**Department of Human Services: Arlington Data System Notes**

**M weights**

| System | Last name | First name | gender | ssn | street | zip | birth_yr | birth_mo | birth_dy |
|---|---|---|---|---|---|---|---|---|---|
| Anasazi | 0.95 | 0.9 | 0.85 | 0.5 | 0.4 | 0.6 | 0.95 | 0.95 | 0.85 |
| Eto | 0.95 | 0.9 | 0.85 | 0.9 | 0.4 | 0.6 | 0.95 | 0.95 | 0.85 |
| Web Vision | 0.95 | 0.9 | 0.85 | 0.9 | 0.4 | 0.6 | 0.96 | 0.95 | 0.85 |
| Combined | 0.95 | 0.9 | 0.85 | 0.5 | 0.4 | 0.6 | 0.95 | 0.95 | 0.85 |

| System | Starting n | # reduced | Final n | Weight cutoff | Notes/observations |
|---|---|---|---|---|---|
| **Anasazi/CERNER**<br>**Note to Sayali: make street into address** | 57065<br>56558 if you do unique | 691 | 56374 | 40 | Some interesting features:<br>- case_no = 0 were crazy<br>- 'COMBINED A INTO B' formats<br>- SSN m-weight set to 0.5 instead of 0.9 to account for missingness (~50%)<br>- NO matching IDs (in this case, ID is case number)<br>- This system has encounter data; so should integrate it to look for most recent demographics<br>- Weights:<br>  - Weight 50-52 has suspected twins of the same gender<br>  - Weight ~47 has suspected twins of opposite gender<br>  - Weights in between give funky matches! |
| **ETO** | 25619<br>25437 if you do unique | 460 | 25159 | -37 cut off point. Below 37, SSN matches are either missing or way off: Lower SSN? | Some interesting features<br>- Eto has merged record ID field that is completely missing<br>- Missing SSN is at 28%. Should we lower m weight to 0.7 instead of 0.9?<br>- 309 cases of system IDs mismatch whom are the same person |

| | | | | | |
|---|---|---|---|---|---|
| | | | | -For exact matched IDs:<br>142 system IDs duplicated across the system | - 142 system IDs duplicated across the system |
| **WVS** | 64927:<br>58115 if u do, unique: | 21595 | 43332 | - 50 for non-exact match IDs<br>- Between 40-50:<br>- Cases of twins? Sometimes birthdates are off!<br>- Exact matched IDs went down till 21 | - Is it possible that there are twins here as well? Look particularly at 30.33667<br>- Also, i think we should go down to 31 for non-exact-id-match sets as well |
| **Joint** | 147,611:<br>Systemwide reduction to:<br>124,868 | 4765 | 120103 | First cut: 33<br>Second cut: 34 | - As we improve the matches from each of the three systems, the initial number may change<br>- Again, plenty missing SSNs; so m-weight 0.5<br>- asz eto wvs<br>asz 1183 2816 696<br>eto 0 308 1888<br>wvs 0 0 1609<br>-<br>- Off-diagonal entries on the table above tell us how many across-system matches to expect. The current total is 5700<br>- After unique: 5,000 of them are shared across systems<br>- Around 33.68.., there's definitely a question about twins!<br>- Funnily enough, in this deduplication, the across system matches are significantly better than the within system ones. |

**Coding Steps for merging back**

Web Vision as a case study:

**Overall Process**

1) Group everything by same System ID: Merge the demographic variables (Cut off: 21-80)
2) Group everything by row ID: Get the different system id matches and merge the demographic variables (Cut off: 50-80)
3) Find the overlaps between #1 and #2. Merge the demo variables there.
4) #1 - overlap = Remaining 1
5) #2 - overlap = Remaining 2
6) dedup_total = overlap + Remaining 1 + Remaining 2
7) Undedup_total = All data set - (dedup_total + linked system Id pairs from 2)
8) Final data set = dedup_total + Undedup_total

Logic that could improve
1) Where majority rule fails (Pls refer to email over weekend)
2) Cases where : (27, 28,29): (29,31) where 31 is not considered. Code is not deliberately written to take care of it.

**Remaining Tasks**

For Final Joined Table:

- Complete step 7 and step 8

Create a systems ID pairs table