"OUT OF THE BOX" SCANS FOR SENSITIVE DATA

Easy Solution to a Difficult Problem?

DiscoverReady®

Philip Richards
CTO
DiscoverReady

AGENDA

1. **Define the problem**
   How big is the problem?
2. **Our data set**
3. **Analysis Process**
4. **Preliminary findings**
5. **Conclusions**
   Where do we go from here?

WHAT IS SENSITIVE DATA?

- **Two basic types of sensitive data**
  - Personal Data
    - PII, PCI, PHI, financial information, etc.
  - Corporate Sensitive Data
    - Proprietary data, financial data, strategy documents, M&A info, system info

Every corporation, every individual, every organization has sensitive data

DATA EXHAUST

- **Where should sensitive data reside (generally)?**
  - Protected systems that need the information
  - Systems of record
  - Vaults, repositories
- **Where should sensitive data not exist (generally)?**
  - Unprotected systems that don't need the information
  - File shares
  - Email
  - Personal systems (for businesses)

When data moves to an unprotected system, it is difficult to contain. We call it:
DATA EXHAUST

DATA EXHAUST QUESTIONS

How big is the exhaust problem?

How much exhaust exists?

How well can industry tools identify exhaust?

Can we improve our ability to identify exhaust?

AGENDA

1. **Define the problem**

   How big is the problem?

2. **Our data set**
3. **Analysis Process**
4. **Preliminary findings**
5. **Conclusions**

   Where do we go from here?

OUR DATA SET

- **We obtained permission to use some data in our possession to answer these questions**
  - 8 Companies
  - 6 Industries
  - Input data
    - 5,030,759 files
    - 15,794 GBs
  - Extracted data
    - 101,074,922 files
    - 18,929 GBs
  - 510 Custodians

- **Data set considerations**
  - Data is not a random sample of companies
  - Data is not a random sample of data from the companies
  - Data is not a random sample of industries
  - Data is not a random sample of people

- We have a lot of data
- Data can tell us where to look next
- Data can help us develop theories that can be proven or disproven

AGENDA

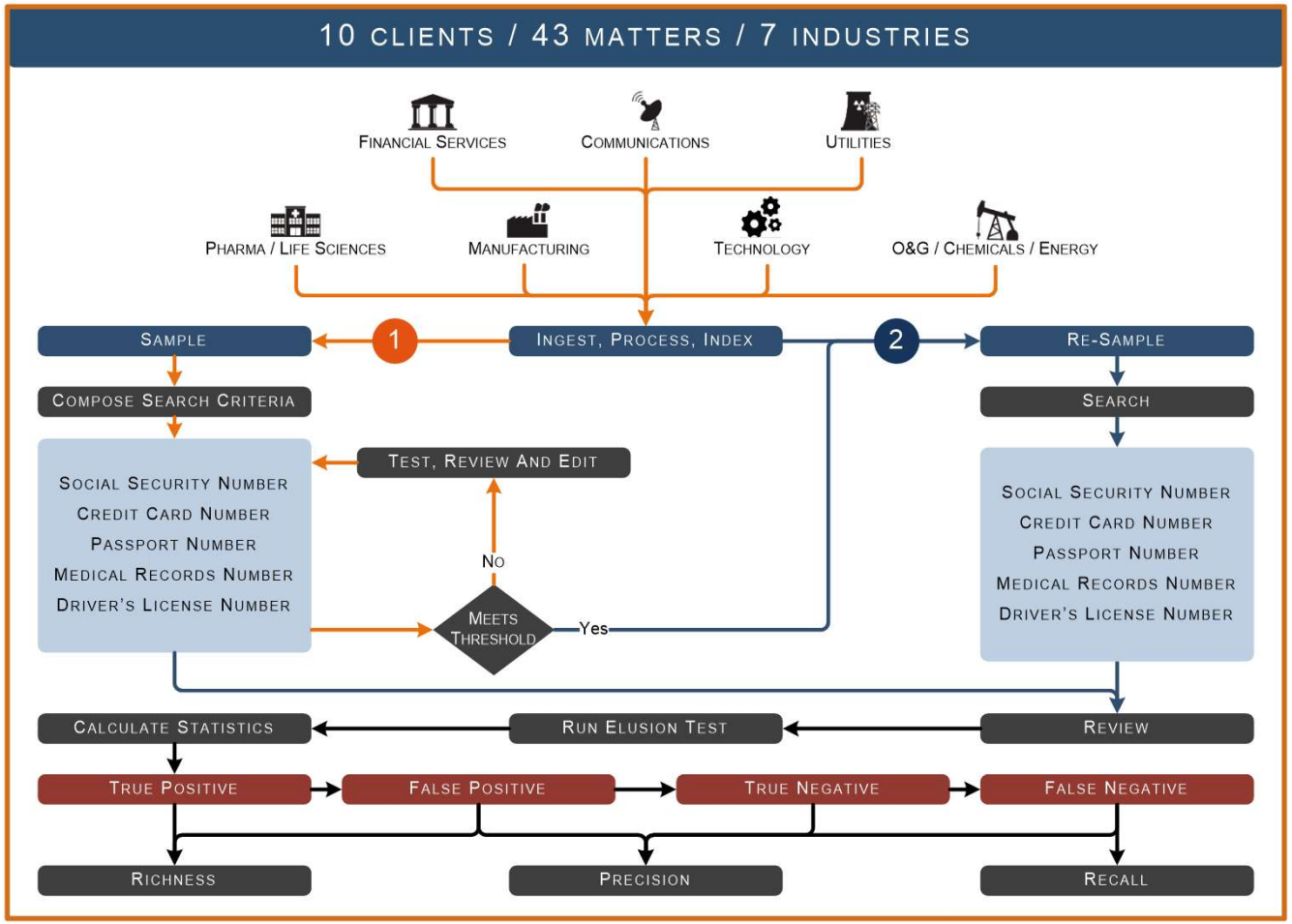
1. **Define the problem**

   How big is the problem?

2. **Our data set**
3. **Analysis Process**
4. **Preliminary findings**
5. **Conclusions**

   Where do we go from here?

ANALYSIS PROCESS



- **Test set 1 –**
  - Data set used for Richness and out-of-the-box scan results
  - Data set used for optimizing out-of-the-box scan results
- **Test set 2 –**
  - Data set used to test optimized criteria

OUT OF THE BOX SCANS

Sensitive Data Type	Standard Scan	Reference	Notes
Driver's License Numbers	Python Library Regex Repository	http://adr-inc.com/PDFs/State_DLFormats.pdf. https://pypi.python.org/pypi/DLNValidation/0.1.4 https://ntsi.com/drivers-license-format/ Known to be valid as of 2016-04-16. https://github.com/adambullmer/USDLRegex	High number of possible permutations
Passport Number	Reference Text	https://www.uscis.gov/e-verify/about-program/e-verify-enhancements/us-passport-and-visa-number-entry	
Medical Records Number	Content Control List, Regex	https://community.sophos.com/kb/en-us/112192	No widely adopted standards
Social Security Number	Leading commercial product	Standard regexes	Variations on standards
Credit Card Number	Leading commercial product	Standard regexes	Variations on standards, verified Luhn check

AGENDA

1. **Define the problem**

   How big is the problem?
2. **Our data set**
3. **Analysis Process**
4. **Preliminary findings**
5. **Conclusions**

   Where do we go from here?

PRELIMINARY FINDINGS – RICHNESS

Full random Sample Results

Measurement Type	Hits	Sample Size	Richness Lower Bound	Richness Upper Bound	Confidence Level
Social Security Number	0	720	0.00%	0.51%	95%
Credit Card Number	0	419	0.00%	0.88%	95%
Medical Record Number	0	186	0.00%	1.96%	95%
Driver's License Number	0	156	0.00%	2.34%	95%
Passport Number	0	1578	0.00%	0.23%	95%

No hits in random sample results.
Low richness makes evaluation of recall difficult.

PRELIMINARY FINDINGS – OUT-OF-THE-BOX PRECISION

Precision sample based on hit results

Data Type	Basic Finding	Sample Size (Hits)	Precision Lower Bound	Precision Upper Bound	Confidence Level
Social Security Number	Data identified by searches	120	4%	15%	95%
Credit Card Number	Data identified by searches	118	2%	11%	95%
Medical Record Number	Data identified by searches	13	2%	45%	95%
Driver's License Number	No true positives in the data	90	0%	4%	95%
Passport Number	No true positives in the data	90	0%	4%	95%

Low precision measured across all out-of-the-box scans

AGENDA

1. **Define the problem**

   How big is the problem?
2. **Our data set**
3. **Analysis Process**
4. **Preliminary findings**
5. **Conclusions**

   Where do we go from here?

PRELIMINARY CONCLUSIONS

- **How big is the exhaust problem?**
  - We found PII for all companies in every industry in every data set
  - None of the data sets should contain sensitive data
- **How much sensitive data exhaust exists?**
  - Ubiquitous in low proportions, but even at low proportions there are large quantities
  - Some types (such as SSN and CCN) exist more commonly than other types (such as Passport and Driver's License)
- **How well can out-of-the-box tools identify exhaust?**
  - Precision is low, usually less than 30% (upper confidence interval bound)
- **Can we improve our ability to identify exhaust?**
  - Yes, but it requires specific (not general) customizations

All companies should be worried about sensitive data exhaust.
We have not found a "silver bullet" for identifying exhaust.

FUTURE WORK

- **Determine recall and how it changes with tuning of searches for higher precision**
  - How to resolve in the common situation of low richness?
- **More sampling for tighter bounds on measurements of precision, recall and richness**
- **Study the distribution of exhaust across industries**
- **Study the distribution of exhaust across file types**
- **Determine how much improvement should be expected when tailoring searches to companies and industries**
- **Include more data types**
- **Include evaluation of additional industry tools**