VLDS Insights

Data: Security & Technology Panel

David Ihrie, Center for Innovative Technology (Moderator)
Will Goldschmidt, Virginia Department of Education
Aaron Schroeder, Virginia Tech Institute for Policy & Governance
# Principles of Good Data Programs

<table>
<thead>
<tr>
<th></th>
<th>Facebook/Google</th>
<th>NSA</th>
<th>VLDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utility (Why)</td>
<td>?</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Transparency (How, When, Where)</td>
<td>?</td>
<td>X</td>
<td>?</td>
</tr>
<tr>
<td>Security (Appropriate)</td>
<td>?</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Boundary (What)</td>
<td>X</td>
<td>?</td>
<td>✓</td>
</tr>
<tr>
<td>Accountability (Who)</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
</tbody>
</table>
VLDS Insights

Data: Security & Technology Panel

David Ihrie, Center for Innovative Technology (Moderator)
Will Goldschmidt, Virginia Department of Education
Aaron Schroeder, Virginia Tech Institute for Policy & Governance
Privacy Protecting Federated query

• Two Steps:
  1. Identity Resolution Process
  2. Query Execution Process

• Three Components
  1. Shaker
  2. Data Adapter
  3. Exposure Database
Getting Data Ready for “De-Identified Federation”

• How De-Identification Works in VLDS System

The Data Adapters prepare the data (including de-identification) before leaving the agency firewall.

Agency Firewalls

Agency 1 Exposure Database -> DATA ADAPTER -> VLDS Shaker -> DATA ADAPTER -> Agency 2 Exposure Database
Getting Data Ready for “De-Identified Federation”

- **What the Data Adapter Does:**

  De’ Smith-Barney IV
  
<table>
<thead>
<tr>
<th>Action</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remove Suffix</td>
<td>Removes suffix from the data.</td>
</tr>
<tr>
<td>Remove Symbols</td>
<td>Removes symbols from the data.</td>
</tr>
<tr>
<td>Substitution Cipher</td>
<td>Applies a substitution cipher to the data.</td>
</tr>
<tr>
<td>Order Hashing</td>
<td>Orders the data based on a hashing key.</td>
</tr>
</tbody>
</table>

  **DeSmithBarney**
  
  - **Substitution Alphabet** – Dynamically Generated using Hashing Key (below)
  - **Hashing Key** – Dynamically Generated and sent from Matching Engine

  - **DeSmithBarney**
    - **YDXWKQTAGOLCNSVEFHMRJPBZUI**
    - **b164f11d-aa37-44ca-93c3-82d3e0155061**
  
  - **C57S78XCEBF9WECP2AA9DK59N1CO27QBES54HFD**

  **CLEANED & ENCODED for TRANSPORT**
**Cleaned and Encoded Matching Data (Internal ID, First and Last Name)**

<table>
<thead>
<tr>
<th>INTERNAL_ID_HASHED</th>
<th>FIRST_NAME</th>
<th>LAST_NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>044AA90CE74E2ED3B6B080CFE93F8ED263B73050</td>
<td>F11BDAE3EM86EA8</td>
<td>J11EDAV3E</td>
</tr>
<tr>
<td>044AA90CE74E2ED3B6B080CFE93F8ED263B73050</td>
<td>F11BDAE3EM86EA8</td>
<td>J11EDAV3E</td>
</tr>
<tr>
<td>044AA90CE74E2ED3B6B080CFE93F8ED263B73050</td>
<td>F11BDAE3EM86EA8</td>
<td>J11EDAV3E</td>
</tr>
<tr>
<td>044AA90CE74E2ED3B6B080CFE93F8ED263B73050</td>
<td>F11BDAE3EM86EA8</td>
<td>J11EDAV3E</td>
</tr>
</tbody>
</table>

**INTERNAL_ID is the same**

**LAST_NAME is NOT**

Many agencies DO NOT have an Index of unique individuals. There can be many representations of that individual.

**What do we do?**

**Statistical Log Analysis and Reduction**

We dynamically build a new “virtual” record made up of “most likely” demographics.
Probabilistic Linkage Process (Creating a Linking Directory)
(After we have a unique person index for each agency dataset)

- Linkage Determination – A Cutoff score needs to be set for each blocked comparison, below which a link is not accepted as a real “link”
- The best method of establishing this cutoff is for the system operator to work with a content-area expert to determine the peculiarities of data for that content-area
- In some data sets in may be very unlikely that a birthdate was entered incorrectly, while in another, it may happen very regularly – a computer can not automatically know this
- Once these cutoffs are set, they don’t need to be changed unless something drastic occurs to change the nature of the dataset