RED TEAM CONFIDENTIAL

EXAMPLE - FOR DEMO PURPOSES ONLY

Contact : Andreas Strøbek, ABS@dataandmore.com

DATA & MORE

DATA PRIVACY & SECURITY

RISK ASSESSMENT

for

MUNSTER INC.

MUNSTER INC
SECURITY SOFTWARE

Data generated on:
20-05-2025

TABLE OF CONTENTS

This report has been produced by Data & More ApS based on data from Munster Inc.. It is considered confidential between the two parties. The report does not contain personal sensitive data but may disclose information about behaviour of Munster Inc. employees. Parts of the report has been generated with the help of AI.

KEY FIGURES

Privacy data at Risk
6.902 | 0,89%

Security data at Risk
1.310 | 0,17%

Data growth forecast
8.697 | 26% anually

Average age of data at Risk
2 years, 9 months, 7 days

Oldest document at risk
21.10.2016

Super Toxic Documents (25+ Data Subjects)
170

Data Subjects at Risk
92 internal | 460 external | 552 total

Outgoing Privacy Data (sent by mail)
2.107

Externally Shared Privacy Data (from drives)
0

Annual Estimated Data Risk Premium
€142.285

EXECUTIVE SUMMARY

This **Privacy Risk & Compliance Report** is based on the scanning of selected email accounts from Munster Inc. with the Data & More Privacy Platform.

Our platform has scanned **777.175** data sets in **35** accounts for a total of **335,70 GB** of data.

The scanning shows a total of **6.902** illegal privacy documents, and **1.310** documents containing security information. In tottal we have **6.902** doucments at **Risk**

Overall, the current trend in the production of data suggests that the amount of data at risk will **grow** by **26** percent annually.

The oldest non-compliant document we found is created **21.10.2016**

Based on the scanning of **35** accounts, the average non-compliant data per account is **248** out of **22.205**. This means a percentage of data that is noncompliant of **1,12%**. The global average with our customers is **0,75%**.

We have identified **552** different Data Subjects in the privacy classified documents. That means Data Subjects for which, you are storing personal sensitive data.

Our security scanning also shows that a total of **1.310** documents contains security information such as passwords, encryption keys and code that constitutes a significant security risk.

The above findings have uncovered a substantial financial, reputation, and operational risk in **Munster Inc.'s** current management of unstructured data.

Based on data from the European Data Protection Board and EU benchmark, the annual Data Privacy Risk Premium (the amount that should be set aside to cover the cost of losing the data to bad actors or auditing) would be €142.285 per year. For the calculation method and methodology please refer to The Cost of Doing Nothing.

This risk and associated cost would be substantially reduced by implementing automatic data compliance as suggested by Data & More

Document class	# Docs	Average age in days
test label	2669	1354
National ID number	1600	1220
Passport	1255	1091
Passwords & Secrets	1090	546
Recruitment	1070	1380
Travel info	665	385
Drivers license	507	841
Misc. ID	471	1033
Health card	383	1183
Employment info	361	430
Health info	282	650
Certificates / Permit	249	433
National ID Card	150	721
Source Code	129	1135
Salary / financial info	112	540
Employee Termination	85	604
Infrastructure Config	59	595
Ethnic origin	51	498
Employee warning	47	284
Educational info	43	335
Sexual orientation	34	395
Vulnerability assessment	32	1197
Tax Info	30	904
Insurance info	23	467
Religious Orientation	19	466
Payment Card	17	1084
Union Membership	17	438
Political Orientation	9	965
test	8	507
test gl	2	50
Wills	2	634

Table 1: Frequency and average age per document class

Document path	# Data Subjects
dj@365.dataandmore.com/GDPR_TEST_DATA/General _GDPR/GDPR Poc PDF.msg	275
dj@365.dataandmore.com/General _GDPR/GDPR Poc PDF.msg	275
abs@365.dataandmore.com/deles/ProjectGDPR/ProjectGDPR/GDPR Poc PDF.msg	275
abs@365.dataandmore.com/deles/ProjectGDPR/GDPR Poc PDF.msg	275
katarina.raguz@365.dataandmore.com/GDPR_TEST_DATA/General _GDPR/GDPR Poc PDF.msg	275
dj@365.dataandmore.com/GDPR_TEST_DATA/General _GDPR/GDPR Poc PDF.msg	275
abs@365.dataandmore.com/deles/ProjectGDPR/GDPR Poc PDF.msg	275
abs@365.dataandmore.com/deles/ProjectGDPR/ProjectGDPR/GDPR Poc PDF.msg	275
dj@365.dataandmore.com/General _GDPR/GDPR Poc PDF.msg	275
dj@365.dataandmore.com/General _GDPR/GDPR Poc - ZIP.msg	254

Table 2: Documents with the highest # of data subjects

ON THE POC SCOPE

Our PoC scanned **777.175** data sets or **335,70 GB**.

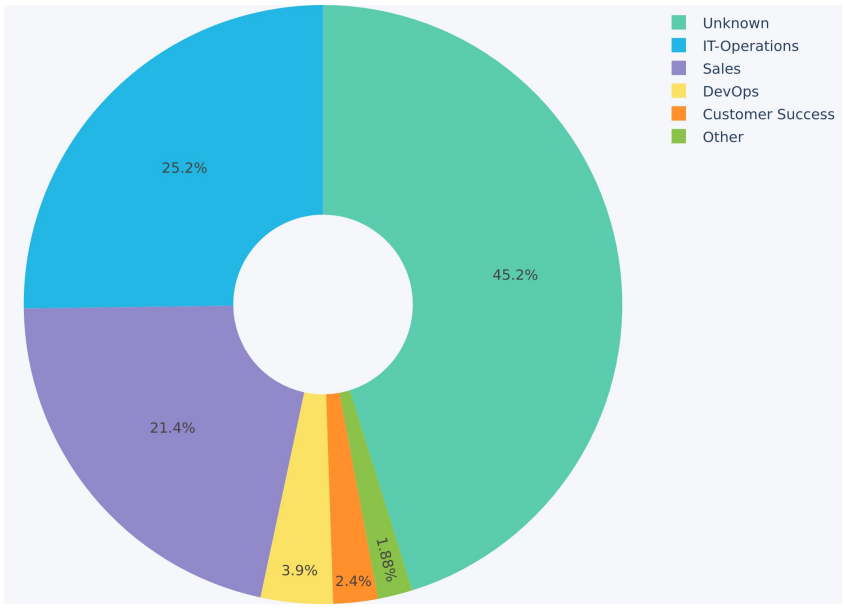

Figure 1: Documents with privacy and security data by Department

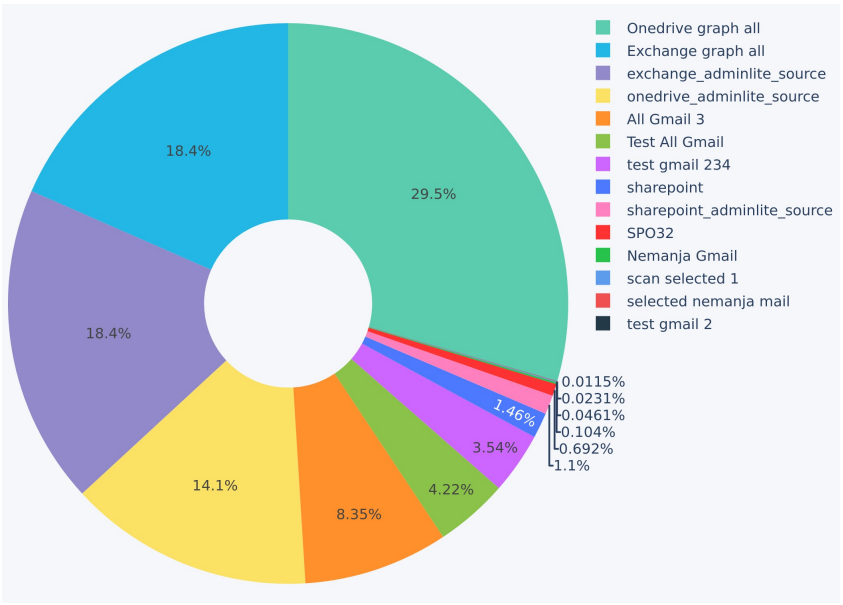

Figure 2: Documents with privacy and security data by source type

Source	Account	Department	# Docs with privacy data
All Gmail 3	dj@dataandmore.com	IT-Operations	164 (0,75%)
All Gmail 3	kbb@dataandmore.com	Unknown	95 (2,44%)
All Gmail 3	milica.stojnic@dataandmore.com	Unknown	90 (3,01%)
All Gmail 3	kbhh@dataandmore.com	Unknown	64 (0,23%)
All Gmail 3	dt@dataandmore.com	Unknown	56 (3,12%)
All Gmail 3	katarina.raguz@dataandmore.com	DevOps	51 (11,43%)
All Gmail 3	ll@dataandmore.com	Unknown	51 (2,21%)
All Gmail 3	svf@dataandmore.com	Unknown	41 (3,21%)
All Gmail 3	men@dataandmore.com	Unknown	38 (0,43%)
All Gmail 3	jm@dataandmore.com	Unknown	35 (0,36%)
All Gmail 3	mas@dataandmore.com	Unknown	22 (12,22%)
All Gmail 3	julia@dataandmore.com	Unknown	21 (0,75%)
All Gmail 3	fhh@dataandmore.com	Unknown	12 (1,29%)
All Gmail 3	finance@dataandmore.com	Unknown	9 (0,24%)
All Gmail 3	mihajlo@dataandmore.com	Unknown	9 (1,85%)
All Gmail 3	nemanja@dataandmore.com	Unknown	9 (1,33%)
All Gmail 3	stjepan@dataandmore.com	Unknown	8 (0,23%)
All Gmail 3	abs@dataandmore.com	Unknown	7 (0,49%)
All Gmail 3	dm@dataandmore.com	Unknown	7 (2,44%)
All Gmail 3	jc@dataandmore.com	Unknown	4 (0,11%)
All Gmail 3	benjamin.sokic@dataandmore.com	Unknown	2 (0,15%)

Source	Account	Department	# Docs with privacy data
All Gmail 3	info@dataandmore.com	Unknown	1 (0,56%)
All Gmail 3	jan.ivica@dataandmore.com	Unknown	1 (0,49%)
All Gmail 3	lt@dataandmore.com	Unknown	1 (0,71%)
All Gmail 3	maw@dataandmore.com	Unknown	1 (25,00%)

Table 3: Risk distribution per account

PRIVACY AND SECURITY DATA SUMMARY

AGE OF DATA

The following visual shows the distribution of non-compliant data sets in terms of age since the data was created in the source - be it a mailbox, OneDrive or a fileshare.

We define non-compliant data as data that contains personal sensitive information according to GDPR and which is more than 3 months old. This limit is a practical time limit set by us based on experience to capture non-compliance. You may have personal sensitive data in your mail or OneDrive where there is a valid reason for storing it. In most cases such data should be kept in a structured data store such as an HR-system rather than kept in e-mails or OneDrive.

The distribution of document age can be consolidated in an average age of the non-compliant data.

The average age of the non-compliant data scanned for this PoC is: **2 years, 9 months, 7 days** .

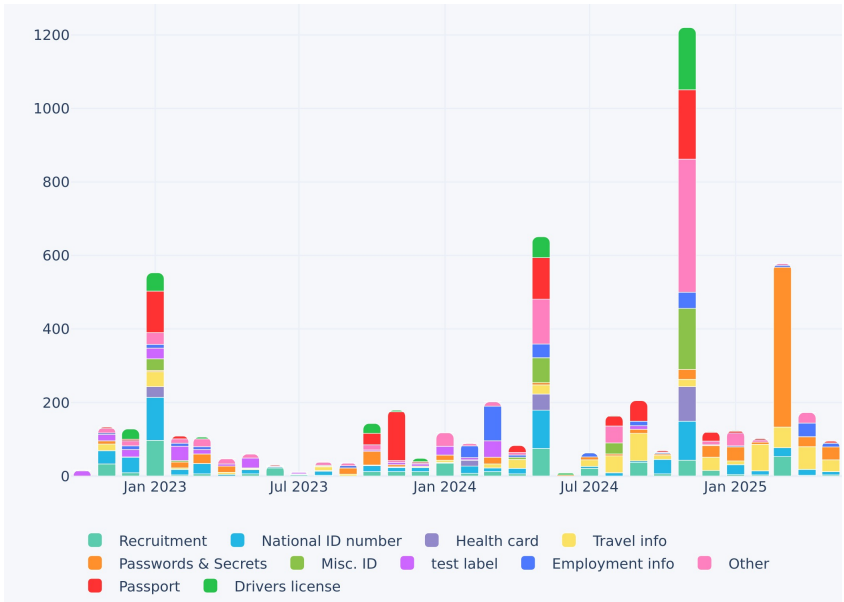

Figure 3: Distribution of age of the data scanned with privacy and security data.

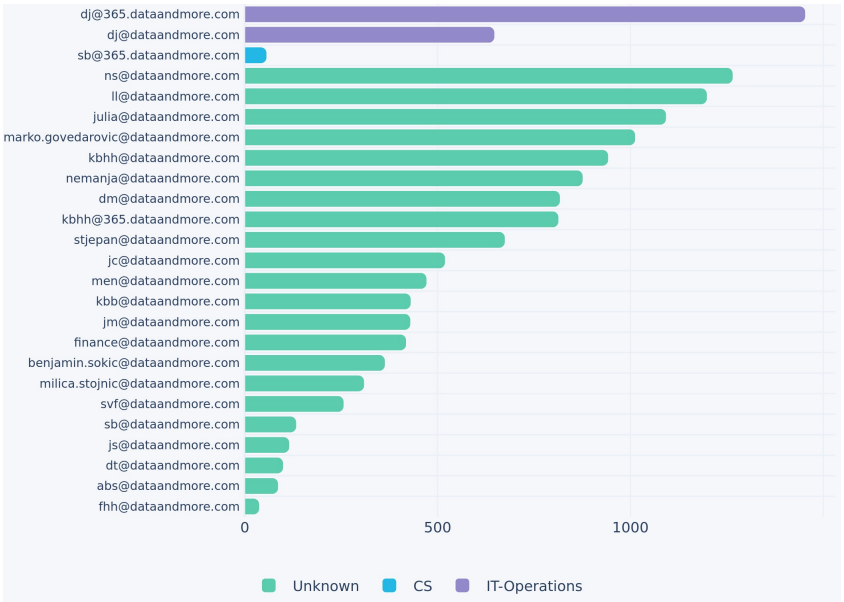

Figure 4: Average age of non-compliant data by account scanned for the PoC.

TRENDS IN DATA PRODUCTION

The following sections analyzes how the creation of data is evolving. It looks at both the trends in data created by day on an overall basis as well as for personal sensitive data. This gives a clear indication of whether your issue with managing personal sensitive data is increasing or decreasing. The graph depicts the data created that was scanned as well as a forecast of how it will evolve given the current trends.

Overall the current trend in the production of data suggests that the amount of data will grow by **25%** annually.

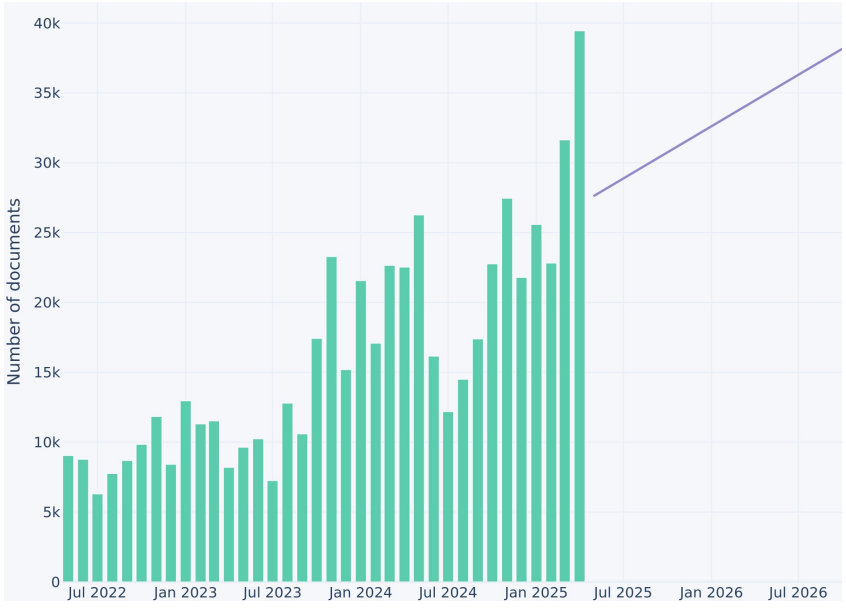

Figure 5: History and trend in the production of data in the scanned accounts.

The graph illustrates the number of emails and documents created per month in Outlook, spanning from July 2022 to July 2026. A notable upward trend is evident, with a significant increase in document creation over time. The data exhibits a steady growth pattern, with some fluctuations. Notably, there are no outliers or anomalies present in the dataset. The linear regression line indicates that this trend will continue into the future, suggesting an ongoing increase in document creation. This analysis provides valuable insights for organizations seeking to manage their personal sensitive data effectively and ensure compliance with GDPR regulations.

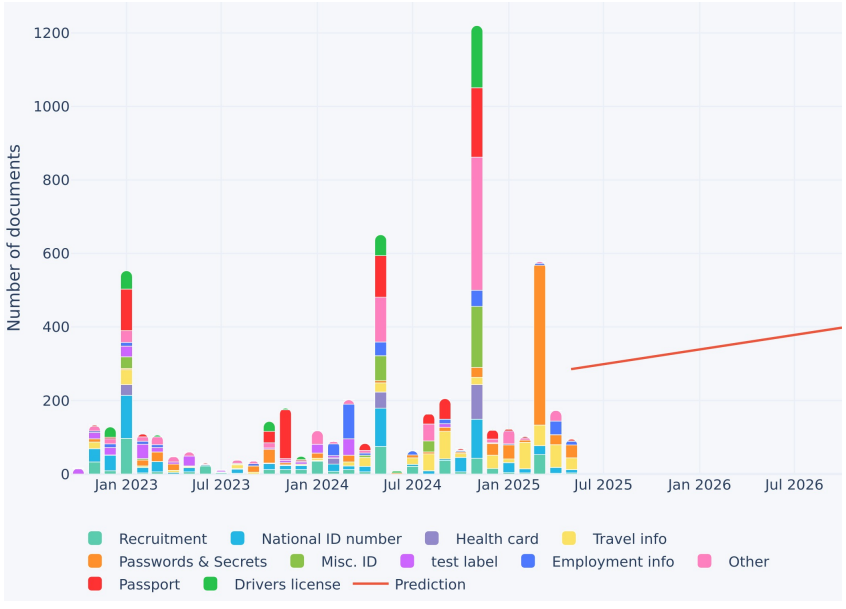

Figure 6: History and trend in the production of non-compliant data in the scanned accounts.

The graph illustrates the number of personal sensitive data created each month in an organization, with a clear upward trend over time. The linear regression line indicates that the rate of data creation has been steadily increasing since January 2023, with a notable spike in July 2024. This suggests that the organization's data management practices may not be keeping pace with the growing volume of sensitive information being generated. To mitigate this risk, it is essential to implement effective data protection measures and ensure compliance with GDPR regulations.

PRIVACY AND SECURITY DATA ANALYSIS

DATA PER ACCOUNT

GDPR does not specify a generic grace period but states that personal data should be deleted when there is not a valid reason to keep it. For practical purposes, Data&More recommends a 3-month grace period. This means that personal data that has been residing in e.g. a users mailbox or a fileshare for three months, should be considered non-compliant by default.

Our average scanned account holds around 247,69 data sets of non-compliant data with the majority sitting in mailboxes.

For the accounts scanned for the PoC we observed the following average number of non-compliant data sets:

Munster Inc. Metrics	Value
Non-compliant data per account	248
Total data per account	22.205
% of Data that is noncompliant	1,12%

Global comparisons:

Global Averages	Value
Non-compliant data per account	336 data sets
Total data per account	41.474
% of Data that is noncompliant	0,75%

TYPES OF PRIVACY DATA AT RISK

Our privacy classification consists of 21 different classes of documents that are considered important according to the GDPR and assessed as important from a security perspective.

The visual below defines the most prevalent classes of privacy and security documents in your organization.

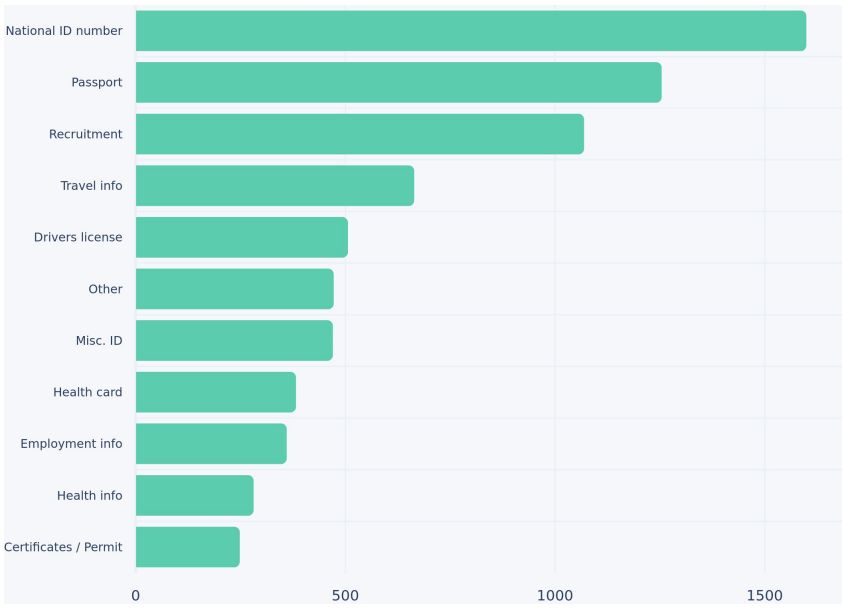

Figure 7: Privacy data by document class (see appendix for definitions).

The graph illustrates the frequency of various document classes, with 'Other' encompassing less frequent categories.

The majority of documents and e-mails are classified as personal sensitive data, categorized by type. This information pertains to a specific organization's personal data compliance context.

Privacy data are typically distributed very unevenly between accounts and the visual below shows the accounts with highest number of privacy data.

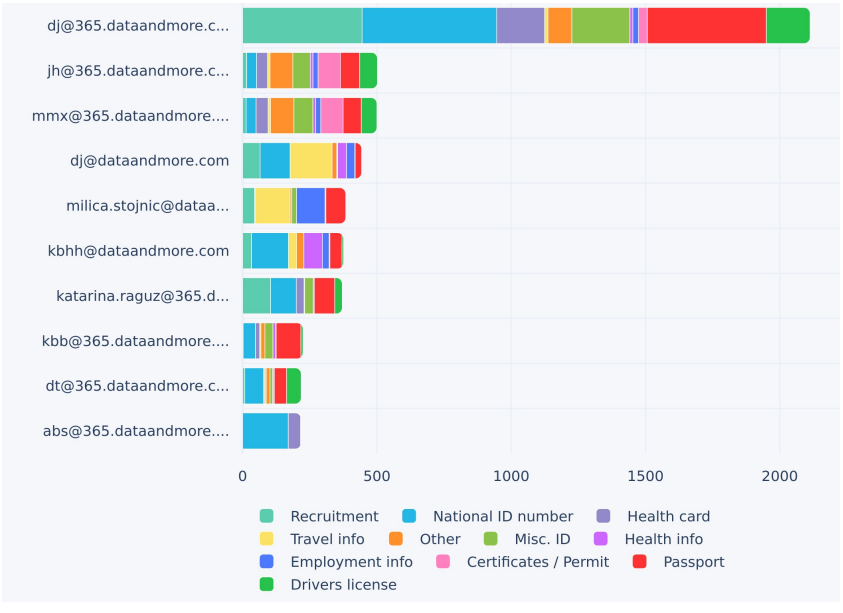

Figure 8: 10 Accounts with the largest amount of non-compliant privacy data.

TYPES OF SECURITY DATA

Our security classification consists of classes of documents that are considered important and potentially problematic from a security perspective.

The visual below defines the most prevalent classes of security documents in your organization.

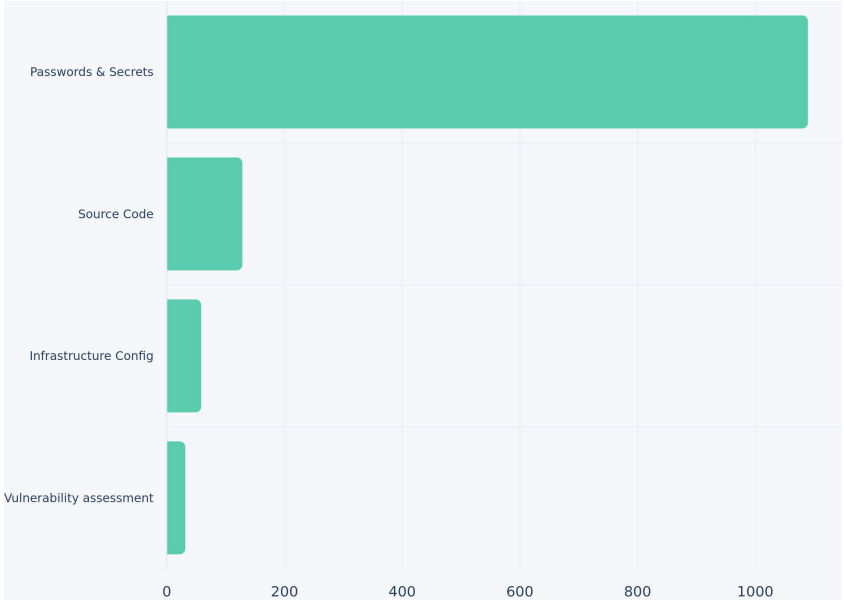

Figure 9: Security data by document class (see appendix for definitions).

The graph displays the frequency of documents and e-mails analyzed by their document security class. The classes are:

Passwords & Secrets: 1000

Source Code: 150

Infrastructure Config: 50

Vulnerability Assessment: 50

The majority of the data is related to passwords and secrets, indicating a high level of security risk in this area.

Security data are typically distributed very unevenly between accounts and the visual below shows the accounts with highest number of security data.

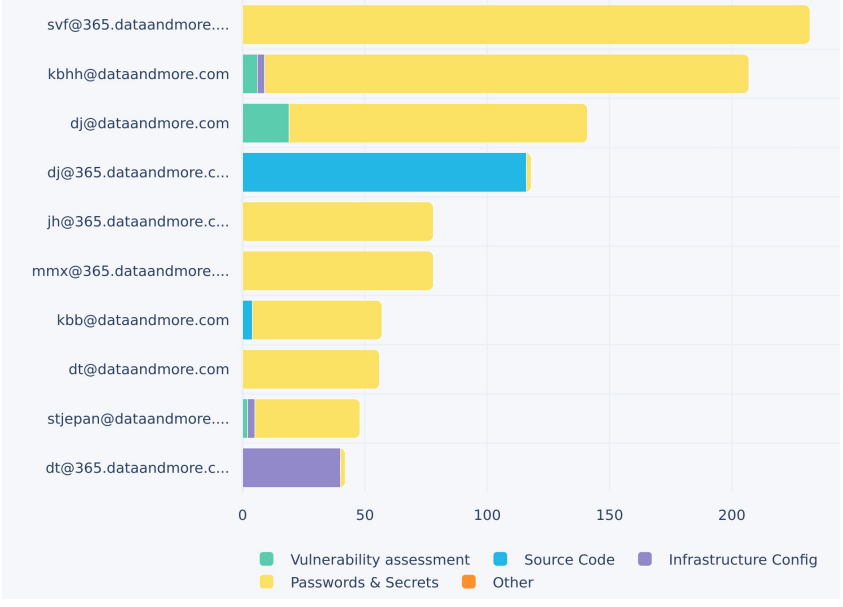

Figure 10: 10 Accounts with the largest amount of non-compliant security data.

FILETYPES WITH NON-COMPLIANT DATA

For your organisation the following filetypes are the ones most often containing personal data that requires management.

The graph shows all data that have been classified as privacy or security data by file type:

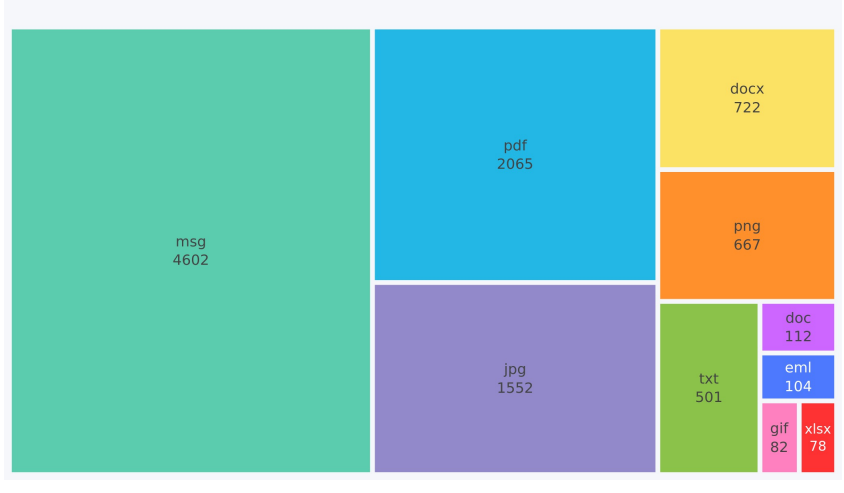

Figure 11: Non-compliant data by filetypes.

The graph illustrates the distribution of classified sensitive data across various file types within an organization, categorized for GDPR compliance purposes. The most prevalent file type is msg (4602), followed by jpg (1552) and pdf (2065). Notably, the majority of files are text-based, with txt (501) being the least common among the top five. This suggests that the organization primarily handles image and document data, with a smaller proportion of text-based files. The graph provides valuable insights into the types of sensitive data handled by the organization, which can inform strategies for ensuring GDPR compliance.

In appendix is a textual definition of the document classes used in our privacy and security classification.

ABOUT DATA & MORE

At Data & More, we protect digital **privacy** and **security** by minimizing the risks, costs, and liabilities associated with processing unstructured data.

We were **founded in 2017**, coinciding with the EU's establishment of the first digital freedom rights in history, known as GDPR. GDPR is based on the simple belief that **Personal Data belongs to the individual** and that companies, organizations, and institutions are not owners, but mere **caretakers of this data.**

We support the EU's idea and intention behind data privacy and individual freedom by creating the world's best tool for identifying and removing data that threatens organizational and individual digital privacy and security.

As responsible caretakers, we must delete data that we no longer have a legitimate interest in, and that has become illegal to retain - we assist organizations in doing just that.

More than 175.000 Data Owners use Data & More to protect the rights of **Data Subjects** in the EU, Canada, and the US. You can meet us in Copenhagen, Stockholm, Calgary, Dusseldorf, and Zagreb or visit us at dataandmore.com

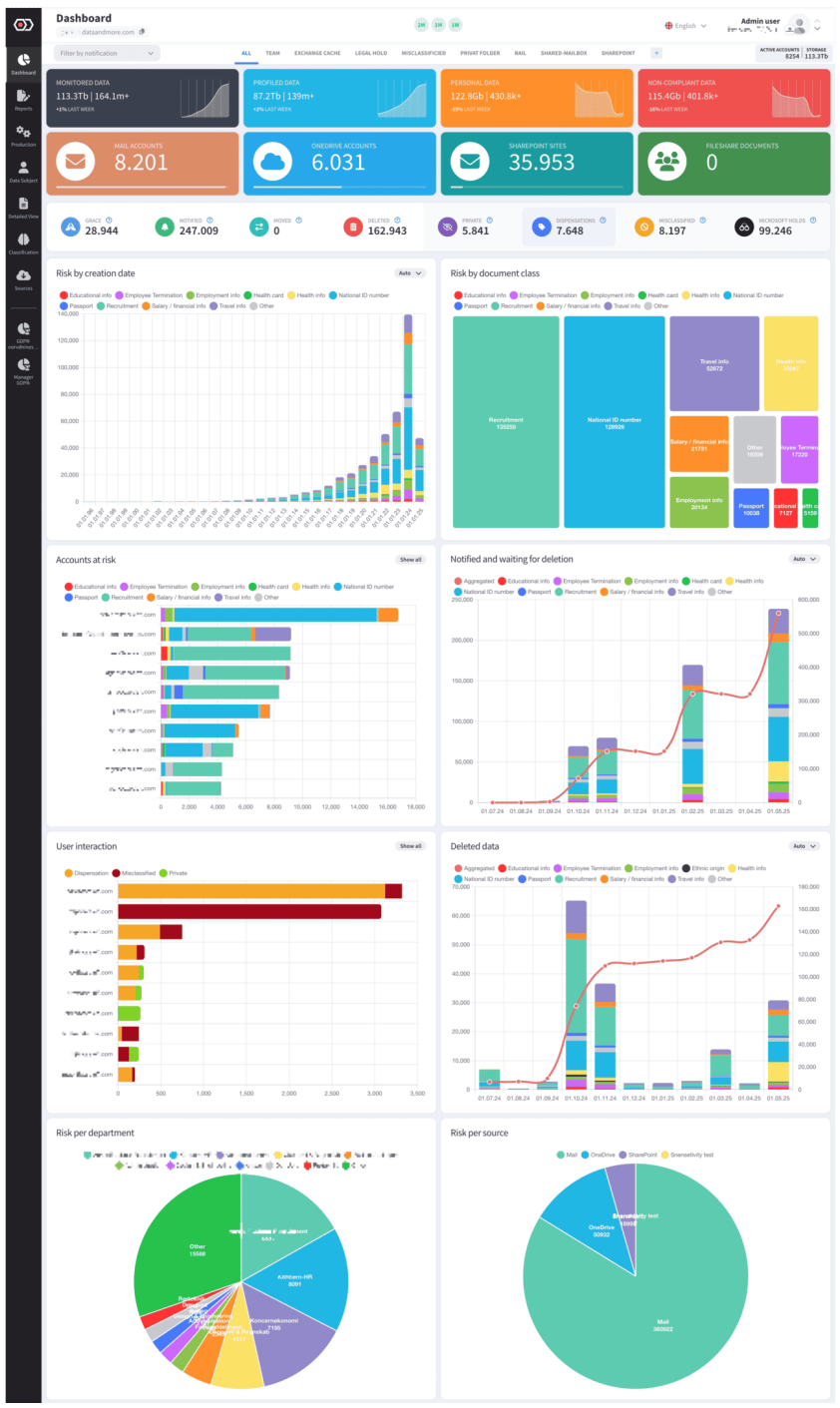

Figure 12: Data & More | Privacy & Security Monitor

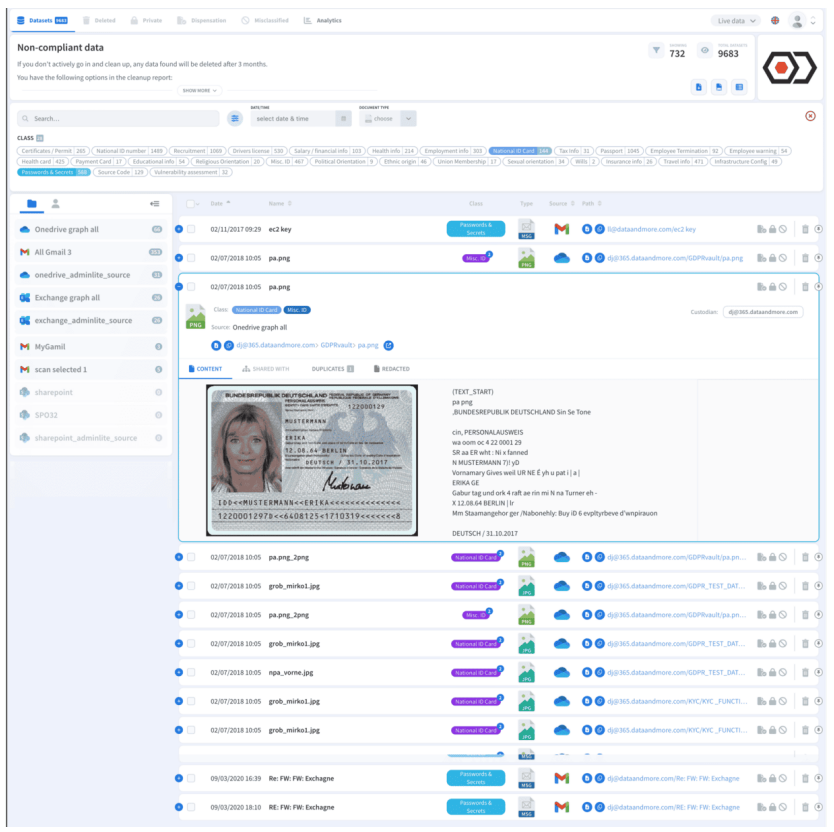

Figure 13: Data & More | Data Minimization Manager

APPENDIX: OUR METHOD

We classify documents with the help of multiple methods and using principles that we have evolved and matured since our inception in 2016.

The task at hand when classifying documents consists of a number of steps. We receive a document from our scanners and based on the contents and metadata in the document, we try to predict whether the document belongs to one of the classes that we have defined. On average, 1 - 1,5% of the documents we scan belong to one of the classes. The remaining 98,5% - 99% are not relevant for our classification. This distribution produces a challenging task for document classification and rules out simple statistical methods based on machine learning. We have broken the classification process down into the following steps: **Data Extraction**, **Text Analysis** and **Classification**.

We use a number of methods for identifying privacy and security data that we have found to be needed in order to ensure the quality of our reports.

ON DATA EXTRACTION:

Capturing file meta data: We capture basic file metadata such as name, type, age, size, creator, etc. that we use for governing the downstream analysis and manage the sendout of our reports.

Classify images: We are using multiple ML based models for identifying various identity cards and to identify images with faces of single persons that are used in id-cards, cv's, etc.

Extract text from images: We use multiple methods for extracting text from images from e.g. id cards.

Extract text from spreadsheets: The tabular data and spreadsheets require a special approach to text extraction to capture relationships e.g. between person names and their id numbers in adjacent columns.

ON TEXT ANALYSIS:

Language recognition: we determine the language of the text as this will guide the downstream analysis of text.

Explicit pattern matching: we apply more than 200 patterns for matching strings with regular expressions. All regular expressions are combined with start- -and stopwords, ensuring that we don't mistake a Danish CPR number for a product number that happens to match the same numbers.

Statistical methods: NER (named entity recognition) is used for identifying persons. We use models trained by third-parties in most European languages.

Syntactic analysis: we use SEMGREX operators to identify sentence patterns which are considered personal sensitive (“My doctor told me that I had pneumonia”, “I was diagnosed with cancer”). This analysis takes place in the most important languages (English, German, French, Spanish, Dutch, Danish, Swedish, Finnish) and is gradually expanded with need.

Identifying False Positives: We adopt the strategy of positively identifying common false positives such as newsletters, templates, notifications, invites, etc. These false positives are used in combination with our positive classification to achieve greater precision.

ON CLASSIFICATION:

Our classification system uses the metadata produced from data extraction and text analysis to identify different privacy classes that are important for GDPR compliance. In the appendix are our current list of document classes for data privacy. The classification system uses Boolean logic to build the classes from available metadata. That means we combine different metadata to predict a document classification.

We have drawn up the entire process in the following classification process overview:

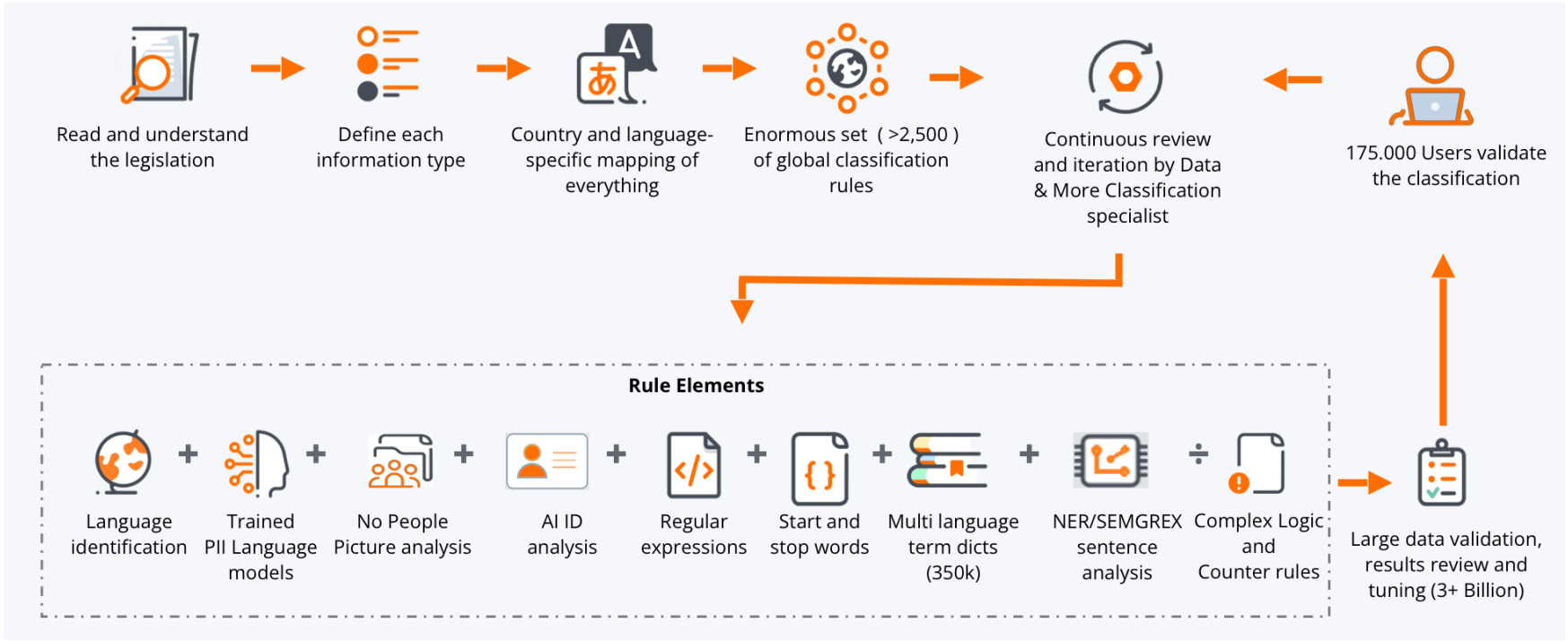

Figure 14: Classification process overview

In order to provide an impression of the complexity of the classification task, we have built a network graph of all the logical elements in the classification, from regular expressions and other labels to the document classes that we show to the user. This graph consists of 822 nodes and 1.528 edges. In the following we present two details from this graph that gives an impression of the complexity.

The passport class and Health info classes are illustrated in the following network visualization of all classification elements that are used in the two document classes.

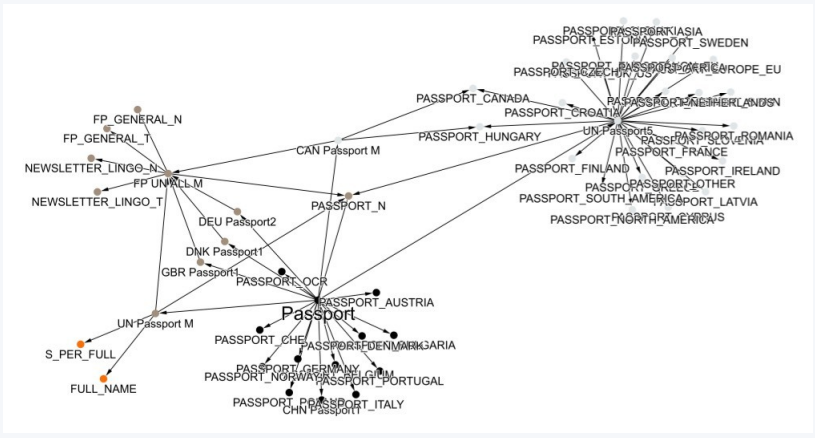

Figure 15: Passport class overview

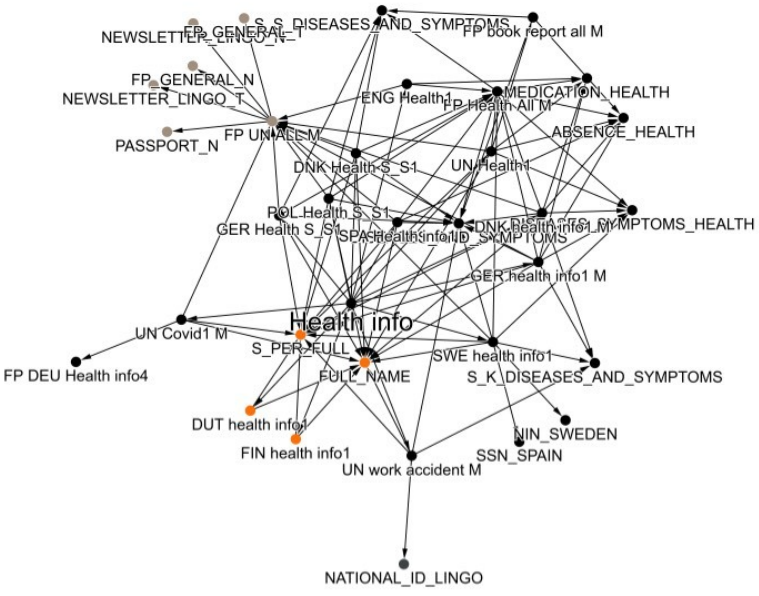

Figure 16: Health Info class overview

The way we define our classes differs with what we are looking for. At one end of the spectrum, the passport class is based on scanned passports using our specialized ocr-engine. We combine that with patternmatching to identify the country of the passport and these metadata are used directly to predict the class.

At the other end of the spectrum we have a document class like “Health Info”. This class picks up health information about natural persons in any kind of context such as e-mails, case files, case worker notes, etc. This class is mostly based on syntactic analysis and phrase libraries.

Our classes are multilingual and work across jurisdictions meaning that a German and a Danish passport belong to the same passport class, enabling us to easily support global multilingual companies.

APPENDIX: DATA PRIVACY AND SECURITY DOCUMENT CLASSES

Our privacy document classes are built on the basis of guidance from authorities on which types of data are important to manage from a GDPR. This includes types of person-identifiable information such as personal numbers on identity cards, as well as documents that convey sensitive information about a natural person. Our classification includes a class for each of the five sensitive areas defined in GDPR: health data, sexual orientation, political orientation, religous orientation, ethnic origin, union membership.

Name	Description of privacy document class
Payment card	Data containing information about a person's credit card. This class is based on explicit pattern matching of credit card numbers whether present on scanned images of cards or credit card numbers in receipts, order, etc. This includes the most common credit cards used globally.
Criminal record	Criminal records on a natural person. This class is predicted on the basis of specific phrases used in criminal records from mutiple countries.
Misc. ID	Various ID documents that we do not have detailed information about. This class picks up ID cards with natural persons that our statistical methods have identified but where we do not have specific information about e.g. the number on the ID card.
Driver's license	Data on drivers licenses that belong to a person. This class uses a combination of statistical methods for identifying scanned drivers licenses, where face, name and number is present. It combines this with explicit pattern matching to also include e.g. car rental agreements where the number of the driver's license of the user is revealed.
Ethnic origin	Predicts information about the ethnic orientation of one or more persons. This class is based on syntactic analysis and phrase libraries to capture this information in any context.
Grant application	Personally identifiable data that appears in applications to foundations for financial support. This class is based on phrase libraries and statistical analysis and explicit pattern matching to find data on natural persons that are applying for or receiving grants.
Health card	Health cards used in different countries as ID when receiving healthcare. This class is primarily based on our image classification and explicit pattern matching so that we predict scanned images of health cards. The class also finds mentions of health card numbers where these are used in conjunction with a natural person such as a case file or e-mail.
Health info	Predicts information related to the health of one or more natural persons. This class is based on syntactic analysis and phrase libraries to capture this information in any context. This includes information about conditions, symptoms, diagnosis and medication associated with a person.
Union membership	Predicts information about the ethnic orientation of one or more persons. This class is based on syntactic analysis and phrase libraries to capture this information in any context. It includes libraries of unions in most EU countries.
National ID number	Captures when National ID numbers appears in a document. This class is primarily based on explicit pattern matching found in multiple kinds of documents such as e-mails, forms, etc.
National ID card	Data for ID cards belonging to identifiable persons from different countries. This class is using the image classification and data extraction from ocr to determine whether a document is a scanned ID card from any of the countries we cover.
Passport	Documents that contain scanned passports or personal passport information. This class is using the image classification and data extraction from ocr to determine whether a document is a scanned passport from any of the countries we cover. Passports from all nationalities are predicted.
Political orientation	Predicts information about the political orientation of one or more persons. This class is based on syntactic analysis and phrase libraries to capture this information in any context. It includes libraries of political parties and movements in most EU countries.
Recruitment	

Name	Description of privacy document class
	Personal data that appear in solicited or unsolicited applications, as well as in CVs as well as rejections of job applications. This class uses statistical methods for identifying natural persons with phrase libraries for wording that appear in e-mails, notes, documents, etc.
Religious orientation	Predicts information about the religious orientation of one or more persons. This class is based on syntactic analysis and phrase libraries to capture this information in any context. It includes libraries of churches and other religious movements in most EU countries.
Salary / financial info	Data that contain information about a person's salary, for example payslips and fee papers. Also information about bonus schemes is searched for. This class uses statistical methods for identifying natural persons with phrase libraries for wording that appear in e-mails, notes, documents, etc. It is also one of the classes relying on false positives to exclude commercial contracts between legal persons.
Sexual orientation	Predicts information about the sexual orientation of one or more persons. This class is based on syntactic analysis and phrase libraries to capture this information in any context. It includes libraries of sexual orientations in most EU countries.
Tax info	Data that contain information about a person's tax information, especially in the form of annual statements. This class combines phrase libraries with explicit pattern matching to identify tax statements that are mostly in pdf format and this based on ocr.
Employee termination	Dataset with information about the termination of an employee's employment within an organization, including resignations, departures, layoffs, and more. This class uses statistical methods for identifying natural persons with phrase libraries for wording that appear in e-mails, notes, documents, etc.
Employment info	Data for employment agreements between employee and employer, whether the terms are described in documents or in a written communication. This class uses statistical methods for identifying natural persons with phrase libraries for wording that appear in e-mails, notes, documents, etc. It is also one of the classes relying on false positives to exclude commercial contracts between legal persons.
Travel info	Data that contain information about a person travelling at specific times, such as hotel, airline and restaurant bookings. This class uses statistical methods for identifying natural persons with phrase libraries for wording that appear in e-mails, notes, documents, etc.
Employee warning	Data concerning internal warnings to one or more individuals due to actions that violate the specific organization's guidelines. his class uses statistical methods for identifying natural persons with phrase libraries for wording that appear in e-mails, notes, documents, etc.
Wills	Wills from natural persons. This class uses statistical methods for identifying natural persons with phrase libraries for wording that appear in e-mails, notes, documents, etc.
Certificates / Permits	Each country issues a variety of official documents for purposes such as naming, marriage, birth, partnerships, and more. These documents are unique in both their type and name, serving as essential legal records for individuals within each respective nation. This class uses statistical methods for identifying natural persons with phrase libraries for specific phrases found in these official documents that follow a fixed format per country.
Work absence	Data concerning cases where an employee fails to work on scheduled days. This class uses statistical methods for identifying natural persons with phrase libraries for specific phrases.
Referral consent	Data related to personal consent, where a person gives consent for their personal information to be shared with an organization or similar types of consent. This class uses statistical methods for identifying natural persons with phrase libraries for specific phrases.
Insurance info	Data for insurance documents that describe how one or more persons are insured, such as home insurance policies and accident insurance policies. This class uses statistical methods for identifying natural persons with phrase libraries for specific phrases.

Our security classes have been pragmatically created on the basis of their potential implications from a security standpoint. We are gradually building out the security classification in an ongoing dialogue with security professionals from our customers.

Table 4: Privacy document classes

Name	Description of security document class
Passwords & Secrets	Passwords and login information for and-user access to systems as well as keys used for encryption of communication and for machine-to-machine communication.
Source code	Data that expose secrets and other information that can potentially help malicious actors get access to systems and data.
Log files	Log files from application systems or servers.
Infrastructure config	Various infrastructure configuration information, including infrastructure automation such as Ansible scripts.
Vulnerability Assessments	Documents assessing security of infrastructure and applications including assessing CVE vulnerabilities and results from penetration-testing.

Table 5: Security document classes