Out of the Breach and Into The Fire!

The Importance of Discovery Intelligence in Sensitive Data Discovery

Educational Webcast

1 | 12 | 22

HAYSTACK®

Michael Sarlo

Chief Innovation Officer, President of Global Investigations & Cyber Discovery Services for HaystackID

As Chief Innovation Officer and President of Global Investigations and Cyber Discovery Services for HaystackID, Mike leads worldwide innovation and service development efforts to support the challenges of cyber, data, and legal discovery.

John Brewer

Chief Data Scientist for HaystackID

As Chief Data Scientist, John serves as the Head of Advanced Technology Services for HaystackID and constructs data science approaches and frameworks to help solve complex discovery and analysis challenges for cybersecurity, information governance, and eDiscovery experts supporting cyber, data, and legal discovery needs.

Anya Korolyov

Vice President of Cyber Discovery Services and Custom Solutions for HaystackID

As Vice President of Cyber Discovery Services and Custom Solutions for HaystackID, Anya defines, develops, and deploys artificial intelligence, data science, machine learning technology processes, and workstream protocols to solve cybersecurity, information governance, and eDiscovery challenges.

Matthew Miller

Senior Vice President of Information Governance & Data Privacy for HaystackID

As Senior Vice President of Information Governance and Data Privacy for HaystackID, Matthew is an expert in assessing, evaluating, and supporting privacy, compliance, and governance solutions for corporations and governments worldwide.

Agenda

1. The Challenge of Sensitive Data
2. The Importance of Discovery Intelligence
3. Detecting Sensitive Data Presence and Scope
4. Identifying Breach or Disclosure Impact and Options
5. Analyzing Individuals and Entities of Breach or Disclosure
6. Notifying Individuals and Entities of Breach or Disclosure
7. A Synergistic Approach to Sensitive Data Discovery

The Challenge of Sensitive Data

Does your Organization

Know all data it retains and processes?

Know where all data is located and stored?

Have visibility into all metadata and content?

Control who can and should have access to which data?

Know how to retrieve data rapidly when needed?

Only retain data as long as necessary?

Dispose, deidentify, or encrypt data regularly?

Processing PII *= creation, collection, use, processing, storage, maintenance, dissemination, disclosure, or disposal (NIST.SP.800-53r5, PT-3, Sept. 2020)*

Confronting Your Data Reality…

Unstructured Data – Consists of data that resides on devices that are in the direct control of a custodian or centrally located and managed by IT

Examples:

- Desktops/Laptops
- PDA's
- Cell phones
- Printers
- Paper
- CD/DVD
- Thumb drive
- MS O365

- Network File Shares
- E-mail Servers
- Backup tapes
- Web Servers
- Content Management Servers
- Archives
- Voice Mail
- Video

Structured Data – Consists of data that resides in a structured table format and is often dynamic in nature

Examples:

- SQL On-premises
- Azure Data Lake
- SAP ERP
- Oracle Data Warehouse

SaaS – Consists of data that resides in 3rd party hosted solutions

Examples:

- Microsoft Dynamics 365
- Salesforce
- ServiceNow
- Workday

Approved File Hosting Services or Shadow IT – Consists of enterprise departments or personnel conducting their own tech initiatives without the knowledge of the actual IT department or where the IT Security team is not part of the vetting/approval process

Examples:

- DropBox
- Box
- Google Drive
- Microsoft OneDrive

- PC Review States: 90% of all data in existence was created in the past 2 years
- 451 Research Estimates: In 2020, 90% of all data generated will be unstructured and more challenging to analyze because it has no predefined format or organization

Data Privacy & Cybersecurity Obligations

Beginning in 2018 with **GDPR superseding the EU Data Protection Directive**, and the continued introduction of new **Data Privacy laws and regulations** both internationally and in the US, have **penalties** related to PII/PHI/PCI cyber-incidents, in addition to **notification and disclosure requirements**, and potential **subsequent civil litigation.**

The Sedona Conference, Commentary on a Reasonable Security Test, 22 Sedona Conf. J. 345 (forthcoming 2021) posits when a data breach has occurred, **did the organization satisfy their legal obligation to provide "reasonable security" for personal information?**

- Evidence of noncompliance with a statute, regulation, or ordinance, or an "industry custom" that required specific security controls for PII, both establish that security for that personal information was not reasonable.

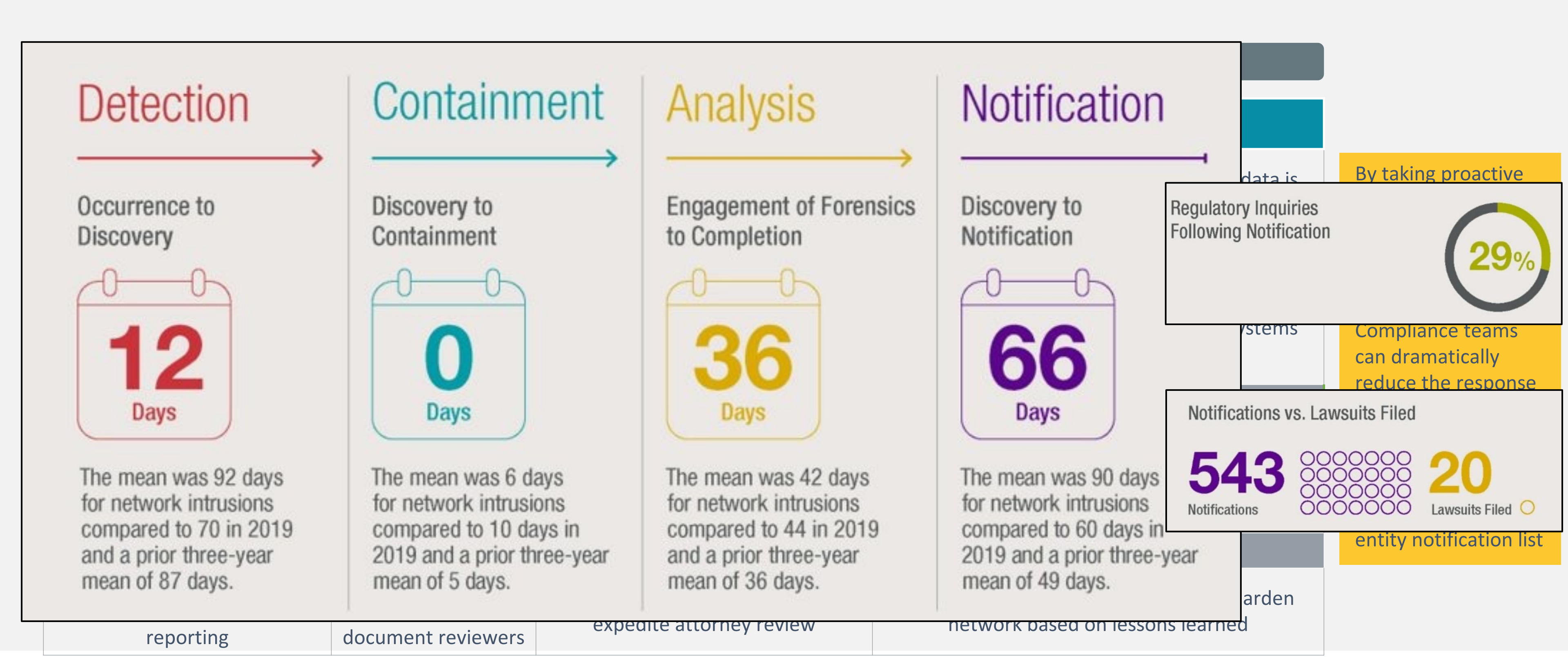
CCPA – CALIFORNIA
CPRA – CALIFORNIA
CDPA – VIRGINIA
CPA – COLORADO
Many more on the horizon….

Canada
Data Minimization Standards

GDPR and **UK GDPR**
Data Minimization Standards

HAYSTACK

Example of a Cyber Incident Response Remediation Timeline



Detection

Occurrence to Discovery

12 Days

The mean was 92 days for network intrusions compared to 70 in 2019 and a prior three-year mean of 87 days.

Containment

Discovery to Containment

0 Days

The mean was 6 days for network intrusions compared to 10 days in 2019 and a prior three-year mean of 5 days.

Analysis

Engagement of Forensics to Completion

36 Days

The mean was 42 days for network intrusions compared to 44 in 2019 and a prior three-year mean of 36 days.

Notification

Discovery to Notification

66 Days

The mean was 90 days for network intrusions compared to 60 days in 2019 and a prior three-year mean of 49 days.

Regulatory Inquiries Following Notification

29%

Notifications vs. Lawsuits Filed

543 Notifications

20 Lawsuits Filed

The Importance of Discovery Intelligence

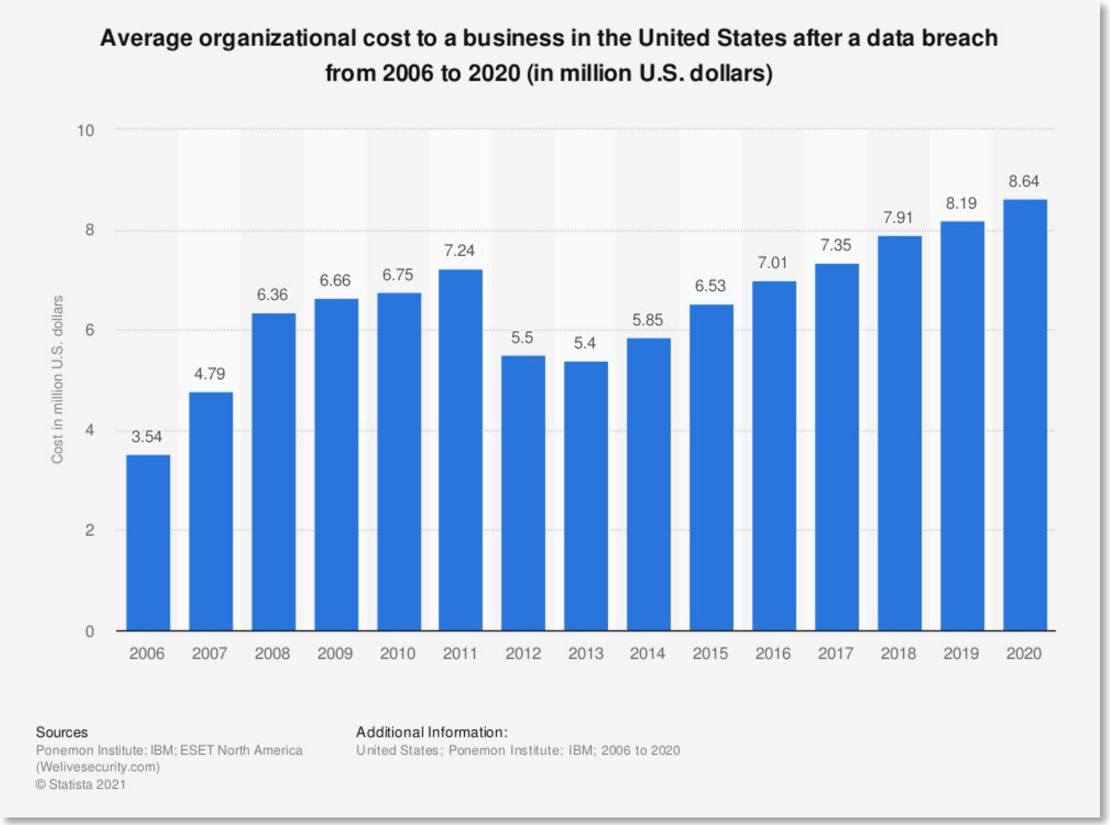
The Average Cost of a Breach

The average cost per data breach:

- United States = **$8.64 million** USD in 2020
  - → **$9.05 million** USD in 2021
- Globally = **$3.86 million** USD in 2020
  - → **$4.24 million** USD in 2021

Total breach costs include:

- **Lost business** resulting from diminished trust or confidence of customers
- Costs related to **detection, escalation, and notification** of the breach
- **Post response activities**, such as credit report monitoring



Average organizational cost to a business in the United States after a data breach from 2006 to 2020 (in million U.S. dollars)

Sources
Ponemon Institute; IBM; ESET North America
(Welivesecurity.com)
© Statista 2021

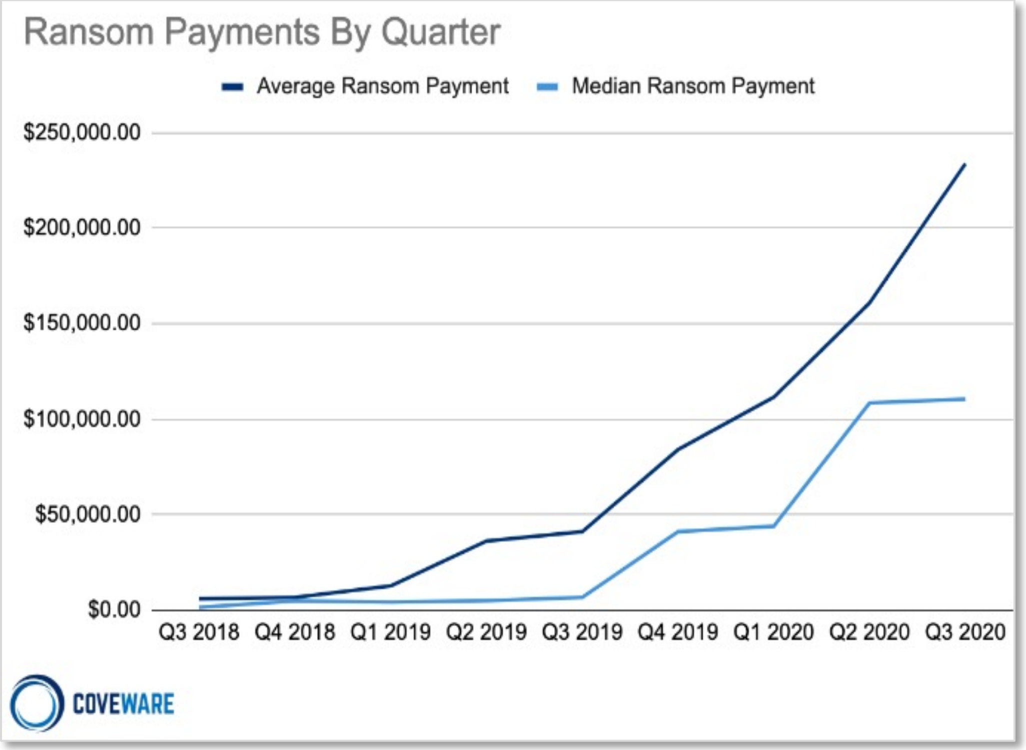
Additional Information:
United States; Ponemon Institute; IBM; 2006 to 2020

Average Ransom Payment Sizing

Cybercriminals continue to be less interested in stealing consumers' personal information.

Ransomware and phishing attacks directed at organizations are now the preferred method of data theft by cyberthieves:

- Require less effort
- Largely automated
- Generate much higher payouts
  - > **$233,000** per event in Q4 2020



In 2020, ransomware attacks increased:

- **471%** in the U.K.
- **150%** in Australia
- **75%** in Singapore
- **70%** in the U.S.

In 2020, phishing was the most favored type of attack globally:

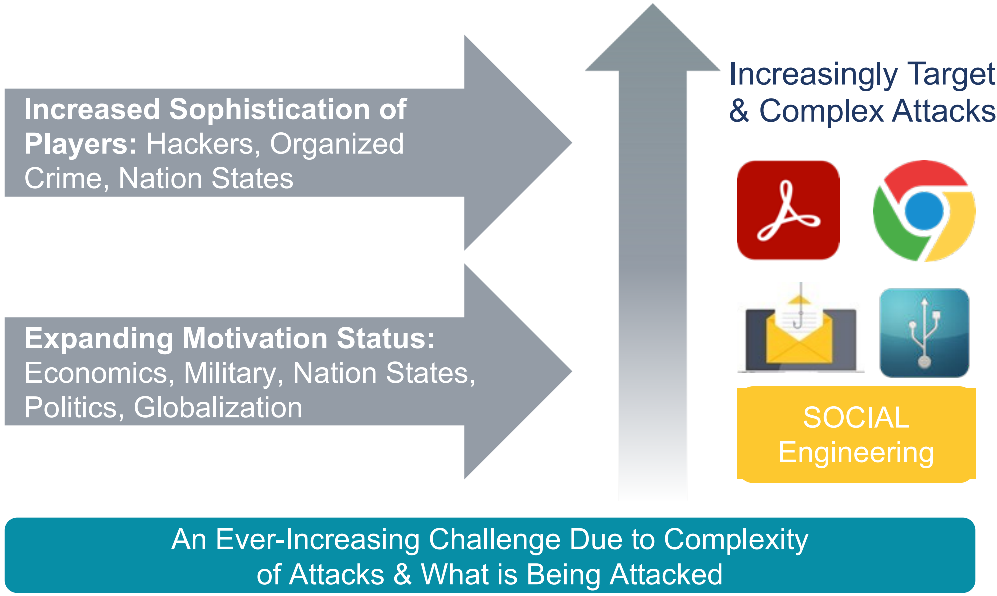
- **25%** of U.S. attacks
- **36%** in Australia
- **75%** increase in Singapore
- Most common data breach type in the U.K. & Germany

Cybersecurity Lessons Learned Since COVID-19

Proactive, Defensive Technology Measures

- Back-up, Disaster Recovery, Resiliency Planning
- Identify and classify sensitive data (PII/PHI/PCI) to enforce data protection and remediation
- Updates and patches
- Only use licensed software
- Only use WiFi networks that are password protected
- 2FA, MFA, VPN
- Employee Training – Please Don't Click
- Add the Phish Alert Report plug-in to M365

In 2020, Cybercriminals preyed on consumers with false information about the COVID-19 pandemic, stimulus payments, and lockdowns via sophisticated spear phishing attacks.



Increased Sophistication of Players: Hackers, Organized Crime, Nation States

Expanding Motivation Status: Economics, Military, Nation States, Politics, Globalization

Increasingly Targeted & Complex Attacks

SOCIAL Engineering

An Ever-Increasing Challenge Due to Complexity of Attacks & What is Being Attacked

Detecting Sensitive Data Presence and Scope

Multiple Approaches

We use a variety of search tools to find sensitive breach information including:

Off-the-Shelf NLP Models

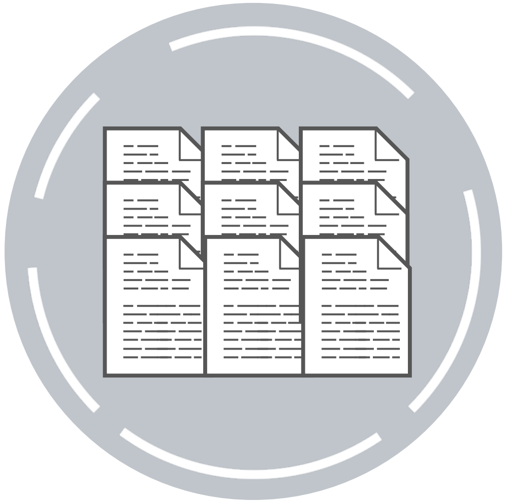
Regular Expressions

Internal Machine Learning

PII/PHI/Entity Density

Comparative analysis of data and entities allows us to break the workstream into two broad categories of documents for review.

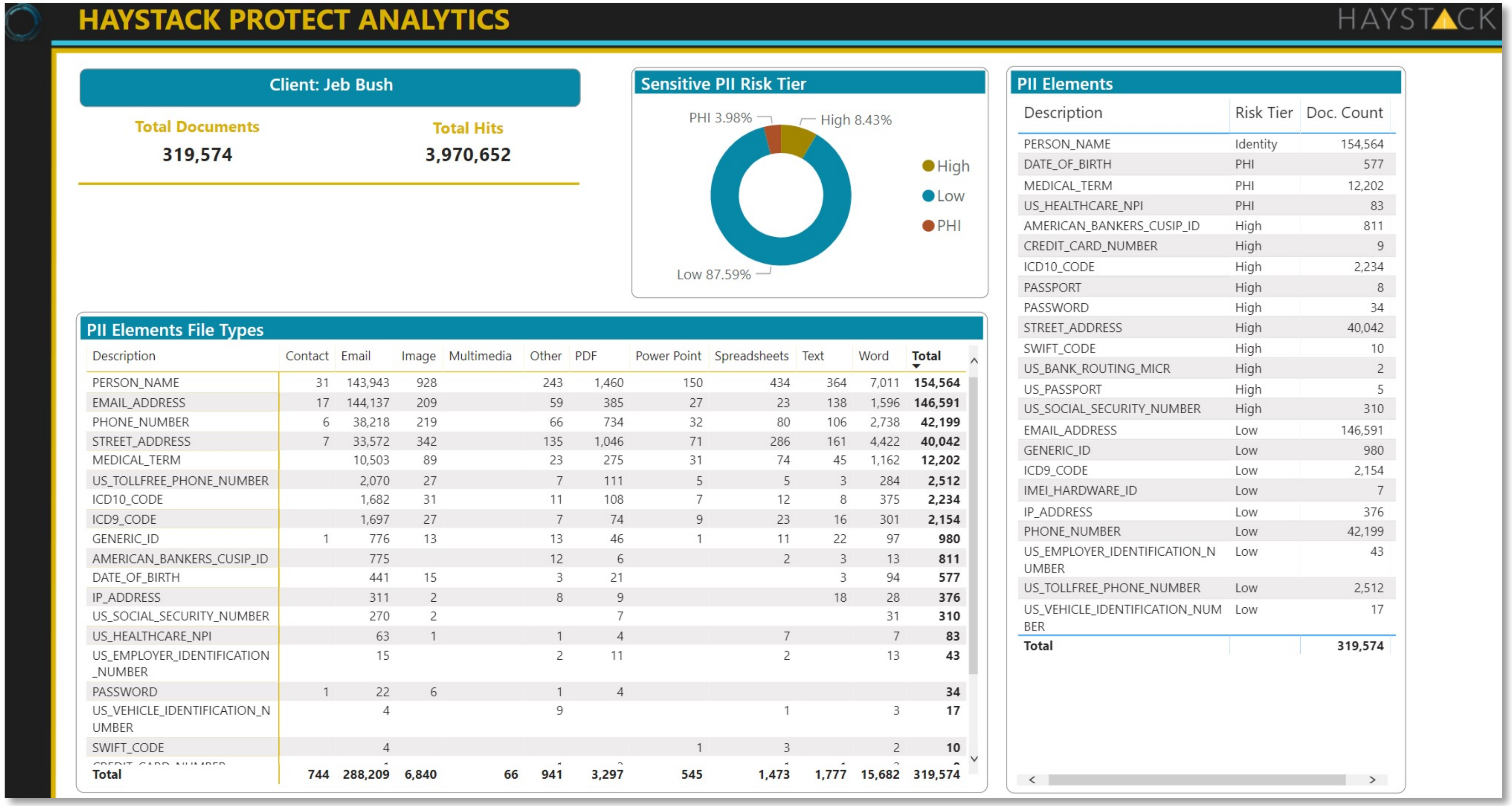
Sparse Documents



Dense Documents

ReviewRight® Protect

AI Model Output & Validation



HAYSTACK PROTECT ANALYTICS

HAYSTACK

Client: Jeb Bush

Total Documents	Total Hits
319,574	3,970,652

Sensitive PII Risk Tier

PII Elements File Types

Description	Contact	Email	Image	Multimedia	Other	PDF	Power Point	Spreadsheets	Text	Word	Total
PERSON_NAME	31	143,943	928		243	1,460	150	434	364	7,011	**154,564**
EMAIL_ADDRESS	17	144,137	209		59	385	27	23	138	1,596	**146,591**
PHONE_NUMBER	6	38,218	219		66	734	32	80	106	2,738	**42,199**
STREET_ADDRESS	7	33,572	342		135	1,046	71	286	161	4,422	**40,042**
MEDICAL_TERM		10,503	89		23	275	31	74	45	1,162	**12,202**
US_TOLLFREE_PHONE_NUMBER		2,070	27		7	111	5	5	3	284	**2,512**
ICD10_CODE		1,682	31		11	108	7	12	8	375	**2,234**
ICD9_CODE		1,697	27		7	74	9	23	16	301	**2,154**
GENERIC_ID	1	776	13		13	46	1	11	22	97	**980**
AMERICAN_BANKERS_CUSIP_ID		775			12	6		2	3	13	**811**
DATE_OF_BIRTH		441	15		3	21			3	94	**577**
IP_ADDRESS		311	2		8	9			18	28	**376**
US_SOCIAL_SECURITY_NUMBER		270	2			7				31	**310**
US_HEALTHCARE_NPI		63	1		1	4		7		7	**83**
US_EMPLOYER_IDENTIFICATION_NUMBER		15			2	11		2		13	**43**
PASSWORD	1	22	6		1	4					**34**
US_VEHICLE_IDENTIFICATION_NUMBER		4			9			1		3	**17**
SWIFT_CODE		4					1	3		2	**10**
Total	**744**	**288,209**	**6,840**	**66**	**941**	**3,297**	**545**	**1,473**	**1,777**	**15,682**	**319,574**

PII Elements

Description	Risk Tier	Doc. Count
PERSON_NAME	Identity	154,564
DATE_OF_BIRTH	PHI	577
MEDICAL_TERM	PHI	12,202
US_HEALTHCARE_NPI	PHI	83
AMERICAN_BANKERS_CUSIP_ID	High	811
CREDIT_CARD_NUMBER	High	9
ICD10_CODE	High	2,234
PASSPORT	High	8
PASSWORD	High	34
STREET_ADDRESS	High	40,042
SWIFT_CODE	High	10
US_BANK_ROUTING_MICR	High	2
US_PASSPORT	High	5
US_SOCIAL_SECURITY_NUMBER	High	310
EMAIL_ADDRESS	Low	146,591
GENERIC_ID	Low	980
ICD9_CODE	Low	2,154
IMEI_HARDWARE_ID	Low	7
IP_ADDRESS	Low	376
PHONE_NUMBER	Low	42,199
US_EMPLOYER_IDENTIFICATION_NUMBER	Low	43
US_TOLLFREE_PHONE_NUMBER	Low	2,512
US_VEHICLE_IDENTIFICATION_NUMBER	Low	17
Total		**319,574**

Identifying Breach or Disclosure Impact and Options

Identifying and Protecting the “Crown Jewel” Data is of High Value to the Organization

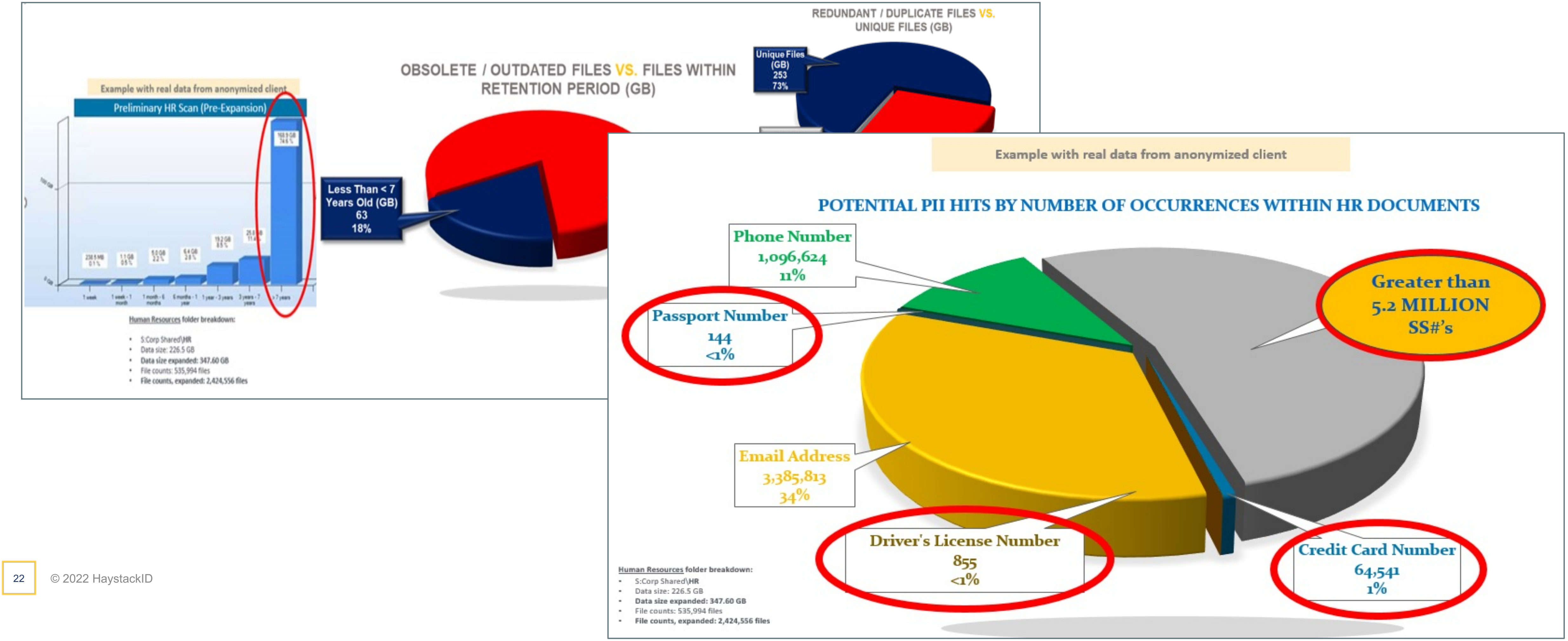
By: IBM Security Board of Advisors

Data Value

Data Type	Examples		% of Sensitive Data
Enterprise Critical	• Critical Intellectual Property	• Top-Secret Plans and Formulas	0.01-0.1%
Executive	• Acquisition and Divestiture Plans	• Executives and Board Deliberations	0.1 - 2%
Regulated	• SPI & PII • Sarbanes-Oxley	• HIPAA • ITAR • Quarterly results	1-50%
Business Strategic	• External Audit Results	• Alliance & Joint Venture, Partner Data • Business Strategic Plans	1-5%
Business Unit Critical	• Design Documents • R & D Results	• Customer records • Pricing Data • Security Data	10-20%
Operational	• Project Plans • Contracts	• Salaries & Benefits Data • Accounts Receivable	20-80%
Near-Public	• List of Partners • Revenue Growth	• Market Intelligence • Pay Compensation Data	10-80%

Operationalize Policies

Implement Defensible Disposition & Remediation – Use Case



Example with real data from anonymized client

Preliminary HR Scan (Pre-Expansion)

Human Resources folder breakdown:

- S:Corp Shared\HR
- Data size: 226.5 GB
- **Data size expanded: 347.60 GB**
- File counts: 535,994 files
- **File counts, expanded: 2,424,556 files**

OBSOLETE / OUTDATED FILES VS. FILES WITHIN RETENTION PERIOD (GB)

REDUNDANT / DUPLICATE FILES VS. UNIQUE FILES (GB)

Example with real data from anonymized client

POTENTIAL PII HITS BY NUMBER OF OCCURRENCES WITHIN HR DOCUMENTS

Human Resources folder breakdown:

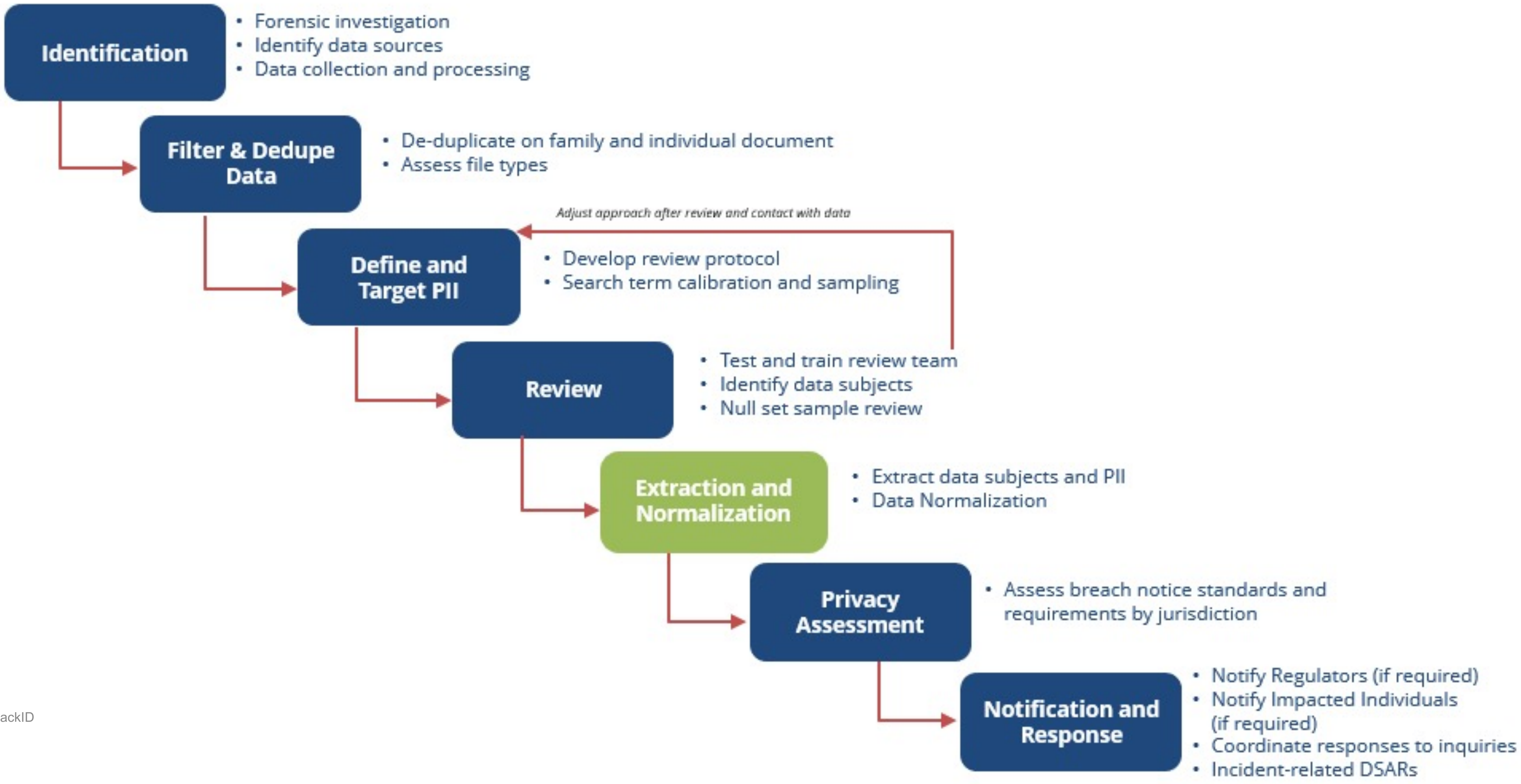
- S:Corp Shared\HR
- Data size: 226.5 GB
- **Data size expanded: 347.60 GB**
- File counts: 535,994 files
- **File counts, expanded: 2,424,556 files**

Analyzing Individuals and Entities of Breach or Disclosure

Post Data Breach Discovery and Review Services

A combination of advanced data detection technologies and processes, extensive legal and regulatory compliance expertise, and proven notification and reporting procedures that harness the power of the world's leading legal discovery and review services and orients them directly on the detection, identification, review, and notification of sensitive data-related breaches and disclosures.

Cyber Discovery Review Process



Identification

- Forensic investigation
- Identify data sources
- Data collection and processing

Filter & Dedupe Data

- De-duplicate on family and individual document
- Assess file types

Adjust approach after review and contact with data

Define and Target PII

- Develop review protocol
- Search term calibration and sampling

Review

- Test and train review team
- Identify data subjects
- Null set sample review

Extraction and Normalization

- Extract data subjects and PII
- Data Normalization

Privacy Assessment

- Assess breach notice standards and requirements by jurisdiction

Notification and Response

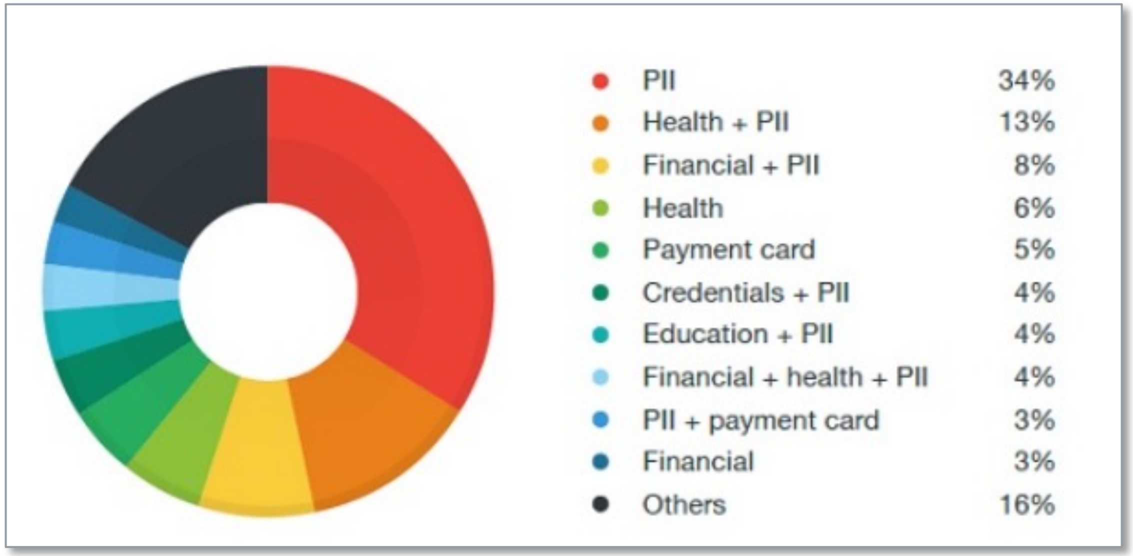
- Notify Regulators (if required)
- Notify Impacted Individuals (if required)
- Coordinate responses to inquiries
- Incident-related DSARs

Sensitive Data Breach Assessment Reporting

Automated Customizable Impact Assessment Reporting

AI Engines and Search Workflows Allow for Creation of Customized Reporting that Includes:

- Count of Sensitive Data by Type
- Count of Sensitive Data by Source
- Count of Unique Person Names and Organizations
- Count of Unique Persona Names and Organizations that Overlap with Sensitive Data Types
- Count of Sensitive Data by Range of Confidence Scores
- Count of Document Types within the Above Categories
- Count of Sensitive Data by Type over Custom Date Ranges
- Roll up reporting of Top Folder Locations
- General Dataset Statistics
- Visual Reporting via Customizable Dashboards
- Exception Reporting
- Deduplication Statistics



Large Document Extraction

Multi-Phased Approach Leveraging Technologists & 1L Extraction Teams

LDS Process (Large Data Subject)

Attorney Data Analysts

1. LDS Identification

a. Is Document In Scope
b. Does document contain more than 20 Data Subjects
c. Document Type – Spreadsheet, PDF, Email

2. LDS Extraction

a. Review of each LDS document and extraction of in scope PII into a standardized format
b. Quality control of extraction
c. Confirmation of all data subjects remaining in scope

Full Name	DOB	SNN	Full Address
John Roger Smith II	1/02/1950	123456789	123 Main St Apartment 12, Home Town, New York, 12345-6789
Smith II, John Roger	1950-1-2	123-45-6789	123 Main St Apt#12, Home Town, NY, 12345

3. LDS Normalization and Quality Control

a. Ensuring all extracted contents are formatted consistently regardless of source format
   i. Names: Last Name Suffix, First Name Middle Name
   ii. DOB: MM/DD/YYYY
   iii. SSN: 123456789
   iv. Address: House # Street Name, City, State Zip Code
   v. Leading, trailing, and double spacing
b. Quality control of normalization report – memo outlining issues/anomalies identified
   a. Jurisdiction matching elements extracted
   b. Name and address parsing
   c. Potential data extraction issues – dummy SSNs
   d. Normalization issues

First Name	Middle Name	Last Name	Suffix	DOB	SNN
John	Roger	Smith	II	01/02/1950	123456789
John	Roger	Smith	II	01/02/1950	123456789

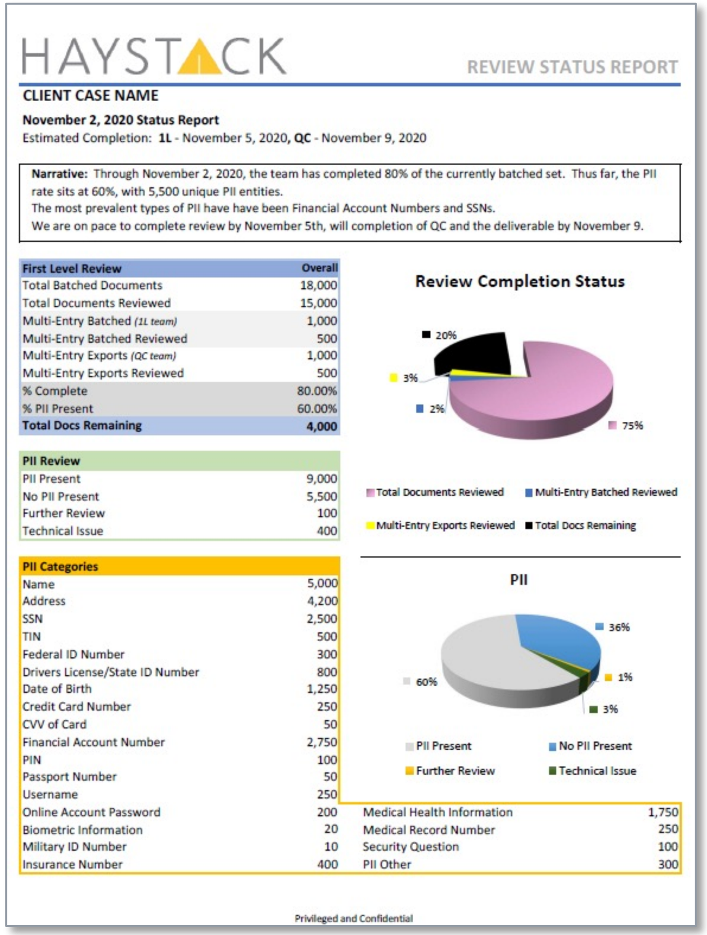
Address 1	Address 2	City	State	Zip Code
123 Main Street	Apt 12	Home Town	NY - New York	12345
123 Main Street	Apt 12	Home Town	NY - New York	12345

Data Breach Notification

PII Review Reporting

Expect customized project review metrics for:

- Up-to-date issue log
- All coding fields and choices
- Unique entity counts
- Estimated completion dates
- QC metrics
- Individual and team pace and overturn rates
- A detailed narrative that provides key information as to the status of the review



HAYSTACK REVIEW STATUS REPORT

CLIENT CASE NAME

November 2, 2020 Status Report
Estimated Completion: **1L** - November 5, 2020, **QC** - November 9, 2020

Narrative: Through November 2, 2020, the team has completed 80% of the currently batched set. Thus far, the PII rate sits at 60%, with 5,500 unique PII entities.
The most prevalent types of PII have have been Financial Account Numbers and SSNs.
We are on pace to complete review by November 5th, will completion of QC and the deliverable by November 9.

First Level Review	Overall
Total Batched Documents	18,000
Total Documents Reviewed	15,000
Multi-Entry Batched *(1L team)*	1,000
Multi-Entry Batched Reviewed	500
Multi-Entry Exports *(QC team)*	1,000
Multi-Entry Exports Reviewed	500
% Complete	80.00%
% PII Present	60.00%
Total Docs Remaining	**4,000**

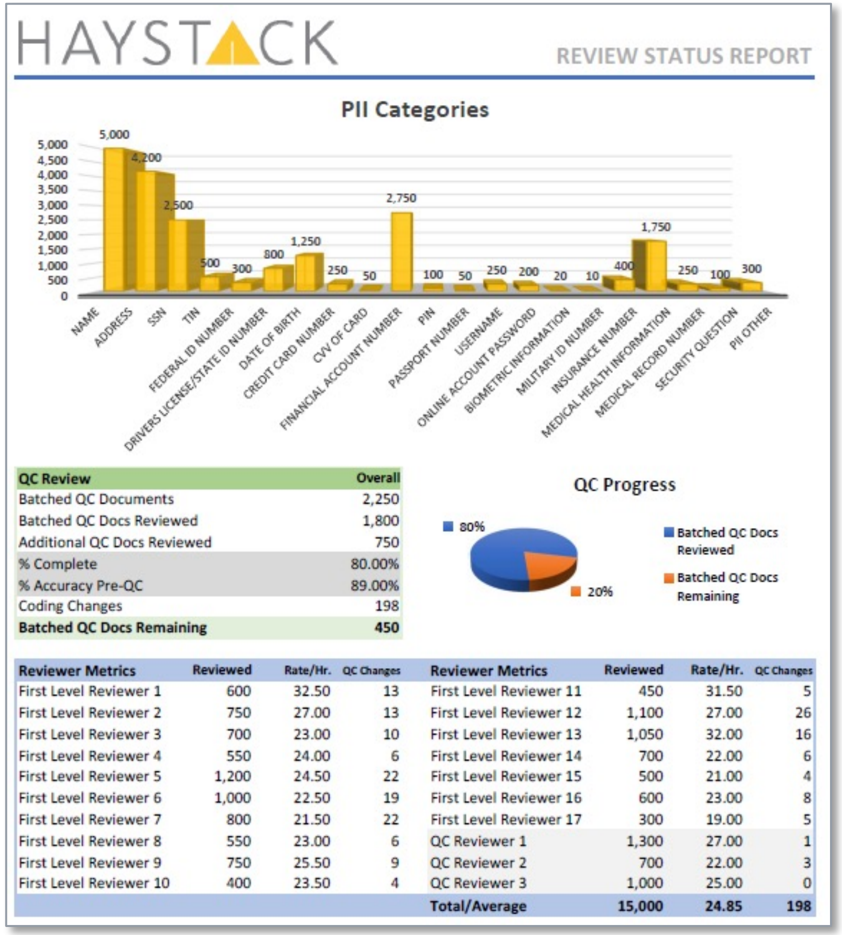
Review Completion Status

PII Review	
PII Present	9,000
No PII Present	5,500
Further Review	100
Technical Issue	400

PII

PII Categories	
Name	5,000
Address	4,200
SSN	2,500
TIN	500
Federal ID Number	300
Drivers License/State ID Number	800
Date of Birth	1,250
Credit Card Number	250
CVV of Card	50
Financial Account Number	2,750
PIN	100
Passport Number	50
Username	250
Online Account Password	200
Biometric Information	20
Military ID Number	10
Insurance Number	400
Medical Health Information	1,750
Medical Record Number	250
Security Question	100
PII Other	300

Privileged and Confidential



HAYSTACK REVIEW STATUS REPORT

PII Categories

QC Review	Overall
Batched QC Documents	2,250
Batched QC Docs Reviewed	1,800
Additional QC Docs Reviewed	750
% Complete	80.00%
% Accuracy Pre-QC	89.00%
Coding Changes	198
Batched QC Docs Remaining	**450**

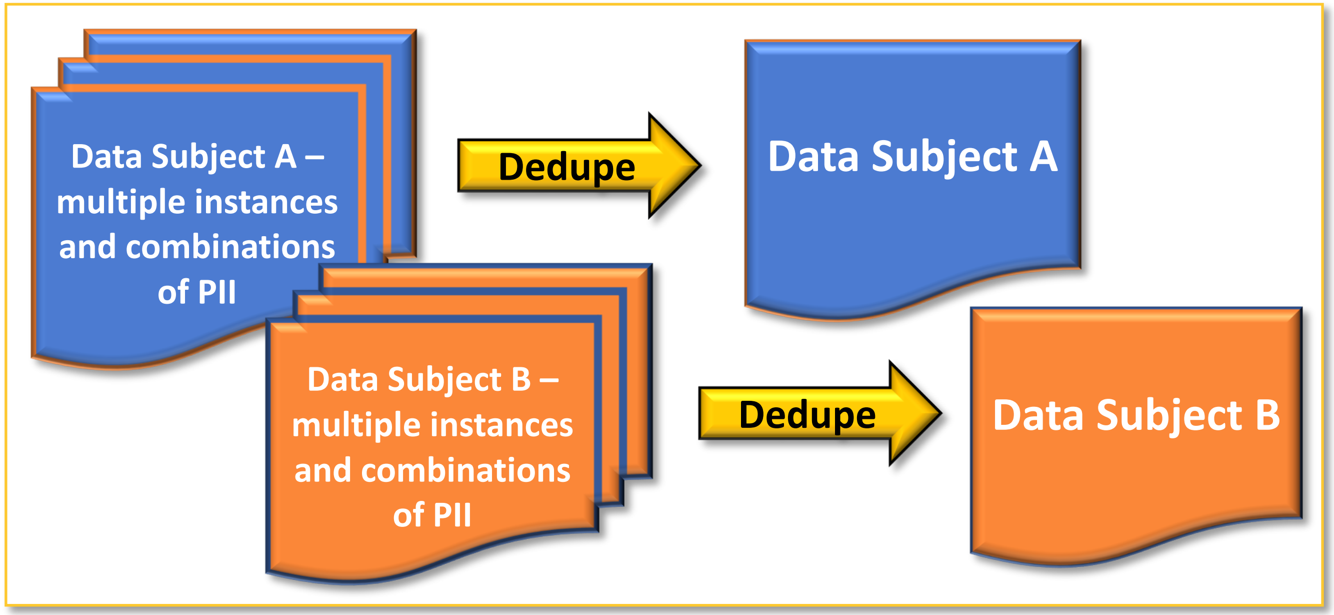
QC Progress

Reviewer Metrics	Reviewed	Rate/Hr.	QC Changes	Reviewer Metrics	Reviewed	Rate/Hr.	QC Changes
First Level Reviewer 1	600	32.50	13	First Level Reviewer 11	450	31.50	5
First Level Reviewer 2	750	27.00	13	First Level Reviewer 12	1,100	27.00	26
First Level Reviewer 3	700	23.00	10	First Level Reviewer 13	1,050	32.00	16
First Level Reviewer 4	550	24.00	6	First Level Reviewer 14	700	22.00	6
First Level Reviewer 5	1,200	24.50	22	First Level Reviewer 15	500	21.00	4
First Level Reviewer 6	1,000	22.50	19	First Level Reviewer 16	600	23.00	8
First Level Reviewer 7	800	21.50	22	First Level Reviewer 17	300	19.00	5
First Level Reviewer 8	550	23.00	6	QC Reviewer 1	1,300	27.00	1
First Level Reviewer 9	750	25.50	9	QC Reviewer 2	700	22.00	3
First Level Reviewer 10	400	23.50	4	QC Reviewer 3	1,000	25.00	0
				Total/Average	**15,000**	**24.85**	**198**

Notifying Individuals and Entities of Breach or Disclosure

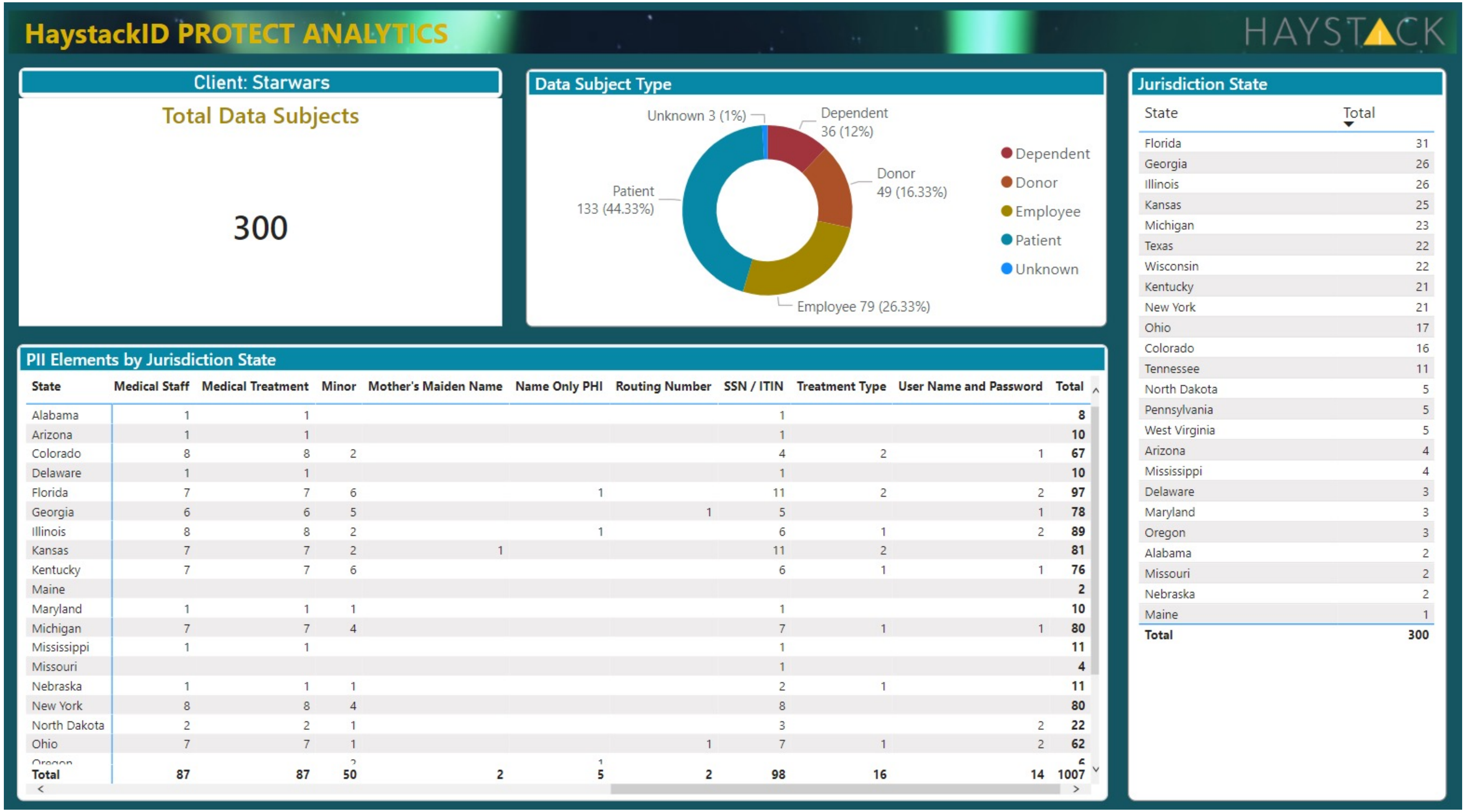
Data Subject Deduplication

- Identification of unique data subjects based off normalized outputs
- Ultra-scalable and tiered deduplication process
- Report of all elements extracted


Data Subject A – multiple instances and combinations of PII
Dedupe
Data Subject A
Data Subject B – multiple instances and combinations of PII
Dedupe
Data Subject B

- HSID recommendations on deduplication rules
  a) Last Name + First Name + SSN
  b) Last Name + First Name + MRN
  c) Last Name + First Name + DOB
  d) Last Name + First Name
  e) Last Name + Fuzzy First Name
- Deduplication test runs and introduction of knockout rules

Notification List and Dashboards



HaystackID PROTECT ANALYTICS

HAYSTACK

Client: Starwars

Total Data Subjects

300

Data Subject Type

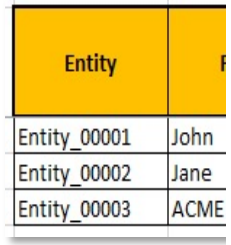
Jurisdiction State

State	Total
Florida	31
Georgia	26
Illinois	26
Kansas	25
Michigan	23
Texas	22
Wisconsin	22
Kentucky	21
New York	21
Ohio	17
Colorado	16
Tennessee	11
North Dakota	5
Pennsylvania	5
West Virginia	5
Arizona	4
Mississippi	4
Delaware	3
Maryland	3
Oregon	3
Alabama	2
Missouri	2
Nebraska	2
Maine	1
Total	**300**

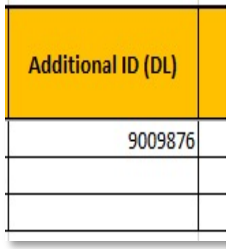
PII Elements by Jurisdiction State

State	Medical Staff	Medical Treatment	Minor	Mother's Maiden Name	Name Only PHI	Routing Number	SSN / ITIN	Treatment Type	User Name and Password	Total
Alabama	1	1					1			**8**
Arizona	1	1					1			**10**
Colorado	8	8	2				4	2	1	**67**
Delaware	1	1					1			**10**
Florida	7	7	6		1		11	2	2	**97**
Georgia	6	6	5			1	5		1	**78**
Illinois	8	8	2		1		6	1	2	**89**
Kansas	7	7	2	1			11	2		**81**
Kentucky	7	7	6				6	1	1	**76**
Maine										**2**
Maryland	1	1	1				1			**10**
Michigan	7	7	4				7	1	1	**80**
Mississippi	1	1					1			**11**
Missouri							1			**4**
Nebraska	1	1	1				2	1		**11**
New York	8	8	4				8			**80**
North Dakota	2	2	1				3		2	**22**
Ohio	7	7	1			1	7	1	2	**62**
Oregon			2		1					**6**
Total	**87**	**87**	**50**	**2**	**5**	**2**	**98**	**16**	**14**	**1007**

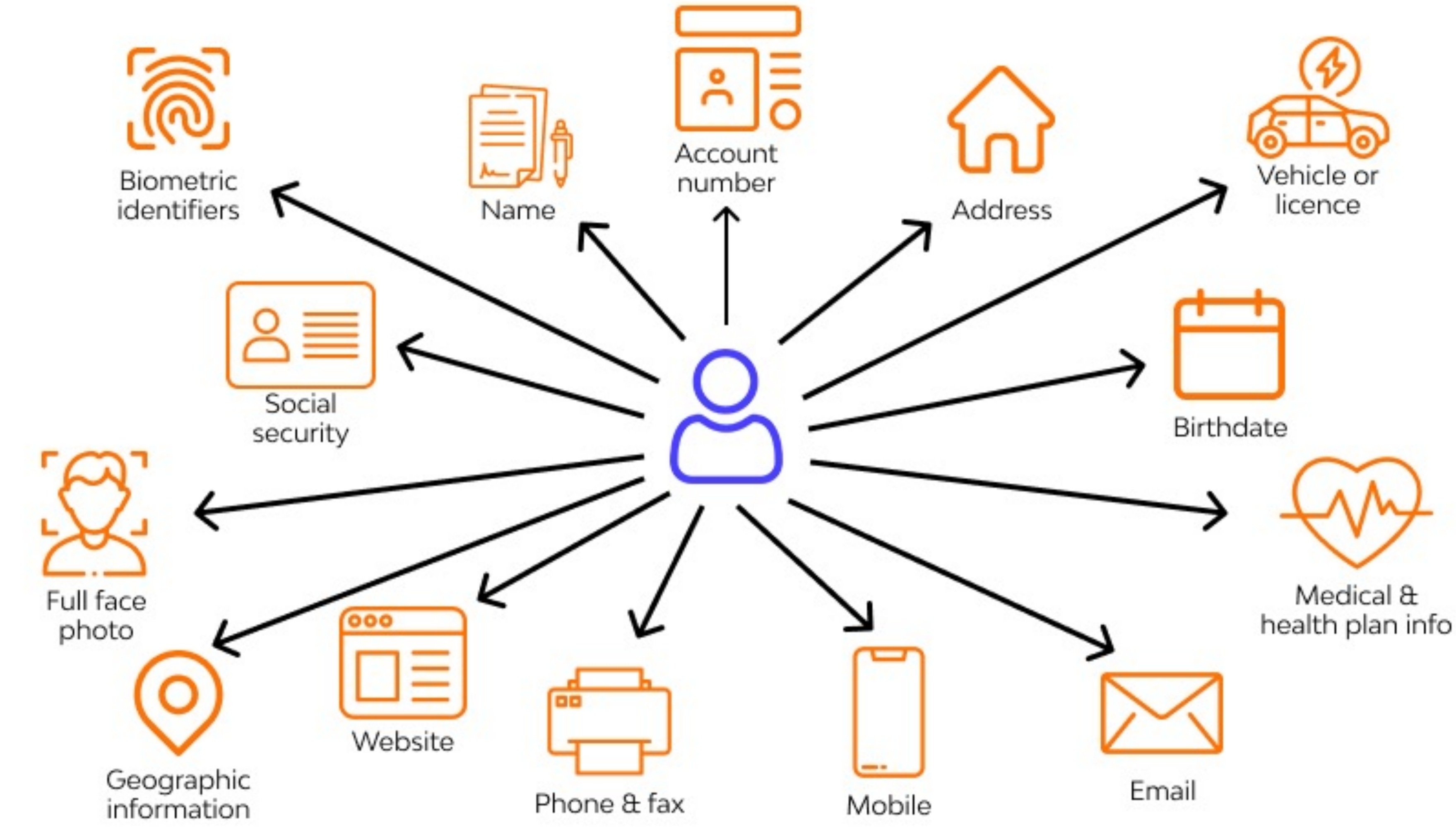
Consolidated Entity Notification List

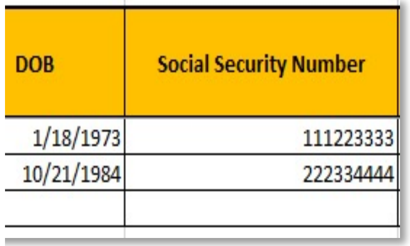


Entity	F
Entity_00001	John
Entity_00002	Jane
Entity_00003	ACME

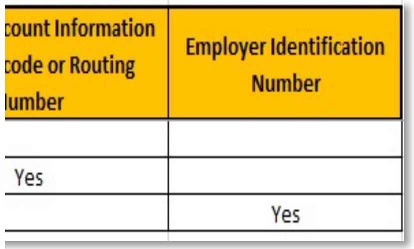


Additional ID (DL)	
9009876	





DOB	Social Security Number
1/18/1973	111223333
10/21/1984	222334444

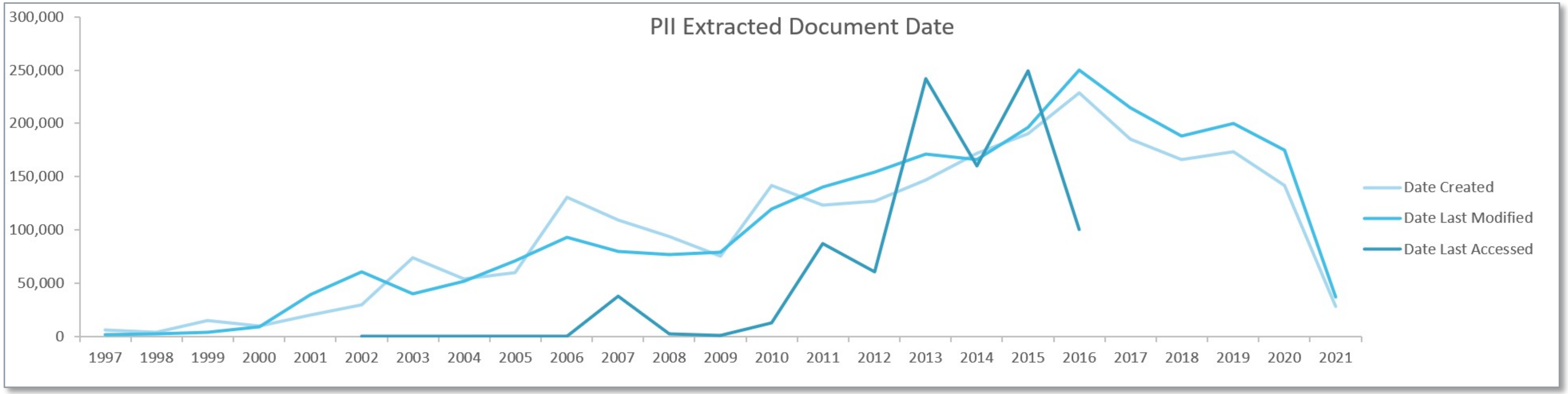


count Information code or Routing umber	Employer Identification Number
Yes	
	Yes

https://www.wallarm.com/what/securing-pii-in-web-applications

Post Notification Support and DSAR Workflow

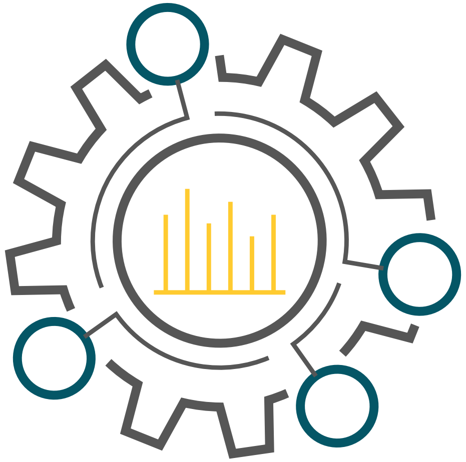
- Identification of Documents for the Data Subject
- Review and Redaction of documents requested by Data Subject
- Heat maps of PII data location and dates



HAYSTACK

A Synergistic Approach to Sensitive Data Discovery

HaystackID’s Information Governance Methodology



Step 1

Solidify Foundational Elements
Maturity Level, Program, Data Map

Step 2

Identify, Classify & Inventory Data
Critical, Sensitive, and ROT

Step 3

Operationalize Policies
Implement Defensible Disposition and Remediation

Step 4

Enable Automated Continuous Data Supervision

Step 5

Ensure Cyber Incident Preparedness
Proactive and Post-Breach Approaches

Step 6

Exceed Reasonable Security Measures

Cybersecurity Consulting Services

These services help you protect your data throughout the information lifecycle.

- **Cyber Discovery** (Post-Data Breach Discovery and Review)
- **Information Governance** (Compliance, Privacy, and Protection)
- **Incident Response & Advisory** (Remediation, Analysis, Post Breach Plan Design)

These services can be employed in **data discovery** and **legal discovery** situations from the point of unstructured data creation and interrogation throughout the information (and litigation) lifecycle.

Cybersecurity:
Protection against the criminal or unauthorized use of electronic data.

Cyber Risk

Information Governance

Incident Response

How can we help you?

Learn how our infinite capabilities can help you at HaystackID.com or reach out to us at info@HaystackID.com / 877.942.9782