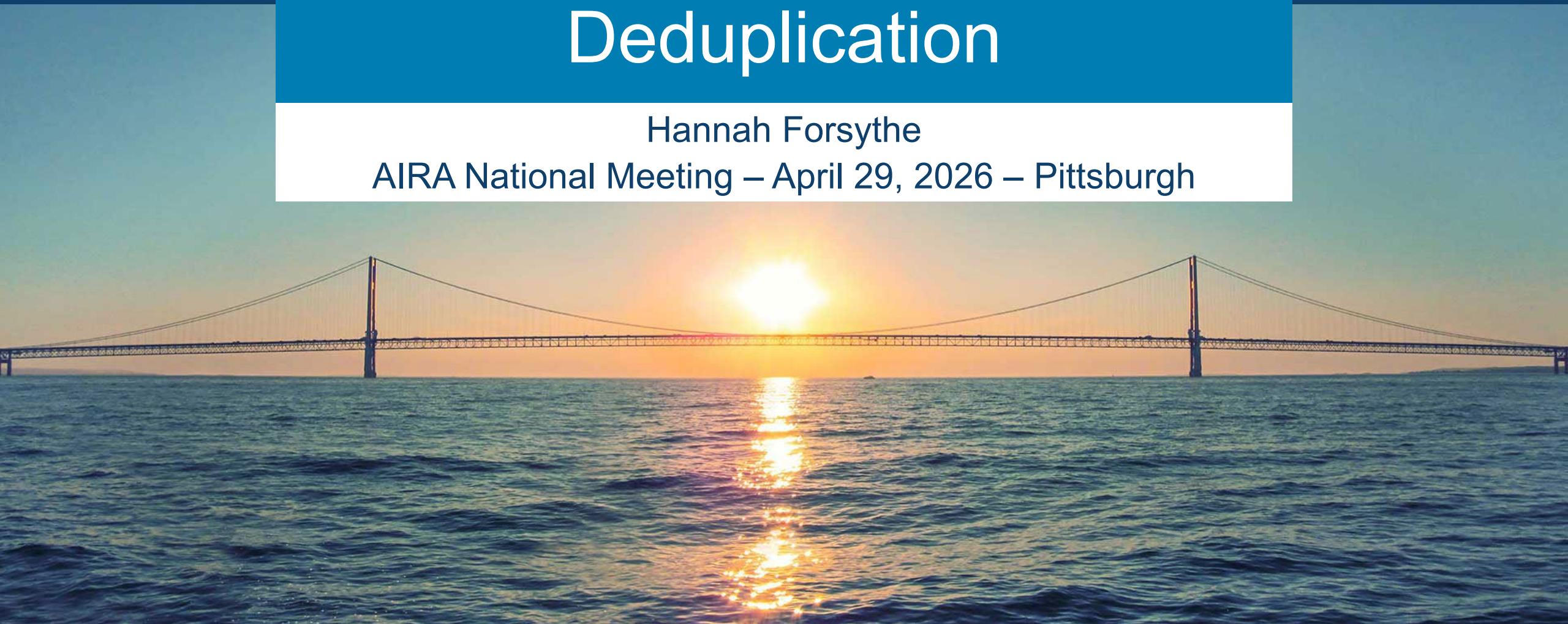
MDHHS
Michigan Department of Health & Human Services

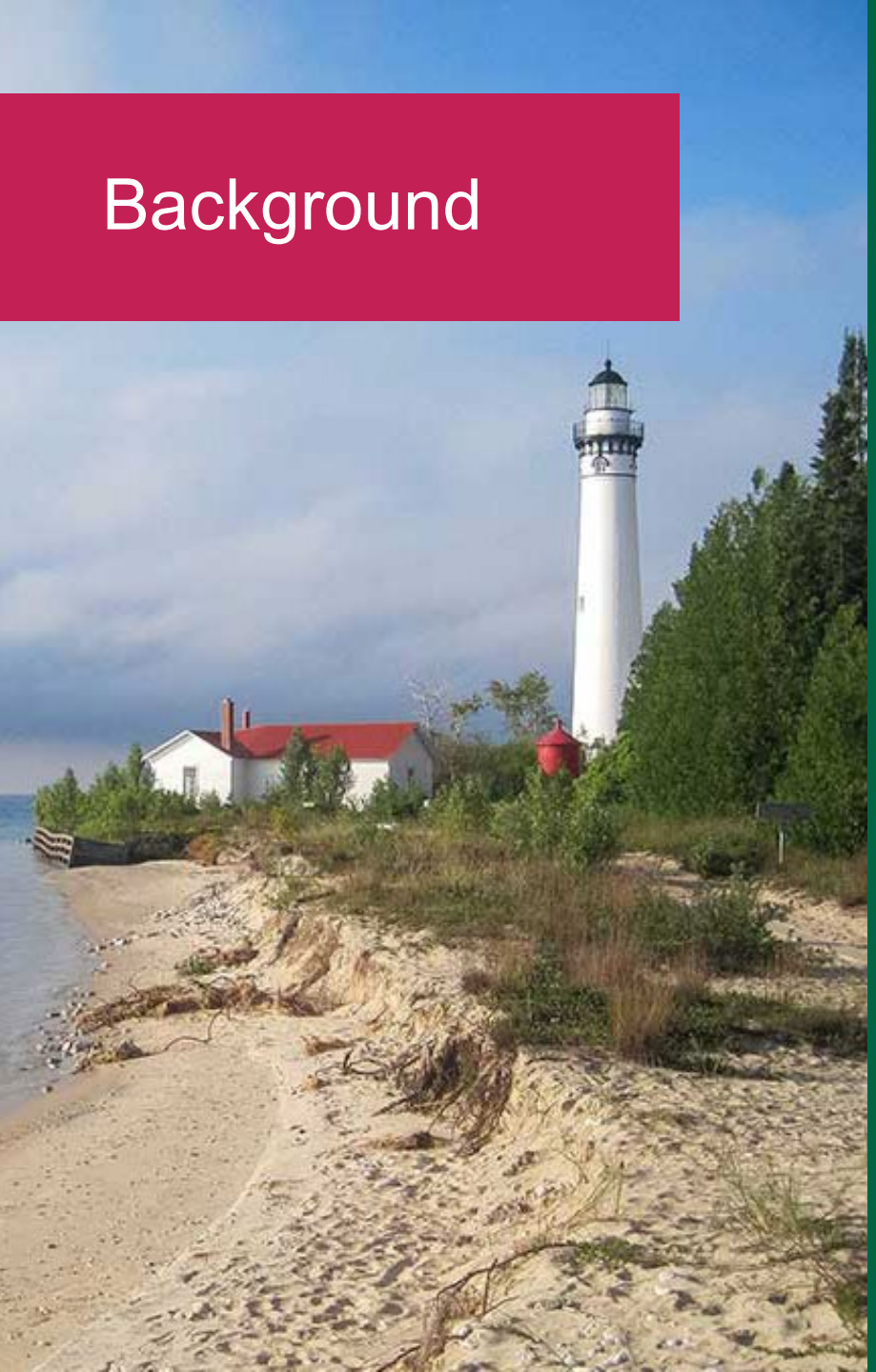
A Strategy for Optimizing Bulk Deduplication

Hannah Forsythe

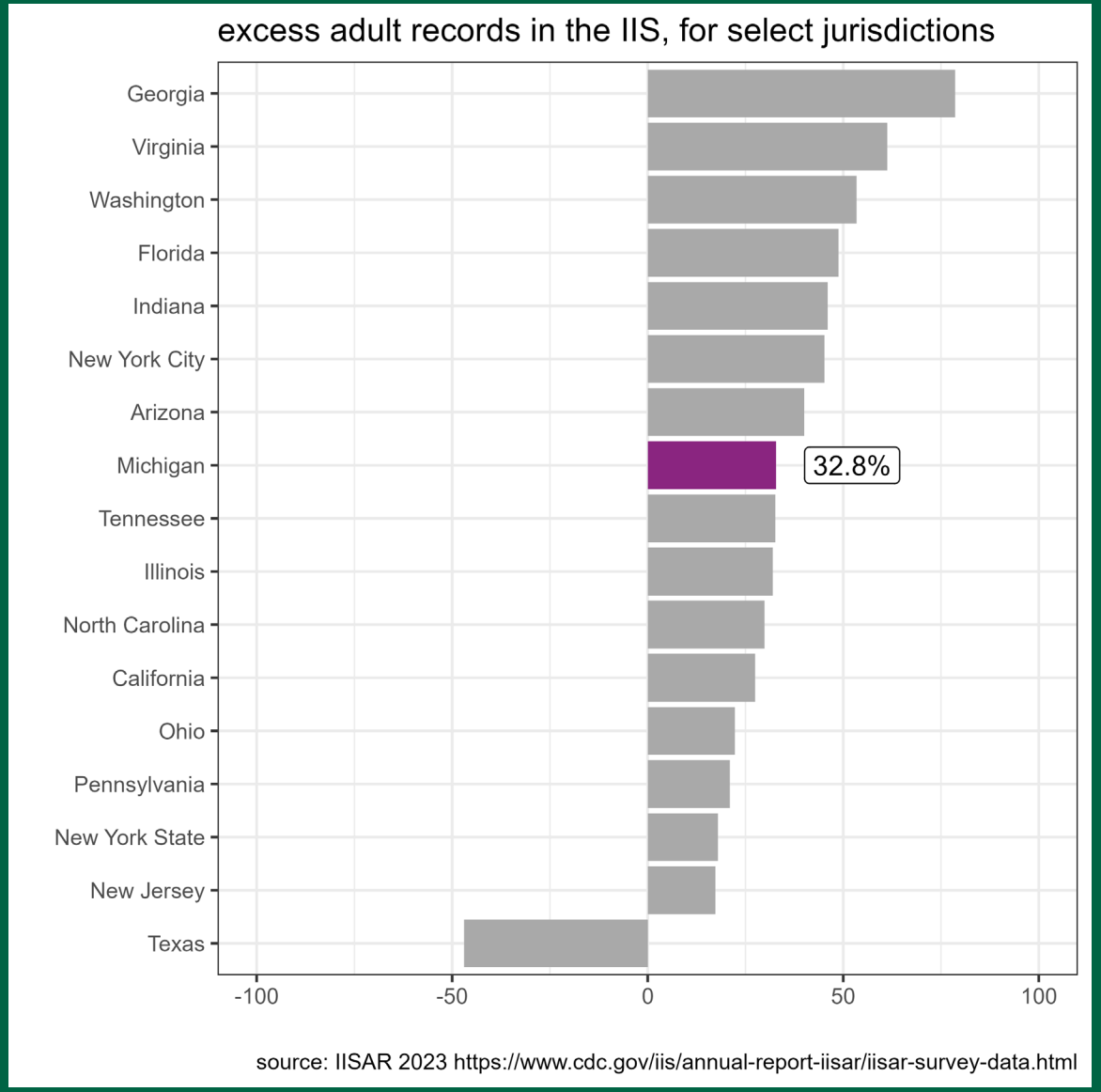
AIRA National Meeting – April 29, 2026 – Pittsburgh

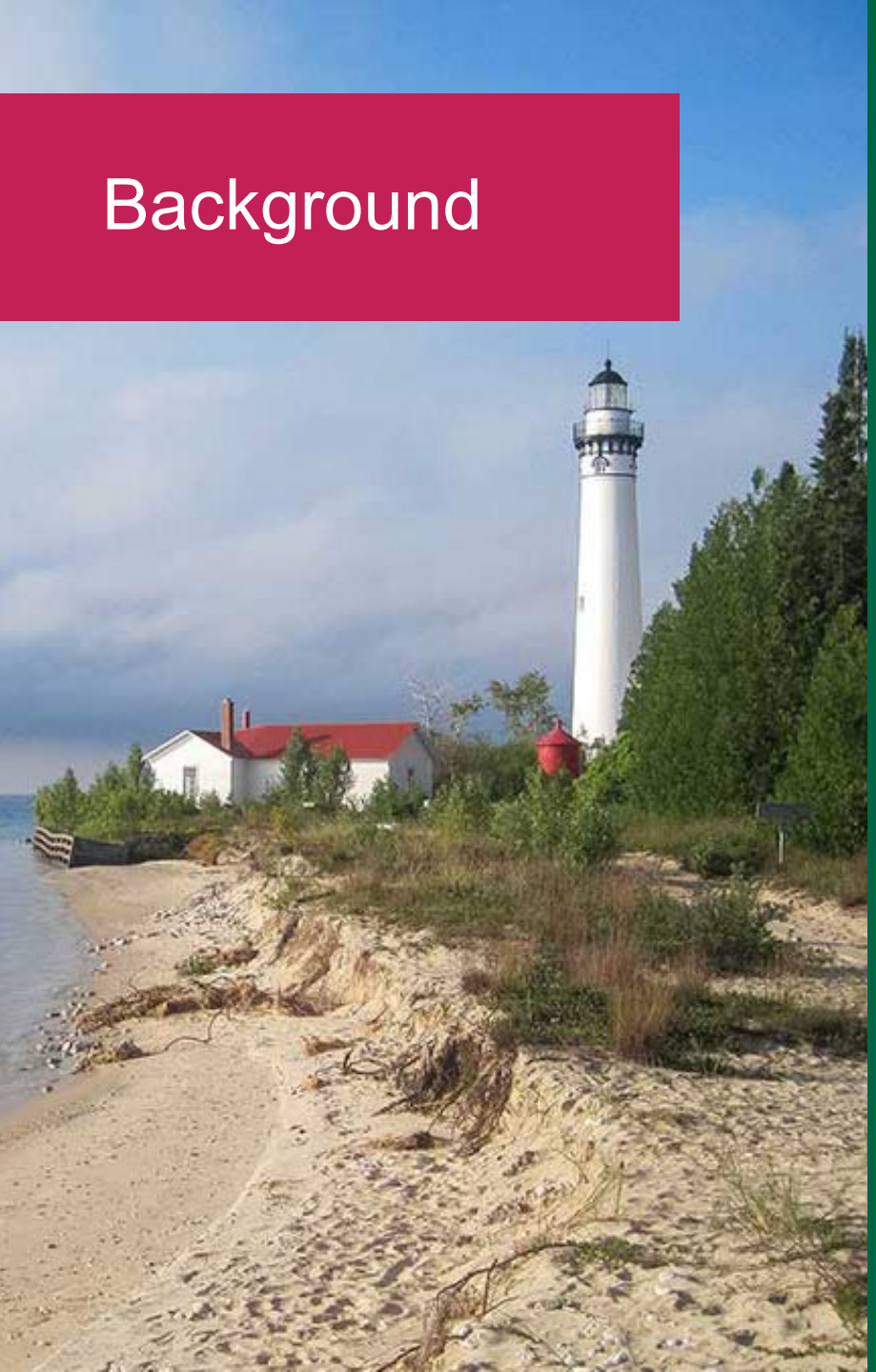


Background



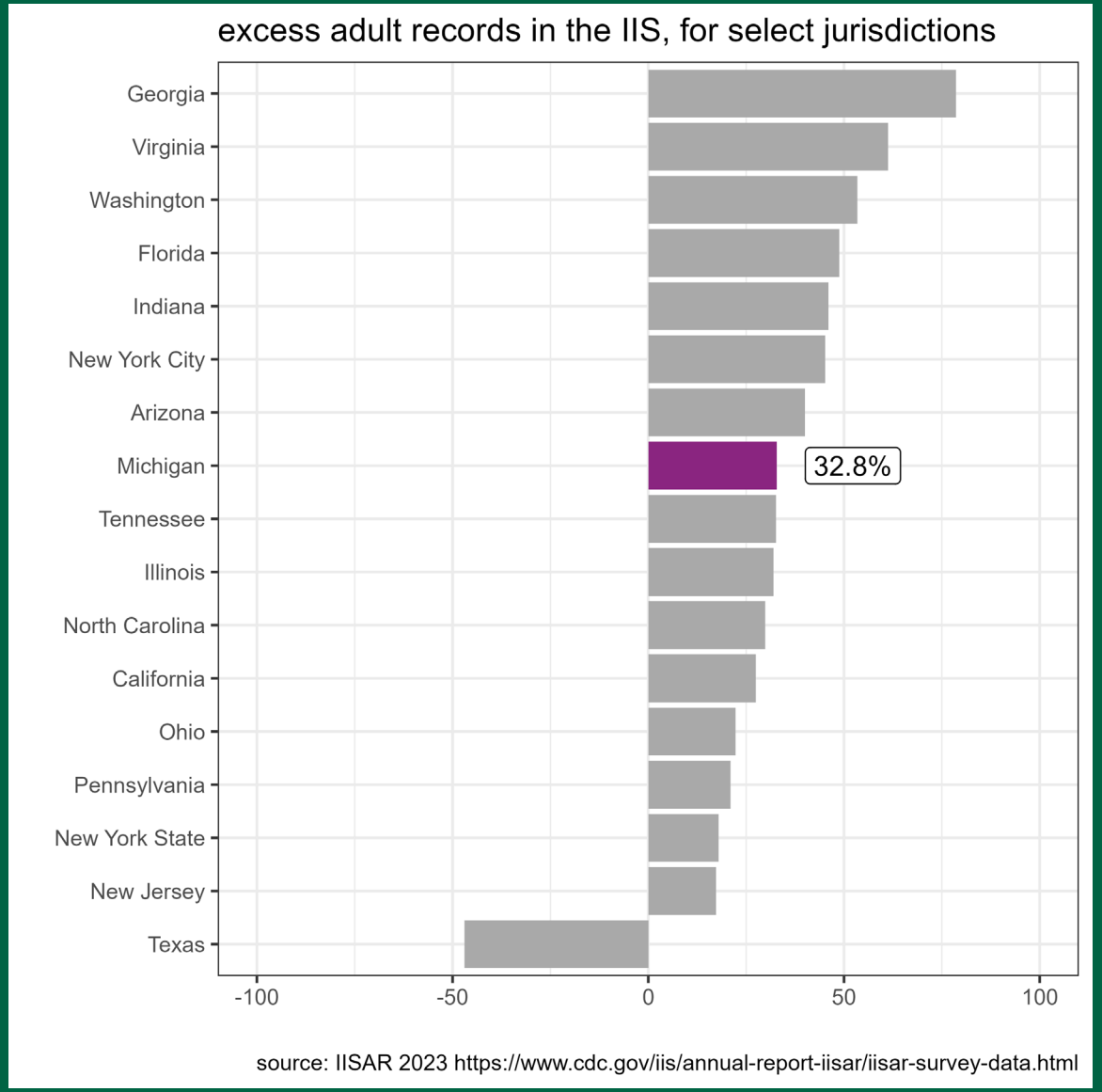
Denominator inflation is a problem for most Immunization Information Systems.



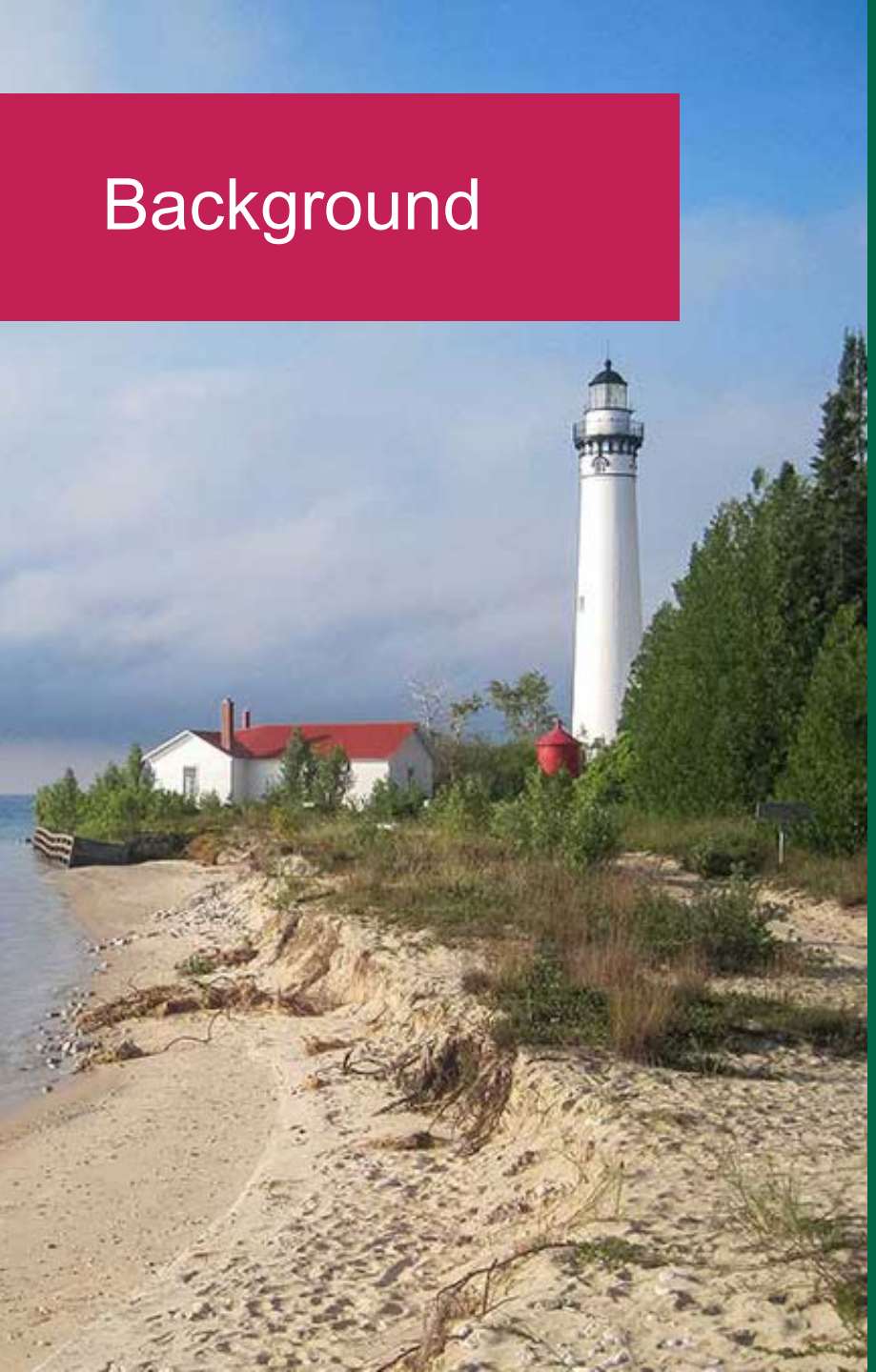


Background

Contributors to Denominator Inflation

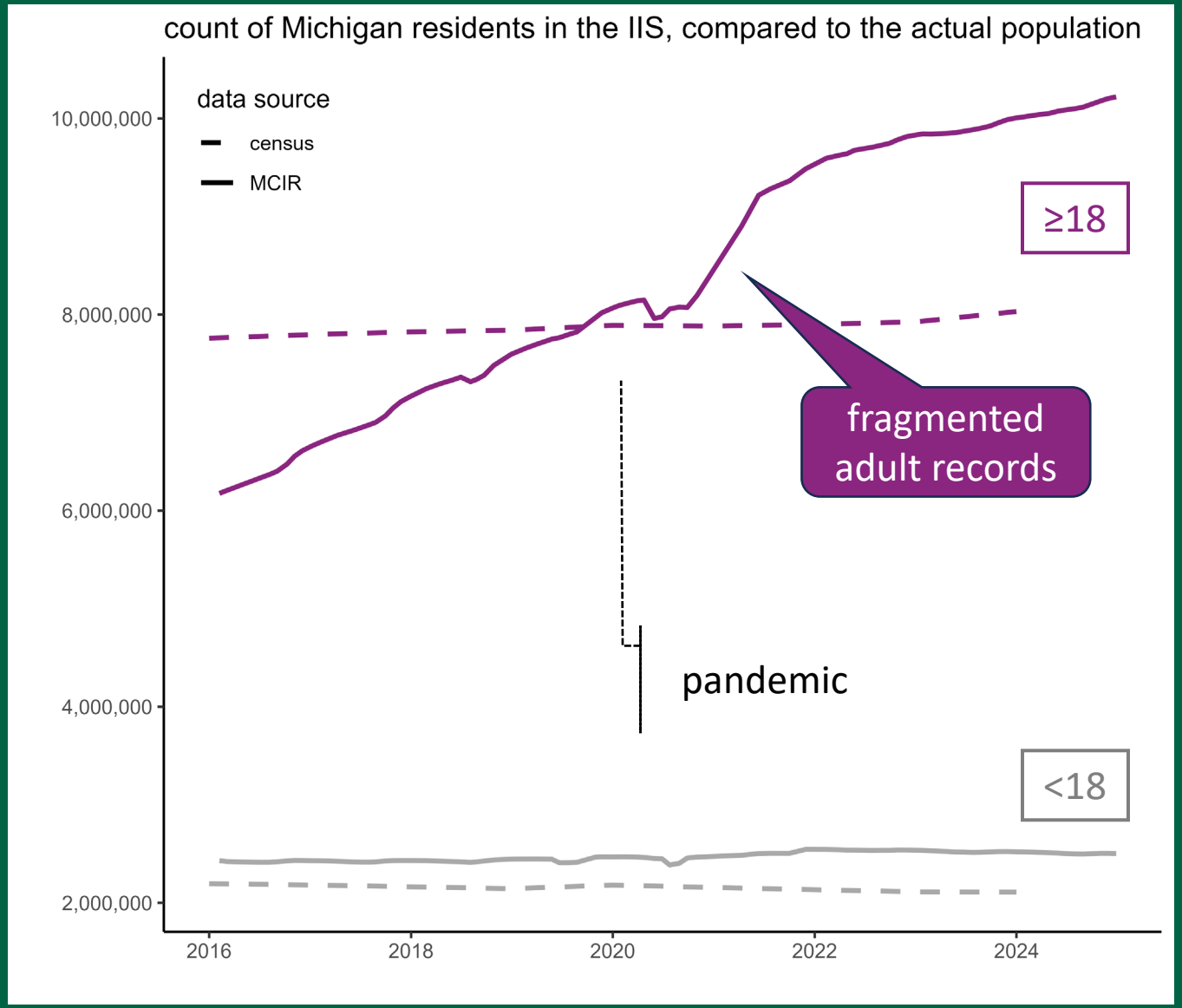


- stale addresses
- unnamed infant records
- untracked last name changes
- aliases, nick names, typos, etc.

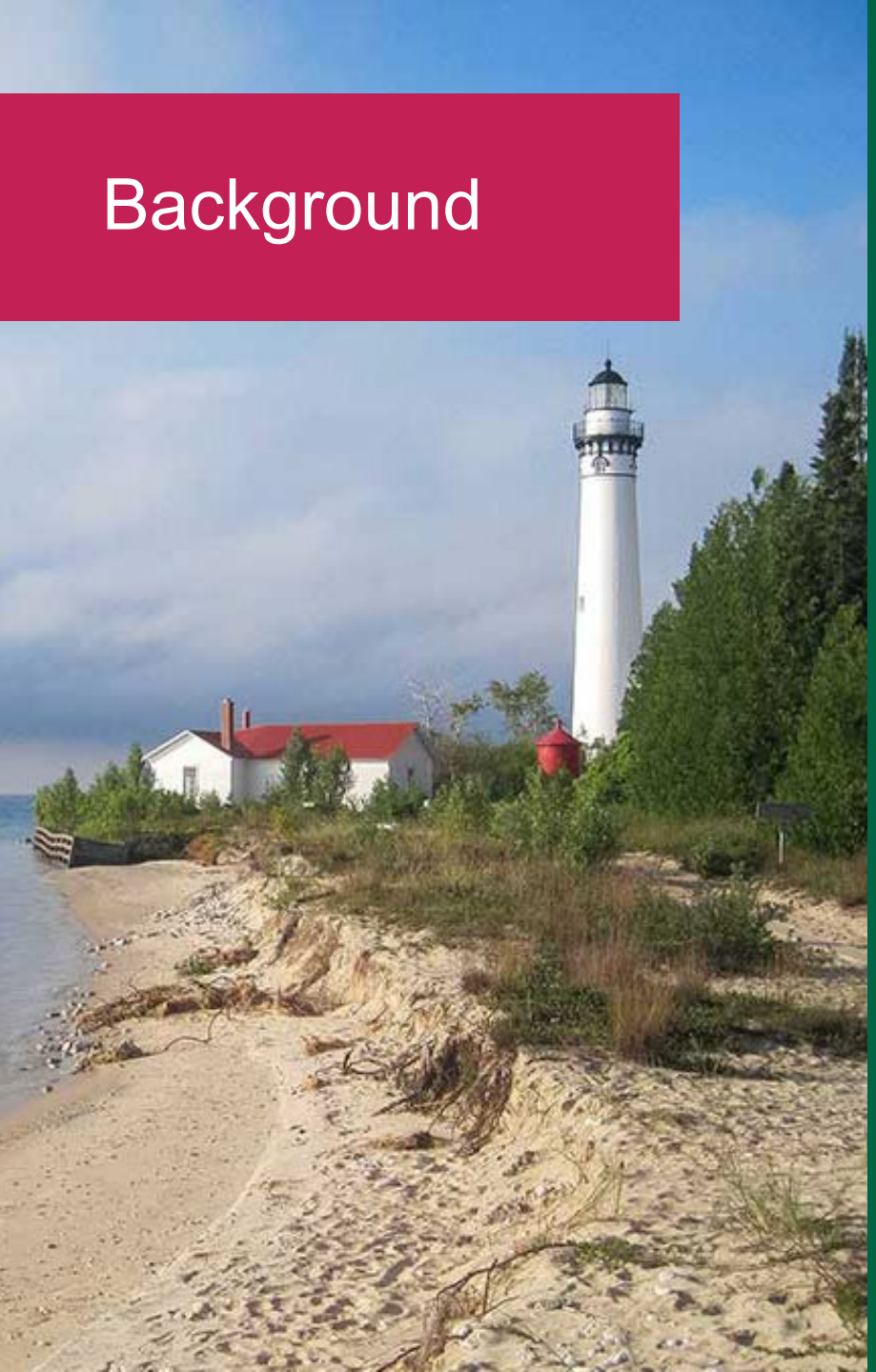


Background

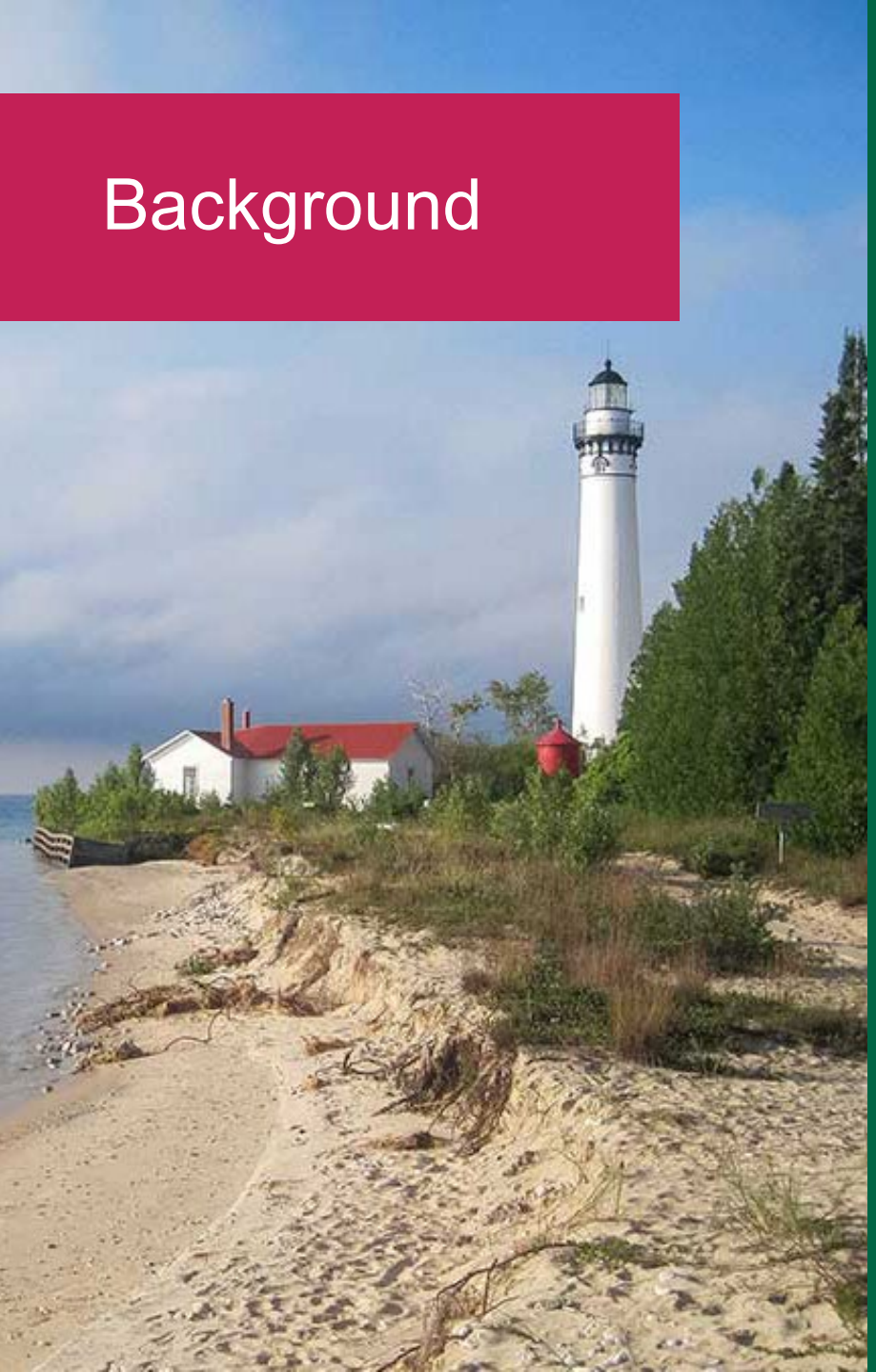
Contributors to Denominator Inflation



Background



We need to deduplicate accurately, at scale.



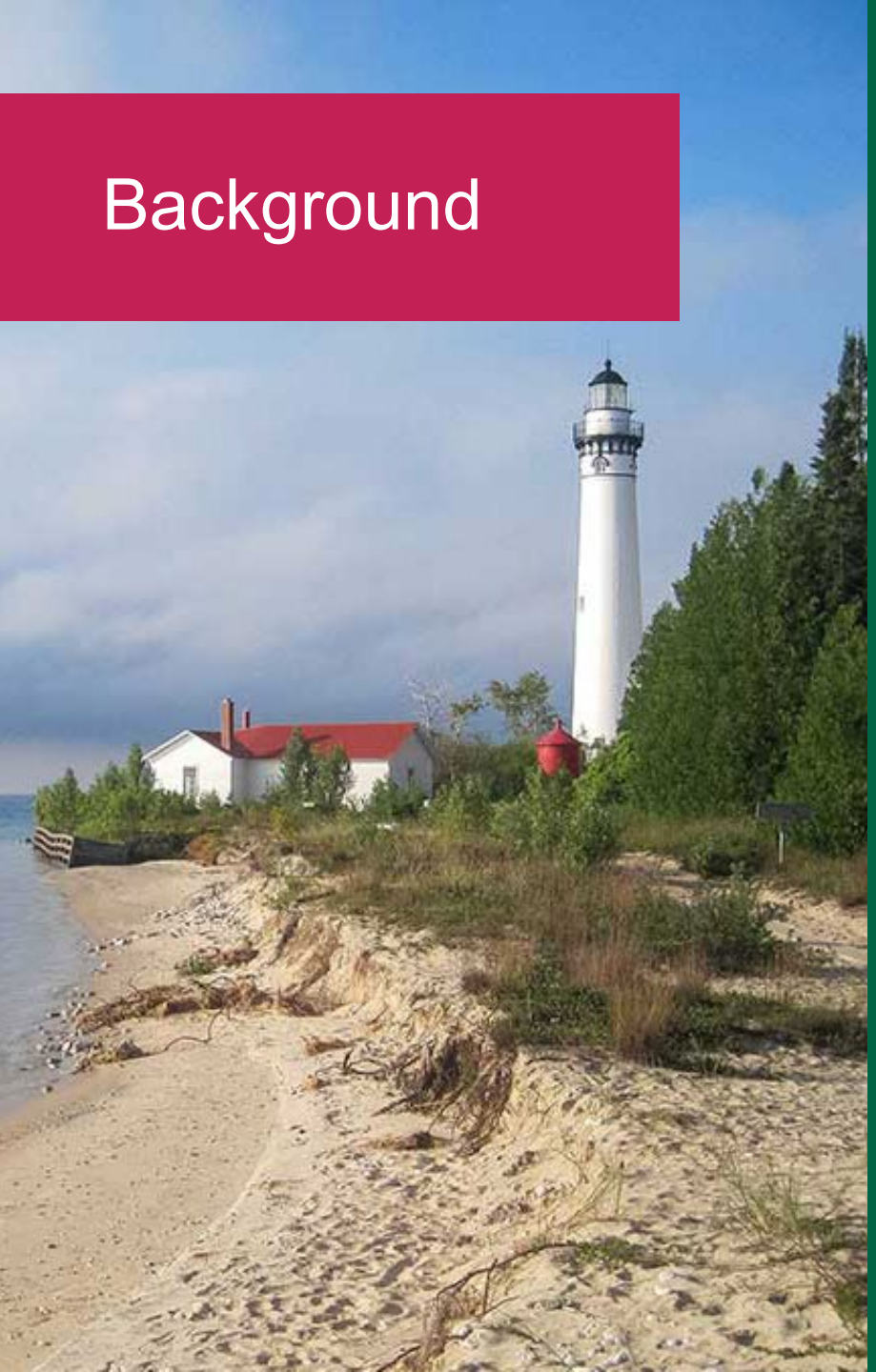
Background

Deduplication strategies

- Prevention
- Auto-link records
- Analytical adjustment
- Manual correction

Best

Worst



Background

Deduplication strategies

Can we combine the best of both?

Auto-link records

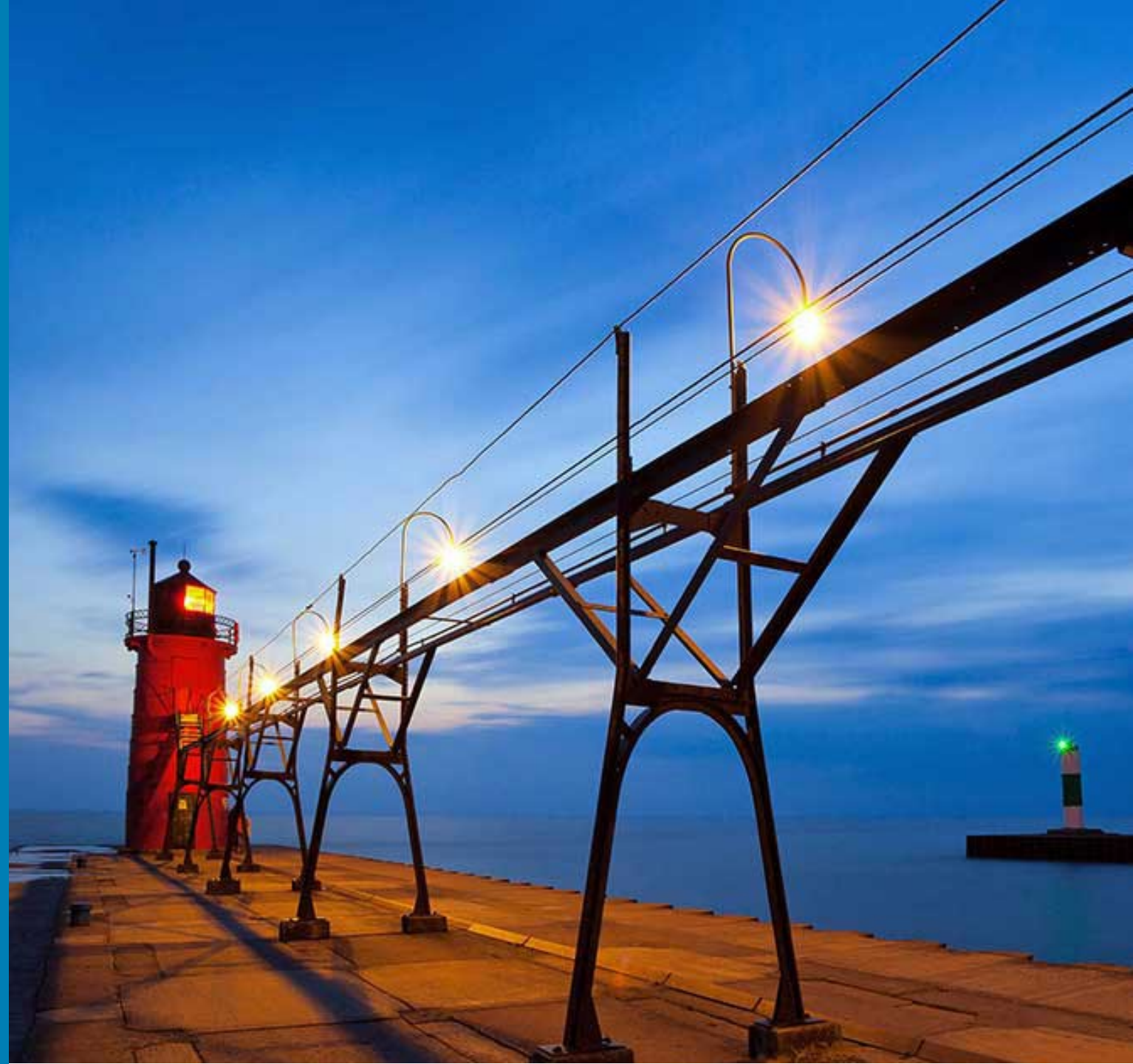
High volume

Manual correction

High quality

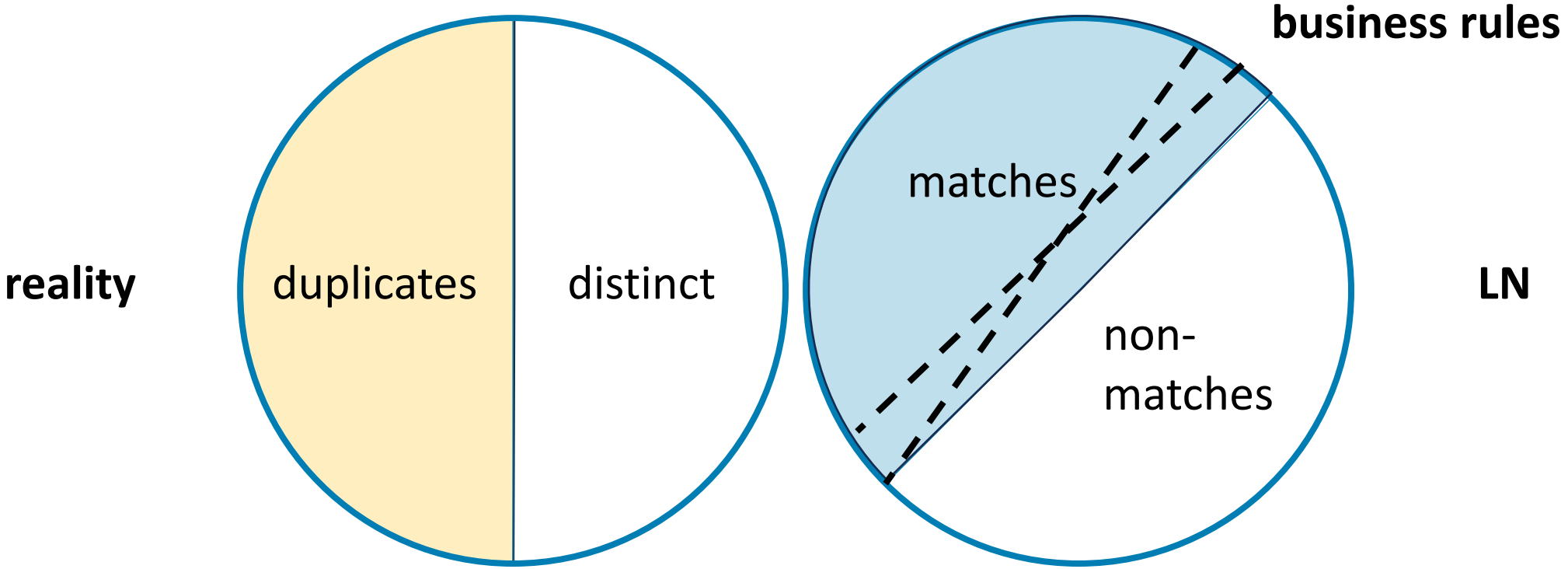
Methods

MDHHS
Michigan Department of Health & Human Services



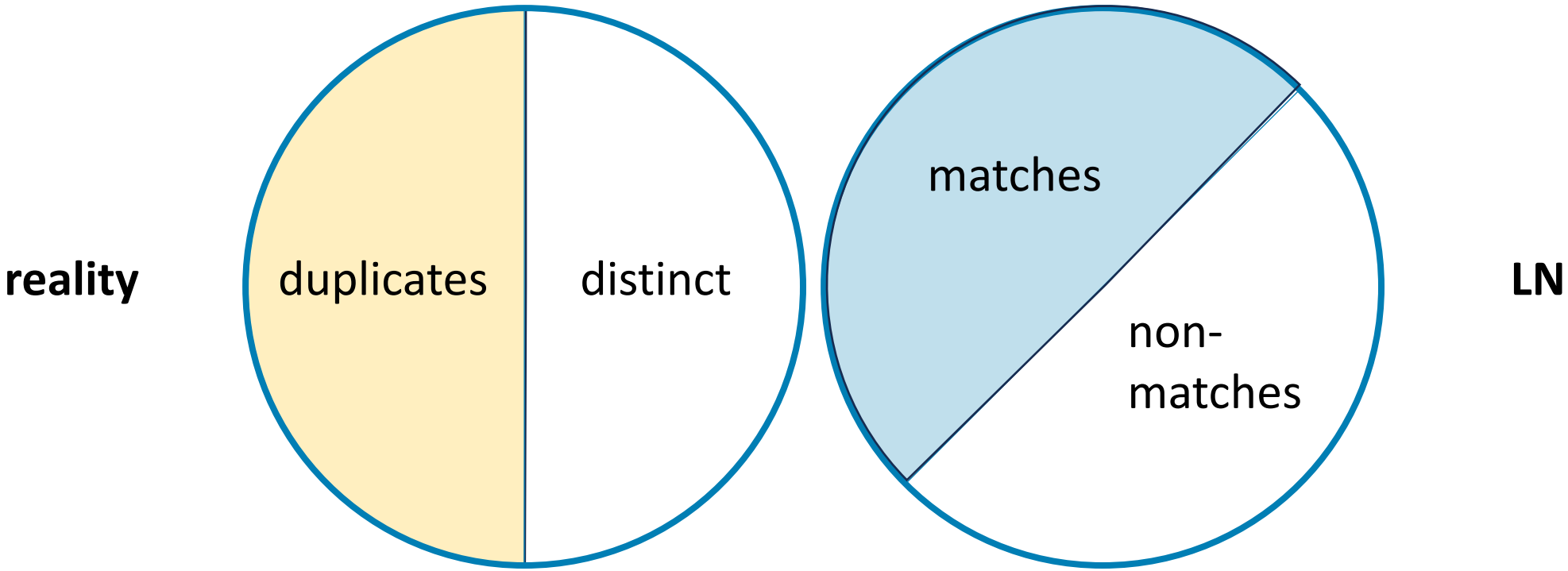
Methods

- Lexis Nexis suggests merges.
- Develop business rules.
  - Perform exploratory data analysis.
  - Propose rules for accepting/rejecting.
  - Randomly sample and review by hand.



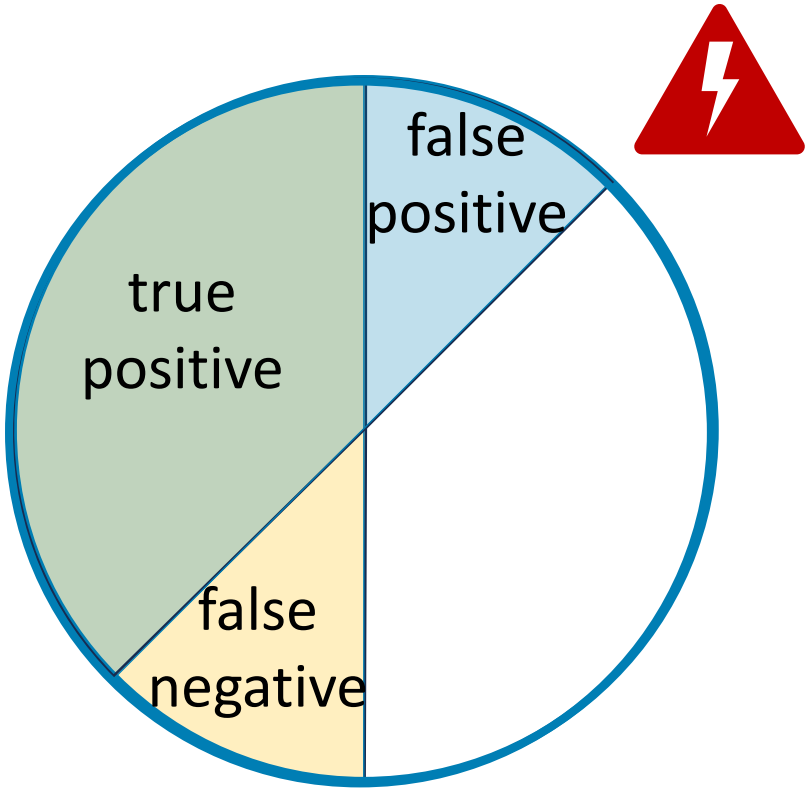
Methods

- Lexis Nexis suggests merges.
- Develop business rules.
  - Perform exploratory data analysis.
  - Propose rules for accepting/rejecting.
  - Randomly sample and review by hand.



Methods

- Lexis Nexis suggests merges.
- Develop business rules.
  - Perform exploratory data analysis.
  - Propose rules for accepting/rejecting.
  - Randomly sample and review by hand.



reality

LN

Methods

- Lexis Nexis suggests merges.
- Develop business rules.
  - Perform exploratory data analysis.
  - Propose rules for accepting/rejecting.
  - Randomly sample and review by hand.
- Evaluation metrics:

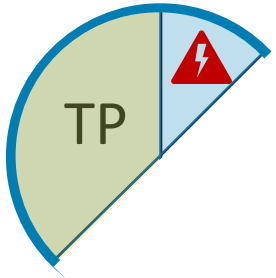
$$precision = \frac{True\ matches}{all\ suggested\ matches} = \frac{TP}{TP+FP}$$

$$sensitivity = \frac{True\ matches}{all\ real\ duplicates} = \frac{TP}{TP+FN}$$

Methods

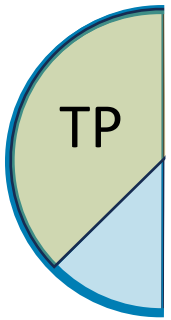
- Lexis Nexis suggests merges.
- Develop business rules.
  - Perform exploratory data analysis.
  - Propose rules for accepting/rejecting.
  - Randomly sample and review by hand.
- Evaluation metrics:

$$precision = \frac{TP}{TP + FP}$$



must exceed 0.99

$$sensitivity = \frac{TP}{TP + FN}$$



higher is better

Results

MDHHS
Michigan Department of Health & Human Services

Lexis Nexis data processing

Input: 14.7 million person records

Processed: 14.4 million

Output: 13.3 million distinct records

Suggested merges

- 881,145 pairs (2 records)
- 66,694 clusters (3+ records)

- 23,948 pairs are known false positives *

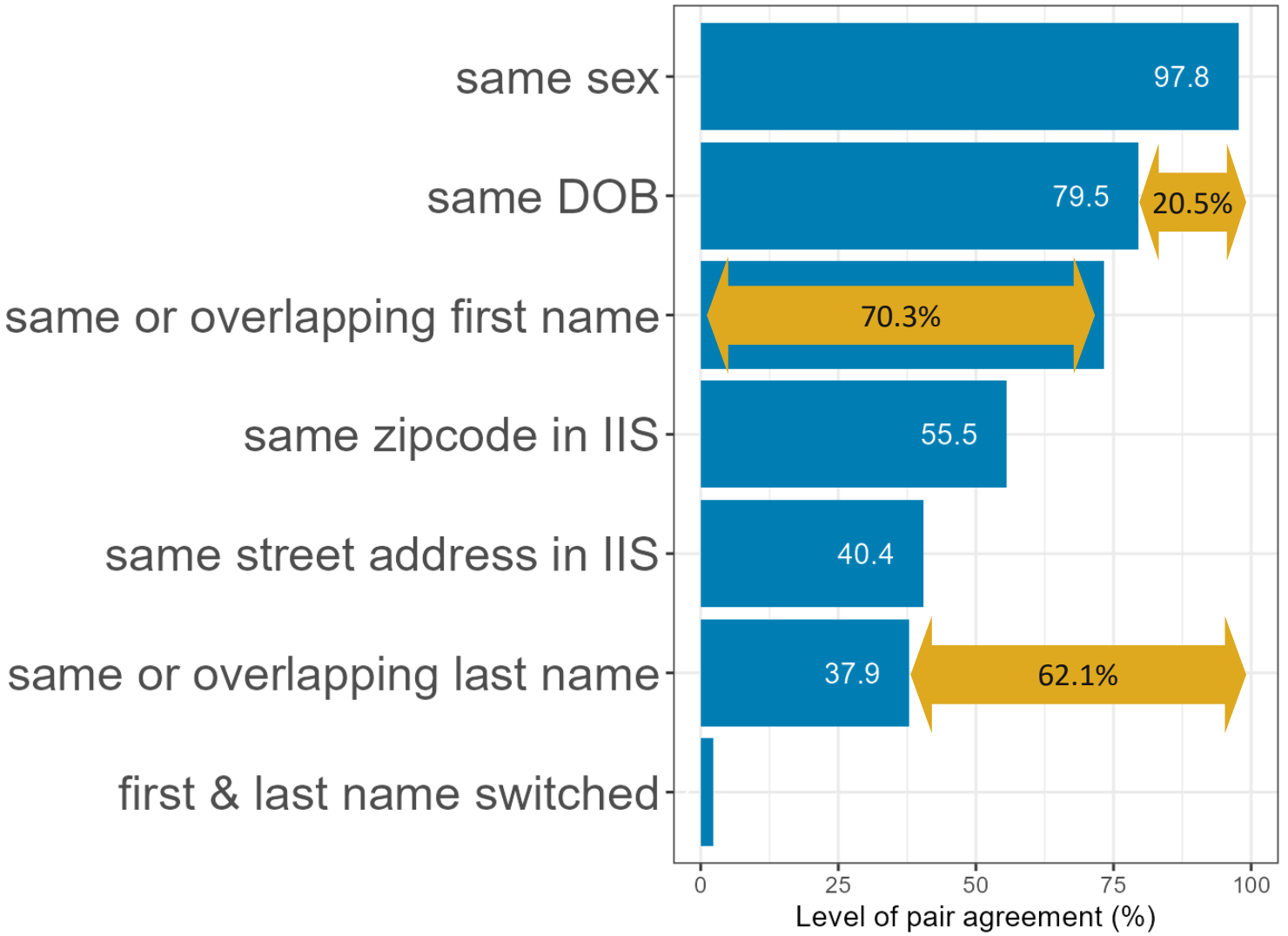
*Manually flagged in IIS; represents 2.7% of 873,511 non-deleted pairs as of March 30, 2026.

Results

Exploratory Data Analysis of Suggested Merges

849,563 possible pairs remaining

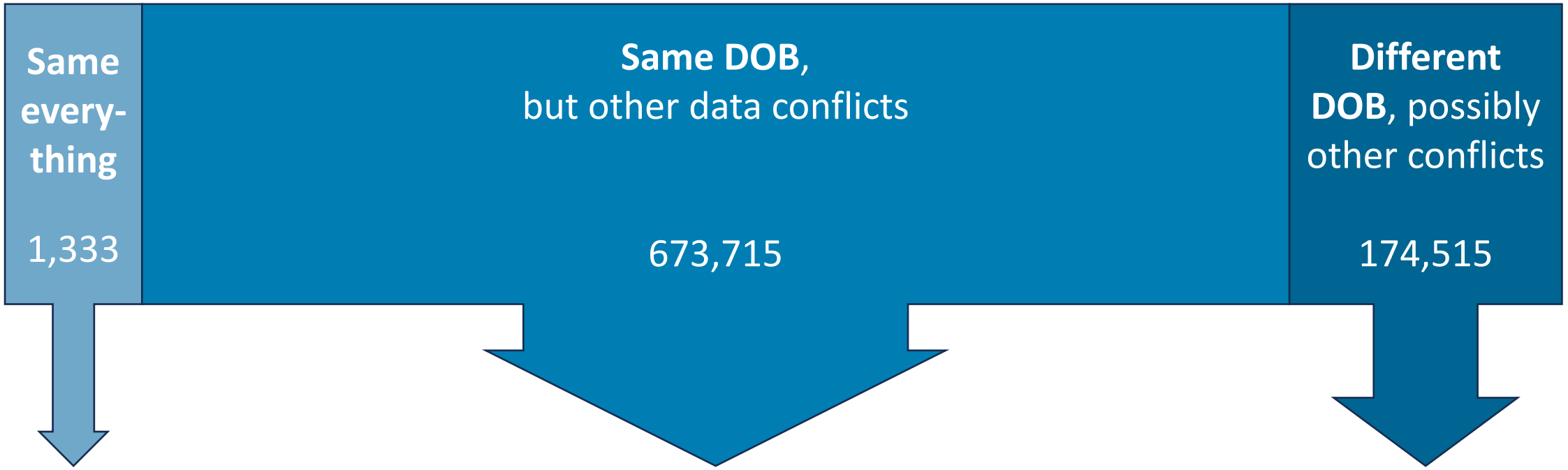
- Nearly all same sex.
- 20% different DOB.
- Majority have same first name, different last name.
- Address data at rest in the IIS not very decisive.



Results

Exploratory Data Analysis of Suggested Merges

849,563 possible pairs remaining



Least sensitive:
OK to merge.

Most sensitive:
Test business rules here first.

Results

Exploratory Data Analysis of Sensitive Merges

174,515 possible pairs

Most sensitive: Test business rules here first.

Results

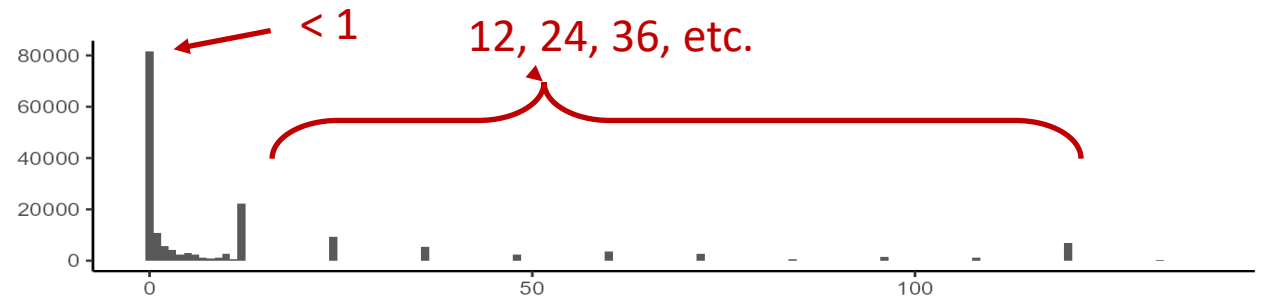
Exploratory Data Analysis of Sensitive Merges

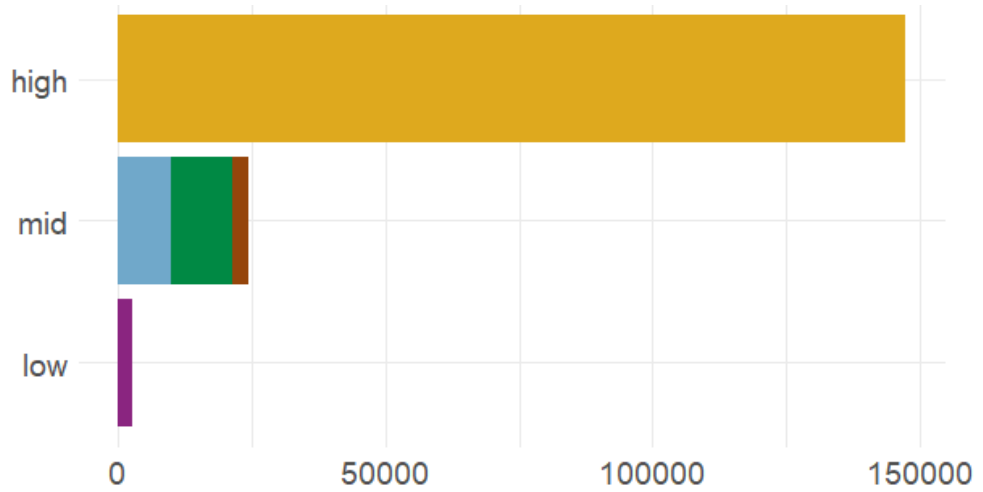
174,515 DOB conflicts

- # different digits: mean = 1.27
- Anagrams: 6.8%
- Similarity of other attributes

high (same sex/first/last)
mid (same first/last)
mid (same sex/first)
mid (same sex/last)
low (< 2 same values)

- Difference in months





Results

Propose Business Rules for Sensitive Merges

174,515 DOB conflicts

Guiding Principle

- Pairs with highly different DOBs are probably not the same person.
- Choosing the wrong DOB has worse consequences for younger people.
- Higher similarity pairs more likely to be the same person.

Business Rule

- BLOCK merges where DOBs are 3-8 digits different (unless anagram).
- Require smaller date differences for younger age groups.
- Test business rules on each similarity level separately.

Results

Propose Business Rules for Sensitive Merges

174,515 DOB conflicts

Blocker

Smaller date differences for younger groups.

```
%macro apply_br;
    set input;
    *blocker: NEVER merge when >2 digits differ (except for anagrams);
    if digit_diff > 2 and anagram_fl = 0 then merge_fl = 0;
        *ALWAYS merge older folks;
        else if age_grp = '90+' then merge_fl = 1;
        *MAYBE merge depending on age and DOB discrepancy;
        else if age_grp = '65-89' and (datediff_mo <= 120) then merge_fl = 1;
        else if age_grp = '18-64' and (datediff_mo <= 24) then merge_fl = 1;
        else if age_grp = '13-17' and (datediff_mo <= 12) then merge_fl = 1;
        else if age_grp = '0-12' and (datediff_mo <= 6) then merge_fl = 1;
    else merge_fl = 0;
mend;
```

Results

Evaluate Business Rules for Sensitive Merges

174,515 DOB conflicts

Q: How do we know this is the right implementation?

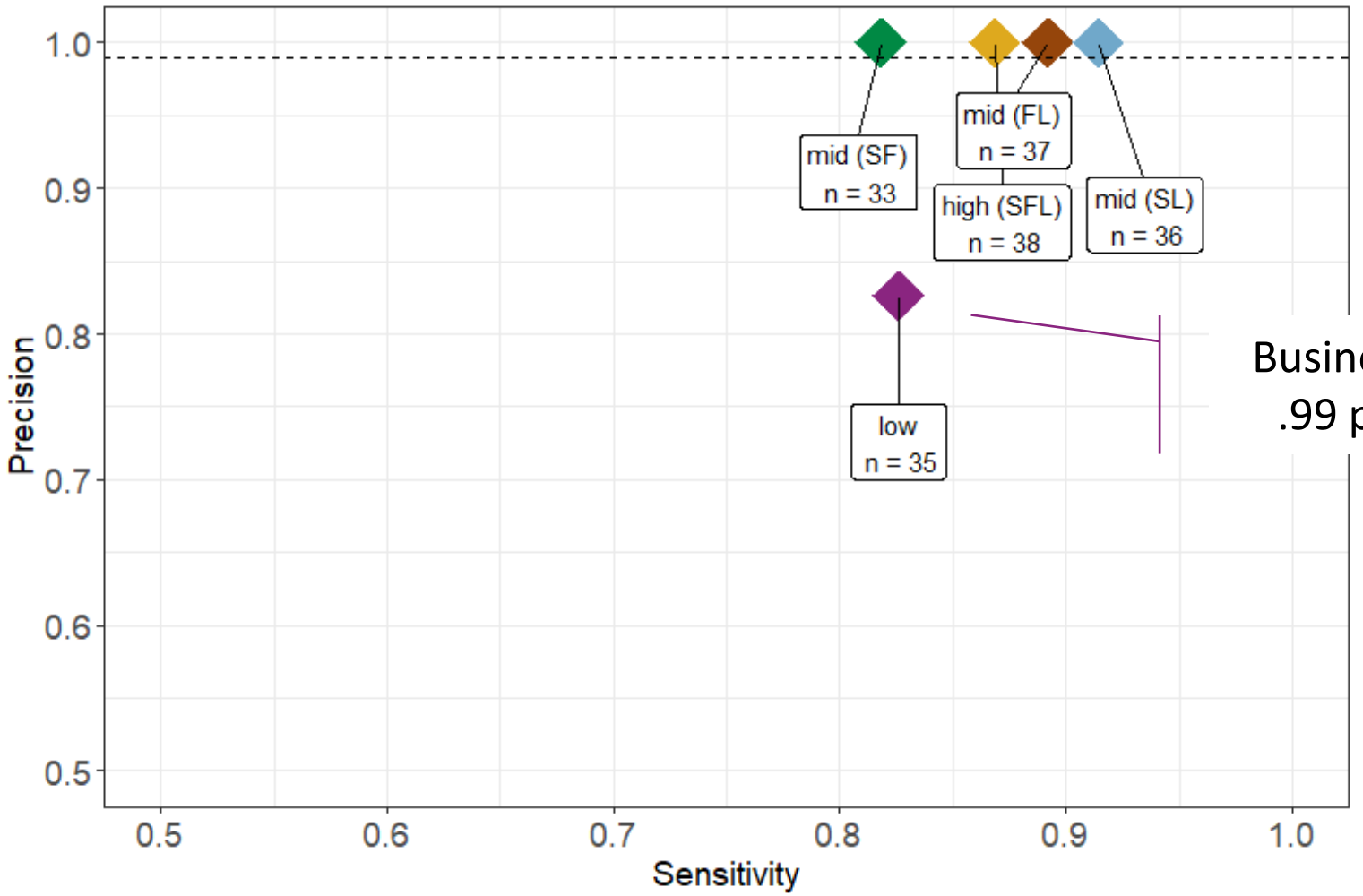
A: Test rules on a hand-checked **Random Sample** (N = 200)

- Stratified by similarity level (40 pairs each).
- Total human-validated matches (TP): 166
- Total human-validated non-matches (TN+FN): 13
- Ambiguous pairs excluded: 21

Results

Evaluate Business Rules for Sensitive Merges

174,515 DOB conflicts



Business rules fail to meet .99 precision threshold.

Results

Selectively Apply Business Rules

174,515 DOB conflicts

- Apply rules only when similarity level is mid to high.
- Total pairs approved for merging: 151,333 (86.7%)

Discussion

MDHHS
Michigan Department of Health & Human Services

Limitations

- Human review is not perfect (10.5% indeterminate).
- These rules determine merges—not which record is right.
  - Lexis Nexis often recommended values from both records.
- Input to Lexis Nexis included deceased persons, lowering the value-per-dollar.

Future directions

Business

- Develop business rules for handling approved merges.
- Develop business rules for remaining pairs.
  - 673,715 same DOB but other data conflicts.
  - Focus: preventing twin merges.
- Update addresses of singleton records.
- Develop rules for ingesting deceased flags.

Research

- Estimate overall denominator inflation.
- Investigate root causes.

Future directions

Business

- Develop business rules for handling approved merges.
- Develop business rules for remaining pairs.
  - 673,715 same DOB but other data conflicts.
  - Focus: preventing twin merges.
- Update addresses of singleton records.
- Develop rules for ingesting deceased flags.

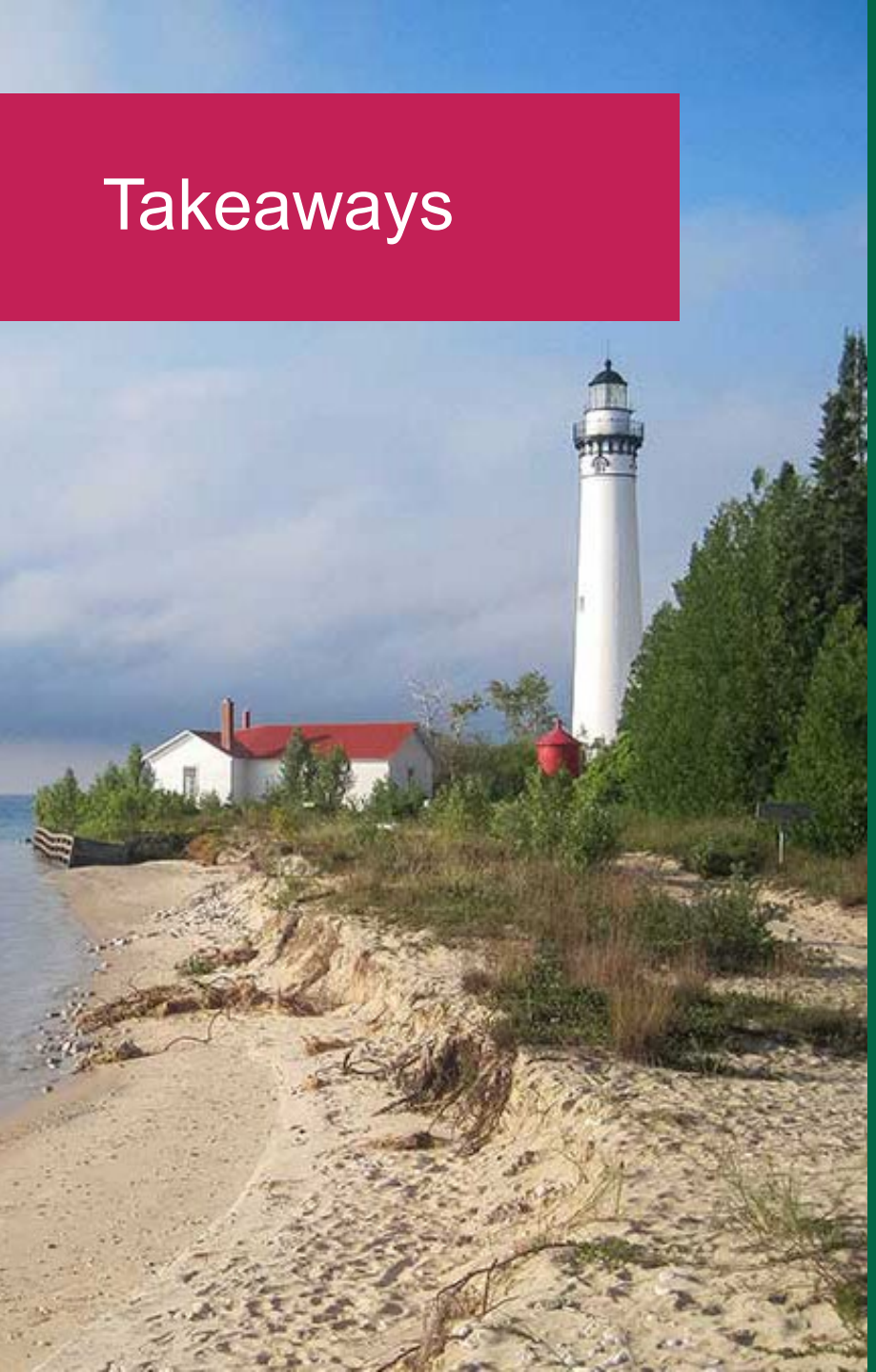
Research

- Estimate overall denominator inflation.
- Investigate root causes.

Lessons learned

Organizational barriers to business implementation

- IIS transition to new vendor.
  - → Shortened timeline.
  - → Legacy vendor winding down.
- Limited staff capacity.
- Institutional sluggishness.
  - → stale addresses.



Takeaways

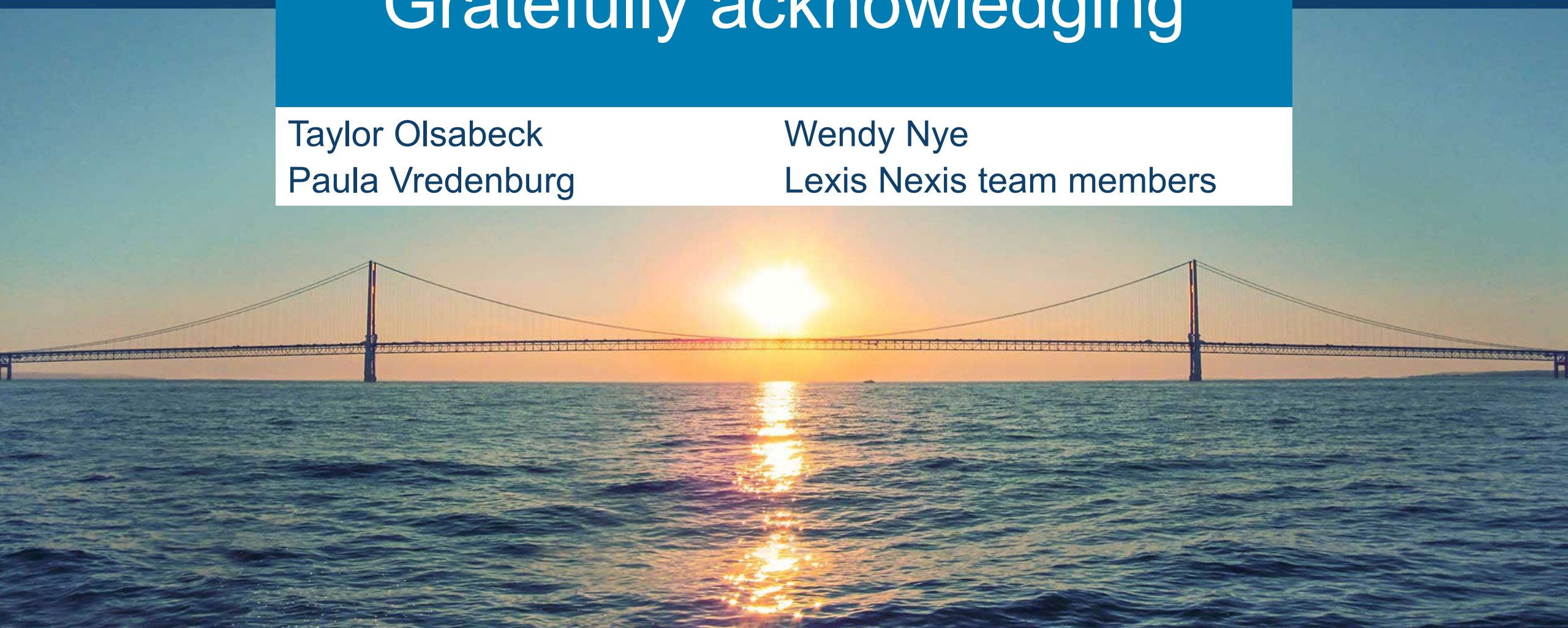
- No third-party dataset is magic.
  - Understand the limitations and their consequences for your IIS.
  - Decide your tolerance for error.
- Stratified random sampling can guide how you handle large, imperfect datasets.

Ask questions and plan your approach before buying a dataset.

MDHHS
Michigan Department of Health & Human Services

Gratefully acknowledging

Taylor Olsabeck
Paula Vredenburg
Wendy Nye
Lexis Nexis team members



Results

Code for Selective Application of Business Rules

174,515 possible pairs

```
%macro apply_br;
          set input;
          *blocker: NEVER merge when >2 digits differ (except for anagrams);
          if digit_diff > 2 and anagram_fl = 0 then merge_fl = 0;
                    *ALWAYS merge older folks;
                    else if age_grp = '90+' then merge_fl = 1;
                    *MAYBE merge when...;
                    else if age_grp = '65-89' and (datediff_mo <= 120) then merge_fl = 1;
                    else if age_grp = '18-64' and (datediff_mo <= 24) then merge_fl = 1;
                    else if age_grp = '13-17' and (datediff_mo <= 12) then merge_fl = 1;
                    else if age_grp = '0-12' and (datediff_mo <= 6) then merge_fl = 1;
          else merge_fl = 0;
mend;

data mydata_br;
          set mydata;
          if similarity_lvl in ('high (SFL)', 'mid (FL)', 'mid (SF)', 'mid (SL)') then
                    do;
                    %apply_br;
                    end;
          else merge_fl = .;
run;
```