

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



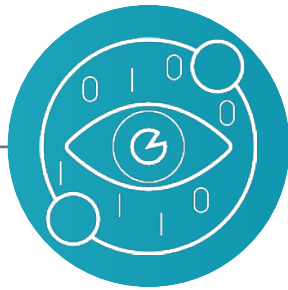
Does your Organization



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?

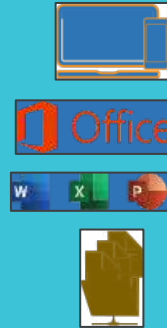
Processing PII = creation, collection, use, processing, storage, maintenance, dissemination, disclosure, or disposal (NIST.SP.800-53r5, PT-3, Sept. 2020)

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....

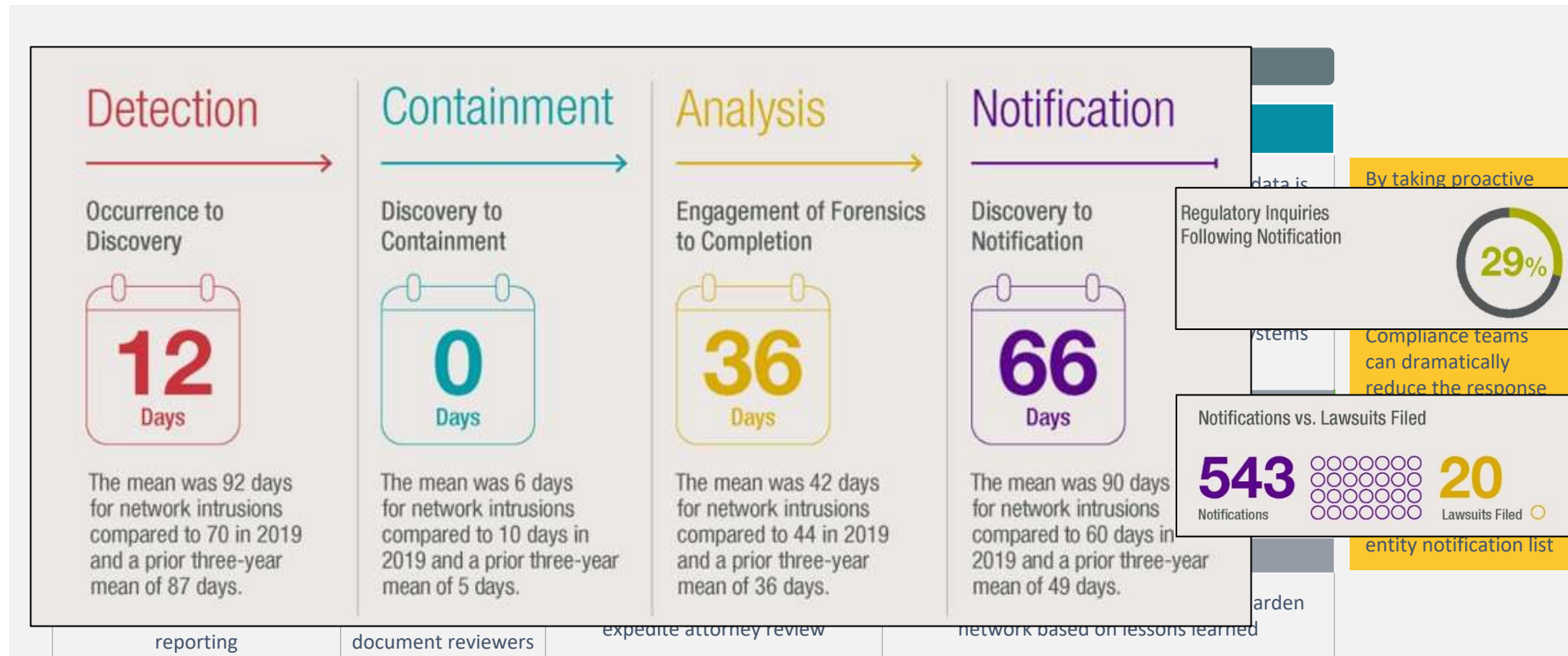


Canada
Data Minimization Standards



GDPR and UK GDPR
Data Minimization Standards

Example of a Cyber Incident Response Remediation Timeline



The Importance of Discovery Intelligence

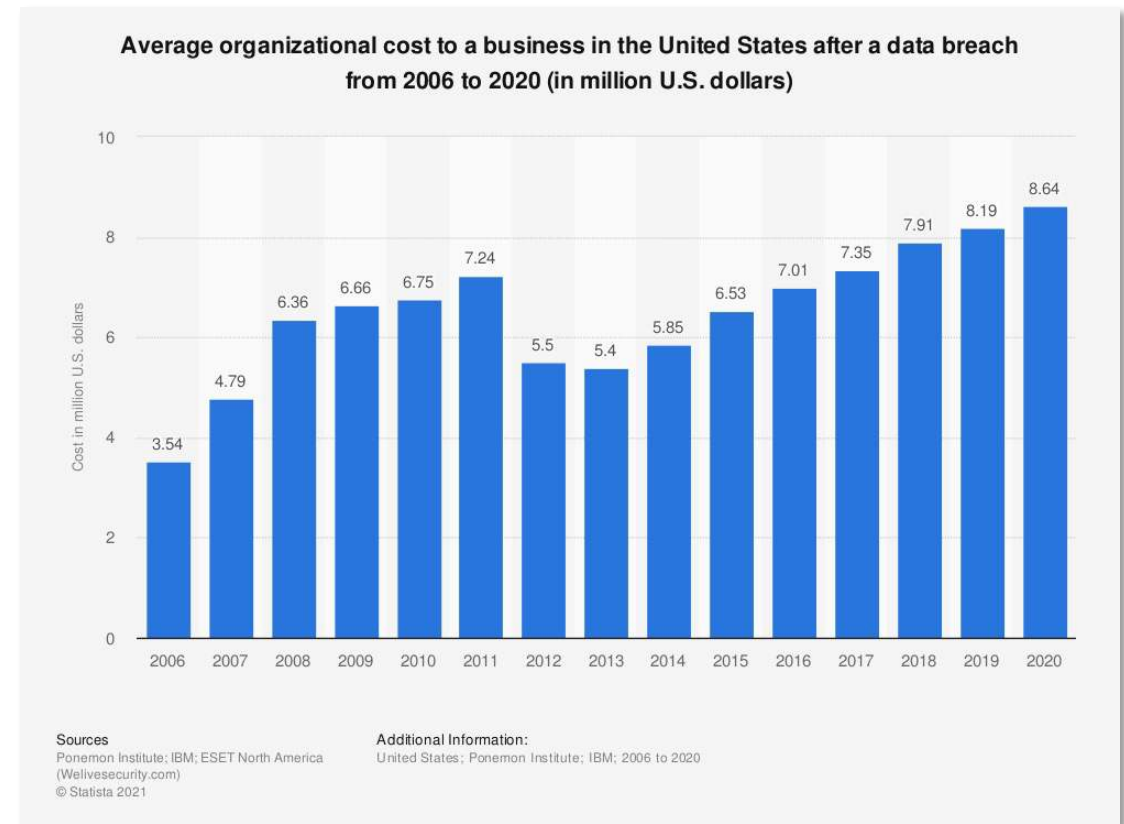
The Average Cost of a Breach

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 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:

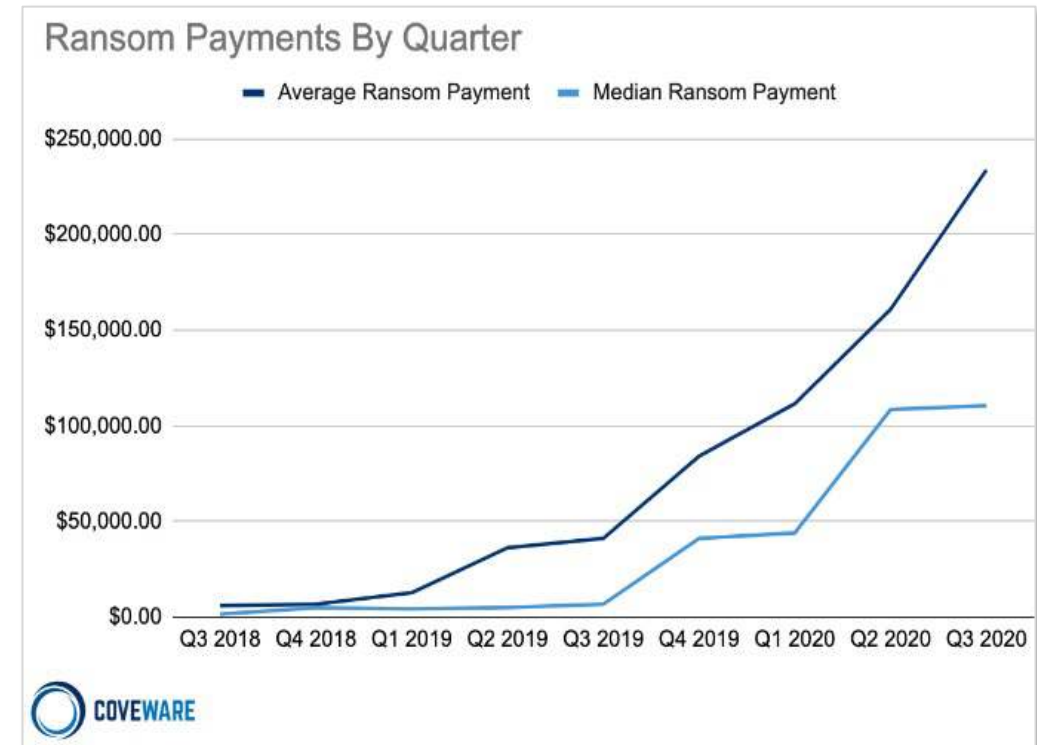
- Require less effort
- Largely automated
- Generate much higher payouts
 - > **\$233,000** per event in Q4 2020

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:

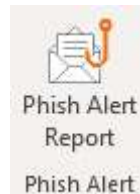
- 25% of U.S. attacks
- 36% in Australia
- 75% increase in Singapore
- Most common data breach type in the U.K. & Germany



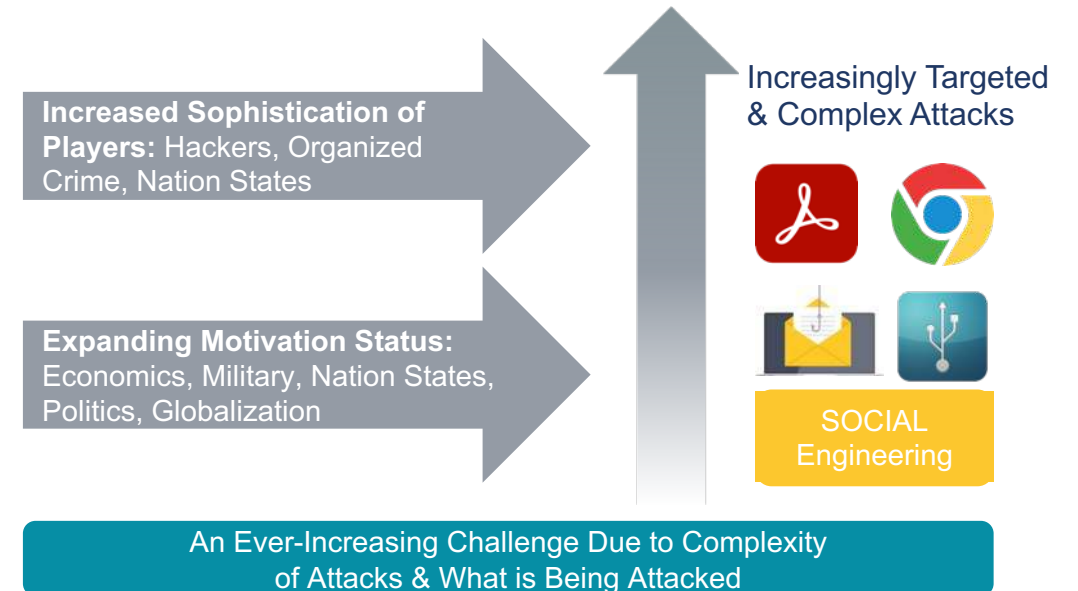
Cybersecurity Lessons Learned Since COVID-19

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.



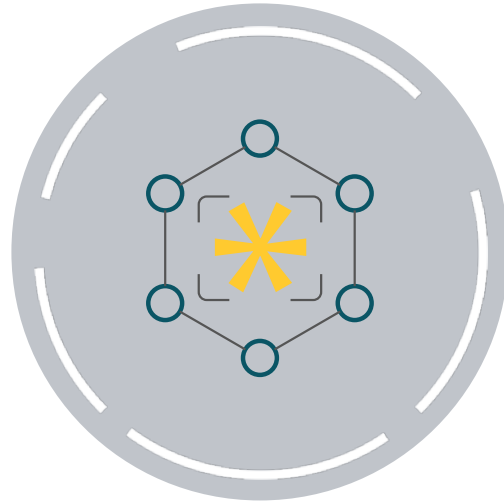
Detecting Sensitive Data Presence and Scope

Multiple Approaches

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



Off-the-Shelf NLP Models



Regular Expressions



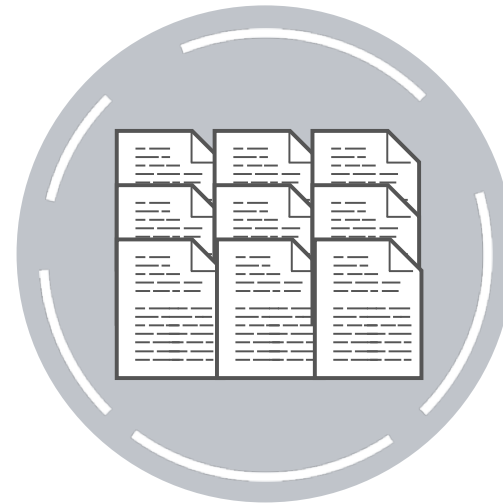
Internal Machine Learning

PII/PHI/Entity Density

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



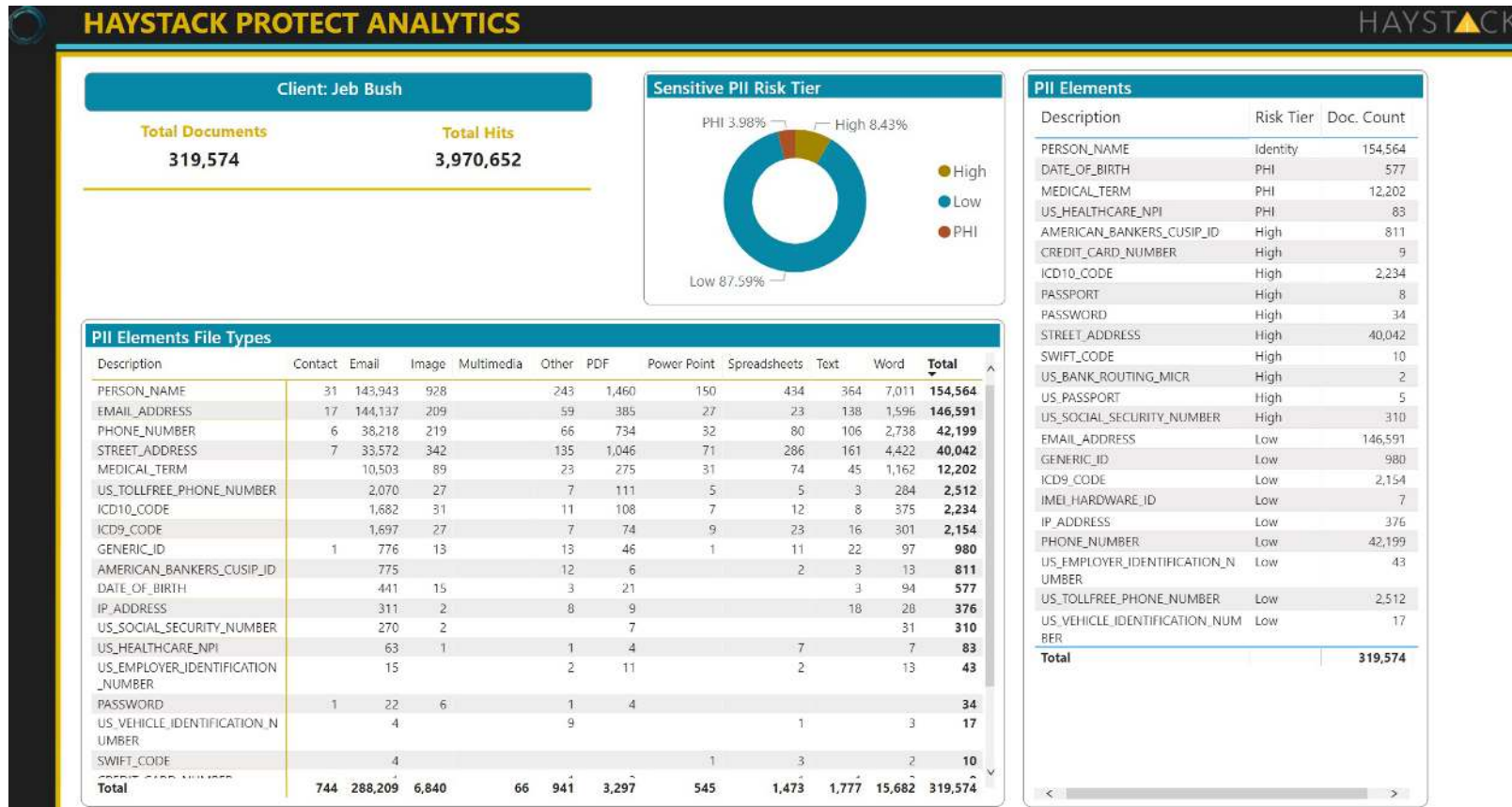
Sparse Documents



Dense Documents

ReviewRight® Protect

AI Model Output & Validation



Identifying Breach or Disclosure Impact and Options

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

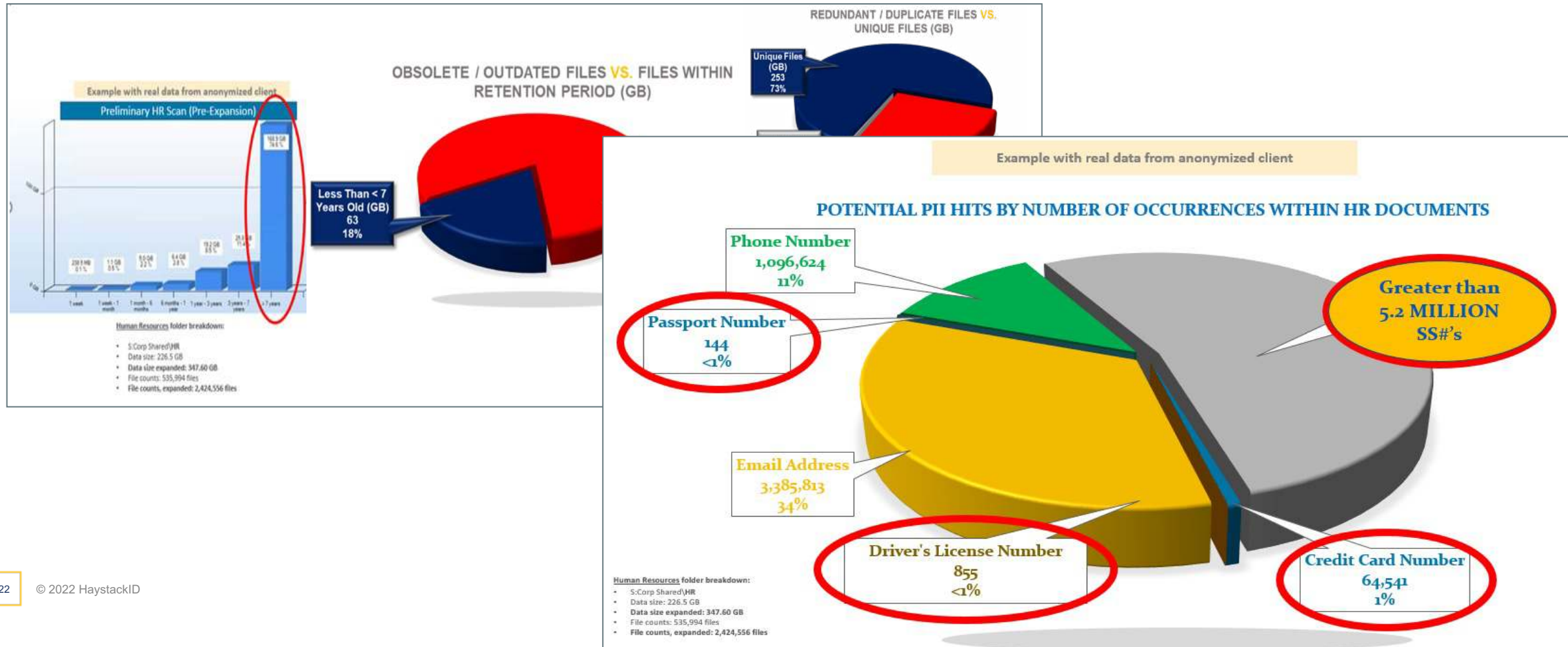
By: IBM Security Board of Advisors

Data Value ↑

Data Type	Examples		% of Sensitive Data
Enterprise Critical	<ul style="list-style-type: none"> Critical Intellectual Property 	<ul style="list-style-type: none"> Top-Secret Plans and Formulas 	0.01-0.1%
Executive	<ul style="list-style-type: none"> Acquisition and Divestiture Plans 	<ul style="list-style-type: none"> Executives and Board Deliberations 	0.1 - 2%
Regulated	<ul style="list-style-type: none"> SPI & PII Sarbanes-Oxley 	<ul style="list-style-type: none"> HIPAA ITAR Quarterly results 	1-50%
Business Strategic	<ul style="list-style-type: none"> External Audit Results 	<ul style="list-style-type: none"> Alliance & Joint Venture, Partner Data Business Strategic Plans 	1-5%
Business Unit Critical	<ul style="list-style-type: none"> Design Documents R & D Results 	<ul style="list-style-type: none"> Customer records Pricing Data Security Data 	10-20%
Operational	<ul style="list-style-type: none"> Project Plans Contracts 	<ul style="list-style-type: none"> Salaries & Benefits Data Accounts Receivable 	20-80%
Near-Public	<ul style="list-style-type: none"> List of Partners Revenue Growth 	<ul style="list-style-type: none"> Market Intelligence Pay Compensation Data 	10-80%

Operationalize Policies

Implement Defensible Disposition & Remediation – Use Case



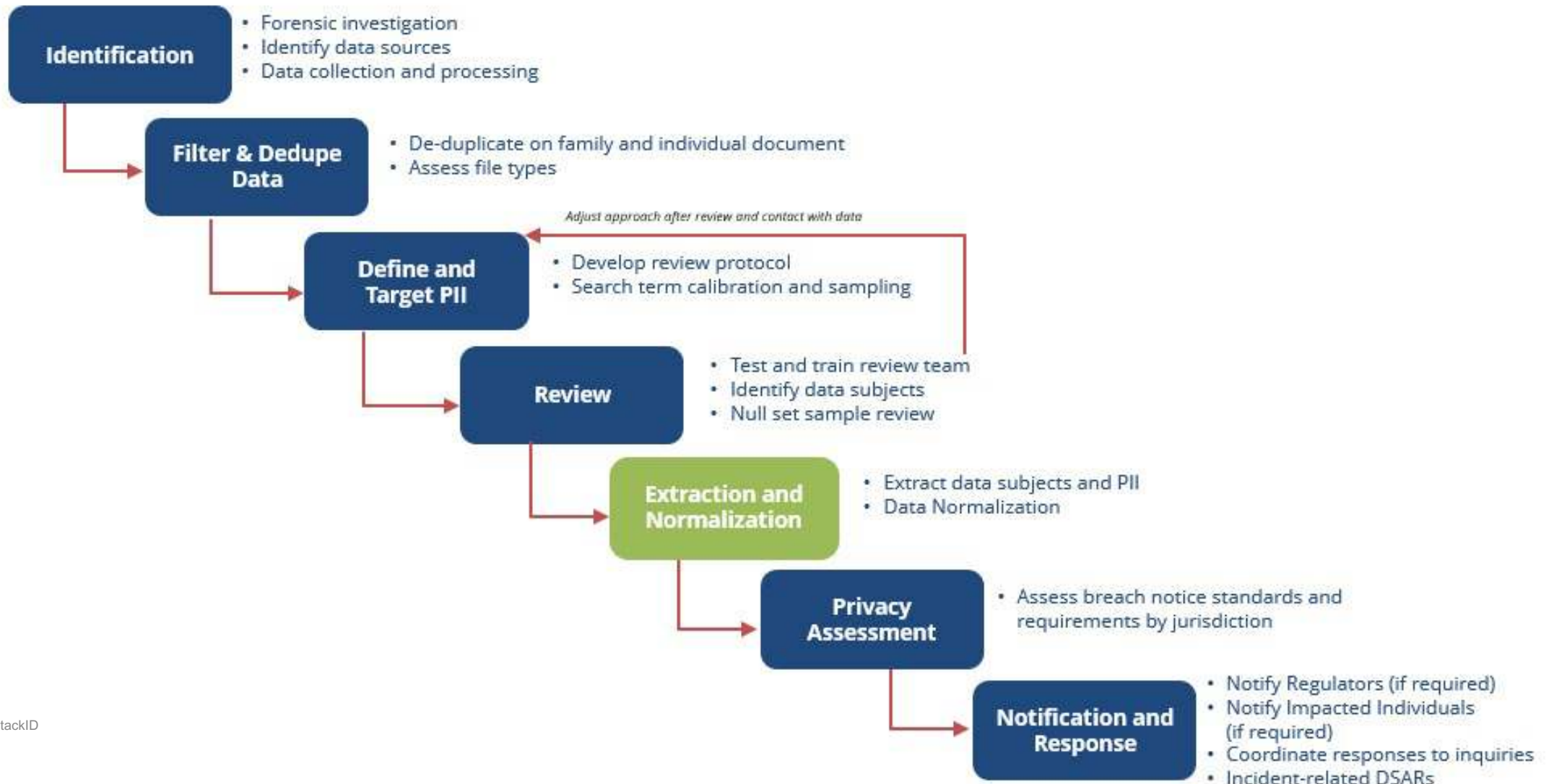
Analyzing Individuals and Entities of Breach or Disclosure

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

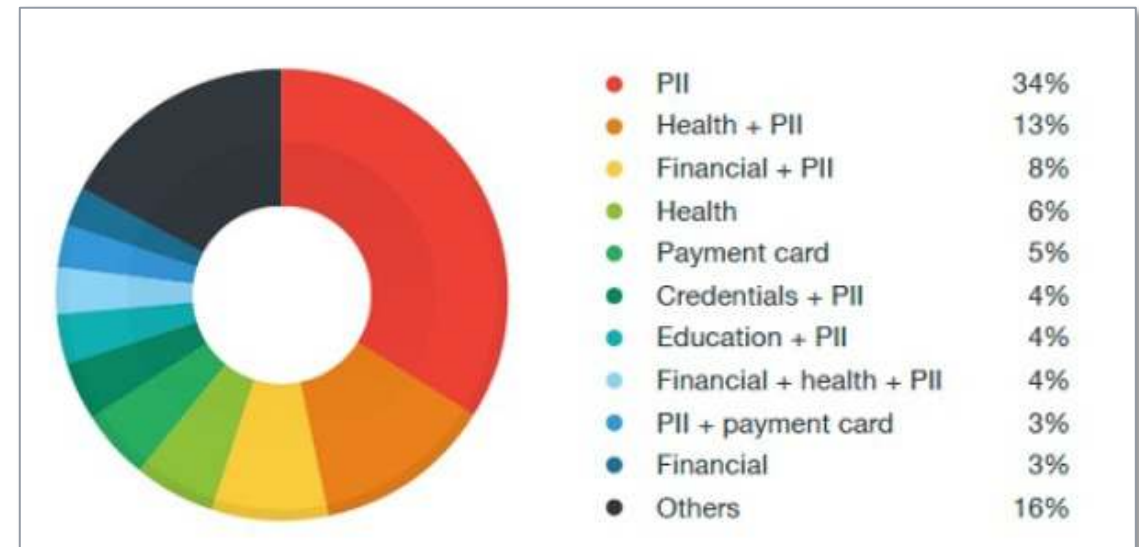


Sensitive Data Breach Assessment Reporting

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

- Is Document In Scope
- Does document contain more than 20 Data Subjects
- Document Type – Spreadsheet, PDF, Email

2. LDS Extraction

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

3. LDS Normalization and Quality Control

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

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

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

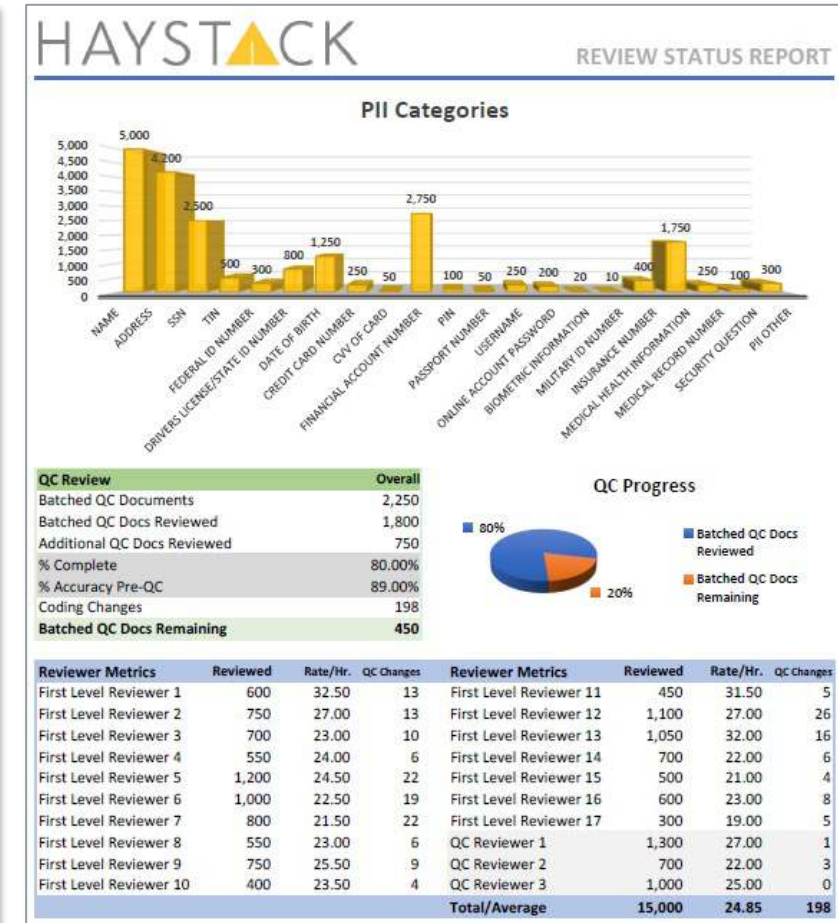
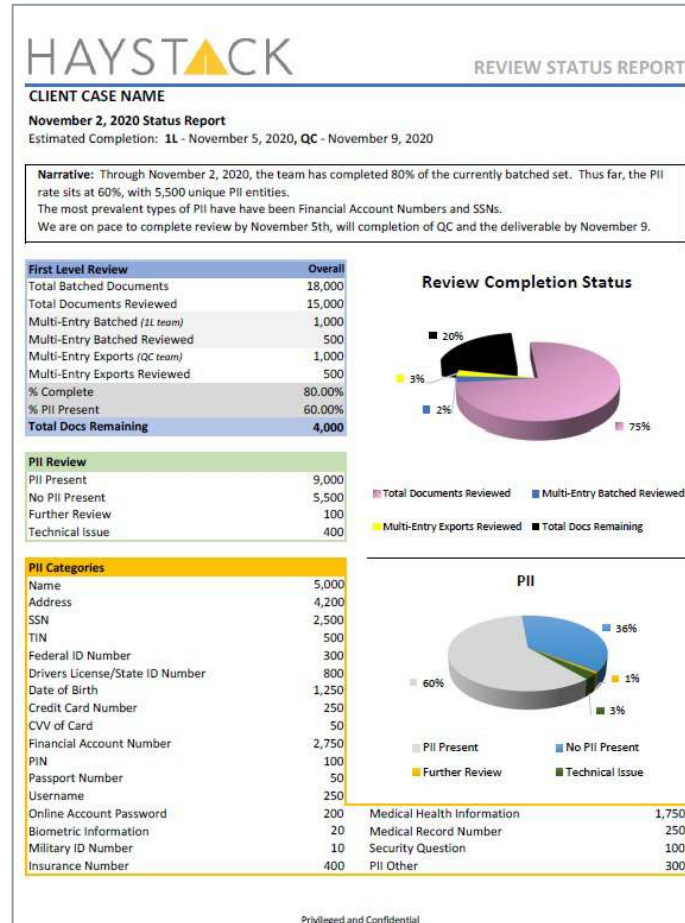
Address 1	Address 2	City	State	Zip Code
123 Main Street	Apt 12	Home Town	NY - New York	12345
123 Main Street	Apt 12	Home Town	NY - New York	12345

Data Breach Notification

PII Review Reporting

Expect customized project review metrics for:

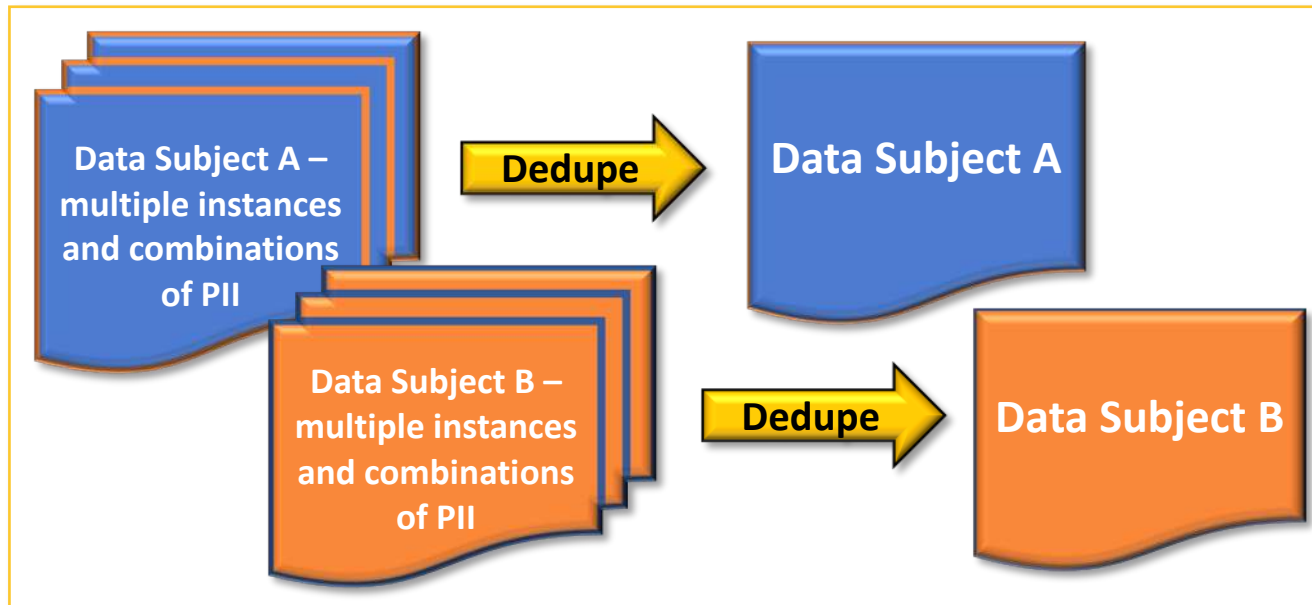
- Up-to-date issue log
- All coding fields and choices
- Unique entity counts
- Estimated completion dates
- QC metrics
- Individual and team pace and overturn rates
- A detailed narrative that provides key information as to the status of the review



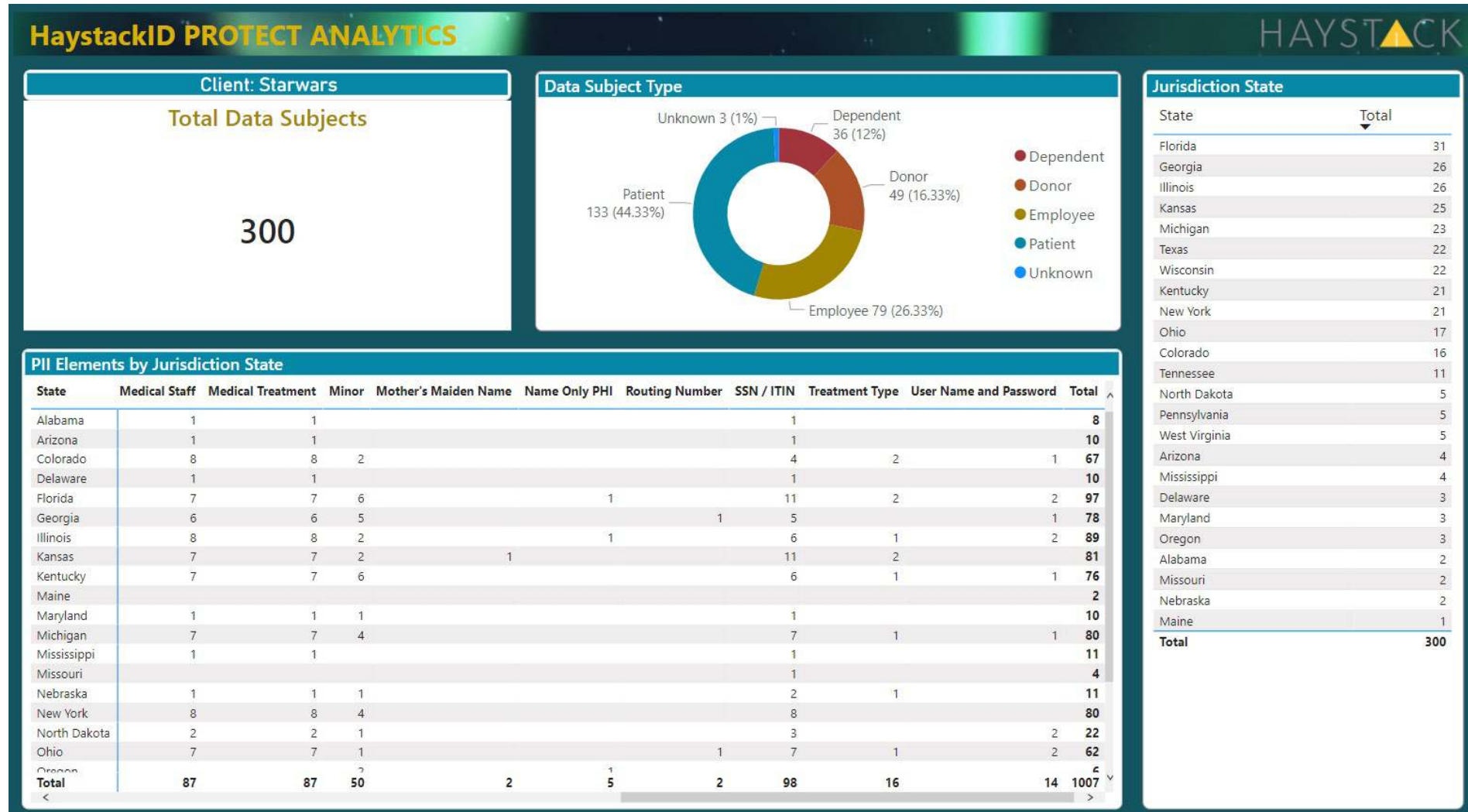
Notifying Individuals and Entities of Breach or Disclosure

Data Subject Deduplication

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



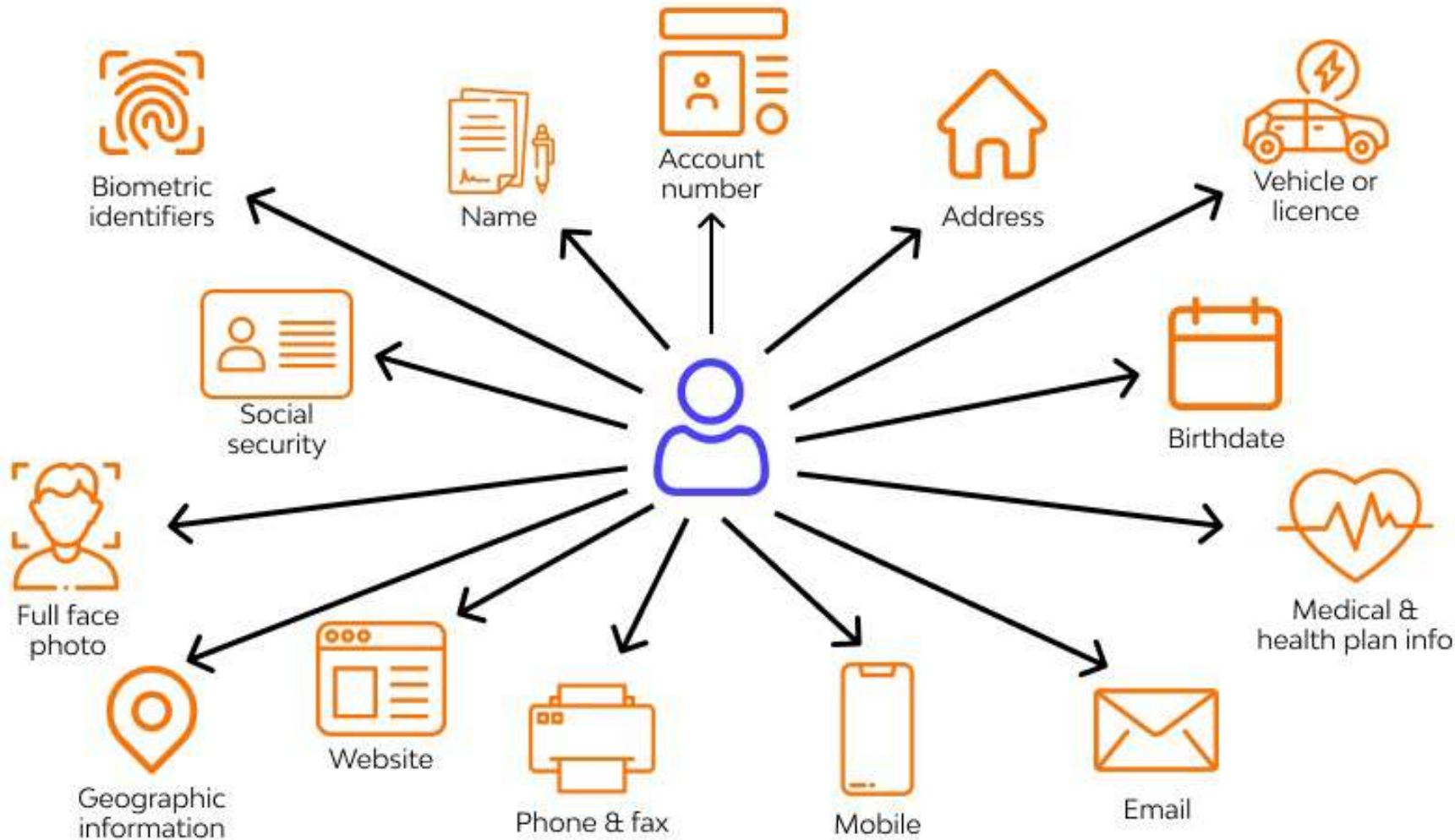
Notification List and Dashboards



Consolidated Entity Notification List

Entity	
Entity_00001	John
Entity_00002	Jane
Entity_00003	ACME

Additional ID (DL)
9009876

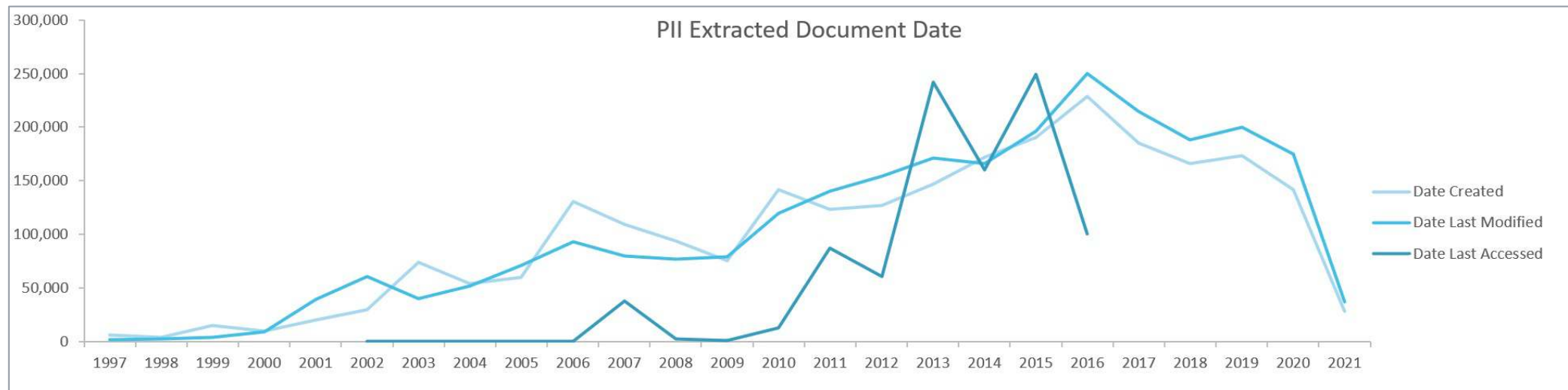


DOB	Social Security Number
1/18/1973	111223333
10/21/1984	222334444

Account Information code or Routing Number	Employer Identification Number
Yes	
	Yes

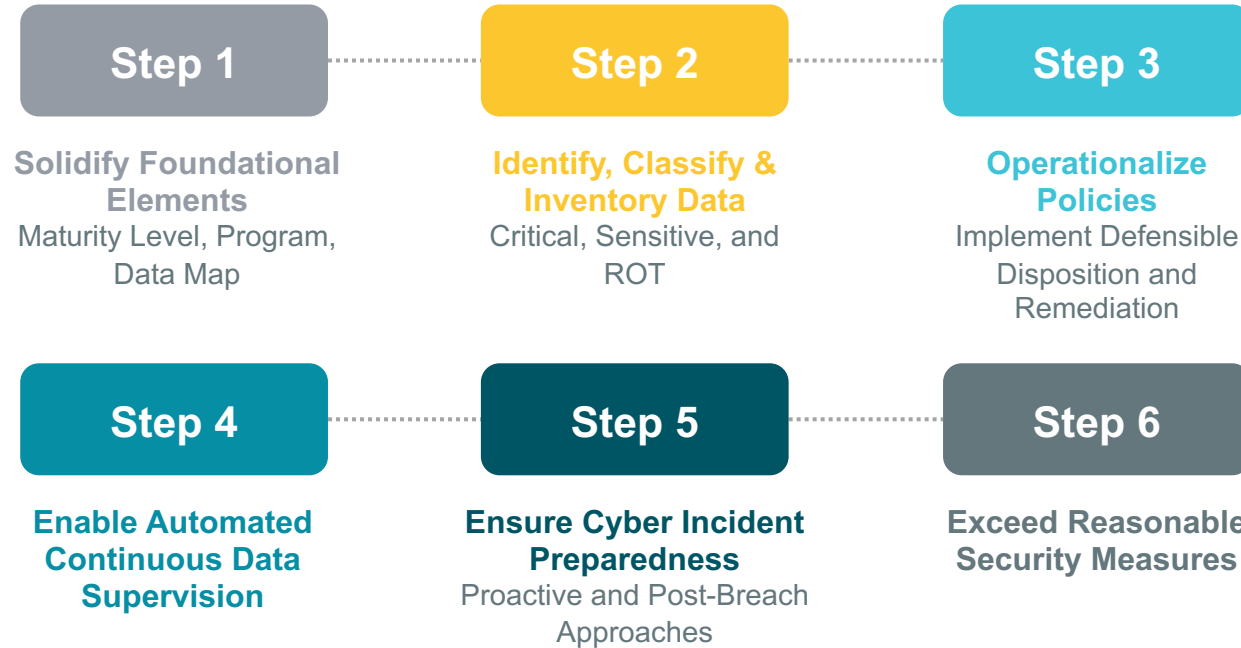
Post Notification Support and DSAR Workflow

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



A Synergistic Approach to Sensitive Data Discovery

HaystackID's Information Governance Methodology



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](tel:877.942.9782)