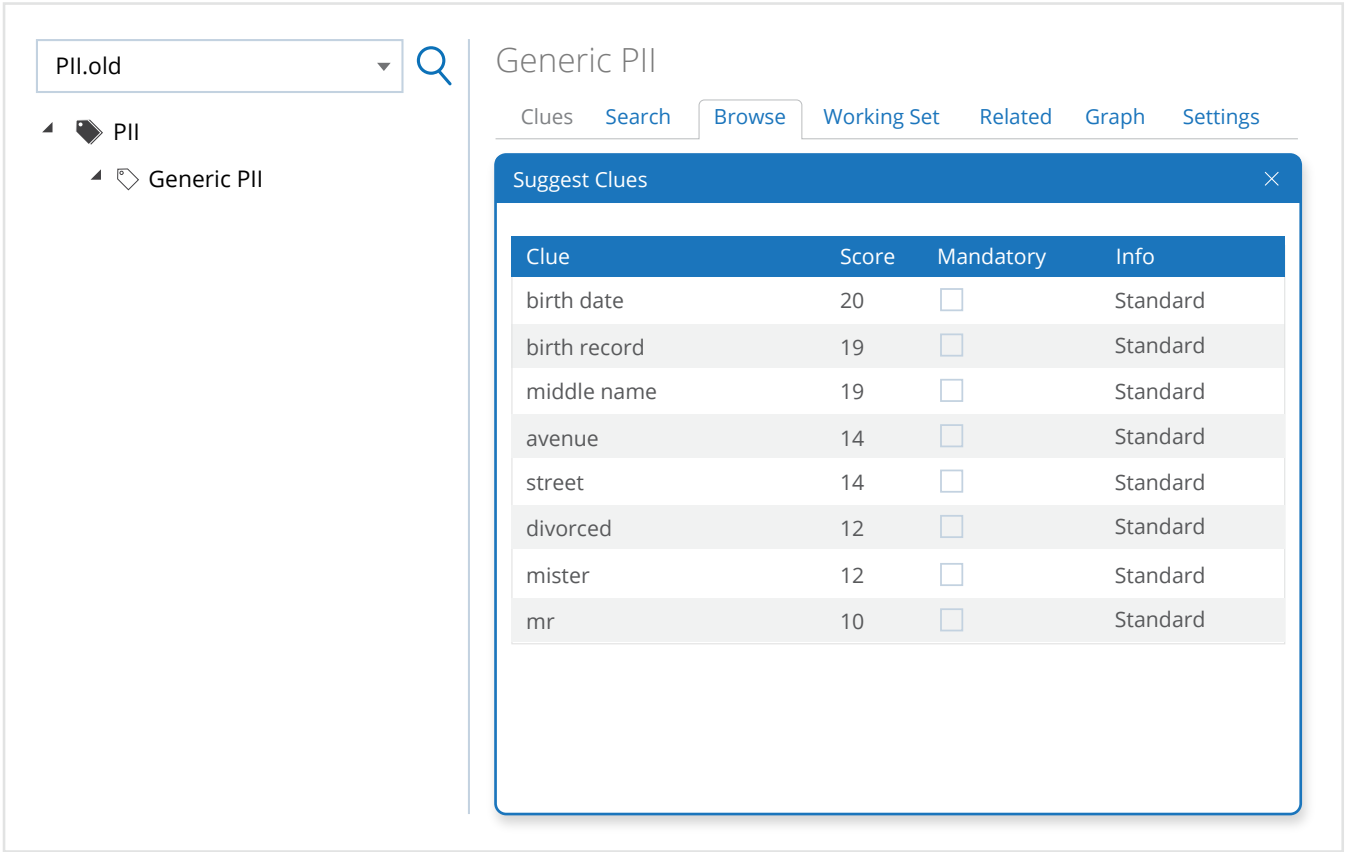
On the flip side, you might discover that some of the near misses are clearly not relevant. Netwrix Auditor enables you to determine how they got their high scores, so you can delete excessive keywords from your classification rules, or specify negative keywords (which prohibit a file containing them to be classified by the rule).



# 8. Derive new keywords from classified files

The traditional approach to automated data classification is to think about what content might be in a sensitive file and then define keywords and RegExes to spot it. If you fail to include certain keywords and RegExes, you're going to miss files with sensitive data. Netwrix Auditor closes this gap by empowering you to take a more empirical approach to identifying sensitive files.

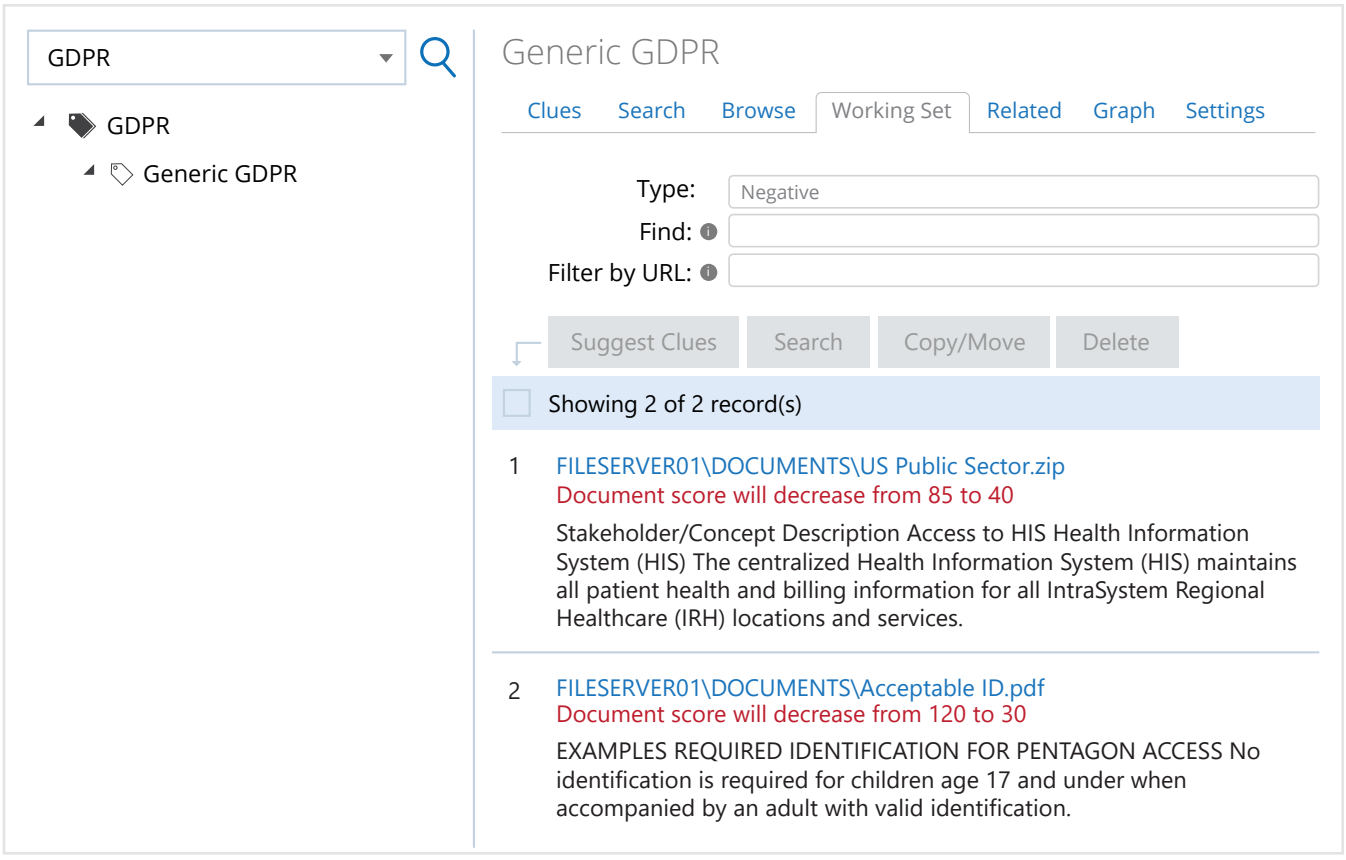
For example, suppose you're looking for documents containing intellectual property. The rules you created worked to some extent — you found some documents and verified that they contain IP. But you can't be sure that you haven't missed anything. Netwrix Auditor can analyze the content of that files that were already classified as IP and automatically suggest and weight new keywords and phrases that those documents have in common, and you can use those terms to find even more documents containing IP. By doing this on regular basis, you'll maintain precise and up-to-date classification rules that will find your IP whenever and wherever it surfaces.



# 9. Validate changes to classification rules before implementing them

Legacy data classification tools leave little to no room for error when it comes to editing and creating new classification rules. Users hope that their modifications will improve precision, but they can't be sure until the solution actually crawls through all their documents, which can take days. If the changes didn't have the desired effect, it's back to the drawing board again. This lack of transparency discourages many organizations from even trying to improve their rules.

Netwrix Auditor's Data Discovery and Classification, on the other hand, can simulate the changes you made and show how they will affect the files that have already been classified. You simply select a representative sample of files and split it into a "good" working set that consists of files that were classified correctly, and a "bad" working set that consists of the false positives. Then you can play around with your classification rules and see whether your changes push the false positives below the threshold without decreasing the scores of the documents in the "good" working set. After you are satisfied with the results, you can implement those changes at scale.



# 10. Repeat strategies 1 through 9 on a regular basis

Although this point may seem obvious, it is the most important one: Data classification is an ongoing effort. If your solution is not transparent and flexible, or if you require assistance from a vendor's engineer every time you need to tweak a rule, then you can't do automated data discovery and classification properly. Nobody but you knows how best to classify your data. The process should be continuous, and your solution should support these continuous workflows and encourage you to customize and tweak things to your specific requirements.

The iterative nature of data classification goes beyond improving precision. As both the data you store and your business objectives evolve, so should your data classification strategy. Tomorrow there might be a new type of sensitive data that you need to be able to identify so you can secure it. Maybe your R&D team will need help organizing their documents for better efficiency, and you'll need to work with them to come up with the most useful categories and classification rules. To meet these new business challenges, your data classification solution should not only be precise but flexible.

Legacy data classification tools don't do that. Netwrix Auditor does.

# Conclusion

Automated data classification is an essential part of any organization's data protection and compliance strategy. After all, you need to know exactly what types of sensitive data you have in order to protect it appropriately and demonstrate to auditors that you have proper controls in place. You need to spend your limited time and budget dollars on the data that matters most.

Unfortunately, many automated data classification tools rely on exact keywords and rigid rules, and prevent you from peeking under the covers to see why documents were classified the way they were in order to learn how to improve the results. Netwrix Auditor gives you the flexibility and visibility you need to implement a continuous data discovery and classification process that gets increasingly precise over time. It also helps you answer the following questions to get the full context around your sensitive data:

- Is there any sensitive data in unsecure locations?
- Who can access each piece of sensitive data?
- Who owns each piece of sensitive data?
- Was this piece of sensitive data breached?

To learn more and see this powerful functionality in action, please visit netwrix.com/classification.

However, discovering and classifying data is just the first step in protecting it. You also need to think about workflows to deal with the data you discover. Do you want to quarantine these files or remove them from an overexposed location altogether? Will your business processes be affected by such drastic measures? Or would you prefer to a less radical approach? Would you want to extract the sensitive content, but leave the redacted document in the original location? Or perhaps restrict the access to only trusted groups? Or simply alert the data protection officer that an action needs to be taken? These are just some of the choices that you have.

Take all risks into consideration before making that decision and as was the case with data discovery and classification itself, think about the benefits that automating some of these workflows will bring.

# About Netwrix

Netwrix Corporation is a software company focused exclusively on providing IT security and operations teams with pervasive visibility into user behavior, system configurations and data sensitivity across hybrid IT infrastructures to protect data regardless of its location. Over 9,000 organizations worldwide rely on Netwrix to detect and proactively mitigate data security threats, pass compliance audits with less effort and expense, and increase the productivity of their IT teams.

Founded in 2006, Netwrix has earned more than 140 industry awards and been named to both the Inc. 5000 and Deloitte Technology Fast 500 lists of the fastest growing companies in the U.S.

Netwrix Auditor is a visibility platform for user behavior analysis and risk mitigation that enables control over changes, configurations and access in hybrid IT environments to protect data regardless of its location. The platform provides security intelligence to identify security holes, detect anomalies in user behavior and investigate threat patterns in time to prevent real damage.

Netwrix Auditor includes applications for Active Directory, Azure AD, Exchange, Office 365, Windows file servers, EMC storage devices, NetApp filer appliances, SharePoint, Oracle Database, SQL Server, VMware and Windows Server. Empowered with a RESTful API and user activity video recording, the platform delivers visibility and control across all of your on-premises and cloud-based IT systems in a unified way.

For more information, visit www.netwrix.com

**On-Premises Deployment**

Download a
free 20-day trial

netwrix.com/go/freetrial

**Virtual Appliance**

Download
our virtual machine image

netwrix.com/go/appliance

**Cloud Deployment**

Deploy NetwrixAuditor
in the Cloud

netwrix.com/go/cloud

**Corporate Headquarters:** 300 Spectrum Center Drive, Suite 1100, Irvine, CA 92618

**Phone:** 1-949-407-5125 **Toll-free:** 888-638-9749 **EMEA:** +44 (0) 203-588-3023

**netwrix.com/social**