Building a Cloud-Based MIDI Pipeline

Ben Kopchick: bkopchick@deloitte.com
DE-IDENTIFICATION PIPELINE OVERVIEW

De-identification methodology made to follow TCIA protocols, pipeline is customizable to their level and improves the process through automation

Approach:
To address the MIDI need, Deloitte developed a Google Cloud based workflow to de-identify imaging data and test the performance of underlying algorithms.

- Multi-modal (MRI/PET/X-RAY/CT) image support
- Processing of DICOM meta-data and image-embedded data
- Context awareness to identify Research Critical Tags (RCTs) and potential PII/PHI burnt into the image
- Framework to measure performance of workflow with ability to utilize multiple algorithms developed using in-house tools (e.g., GCP-native vs. externally developed ML/AI based methods)
- Report with detailed information about identified PHI/PII and action taken
- Test dataset with synthetic PHI/PII from TCIA is used for benchmarking
MIDI PIPELINE TECHNICAL ARCHITECTURE

Cloud pipelines offer configurable systems that are scalable for large and growing datasets.

Pipeline Stages
1. Image Upload
2. Pipeline Triggered
3. Pre-process Images
4. Image to Dicom Store
5. Image De-Identified
6. De-Identification Logs Captured
7. Logs Searched
8. Analysis Performed
DE-IDENTIFICATION PROCESS

- Create datasets using GCP Cloud Healthcare API
  
  **Tag Options:**
  - Keep Tags
  - Remove Tags
  - Reset Tags
  - Clean Text Tags
  - Clean Image Tags
  - Recurse Tags

  **Text Transformation Options:**
  - Replace With Info Type Config
  - Date Shift Config

  **Tag Specification Options:**
  - Name
  - Hexadecimal ID
  - Value Representation (VR)

- Create DICOM store within dataset
TCIA SAMPLE DATA SET
The MIDI Pipeline was tested with multiple data sets to confirm accuracy in de-identification

<table>
<thead>
<tr>
<th>Tag</th>
<th>Orig</th>
<th>De-Id</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOP Instance UID</td>
<td>2.25.112784503178059210578740147414000844578</td>
<td>1.3.6.1.4.1.11129.5.1.116018550429166427175869...</td>
</tr>
<tr>
<td>Study Date</td>
<td>20130713</td>
<td>20130414</td>
</tr>
<tr>
<td>Series Date</td>
<td>20130713</td>
<td>20130414</td>
</tr>
<tr>
<td>Acquisition Date</td>
<td>20130713</td>
<td>20130414</td>
</tr>
<tr>
<td>Content Date</td>
<td>20130713</td>
<td>20130414</td>
</tr>
<tr>
<td>Accession Number</td>
<td>20130714E864535</td>
<td></td>
</tr>
<tr>
<td>Institution Name</td>
<td>Scott Community Hospital</td>
<td></td>
</tr>
<tr>
<td>Institution Address</td>
<td>334 Michael Manor Saratview, PA 56560</td>
<td></td>
</tr>
<tr>
<td>Referring Physician’s Address</td>
<td>0344 Green Inlet Jeffreyland, HI 66060</td>
<td></td>
</tr>
<tr>
<td>Study Description</td>
<td>XR CHEST AP PORTABLE for Douglas Davidson</td>
<td>XR CHEST AP PORTABLE for [PERSON_NAME]</td>
</tr>
<tr>
<td>Performing Physician’s Name</td>
<td>(E, R, C, W, N, A, P, E, T, E, R)</td>
<td></td>
</tr>
<tr>
<td>Patient’s Name</td>
<td>(D, A, V, I, D, S, C, N, A, D, O, U, G, L, A, S)</td>
<td></td>
</tr>
</tbody>
</table>

- Two data sets from TCIA have been run
  - The first contains 1,836 DICOM images and an accompanying answer key to validate our pipeline’s work
  - The second data set contains 23,921 images and was validated by a third party with TCIA answer key
The MIDI De-Identification Pipeline is performing at above 98% accuracy per action and at a fast rate.

RESULTS OF BENCHMARK (PRELIMINARY)

Time to de-identify 14,372 image slices

For 93 Patients, 14,372 image slices (4.5 GB)

0.017 sec/Image Average De-Identify Time

4 min 6 sec Average Run Time Total
RESULTS OF BENCHMARK (PRELIMINARY)

The MIDI De-Identification Pipeline is performing at above 98% accuracy per action and at a fast rate.

<table>
<thead>
<tr>
<th>Action Taken</th>
<th>Percent Correct</th>
<th>Dataset 1</th>
<th>Dataset 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Retained</td>
<td>99.5%</td>
<td>99.2%</td>
<td></td>
</tr>
<tr>
<td>Text Not Null</td>
<td>99.5%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Pixels Hidden</td>
<td>99.5%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Date Shifted</td>
<td>100%</td>
<td>98.3%</td>
<td></td>
</tr>
<tr>
<td>Text Removed</td>
<td>99.5%</td>
<td>84.7%</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>99.7%</strong></td>
<td><strong>98.7%</strong></td>
<td></td>
</tr>
</tbody>
</table>

Accuracy for Actions (Dataset 1)
RESULTS OF BENCHMARK

All PHI/PII pixels were correctly identified and removed.

- **True Positive Image De-Identification**
- Name and dates correctly identified as PHI and removed
- Non-PHI data correctly retained
RESULTS OF BENCHMARK

Two false positives in the burnt-in image data (i.e., data was removed unnecessarily) were identified.

- False Positive Image
- Incorrectly identified PHI partially covered up image
## RESULTS OF BENCHMARK

### False Negative
- Text failed to be removed (fixed in pre-processing)

### False Positives
- Software version mistaken as IP address

### Name Issues
- Names containing underscore not correctly identified: 
  e.g., A_John Doe
- Non-names that can be mistaken for names: 
  e.g., MR Header
- Non-western and atypical names: 
  e.g., Bhavani Singh

### Date Issues
- Dates are not easily recognized in non-Date fields 
  (fixed in pre-processing)
DISCOVERIES DURING DE-ID PROCESS

- The use of crypto hashes can lead to failure in following the DICOM format
  - Many tags data elements have character limits that this fails to follow
  - Other options include using a placeholder ("[PERSON_NAME]") or erasing text
- Addresses and some names appeared to be partially de-identified
  - This is due to Google’s NLP searching for real addresses
  - In the provided TCIA dataset, addresses were not real

<table>
<thead>
<tr>
<th>action</th>
<th>tag</th>
<th>origval</th>
<th>deidval</th>
</tr>
</thead>
<tbody>
<tr>
<td>text_removed</td>
<td>&lt;PatientAddress&gt;</td>
<td>7296 Wyatt Light Suite 457 Port Kristi, CO 16956</td>
<td>7296 6U0oTcqvdvXfuuEHLvdIfu2PwXAJV1dauuQLLQ/2Fw= Light Suite 457 Port HdwdWsEF/O6I/EzdzgXDSvr8WhoyTA8PpT9fGnyQ=, CO 16956</td>
</tr>
<tr>
<td>text_removed</td>
<td>&lt;PatientAddress&gt;</td>
<td>7025 James Ford Suite 835 South Edwardfurl, NM 87894</td>
<td>7025 HHlsDiKdCsb9xMrDZHdGRVJDPAdC90Wbkp5uGgQ= Suite 835 South Edwardfurl, NM 87894</td>
</tr>
</tbody>
</table>
DISCOVERIES DURING DE-ID PROCESS

- The use of crypto hashes can lead to failure in following the DICOM format
  - Many tags data elements have character limits that this fails to follow
  - Other options include using a placeholder ("[PERSON_NAME]") or erasing text
- Addresses and some names appeared to be partially de-identified
  - This is due to Google’s NLP searching for real addresses
  - In the provided TCIA dataset, addresses were not real

<table>
<thead>
<tr>
<th>action</th>
<th>tag</th>
<th>origval</th>
<th>deidval</th>
</tr>
</thead>
<tbody>
<tr>
<td>text_removed</td>
<td>&lt;PatientAddress&gt;</td>
<td>7296 Wyatt Light Suite 457 Port, Kristi, CO 16956</td>
<td>7296 6U0oTcqhdvXfUQrEHlvdfu2PwXAJV1dauuQLLQ/2Fw= Light Suite 457 Port HdwdsEF/O6I/EzdgzXDSvrD8WhotyTA8PpTy9fGnyQ=, CO 16956</td>
</tr>
<tr>
<td>text_removed</td>
<td>&lt;PatientAddress&gt;</td>
<td>7025 James Ford Suite 835, South Edwardfurt, NM 87894</td>
<td>7025 HHlsDlKdCsb9xMrDZSHHzGvJDPAedC90Wbwp5uHgQ= Suite 835 South Edwardfurt, NM 87894</td>
</tr>
</tbody>
</table>

**Zip Codes:**
16946 => Sweden
87894 => Mexico
The use of crypto hashes can lead to failure in following the DICOM format
- Many tags data elements have character limits that this fails to follow
- Other options include using a placeholder ("[PERSON_NAME]") or erasing text

Addresses and some names appeared to be partially de-identified
- This is due to Google’s NLP searching for real addresses
- In the provided TCIA dataset, addresses were not real

<table>
<thead>
<tr>
<th>action</th>
<th>tag</th>
<th>origval</th>
<th>deidval</th>
</tr>
</thead>
<tbody>
<tr>
<td>text_removed</td>
<td>&lt;PatientAddress&gt;</td>
<td>7296 Wyatt Light Suite 457 Port Kristi, CO 16956</td>
<td>7296 6U0oTcqhdvXfUqrEHLvlIfu2PwXAJV1dauuQLLQ/2Fw= Light Suite 457 Port HdwdWsEF/O6I/EzdgzXDSvrd8WhotyTA8PpTy9fGnyQ=, CO 16956</td>
</tr>
<tr>
<td>text_removed</td>
<td>&lt;PatientAddress&gt;</td>
<td>7025 James Ford Suite 835 South Edwardfurt, NM 87894</td>
<td>7025 HHlsDlKdCsb9xMrDZSHHzGRvJDPAedC90Wb kp5uHgQ= Suite 835 South Edwardfurt, NM 87894</td>
</tr>
</tbody>
</table>

Zip Codes:
- 16946 => Sweden
- 87894 => Mexico

Cities:
- Do not exist
DISCOVERIES DURING DE-ID PROCESS

• The use of crypto hashes can lead to failure in following the DICOM format
  • Many tags data elements have character limits that this fails to follow
  • Other options include using a placeholder (“[PERSON_NAME]”) or erasing text

• Addresses and some names appeared to be partially de-identified
  • This is due to Google’s NLP searching for real addresses
  • In the provided TCIA dataset, addresses were not real

<table>
<thead>
<tr>
<th>action</th>
<th>tag</th>
<th>origval</th>
<th>deidval</th>
</tr>
</thead>
<tbody>
<tr>
<td>text_removed</td>
<td>&lt;PatientAddress&gt;</td>
<td>7296 Wyatt Light Suite 457 Port 7296 U0oTcqhdvXfUQrEHLvdfu2PwXAJV1dauuQLLQ/2Fw= Light Suite 457 Port HwdWsEF/O6I/EzdgzXDSvrd8WhotyTA8PpTy9fGnyQ=, CO 16956</td>
<td></td>
</tr>
<tr>
<td>text_removed</td>
<td>&lt;PatientAddress&gt;</td>
<td>7025 James Ford Suite 835 South Edwardfurt, NM 87894</td>
<td>7025 HHlsDlKdCsb9xMrDZSHHzGrVJDPAdC90Wbkp5uHgQ= Suite 835 South Edwardfurt, NM 87894</td>
</tr>
</tbody>
</table>

Zip Codes:
16946 => Sweden
87894 => Mexico

Names:
Wyatt, Kristi, James Ford recognized as names

Cities:
Do not exist
CONCLUSIONS

The Google Healthcare API DICOM De-Identification service shows great promise as a viable option and further testing is recommended before being deployed in a production environment.

- Many of the tools used are in Open Beta
  - Further software changes could be made that could improve the pipeline and need to be tested on release
- Automated analysis of pixel removal can be used to identify false-positives
- Pre- and post-processing can catch many errors we currently find
- Can implement other solutions on top of Healthcare API, the cloud will allow other software to be used in pipeline
- A human-in-the-loop is still recommended to Quality Check images
  - Combining the efforts of a human expert and de-identification service will increase the accuracy (compared to using either alone) and speed up the process
Acknowledgements

Keyvan Farahani  David Clunie
Ulrike Wagner    Qinyan Pan
Michael Rutherford Scott Gustafson
Fred Prior       Juergen Klenk
Laura Opsahl-Ong Bhavani Singh
Bob Lou          Cheryl Corman
David Belardo    Anne Billak
Kathryn Johnson

Ben Kopchick: bkopchick@deloitte.com