Agency for Healthcare Research and Quality
Advancing Excellence in Health Care

USING MACHINE LEARNING TO ABSTRACT HEALTH INSURANCE BOOKLET COST SHARING FOR THE MEDICAL EXPENDITURE PANEL SURVEY

April 13, 2021

Medical Expenditure Panel Survey Components Related to this Project

MEPS Household Component - HC

Survey of ~ 15,000 households conducted via CAPI (supplemented by CATI in 2020 and 2021) in multiple rounds over several years

MEPS Medical Provider Component - MPC

Collection of administrative records from medical providers (hospitals (inpatient, emergency room and outpatient care)), office based providers, home health care agencies, and pharmacies who MEPS HC respondents had seen for health care in the preceding year

.

Together these provide national estimates of healthcare use, expenditures, insurance coverage, sources of payment, access to care and healthcare quality

Project Overview

PURPOSE

Develop an abstraction tool and database of health insurance policy booklet cost-sharing data collected from households participating in the Medical Expenditure Survey – Household Component (MEPS-HC).

GOAL

Link data to the MEPS-HC so that information about health insurance coverage generosity (deductibles, service cost-sharing) may be used to enrich research conducted with the MEPS.

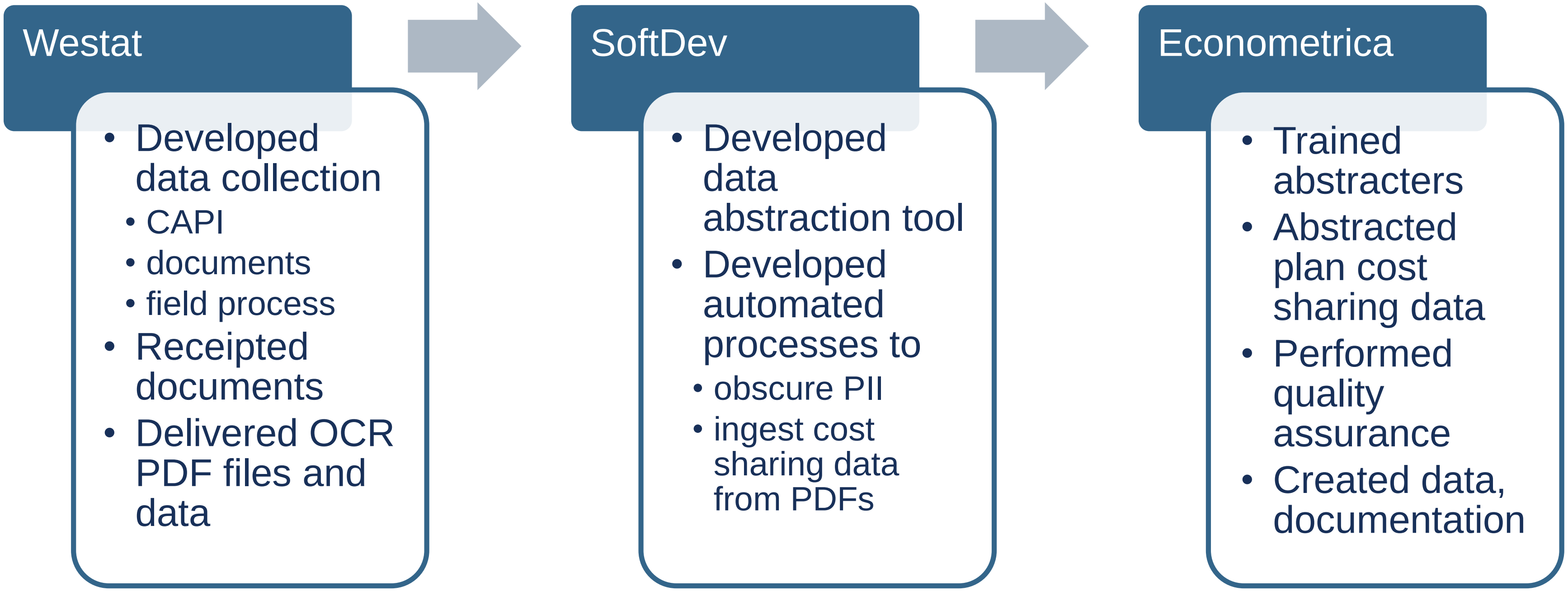
Insurance Booklet Types

Medicare Advantage (Part C) or prescription drug benefits (Part D).
Private insurance through an employer without Medicare coverage.
Private insurance (employer or non-employer) with Medicare coverage.
All other types of eligible private insurance plans. • Private plans that are not through an employer without Medicare. • Private insurance (employer or non-employer) that covers any person in the household even when policyholder is not living in the household.

Documents Requested for Abstraction

- What did MEPS-HC request?
  - Evidence of Coverage (EOC)
  - Summary of Benefits and Coverage (SBC)
- What did MEPS-HC accept?
  - Any insurance document with cost sharing information

Overview of Process



Westat

- Developed data collection
  - CAPI
  - documents
  - field process
- Receipted documents
- Delivered OCR PDF files and data

SoftDev

- Developed data abstraction tool
- Developed automated processes to
  - obscure PII
  - ingest cost sharing data from PDFs

Econometrica

- Trained abstracters
- Abstracted plan cost sharing data
- Performed quality assurance
- Created data, documentation

Handling of Survey Responses

- Files were scanned for the following:
  - Document type
  - Presence of a plan name
  - Plan name matches with plan from interview
  - Presence of coverage period on document
  - Whether coverage period is current
  - Presence of cost-sharing information
  - Whether document appears complete

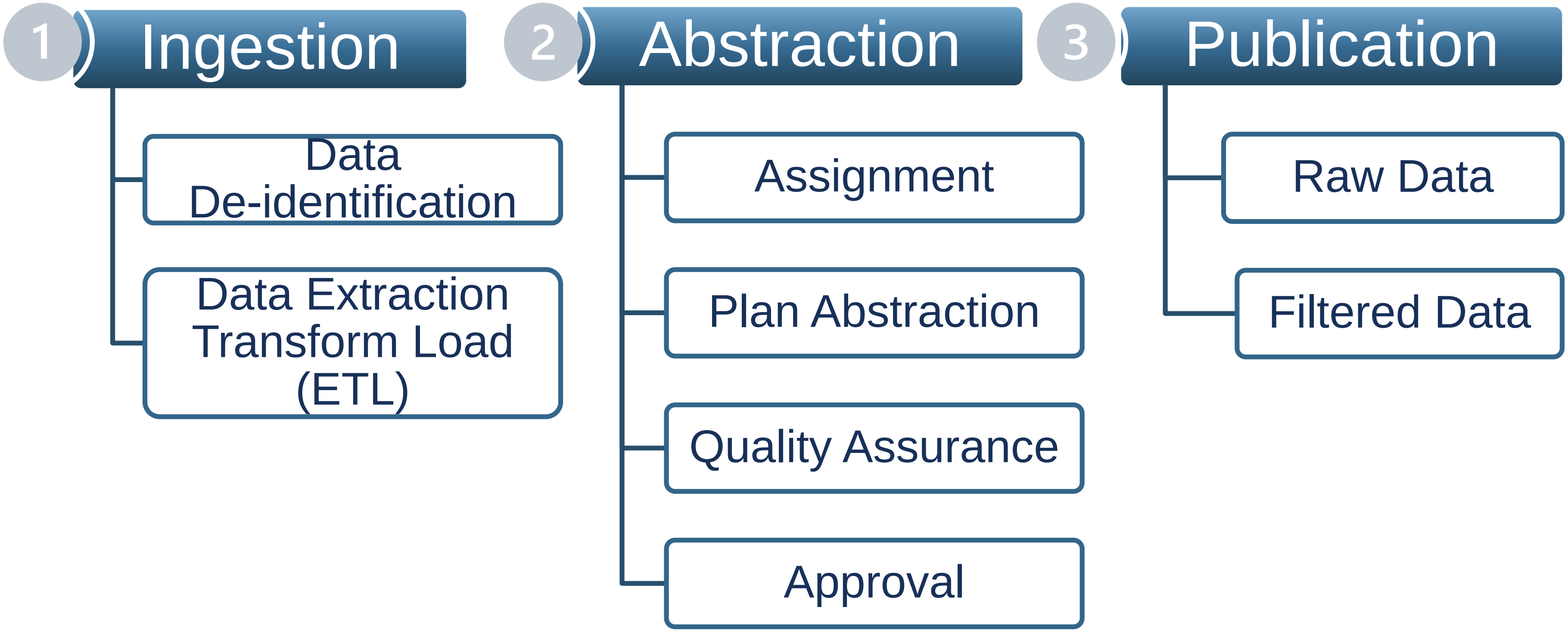
Policy Booklet Response Rates

Protocol Folder	Total Number		Response Rate (%)
	Requested	Received	
Medicare Parts C and D	3386	1142	33.73
Private w/o Medicare	6049	1826	30.19
Private with Medicare	1490	450	30.20
Other Private	1525	370	24.26
Total	**12450**	**3788**	**30.43**

SOFTDEV

MEPS-HC Health Policy Booklet (HPB) Abstraction Tool (MAT) Development

Three Core Functions



1 Ingestion
- Data De-identification
- Data Extraction Transform Load (ETL)

2 Abstraction
- Assignment
- Plan Abstraction
- Quality Assurance
- Approval

3 Publication
- Raw Data
- Filtered Data

Ingestion: Data De-identification

CHALLENGES

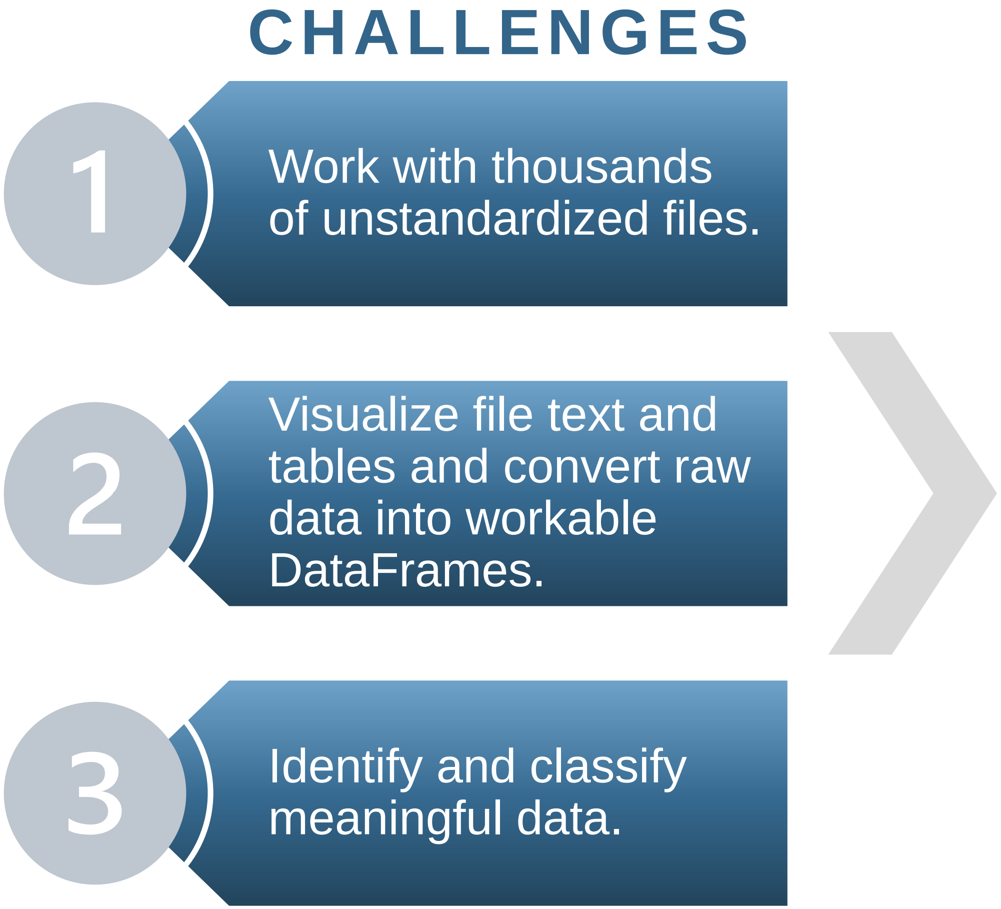
1. Protect the identity of all survey respondents.
2. Comply with FedRAMP moderate standards.
3. Reduce risk of inadvertently exposing sensitive information.

SOLUTION

Leverage the Data Loss Prevention (DLP) service to develop an automated data transformation pipeline to de-identify sensitive data like personally identifiable information (PII) to

- preserve the utility of the source data for joining analytics or
- reduce the risk of handling the data by obfuscating the raw sensitive identifiers.

Ingestion: Data Extraction Transform and Load (ETL)



CHALLENGES

1 Work with thousands of unstandardized files.

2 Visualize file text and tables and convert raw data into workable DataFrames.

3 Identify and classify meaningful data.

SOLUTION

Camelot is a Python library that makes it easy for developers to extract tables from PDF files:

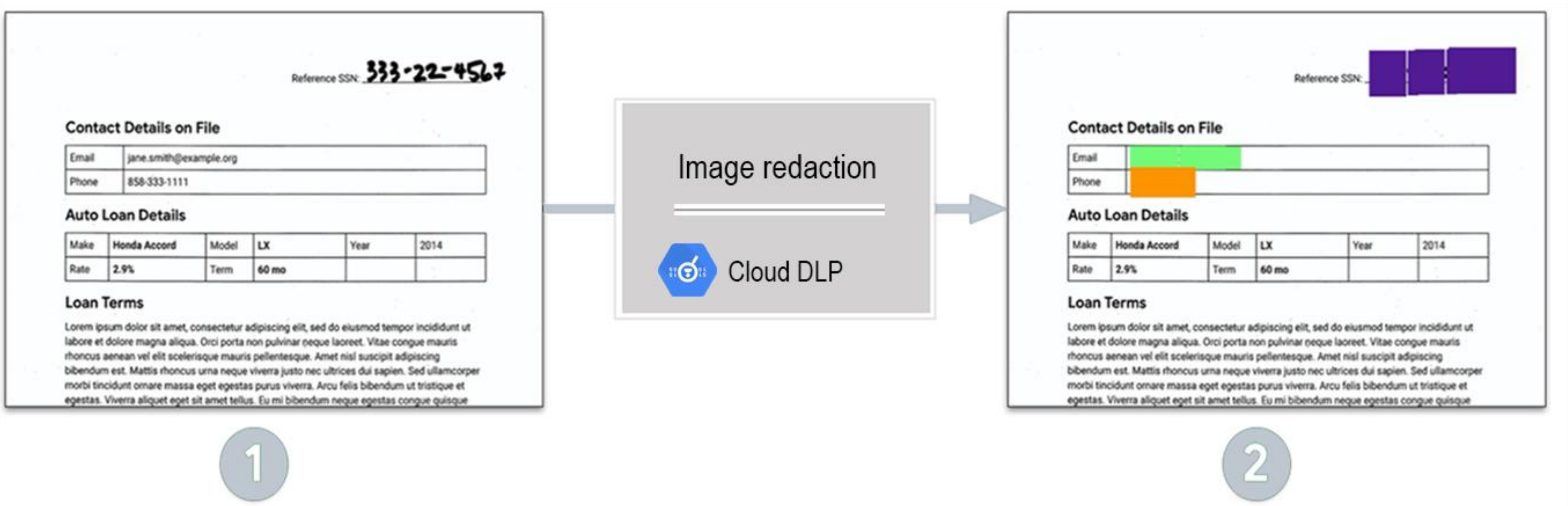
- *Bad* tables can be discarded based on **metrics** like accuracy and whitespace, without ever having to manually look at each table.
- Each table is a pandas DataFrame, which seamlessly integrates into ETL and data analysis workflows.
- Export to multiple formats, including JSON, Excel, HTML and SQLite.

Data Loss Prevention (DLP)

The process of discovering sensitive data, enforcing protective measures such as encryption and redaction methods, and preventing it from leaving the enterprise in unwanted or noncompliant ways.

DLP: What It Does

- Cloud DLP can redact sensitive text from an image.
- Using InfoType detectors and Computer Vision, Cloud DLP inspects a PDF file for text, detects sensitive data within the text, and then redacts any matching sensitive data.



Reference SSN: 333-22-4567

Contact Details on File

Email	jane.smith@example.org
Phone	858-333-1111

Auto Loan Details

Make	**Honda Accord**	Model	**LX**	Year	2014
Rate	**2.9%**	Term	**60 mo**		

Loan Terms

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Orci porta non pulvinar neque laoreet. Vitae congue mauris rhoncus aenean vel elit scelerisque mauris pellentesque. Amet nisl suscipit adipiscing bibendum est. Mattis rhoncus urna neque viverra justo nec ultrices dui sapien. Sed ullamcorper morbi tincidunt ornare massa eget egestas purus viverra. Arcu felis bibendum ut tristique et egestas. Viverra aliquet eget sit amet tellus. Eu mi bibendum neque egestas congue quisque

Reference SSN: [redacted]

Contact Details on File

Email	[redacted]
Phone	[redacted]

Auto Loan Details

Make	**Honda Accord**	Model	**LX**	Year	2014
Rate	**2.9%**	Term	**60 mo**		

Loan Terms

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Orci porta non pulvinar neque laoreet. Vitae congue mauris rhoncus aenean vel elit scelerisque mauris pellentesque. Amet nisl suscipit adipiscing bibendum est. Mattis rhoncus urna neque viverra justo nec ultrices dui sapien. Sed ullamcorper morbi tincidunt ornare massa eget egestas purus viverra. Arcu felis bibendum ut tristique et egestas. Viverra aliquet eget sit amet tellus. Eu mi bibendum neque egestas congue quisque

DLP: InfoType Detectors

InfoType Detectors are defined using:

- Dictionaries - A component of a Natural Language Processing (NLP) system that contains information (semantic, grammatical) about individual words or word strings.

- Regular Expressions - Instruction given to a function on what and how to match or replace a set of strings within a defined pattern. Example: MM/DD/YYYY

DLP: InfoType Detectors

InfoType	InfoType Description
AGE	An *age* measured in months or years.
DATE_OF_BIRTH	A *date of birth*. Note: Not recommended for use during latency sensitive operations.
EMAIL_ADDRESS	An *email address* identifies the mailbox that emails are sent to or from. The maximum length of the domain name is 255 characters, and the maximum length of the local-part is 64 characters.
FIRST_NAME	A *first name* is defined as the first part of a PERSON_NAME. Note: Not recommended for use during latency sensitive operations.
GENDER	A person's *gender identity*.
ICD9_CODE	The *International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) lexicon* is used to assign diagnostic and procedure codes associated with inpatient, outpatient, and physician office use in the United States. The US National Center for Health Statistics (NCHS) created the ICD-9-CM lexicon. It is based on the ICD-9 lexicon but provides for more morbidity detail. The ICD-9-CM lexicon is updated annually on October 1.
ICD10_CODE	Like ICD-9-CM codes, the *International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) lexicon* is a series of diagnostic codes. The World Health Organization (WHO) publishes the ICD-10-CM lexicon to describe causes of morbidity and mortality.

DLP: Computer Vision

- A service that allows developers to analyze the content of an image through extracted data. For this purpose, utilizes machine learning models trained on a large dataset of images.
- The higher quality data you deliver and the better the design of the model you use, the smarter outcome will be produced. With machine learning Application Programming Interfaces (APIs), Artificial Intelligence (AI) applications can be easily incorporated.

Machine Learning

- The ability of statistical models to develop capabilities and improve their performance over time without the need to follow explicitly programmed instructions.

- Most cognitive technologies are based on machine learning and its more complex progeny, deep learning. That includes computer vision and NLP.

Architect Objectives: Fast, Flexible and Customizable

- Python - The most preferred programming language in data science and machine learning:
  - Quick implementation of Python code for solving complicated mathematic, and other advanced problems.
  - Vast number of packages to solve machine learning problems. There are packages for everything, which include images, audio, text, deep learning, records abstraction, scientific computing
- Hybrid Database – Leverages Relational and Flat Files database concepts.
  - Extends the benefits of both concepts
  - Reduces the limitation of either approach.
- Serverless Services - Serverless architectures offer greater scalability, more flexibility, and quicker time to release.

Ingestion: Machine Learning

- MEPS Abstraction Tool leverages learning capabilities to enhance image recognition and computer vision and natural language processing that all come into play to seek out overlooked or unexpected sensitive data and automatically redact it.

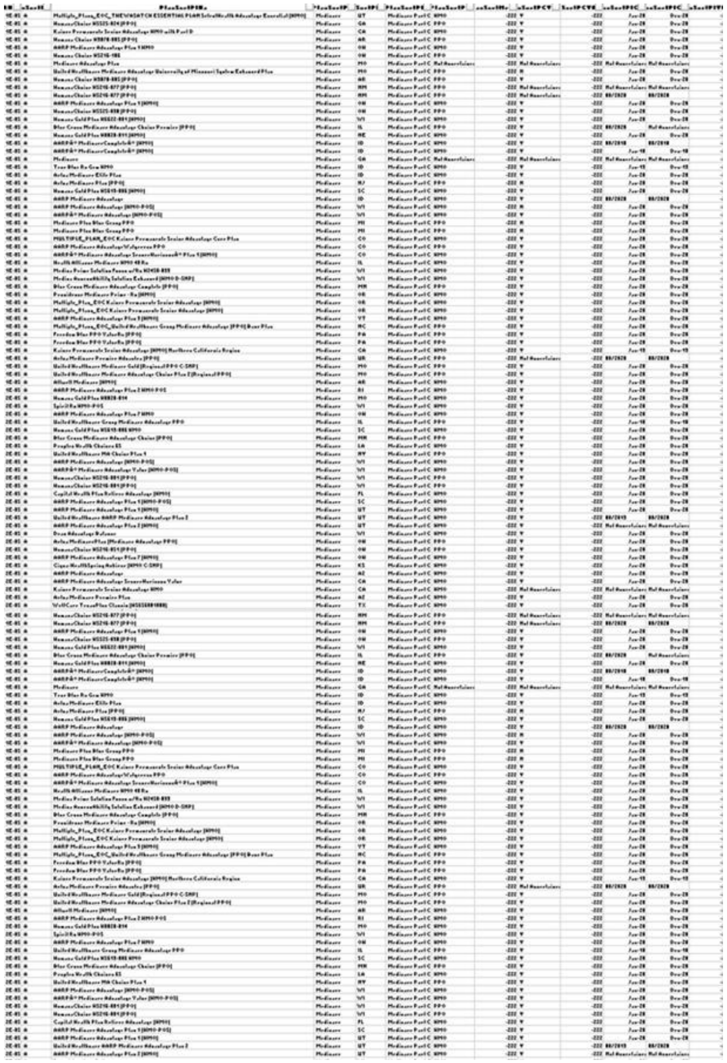
- A computing device receives a document that was incorrectly classified as sensitive data based on a machine learning-based detection (MLD) profile.

- The computing device modifies a training data set that was used to generate the MLD profile by adding the document to the training data set as a negative example of sensitive data to generate a modified training data set.

- The computing device then analyzes the modified training data set using machine learning to generate an updated MLD profile.

Abstraction Outputs and Deliverables

- Abstracted and QA completed:
  - Over 5,000 plan documents
  - 3,382 unique insurance plans
  - ~500,000 data points with QA on 100% of output
- Produced:
  - Unified dataset
  - Datasets by plan type
  - Supporting documentation (e.g., codebook)
  - Summary statistics



Thank you!

AHRQ Agency for Healthcare Research and Quality
Advancing Excellence in Health Care

Monica.Wolford@AHRQ.HHS.gov
301-427-1651

SoftDev

Sandy.Pope@softdevconsulting.com
919-246-4386