# Retrieving a Sorted List of Hundreds of Closest Relatives from FamilySearch Family Tree in Seconds

Family History Technology Workshop
Brigham Young University
March 20, 2014

Ben Baker
bakerb@familysearch.org

FamilySearch

# Weighted Relationship Distance

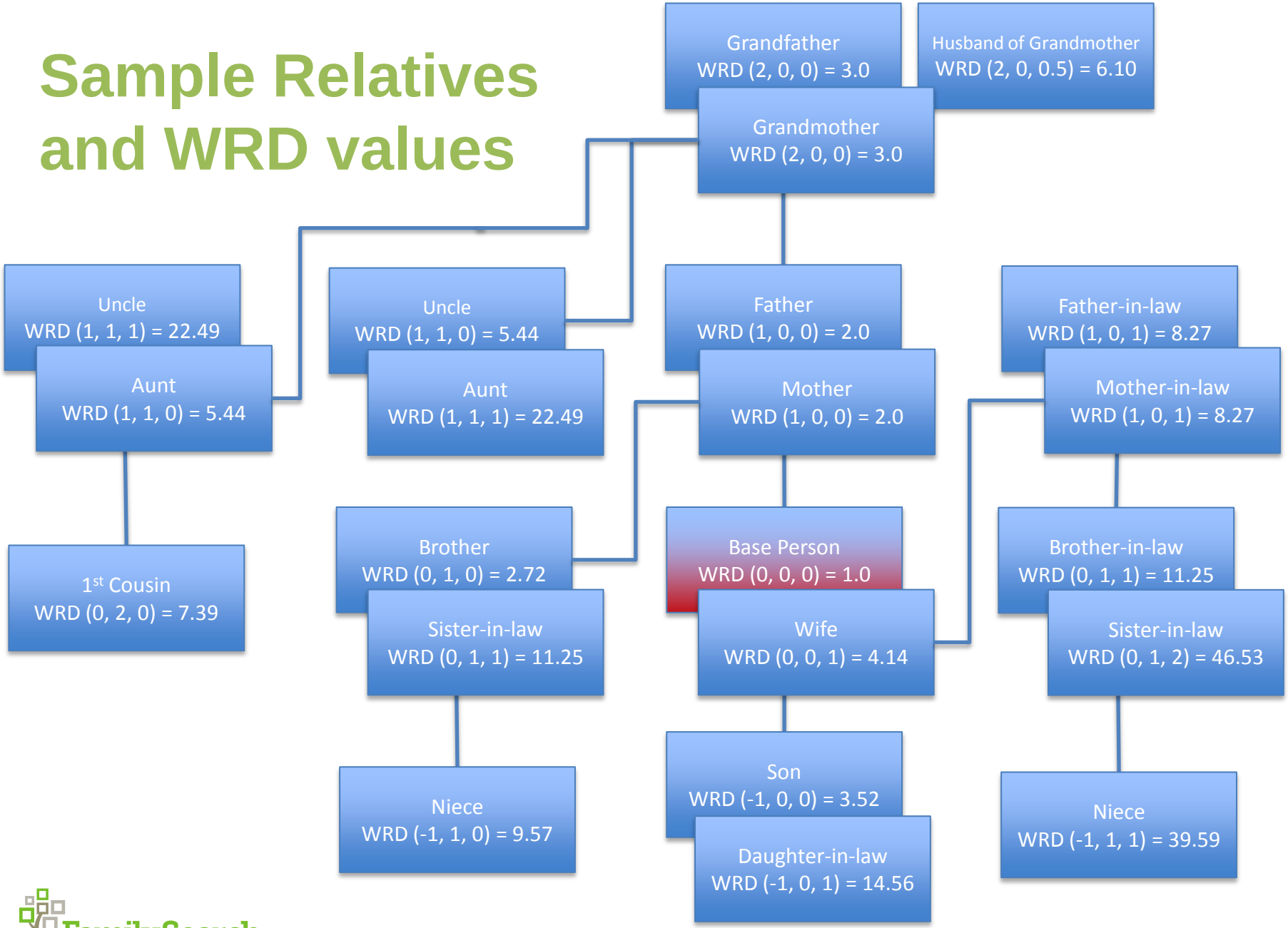
$$WRD_{(g,\ c,\ m)} = \alpha(|g| + 1)\ e^{\beta c}\ e^{\gamma m}$$

g – Generational or “vertical” distance
Number of generations from base person
c – Collateral or “horizontal” distance
Minimum generations to a closest common ancestor
m – Marriage distance
Number of marriages between base person
α, β, γ – Weighting factors to control growth rates

2013 Family History Technology Workshop – *Beyond the Relationship Calculator*
*Using a Weighted Relationship Distance Metric to Prioritize, Categorize and Visualize Relatives*
http://fht.byu.edu/prev_workshops/workshop13/papers/baker-beyond-fhtw2013.pdf

FamilySearch

# Sample Relatives and WRD values

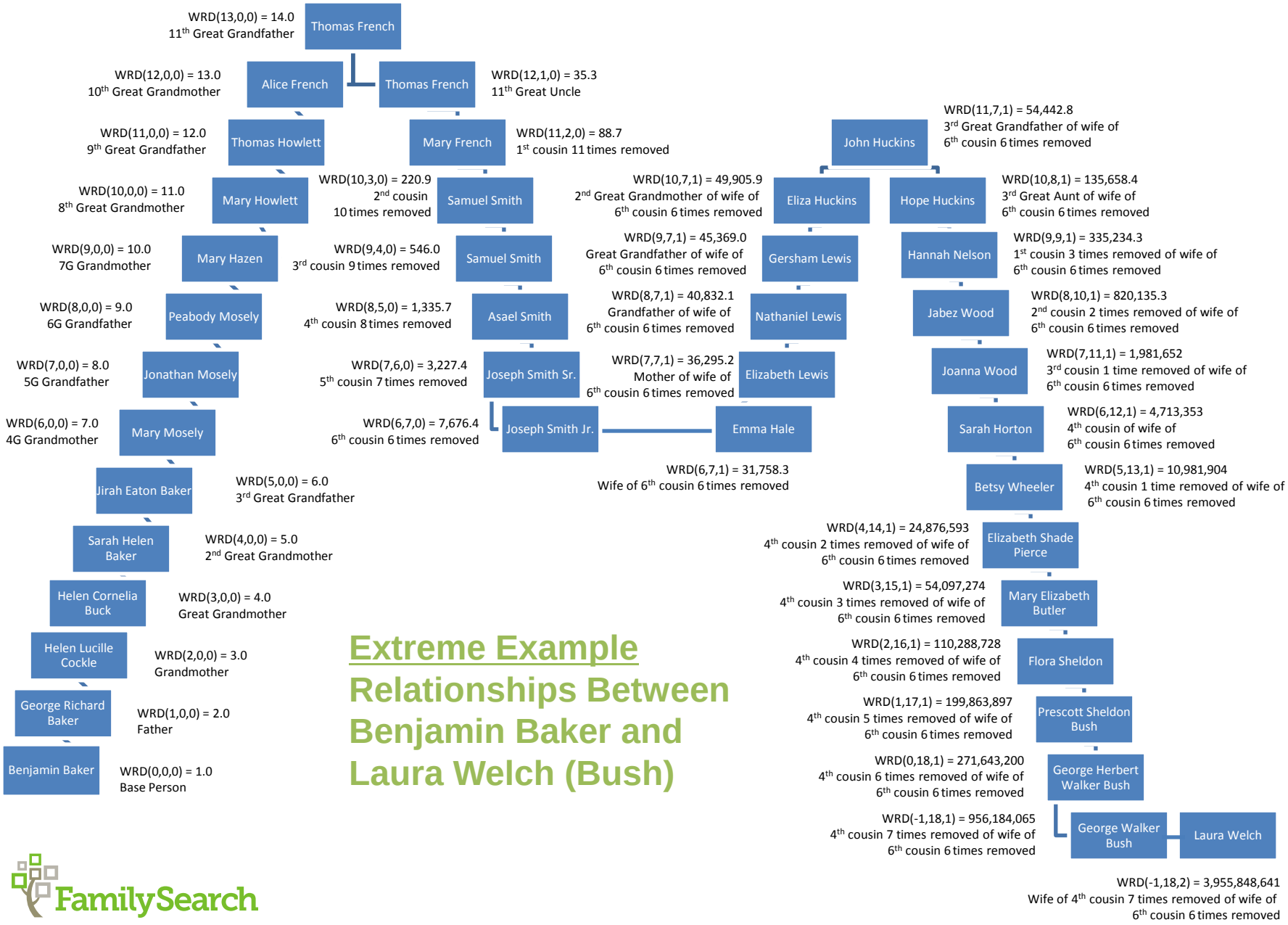



FamilySearch

# Same Sample Relatives in Table Format in Sorted WRD Order

| Relative Description | g | c | m | WRD |
|---|---|---|---|---|
| Base Person | 0 | 0 | 0 | 1.0 |
| Father/Mother | 1 | 0 | 0 | 2.0 |
| Brother | 0 | 1 | 0 | 2.72 |
| Grandfather/Grandmother | 2 | 0 | 0 | 3.0 |
| Son | -1 | 0 | 0 | 3.52 |
| Wife | 0 | 0 | 1 | 4.14 |
| Aunt/Uncle (sibling of parent) | 1 | 1 | 0 | 5.44 |
| Husband of Grandmother | 2 | 0 | 0.5 | 6.10 |
| 1st cousin | 0 | 2 | 0 | 7.39 |
| Father/mother-in-law | 1 | 0 | 1 | 8.27 |
| Niece (daughter of brother) | -1 | 1 | 0 | 9.57 |
| Brother/sister-in-law (spouse of sibling or sibling of spouse) | 0 | 1 | 1 | 11.25 |
| Daughter-in-law | -1 | 0 | 1 | 14.56 |
| Aunt/Uncle (spouse of sibling of parent) | 1 | 1 | 1 | 22.49 |
| Niece (daughter of spouse's brother) | -1 | 1 | 1 | 39.59 |
| Sister-in-law (wife of spouse's brother) | 0 | 1 | 2 | 46.53 |

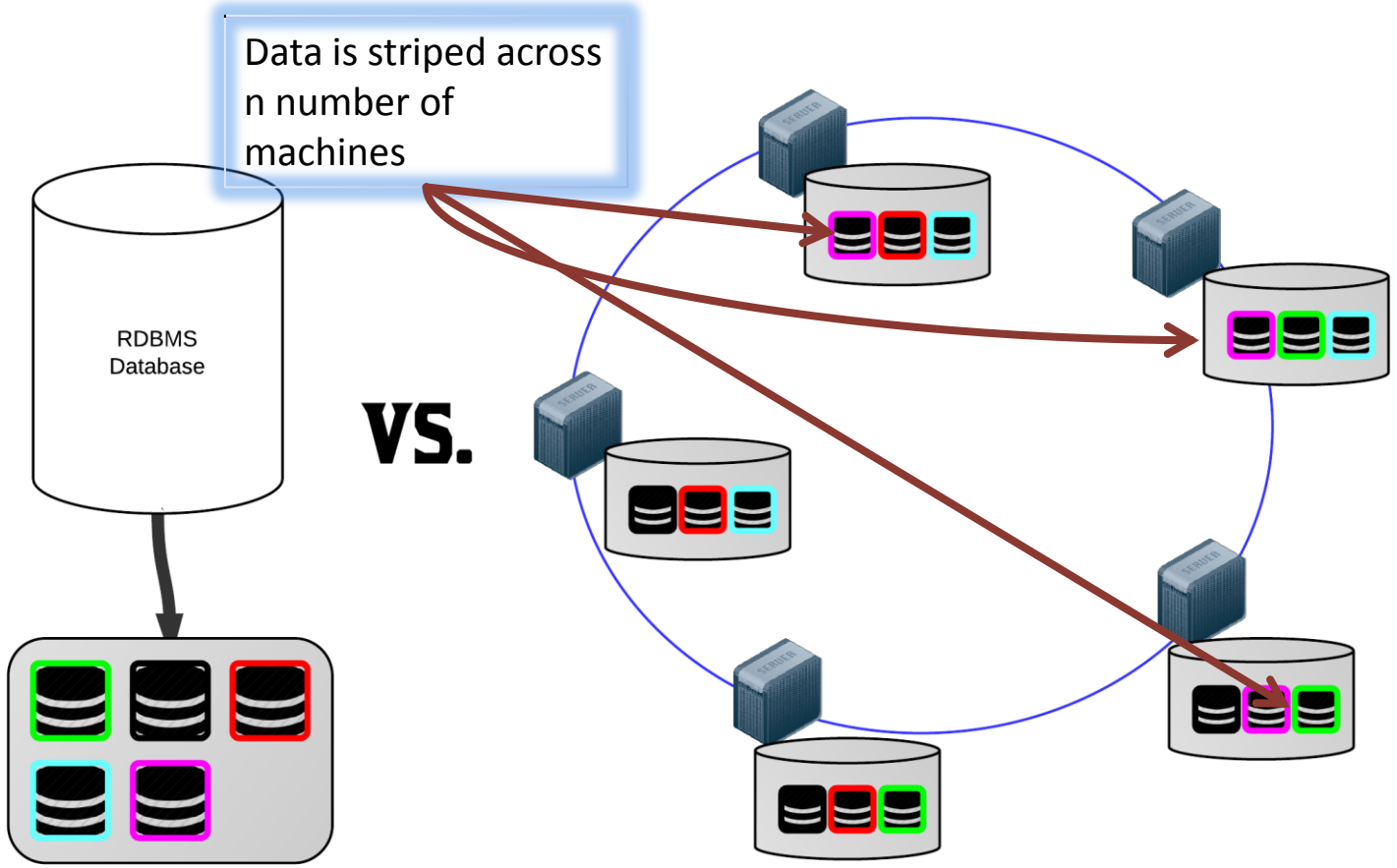
FamilySearch


Extreme Example
Relationships Between
Benjamin Baker and
Laura Welch (Bush)
WRD(13,0,0) = 14.0
11th Great Grandfather
Thomas French
WRD(12,0,0) = 13.0
10th Great Grandmother
Alice French
Thomas French
WRD(12,1,0) = 35.3
11th Great Uncle
WRD(11,0,0) = 12.0
9th Great Grandfather
Thomas Howlett
Mary French
WRD(11,2,0) = 88.7
1st cousin 11 times removed
WRD(10,0,0) = 11.0
8th Great Grandmother
Mary Howlett
WRD(10,3,0) = 220.9
2nd cousin
10 times removed
Samuel Smith
WRD(9,0,0) = 10.0
7G Grandmother
Mary Hazen
WRD(9,4,0) = 546.0
3rd cousin 9 times removed
Samuel Smith
WRD(8,0,0) = 9.0
6G Grandfather
Peabody Mosely
WRD(8,5,0) = 1,335.7
4th cousin 8 times removed
Asael Smith
WRD(7,0,0) = 8.0
5G Grandfather
Jonathan Mosely
WRD(7,6,0) = 3,227.4
5th cousin 7 times removed
Joseph Smith Sr.
WRD(6,0,0) = 7.0
4G Grandmother
Mary Mosely
WRD(6,7,0) = 7,676.4
6th cousin 6 times removed
Joseph Smith Jr.
Jirah Eaton Baker
WRD(5,0,0) = 6.0
3rd Great Grandfather
Sarah Helen Baker
WRD(4,0,0) = 5.0
2nd Great Grandmother
Helen Cornelia Buck
WRD(3,0,0) = 4.0
Great Grandmother
Helen Lucille Cockle
WRD(2,0,0) = 3.0
Grandmother
George Richard Baker
WRD(1,0,0) = 2.0
Father
Benjamin Baker
WRD(0,0,0) = 1.0
Base Person
WRD(10,7,1) = 49,905.9
2nd Great Grandmother of wife of
6th cousin 6 times removed
WRD(9,7,1) = 45,369.0
Great Grandfather of wife of
6th cousin 6 times removed
WRD(8,7,1) = 40,832.1
Grandfather of wife of
6th cousin 6 times removed
WRD(7,7,1) = 36,295.2
Mother of wife of
6th cousin 6 times removed
Emma Hale
WRD(6,7,1) = 31,758.3
Wife of 6th cousin 6 times removed
John Huckins
WRD(11,7,1) = 54,442.8
3rd Great Grandfather of wife of
6th cousin 6 times removed
Eliza Huckins
Hope Huckins
WRD(10,8,1) = 135,658.4
3rd Great Aunt of wife of
6th cousin 6 times removed
Gersham Lewis
Hannah Nelson
WRD(9,9,1) = 335,234.3
1st cousin 3 times removed of wife of
6th cousin 6 times removed
Nathaniel Lewis
Jabez Wood
WRD(8,10,1) = 820,135.3
2nd cousin 2 times removed of wife of
6th cousin 6 times removed
Elizabeth Lewis
Joanna Wood
WRD(7,11,1) = 1,981,652
3rd cousin 1 time removed of wife of
6th cousin 6 times removed
Sarah Horton
WRD(6,12,1) = 4,713,353
4th cousin of wife of
6th cousin 6 times removed
Betsy Wheeler
WRD(5,13,1) = 10,981,904
4th cousin 1 time removed of wife of
6th cousin 6 times removed
WRD(4,14,1) = 24,876,593
4th cousin 2 times removed of wife of
6th cousin 6 times removed
Elizabeth Shade Pierce
WRD(3,15,1) = 54,097,274
4th cousin 3 times removed of wife of
6th cousin 6 times removed
Mary Elizabeth Butler
WRD(2,16,1) = 110,288,728
4th cousin 4 times removed of wife of
6th cousin 6 times removed
Flora Sheldon
WRD(1,17,1) = 199,863,897
4th cousin 5 times removed of wife of
6th cousin 6 times removed
Prescott Sheldon Bush
WRD(0,18,1) = 271,643,200
4th cousin 6 times removed of wife of
6th cousin 6 times removed
George Herbert Walker Bush
WRD(-1,18,1) = 956,184,065
4th cousin 7 times removed of wife of
6th cousin 6 times removed
George Walker Bush
Laura Welch
WRD(-1,18,2) = 3,955,848,641
Wife of 4th cousin 7 times removed of wife of
6th cousin 6 times removed
FamilySearch

# Relationship Calculator Deficiencies

1. A relationship calculation must be initiated between two persons in an ad hoc manner and repeated for a different set of two persons
2. It is not possible to sort a list of arbitrarily related persons by closeness
3. Relationship calculations through marriages are not possible
4. Data to perform the relationship calculations must be pre-calculated and stored to be performant enough for on-demand calculation in an enterprise system

# Apache Cassandra-Based FamilySearch Family Tree




FamilySearch

# Proposed System

RESTful web service endpoint added to 5-node Cassandra based test system

**Method / Description**

*GET relatives/{id}*

Retrieve the closest relatives to a person up to a specified maximum weighted relationship distance.

**Parameters**

*double maxDistance* – Optional query parameter (default 10.0) to specify the maximum distance to return relatives of the person up to.

**Returns**

A list of RelativePerson objects, sorted by increasing distance

Future methods planned to retrieve closest common ancestor and closest relation between two people.

FamilySearch

# RelativePerson Object


```
{
  "name": "George Dean Cockle",
  "id": "KWJX-VMC",
  "lifespan": "1901 - 1970",
  "fatherIds": [
    "K2V7-PV5"
  ],
  "motherIds": [
    "K2V7-P92"
  ],
  "spouseIds": [
    "MM8P-MGS",
    "KWBB-7G9"
  ],
  "familyName": "Cockle",
  "givenName": "George Dean",
  "gender": "MALE",
  "childIds": [
  ],
    "relative": {
      "id": "K2V7-PV5",
      "name": "John Cockle",
      "lifespan": "1852 - 1921",
      "relativeRole": "CHILD"
    },
    "weightedRelationshipDistance": {
      "starRanking": 8,
      "simpleRelationshipDistance": "3",
      "weightedRelationshipDistance": "8.155",
      "generationDistance": 2,
      "collateralDistance": 1,
      "marriageDistance": "0"
    },
    "relativeDescription": "Great Uncle"
}
```


FamilySearch

# Performance Results

| Maximum WRD | Num Persons | Mean Execution Time |
|---|---|---|
| 5.0 | 43 | 1.95s |
| 10.0 | 167 | 22.1s |
| 15.0 | 554 | 89.8s |

Performance tuning is still necessary to make this service more useful for production use.

For comparison purposes, the average time to retrieve data for a pedigree containing 10-15 persons in FamilySearch Family Tree is typically about 3-5 seconds. The Eureka team has show sub-second pedigree load time involving the same number of persons and loading up to 12 generations in several seconds.

FamilySearch

# Live Demo

FamilySearch

# Future Applications Enabled

- Identifying the closest relatives where historical record hints have been identified but not attached yet as sources.
- Promoting e-mail campaigns to point out what others have added to your relatives such as new photos, sources, stories, etc. to draw users back to the site.
- Easily identifying end of line relatives and likely places for successful descendancy research.
- Facilitate LDS temple work for closest relatives first and sharing more distantly related people with others.
- Producing to-do lists sorted by closeness of relation on any task a user may want to undertake (Ex. fixing data anomalies, merging possible duplicates, providing missing data, etc.)
- Automatic watching of relatives via e-mail alerts based on closeness.
- Applications across a set of users (Ex. closeness of relation to the user for all LDS temple submissions or photo uploads on FamilySearch)
- Sorting items such as memories, watched persons, LDS temple reservation lists, etc. in order of those closest to the user.
- Pointing out relatives who have been identified as prominent or have participated in significant events.
- More . . .

# Next Steps

- Full transition from Oracle to Cassandra expected to take a while, but expect one-way synchronization sooner
- Enables read-only operations on Family Tree data, including this work
- Intend to petition to utilize this work in production before transition is complete
- May implement a smaller-scale solution in the current system to prove value of WRD metric over current scope of interest service

# Additional Q & A

FamilySearch