Record Linkage

- Identifying multiple records that refer to the same person.
- Purposes:
  - Build more complete and concise picture of individual
  - Avoid duplication of ordinances
- Use names, dates, places, relatives, and other data to decide.
Limitations of exact matching

• Non-overlapping data
  – Alex Gray, b. 2 Jan 1802, VA; son of William Gray & Mary Turner
  – Alex Gray, m. 19 Aug 1830 to Susannah Robinhold.

• Data variation
  – Alexander Grey, b. about 1805, Virg.; Son of Bill & Polly Grey.
Name Variations

• Nicknames *(Margaret/Peggy, Mary/Polly)*
• Transcription or typographical errors *(James/Jarnes, Alexander/Alexadner)*
• Abbreviations *(William/Wm./W.)*
• Translation/immigration name changes *(Schmidt/Smith, Müller/Mueller/Miller)*
• Same-sounding spelling variations *(Barns/Barnes)*
• Minor changes to names over time *(Speak/Speake/Speaks/Speakes)*
Name Standardization
*Bringing together similar names*

- Name Encoding Algorithms
  - Soundex
  - NYSIIS
  - Metaphone/Double Metaphone
- Name Catalogs
- Name comparison functions
  - Edit Distance
  - Jaro-Winkler
Soundex (1918)

First letter + 3 digits. Drop vowels (+w,h,y),
combine double letters, map letters to digits:

1 b,f,p,v
2 c,g,j,k,q,s,x,z
3 d,t
4 l
5 m, n
6 r

Miller = M460
Mueller = M460
NYSIIS (1970)

1) Translate first characters of name:
   MAC => MCC, KN => NN, K => C, PH => FF, PF => FF, SCH => SSS

2) Translate last characters of name:
   EE  => Y; IE  => Y; DT,RT,RD,NT,ND => D

3) First character of key = first character of name.

4) Translate remaining characters by following rules, incrementing by one character each time:
   a. EV => AF else A,E,I,O,U => A
   b. Q => G, Z => S, M => N
   c. KN => N, else K => C
   d. SCH => SSS, PH => FF
   e. H => If previous or next is non-vowel, previous
   f. W => If previous is vowel, previous
      Add current to key if current ≠ last key character

5) If last character is S, remove it
6) If last characters are AY, replace with Y
7) If last character is A, remove it
Metaphone, Double Metaphone

• Map letters to 16 consonants
  – Bender => BNTR

• Double Metaphone has primary + “alternate” encoding for some names
  – Schneider => XNTR, SNTR
  – Thomas => TMS
Name Catalogs

- ODM (Ordinance Data Management) catalog
  - Developed since about 1969
  - 20 regional catalogs (North America, British Isles, Norway, Central America, etc.)
- Manually built, largely as needed
  - Maggie, Peggy, Margret => MARGARET
- Can map same name to different standards
  - John => JOHAN (Germany), John=>JOHN (NA)
Catalog Variants

- **“Universal”** catalog
  - All regions in one catalog
  - “Bucket IDs” instead of standard spellings
  - Spelling can appear in multiple “buckets”

- Cultural catalog (region-specific bucket IDs)
  - *Default* culture (North America catalog)
  - Culture based on *person* events
  - Culture based on person’s and *relatives* events

- **Edit Distance** catalog
  - All names in database within edit distance of 0.95.
Labeled Data

- 178,880 individuals in sample database
- About 25,000 pairs identified as matches
- Build Lucene index using each name standardization method
- Issue query using each method
  - `given:john given:alan`
  - `surname:gray`
  - `soundex_given:J250 soundex_given:A450`
Recall vs. “Cost”

- **Recall**: % of known matches that are “brought together” by a given standardization technique.

- **Cost**: Average number of “hits” per individual in queries using given standardization technique.
Cost/Recall example

- **Recall:**
  - 85% of matched pairs had an original surname in common
  - 89% of matched pairs had a Soundex surname in common

- **Cost:**
  - Avg. of 61 people (from 178,880) had same surname as each individual.
  - Avg. of 261 people had same Soundex surname

- So Soundex has “better” recall but “worse” cost, because it casts a broader net.
Given Name Cost vs. Recall

Cost (number of hits)

- **Others**
- **Winners**
Given Name Cost vs. Recall

- **Cost (number of hits)**
- **Recall**

- **Others**
- **Winners**
<table>
<thead>
<tr>
<th>Given Name Fields</th>
<th>Recall</th>
<th>AvgHits</th>
<th>% of Db</th>
</tr>
</thead>
<tbody>
<tr>
<td>Universal + Orig</td>
<td>99.08</td>
<td>9689</td>
<td>5.42%</td>
</tr>
<tr>
<td>Universal</td>
<td>98.67</td>
<td>9689</td>
<td>5.42%</td>
</tr>
<tr>
<td>ODM + Orig</td>
<td>98.62</td>
<td>3771</td>
<td>2.11%</td>
</tr>
<tr>
<td>Soundex</td>
<td>98.31</td>
<td>5761</td>
<td>3.22%</td>
</tr>
<tr>
<td>Culture_default + Orig</td>
<td>98.16</td>
<td>4712</td>
<td>2.63%</td>
</tr>
<tr>
<td>Double Metaphone</td>
<td>98.09</td>
<td>6595</td>
<td>3.69%</td>
</tr>
<tr>
<td>Culture_relative + Orig</td>
<td>97.81</td>
<td>2292</td>
<td>1.28%</td>
</tr>
<tr>
<td>ODM</td>
<td>97.72</td>
<td>3620</td>
<td>2.02%</td>
</tr>
<tr>
<td>Metaphone</td>
<td>97.65</td>
<td>4771</td>
<td>2.67%</td>
</tr>
<tr>
<td>Culture_person + Orig</td>
<td>97.57</td>
<td>2828</td>
<td>1.58%</td>
</tr>
<tr>
<td>Edit</td>
<td>97.40</td>
<td>3280</td>
<td>1.83%</td>
</tr>
<tr>
<td>NYSIIS</td>
<td>97.21</td>
<td>5847</td>
<td>3.27%</td>
</tr>
<tr>
<td>Culture_default</td>
<td>96.96</td>
<td>4712</td>
<td>2.63%</td>
</tr>
<tr>
<td>Orig</td>
<td>94.32</td>
<td>1895</td>
<td>1.06%</td>
</tr>
<tr>
<td>Culture_relative</td>
<td>90.30</td>
<td>1191</td>
<td>0.67%</td>
</tr>
<tr>
<td>Culture_person</td>
<td>83.11</td>
<td>1875</td>
<td>1.05%</td>
</tr>
<tr>
<td>Surname Fields</td>
<td>Recall</td>
<td>AvgHits</td>
<td>% of Db</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>ODM + Orig</td>
<td>93.41</td>
<td>99.9</td>
<td>0.06%</td>
</tr>
<tr>
<td>ODM</td>
<td>92.92</td>
<td>94.8</td>
<td>0.05%</td>
</tr>
<tr>
<td>Universal + Orig</td>
<td>89.39</td>
<td>332.4</td>
<td>0.19%</td>
</tr>
<tr>
<td>Double Metaphone</td>
<td>89.24</td>
<td>264.4</td>
<td>0.15%</td>
</tr>
<tr>
<td>Soundex</td>
<td>89.22</td>
<td>260.5</td>
<td>0.15%</td>
</tr>
<tr>
<td>Culture_relative + Orig</td>
<td>88.79</td>
<td>75.1</td>
<td>0.04%</td>
</tr>
<tr>
<td>NYSIIS</td>
<td>88.57</td>
<td>150.3</td>
<td>0.08%</td>
</tr>
<tr>
<td>Metaphone</td>
<td>88.35</td>
<td>181.0</td>
<td>0.10%</td>
</tr>
<tr>
<td>Culture_person + Orig</td>
<td>88.02</td>
<td>72.2</td>
<td>0.04%</td>
</tr>
<tr>
<td>Culture_default + Orig</td>
<td>87.79</td>
<td>95.5</td>
<td>0.05%</td>
</tr>
<tr>
<td>Edit</td>
<td>87.59</td>
<td>123.2</td>
<td>0.07%</td>
</tr>
<tr>
<td>Universal</td>
<td>86.28</td>
<td>332.1</td>
<td>0.19%</td>
</tr>
<tr>
<td>Orig</td>
<td>84.62</td>
<td>61.1</td>
<td>0.03%</td>
</tr>
<tr>
<td>Culture_default</td>
<td>79.58</td>
<td>94.7</td>
<td>0.05%</td>
</tr>
<tr>
<td>Culture_relative</td>
<td>73.50</td>
<td>53.1</td>
<td>0.03%</td>
</tr>
<tr>
<td>Culture_person</td>
<td>67.60</td>
<td>48.3</td>
<td>0.03%</td>
</tr>
<tr>
<td>Given + Surname Fields</td>
<td>Recall</td>
<td>AvgHits</td>
<td>% of Db</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>--------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>ODM + Orig</td>
<td>99.68</td>
<td>3850</td>
<td>2.15%</td>
</tr>
<tr>
<td>Soundex</td>
<td>99.54</td>
<td>5998</td>
<td>3.35%</td>
</tr>
<tr>
<td>Universal + Orig</td>
<td>99.42</td>
<td>9990</td>
<td>5.58%</td>
</tr>
<tr>
<td>NYSIIS</td>
<td>99.41</td>
<td>5976</td>
<td>3.34%</td>
</tr>
<tr>
<td>Culture_relative + Orig</td>
<td>99.35</td>
<td>2348</td>
<td>1.31%</td>
</tr>
<tr>
<td>Culture_person + Orig</td>
<td>99.25</td>
<td>2882</td>
<td>1.61%</td>
</tr>
<tr>
<td>Metaphone</td>
<td>99.20</td>
<td>4931</td>
<td>2.76%</td>
</tr>
<tr>
<td>Double Metaphone</td>
<td>99.20</td>
<td>6835</td>
<td>3.82%</td>
</tr>
<tr>
<td>Culture_default + Orig</td>
<td>99.16</td>
<td>4788</td>
<td>2.68%</td>
</tr>
<tr>
<td>Orig + Swap</td>
<td>98.53</td>
<td>2135</td>
<td>1.19%</td>
</tr>
<tr>
<td>Orig</td>
<td>98.00</td>
<td>1939</td>
<td>1.08%</td>
</tr>
<tr>
<td>Edit + Orig</td>
<td>98.00</td>
<td>1939</td>
<td>1.08%</td>
</tr>
</tbody>
</table>
Overall Improvement

ODM+Orig:

• Given: 94.32 to 98.62 => 75% reduction in misses.
• Surname: 84.62% to 93.41% => 57% reduction in misses.
• Combined: 98% to 99.68% => 84% reduction in misses.

at a cost of about twice as many hits.
Conclusions

• Standardization significantly improves recall.
• Catalog-based methods gave better recall at lower number of hits than algorithmic methods (except “universal”)
• Using culture (and using relatives to help select culture) improved accuracy of catalogs.
• Still, algorithmic methods like Soundex had reasonable recall and are inexpensive to implement.