Genealogical Record Linkage: Features for Automated Person Matching

Randy Wilson
wilsonr@familysearch.org
Record Linkage definition

• **Record linkage** is the process of identifying multiple records that refer to the same thing (in our case, the same real-world person).

• **Blocking**: Finding potentially matching records.

• **Scoring**: Evaluating potentially matching records to see how likely they are to match.
Reasons for identifying matches

- Identify duplication individuals
- Find additional source information on a person.
- Build more complete picture of individuals and families
- Avoid duplicate research efforts
<table>
<thead>
<tr>
<th>Person ID</th>
<th>11157365495</th>
<th>1100911109</th>
</tr>
</thead>
</table>
| Name        | [1] Jakob Balli  
 Jacob Balli | [2] JAKOB BALLI  
 Jakob Balli  
 Jakob Balli junr  
 Jacob Balli |
|             | 11157365495 [hasOrd] | 1100911109 [hasOrd] |
| Birth       | 12 Mar 1809  
 12 Mar 1809  
 12 Mar 1809  
 Matten Interlaken, Bern, Switzerland  
 Matten, Bern, Switz.  
 Matten, Bern, Switzerland | 18330118  
 16Jun1833  
 18 Jan 1833  
 18 Jan 1833  
 18 Jan 1833  
 Matten, Bern, Switzerland  
 Matten, Bern, Switzerland  
 Matten, Bern, Switzerland  
 Matten, Bern, Switzerland  
 Interlaken, Bern, Switz.  
 Matten, Bern, Switzerland |
| Death       | 6 May 1848  
 6 May 1848  
 Matten, Bern, Switz. | 19130825  
 26Aug1913  
 26Aug1913  
 26 Aug 1913  
 26 Aug 1913  
 Salt Lake City, Utah  
 Salt Lake City, Utah  
 Salt Lake City, Utah  
 Salt Lake City, Utah  
 Salt Lake City, Utah |
| Burial      | 29Aug1913  
 29 Aug 1913  
 Salt Lake City, Utah  
 Salt Lake City, Utah  
 Salt Lake City, Utah | 18670308  
 1867  
 8Mar1867  
 8 Mar 1867  
 Matten, Bern, Switzerland  
 Matten, Bern, Switzerland  
 Matten, Bern, Switzerland |
| Marriage    | 3 Johanna Caspar Balli  
 Johannes Caspar Balli  
 Johann Caspar Balli | 11157365495 [hasOrd] |
 Johann Caspar Balli  
 Johann Caspar Balli | [1] Jakob Balli  
 Jacob Balli |
Measuring accuracy

- **Precision**
  - Percent of a system’s matches that are correct.
    \[=\text{correct match} / (\text{correct match} + \text{false match})\]

- **Recall**
  - Percent of available matches that the system finds.
    \[=\text{correct match} / (\text{correct match} + \text{false differ})\]
P/R Example

<table>
<thead>
<tr>
<th></th>
<th>True Match</th>
<th>True Differ</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output Match</td>
<td>90</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Output Differ</td>
<td>30</td>
<td>290</td>
<td>320</td>
</tr>
<tr>
<td>Total</td>
<td>120</td>
<td>300</td>
<td>420</td>
</tr>
</tbody>
</table>

- Recall = True Matches/Total Matches = 90/120 = 75%
- Precision = True Matches/Output Matches = 90/100 = 90%
- (Missed match rate = 25% = false negative rate)
- (False match rate = 10% = false positive rate)
More P/R definitions

Pick whichever definition makes the most sense to you.

- **Precision**: 
  Percent of matches that a system comes up with that are correct.
  \[
  = 100\% \times \frac{\text{#correct matches}}{\text{#correct matches} + \text{#incorrect matches}}
  
  = 100\% \times \frac{\text{#correct matches}}{\text{total #matches found}}
  
  = 100\% - \text{(Percent of matches that a system comes up with that are wrong)}
  
  = 100\% - \text{(false match rate)}
  
- **Recall**: 
  Percent of true matches in the data that the system comes up with.
  \[
  = 100\% \times \frac{\text{#correct matches found}}{\text{#correct matches found} + \text{#correct matches not found}}
  
  = 100\% \times \frac{\text{#correct matches found}}{\text{#matches available in the data}}
  
  = 100\% - \text{(Percent of matches that the system failed to find)}
Histogram: P/R Trade-off
P/R Curves and Thresholds
Better precision => worse recall, and vice-versa
Improving the trade-off
Example: Learning algorithm
Areas of improvement

- Better training data
  - More data
  - More representative of target usage
- Better learning algorithm
  - Neural networks, machine learning
- Better blocking
  - Multiple blocking passes to get highest recall with fewest total hits.
- Better features
Matching in New FamilySearch

- Select random individuals
- Do [Lucene] query to find potential matches
- Select pairs across score range
- Show pairs to experts for labeling
- Audit labels, especially outliers
- Develop matching features
- Train feature weights using neural networks
- Pick thresholds with least objectionable P/R
## Thresholds for star ratings

The first record listed below is the record from your family tree:

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Birth or Christening</th>
<th>Death or Burial</th>
<th>Spouse</th>
<th>Parents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Drusilla Dorris</td>
<td>20 May 1881</td>
<td>James Hendricks</td>
<td>William Dorris</td>
<td>Catherine Frost</td>
</tr>
</tbody>
</table>

The records listed below are the possible duplicates:

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Birth or Christening</th>
<th>Death or Burial</th>
<th>Spouse</th>
<th>Parents</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Drusilla Dorris</td>
<td>about 1813</td>
<td>James Hendricks</td>
<td>William Dorris</td>
<td>Catherine Frost</td>
</tr>
<tr>
<td>3</td>
<td>Drusilla Dorris</td>
<td>1805</td>
<td>James Hendricks</td>
<td>William Dorris</td>
<td>Catherine Frost</td>
</tr>
<tr>
<td>4</td>
<td>Drusilla Dorris</td>
<td>about 1803</td>
<td>James Hendricks</td>
<td>William Dorris</td>
<td>Catherine Frost</td>
</tr>
<tr>
<td>5</td>
<td>Rebecca Dorris</td>
<td>22 February 1793</td>
<td>James Hendricks</td>
<td>William Dorris</td>
<td>Catherine Frost</td>
</tr>
<tr>
<td>6</td>
<td>Rebecca Dorris</td>
<td>about 1796</td>
<td>James Hendricks</td>
<td>William Dorris</td>
<td>Catherine Frost</td>
</tr>
<tr>
<td>7</td>
<td>Tabitha Dorris</td>
<td>12 January 1804</td>
<td>James Hendricks</td>
<td>William Dorris</td>
<td>Catherine Frost</td>
</tr>
</tbody>
</table>
Matching Features

• How well does given name agree?
• How well does surname agree?
• Birth date? Birth place?
• Marriage/death/burial?
• Father/Mother/Spouse names?
Person-matching Features

- **Features**
  - Names
  - Dates
  - Places
  - Misc
- **Feature values**
  - Levels of feature agreement
- **Weights**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IndGivenName=-1</td>
<td>-2.2224</td>
</tr>
<tr>
<td>IndGivenName=1</td>
<td>0.5968</td>
</tr>
<tr>
<td>IndGivenName=2</td>
<td>0.687</td>
</tr>
<tr>
<td>IndGivenName=3</td>
<td>0.0743</td>
</tr>
<tr>
<td>IndGivenName=4</td>
<td>1.5611</td>
</tr>
<tr>
<td>IndGivenName=5</td>
<td>0.686</td>
</tr>
<tr>
<td>IndGivenName=6</td>
<td>0.4946</td>
</tr>
<tr>
<td>IndGivenName=7</td>
<td>1.2099</td>
</tr>
<tr>
<td>IndCommonGivenName=1</td>
<td>1.0244</td>
</tr>
<tr>
<td>IndCommonGivenName=2</td>
<td>1.0773</td>
</tr>
<tr>
<td>IndCommonGivenName=3</td>
<td>1.1974</td>
</tr>
<tr>
<td>IndCommonGivenName=4</td>
<td>1.4942</td>
</tr>
<tr>
<td>IndSurname=-1</td>
<td>-1.8169</td>
</tr>
<tr>
<td>IndSurname=1</td>
<td>1.4038</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-5.0982</td>
</tr>
</tbody>
</table>
Names: Name variations

- Upper/lower case. ("MARY", "Mary", "mary")
- Maiden vs. married name. ("Mary Turner"/"Mary Jacobs").
- Husband’s name ("Mrs. John Smith" / "Mary Turner")
- Nicknames. ( "Mary"/"Polly"; "Sarah"/"Sally"; "Margaret"/"Peggy")
- Spelling variations ("Elizabeth" vs. "Elisabeth"; "Speak"/"Speake"/"Speaks"/"Speakes")
- Initials ("John H. Smith" / "John Henry Smith")
- Abbreviations ("Wm."/"William", "Jas"/"James")
- Cultural changes (e.g., "Schmidt" -> "Smith").
- Typographical errors ("John Smith"/"John Smiht")
- Illegible handwriting (e.g., "Daniel" and "David").
More name variations

- **Spacing** ("McDonald"/ "Mc Donald")
- **Articles** ("de la Cruz" / "Cruz")
- **Diacritics** ("Magaña", "Magana")
- **Script changes** (e.g., “津村”, “タカハシ”, “Takahashi”).
- **Name order variations.** ("John Henry", “Henry John”).
- **Given/surname swapped.** (Kim Jeong-Su, Jeong-Su Kim)
- **Multiple surnames** (e.g., “Juanita Martinez y Gonzales”)
- **Patronymic naming.** (“Lars Johansen, son of Johan Svensen”, “Lars Svensen”).
- **Patriarchal naming.** (e.g., “Fahat Yogol”, “Fahat Yogol Maxmud”, “Fahat Maxmud”)

Names: Normalization

• Remove punctuation:
  Mary “Polly” ➔ mary polly
• Convert diacritics (Magaña ➔ magana)
• Lower case
• Remove prefix/suffix (Mr., Sr., etc.)
• Separate given and surname pieces
Names: Comparing pieces

• Name piece agreement:
  – Exact ("john", "john")
  – Near: Jaro-Winkler > 0.92 ("john", "johan")
  – Far:
    • Jaro-Winkler > 0.84
    • One “starts with” the other ("eliza", "elizabeth")
    • Initial match ("e", "e")
  – Differ: ("john", "henry")

<table>
<thead>
<tr>
<th></th>
<th>john</th>
<th>henry</th>
</tr>
</thead>
<tbody>
<tr>
<td>johan</td>
<td>Near</td>
<td>Differ</td>
</tr>
<tr>
<td>h</td>
<td>Differ</td>
<td>Far</td>
</tr>
</tbody>
</table>
## Names: Piece alignment

<table>
<thead>
<tr>
<th></th>
<th>john</th>
<th>henry</th>
</tr>
</thead>
<tbody>
<tr>
<td>johan</td>
<td>Near</td>
<td>Differ</td>
</tr>
<tr>
<td>h</td>
<td>Differ</td>
<td>Far</td>
</tr>
</tbody>
</table>

- **johan** and **john** are **Near**.
- **h** and **henry** are **Far**.

### Missing Value

- **johan** and **<none>** are **Near**.
- **<none>** and **henry** are **<Missing>**.
Full name agreement levels

7: One “exact” name piece agreement, and at least one more piece that is exact or at least near. No “missing” pieces.
6: One “exact” name piece agreement, and at least one more piece that is exact or at least near. At least one “missing” piece.
5: One “exact”, no “missing”.
4: At least one “near”, no “missing”.
3: One “exact”, at least one “missing”.
2: At least one “far”; no “missing”
1: At least one “far” or “near”; at least one “missing”
0: No data: At least one name has no name at all.
-1: Conflict: At least one “differ”
Name frequency (odds)

• Given names
  1: Odds $\leq 40$ (very common: John is 1 in 25)
  2: $40 < $Odds $\leq 300$
  3: $300 < $Odds $\leq 1500$
  4: $\text{Odds} > 1500$ (rare: name not in the list)

• Surnames
  1: $\text{Odds} \leq 4000$ (common)
  2: $4000 < $Odds $\leq 10,000$
  3: $10,000 < $Odds $\leq 100,000$
  4: $\text{Odds} > 100,000$ (rare: name not in the list)
Dates: Date variations

- Estimated years. (e.g., “3 Jun 1848” vs. “about 1850”)
- Auto-estimated years. (“<1852>”)  
- Errors in original record. (Census age, “round to nearest 5 years”)
- Confusion between similar events (birth/christening, etc.)
- Lag between event and recording of event. (birth, civil registration; date of event vs. recording)
- Entry or typographical errors. (“1910”/“1901”; “1720”/“172”)
- Calendar changes. (Julian vs. Gregorian calendar, 1582-1900s)
Dates: Levels of Agreement

3: **Exact.** Day, month, year agreement.

2: **Year.** Year agrees; no day/month (or within 1 day)

1: **Near.** Within 2 years; no day/month conflict (agree or missing)

0: **Missing.**

-1: **Differ.** Year off by > 2, or day/month off by more than 1.
Date propagation features

• Child date difference
  – Closest child is <10, <16, <22, <30, >=30 years apart.

• Early child birth: age at other’s child’s birth
  – <5, <15, <18, >= 18

• Late child birth
  – < 45, <55, <65, >=65
Place variation

• Place differences for an event
  – Different places for similar events. (birth/christening)
  – Multiple marriages (in different places)
  – Estimated places. (“of Tennessee”)
  – Data errors.
Place name differences

- Text differences for same place
  - Abbreviations ("VA" vs. "Virginia")
  - Different numbers of levels.
    - ("Rose Hill, Lee, Virginia, USA", "Virginia").
  - Inclusion of place level indicators such as "county" or "city"
    - ("Lee, VA", "Lee Co., VA")
  - Inclusion of commas to indicate "missing levels".
    - (", Lee, VA" vs. "Lee, VA").
  - Changing boundaries.
  - Place name change. (Istanbul/Constantinople. New York/New Amsterdam)
Place agreement levels

- 8: Agreed down to level 4 (i.e., levels 1, 2, 3 and 4 all have the same place id).
- 7: Agreed down to level 3, disagreed at level 4.
  (“Riverton, Salt Lake, Utah, USA” vs. “Draper, Salt Lake, Utah, USA”)
- 6: Agreed down to level 3, no data at level 4.
  (“Rose Hill, Lee, VA, USA” vs. “Lee, VA, USA”)
- 5: Agreed down to level 2, disagreed at level 3.
- 4: Agreed down to level 2, no data at level 3.
- 3: Agreed at level 1 (country), disagreed at level 2 (e.g., state)
- 2: Agreed at level 1 (country), no data at level 2 (i.e., at least one of the places had only a country)
- 1: Disagree at level 1 (i.e., country disagrees)
- 0: Missing data (no effect)
Cross-event place agreement

• "Spouse family" places
  – Individual or spouse’s birth or christening vs.
  – Other person’s marriage or child birth places.

• "All places"
  – All places of one person and their relatives vs.
  – All places of the other person
  – “Did they cross paths?”
Miscellaneous features

• Gender. Hard-coded weight.
• Own ancestor.
• Siblings (matching parent ID)
• No names penalty
Empirical results

• Features:
  – Simple field agreement features
  – Vs. complex multi-valued features

• Weight generation algorithm
  – Probabilistic Record Linkage (Naïve Bayes)
  – vs. Neural Network (Perceptron)

• Train on 48,000 pairs, test on 32,000 pairs.
Empirical Results
## Empirical Results

<table>
<thead>
<tr>
<th>Precision</th>
<th>Simple Fields</th>
<th></th>
<th>Full Features</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>PRL</strong></td>
<td><strong>NN</strong></td>
<td><strong>PRL</strong></td>
<td><strong>NN</strong></td>
</tr>
<tr>
<td>90</td>
<td>77.3</td>
<td>85.5</td>
<td>93.9</td>
<td>98.6</td>
</tr>
<tr>
<td>91</td>
<td>76.4</td>
<td>84.1</td>
<td>93.5</td>
<td>98.5</td>
</tr>
<tr>
<td>92</td>
<td>75.4</td>
<td>82.6</td>
<td>93.2</td>
<td>98.2</td>
</tr>
<tr>
<td>93</td>
<td>74.5</td>
<td>81.2</td>
<td>92.8</td>
<td>98.0</td>
</tr>
<tr>
<td>94</td>
<td>73.5</td>
<td>79.7</td>
<td>92.5</td>
<td>97.7</td>
</tr>
<tr>
<td>95</td>
<td>72.5</td>
<td>77.3</td>
<td>91.0</td>
<td>97.2</td>
</tr>
<tr>
<td>96</td>
<td>71.6</td>
<td>74.9</td>
<td>89.8</td>
<td>96.7</td>
</tr>
<tr>
<td>97</td>
<td>60.7</td>
<td>64.2</td>
<td>86.8</td>
<td>95.5</td>
</tr>
<tr>
<td>98</td>
<td>49.8</td>
<td>55.7</td>
<td>83.6</td>
<td>92.9</td>
</tr>
<tr>
<td>99</td>
<td>40.6</td>
<td>45.1</td>
<td>68.9</td>
<td>90.7</td>
</tr>
<tr>
<td>100</td>
<td>5.9</td>
<td>34.7</td>
<td>32.6</td>
<td>81.6</td>
</tr>
</tbody>
</table>
Research Features

• Scandinavian name stemming
  (-sen, -son, -se, -sdotter, etc. => son)
• Name standard ids
• Generic date propagation
  – Compare birth, marriage, death ranges
• 14-day “near” match

Other areas of research
• Graph-matching
• Family reconstitution / full population matching
Conclusions

• Feature development crucial to accurate matching
• These features can serve as a starting point
• Focus further feature development on cases where errors are most common.
Questions?

Randy Wilson

wilsonr@familysearch.org