Name Standardization for Genealogical Record Linkage

Randy Wilson

Family & Church History Department

The Church of Jesus Christ of Latter-Day Saints

WilsonR@ldschurch.org

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*

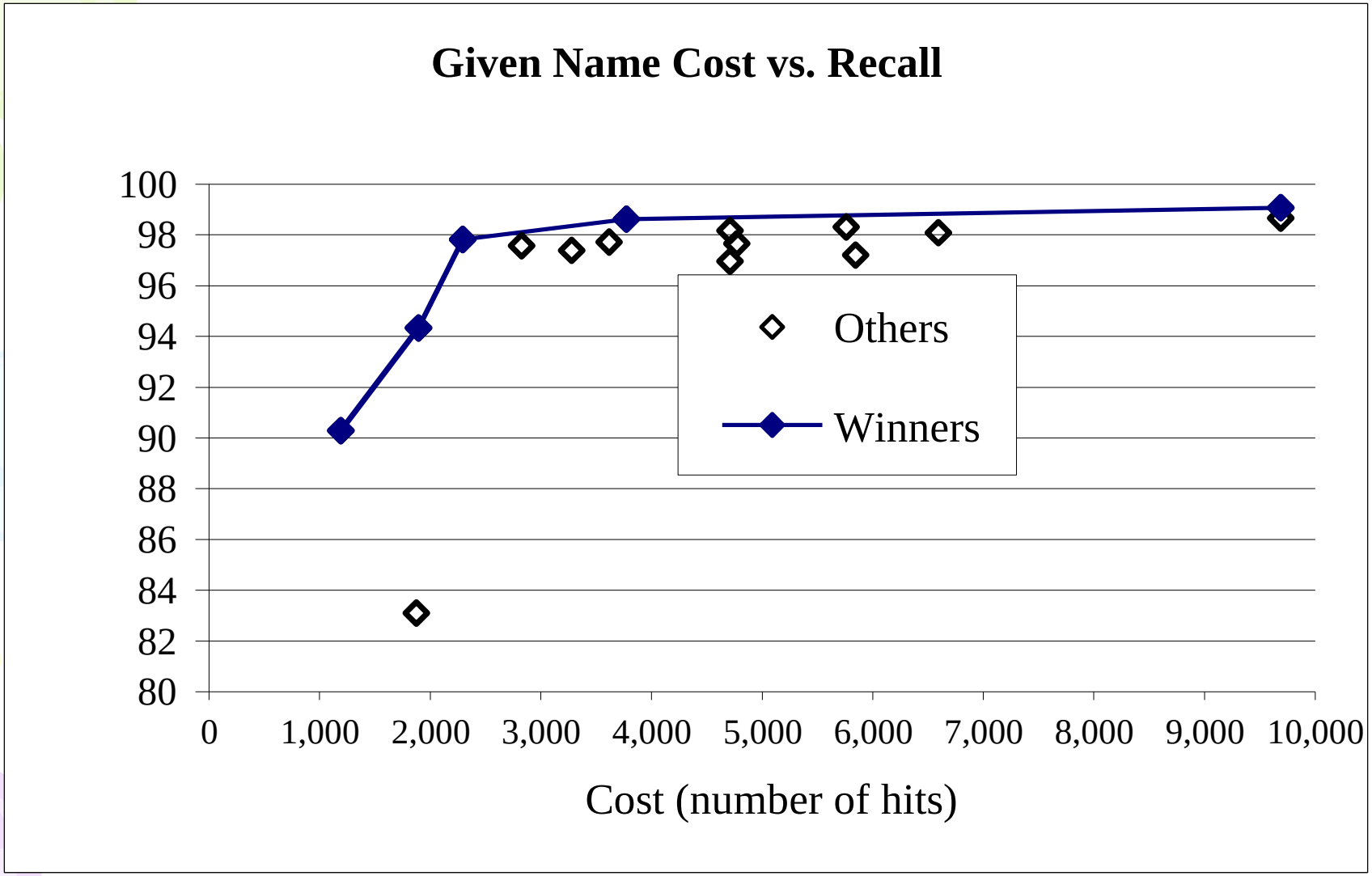
Recall vs. “Cost”

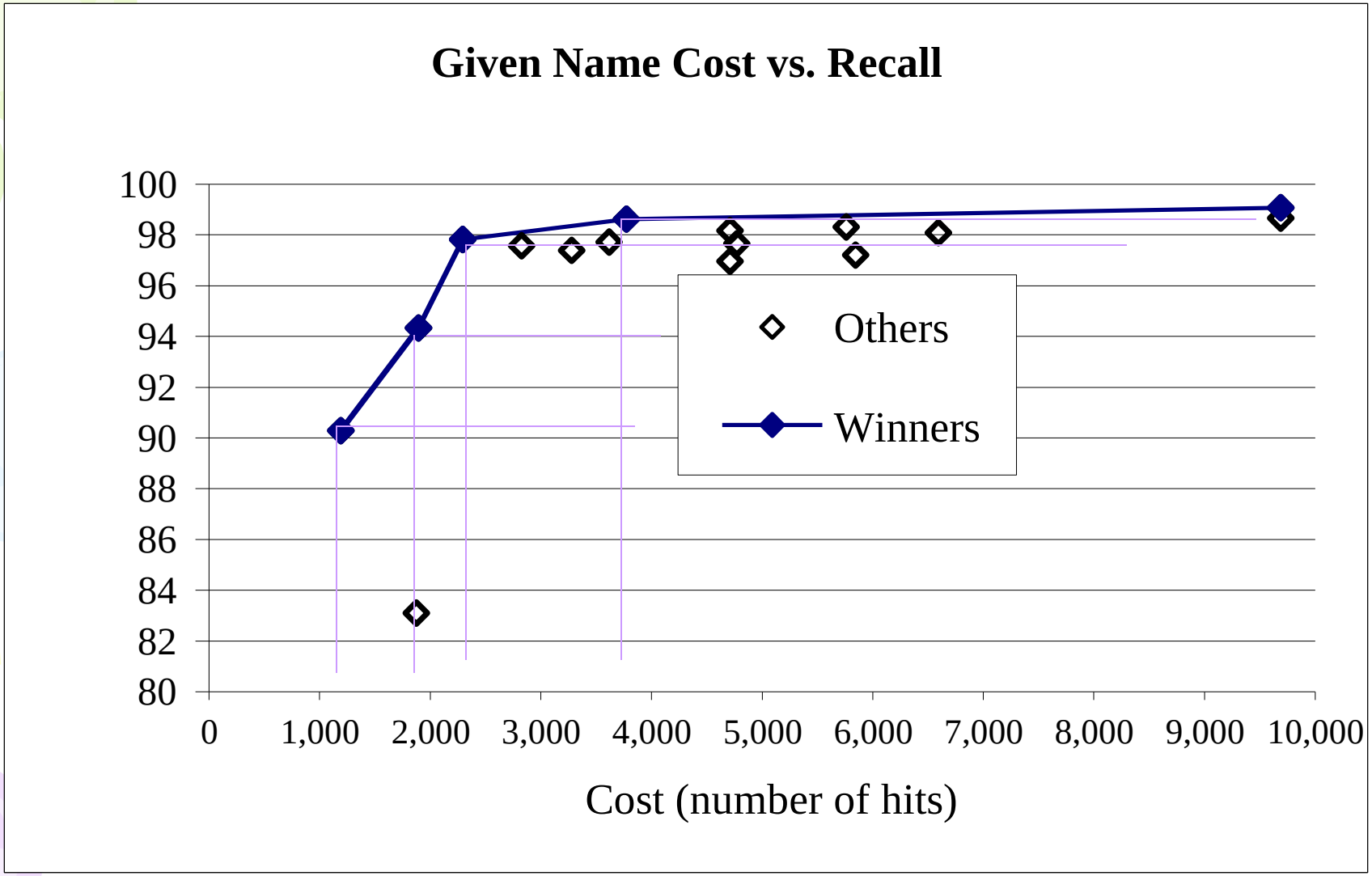
- **Recall:** % of known matches that are “brought together” by a given standardization technique.

- **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.

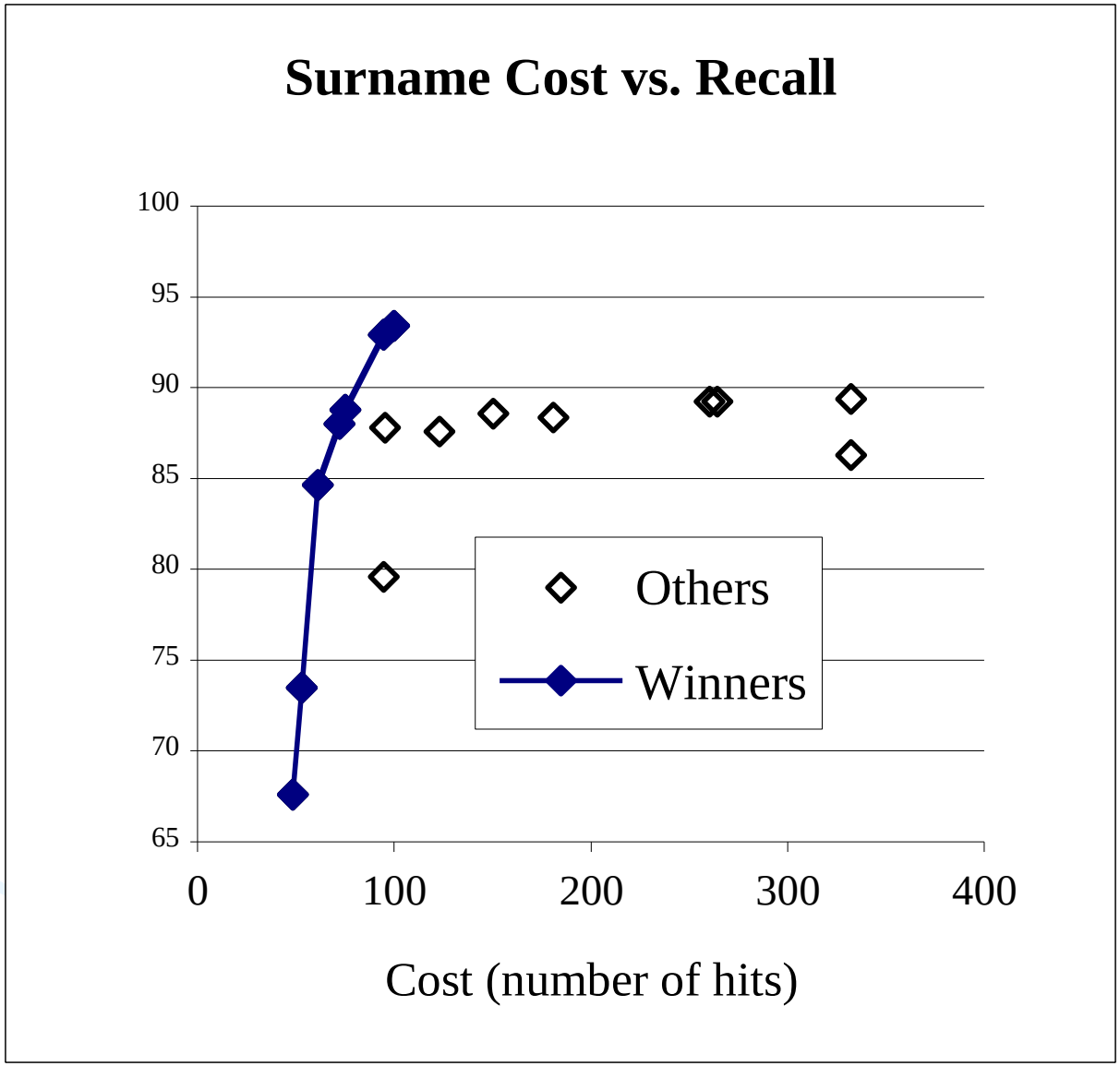

Given Name Cost vs. Recall
100
98
96
94
92
90
88
86
84
82
80
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,000
Cost (number of hits)
Others
Winners


Given Name Cost vs. Recall
100
98
96
94
92
90
88
86
84
82
80
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,000
Cost (number of hits)
Others
Winners

Given Names

Given Name Fields	Recall	AvgHits	% of Db
Universal + Orig	**99.08**	**9689**	**5.42%**
Universal	98.67	9689	5.42%
ODM + Orig	**98.62**	**3771**	**2.11%**
Soundex	98.31	5761	3.22%
Culture_default + Orig	98.16	4712	2.63%
Double Metaphone	98.09	6595	3.69%
Culture_relative + Orig	**97.81**	**2292**	**1.28%**
ODM	97.72	3620	2.02%
Metaphone	97.65	4771	2.67%
Culture_person + Orig	97.57	2828	1.58%
Edit	97.40	3280	1.83%
NYSIIS	97.21	5847	3.27%
Culture_default	96.96	4712	2.63%
Orig	**94.32**	**1895**	**1.06%**
Culture_relative	**90.30**	**1191**	**0.67%**
Culture_person	83.11	1875	1.05%

Surnames


Surname Cost vs. Recall
100
95
90
85
80
75
70
65
0
100
200
300
400
Others
Winners
Cost (number of hits)

Surname

Surname Fields	Recall	AvgHits	% of Db
ODM + Orig	**93.41**	**99.9**	0.06%
ODM	**92.92**	**94.8**	0.05%
Universal + Orig	89.39	332.4	0.19%
Double Metaphone	89.24	264.4	0.15%
Soundex	89.22	260.5	0.15%
Culture_relative + Orig	**88.79**	**75.1**	**0.04%**
NYSIIS	88.57	150.3	0.08%
Metaphone	88.35	181.0	0.10%
Culture_person + Orig	**88.02**	**72.2**	**0.04%**
Culture_default + Orig	87.79	95.5	0.05%
Edit	87.59	123.2	0.07%
Universal	86.28	332.1	0.19%
Orig	**84.62**	**61.1**	**0.03%**
Culture_default	79.58	94.7	0.05%
Culture_relative	**73.50**	**53.1**	**0.03%**
Culture_person	**67.60**	**48.3**	**0.03%**

Given + Surname Fields	Recall	AvgHits	% of Db
ODM + Orig	**99.68**	**3850**	**2.15%**
Soundex	99.54	5998	3.35%
Universal + Orig	99.42	9990	5.58%
NYSIIS	99.41	5976	3.34%
Culture_relative + Orig	**99.35**	**2348**	**1.31%**
Culture_person + Orig	99.25	2882	1.61%
Metaphone	99.20	4931	2.76%
Double Metaphone	99.20	6835	3.82%
Culture_default + Orig	99.16	4788	2.68%
Orig + Swap	**98.53**	**2135**	**1.19%**
Orig	**98.00**	**1939**	**1.08%**
Edit + Orig	98.00	1939	1.08%

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.