Deloitte.

NIH NATIONAL CANCER INSTITUTE

Building a Cloud-Based MIDI Pipeline

Ben Kopchick: bkopchick@deloitte.com

Premier Partner

Google Cloud

DE-IDENTIFICATION PIPELINE OVERVIEW

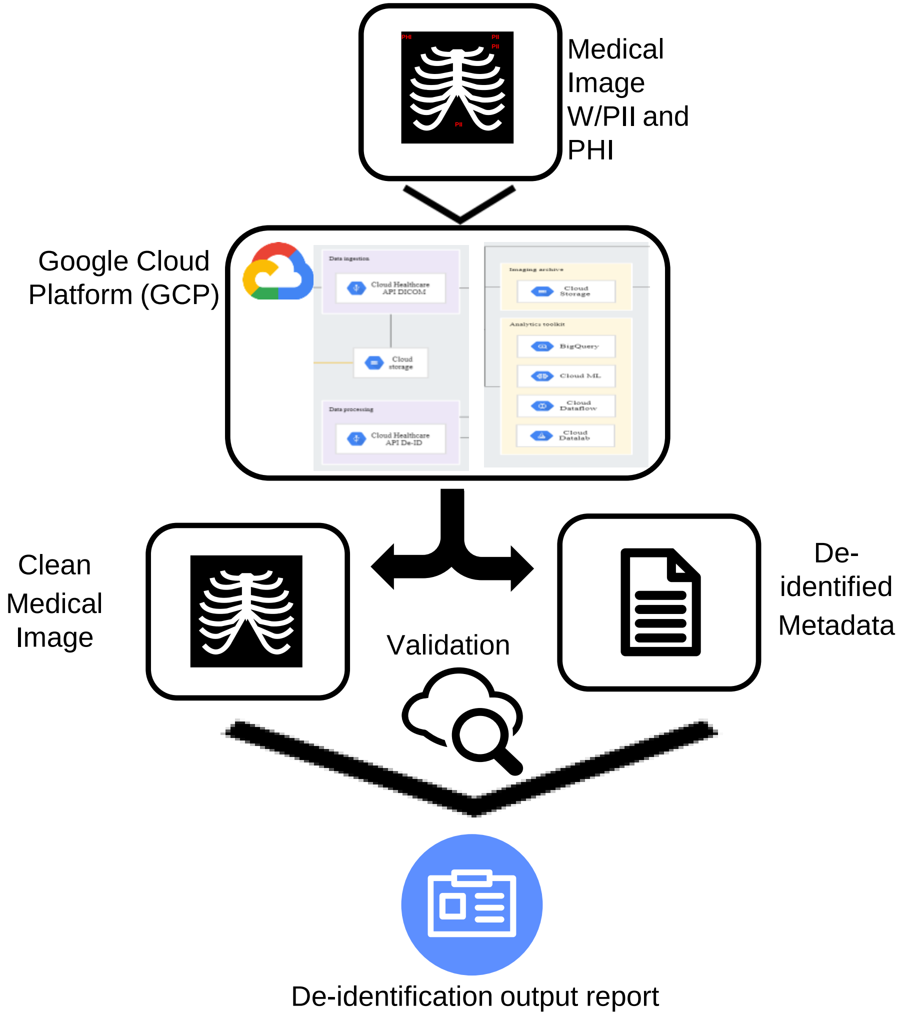
De-identification methodology made to follow TCIA protocols, pipeline is customizable to their level and improves the process through automation

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.

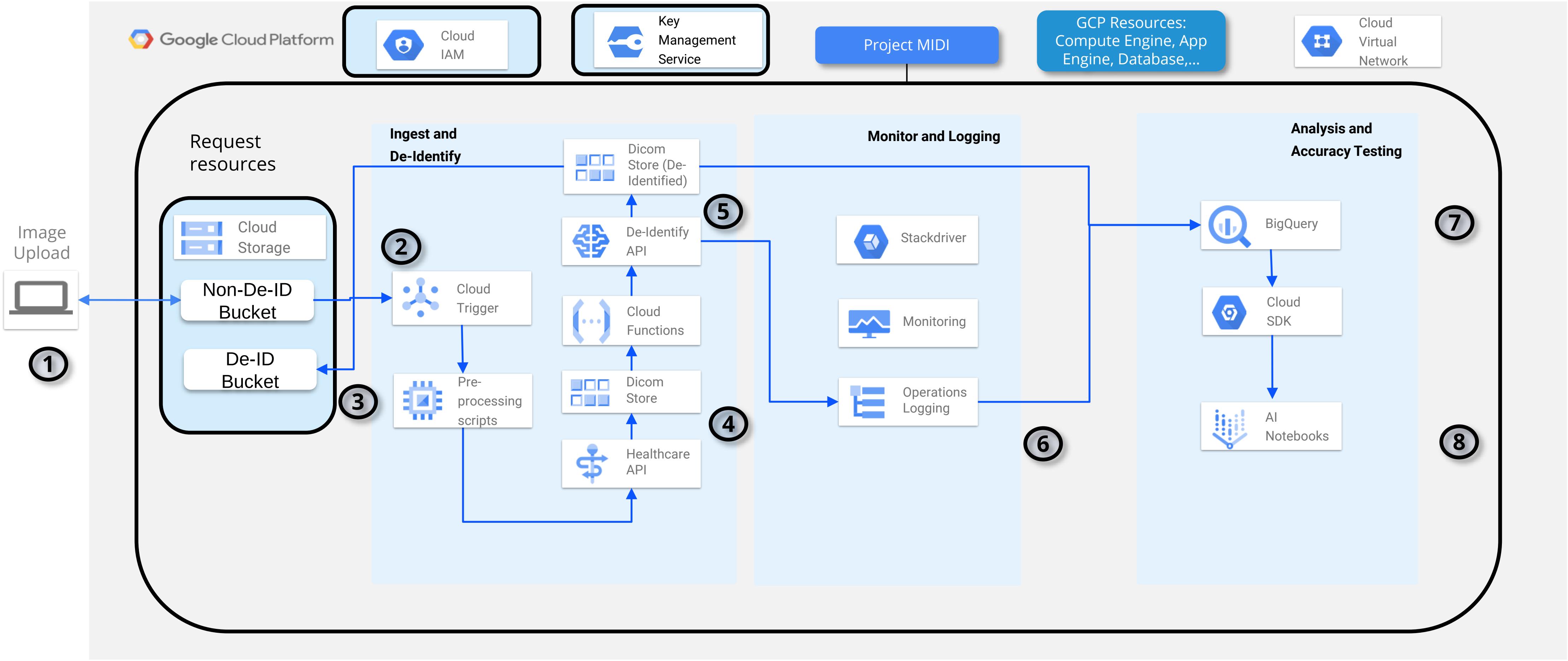
- Multi-modal (MRI/PET/X-RAY/CT) image support
- Processing of DICOM meta-data and image-embedded data
- Context awareness to identify Research Critical Tags (RCTs) and potential PII/PHI burnt into the image
- Framework to measure performance of workflow with ability to utilize multiple algorithms developed 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

Medical Image De-Identification Pipeline



MIDI PIPELINE TECHNICAL ARCHITECTURE

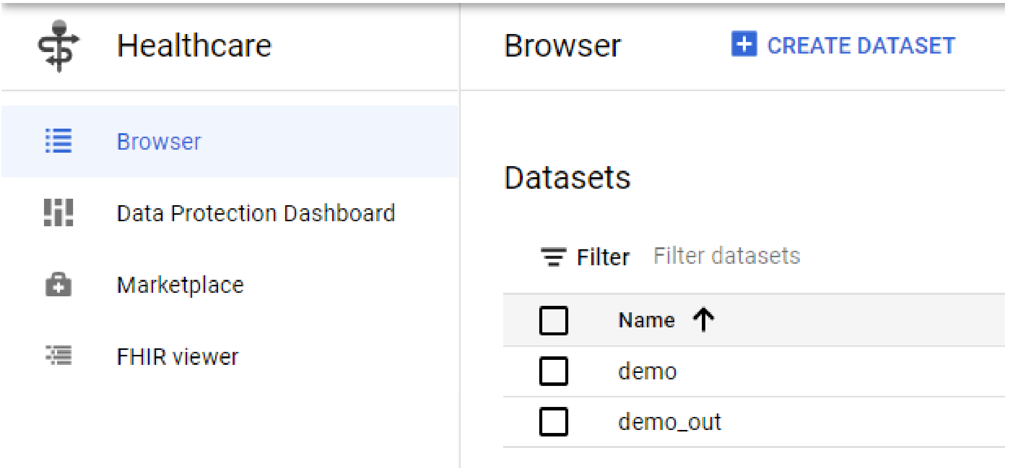
Cloud pipelines offer configurable systems that are scalable for large and growing datasets.



Pipeline Stages

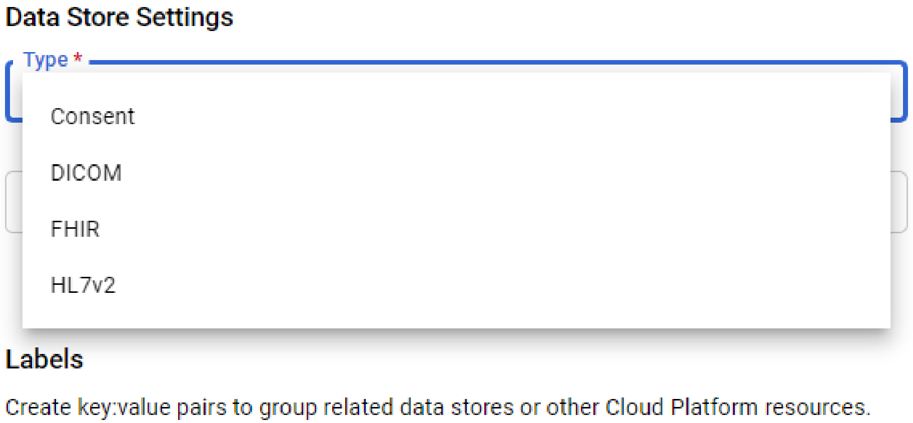
1. Image Upload
2. Pipeline Triggered
3. Pre-process Images
4. Image to Dicom Store
5. Image De-Identified
6. De-Identification Logs Captured
7. Logs Searched
8. Analysis Performed

DE-IDENTIFICATION PROCESS



☐	Name ↑
☐	demo
☐	demo_out

- Create datasets using GCP Cloud Healthcare API



Create key:value pairs to group related data stores or other Cloud Platform resources.

- Create DICOM store within dataset

Tag Options:

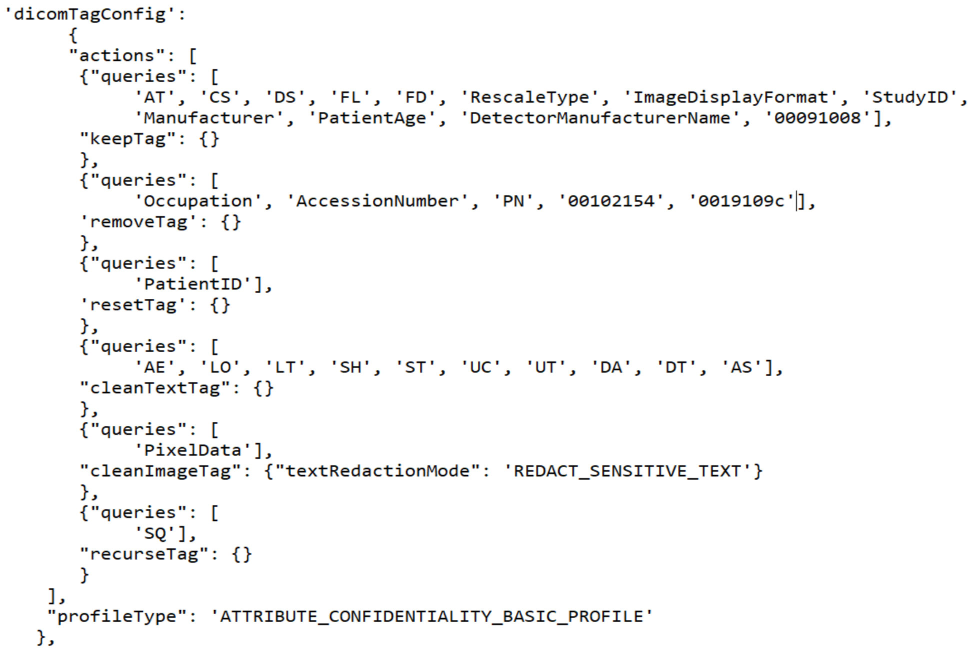
- Keep Tags
- Remove Tags
- Reset Tags
- Clean Text Tags
- Clean Image Tags
- Recurse Tags

Text Transformation Options:

- Replace With Info Type Config
- Date Shift Config

Tag Specification Options:

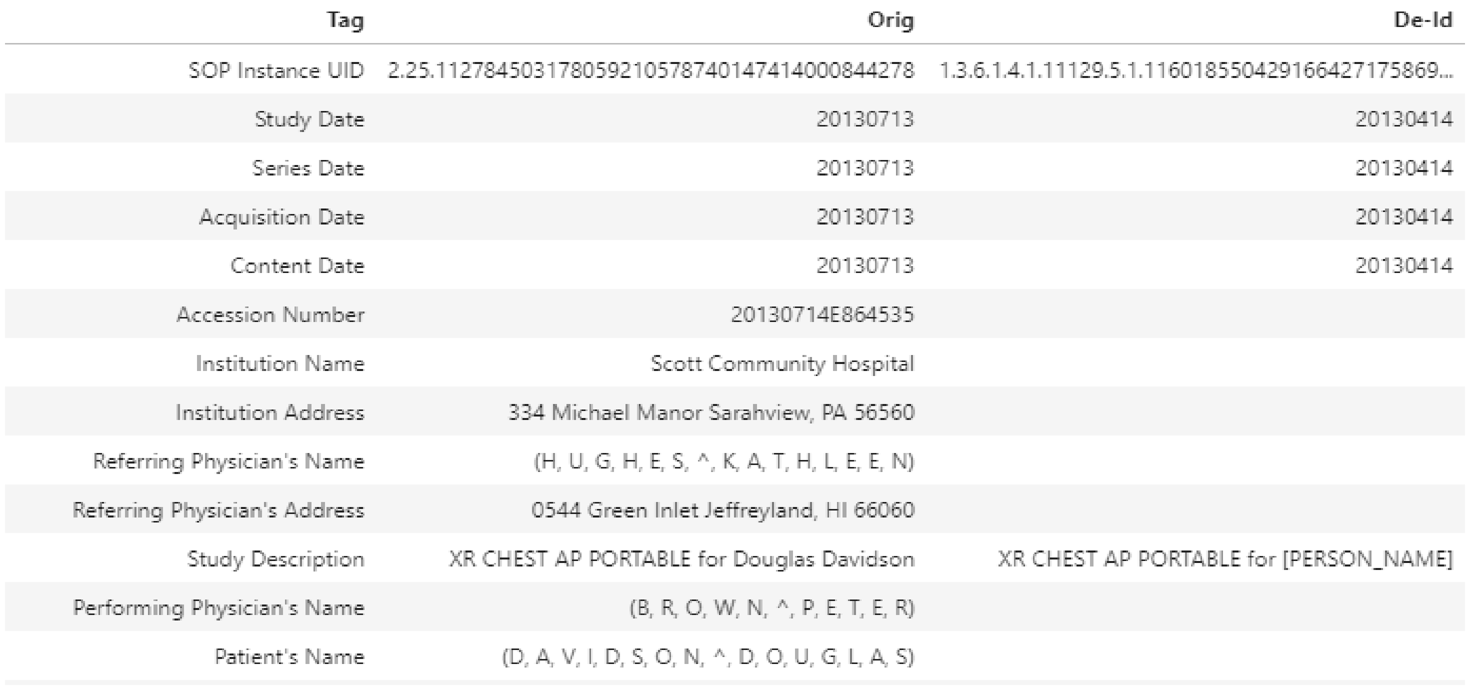
- Name
- Hexadecimal ID
- Value Representation (VR)



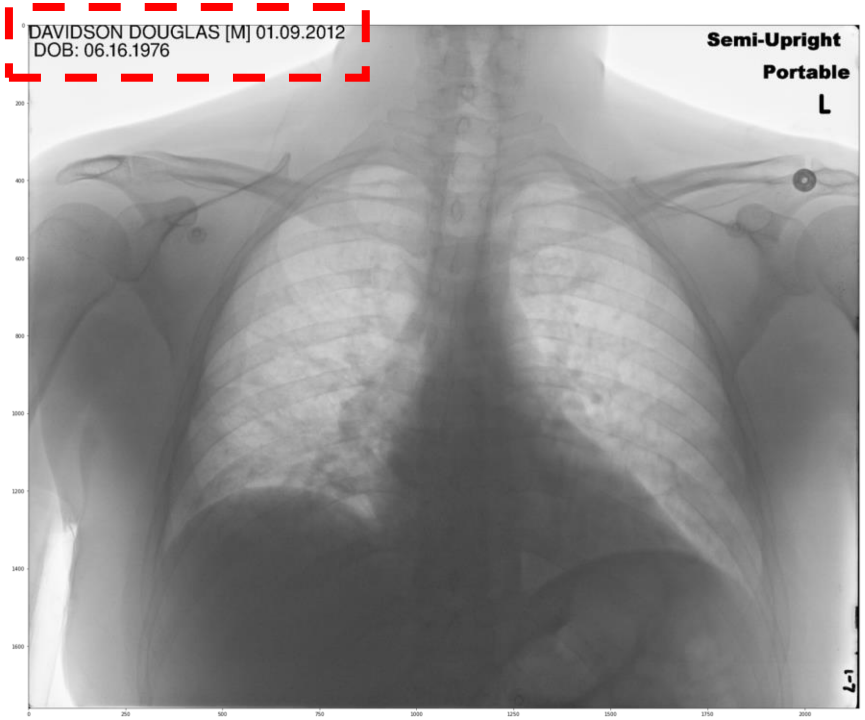
```
'dicomTagConfig':
      {
      "actions": [
       {"queries": [
            'AT', 'CS', 'DS', 'FL', 'FD', 'RescaleType', 'ImageDisplayFormat', 'StudyID',
            'Manufacturer', 'PatientAge', 'DetectorManufacturerName', '00091008'],
       "keepTag": {}
       },
       {"queries": [
            'Occupation', 'AccessionNumber', 'PN', '00102154', '0019109c'],
       'removeTag': {}
       },
       {"queries": [
            'PatientID'],
       'resetTag': {}
       },
       {"queries": [
            'AE', 'LO', 'LT', 'SH', 'ST', 'UC', 'UT', 'DA', 'DT', 'AS'],
       "cleanTextTag": {}
       },
       {"queries": [
            'PixelData'],
       "cleanImageTag": {"textRedactionMode": 'REDACT_SENSITIVE_TEXT'}
       },
       {"queries": [
            'SQ'],
       "recurseTag": {}
       }
     ],
     "profileType": 'ATTRIBUTE_CONFIDENTIALITY_BASIC_PROFILE'
    },
```

TCIA SAMPLE DATA SET

The MIDI Pipeline was tested with multiple data sets to confirm accuracy in de-identification



Tag	Orig	De-Id
SOP Instance UID	2.25.112784503178059210578740147414000844278	1.3.6.1.4.1.11129.5.1.116018550429166427175869...
Study Date	20130713	20130414
Series Date	20130713	20130414
Acquisition Date	20130713	20130414
Content Date	20130713	20130414
Accession Number	20130714E864535	
Institution Name	Scott Community Hospital	
Institution Address	334 Michael Manor Sarahview, PA 56560	
Referring Physician's Name	(H, U, G, H, E, S, ^, K, A, T, H, L, E, E, N)	
Referring Physician's Address	0544 Green Inlet Jeffreyland, HI 66060	
Study Description	XR CHEST AP PORTABLE for Douglas Davidson	XR CHEST AP PORTABLE for [PERSON_NAME]
Performing Physician's Name	(B, R, O, W, N, ^, P, E, T, E, R)	
Patient's Name	(D, A, V, I, D, S, O, N, ^, D, O, U, G, L, A, S)	

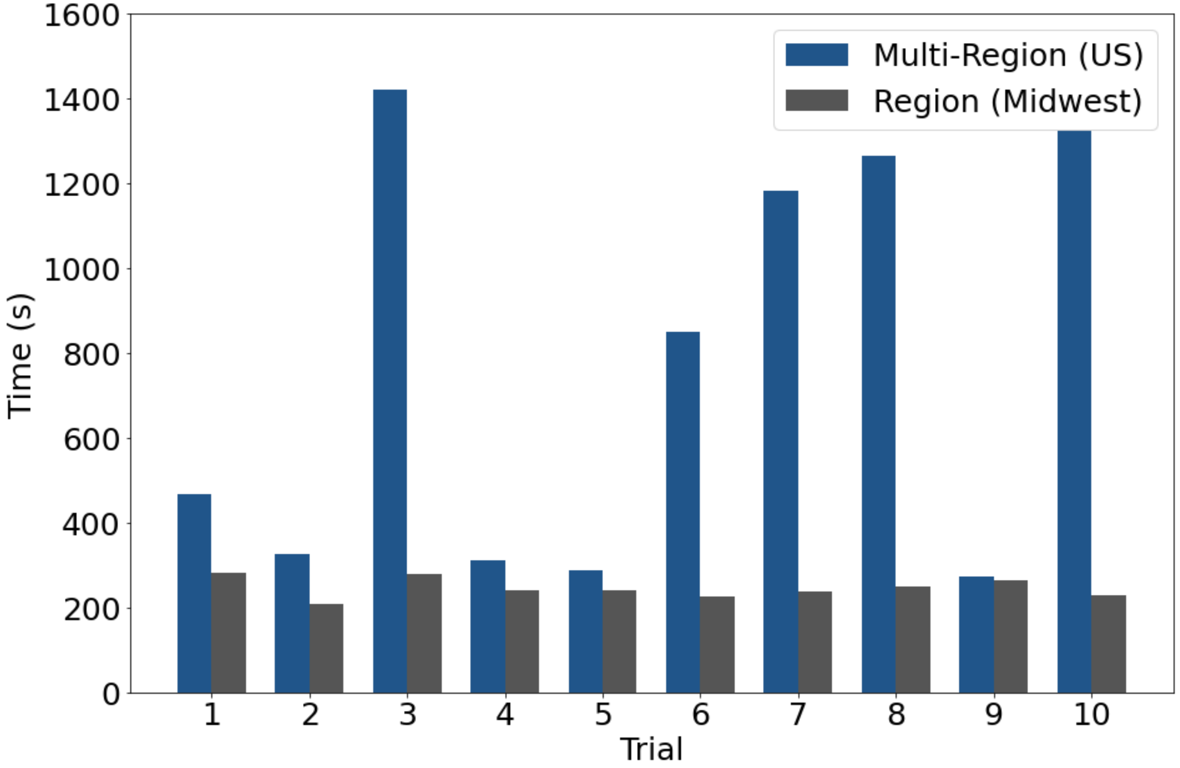


- Two data sets from TCIA have been run
  - The first contains 1,836 DICOM images and an accompanying answer key to validate our pipeline’s work
  - The second data set contains 23,921 images and was validated by a third party with TCIA answer key

RESULTS OF BENCHMARK (PRELIMINARY)

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

Time to de-identify 14,372 image slices



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

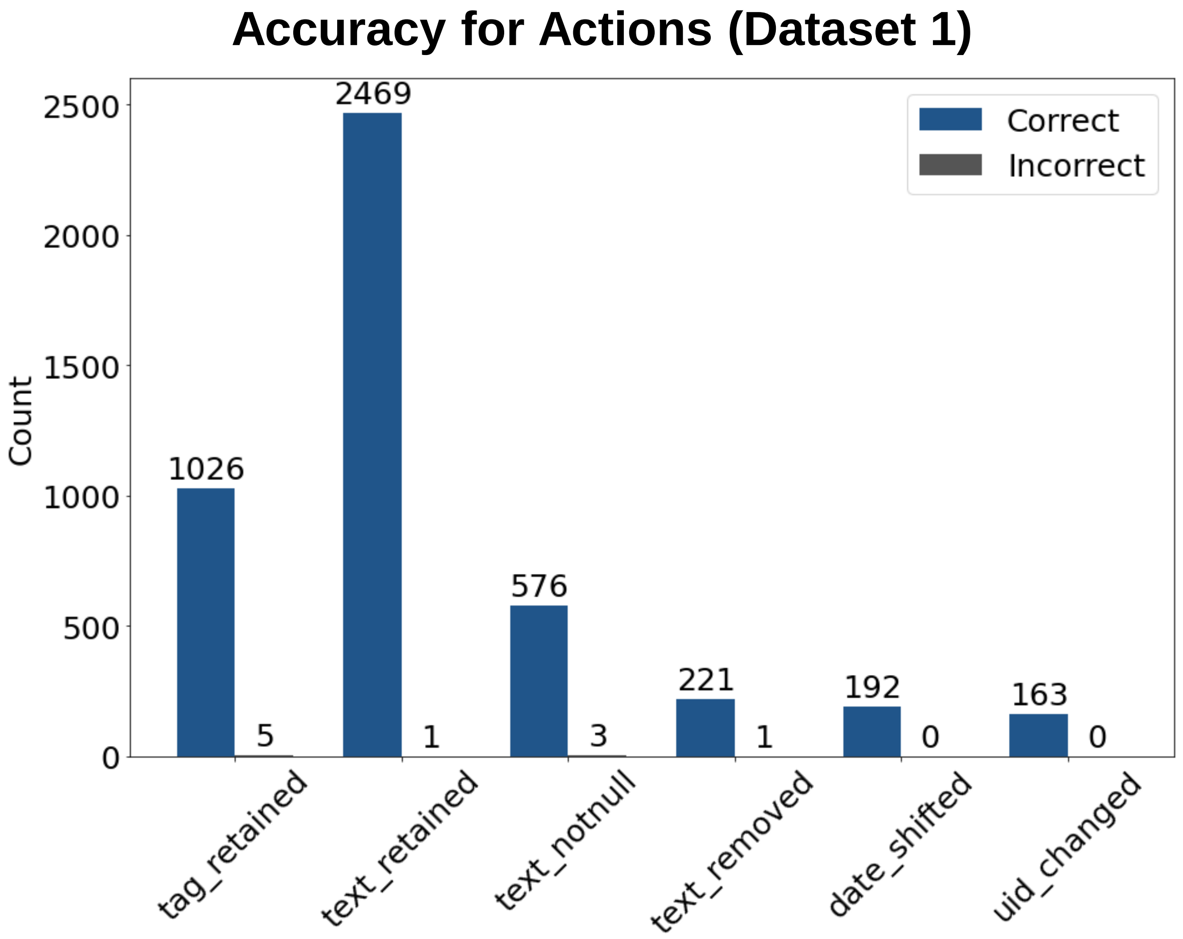
.017 sec/Image Average De-Identify Time

4 min 6 sec Average Run Time Total

RESULTS OF BENCHMARK (PRELIMINARY)

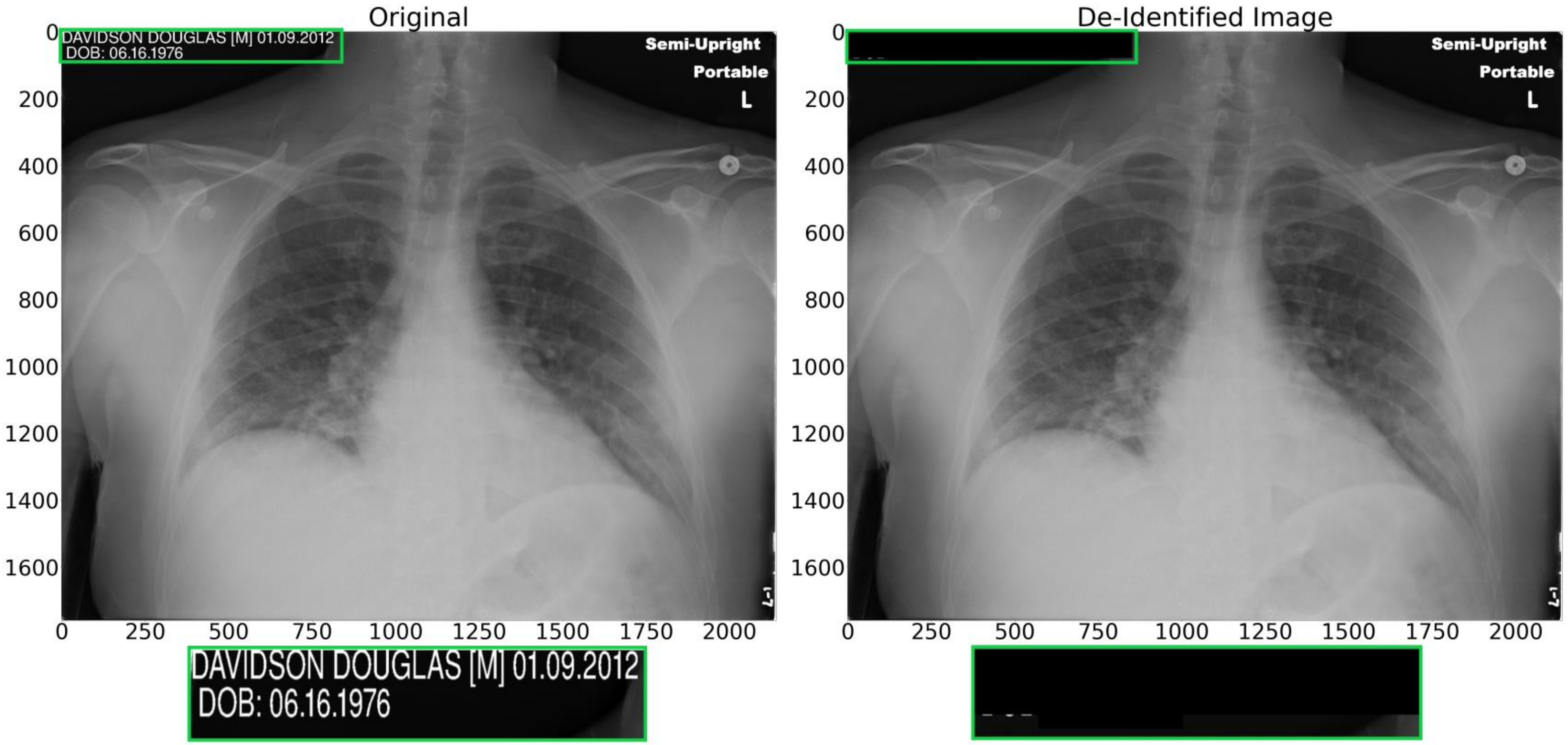
The MIDI De-Identification Pipeline is performing at above 98% accuracy per action and at a fast rate.

Action Taken	Percent Correct	
	Dataset 1	Dataset 2
Text Retained	99.5%	99.2%
Text Not Null	99.5%	100%
Pixels Hidden	99.5%	100%
Date Shifted	100%	98.3%
Text Removed	99.5%	84.7%
Total	**99.7%**	**98.7%**



RESULTS OF BENCHMARK

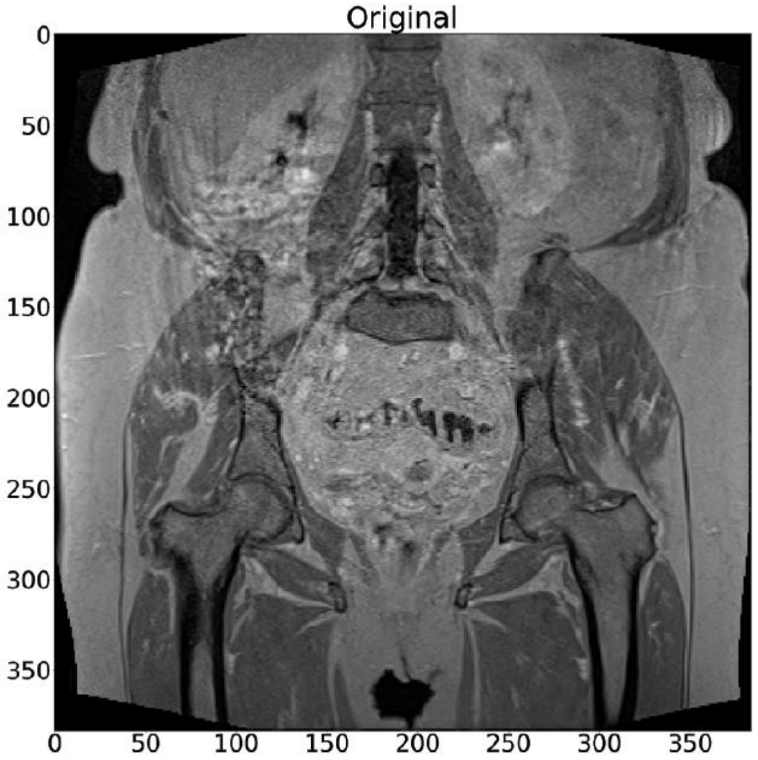
All PHI/PII pixels were correctly identified and removed.

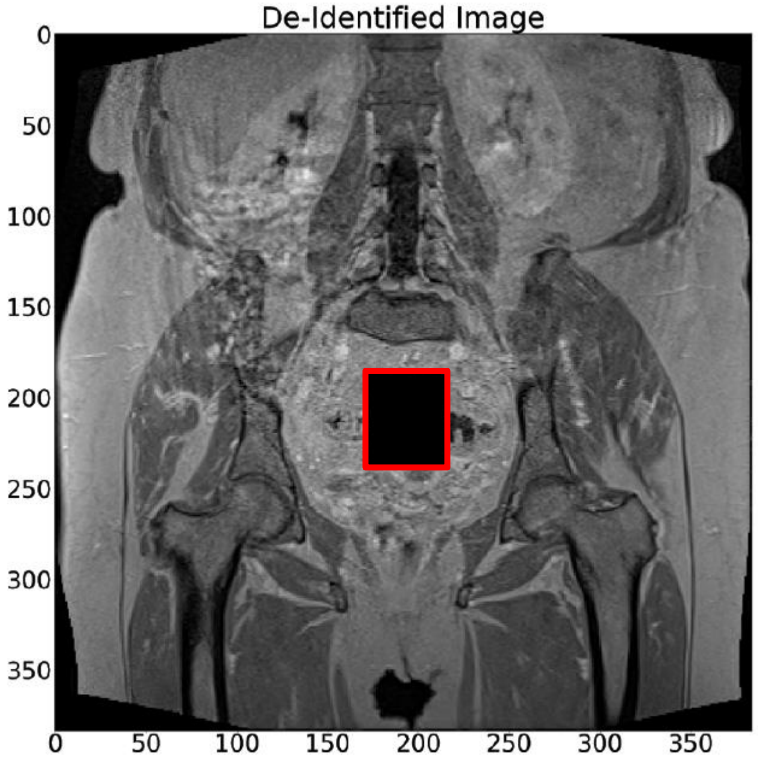


- **True Positive Image De-Identification**
- Name and dates correctly identified as PHI and removed
- Non-PHI data correctly retained

RESULTS OF BENCHMARK

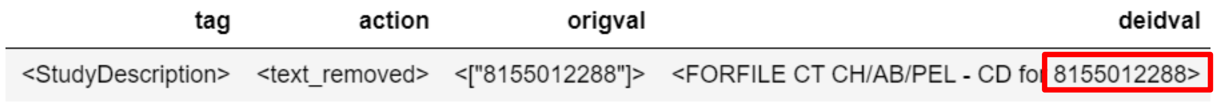
Two false positives in the burnt-in image data (i.e., data was removed unnecessarily) were identified.





- **False Positive Image**
- Incorrectly identified PHI partially covered up image

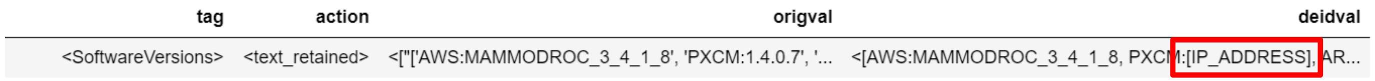
RESULTS OF BENCHMARK



tag	action	origval	deidval
<StudyDescription>	<text_removed>	<["8155012288"]>	<FORFILE CT CH/AB/PEL - CD for 8155012288>

False Negative

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



tag	action	origval	deidval
<SoftwareVersions>	<text_retained>	<["['AWS:MAMMODROC_3_4_1_8', 'PXCM:1.4.0.7', '...	<[AWS:MAMMODROC_3_4_1_8, PXCM:[IP_ADDRESS], AR...

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:

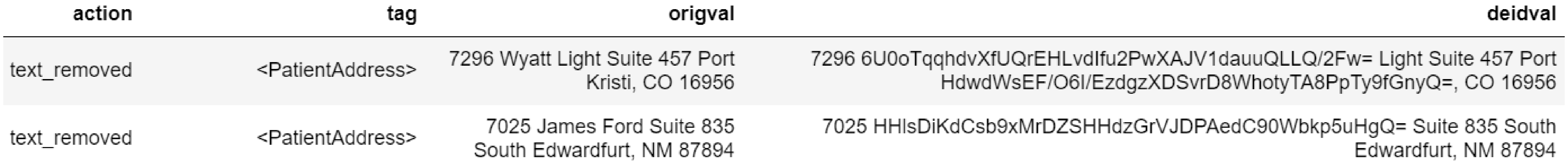
  e.g., Bhavani Singh

Date Issues

- Dates are not easily recognized in non-Date fields (fixed in pre-processing)

DISCOVERIES DURING DE-ID PROCESS

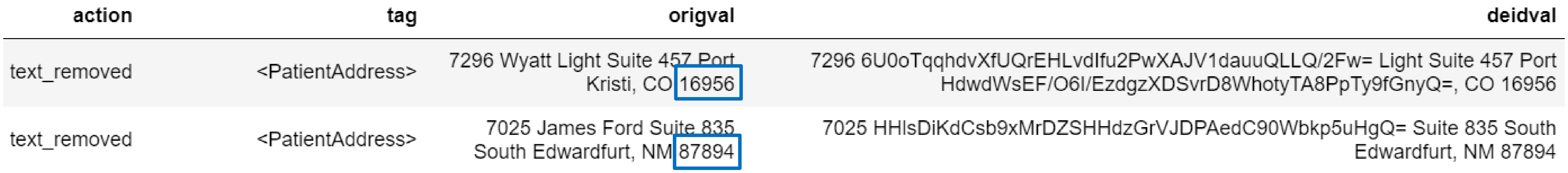
- The use of crypto hashes can lead to failure in following the DICOM format
  - Many tags data elements have character limits that this fails to follow
  - Other options include using a placeholder (“[PERSON_NAME]”) or erasing text
- Addresses and some names appeared to be partially de-identified
 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



action	tag	origval	deidval
text_removed	<PatientAddress>	7296 Wyatt Light Suite 457 Port Kristi, CO 16956	7296 6U0oTqqhdvXfUQrEHLvdIfu2PwXAJV1dauuQLLQ/2Fw= Light Suite 457 Port HdwdWsEF/O6I/EzdgzXDSvrD8WhotyTA8PpTy9fGnyQ=, CO 16956
text_removed	<PatientAddress>	7025 James Ford Suite 835 South Edwardfurt, NM 87894	7025 HHlsDiKdCsb9xMrDZSHHdzGrVJDPAedC90Wbkp5uHgQ= Suite 835 South Edwardfurt, NM 87894

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

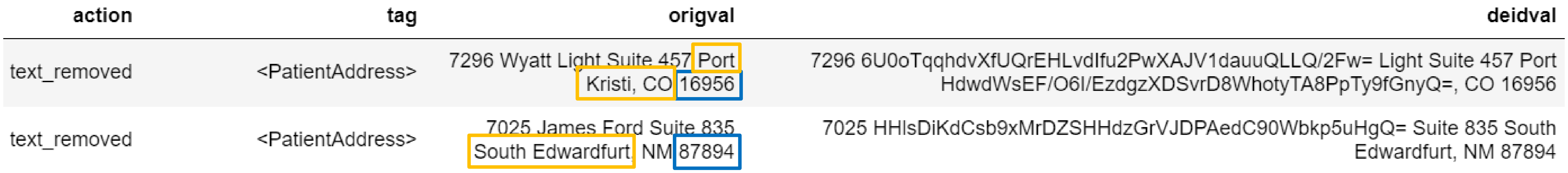


action	tag	origval	deidval
text_removed	<PatientAddress>	7296 Wyatt Light Suite 457 Port Kristi, CO 16956	7296 6U0oTqqhdvXfUQrEHLvdIfu2PwXAJV1dauuQLLQ/2Fw= Light Suite 457 Port HdwdWsEF/O6I/EzdgzXDSvrD8WhotyTA8PpTy9fGnyQ=, CO 16956
text_removed	<PatientAddress>	7025 James Ford Suite 835 South Edwardfurt, NM 87894	7025 HHlsDiKdCsb9xMrDZSHHdzGrVJDPAedC90Wbkp5uHgQ= Suite 835 South Edwardfurt, NM 87894

Zip Codes:
16946 => Sweden
87894 => Mexico

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



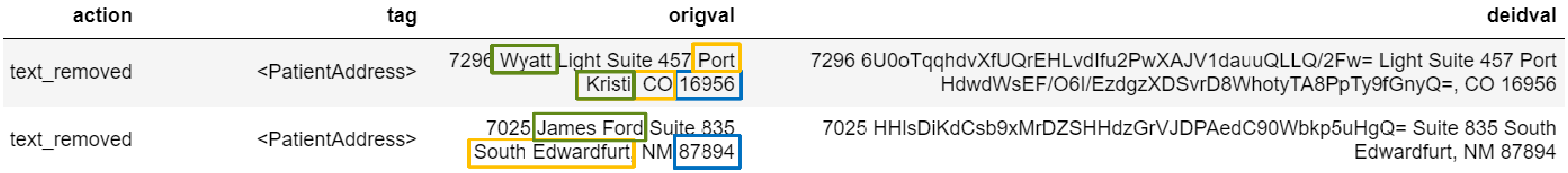
action	tag	origval	deidval
text_removed	<PatientAddress>	7296 Wyatt Light Suite 457 Port Kristi, CO 16956	7296 6U0oTqqhdvXfUQrEHLvdIfu2PwXAJV1dauuQLLQ/2Fw= Light Suite 457 Port HdwdWsEF/O6I/EzdgzXDSvrD8WhotyTA8PpTy9fGnyQ=, CO 16956
text_removed	<PatientAddress>	7025 James Ford Suite 835 South Edwardfurt, NM 87894	7025 HHlsDiKdCsb9xMrDZSHHdzGrVJDPAedC90Wbkp5uHgQ= Suite 835 South Edwardfurt, NM 87894

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



action	tag	origval	deidval
text_removed	<PatientAddress>	7296 Wyatt Light Suite 457 Port Kristi, CO 16956	7296 6U0oTqqhdvXfUQrEHLvdIfu2PwXAJV1dauuQLLQ/2Fw= Light Suite 457 Port HdwdWsEF/O6I/EzdgzXDSvrD8WhotyTA8PpTy9fGnyQ=, CO 16956
text_removed	<PatientAddress>	7025 James Ford Suite 835 South Edwardfurt, NM 87894	7025 HHlsDiKdCsb9xMrDZSHHdzGrVJDPAedC90Wbkp5uHgQ= Suite 835 South Edwardfurt, NM 87894

Zip Codes:
16946 => Sweden
87894 => Mexico

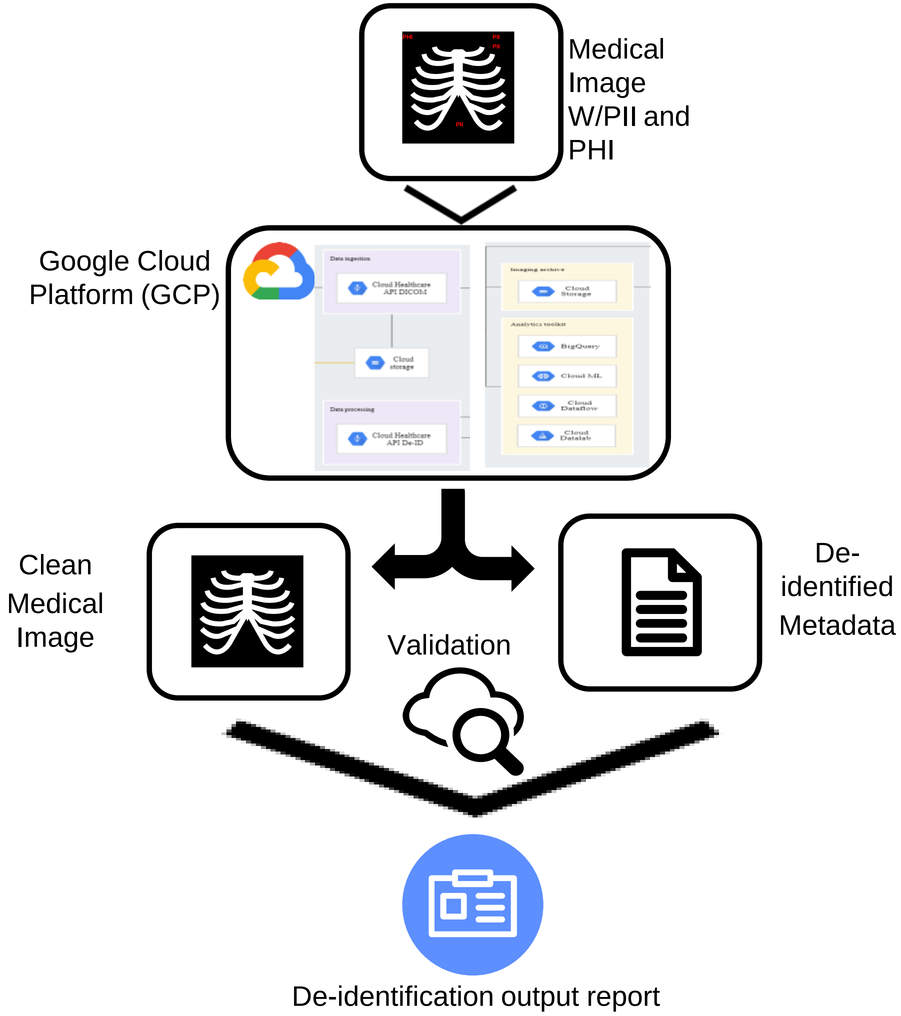
Names:
Wyatt, Kristi, James Ford recognized as names

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

Ulrike Wagner

Michael Rutherford

Fred Prior

Laura Opsahl-Ong

Bob Lou

David Belardo

Kathryn Johnson

David Clunie

Qinyan Pan

Scott Gustafson

Juergen Klenk

Bhavani Singh

Cheryl Corman

Anne Billak

Ben Kopchick: bkopchick@deloitte.com