The Department of Health Master Person Index (DOHMPI)
Introduction and Use Cases

Kyle Lunt / Huaizhong Pan

Office of Health Informatics
Center for Health Data and Informatics
Utah Department of Health
10/27/2021

Outline

- Overview
- DOHMPI Dashboard
- APIs
- Use cases
- QA process
- Challenges
- Future work
- How to Get started

Brief intro to DOHMPI

- Identity Resolution is the process of building a cohesive view of a person across multiple systems. It involves analysis of demographics and identifiers to link records across systems.
- System A:

Name	DOB	Sex
Mary Smith	4/17/1970	F
Robert Jones	9/29/1960	M
Alex Parker	5/14/1990	U

- System B:

Name	DOB	Sex
Mary Smith	4/17/1970	F
Bob Jones	9/29/1960	M
Alexander Parker	5/14/1990	M

- Deterministic vs Probabilistic Matching

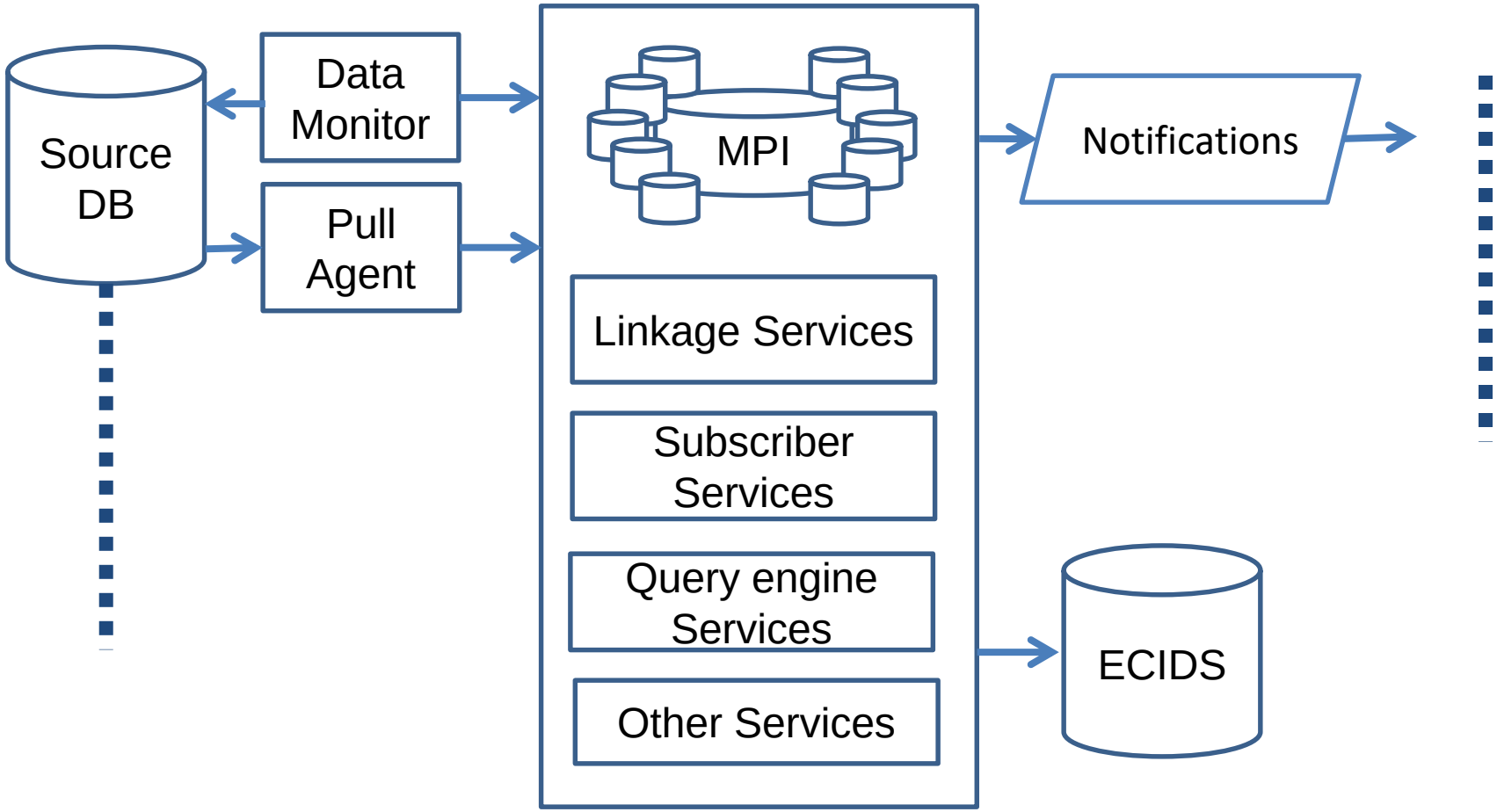
Brief intro to DOHMPI

- The Department of Health Master Person Index (DOHMPI) provides ongoing linkage of multiple public health information systems for both operational and research purposes.
- Proprietary hybrid (probabilistic and deterministic) algorithms, developed by Multidimensional Software Creations (MDSC) of Logan, Utah.

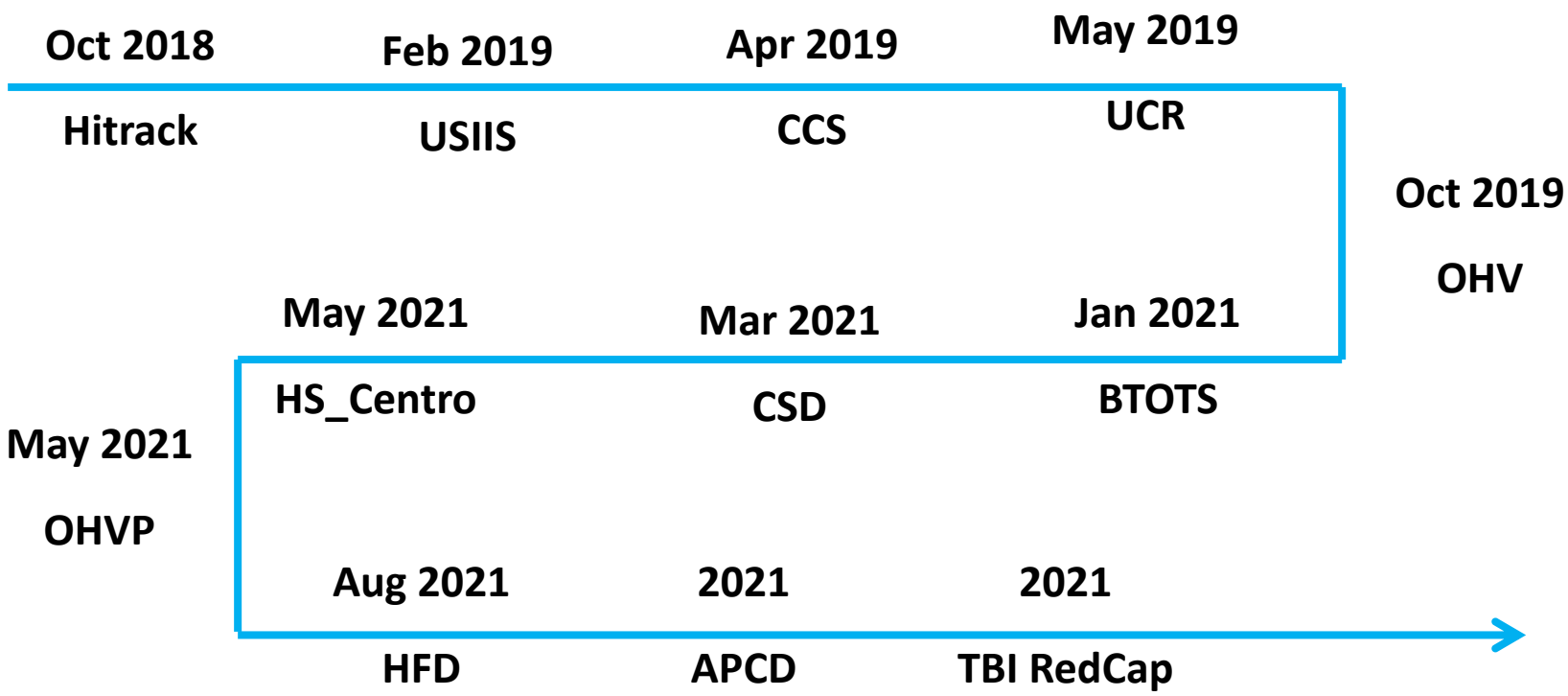
Brief intro to DOHMPI

- More than **18.9 million** records from 16 health information systems.
- More than **12.2 million** linkages.
- More than **9.9 million** unique identities.

DOHMPI Structure



DOHMPI Data Sources



DOHMPI Data Sources

	DOHMPI Code	Data Source Program(s)
1	ASQ	Ages and Stages Questionnaire (ASQ) Screening
2	BTOTS	Early Intervention Part C (Baby Watch)
3	CCS	DWS- Child Care Subsidy
4	CSD	DOPL - Controlled Substance Database
5	EDEN	Utah Death Registry
6	HiTrack	Newborn Hearing Screening
7	MEDICAID	Medicaid
8	OHV Legacy	Office of Home Visiting (legacy system)
9	Professional Licensing	DOPL - Professional Licensing
10	Ucr	Utah Cancer Registry
11	Uintah	Utah Birth Registry
12	Usiis	Utah Immunization Registry
13	WIC	Women Infant and Children
14	HS_Centro	Head Start - Centro de la Familia
15	OHVP	Office of Home Visiting Penelope (current system)
16	HFD	Healthcare Facilities Database

Coming Soon

	DOHMPI Code	Data Source Program(s)
1	APCD	All Payer Claims Database (all Utah payers)
2	HS_DDI	Head Start - DDI Vantage
3	TBI	Traumatic Brain Injury Registry
4	NBS	Newborn Screening

Possible Use Cases:

- Link persons across different visits
- Link persons between HFD and RedCap data for TBI registry
- Link CSD and Birth and Death registries
- Automatically link Birth and Death records
- CSD and WIC
- CSD and APCD
- EMS and ED

DOHMPI Dashboard

Features:

- Data loading status
- Updated the status of data sources
- Updated consumer notifications status
- Updated DOHMPI road map plans.
- Monitor data integration for each data source by select time period.
- Random sample of DOHMPI personIds
- DOHMPI API Logs

UTAH DEPARTMENT OF HEALTH
DOHMPI Dashboard

MPI Dashboard Advanced Matcher Hello @utah.gov!

MPI Dashboard | DOHMPI Status | Roadmap

MPI Dashboard

Unique Identities
9,912,572

Loading Status

Datasource	Last Loaded
DOHMPI	10/23/2021
APCD - Medicaid (All Payers Claims Database)	06/11/2021
ASQ (Ages and Stages Questionnaire)	10/22/2021
BTOTS (Early Intervention Part C)	10/22/2021
CCS (Child Care Subsidy)	08/09/2021
CSD (Controlled Substance Database)	03/18/2021
DOPL (Professional Licensing)	10/23/2021

Recent Volume

Component	Past 24-Hour Volume	Past 7-Day Volume
DOHMPI - Add New Person	17,872	100,042
DOHMPI - Link Person	142,845	1,076,058
DOHMPI - Matches	161,766	1,180,852
DOHMPI - Relink Person	504	2,882
DOHMPI - Unlink Person	225	1,553
Data Monitor - ASQ	0	605
Data Monitor - BTOTS	104	770

UTAH DEPARTMENT OF HEALTH
DOHMPI Dashboard

MPI Dashboard Advanced Matcher

Advanced Statistics

Select options from below to show volume statistics of MPI components.

Filter Results

× EDEN (Death Registry)

Time Frame *

Daily

Date Range *

10/01/2021 to 10/10/2021

Display

Click 'Display' to see statistics.

Results

Data Monitor - EDEN

From	To	Volume
10/1/2021	10/1/2021	231
10/2/2021	10/2/2021	1
10/3/2021	10/3/2021	0
10/4/2021	10/4/2021	142
10/5/2021	10/5/2021	162
10/6/2021	10/6/2021	183
10/7/2021	10/7/2021	158
10/8/2021	10/8/2021	143
10/9/2021	10/9/2021	5
10/10/2021	10/10/2021	0

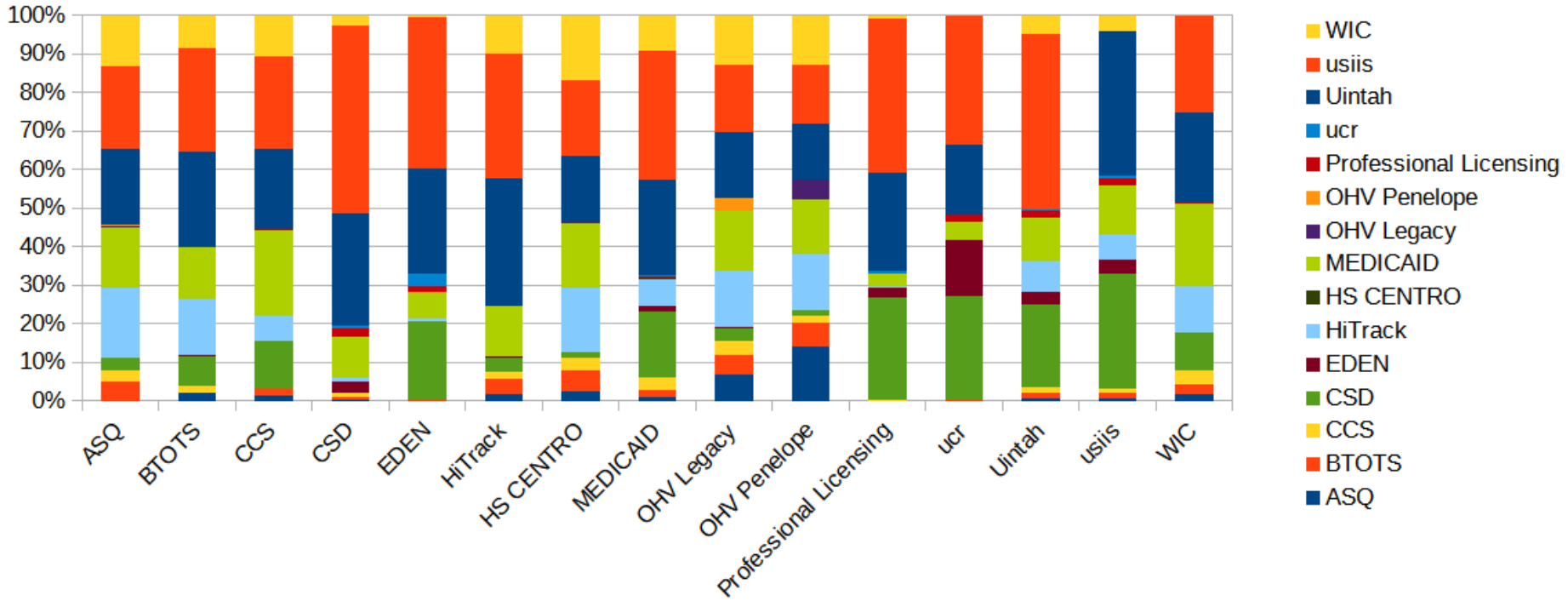
Current Identity Counts

Datasource	Identity Count	SPID Count
asq	35,865	40,517
btots	116,069	122,700
ccs	101,409	104,504
csd	2,515,115	3,318,588
eden_vs	977,104	977,470
hfd	855,213	953,717
hitrack	1,227,519	1,236,014
hs_centro	1,142	1,143
medicaid	968,068	972,082
ohv_legacy	1,607	1,622
ohv_penelope	862	865
professional_licensing	229,175	230,616
ucr	54,013	54,027
uintah	5,075,268	6,139,041
usiis	5,690,584	5,982,747
wic	334,366	335,977

Cross-Over Identity Counts

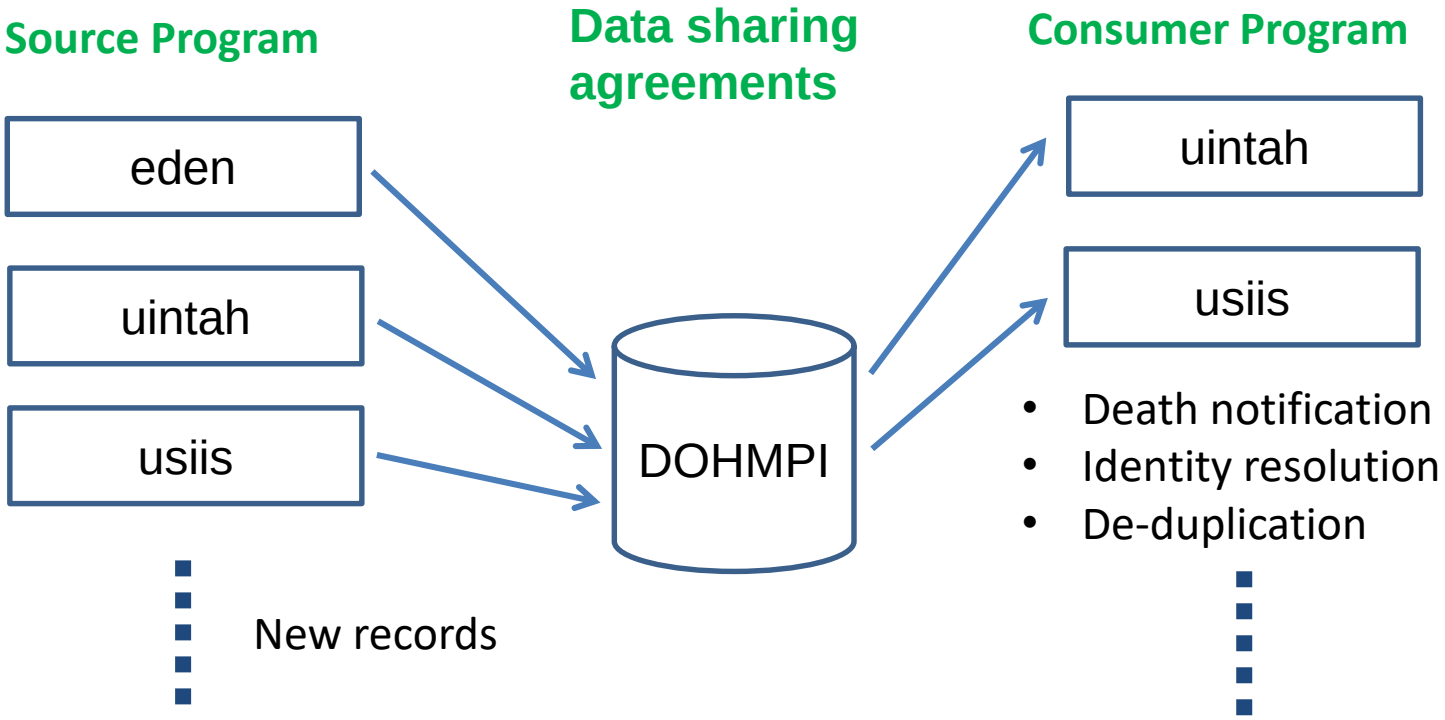
	ASQ	BTOTS	CCS	CSD	EDEN	HFD	HiTrack	HS CENTRO	MEDICAID	OHV Legacy	OHV Penelope	Professional Licensing	ucr	Uintah	usiis	WIC
ASQ	-	7,517	4,660	4,934	285	1,320	29,309	145	24,289	513	722	2	14	30,233	32,973	20,749
BTOTS	7,517	-	7,954	29,819	1,314	6,674	90,016	295	52,049	374	310	55	101	97,465	106,763	33,353
CCS	4,660	7,954	-	47,786	1,015	19,964	44,113	189	86,495	271	103	1,165	192	80,818	93,936	42,725
CSD	4,934	29,819	47,786	-	137,898	420,345	213,403	92	462,259	250	65	93,360	38,530	1,271,471	2,145,967	116,547
EDEN	285	1,314	1,015	137,898	-	103,610	8,965	1	45,769	13	7	10,434	21,071	183,164	263,946	2,536
HFD	1,320	6,674	19,964	420,345	103,610	-	48,118	46	167,205	132	17	25,115	12,973	409,545	592,601	38,655
HiTrack	29,309	90,016	44,113	213,403	8,965	48,118	-	923	331,043	1,220	729	1,392	445	968,316	960,627	176,424
HS CENTRO	145	295	189	92	1	46	923	-	929	7	10	-	1	945	1,083	948
MEDICAID	24,289	52,049	86,495	462,259	45,769	167,205	331,043	929	-	1,163	713	11,541	6,747	671,166	908,421	256,902
OHV Legacy	513	374	271	250	13	132	1,220	7	1,163	-	259	-	1	1,269	1,303	980
OHV Penelope	722	310	103	65	7	17	729	10	713	259	-	-	-	744	772	668
Professional Licensing	2	55	1,165	93,360	10,434	25,115	1,392	-	11,541	-	-	-	2,538	90,636	142,649	3,128
ucr	14	101	192	38,530	21,071	12,973	445	1	6,747	1	-	2,538	-	25,649	47,821	435
Uintah	30,233	97,465	80,818	1,271,471	183,164	409,545	968,316	945	671,166	1,269	744	90,636	25,649	-	2,721,317	286,349
usiis	32,973	106,763	93,936	2,145,967	263,946	592,601	960,627	1,083	908,421	1,303	772	142,649	47,821	2,721,317	-	310,478
WIC	20,749	33,353	42,725	116,547	2,536	38,655	176,424	948	256,902	980	668	3,128	435	286,349	310,478	-

Cross-Over Identity Counts



DOHMPI Consumer Use Cases

- Death notification
- Identity Resolution – de-duplication and linkage



DOHMPI Consumer Use Cases

- **Notifications:** Automated near real-time Data Source specific messages that assist Data Source(s) and Consumer(s) for operational and research use cases:
- **Death Notification:** Data Source specific death message.
- **Pseudo-ID:** Consumer system specific pseudo identifiers for linked identity Data Warehouses that contain “limited” demographic and program data.
- **SPID-SPID linkage:** Notification to a Data Source that multiple distinct identifiers exist for the same identity.

DOHMPI APIs

DOHMPI API: This allows Consumer system(s) to interact with the DOHMPI and download Consumer specific notifications (Death, SPID->SPID linkage) in near real-time and at whatever frequency they desire:

- **Match:** Check if a match exists for some personal identification data
- **Match-Exists:** Match personal identification data to existing persons.
- **List Spid-Spid Link Notifications:** Get SPID-SPID linkage notifications from the specified notification schema.
- **List Death Notifications:** Get death notifications from the specified notification schema.

DOHMPI API Functionality	Available DOHMPI API
Match	Yes
Match-Exists	Yes
List Spid-Spid Link Notifications	Yes
List Authorized Datasources	Yes
List Death Notifications	Yes

DOHMPI API

http://dohmpiapi.utah.gov/swagger/docs/v1

api_key

Explore

DOHMPI API v1

Public interface to programmatically interact with DOHMPI API.

Note: Due to a bug in the library that is used to generate this documentation page, an extra "application/json" object layer is included in the response examples. This object is not present in the actual API responses.

Matcher

Show/Hide | List Operations | Expand Operations

POST	/api/match-exists	Check if a match exists for some personal identification data.
POST	/api/match	Match personal identification data to existing persons.

Metadata

Show/Hide | List Operations | Expand Operations

GET	/api/datasources	Returns all datasources the authorized user has access to as it applies to the match endpoints (/api/match and /api/match-exists).

Notifications

Show/Hide | List Operations | Expand Operations

GET	/api/notifications/{schema}/death	Get death notifications from the specified notification schema.
GET	/api/notifications/{schema}/spid-spid-link	Get SPID-SPID (unique identifier from the specific schema) linkage notifications from the specified notifications schema.

[BASE URL: , API VERSION: v1]

DOHMPI Match API

POST /api/match

Match personal identification data to existing persons.

Implementation Notes

Required minimum parameters include – firstName, lastName, and birthDate

Response Class (Status 200)

OK

Model | Example Value


```
{
  "application/json": {
    "items": [
      {
        "confidence": 0.723255005536254,
        "spids": [
          {
            "datasourceId": 17,
            "spid": "123456"
          },
```


Response Content Type application/json

DOHMPI spid-spid API

GET /api/notifications/{schema}/spid-spid-link

Get SPID-SPID (unique identifier from the specific schema) linkage notifications from the specified notifications schema.

Response Class (Status 200)

OK

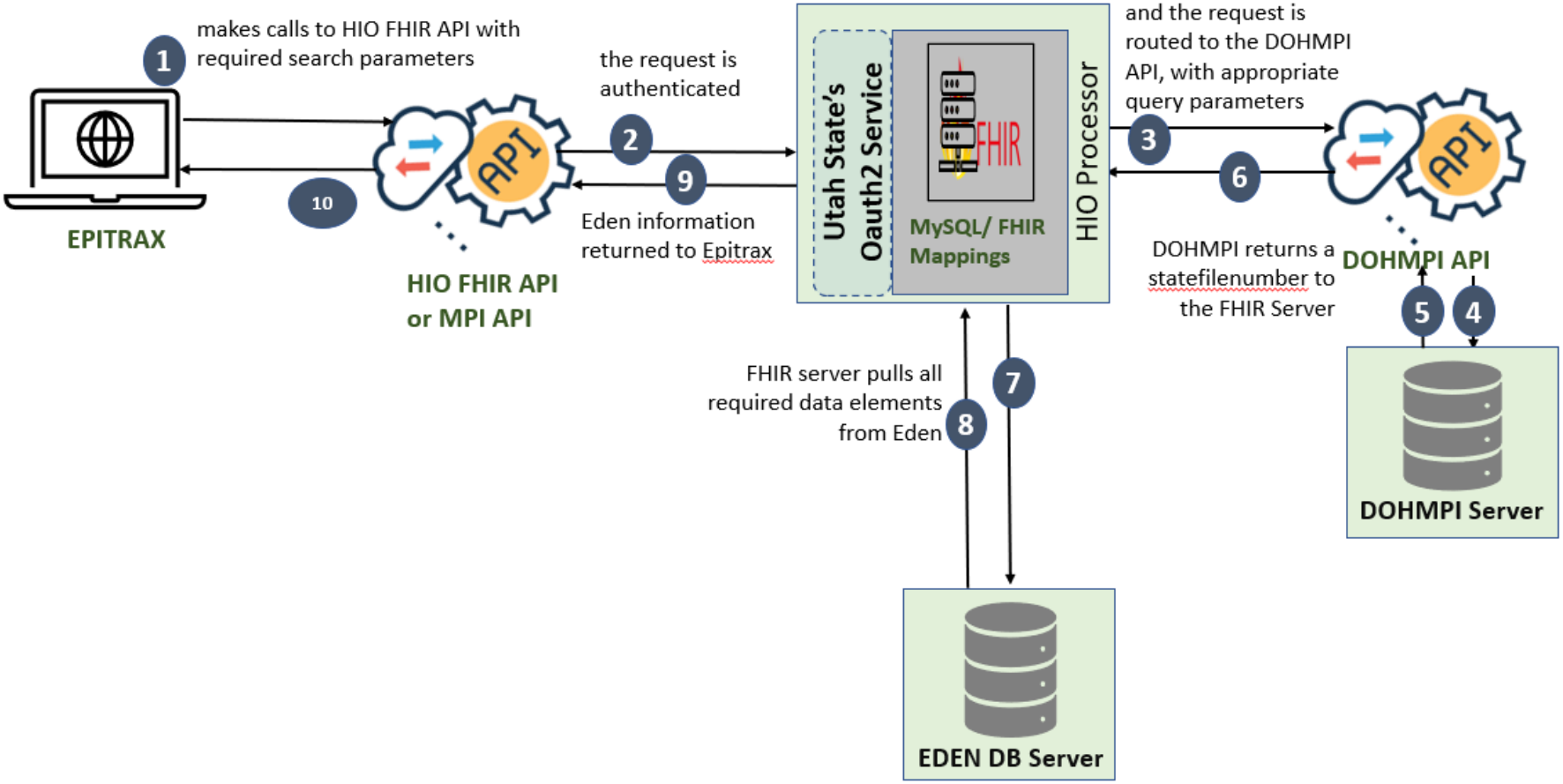
Model | Example Value


```
    "nextPageLink": "https://www.example.com/api/request?page=3&limit=50"
  },
  "items": [
    {
      "datasource": 0,
      "spid": "1352192311",
      "linkDatasource": 28,
      "linkSpid": "1385725825",
      "createdOn": "2020-01-19T00:00:00",
      "isLinked": true
    },
```


Response Content Type application/json

Epitrax MPI API

Version 2 – Probabilistic matching to a single source using MPI API & DOHMPI API



Epitrax MPI API Performance

DOHMPI API Usage

Filters

TABLES

Endpoints

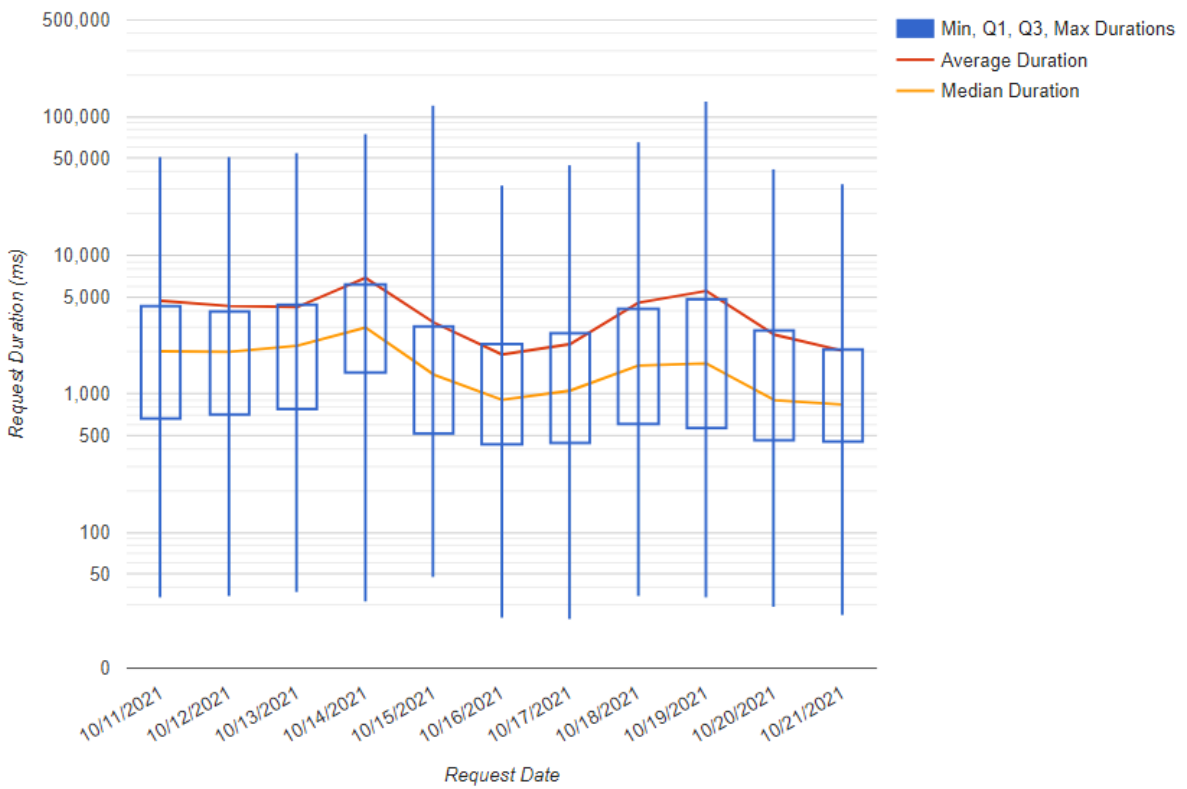
API Keys

CHARTS

Request Duration Over Time

Request Success Over Time

Request Duration Over Time



CONTINUOUS QUALITY IMPROVEMENT - DOHMPI QUALITY ASSURANCE

Goal for DOHMPI QA

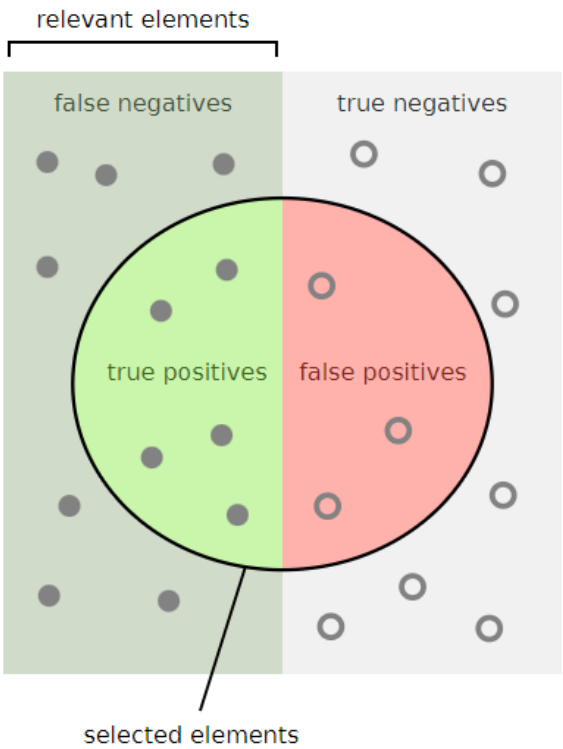
To develop a methodology and automate process system to continuously monitor linkage quality of DOHPMI and provide feedback to improve the linkage quality.

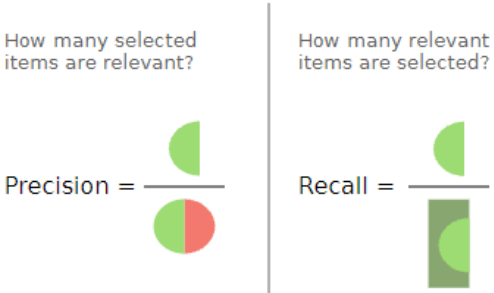
All time – identify overall match precision and recall
Previous month – identify overall match precision and recall (will be able to better determine if significant match decisions changes occur).

Precision/ Positive Predictive Value (PPV)
= predict_correct_matches/All_predict_matches
= TP/(TP + FP)

Recall / Sensitivity
= predict_correct_matches/All_true_matches
= TP/(TP + FN)

F1 value
= 2 x precision x recall / (precision + recall)





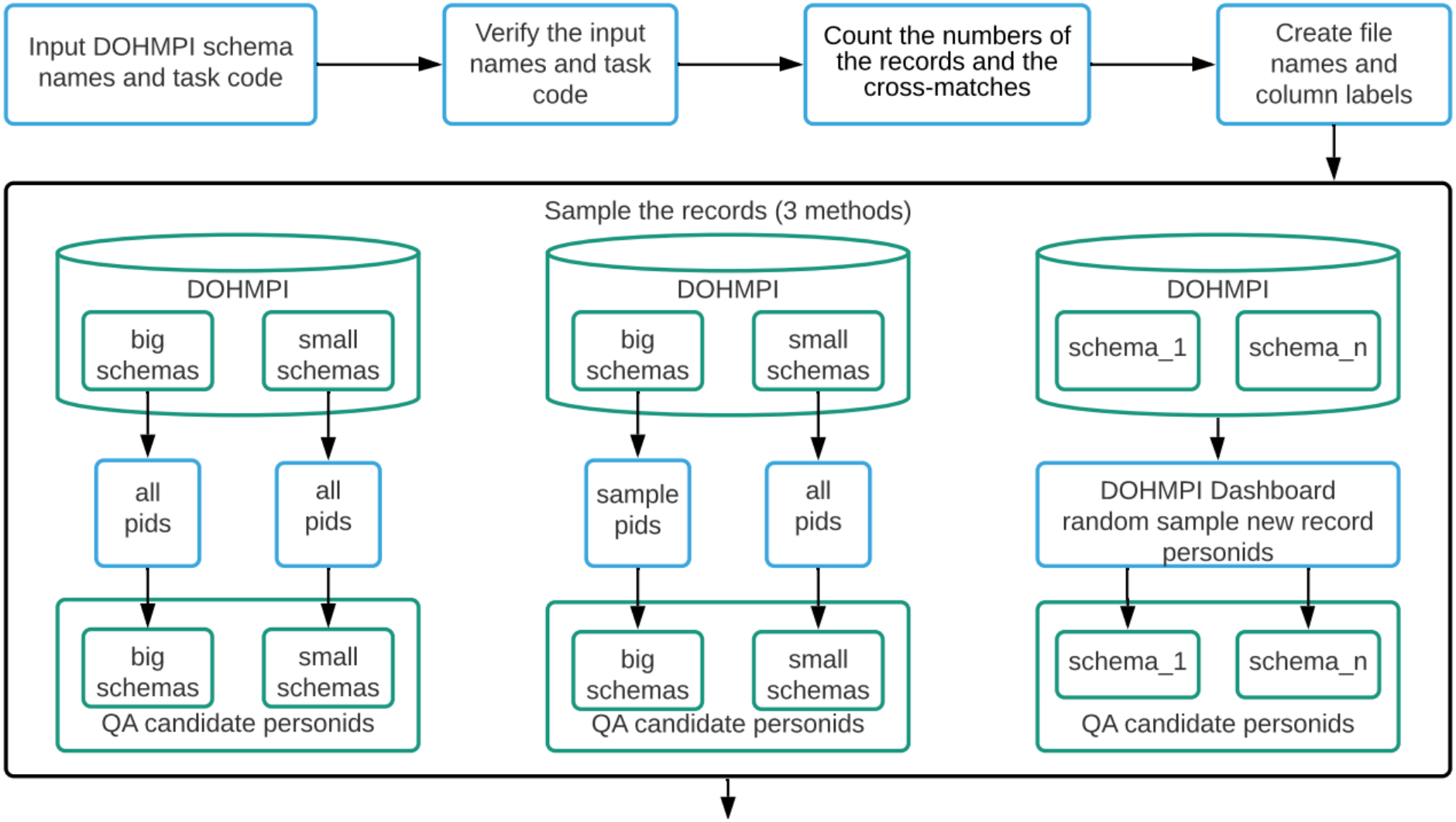
https://en.wikipedia.org/wiki/Precision_and_recall

Record Linkage Methods

- Deterministic—look for exact or nearly exact matches on combinations of variables
- Probabilistic—calculate a score based on probabilities
- Other methods for different scenarios
  - Rule-based approach
  - Bayesian approach
  - Unsupervised and supervised machine-learning approaches

Here we used

- **Deterministic/Probabilistic + Rule-based approach**
- **Probabilistic**


Input DOHMPI schema names and task code
Verify the input names and task code
Count the numbers of the records and the cross-matches
Create file names and column labels
Sample the records (3 methods)
DOHMPI
big schemas
small schemas
all pids
all pids
big schemas
small schemas
QA candidate personids
DOHMPI
big schemas
small schemas
sample pids
all pids
big schemas
small schemas
QA candidate personids
DOHMPI
schema_1
schema_n
DOHMPI Dashboard random sample new record personids
schema_1
schema_n
QA candidate personids

UTAH DEPARTMENT OF HEALTH
DOHMPI Dashboard

MPI Dashboard Advanced Matcher

PersonID Quality Assurance (QA)

Form	
All Data (Time-Period)	Download CSV
Random Sample (All-Time)	Download CSV
Random Sample (Time-Period)	Download CSV

© 2021 - Utah Department of Health

UTAH DEPARTMENT OF HEALTH
DOHMPI Dashboard

MPI Das

PersonID Quality

Form	
All Data (Time-Period)	
Random Sample (All-Time)	
Random Sample (Time-Period)	

© 2021 - Utah Department of Health

Random Sample PersonID (Time-Period)

Sample Count *

50000

From *

09/01/2021

To *

09/30/2021

Download

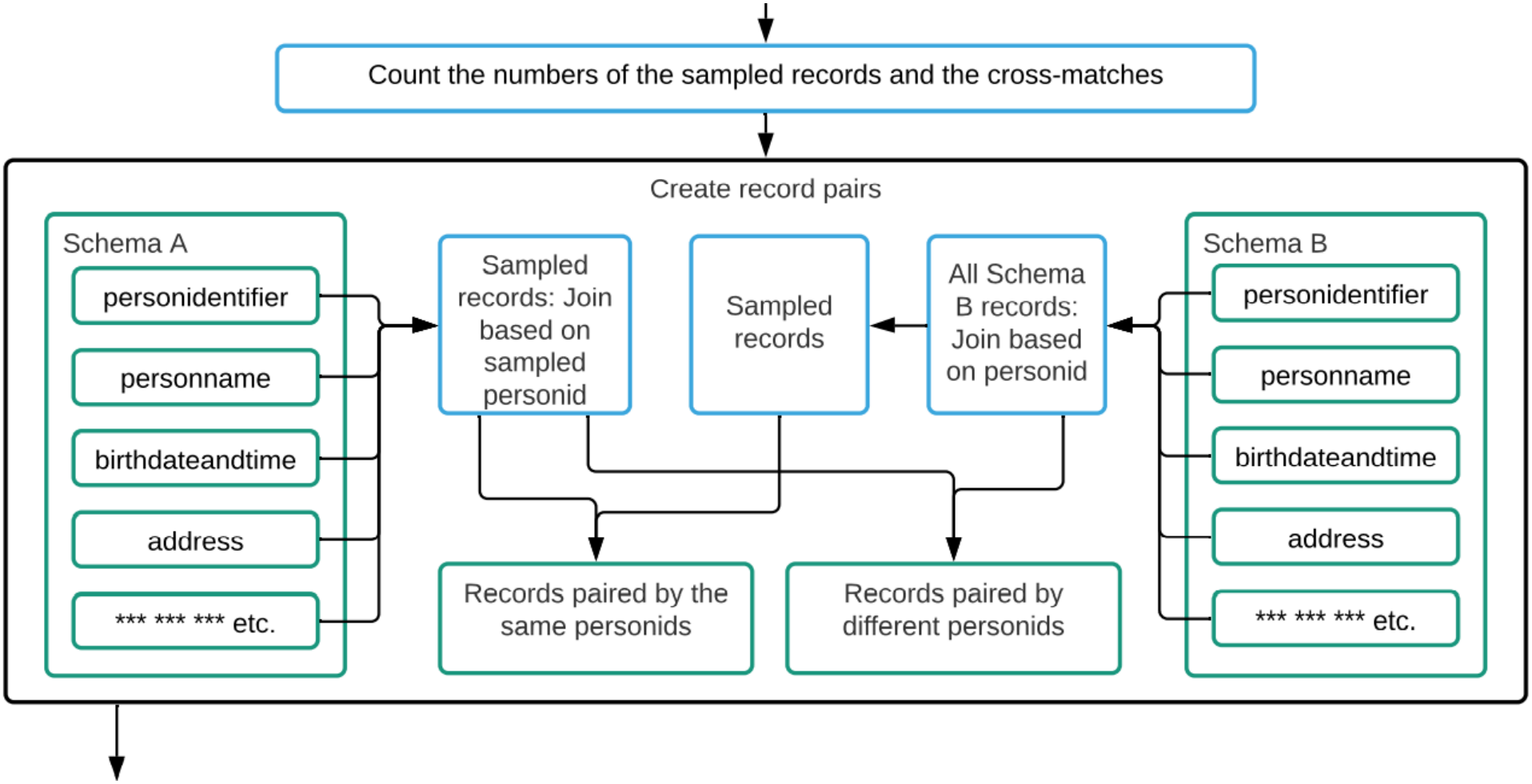
Total Count: 942,160

Confidence Level and Margin of Error

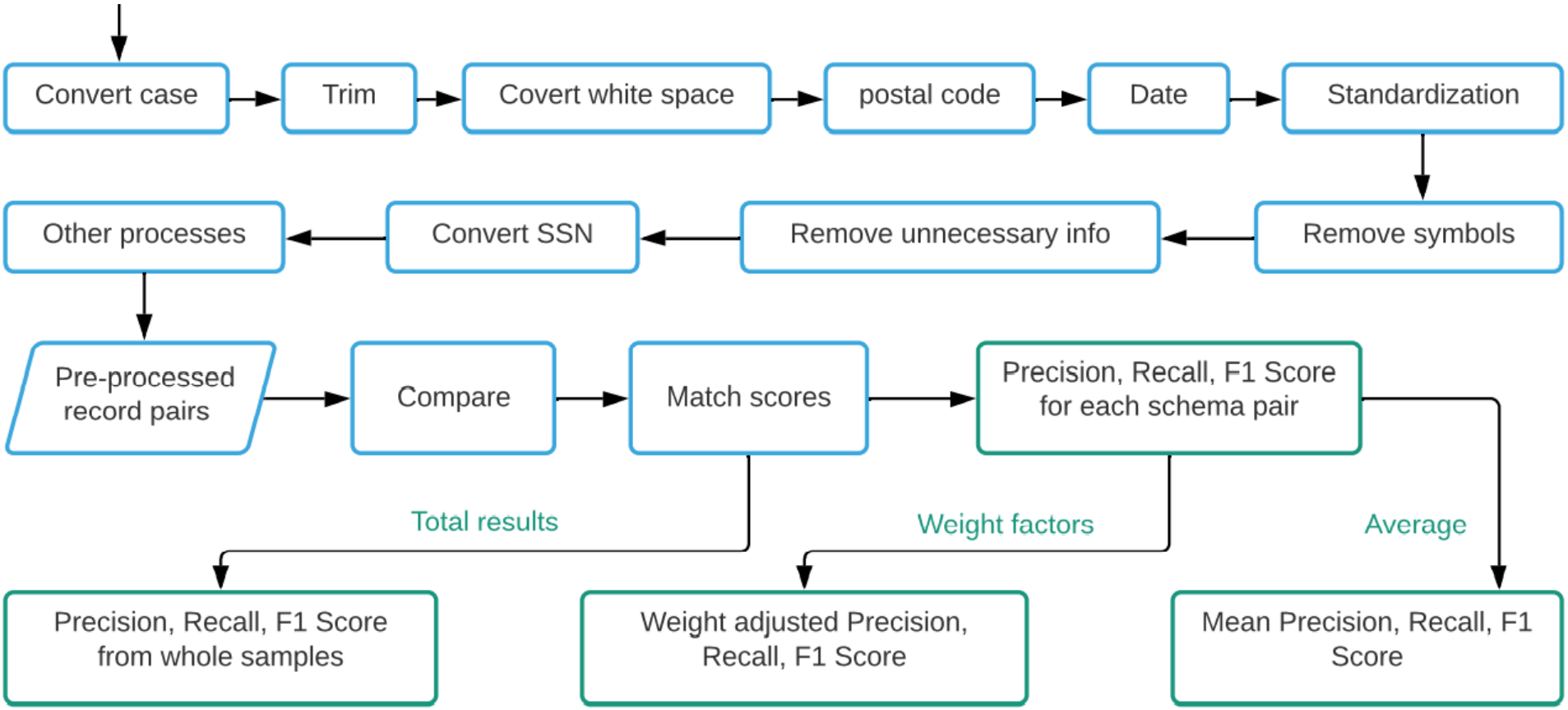
N (Sample Size)	**Margin of Error at 90% confidence**	**Margin of Error at 95% confidence**	**Margin of Error at 99% confidence**
10	26.09%	31%	40.79%
100	8.25%	9.80%	12.90%
1000	2.61%	3.10%	4.08%
5000	1.17%	1.40%	1.82%
10000	0.83%	0.98%	1.29%
20000	0.58%	0.69%	0.91%

Data Source	ASQ	BTOTS	CCS	CSD	EDEN	HFD	HiTrack	HS CENTRO	MEDICAID	OHV Legacy	OHV Penelope	Professional Licensing	ucr	Uintah	usiis	WIC
ASQ	-	7505	4657	4934	284	336	29248	144	24256	513	722	2	14	30172	32905	20714
BTOTS	7505	-	7950	29817	1312	3478	89879	293	51993	374	310	54	101	97313	106606	33285
CCS	4657	7950	-	47781	1011	12362	44103	189	86484	271	103	1163	192	80803	93950	42711
CSD	4934	29817	47781	-	137530	256312	213399	92	462219	250	65	93280	38523	1271188	2145906	116476
EDEN	284	1312	1011	137530	-	48005	8941	1	45620	13	7	10413	21001	182368	263149	2527
HFD	336	3478	12362	256312	48005	-	25418	3	103150	80	6	14506	5955	247269	365183	22495
HiTrack	29248	89879	44103	213399	8941	25418	-	923	330872	1220	729	1380	445	967173	959604	176145
HS CENTRO	144	293	189	92	1	3	923	-	929	7	10	-	1	945	1083	948
MEDICAID	24256	51993	86484	462219	45620	103150	330872	929	-	1163	713	11520	6745	670913	908340	256699
OHV Legacy	513	374	271	250	13	80	1220	7	1163	-	259	-	1	1269	1303	980
OHV Penelope	722	310	103	65	7	6	729	10	713	259	-	-	-	744	772	668
Professional Licensing	2	54	1163	93280	10413	14506	1380	-	11520	-	-	-	2538	90518	142530	3121
ucr	14	101	192	38523	21001	5955	445	1	6745	1	-	2538	-	25590	47788	434
Uintah	30172	97313	80803	1271188	182368	247269	967173	945	670913	1269	744	90518	25590	-	2719609	285899
usiis	32905	106606	93950	2145906	263149	365183	959604	1083	908340	1303	772	142530	47788	2719609	-	309966

QA Process


Count the numbers of the sampled records and the cross-matches
Create record pairs
Schema A
personidentifier
personname
birthdateandtime
address
*** *** *** etc.
Sampled records: Join based on sampled personid
Sampled records
All Schema B records: Join based on personid
Schema B
personidentifier
personname
birthdateandtime
address
*** *** *** etc.
Records paired by the same personids
Records paired by different personids

QA Process


Convert case
Trim
Covert white space
postal code
Date
Standardization
Remove symbols
Remove unnecessary info
Convert SSN
Other processes
Pre-processed record pairs
Compare
Match scores
Precision, Recall, F1 Score for each schema pair
Total results
Weight factors
Average
Precision, Recall, F1 Score from whole samples
Weight adjusted Precision, Recall, F1 Score
Mean Precision, Recall, F1 Score

Clean and Standardize the Data

- Remove control and special characters.
- Remove leading, tailing spaces.
- Remove symbols (SSN 123-45-6789 -> 123456789).
- Keep 5-digit postal code.
- Convert date to same format (YYYY-MM-DD).
- Convert vocabularies to same format (west -> W, slc -> SALT LAKE CITY, avenue -> AVE, road -> RD, etc.)
- Convert P. O. BOX (PO BOX 21, PO BOX 11).
- Remove Apt. xx, rm xx, suite xx, etc.
- Keep NULL/NA records.

Comparison

- stringdist, Jaro–Winkler distance, containing, etc.
- Using nick name table (contains thousands of nick names)
- Check the popularity of the names (high frequency names).
- Different ways to assign the weight factor to all pairs, and sum up all adjusted scores to calculate match.
- According to match score, count the numbers of match, unknown, mismatch, calculate precision, recall and F1 value.

Manual Review

- Focus on pairs containing the select schema at a time.
- R gives the sum of matched fields scores and sorted in DESC order.
  - Check records with high sum scores but low match scores.
  - Check records with low sum scores but high match scores.
  - Check records same sum scores but different match scores.
  - Check all records.

- Provide suggestions to correct the results and modify R rules.

- Report (DOHMPI_QA_report_2021-09-12.html)

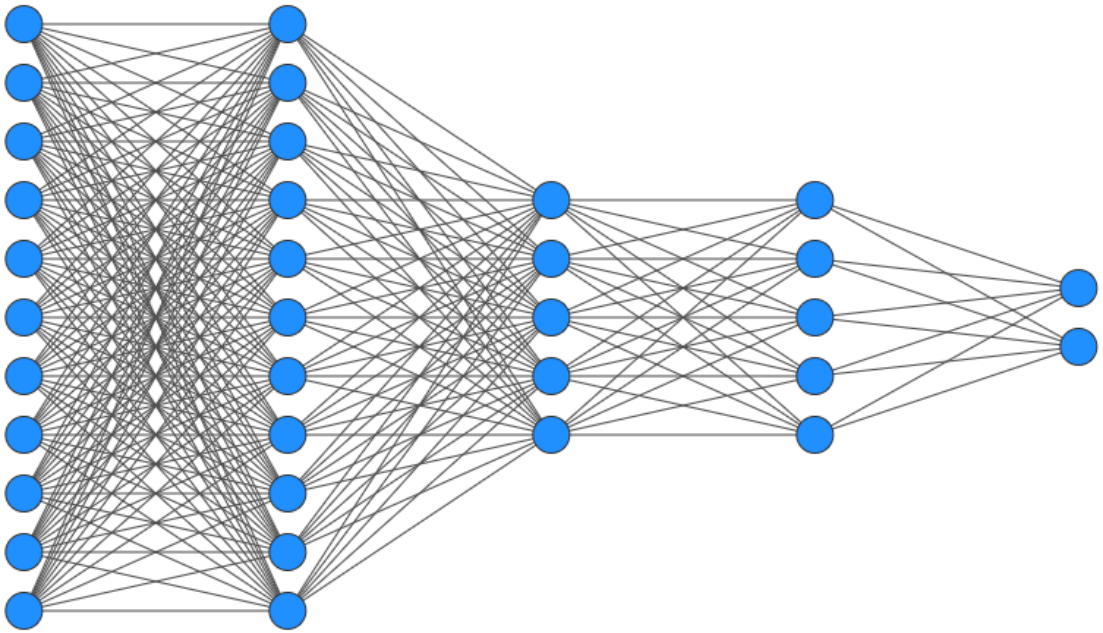
Challenges

- Most records do not have an SSN
- Using other person's SSN/id
- Missing values
- Name change (marriage, adoption)
- Address change
- Different address formats
- Simplified name vs full name, multiple words in first/last names
- Baby A, Boy A, Girl A, Baby Boy, Baby Girl
- Typo or not typo for short names (Ava, Eva)
- When should check sound similarity and when should check string similarity, which algorithm (edition, Jaro-Winkler, containing)
  - Analy Aneliy; Izzabella Isabella; Orin Oren
- Twin:
  - Slightly different spelling in first name may be different people (JADEN KADEN)

Future QA Work

- Continue improve matching rules using manual review results.
- Maintain DOHMPI precision above 99% while improving recall rates to 90%.
- Run DOHMPI QA automatically on the server or AWS.
- Develop a sound similarity comparing tool to compare every syllable.
- Using natural language processing (NLP) methods to compare address.
- Add referential matching.
- Add machine learning/deep learning methods.

Input:
11
features/
variables



Output:
Match/
Not
match

How to Get Started

- Contact HIO staff, Kyle Lunt and Huaizhong Pan
- Discuss your requirements and system
- HIO will contact our vendor for the cost
- HIO will help prepare a Data Sharing Agreement (DSA)
- After the DSA is signed, we work with MDSC to validate the data and begin loading into DOHMPI.

Thank you!

Questions? Put them in the chat or raise your digital hand!