The Department of Health Master Person Index Quality Analysis and Improvement

Pan Huaizhong Pan

Office of Health Informatics
Center for Health Data and Informatics
Utah Department of Health

6/26/2019

Outline

Brief intro to DOHMPI

Why DOHMPI QA

QA Goal

How do DOHMPI QA (procedures)

Results

Apply for other record linkage QA

Challenges

Future work

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.

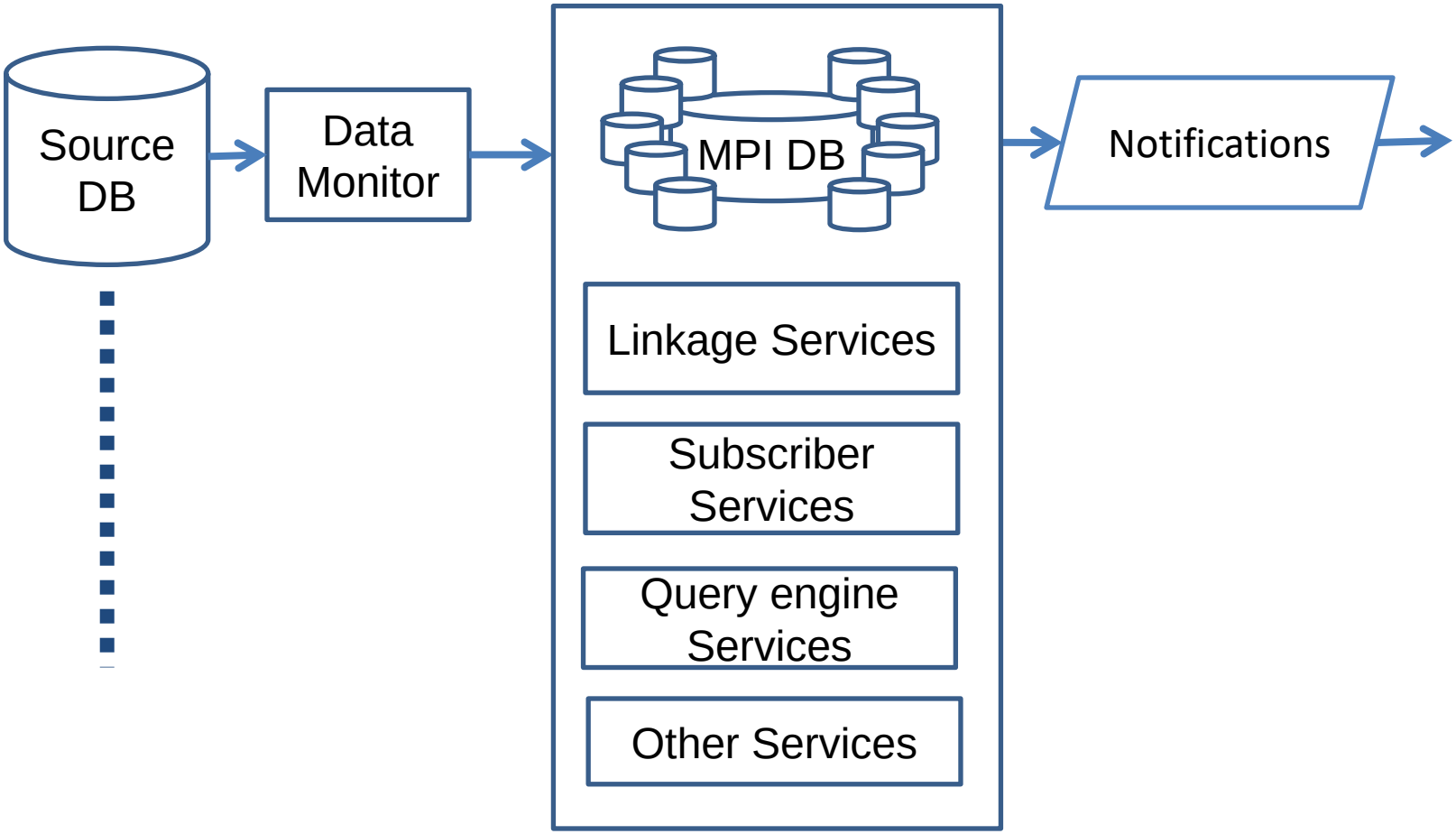
Team

- Proprietary hybrid (probabilistic and deterministic) algorithms (developed by Multidimensional Software Creations (MDSC) of Logan, Utah).
- Kailah Davis: High level guideline, data sharing agreements, loading new datasets, coordination with MDSC.

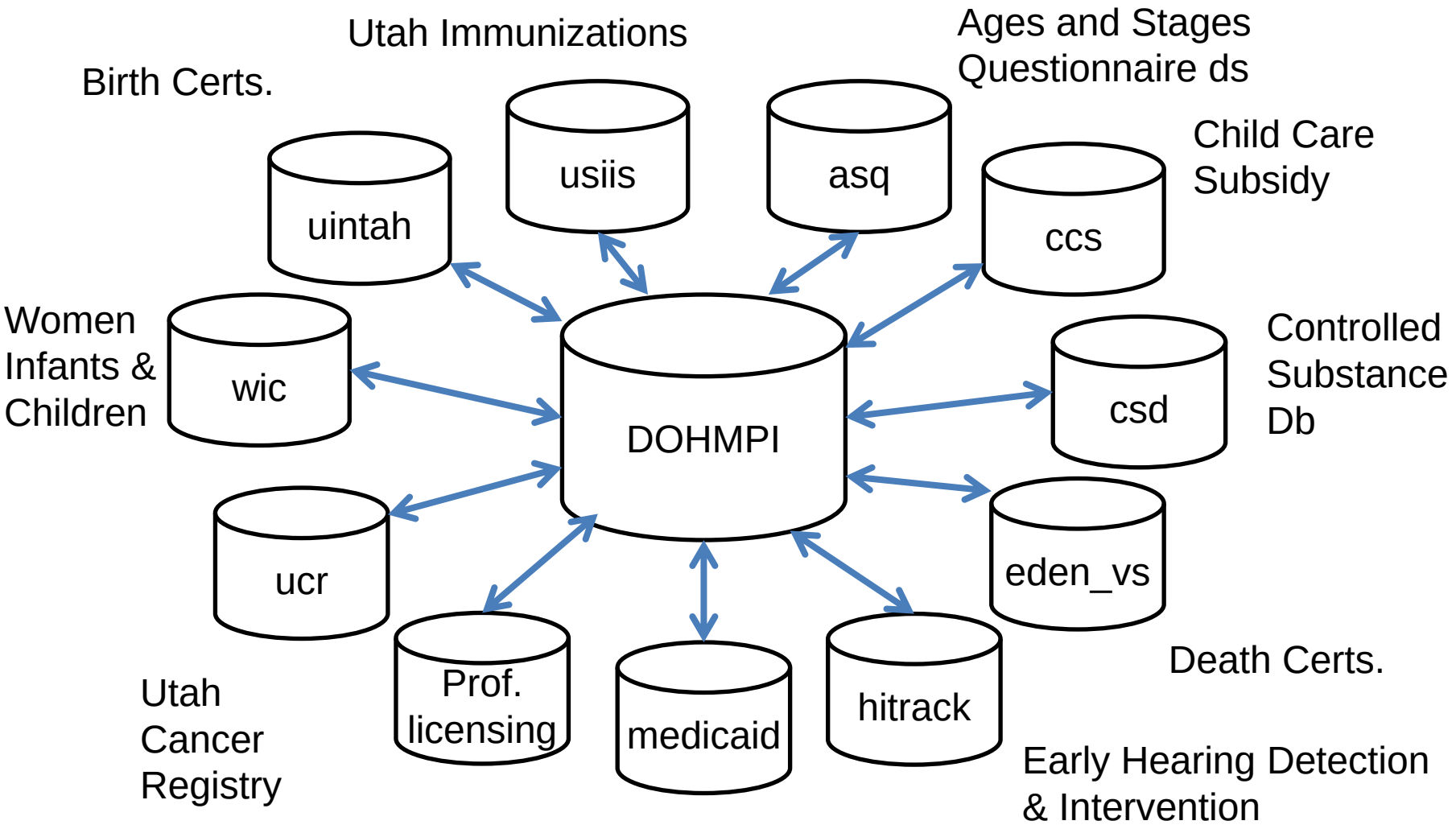
- **Data Integration and Record linkage:**
  - HIO Staff: data transformation, data integration.
  - DTS Staff: data transformation, data integration, and API development.

- **QA:**
  - HP: Modifying/updating R scripts, optimizing matching rules according to the feedback, running monthly QA, providing QA report.
  - Valli, Aihua, Robert, others: Help review and validate matching rules.
  - Humaira Lewon: Manually reviewing R QA system results and providing feedback.

DOHMPI Structure



DOHMPI Structure

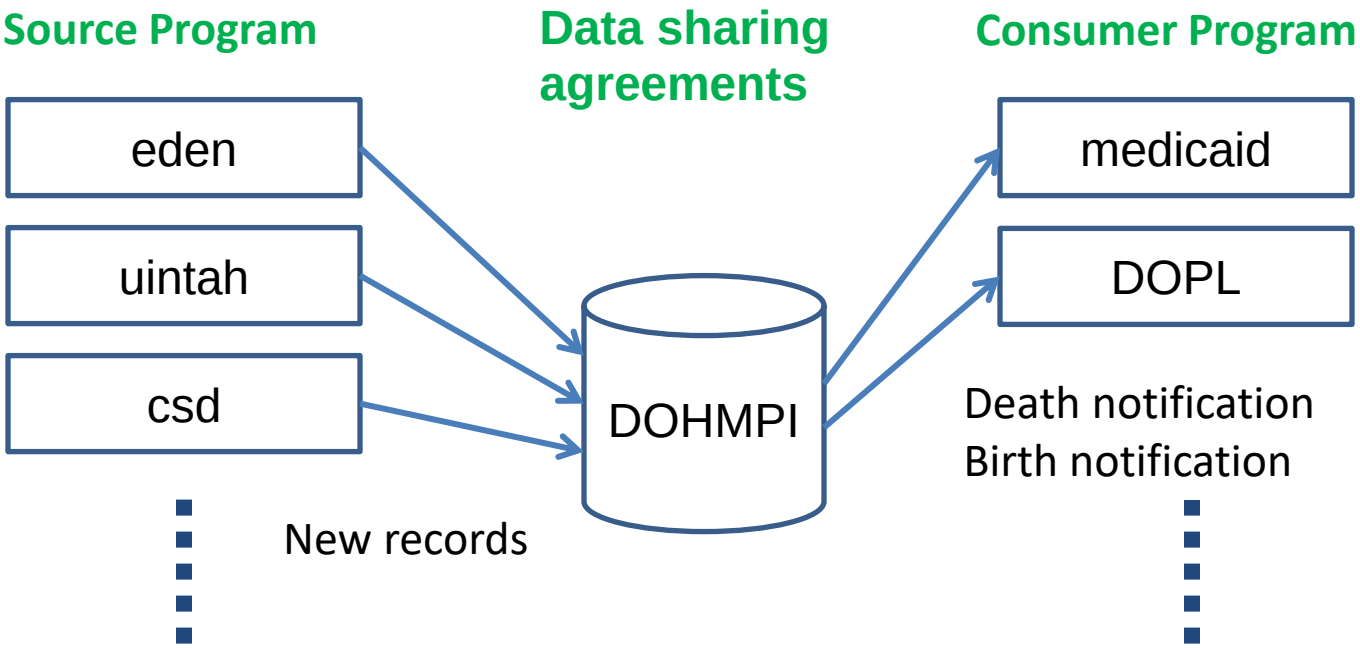


11 Data Sources for June 2019

Demographics data

What DOHMPI can do (use case)

Death notification, DOHMPI notifies source programs (Medicaid) when an individual is linked to a birth/death certificate.



Data quality

Incorrectly linking a (death/any) record to source programs is unacceptable.

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.

Objective of DOHMPI QA

- Monitor DOHMPI linkage precision/recall monthly/weekly.
- Maintain Link Precision > 99% (low false positive links between programs)
- Maintain Link Recall at > 88%.

Precision/ Positive Predictive Value (PPV)

Precision = predict_correct_matches/All_predict_matches = TP/(TP + FP)

Recall / Sensitivity

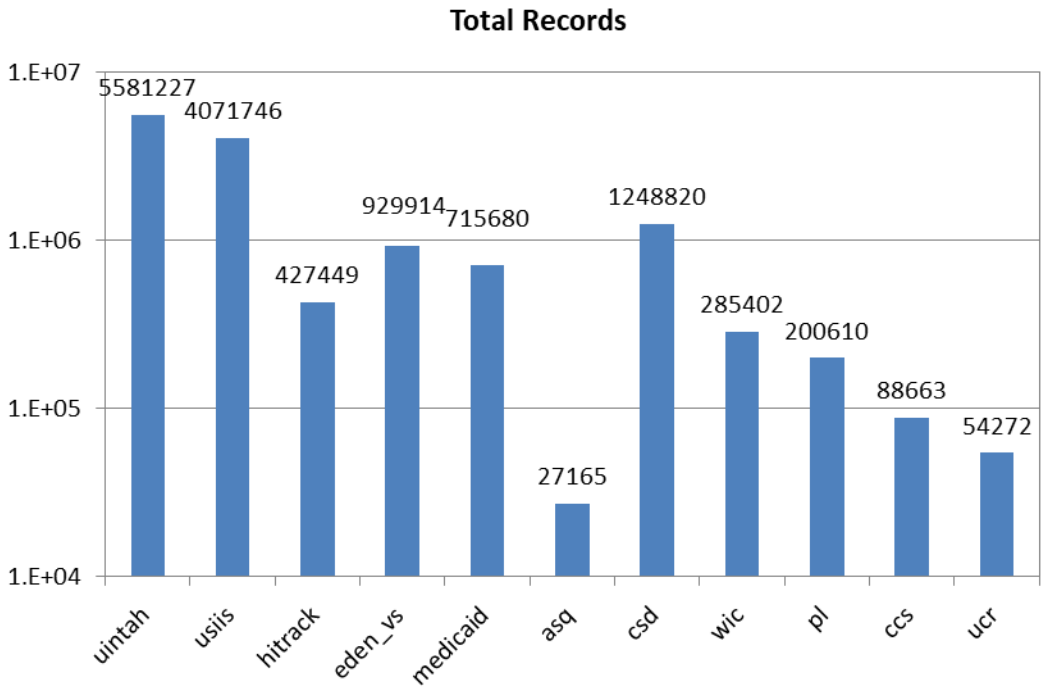
Recall = predict_correct_matches/All_true_matches = TP/(TP + FN)

F1 value

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

Source	Record
uintah	5,581,227
usiis	4,071,746
hitrack	427,449
eden_vs	929,914
medicaid	715,680
asq	27,165
csd	1,248,820*
wic	285,402
pl	200,610
ccs	88,663
ucr	54,272

* pl: professional licensing



Source_1	Source_2	DOHMPI Match
uintah	usiis	1,495,859
uintah	hitrack	397,040
uintah	eden_vs	100,958
uintah	medicaid	418,075
uintah	wic	151,634
… …	… …	… …

Source - Source pairs for QA

FP \ FN	uintah	usiis	hitrack	eden	medicaid	asq	csd	wic	pl	ucr	ccs
uintah		1	1	1	1	1	1	1	1	1	1
usiis	1		1	1	1	1	1	1	1	1	1
hitrack	1	1		1	1	1	1	1	1	1	1
eden	1	1	1		1	1	1	1	1	1	1
medicaid	1	1	1	1		1	1	1	1	1	1
asq	1	1	1	1	1		1	1	1	1	1
csd	1	1	1	1	1	1		1	1	1	1
wic	1	1	1	1	1	1	1		1	1	1
pl	1	1	1	1	1	1	1	1		1	1
ucr	1	1	1	1	1	1	1	1	1		1
ccs	1	1	1	1	1	1	1	1	1	1	

11C2 x 2 = 55 x 2 = 110 schema pairs for QA and manual review.

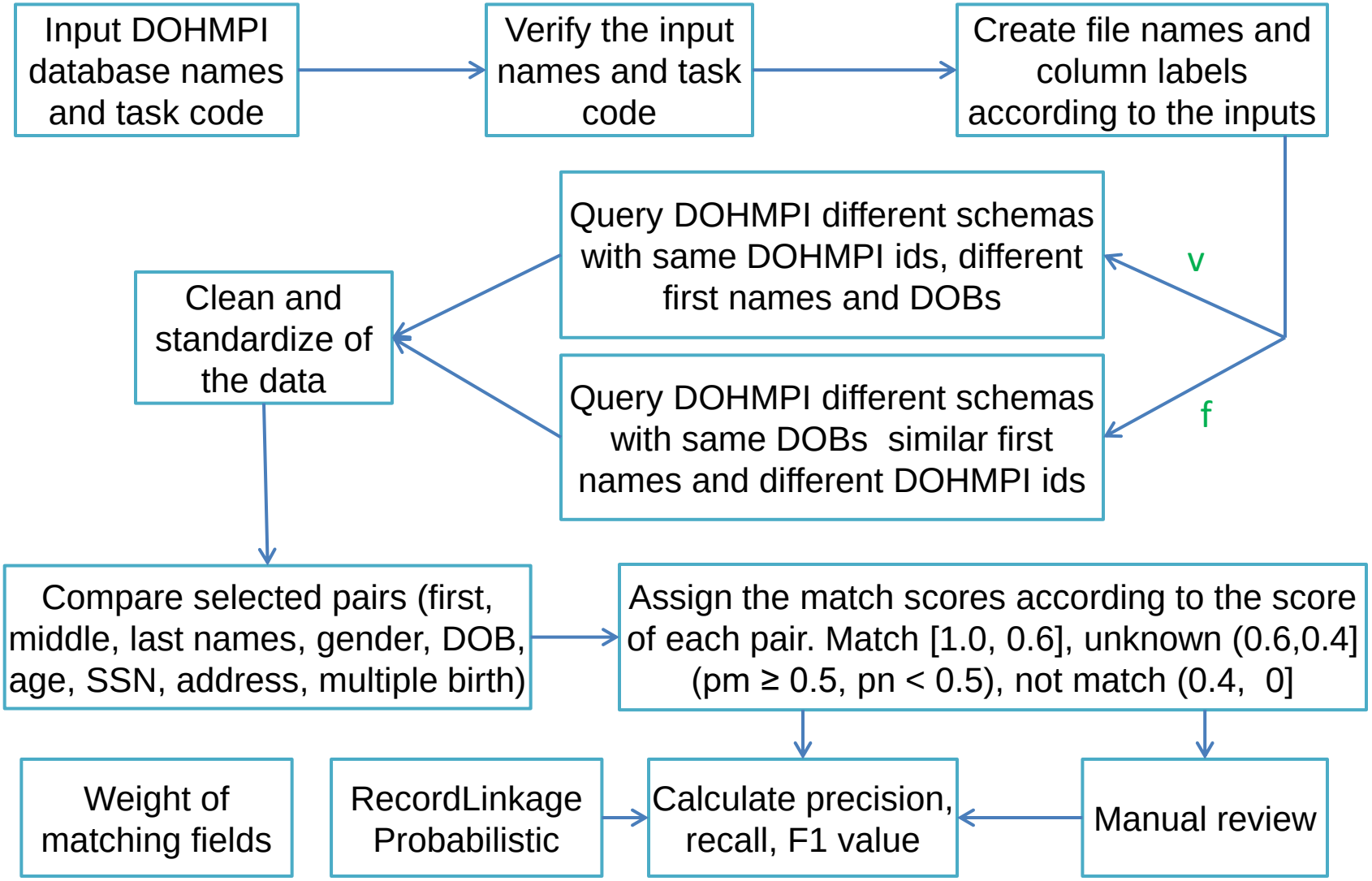
Major Record Linkage Methods

- 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 + Rule-based approach**
- **Probabilistic**

Automate DOHMPI QA process system procedures outline



Methodologies implemented in R

QA procedure overview

- Input, switch, most used functions, 4 SQL query and de-duplicate, clean, standardize, rule-based comparison, probabilistic linkage, report are included in 12 modules (for easy maintenance).

- Big databases use more restrict fuzzy match or exact match, smaller databases use less restricted fuzzy match.
- Provide matching score and the rule number for each record for manual review.
- Create analytical report for improving the precision and recall of DOHMPI.

Input

- Clear user interface. Run part, run all. Extensible.
- Take 2 schema names and a task code (verify DOHMPI match, or find DOHMPI missed match).
- Show the progress of the program running (when wait for hours or more, better show the program is running and estimate the processing time).
- Summarize the numbers of total, match (TP), mismatch (FP), missed match (FN) records.

Switch

- Check user inputs.
- Call the corresponding query modules.

Query

1. Check/verify the schema names, build the output file names according to the input and the date.

   schema1_schema2_date_taskCode_stepCode.txt

   schema1_schema2_date_taskCode_stepCode.csv

   Date: YYYY-MM-DD format for easy sorting.

2. Assign schema source code ids to the column names.
3. Build SQL query according to the inputs.
4. Connect to database.

5. SQL query

(1) Query 2 different DOHMPI schemas with same personids and different first names and DOBs.

- Join, calculate, extract all elements for comparison.

(2) Query 2 different DOHMPI schemas with same DOBs, similar first names and different personids

- Join, calculate, extract all elements for comparison.

Using CTE, subqueries, temp tables optimize queries.

(3) First name, middle name, last name, gender, DOB, age, SSN, street, city, postal code, is multiple birth.

6. Pre-process the query results
   - Replace (|) to (.) and (“) to (‘). We used (|) as delimiter, because there are (,) in our data, and (“”) for field boundary, make sure not remove leading zeros in ids.

7. De-duplication
   - Sort the data first, keep the records with less missing values listed first, recent record listed first, de-duplicate the data, remove same id records with more missing values.

Comparison

1. Clean the data
    - 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.

2. Deterministic comparison

- Compare all pairs (stringdist, Jaro–Winkler distance, is containing, etc.), assign similarity scores. Using nick name table (contains thousands of nick names) to check the nick names. Check the popularity of the names (most popular/common, high frequency names).
- Use comprehensive rules to calculate the match (Other way is to assign the weight factor to all pairs, and sum up all adjusted scores to calculate match, we have this method, but did not be used in manual review).
- According to match score, count the numbers of match, unknown, mismatch, calculate precision, recall and F1 value.

Some match rules

1. All fields are case insensitive, can have 1-2 typos (DOB, SSN will be assigned different scores).
2. SSN + one of (firstname, middlename, lastname, birthdatetime, streetline1).
3. SSN (1-2?) + 3 of (firstname, middlename, lastname, birthdatetime, streetline1).
4. Firstnames + 3 of (middlename, lastname, birthdatetime, streetline1).
5. Firstnames / lastnames + (birthdatetime, streetline1).
   - (4, 5) Multibirth with different firstname will be assigned not match.
6. Firstname, middlename / lastname, birthdatetime + popularity of names in same city/postal code area, or if in the top 400 first name/last name lists.
   - (6) Male and young with different lastname will reduce the score.
7. Only birthdatetime and streetline1 (unknown).
8. Only firstnames and birthdatetime + less popular name in the area (unknown).
   - (8) Male and young with different lastname will reduce the score.
9. Difference ≥ 3 digits in valid SSN + different firstname / lastname (not match).
10. Difference ≥ 3 letters in firstname + true multibirth (not match).

- More popular name, ≥ 66% find people with same firstname/lastname and DOB in the area.
- Less popular name, ≤ 33% find people with same firstname/lastname and DOB in the area.

Probabilistic method using R RecordLinkage package

- Use cleaned data from the previous module.
- Separate the data from the different schemas/datasets.
- Blocking
- Re-pair the records.
- Calculate the matching weights of all pairs.
- Compare with rule-based results.

Manual review R results

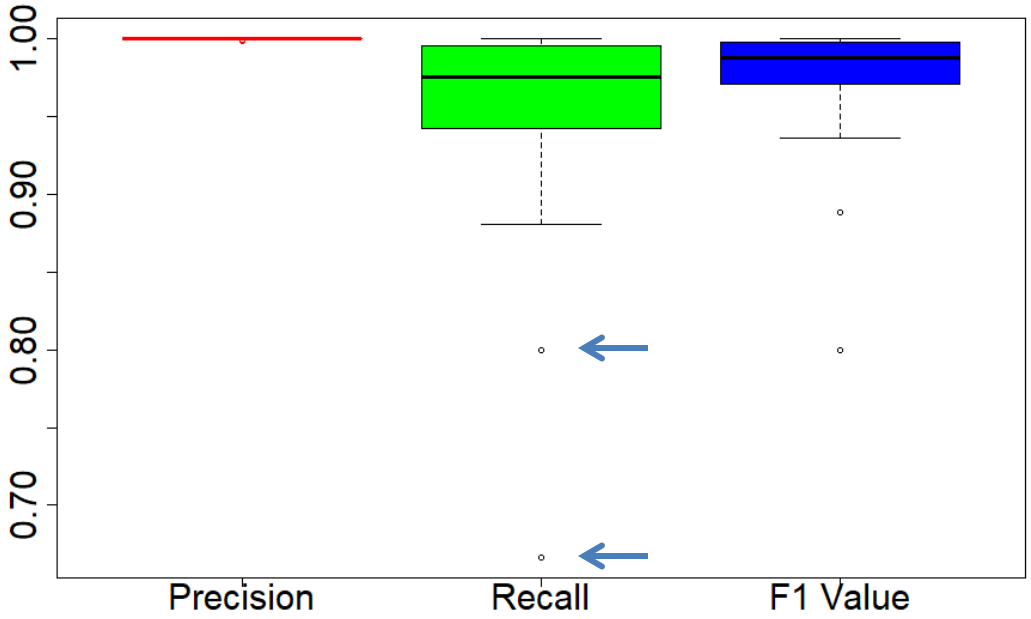
- Focus on pairs containing the select schema at a time.
- R give the sum of matched fields scores and sorted in DESC order.
  - Check records with high sum scores but low match scores.
  - Check records with low sum scores but high match scores.
  - Check records same sum scores but different match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

  - Check all records.

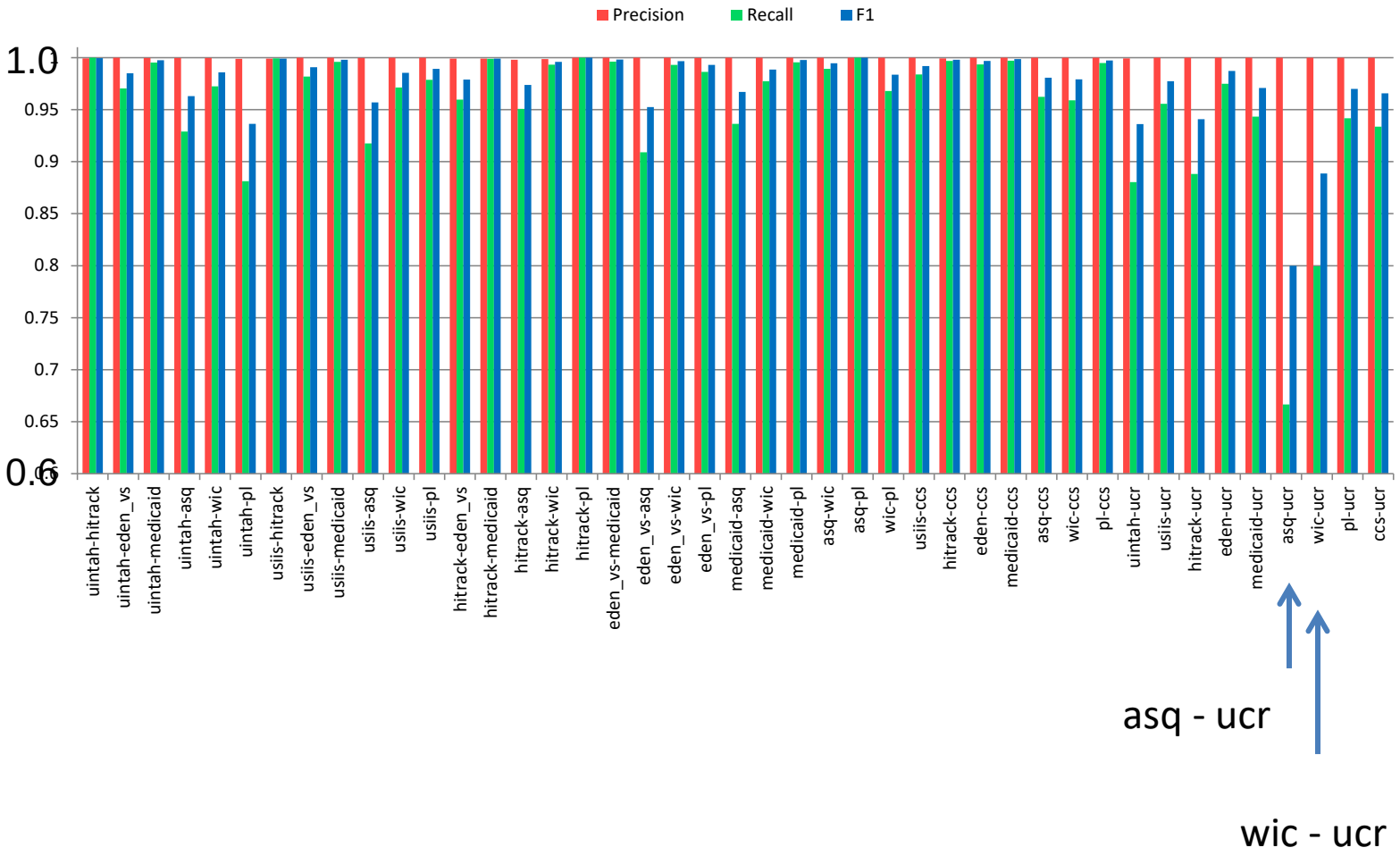
- Provide suggestions for manually correct the results and modify R rules.

Source_1	Source_2	DOHMPI Match	Found New Match	Found Mismatch	Precision	Recall	F1
uintah	hitrack	397040	42	255	99.94%	99.99%	99.96%
uintah	eden_vs	100958	3081	6	99.99%	97.04%	98.49%
uintah	medicaid	418075	1974	88	99.98%	99.53%	99.75%
uintah	asq	12097	926	3	99.98%	92.89%	96.30%
uintah	wic	151634	4289	41	99.97%	97.25%	98.59%
uintah	pl	33275	4477	42	99.87%	88.13%	93.63%
usiis	hitrack	337655	258	307	99.91%	99.92%	99.92%
usiis	eden_vs	133108	2466	1	100.00%	98.18%	99.08%
usiis	medicaid	552038	2241	15	100.00%	99.60%	99.80%
usiis	asq	12509	1124	3	99.98%	91.75%	95.69%
usiis	wic	148621	4403	8	99.99%	97.12%	98.54%
usiis	pl	93386	2044	2	100.00%	97.86%	98.92%
hitrack	eden_vs	2214	93	2	99.91%	95.97%	97.90%
hitrack	medicaid	113391	122	105	99.91%	99.89%	99.90%
hitrack	asq	10371	537	21	99.80%	95.07%	97.38%
hitrack	wic	72237	491	98	99.86%	99.32%	99.59%
eden_vs	medicaid	21931	83	0	100.00%	99.62%	99.81%
eden_vs	pl	6220	87	1	99.98%	98.62%	99.30%
medicaid	asq	8865	603	1	99.99%	93.63%	96.71%
medicaid	wic	112362	2616	5	100.00%	97.72%	98.85%

*** 55 dataset pairs at this time**



	Precision	Recall	F1 Value
Min.	0.9980	0.6667	0.8000
1st Qu.	0.9998	0.9425	0.9704
Median	1.0000	0.9749	0.9873
Mean	0.9997	0.9562	0.9764
3rd Qu.	1.0000	0.9950	0.9974
Max.	1.0000	1.0000	1.0000


Precision
Recall
F1
1.0
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
uintah-hitrack
uintah-eden_vs
uintah-medicaid
uintah-asq
uintah-wic
uintah-pl
usiis-hitrack
usiis-eden_vs
usiis-medicaid
usiis-asq
usiis-wic
usiis-pl
hitrack-eden_vs
hitrack-medicaid
hitrack-asq
hitrack-wic
hitrack-pl
eden_vs-medicaid
eden_vs-asq
eden_vs-wic
eden_vs-pl
medicaid-asq
medicaid-wic
medicaid-pl
asq-wic
asq-pl
wic-pl
usiis-ccs
hitrack-ccs
eden-ccs
medicaid-ccs
asq-ccs
wic-ccs
pl-ccs
uintah-ucr
usiis-ucr
hitrack-ucr
eden-ucr
medicaid-ucr
asq-ucr
wic-ucr
pl-ucr
ccs-ucr
asq - ucr
wic - ucr

Leveraging QA ‘R’ Code

- Modularity and flexibility of the QA system, the ‘R’ code can be leveraged to automatically check the linkage quality for other person matching software such as Informatica Identity Resolution (IIR).
  - Change connection
  - Input format/interface

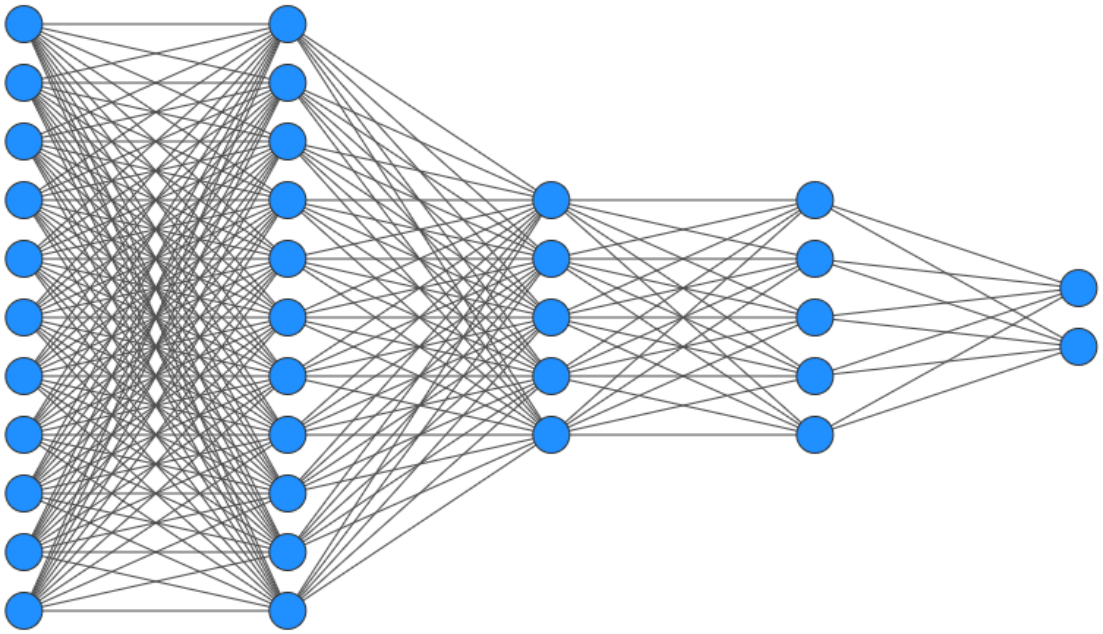
Challenges

- Most records do not have SSN.
- SSN not match, most other fields match (more than 1 SSN?).
- Not enough info, missing values (Not only for research, we cannot ignore).
- Name change, women/men may change their names (marriage, adoption).
- Address change (first name, last name, DOB, gender match, address not match).
- Different address formats (de-identified by replace by 0s)
  - 000 W 00 S, 012 WEST 00 SOUTH
  - 000 N 000 W APT 11-A, 000 NORTH 000 WEST 11A
  - 0033 BROCKET DEER DR00000, 0022 W Brocket Deer Dr
- Simplified name vs full name, multiple words in first/last names.
- Baby A, BA John, 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 algorism (edition, Jaro-Winkler, containing).
  - Analy Aneliy; Izzabella Isabella; Orin Oren (not in nickname, sound?)
- Twin:
  - All same but slightly different in first name may be different people (JADEN KADEN).
  - All different except first names and DOBs may be same person.

Future work

- Continue improve matching rules using manual review results.
- Maintain DOHMPI precision to more than 99% while improving recall rates to 90% or greater for all schema pairs.
- DOHMPI can automatically run on server.
- Develop a sound similarity comparing tool to compare every syllable.
- Using natural language processing (NLP) methods to compare address.
- Add machine learning/deep learning methods.

Input:
11
features/
variables



Output:
Match/
Not
match

ACKNOWLEDGEMENT

Navina Forsythe
Kailah Davis
Humaira Lewon
Matthew Plendl
Robert Wilson
Brantley Scott

Jeffrey Duncan
Valli Chidambaram
Aihua Tong
Zakir Guler
Jon Reid
Mike Jolley

THANK YOU

Q/A