How to Link Agency Client Records and Protect Privacy

Dr. Geoffrey Messier
Department of Electrical & Software Engineering
Schulich School of Engineering
University of Calgary

Land Acknowledgement

I would like to take this opportunity to acknowledge the traditional territories of the people of the Treaty 7 region in Southern Alberta, which includes the Blackfoot Confederacy (comprising the Siksika, Piikani, and Kainai First Nations), as well as the Tsuut'ina First Nation, and the Stoney Nakoda (including the Chiniki, Bearspaw, and Wesley First Nations). The City of Calgary is also home to Métis Nation of Alberta, Region 3.

Supporters and Partners

CALGARY
DROP-IN
CENTRE

trellis
SOCIETY

Calgary Homeless
FOUNDATION

MAKING
THE SHIFT INC

Calgary
Health
Foundation™

Alberta
Government

This study is based in part on data provided by Alberta Seniors, Community and Social Services. The interpretations and conclusions contained herein are those of the researchers and do not necessarily represent the views of the Government of Alberta. Neither the Government of Alberta nor Alberta Seniors, Community and Social Services express any opinion in relation to this study.

The Housing/Homelessness Data Ecosystem

UNIVERSITY OF CALGARY

- Data fragmentation.
  - Many agencies, incompatible IT systems, consent to share often not collected.
- No standard identifier.
  - People identified using names and birthdates.
- Despite these challenges, merging records across agencies has value:
  - **Individual Level:** Better care decisions, improved safety, avoids retelling traumatizing events.
  - **Population Level:** Seeing flow of people through a system of care highlights what programs do/don't work well, improves collaboration.

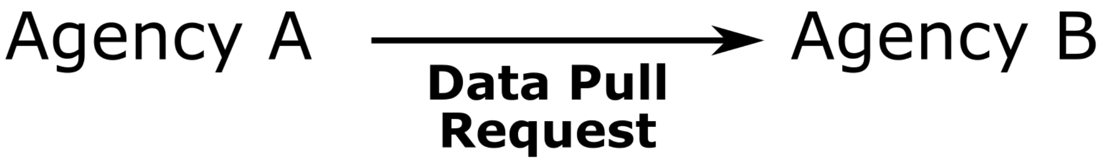
Privacy and Consent

- Default Assumption: No merging of identifying information without consent.
- In reality, the decision to share identifying information is a nuanced, case-by-case process.[1]
- Identity isn't required for population level analysis but names and birthdates are still required to link records.
- This must be done while protecting privacy.

[1] PolicyWise, "Ethical Decision Making Framework for Information Sharing"

Merging at the Individual Level

- Geoff accesses Agency A and gives consent to for staff to also use his data from Agency B.



Client Records (Agency A)

ID	Date	Name	Data
101	Jan 10	Geoff	…
102	Jan 10	Kermit	…
101	Jan 12	Geoff	…
103	Jan 14	Fozzy	…

⋮

Client Records (Agency B)

ID	Date	Name	Data
40	Jan 10	Beaker	…
30	Jan 11	Geoff	…
20	Jan 11	Piggy	…
50	Jan 15	Jeoff	…

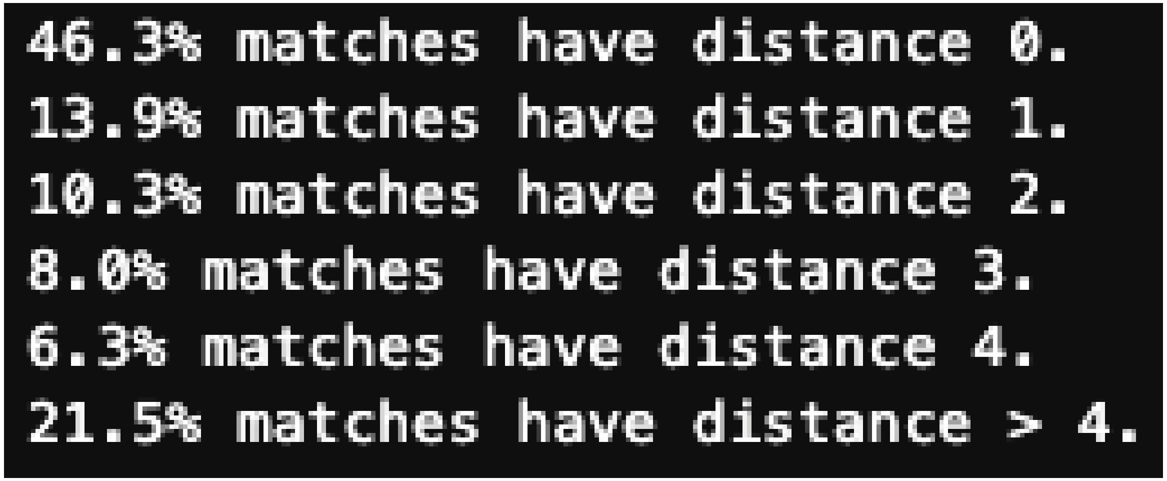
⋮

Messy Names

- A human can spot the similarity between “Geoff” and “Jeoff” but humans can’t scan thousands of records.
- How could a computer do it?
- Calculate a metric that measures word similarity:
  - Levenshtein distance (aka “edit distance”) counts the number of character edits to change one word to another.
  - These edits include insertions, deletions and substitutions.
  - Examples
    - “Geoff” ↔ “Jeoff”: Dist = 1 (1 substitution)
    - “Geoff” ↔ “Jeffrey”: Dist = 5 (1 substitution, 1 deletion, 3 insertions)

Edit Distance in Practice

- Calgary Homeless Foundation (CHF) staff manually matched 769 records from different agencies.

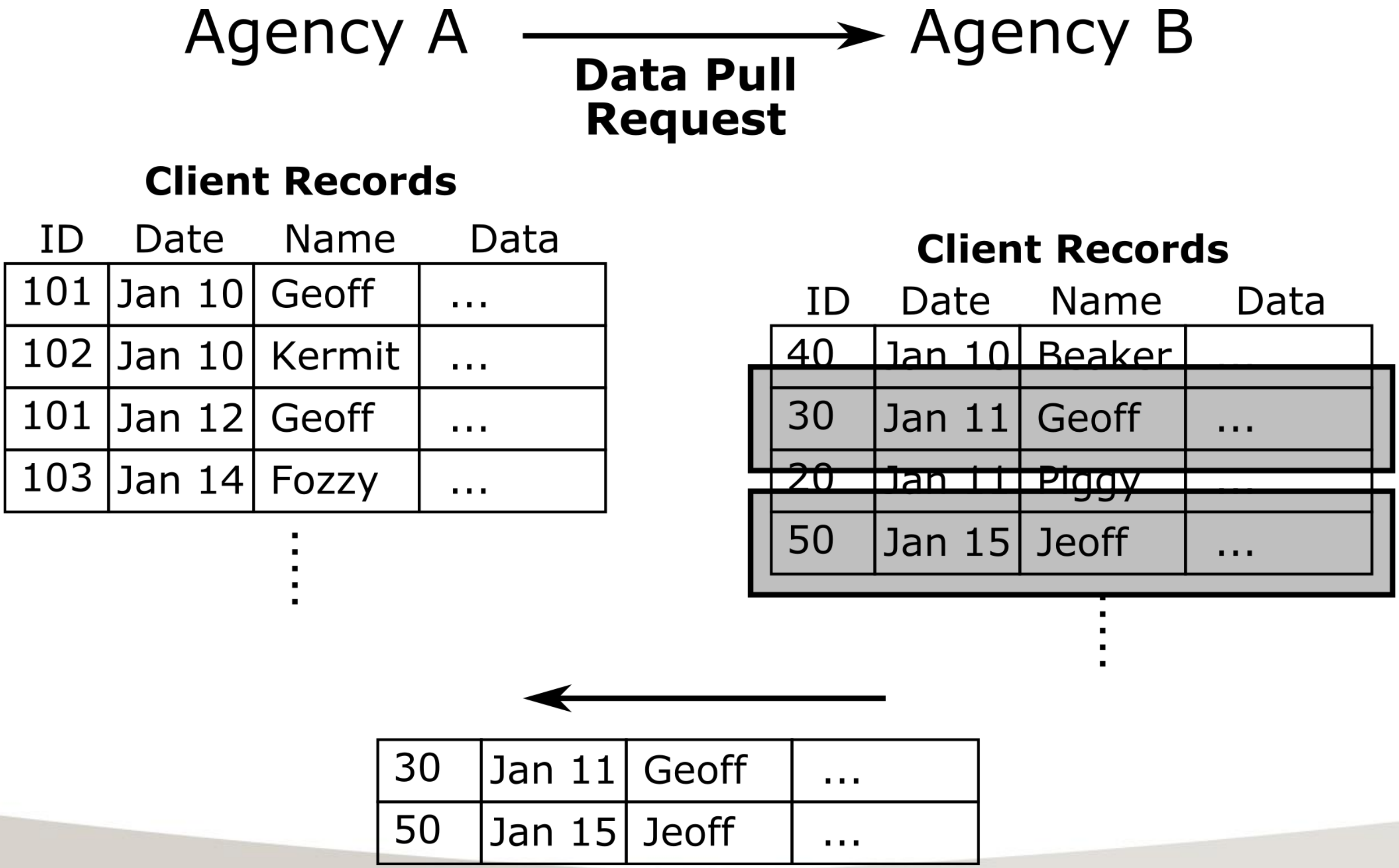


```
46.3% matches have distance 0.
13.9% matches have distance 1.
10.3% matches have distance 2.
8.0% matches have distance 3.
6.3% matches have distance 4.
21.5% matches have distance > 4.
```

- Anecdotally:
    - Low distance matches tended to be spelling/typing mistakes.
    - Higher distance matches were first/last name changes.

Merging at the Individual Level (Made Easier)

- Agency B does an edit distance search for all records similar to “Geoff”.



Agency A

Data Pull Request

Agency B

Client Records

ID	Date	Name	Data
101	Jan 10	Geoff	...
102	Jan 10	Kermit	...
101	Jan 12	Geoff	...
103	Jan 14	Fozzy	...

Client Records

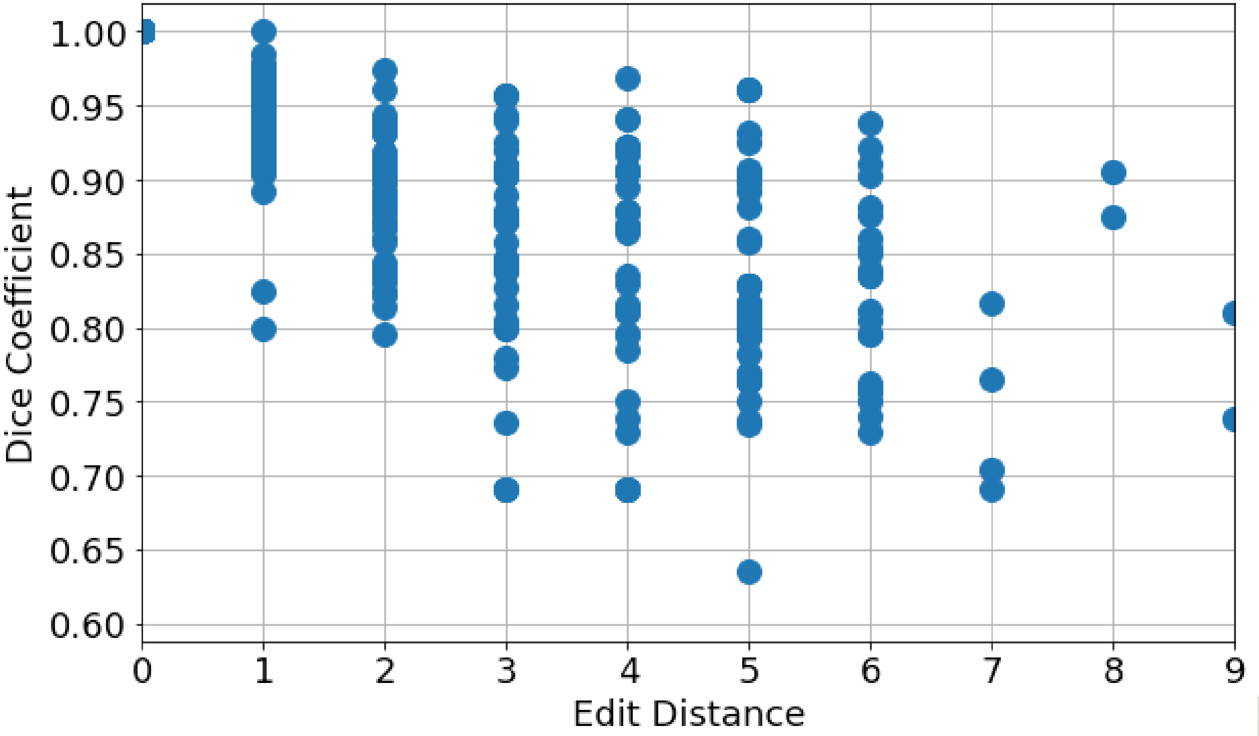
ID	Date	Name	Data
40	Jan 10	Beaker	...
30	Jan 11	Geoff	...
20	Jan 11	Piggy	...
50	Jan 15	Jeoff	...

30	Jan 11	Geoff	...
50	Jan 15	Jeoff	...

Merging at the Population Level

- If consent is not obtained, privacy must be preserved.
- Names and birthdates must be scrambled but we still want to spot "Geoff" and "Jeoff".
- Solution: Bloom Filter scrambling of names and DOBs.

The Dice coefficient measures the difference between words scrambled with Bloom filters.

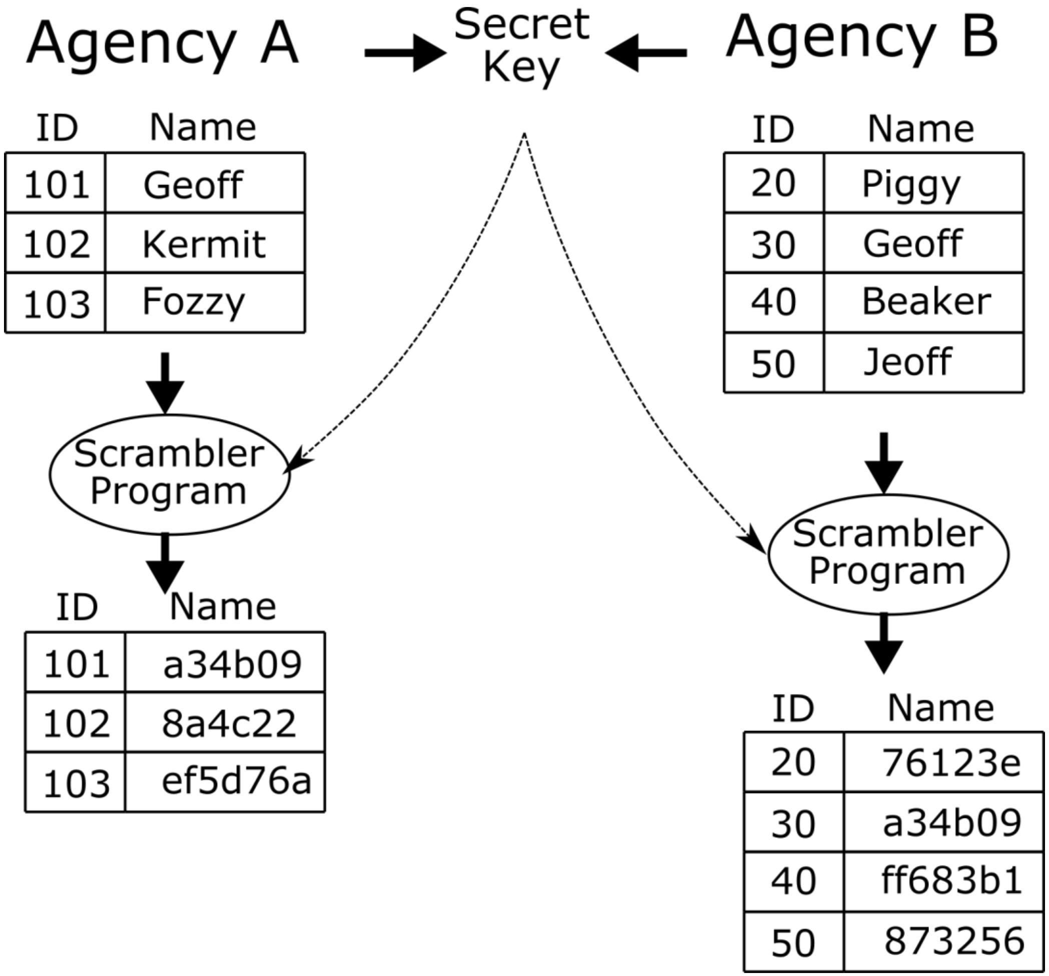


Merging Accuracy

- Declare a positive match between two records if the Dice coefficient is above a certain threshold.
- How well does this work?
  - Dice Threshold: 0.8232
  - Precision: 72% (28% of identified matches are incorrect)
  - Recall: 80% (20% of matches are missed)
- This is reasonable performance:
  - A certain amount of error won’t affect large system studies.
  - Machine learning can be used to improve performance.

Population Merging: Attempt #1

- Agencies A and B want to merge their entire data sets to examine how the same people access their services.
- They have access to a Bloom filter program and agree on a secret key.
- The agencies will now exchange scrambled data…



ID	Name
101	Geoff
102	Kermit
103	Fozzy

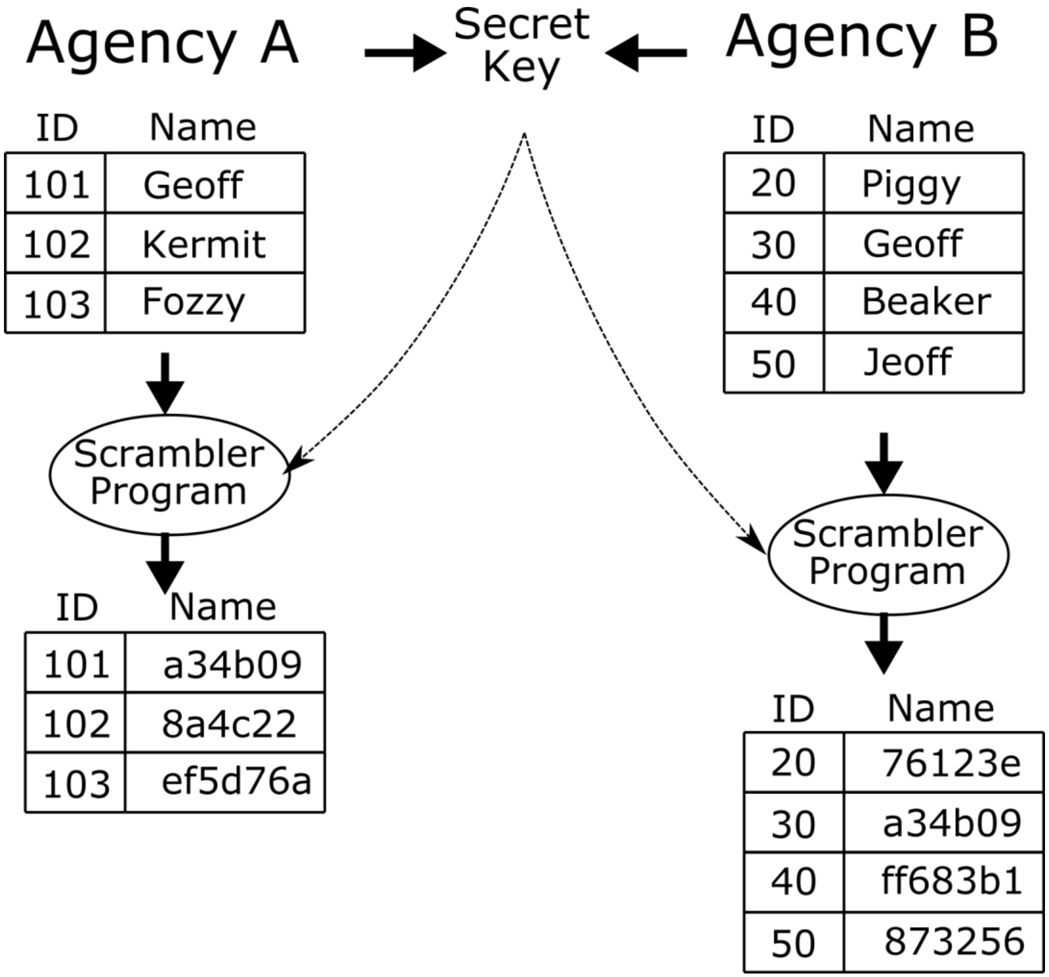
ID	Name
20	Piggy
30	Geoff
40	Beaker
50	Jeoff

ID	Name
101	a34b09
102	8a4c22
103	ef5d76a

ID	Name
20	76123e
30	a34b09
40	ff683b1
50	873256

STOP!!!

- Agency A knows the scrambled version of “Geoff”, “Kermit” and “Fozzy”.
- Any data from Agency B with those scrambled fields can be mapped back to real names.



ID	Name
101	Geoff
102	Kermit
103	Fozzy

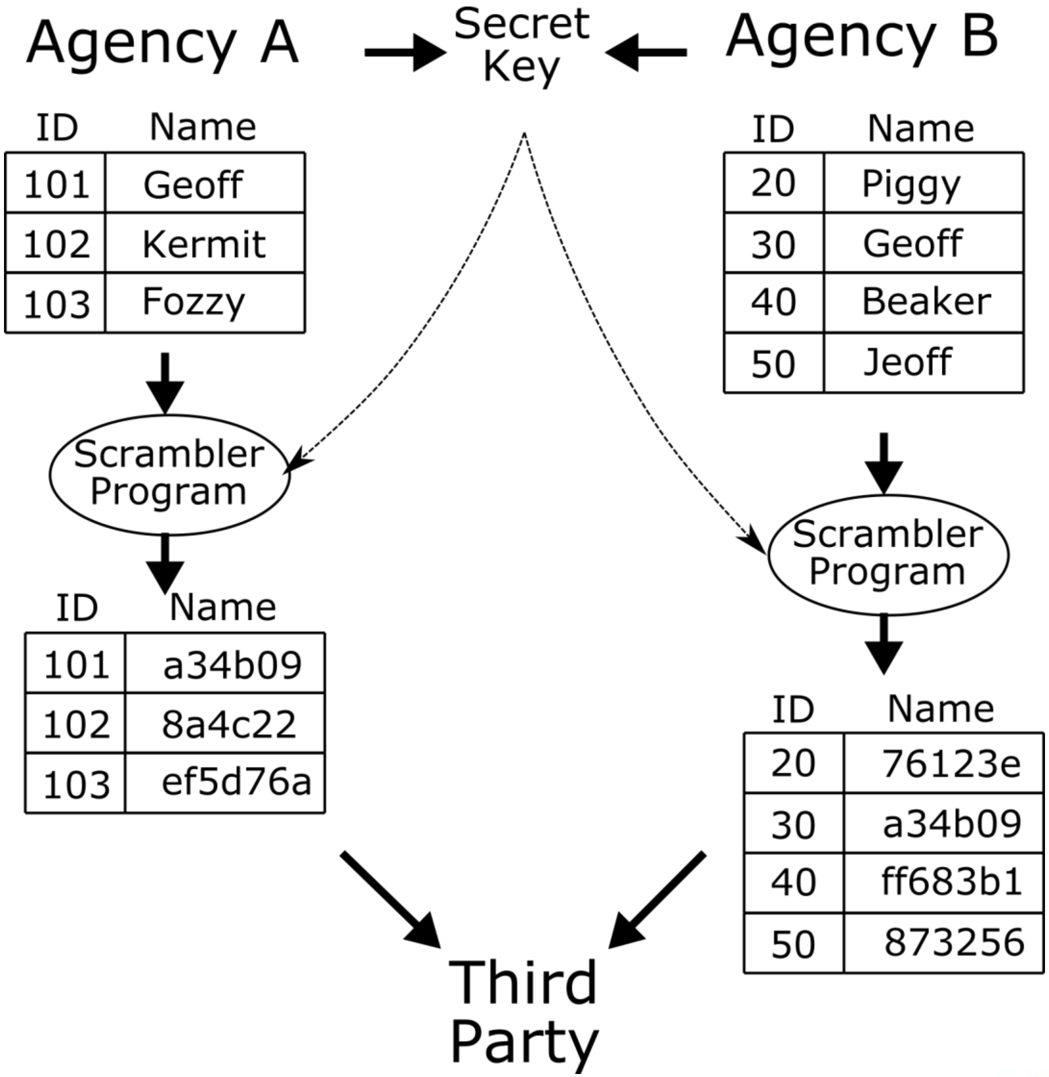
ID	Name
101	a34b09
102	8a4c22
103	ef5d76a

ID	Name
20	Piggy
30	Geoff
40	Beaker
50	Jeoff

ID	Name
20	76123e
30	a34b09
40	ff683b1
50	873256

Population Merging: Attempt #2

- The merging and system analysis must be done by a third party.
- Rules:
  1. The third party can’t have the secret key or unscrambled names.
  2. Agencies can’t have data from other agencies.



ID	Name
101	Geoff
102	Kermit
103	Fozzy

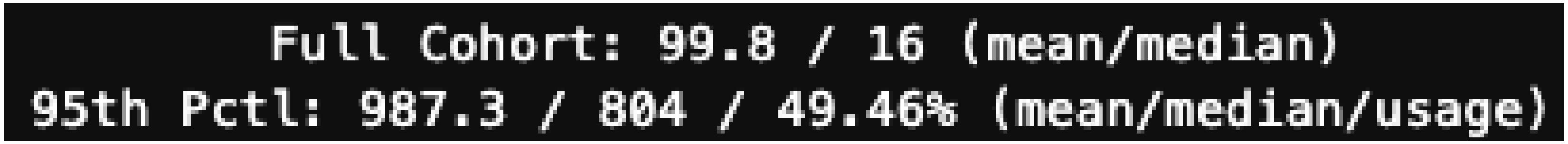
ID	Name
20	Piggy
30	Geoff
40	Beaker
50	Jeoff

ID	Name
101	a34b09
102	8a4c22
103	ef5d76a

ID	Name
20	76123e
30	a34b09
40	ff683b1
50	873256

Case Study: Calgary Housing/Homelessness System of Care Population Merge

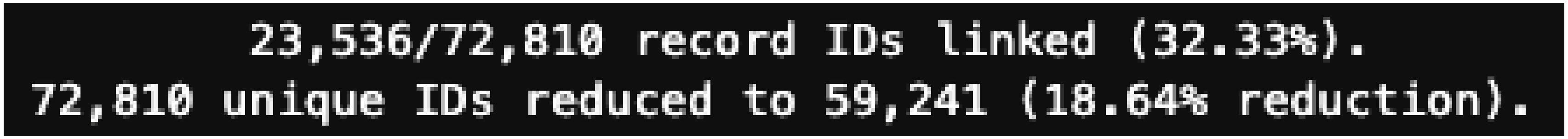
- Calgary Homeless Shelter Data Set:
  - Dates: Jan 1, 2009 – Dec. 31, 2019
  - Number of Shelters: 6
  - Number of Unique Client Shelter Records: 72,810
  - Names and birthdates Bloom filter scrambled by CHF.
- Record Stay Statistics (Before Merge):



```
     Full Cohort: 99.8 / 16 (mean/median)
95th Pctl: 987.3 / 804 / 49.46% (mean/median/usage)
```

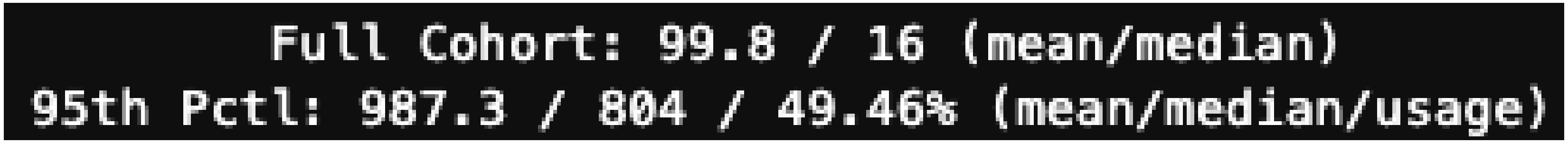
Case Study: Calgary Housing/Homelessness System of Care Population Merge

- Merge Results:



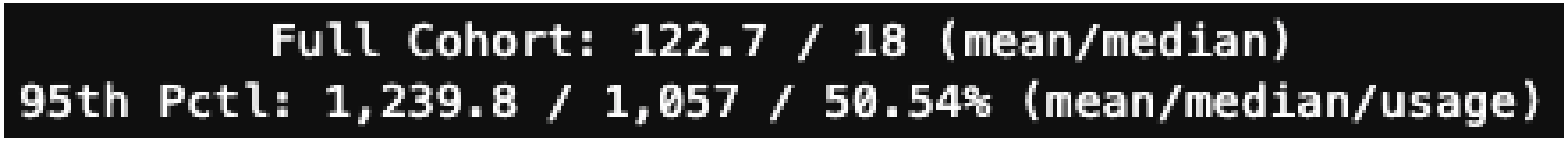
```
      23,536/72,810 record IDs linked (32.33%).
72,810 unique IDs reduced to 59,241 (18.64% reduction).
```

- Shelter Stay Statistics (Before Merge):



```
       Full Cohort: 99.8 / 16 (mean/median)
95th Pctl: 987.3 / 804 / 49.46% (mean/median/usage)
```

- Shelter Stay Statistics (After Merge):



```
         Full Cohort: 122.7 / 18 (mean/median)
95th Pctl: 1,239.8 / 1,057 / 50.54% (mean/median/usage)
```

Conclusions

- Records can be merged efficiently while still respecting privacy.
- This works with a variety of IT setups (a fully integrated database across agencies is **not** required).
- Chat with me if you're interested in trying this!

Geoff's Webpage & Contact Info: