Managing Big Data Reaching Back to the 11th Century

Scott Sorensen

ancestry.com

Ancestry.com mission

ancestry.com™

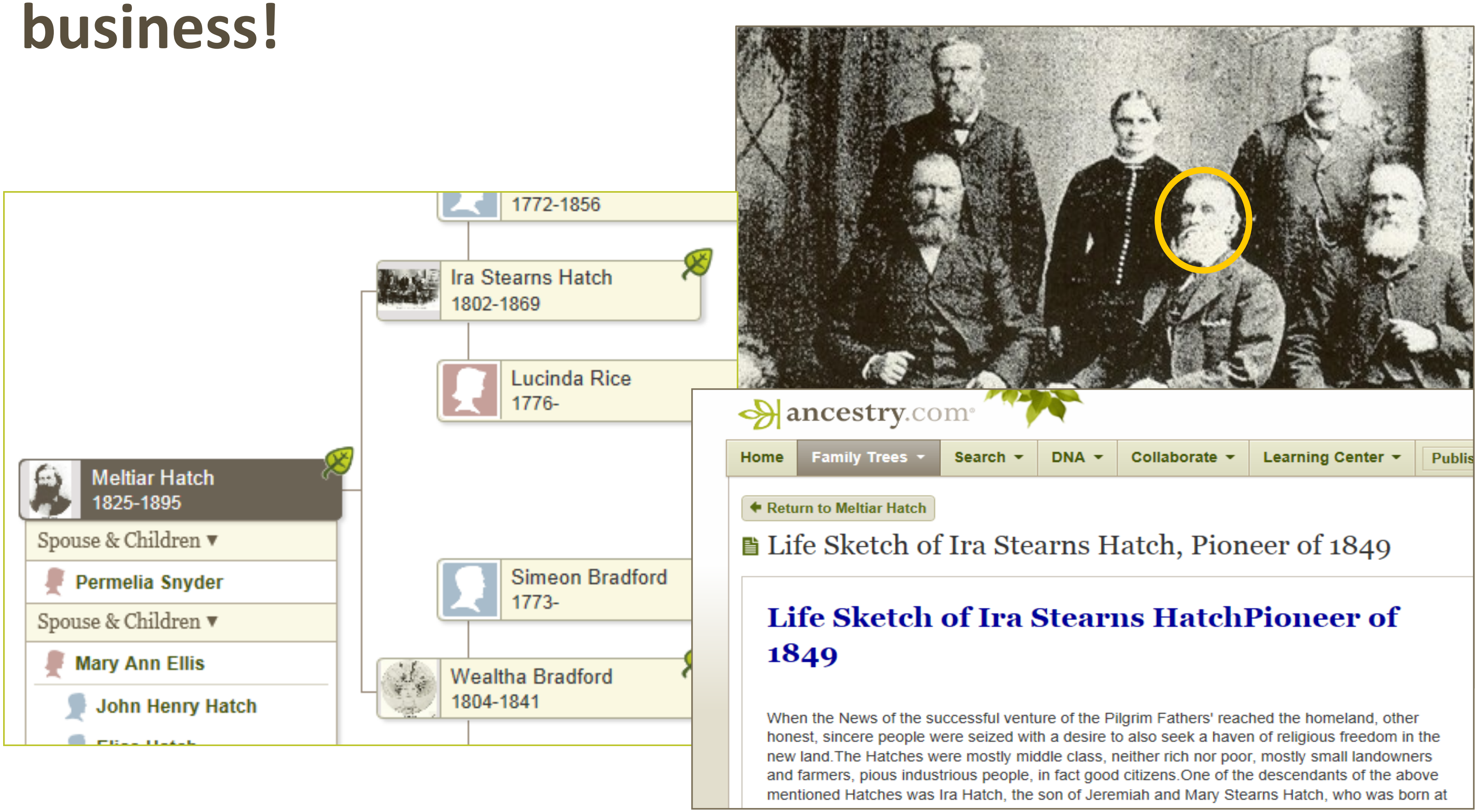
Our mission is to help everyone

discover, preserve and share

their family history.

Data is our product

It's the "aha" moment of a discovery that drives our business!



1772-1856

Ira Stearns Hatch
1802-1869

Lucinda Rice
1776-

Meltiar Hatch
1825-1895

Spouse & Children ▾

Permelia Snyder

Spouse & Children ▾

Mary Ann Ellis

John Henry Hatch

Simeon Bradford
1773-

Wealtha Bradford
1804-1841

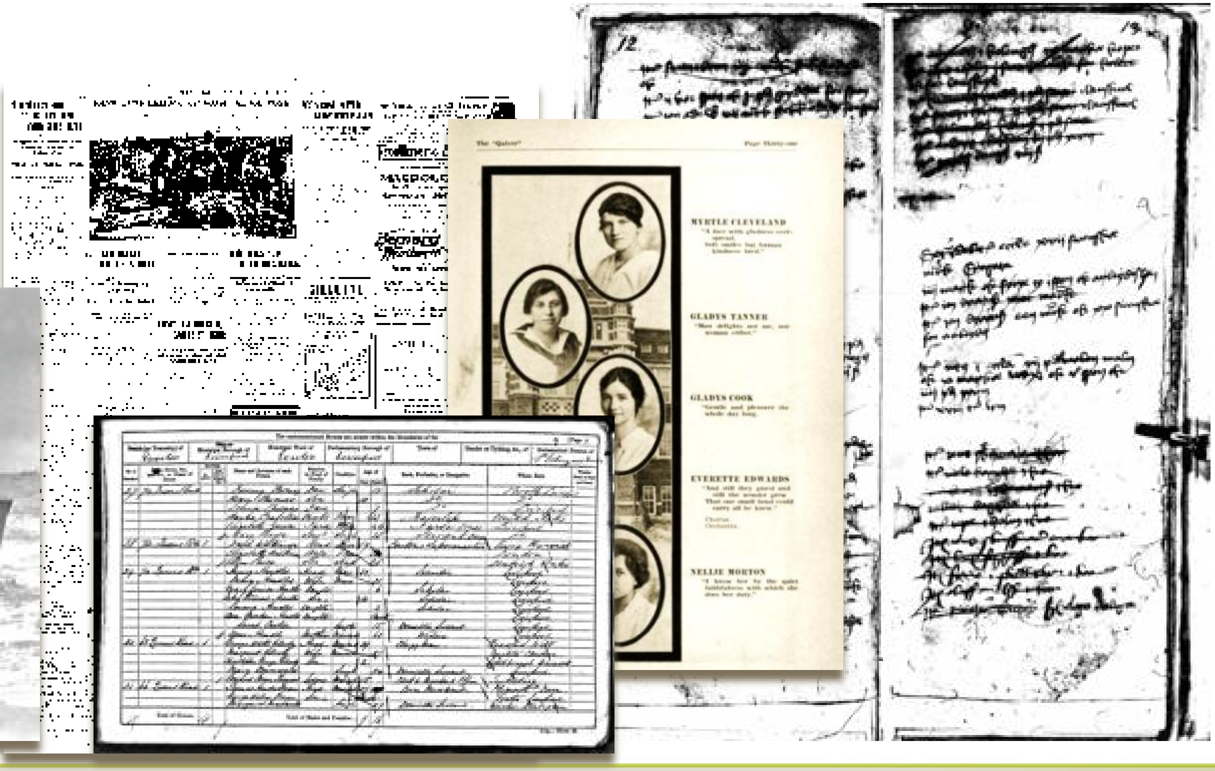
Life Sketch of Ira Stearns Hatch, Pioneer of 1849

Life Sketch of Ira Stearns HatchPioneer of 1849

When the News of the successful venture of the Pilgrim Fathers' reached the homeland, other honest, sincere people were seized with a desire to also seek a haven of religious freedom in the new land.The Hatches were mostly middle class, neither rich nor poor, mostly small landowners and farmers, pious industrious people, in fact good citizens.One of the descendants of the above mentioned Hatches was Ira Hatch, the son of Jeremiah and Mary Stearns Hatch, who was born at

World’s largest online family history resource

- Over 30,000 historical content collections
- Records dating back to 11th century
- 12 billion records and images
- 10 petabytes

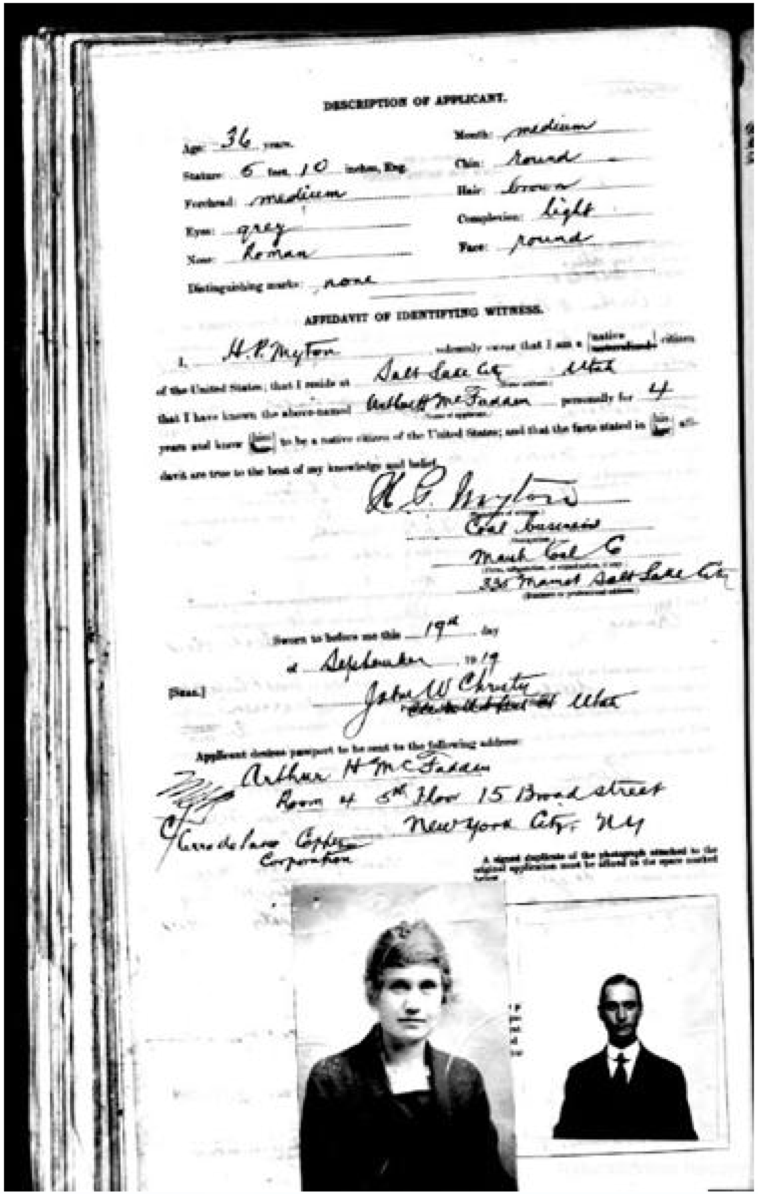

The "Quiver"
Page Thirty-one
MYRTLE CLEVELAND
GLADYS TANNER
GLADYS COOK
EVERETTE EDWARDS
NELLIE MORTON

User contributed content and structure

- 50 million family trees
- More than 5 billion profiles
- 200 million stories and photos

ancestry.com

Behavioral data



DESCRIPTION OF APPLICANT.

Age: 36 years.
Stature: 5 feet, 10 inches, Eng.
Forehead: medium
Eyes: grey
Nose: Roman
Mouth: medium
Chin: round
Hair: brown
Complexion: light
Face: round
Distinguishing marks: none

AFFIDAVIT OF IDENTIFYING WITNESS.

I, H. P. Myton solemnly swear that I am a native ~~naturalized~~ citizen of the United States; that I reside at Salt Lake City Utah; that I have known the above-named Arthur H McFaddan personally for 4 years and know him to be a native citizen of the United States; and that the facts stated in his affidavit are true to the best of my knowledge and belief.

H. P. Myton
Coal business
Mack Coal Co
330 Ma[illegible] Salt Lake City

Sworn to before me this 19th day
of September 1919

[Seal.]

John W Christy
~~[illegible]~~ Utah

Applicant desires passport to be sent to the following address:

Arthur H McFaddan
Room 4 5th Floor 15 Broad street
New York City, NY
Cerro de Pasco Copper Corporation

A signed duplicate of the photograph attached to the original application must be affixed in the space marked below.

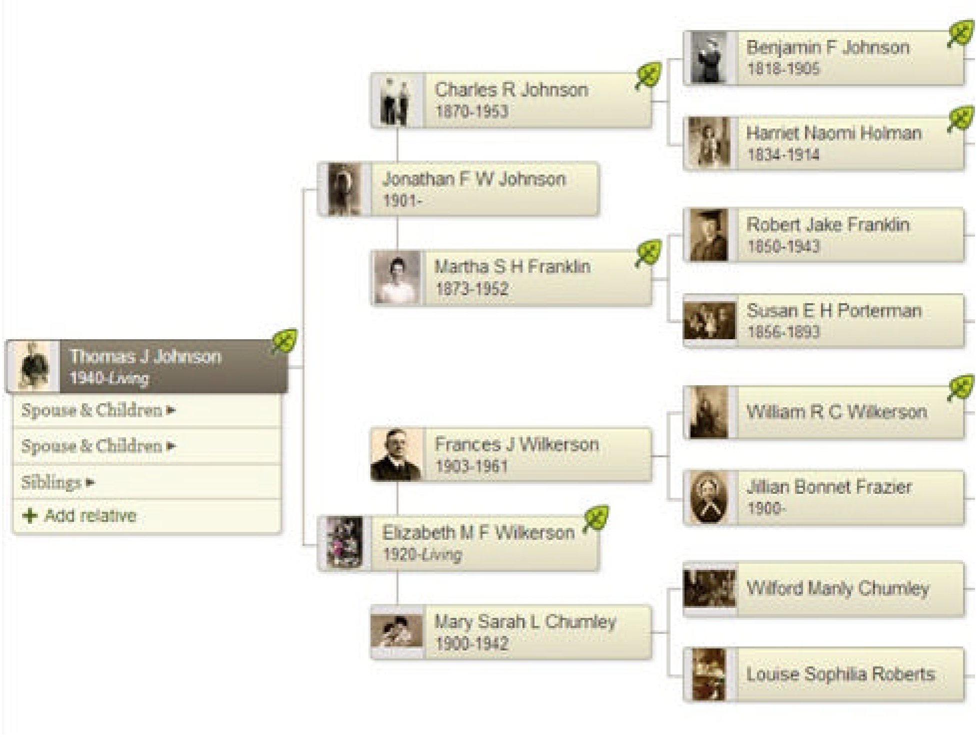
Next Best Discovery Algorithm

- 40 million searches/day
- 10 million people added to trees/day

Alamy

The math behind our big data equation

2,020 nodes
572 marriage edges
2,910 family edges



Thomas J Johnson
1940-*Living*

Spouse & Children ▸
Spouse & Children ▸
Siblings ▸
+ Add relative

Jonathan F W Johnson
1901-

Elizabeth M F Wilkerson
1920-*Living*

Charles R Johnson
1870-1953

Martha S H Franklin
1873-1952

Frances J Wilkerson
1903-1961

Mary Sarah L Chumley
1900-1942

Benjamin F Johnson
1818-1905

Harriet Naomi Holman
1834-1914

Robert Jake Franklin
1850-1943

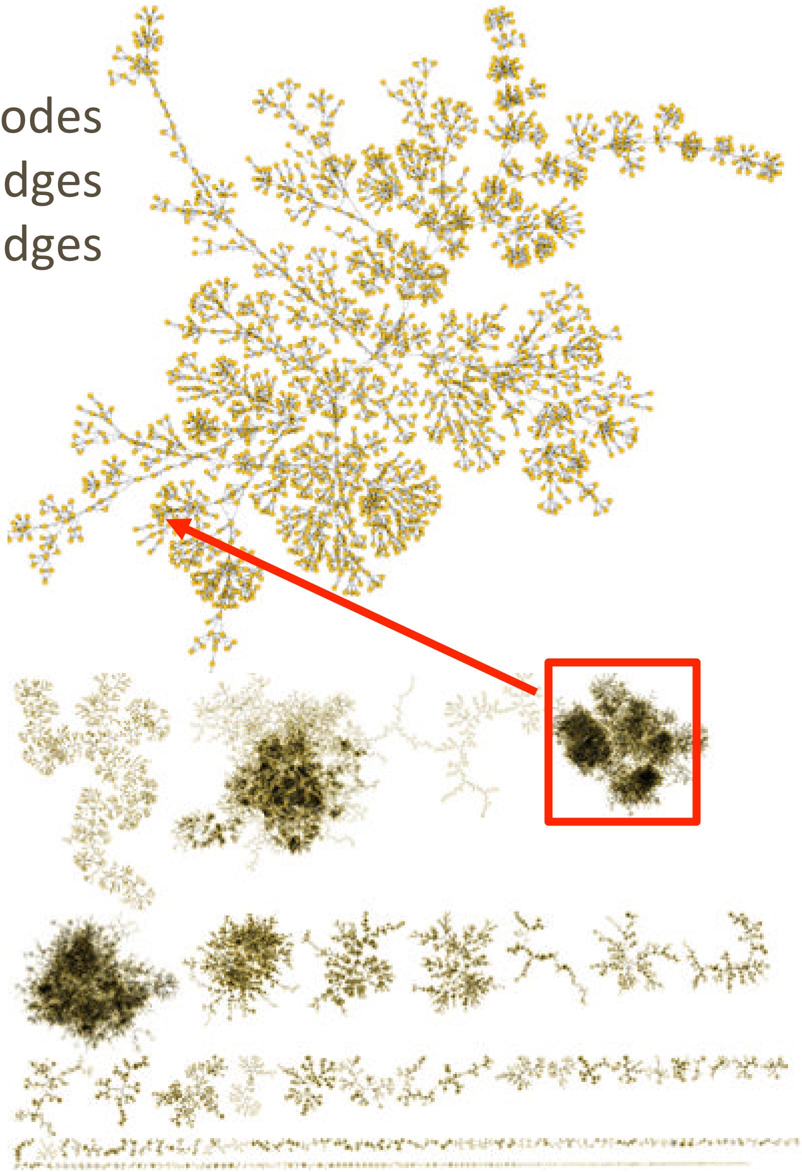
Susan E H Porterman
1856-1893

William R C Wilkerson

Jillian Bonnet Frazier
1900-

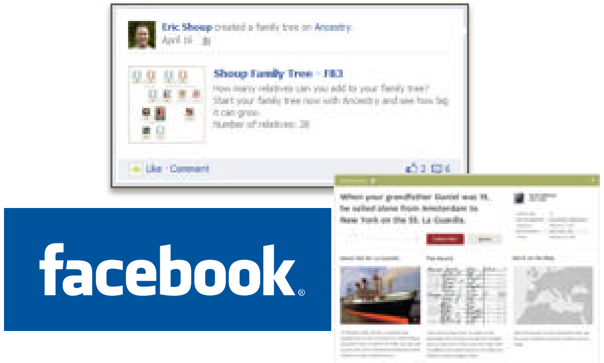
Wilford Manly Chumley

Louise Sophilia Roberts



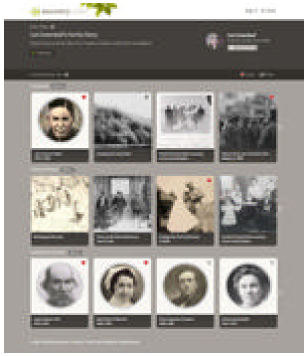
We've barely scratched the surface

- Making the site more social through sharing

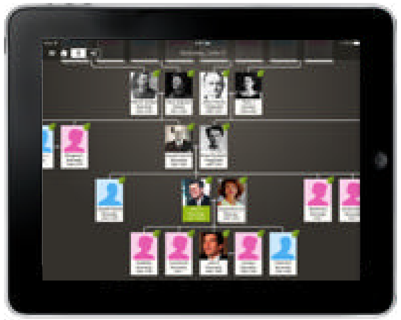

Eric Shoup created a family tree on Ancestry.
April 16

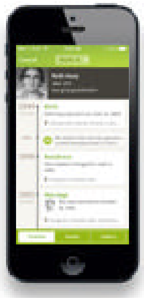
Shoup Family Tree - FB3
How many relatives can you add to your family tree? Start your family tree now with Ancestry and see how big it can grow.
Number of relatives: 28

When your grandfather Daniel was 15, he sailed alone from Amsterdam to New York on the SS. La Guardia.

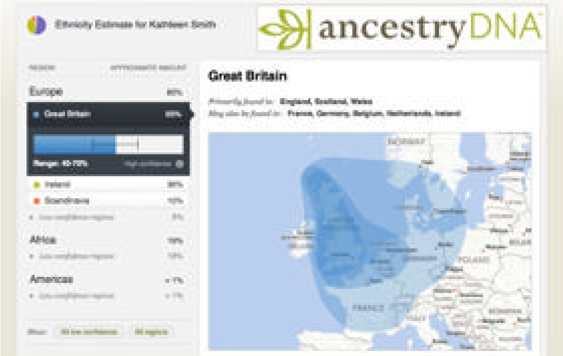


- Mobile extends the core users experience and attracts a new demographic





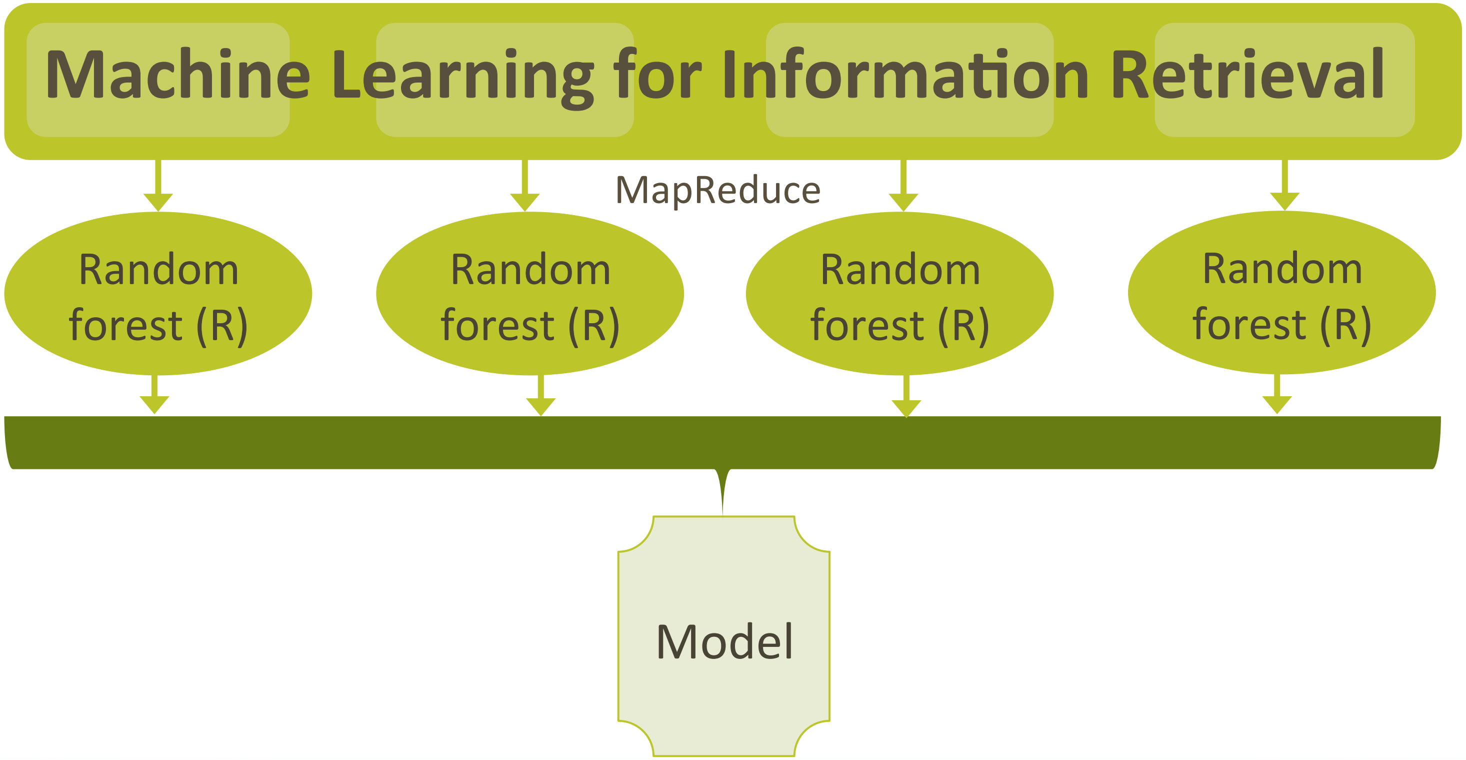
- New experiences like AncestryDNA


Ethnicity Estimate for Kathleen Smith

Great Britain

Primarily found in: England, Scotland, Wales
May also be found in: France, Germany, Belgium, Netherlands, Ireland

Our transition to Hadoop



Machine Learning for Information Retrieval

MapReduce

Random forest (R)

Random forest (R)

Random forest (R)

Random forest (R)

Model

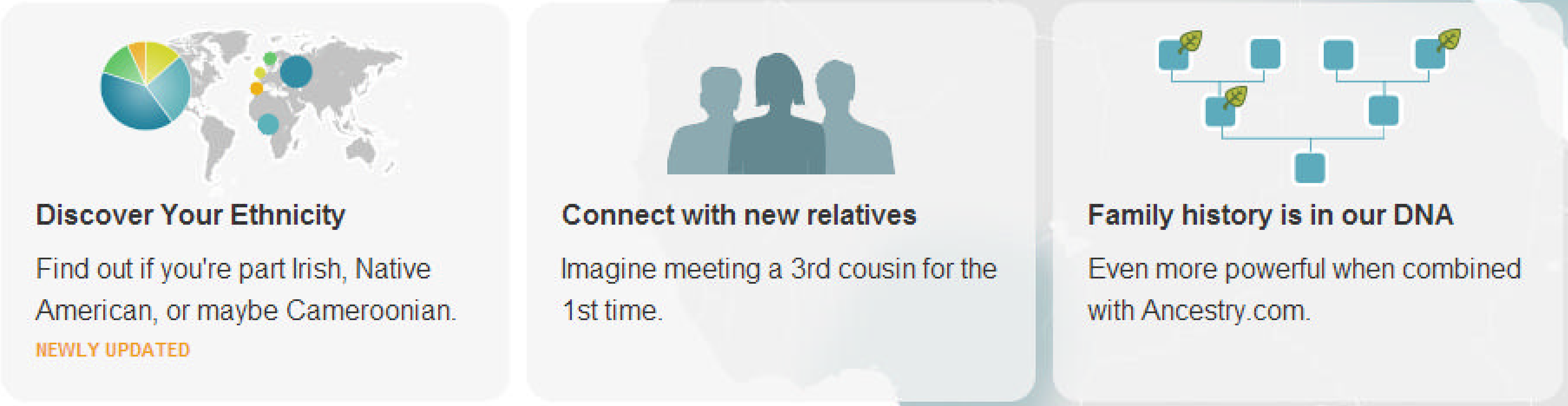
How we're using Hadoop

1. Machine Learning
2. Predictive analytics
3. Natural Language Processing and Entity Extraction
4. DNA Processing

AncestryDNA

Spit in a tube, pay $99, learn your past

- Autosomal DNA tests
- Samples from over 200,000 people
- 700,000 SNPs for each sample
- 10,000,000 4th cousin matches



Discover Your Ethnicity

Find out if you're part Irish, Native American, or maybe Cameroonian.

NEWLY UPDATED

Connect with new relatives

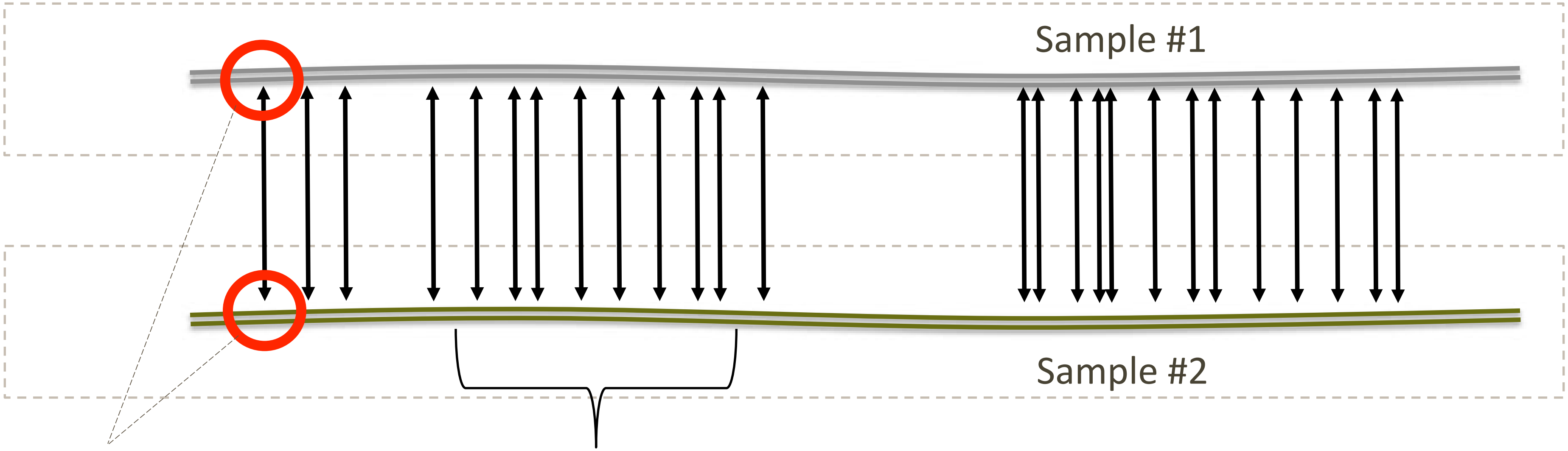
Imagine meeting a 3rd cousin for the 1st time.

Family history is in our DNA

Even more powerful when combined with Ancestry.com.

Estimating IBD (matching)

- We identify "long" DNA segments shared by two individuals.
- These segments are said to be Identical-by-Descent (IBD) and identify recent shared genetic ancestry.



Sample #1

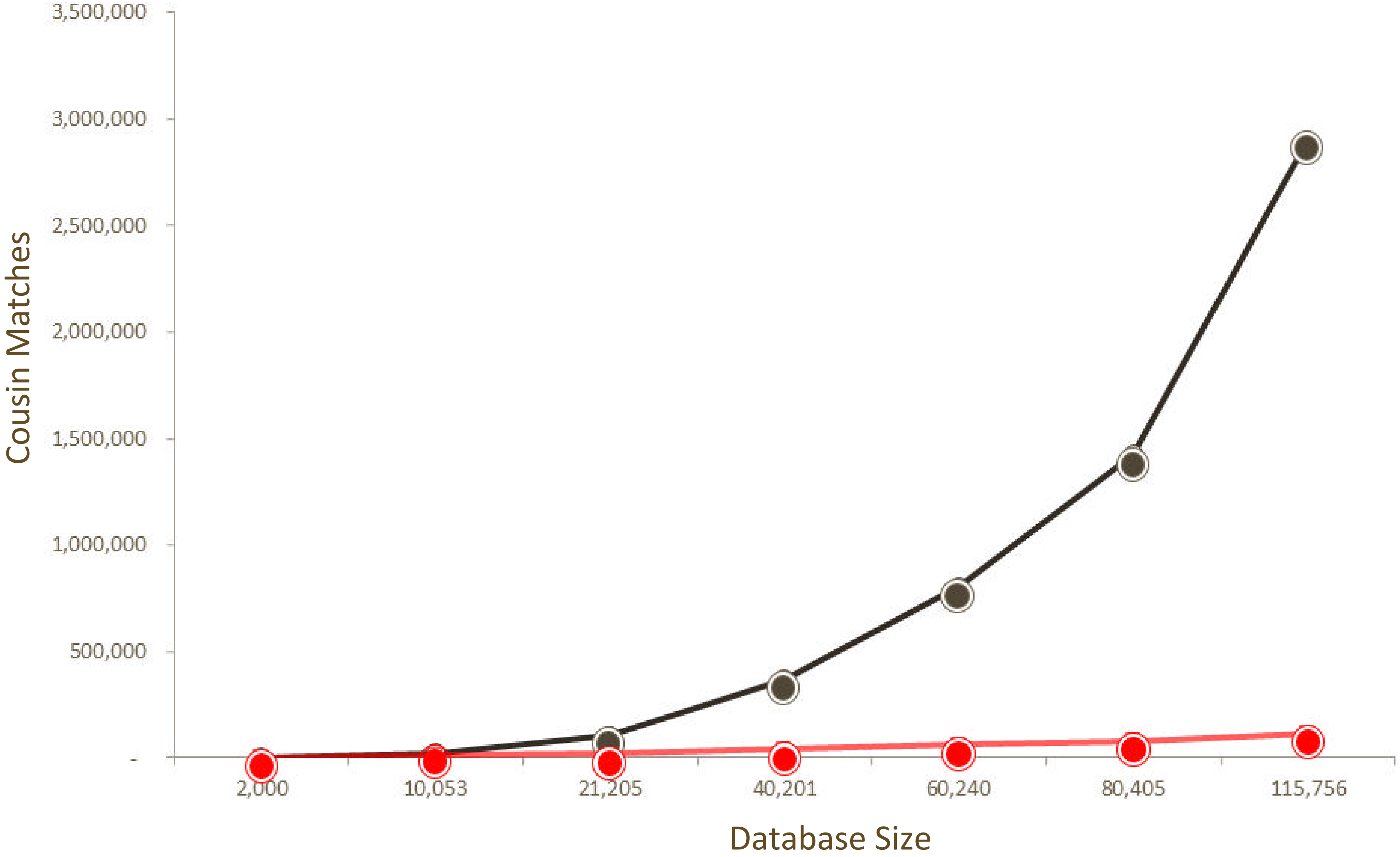
Sample #2

Sites essentially either match or they don't

IBD estimation is based on long sequences of consecutive matches

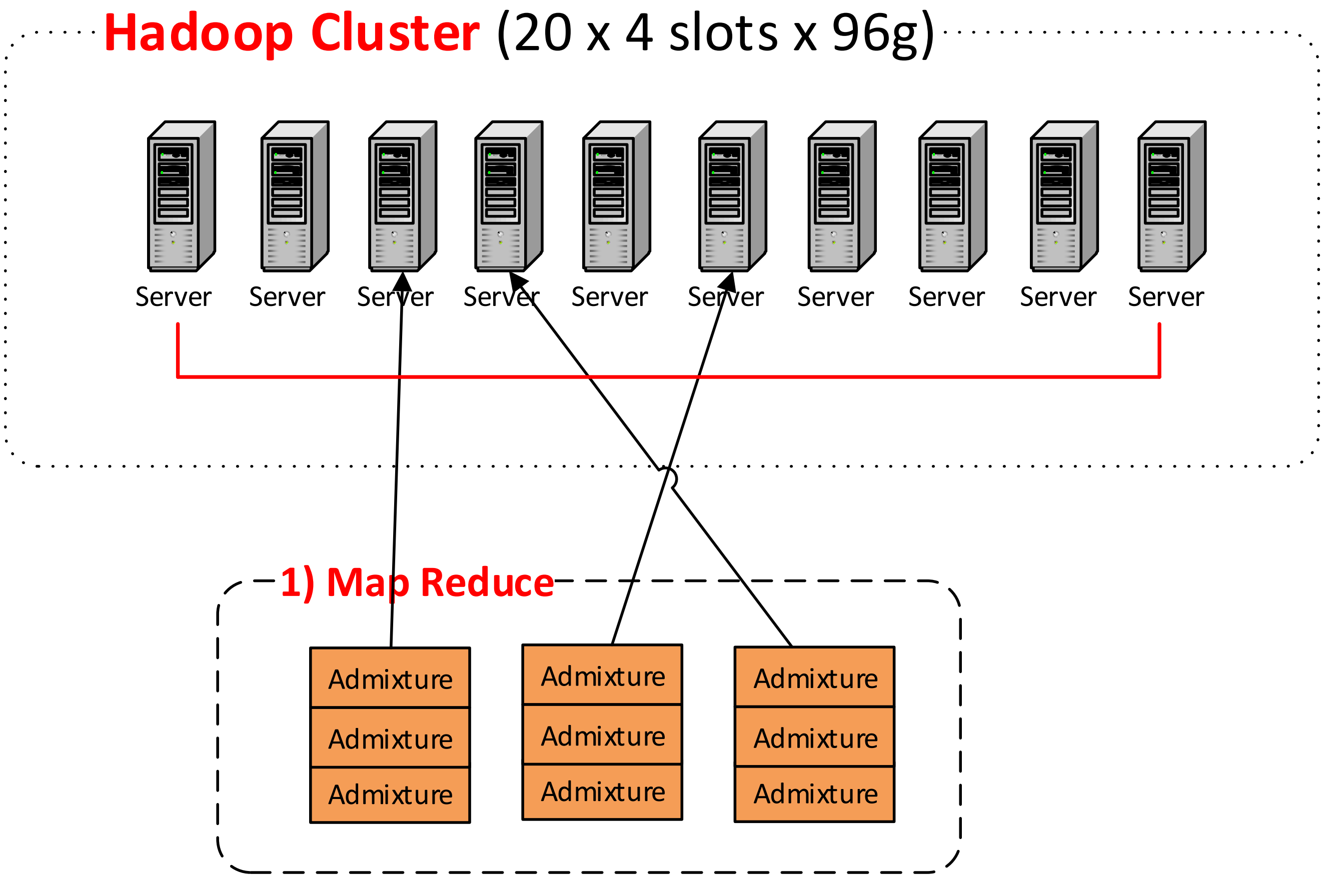
We have a statistical model, based on real pedigree data, that maps match lengths to relationship

Network effect & cousin matches

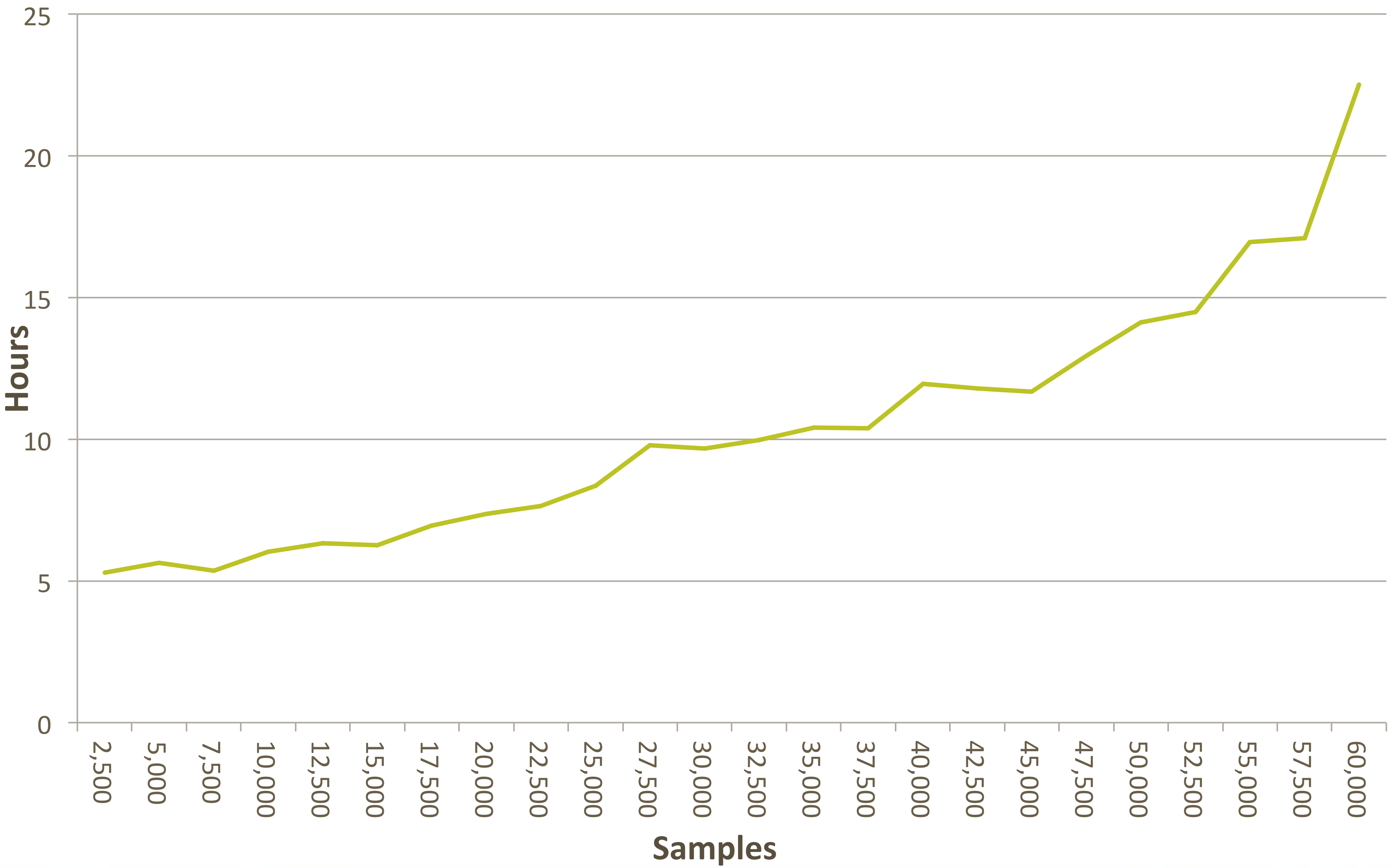


Algorithms in the pipeline

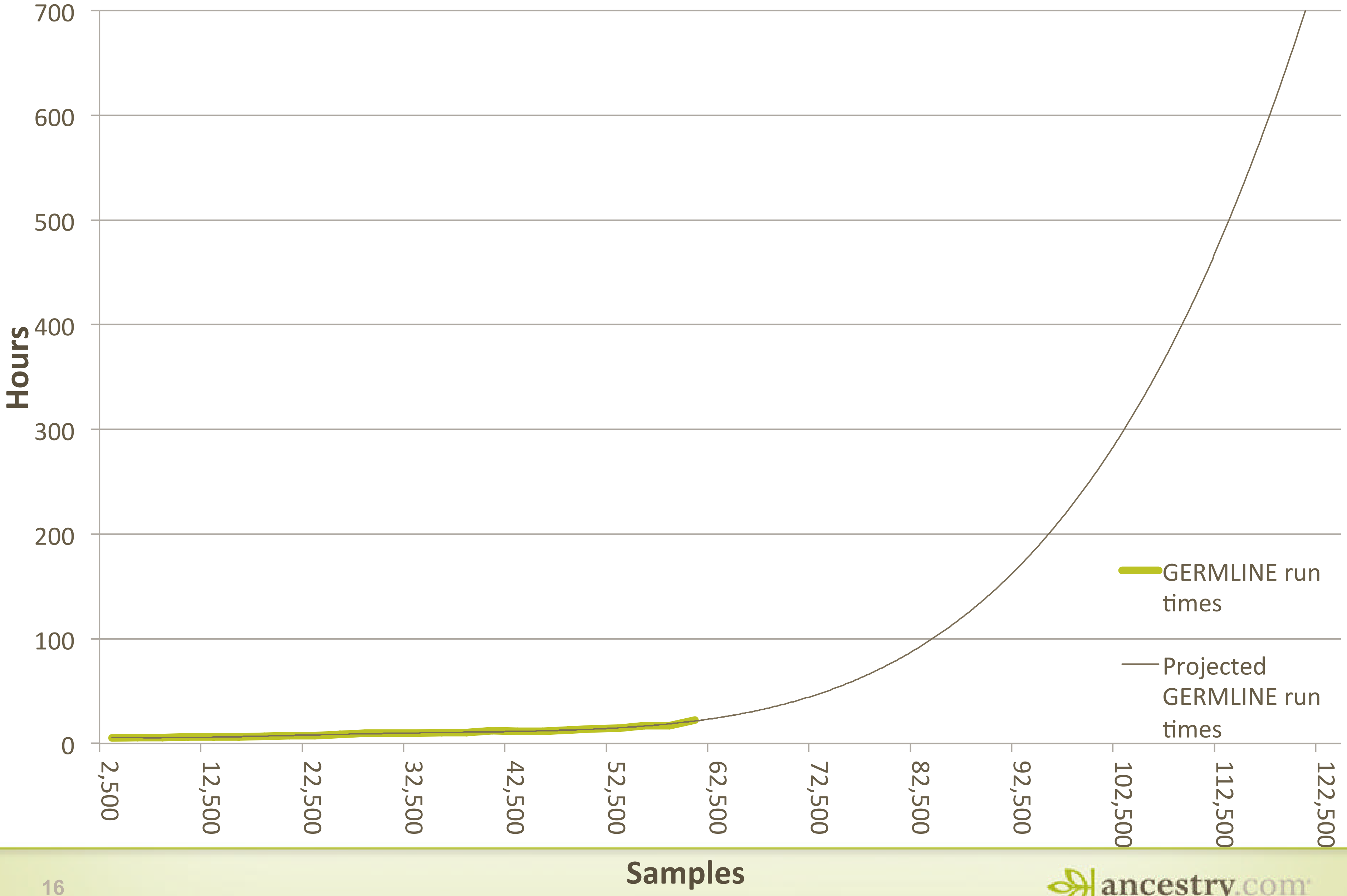
GERMLINE → Jermline



GERMLINE run times (in hours)



Projected GERMLINE run times (in hours)



DNA matching: How it works

The Input

Kara Thrace, aka Starbuck

- Ace viper pilot
- Has a special destiny
- Not to be trifled with

```
Starbuck : ACTGACCTAGTTGAC
Adama    : TTAAGCCTAGTTGAC
```

Admiral Adama

- Admiral of the Colonial Fleet
- Routinely saves humanity from destruction

DNA matching: How it works

Separate into words

0 1 2

Starbuck : ACTGA CCTAG TTGAC

Adama : TTAAG CCTAG TTGAC

DNA matching: How it works

Build the hash table

		0	1	2
Starbuck	:	ACTGA	CCTAG	TTGAC
Adama	:	TTAAG	CCTAG	TTGAC

ACTGA_0	:	Starbuck
TTAAG_0	:	Adama
CCTAG_1	:	Starbuck, Adama
TTGAC_2	:	Starbuck, Adama

DNA matching: How it works

Iterate through genome and find matches

	0	1	2
Starbuck :	ACTGA	CCTAG	TTGAC
Adama :	TTAAG	CCTAG	TTGAC

ACTGA_0 : Starbuck

TTAAG_0 : Adama

CCTAG_1 : Starbuck, Adama

TTGAC_2 : Starbuck, Adama

Starbuck and Adama match from position 1 to position 2

Does that mean they're related?

?

...maybe

But wait… what about Baltar?

Baltar : TTAAGCCTAGGGGCG

Gaius Baltar

- Handsome
- Genius
- Kinda evil

Step one : Update the hash table.

	Starbuck	Adama
2_ACTGA_0	1	
2_TTAAG_0		1
2_CCTAG_1	1	1
2_TTGAC_2	1	1

Already stored in HBase

Baltar : TTAAG CCTAG GGGCG

New sample to add

Key : [CHROMOSOME]_[WORD]_[POSITION]
Cell value : A byte set to 1, denoting that the user has that word at that position on that chromosome

Step two : Find matches, update the results table

	2_Starbuck	2_Adama
2_Starbuck		{ (1, 2), ...}
2_Adama	{ (1, 2), ...}	

Already stored in HBase

Baltar and Adama match from position 0 to position 1
Baltar and Starbuck match at position 1

New matches to add

Key : [CHROMOSOME]_[USER ID]
Cell value : A list of ranges where the two users match on a chromosome

The germline way

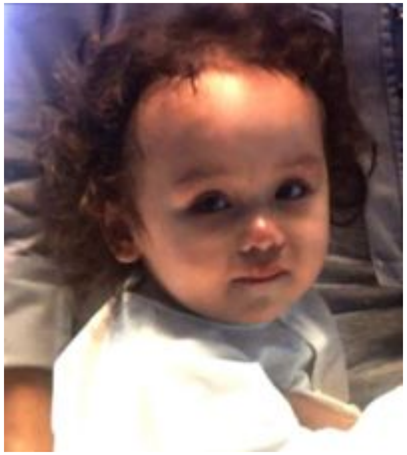
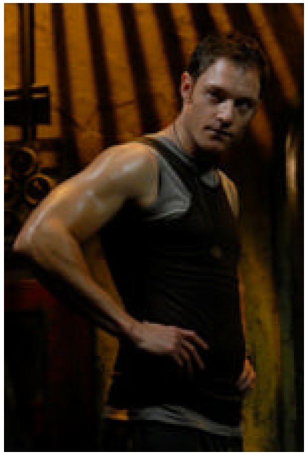
Hash Table

	Starbuck	Adama	Baltar
2_ACTGA_0	1		
2_TTAAG_0		1	1
2_CCTAG_1	1	1	1
2_TTGAC_2	1	1	
2_GGGCG_2			1

Results Table

	2_Starbuck	2_Adama	2_Baltar
2_Starbuck		{ (1, 2), ...}	{ (1), ...}
2_Adama	{ (1, 2), ...}		{ (0,1), ...}
2_Baltar	{ (1), ...}	{ (0,1), ...}	

But wait … what about Zarek, Roslin, Hera, and Helo?

Run them in parallel with Hadoop!

Photo by Benh Lieu Song

Parallelism with Hadoop

- Batches are usually about a thousand people.
- Each mapper takes a single chromosome for a single person.
- MapReduce Jobs :

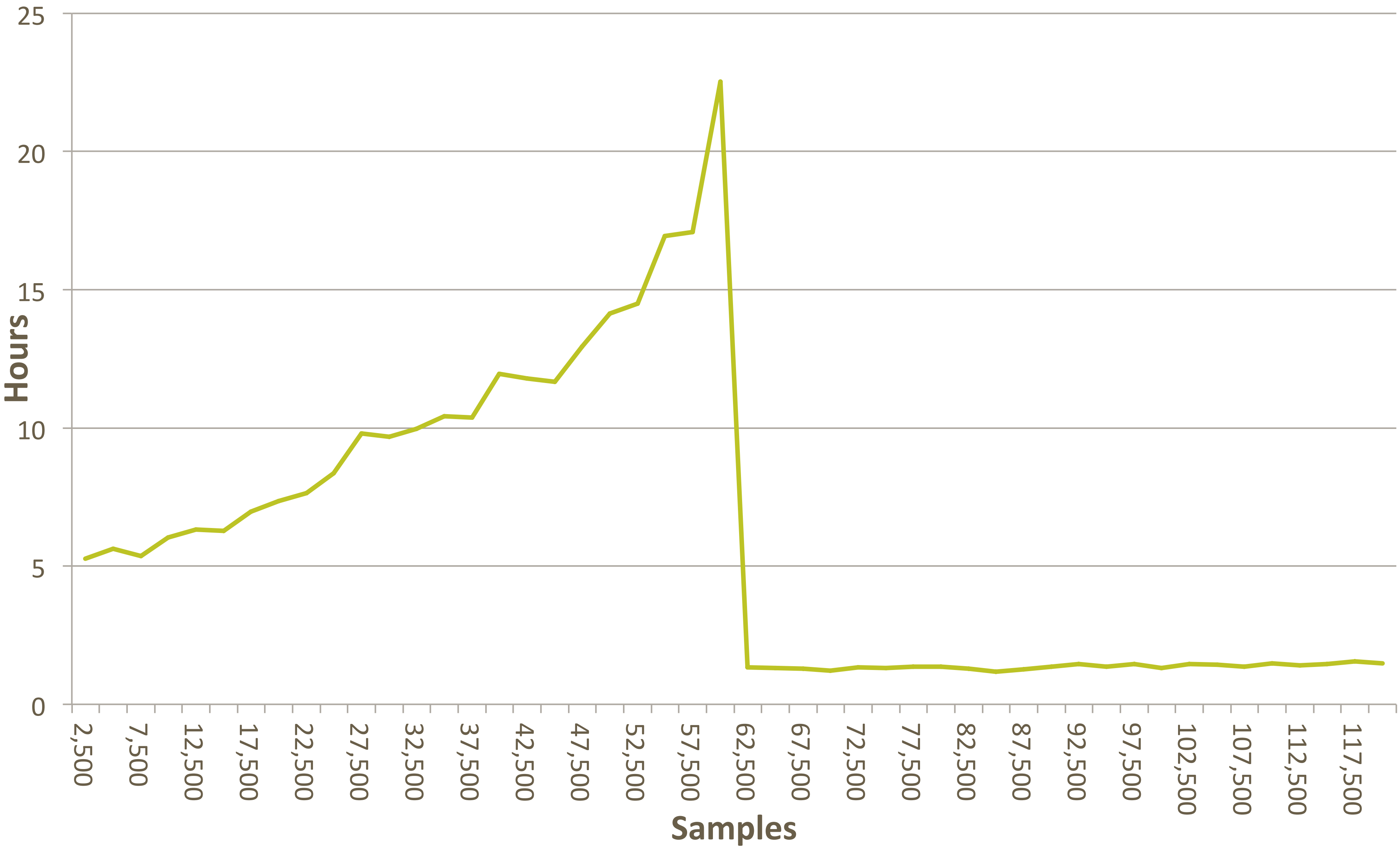
  Job #1 : Match Words

  - Updates the hash table

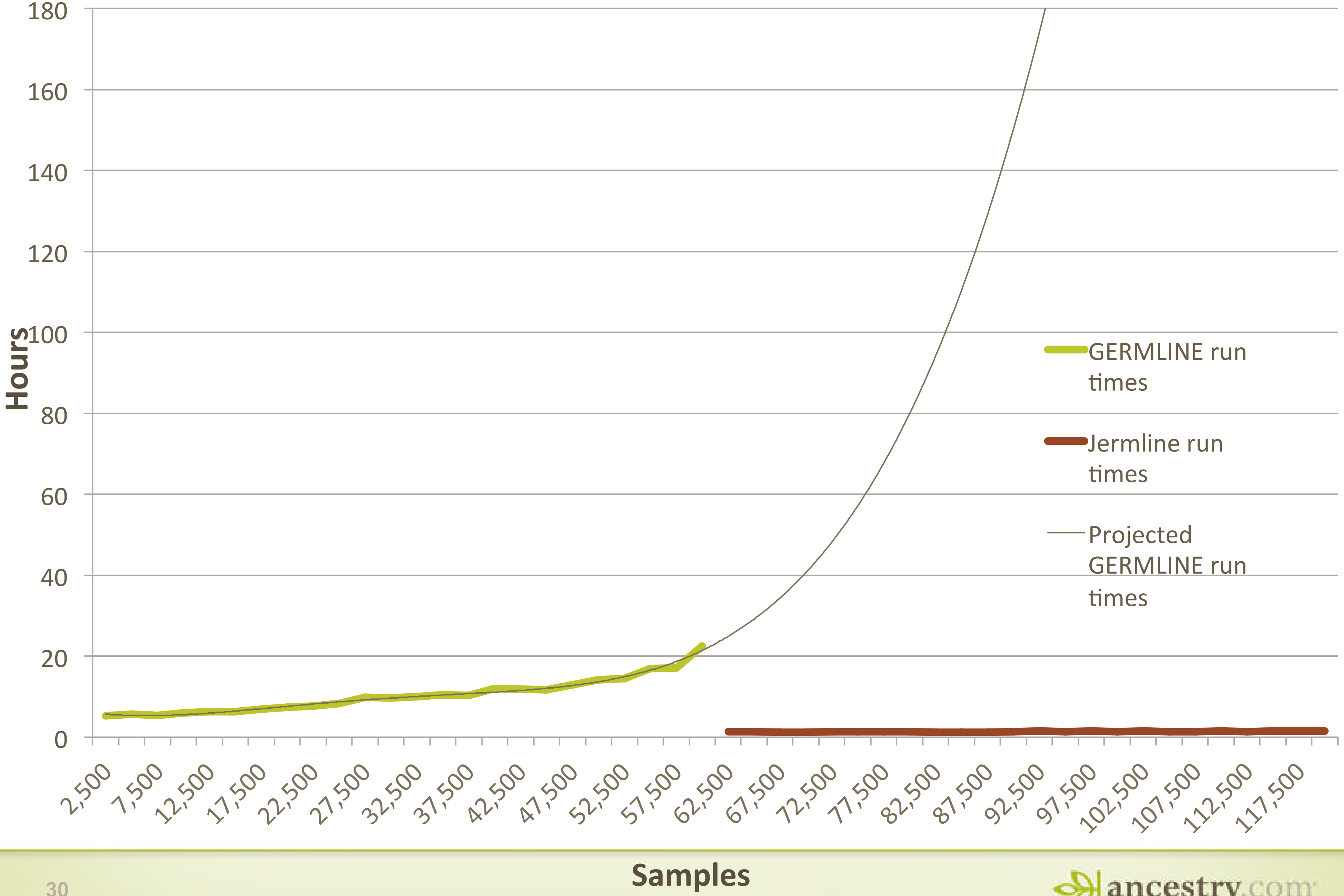
  Job #2 : Match Segments

  - Identifies areas where the samples match

Run times for matching (in hours)



Run times for matching (in hours)



AncestryDNA – Cast of characters

Scientists

Think they can code:

- Linux
- MySQL
- PERL and/or Python

Software Engineers

Think they are Scientists:

- Math
- Statistics
- Read science papers

Pressures of a startup business

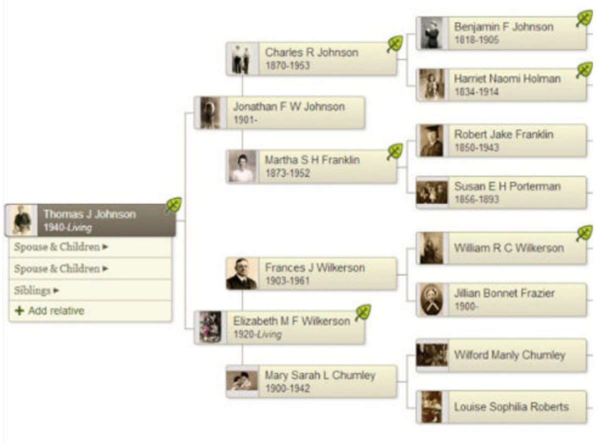
- Release a product, learn, and then scale

Other lessons learned

- Prototyping is key to overcoming resistance to change
- Technical architecture is heavily influenced by people organization
- Developing a team of experienced Hadoop users can often be done using internal employees
- A culture of experimentation and innovation yields the best results

Using Hadoop to drive scalable results

- Machine learning and predictive analytics
- Entity extraction and product development
- DNA pipeline processing



Thomas J Johnson
1940-Living
Spouse & Children ▸
Spouse & Children ▸
Siblings ▸
+ Add relative

Jonathan F W Johnson
1901-

Elizabeth M F Wilkerson
1920-Living

Charles R Johnson
1870-1953

Martha S H Franklin
1873-1952

Frances J Wilkerson
1903-1961

Mary Sarah L Chumley
1900-1942

Benjamin F Johnson
1818-1905

Harriet Naomi Holman
1834-1914

Robert Jake Franklin
1850-1943

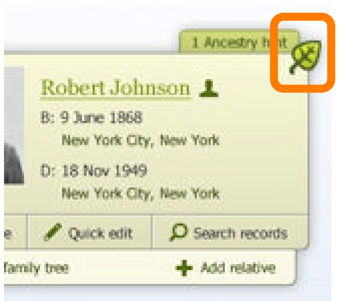
Susan E H Porterman
1856-1893

William R C Wilkerson

Jillian Bonnet Frazier
1900-

Wilford Manly Chumley

Louise Sophia Roberts

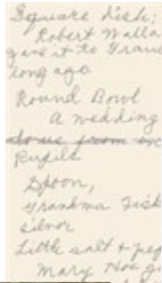


1 Ancestry hint

Robert Johnson

B: 9 June 1868
New York City, New York

D: 18 Nov 1949
New York City, New York



Square dish;
Robert Wallace's mother
gave it to Grandma Ryan
long ago
Round Bowl
a wedding present
~~from~~
Ryll
Spoon,
grandma
silver
Little salt & pep
Mary

Questions?

Tech Roots blog - http://blogs.ancestry.com/techroots/