Genealogical Record Linkage: Features for Automated Person Matching

Randy Wilson

wilsonr@familysearch.org

FAMILYSEARCH
WHERE GENERATIONS MEET

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

Are these the same person?

Person ID	11157365495	11009911109
Name	[1] Jakob Balli Jacob Balli 11157365495 [hasOrd] M	[2] JAKOB BALLI Jakob Balli Jakob Balli junr Jacob Balli 11009911109 [hasOrd] M
Birth	12 Mar 1809 12 Mar 1809 12 Mar 1809 Matten Interlaken, Bern, Switzerland Matten, Bern, Switz. Matten, Bern, Switzerland	18330118 18Jan1833 18Jan1833 18 Jan 1833 18Jan1833 18 Jan 1833 , MATTEN, BERN, SWITZERLAND Matten, Bern, Switzerland Matton, Bern, Switzerland Matten, Bern, Switz. Interlaken, Bern, Switzer. , Matten, Bern, Switzerland
Death	6 May 1848 6 May 1848 Matten, Bern, Switz.	19130826 26Aug1913 26Aug1913 26 Aug 1913 26 Aug 1913 Salt Lake City, Utah Salt Lake City,, Utah , Salt Lake City, Salt Lake, Utah
Burial		29Aug1913 29 Aug 1913 29 Aug 1913 Salt Lake City, Utah Salt Lake City,, Utah , Salt Lake City, Salt Lake, Utah
Marriage		18670308 1867 8Mar1867 8 Mar 1867 8 Mar 1867 MATTEN, BERN, SWITZERLAND Matten, Bern, Switzerland Matten, Bern, Switz. Matten, Bern, Switzerland
Father	[3] John Casper Balli Johann Casper Balli Johann Caspar Balli Johannes Caspar Balli Johann Kaspar Balli 11145735598 [hasOrd] M	[1] Jakob Balli Jacob Balli 11157365495 [hasOrd] M

FAMILYSEARCH
WHERE GENERATIONS MEET

Measuring accuracy

- Precision
  - Percent of a system's matches that are correct.

  [=correct match / (correct match + false match)]
- Recall
  - Percent of available matches that the system finds.

  [=correct match / (correct match + false differ)]

P/R Example

	True Match	True Differ	Total
Output Match	**90**	**10**	**100**
Output Differ	**30**	290	320
Total	**120**	300	420

- 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)

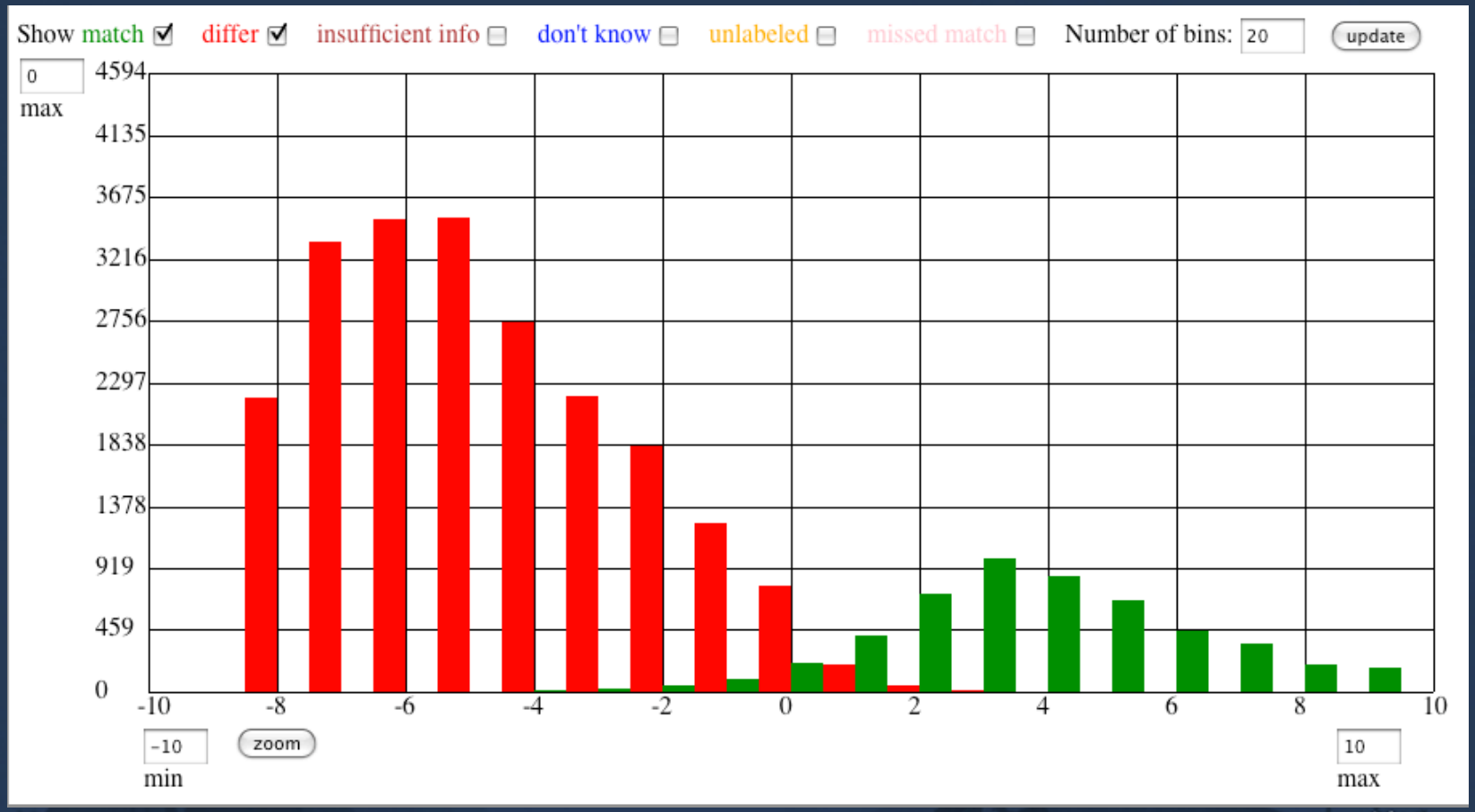
More P/R definitions

Pick whichever definition makes the most sense to you.

- **Precision**:
  Percent of matches that a system comes up with that are correct.
  =100% * (#correct matches) / (#correct matches + #incorrect matches)
  =100% * (#correct matches) / (total #matches found)
  =100% - (Percent of matches that a system comes up with that are wrong)
  =100% - (false match rate)

- **Recall**:
  Percent of true matches in the data that the system comes up with.
  =100% * (#correct matches found) / (#correct matches found + #correct matches not found)
  =100% * (#correct matches found) / (#matches available in the data)
  =100% - (Percent of matches that the system failed to find)

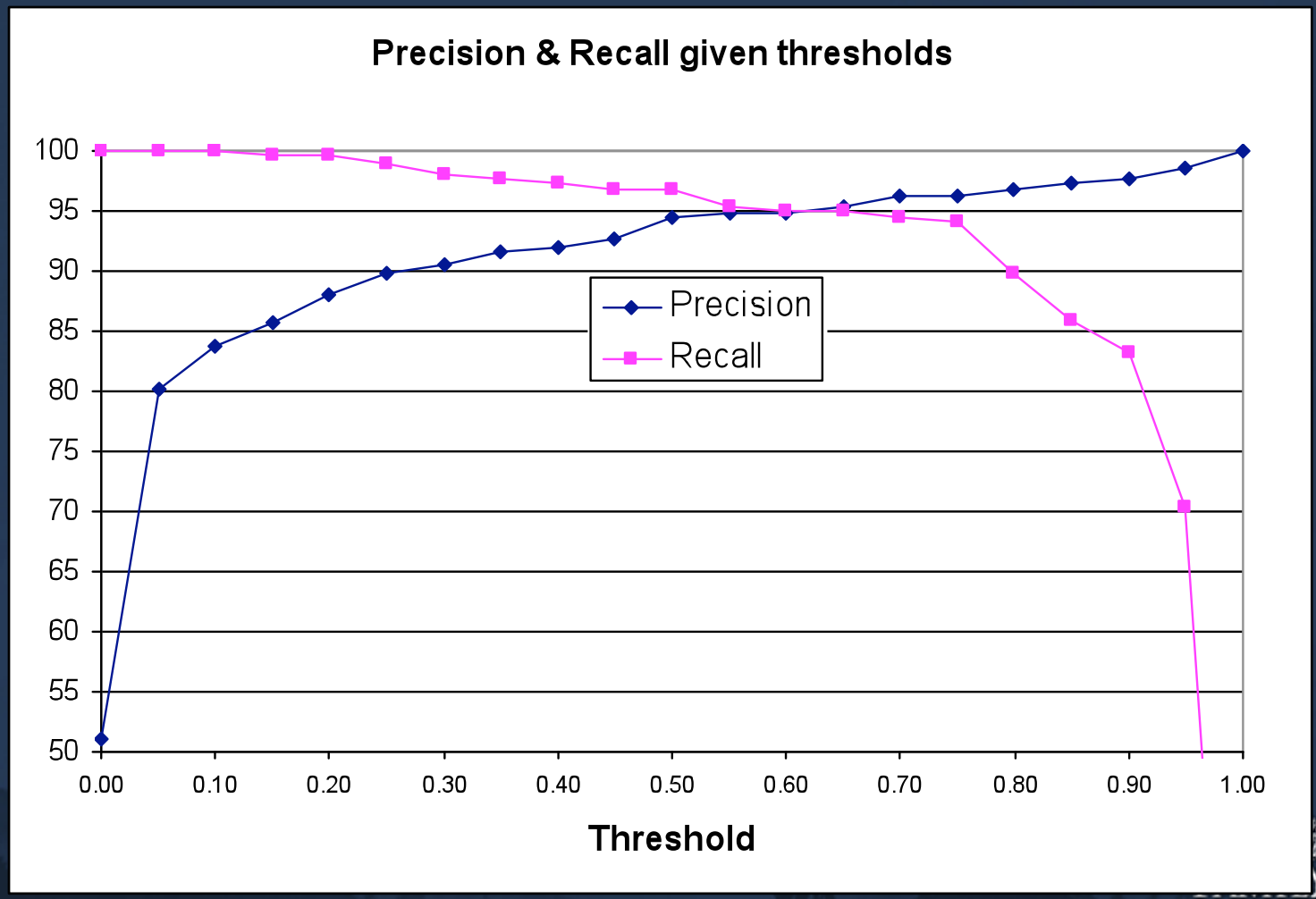
Histogram: P/R Trade-off



FAMILYSEARCH
WHERE GENERATIONS MEET

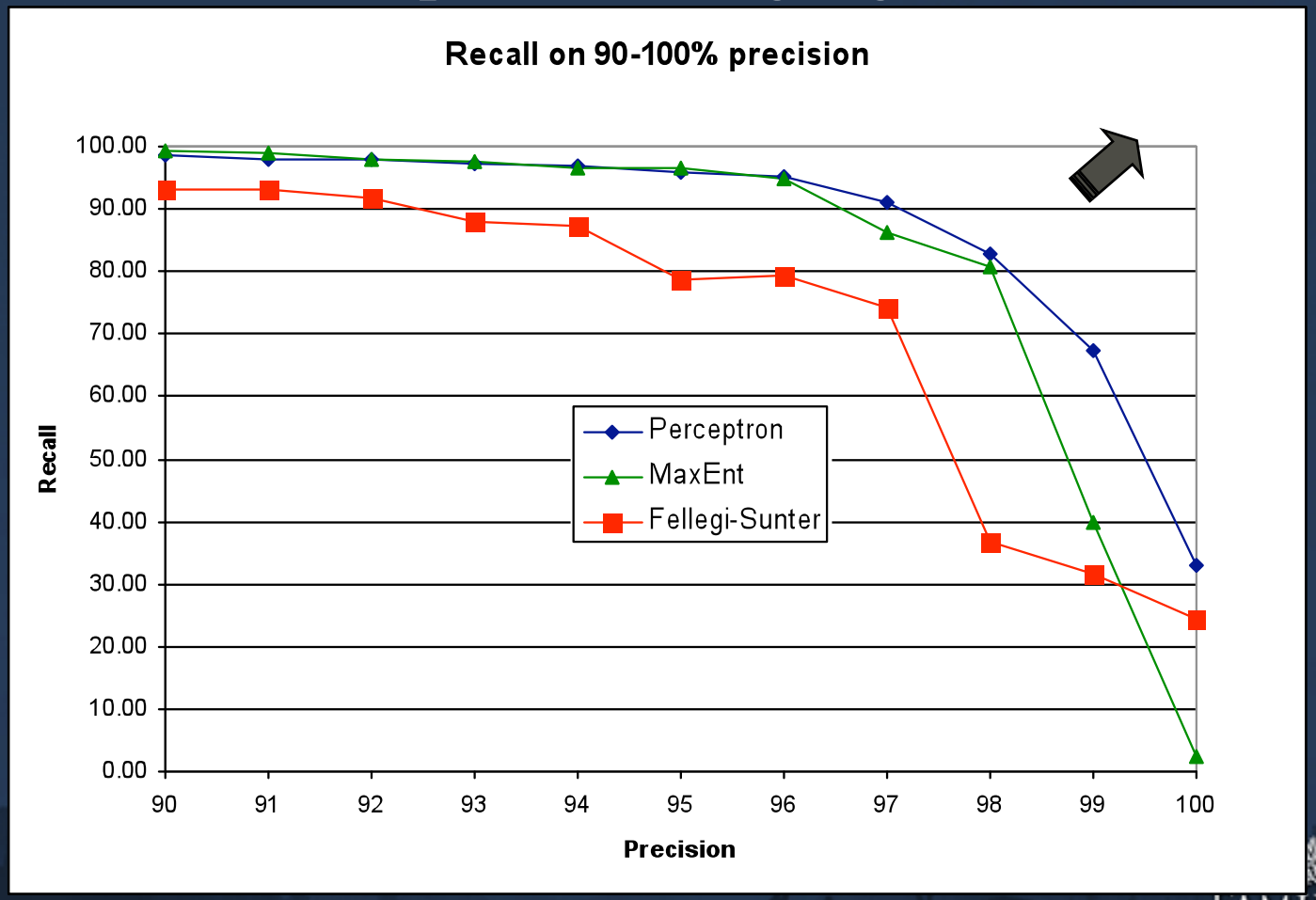
P/R Curves and Thresholds

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

FAMILYSEARCH
WHERE GENERATIONS MEET

Thresholds for star ratings

	Name	Birth or Christening	Death or Burial	Spouse	Parents
The first record listed below is the record from your family tree					
☑ 📁	Drusilla Dorris		20 May 1881 Richmond, Cache, Utah	James Hendricks	William Dorris Catherine Frost
The records listed below are the possible duplicates					
☐ **1.**	Drusilla Dorris ★★★★★		20 March 1881	James Hendricks	William Dorris Catherine Frost
☐ **2.**	Drusilla Dorris ★★★★★	about 1813 Franklin, Simpson, Kentucky			William Dorris Catherine Frost
☐ **3.**	Drusilla Dorris ★★☆☆☆	1805 of, Sumner, Tennessee, USA			William Dorris Catherine Frost
☐ **4.**	Drusilla Dorris ★★☆☆☆	about 1803 of, Turner, Tn			William Dorris Catherine Frost
☐ **5.**	Rebecca Dorris ★★☆☆☆	22 February 1793 , Turner, Tennessee, USA	18 December 1835	Samuel Hendricks	William Dorris Catherine Frost
☐ **6.**	Rebecca Dorris ★☆☆☆☆	about 1796 Sumner, Tennesse, United States		Samuel Hendricks	William Dorris Catherine Frost
☐ **7.**	Tabitha Dorris ★☆☆☆☆	12 January 1804		John Hendricks	William Dorris Catherine Frost

Compare in More Detail

FAMILYSEARCH
WHERE GENERATIONS MEET

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

IndGivenName=-1: -2.2224
IndGivenName=1: 0.5968
IndGivenName=2: 0.687
IndGivenName=3: 0.0743
IndGivenName=4: 1.5611
IndGivenName=5: 0.686
IndGivenName=6: 0.4946
IndGivenName=7: 1.2099
IndCommonGivenName=1: 1.0244
IndCommonGivenName=2: 1.0773
IndCommonGivenName=3: 1.1974
IndCommonGivenName=4: 1.4942
IndSurname=-1: -1.8169
IndSurname=1: 1.4038
...
Bias: -5.0982

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”)

	john	**henry**
johan	Near	Differ
h	Differ	Far

Names: Piece alignment

	john	**henry**
johan	Near	Differ
h	Differ	Far

johan	john	*Near*
h	henry	*Far*

	john	**henry**
johan	Near	Differ

johan	john	*Near*
<none>	henry	*<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 <= 40 (very common: John is 1 in 25)
  - 2: 40 < Odds <= 300
  - 3: 300 < Odds <= 1500
  - 4: Odds > 1500 (rare: name not in the list)
- Surnames
  - 1: Odds <= 4000 (common)
  - 2: 4000 < Odds <= 10,000
  - 3: 10,000 < Odds <= 100,000
  - 4: 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?"

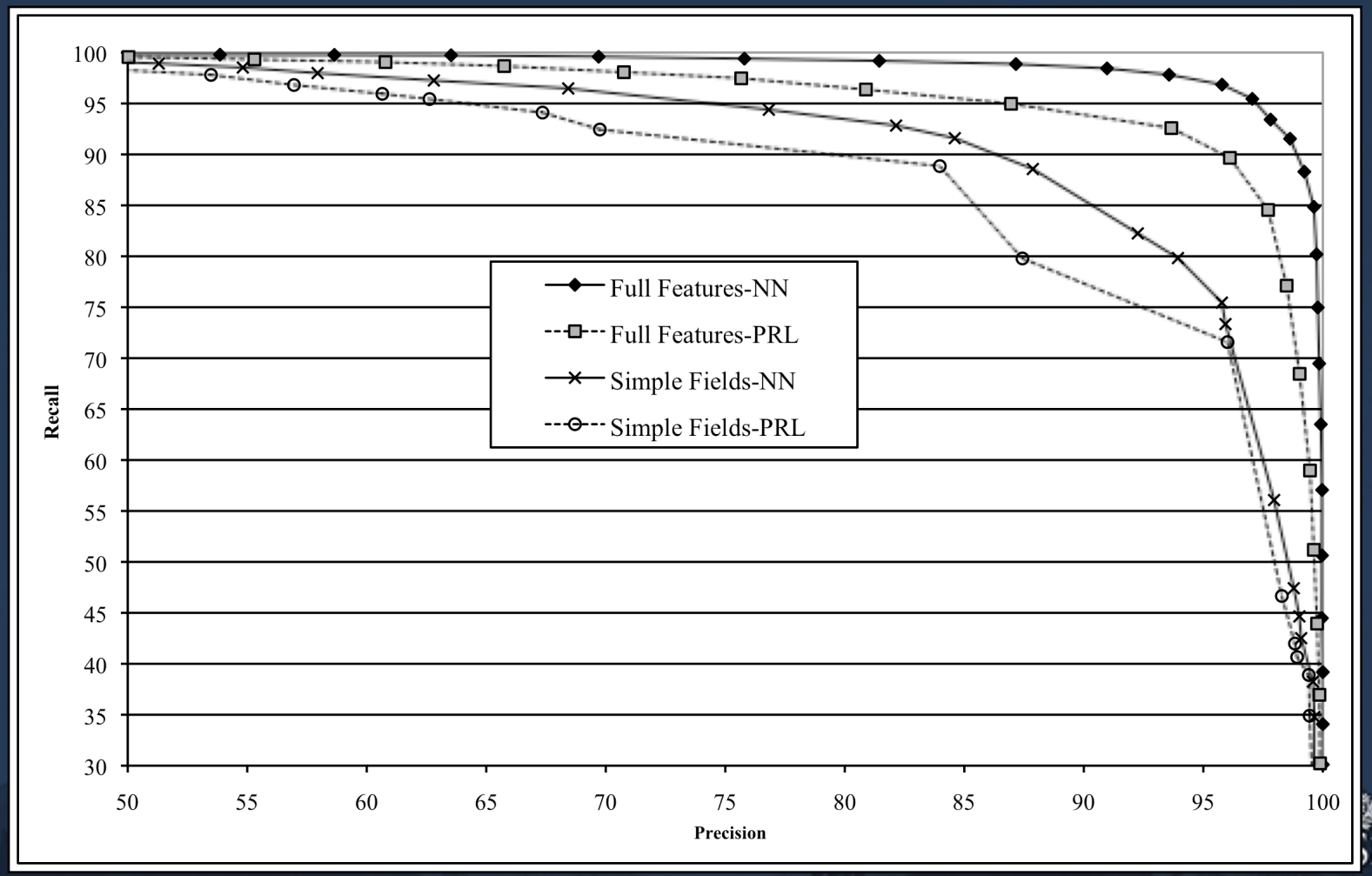
Miscellaneous features

- 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

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

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

FAMILYSEARCH
WHERE GENERATIONS MEET

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

FAMILYSEARCH
WHERE GENERATIONS MEET