"OUT OF THE BOX" SCANS FOR SENSITIVE DATA

Easy Solution to a Difficult Problem?

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
When data moves to an unprotected system, it is difficult to contain. We call it: 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)
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?
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?
• 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

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
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

10 CLIENTS / 43 MATTERS / 7 INDUSTRIES

- **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
## Sensitive Data Type

<table>
<thead>
<tr>
<th>Sensitive Data Type</th>
<th>Standard Scan</th>
<th>Reference</th>
<th>Notes</th>
</tr>
</thead>
</table>
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 |
| 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?
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**

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

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?
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