

# Data Mining of Online Genealogy Datasets for Revealing Lifespan Patterns in Human Population

Michael Fire and Yuval Elovici\*

Telekom Innovation Laboratories at Ben-Gurion University of the Negev  
Department of Information Systems Engineering, Ben-Gurion University

January 5, 2014

## Abstract

Online genealogy datasets contain extensive information about millions of people and their past and present family connections. This vast amount of data can assist in identifying various patterns in human population. In this study, we present methods and algorithms which can assist in identifying variations in lifespan distributions of human population in the past centuries, in detecting social and genetic features which correlate with human lifespan, and in constructing predictive models of human lifespan based on various features which can easily be extracted from genealogy datasets.

We have evaluated the presented methods and algorithms on a large online genealogy dataset with over a million profiles and over 9 million connections, all of which were collected from the WikiTree website. Our findings indicate that significant but small positive correlations exist between the parents' lifespan and their children's lifespan. Additionally, we found slightly higher and significant correlations between the lifespans of spouses. We also discovered a very small positive and significant correlation between longevity and reproductive success in males, and a small and significant negative correlation between longevity and reproductive success in females. Moreover, our machine learning algorithms presented better than random classification results in predicting which people who outlive the age of 50 will also outlive the age of 80.

We believe that this study will be the first of many studies which utilize the wealth of data on human populations, existing in online genealogy datasets, to better understand factors which influence human lifespan. Understanding these factors can assist scientists in providing solutions for successful aging.

**Keywords.** Genealogy Data Mining, Aging, Gerontology, Human Population Lifespan, Lifespan Prediction, Date Mining, Machine Learning, WikiTree

## 1 Introduction

In the last decade, Web 2.0 websites, such as Wikipedia<sup>1</sup> and Reddit,<sup>2</sup> have become extremely popular and widespread. Web 2.0 websites offer Internet users opportunities

---

\*Email:{mickyfi,elovici}@bgu.ac.il

<sup>1</sup><http://en.wikipedia.org>

<sup>2</sup><http://www.reddit.com>

to connect, collaborate, and share information with each other, creating massive datasets with millions of content items. For example, the free encyclopedia Wikipedia has more than 4.3 million articles and more than 127,000 active users who contribute new content to the site on a regular basis [24]. One type of Web 2.0 site which recently became popular is genealogy websites. Genealogy websites, such as MyHeritage,<sup>3</sup> Ancestry,<sup>4</sup> WikiTree,<sup>5</sup> and Familiypedia,<sup>6</sup> have millions of users [16, 2] that use these websites to create, discover, and share their family history by generating and updating online family trees. These online family trees consist of personal data on family members from the last several centuries, and they give many personal details for each family member, such as the member’s date of birth, date of death, ancestors’ details, and children’s details, among others.

The family tree structure and the family members’ personal details that are stored in these genealogy websites create large-scale datasets, which contain billions of entries [10, 2] on human life and death properties. These datasets can be utilized to reveal interesting patterns regarding lifespan changes over the centuries. Additionally, these datasets can also assist in better understanding and identifying characteristics which are correlated with human lifespan changes. For example, these datasets can be explored and utilized to answer the following questions: *Does having more children extend one’s lifespan? Does having long-lived ancestors prolong life? Does getting married lengthen one’s lifespan?* Answering these types of questions can assist scientists in providing insights and solutions for successful aging.

In this study, we present data mining algorithms for analyzing large genealogy datasets in order to examine human population lifespan variations over a substantial length of time (see Section 3.3.1). Moreover, we introduce methods to utilize these types of datasets to identify features which correlate with human lifespan (see Section 3.3.2). Additionally, we also present Machine Learning (ML) algorithms based on features extracted from genealogy datasets, which can assist in predicting if a particular 50-year-old individual will reach the age of 80 (see Section 3.3.4).

To test and evaluate our algorithms, we developed a web crawler which crawled and parsed public profiles from the WikiTree website. WikiTree is a free, collaborative family-history website, which contains more than 5 million user-contributed profiles [5] of individuals who have lived in the past centuries, and many of the profiles contain personal details about each individual. Using the collected data from WikiTree, we were able to construct a dataset (referred to as the WikiTree dataset) of over a million public profiles, out of which at least 416,030 profiles were of individuals who were born in the United States (see Section 3.4).

By analyzing the WikiTree dataset, we calculated various statistics on variations of population lifespan over the last centuries, including specific statistics on the lifespan variations of the United States population (see Section 3.3 and Figures 3 and 4). As a result of this analysis, we discovered several interesting historical lifespan change patterns (see Section 5); for example, we discovered that the average lifespan of females who were born in the United States and lived beyond the age of ten increased sharply in just a half-century: from 62.66 in 1850 to 72.5 in 1900 (see Figure 3).

Using the WikiTree dataset, we constructed a social network directed multigraph which contains over 1.38 million vertices and over 9.19 million links (see Section 3.1 and Table 3). We then analyzed the social network graph and extracted 21 features, such as parents’

---

<sup>3</sup><http://en.wikipedia.org>

<sup>4</sup><http://www.ancestry.com>

<sup>5</sup><http://www.wikitree.com>

<sup>6</sup><http://familiypedia.wikia.com>

and grandparents' ages of death, for each vertex in the graph (see Section 3.2). By using the extracted features and simple linear regression models, we discovered significant correlations with low coefficients of determination between the individuals' ages of death and the ages of death of their siblings, parents, spouses, and grandparents (see Table 5). We also discovered a slightly higher significant correlation between the individual's age of death and the age of death of his or her spouse (see Table 5). Additionally, we constructed multiple linear regression models for predicting an individual's age of death based on various features which were extracted from the individual's personal details. Our multiple linear models were with high significance and Multiple Adjusted R-squared values up to 0.085 (see Table 6).

Our ML classifiers have presented better than random results in predicting which individuals who outlived the age of fifty and passed the age of menopause will also outlive the age of 80 (see Section 4.3).

The remainder of the paper is structured as follows: In Section 2 we give a brief overview of previous relevant studies on characteristics which were found to be correlated with human lifespan. In this section, we also introduce several studies which used similar data mining algorithms as this study. Next, in Section 3 we present the methods and algorithms we developed for studying genealogy datasets. In this section, we also describe our constructed WikiTree dataset. Then, in Section 4 we present our algorithm evaluations results on the WikiTree dataset. Lastly, in Section 5 we discuss our results, and we also offer future research directions.

## 2 Related Work

The factors that influence human lifespan have been thoroughly studied over the past decades [20, 14, 13, 11, 8]. In this section we give a brief overview of recent genealogical studies that are most relevant to this study, pinpointing similar factors. Additionally, we also give a short overview of recent studies in the field of social network analysis and data mining, which used a similar methodology to the one used throughout this study.

In recent years, many studies have tried to find correlations between parents' and childrens' lifespans, as well as correlations between lifespans of parents and their number of children: In 1998, Westendorp and Kirkwood used a historical dataset, from the British aristocracy, to study the connection between longevity and reproductive success. They discovered that longevity was positively correlated with age at first childbirth, and negatively correlated with number of children. In 2000, Thomas et al. [20] studied the connection between longevity and fertility using a statistical dataset of 153 countries. They concluded that "humans who invest heavily in reproduction while young will, on average, pay for this reproductive success with a shortened lifespan." In 2001, Mitchell et al. [14] used genealogical data of Old Order Amish members to estimate the parent-child correlations in lifespan. They also estimated the child age of death as a function of parent age at death. They discovered significant but small correlations between parental and child ages at death.

In 2006, McArdle et al. [13] studied the correlation between the number of children and lifespan using genealogical data of 2,015 individuals who were members of an Old Order Amish community. In their study they discovered lifespans of fathers increased linearly with increasing number of children, while lifespans of mothers increased linearly up to 14 children but decreased with each additional child beyond 14. In 2007, Le Bourg [11] presented a thorough review of studies which researched the relationship between fertility and longevity under various conditions. According to Le Bourg, the review results indi-

cated that “in natural fertility conditions longevity does not decrease when the number of children increases but, in modern populations, mortality could slightly increase when women have more than ca 5 children.” In 2011, Gögele et al. [8] conducted a comprehensive genealogical study with a thorough assessment of the heritability of lifespan and longevity in three villages in Italy. Their research, which included studying more than 50,000 individuals across four centuries, discovered “a general low inheritance of human lifespan, but which increases substantially when considering long-living individuals, and a common genetic background of lifespan and reproduction.”

Many studies found connections between an excess in mortality and bereavement, also known as the “widow effect.” In 1969, Parkes et al. [18] followed 4,486 widowers at the age of 55 for nine years. Out of these widowers, 213 died during their first six months of bereavement, 40% above the expected rate for married men of the same age. In 1996, Martikainen et al. [12] conducted a large scale study of 1,580,000 married Finnish individuals and also discovered excess mortality among the bereaved. In 2008, Elwert and Christakis [6] studied 373,189 elderly married couples in the United States. They discovered that the death of a spouse from almost all causes increased the mortality of the bereaved partner to varying degrees.

In our research we used several regression and ML techniques for lifespan prediction. In order to carry out our work, we mainly used attributes which could be extracted from genealogy datasets in order to construct the genealogy social network and extract features from the network (see Section 3). Similar techniques that involve social network analysis and regression were used by Christakis [4] in researching the spread of obesity, by Altshuler et al. [1] in predicting the individual parameters and social links of smart-phone users, and by Fire et al. [7] in predicting students’ final exam scores.

### 3 Methods and Experiments

To cope with the challenge of analyzing a huge online genealogy dataset with ten of millions of records on individuals’ personal data and their connections, we first chose to convert the dataset into a social network represented by a directed multigraph where vertices represent people and links represent connections among family members (see Section 3.1). Next, we used the constructed social network graph and extracted various features from each vertex, such as the vertex’s number of children, year of birth, and gender (see Section 3.2). We then used the extracted features to determine various statistics on the population lifespan variations over time (see Section 3.3.1). After that, we used linear regression to find the features that significantly influence human lifespan. We also constructed multi-linear regression models for lifespan prediction (see Section 3.3.2). Lastly, we used ML algorithms to construct classifiers which can predict if a person from the United States who outlives the age of fifty will also reach the age of eighty (see Section 3.3.4).

To perform our statistical calculations and to construct our predictive models, we used various datasets that were extracted from a large genealogy dataset. These datasets are defined in Table 1. Additionally, the methods and algorithms we have used throughout this study are summarized in Table 2.

Table 1: Dataset Definitions

Dataset Name	Description	Formal Definition
<b>All-Dataset</b>	All the vertices with valid Age-of-Death feature values.	$\{v \in V   0 \leq \text{Age} - \text{of} - \text{Death}(v) \leq 122\}$
<b>All-Dataset-&lt;Age&gt;</b>	All the vertices with valid Age-of-Death feature values which outlived the selected <Age>.	$\{v \in \text{All-Dataset}   \text{Age} - \text{of} - \text{Death}(v) \geq <\text{Age}>\}$
<b>&lt;Country&gt;-Dataset</b>	All the vertices born in <Country> with valid Age-of-Death feature values.	$\{v \in \text{All-Dataset}   \text{Birth-Country}(v) = <\text{Country}>\}$
<b>&lt;Country&gt;-Dataset-&lt;Age&gt;</b>	All the vertices born in <Country> with valid Age-of-Death feature values which outlived the selected <Age>.	$\{v \in \text{All-Dataset}-<\text{Age}>   \text{Birth-Country}(v) = <\text{Country}>\}$
<b>&lt;Gender&gt;- Dataset</b>	All the vertices with valid Age-of-Death feature values, and Gender feature value equal to male (=1) or to female (=2).	$\{v \in \text{All-Dataset}   \text{Gender}(v) = <\text{Gender}>\}$
<b>&lt;Gender&gt;- Dataset-&lt;Age&gt;</b>	All the vertices with valid Age-of-Death feature values which outlived the age of <Age>, and have Gender feature values equal to male (=1) or female (=2).	$\{v \in <\text{Gender}>-\text{Dataset}   \text{Age} - \text{of} - \text{Death}(v) \geq <\text{Age}>\}$
<b>&lt;Gender&gt;-&lt;Country&gt;-Dataset-&lt;Age&gt;</b>	All the vertices with valid Age-of-Death feature who outlived the selected <Age>, were born in <Country>, and the Gender feature values are equal only to one value (<Gender>): Male (=1) or Female (=2).	$\{v \in \text{Gender-Dataset}   \text{Birth-Country}(v) = <\text{Country}> \wedge \text{Age} - \text{of} - \text{Death}(v) \geq <\text{Age}>\}$
<b>&lt;Feature&gt;-Dataset-&lt;Age&gt;</b>	All the vertices which have valid selected feature (<Feature>) values and Age-of-Death feature values of at least the selected age (<Age>).	$\{v \in \text{All-Dataset}-<\text{Age}>   \exists <\text{feature}>(v)\}$
<b>&lt;Gender&gt;-&lt;Feature&gt;-Dataset-&lt;Age&gt;</b>	All the vertices which have valid selected feature (<Feature>) values, Gender feature values which are equal to only one value (<Gender>): Male (=1) or Female (=2), and Age-of-Death feature values above or equal to the selected age (<Age>).	$\{v \in \text{All-Dataset}-<\text{Age}>   \exists <\text{feature}>(v) \wedge \text{Gender}(v) = <\text{Gender}>\}$
<b>Married-Dataset</b>	All the vertices with valid Age-of-Death feature values, and Spouse-Number of at least 1.	$\{v \in \text{All-Dataset}   \text{Spouse-Number}(v) \geq 1\}$
<b>No-Missing-Dataset-&lt;Age&gt;</b>	All the vertices without missing numeric features, and valid Age-of-Death feature values which are at least the selected <Age> .	$\{v \in \text{All-Dataset}-<\text{Age}>   \forall \text{feature} \in \text{All-Numeric-Features}, \exists \text{feature}(v) \in \mathcal{R}\}$

Table 2: Method and Algorithm Overview

Algorithm Goal	Datasets	Method
<b>Present lifespan variations over time</b>	<i>All-Dataset</i> <i>United-States-Dataset</i>	Calculate the Age-of-Death distribution of all people who were born in each quarter of a century between 1650 and 1900.
<b>Median and average lifespan calculation over time</b>	<i>All-Dataset-10</i> <i>United-States-Dataset-10</i> <i>Male-Dataset-10</i> <i>Female-Dataset-10</i> <i>Male-United-States-Dataset-10</i> <i>Female-United-States-Dataset-10</i>	Calculate, for each dataset, both the average and median lifespans for people who were born in each year between 1650 and 1900, and outlived the age of 10.
<b>Identify inheritance of human lifespans</b>	<i>&lt;Feature&gt;-Dataset-10</i>	Construct a simple linear regression between each selected feature in the extended family features and the Age-of-Death feature.
<b>Identify correlation between spouses' lifespan</b>	<i>Married-Dataset</i>	Construct a simple linear regression between each one of the Spouse-Age-of-Death features (Min/Max/Average), and the Age-of-Death feature.
<b>Identify if longevity is correlated with reproductive success</b>	<i>Children-Number-Dataset-50</i> <i>Male-Children-Number-Dataset-50</i> <i>Female-Children-Number-Dataset-50</i>	Construct a simple linear regression between the Children –Number feature and the Age-of-Death feature on each dataset.
<b>Construct models for predicting individuals' lifespans</b>	<i>No-Missing-Dataset-10</i> <i>No-Missing-Dataset-50</i>	Construct backward stepwise multiple linear regression models for predicting the Age-of-Death using All-Numeric-Features, the Heritage-Features, and the Nuclear-Family-Features sets.
<b>Predict if an individual will outlive the age of 80</b>	<i>United-States-Dataset-50</i> (only vertices in which the birth year was between 1650 and 1900)	Construct various machine learning classifiers for predicting if an individual born in the United States, and outlived the age of 50, will also reach the age of 80.

### 3.1 Constructing the Genealogy Social Networks

We constructed the social network directed multigraph  $G := \langle V, E \rangle$  from the genealogy dataset in the following manner: First, we assembled the graph vertices set  $V$  by adding a new vertex  $v \in V$  for each profile in the genealogy dataset. We then defined  $E$  as the multiset of links in the graph, with each link  $e \in E$  defined to be a tuple  $e = (u, v, t, c) \in E$ , where  $u, v \in V$ ;  $t$  is the link type, which can be one of the following values:  $t \in \{\text{Spouse}, \text{Child}, \text{Parent}, \text{Sibling}\}$ ; and  $c$  is the creation date of the link. For example, if the genealogy dataset contains the profiles of Queen Elizabeth II and Prince Charles, then the social network graph will contain the following vertices: Elizabeth II, Prince Charles  $\in V$ , and the following edges: (Elizabeth II, Prince Charles, Parent, 14 November 1948) and (Prince Charles, Elizabeth II, Child, 14 November 1948). In case the genealogy dataset contains link of type  $t$  between a public profile  $u$  and a private profile, we added to the multigraph a new vertex  $v_{private}$  to  $V$  and added new link  $e = (u, v_{private}, t, \emptyset)$  to  $E$ .<sup>7</sup> Next, for each non-private vertex  $v \in V$  in the network, we added attributes based on the information extracted from each individual's profile page, which is represented by  $v$ .<sup>8</sup> Then, for each vertex  $v$ , we calculated a list of features described in the following subsection. Lastly, we removed the birthday and death date data of any vertices in the graph with the following inconsistencies in their data: (a) profiles with a negative age of death, which can occur from reversing the birth and death dates; (b) profiles shown to be over the age of 122, the maximum recorded age [23]; and (c) profiles shown as having children and also an age of less than five.<sup>9</sup>

### 3.2 Feature Extraction

After constructing the social network graph, we can extract, if possible, three types of features for each vertex: The first type is the vertex *general profile features*, which include basic information about the vertex, such as birth year, gender, and full name. The second type is the *nuclear family features*, which include information about the vertex's children and spouses. The third and last type of features is the *extended family features*, which include information about the vertex's parents, siblings, and grandparents. In this study, we extracted a total of 21 features for each vertex  $v \in V$ . In the remainder of this section, we introduce and give formal definitions for each one of these features.

#### 3.2.1 General Features

1. **Full-Name( $v$ )** - The full name of  $v$ .
2. **Birth-Year( $v$ )** - The birth year of  $v$ .
3. **Death-Year( $v$ )** - The death year of  $v$ .

---

<sup>7</sup>During this study, we have utilized private profiles to calculate public profiles' features, such as Children-Number( $u$ ) and Spouse-Number( $u$ ) more accurately (see Section 3.2). In many cases, we cannot distinguish if two or more private profiles are in fact represent the same single profile in the genealogy dataset. Nevertheless, we can estimate the number of distinct private profiles by utilizing the private profile single link of type  $t$ . Namely, due to the fact that most people have two parents, we can conclude that  $n$  private profiles with single link of type "Child" represent at least  $\frac{n}{2}$  distinct profiles.

<sup>8</sup>In most online genealogy websites, a profile usually contains the following information about each individual: gender, birth and death dates, location of birth, location of death, parents' names, spouses' names, siblings' names, and children's names.

<sup>9</sup>The youngest mother on record was a 5-year-old Peruvian girl [15].

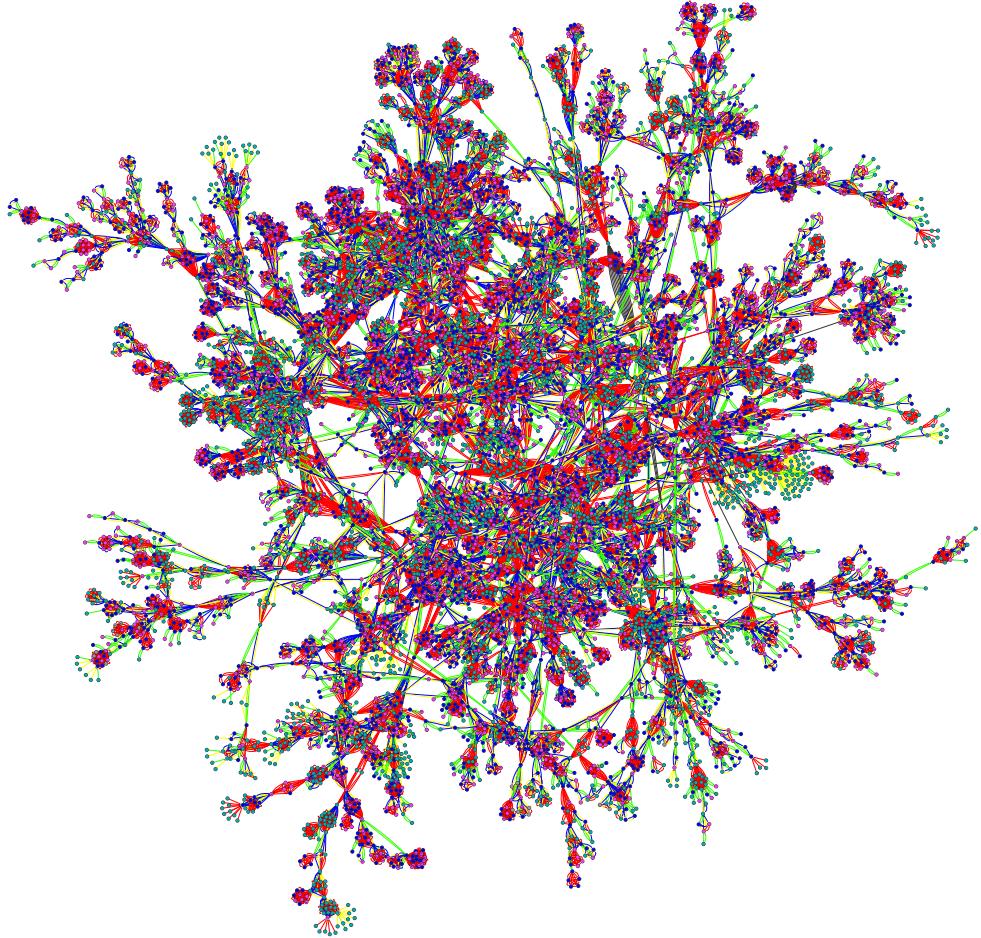


Figure 1: **WikiTree’s directed multigraph (subgraph with 10,000 vertices)**. Each link color represents a connection of a different type: *blue* is a Parent link, *red* is a Sibling link, *green* is a Spouse link, and *yellow* is a Child link. The color of each vertex defines the vertex’s gender: *blue* represents male vertices, *pink* represents female vertices, and *gray* represents unknown gender. Each vertex label, which is visible by zooming into the graph, contains the vertex lifespan if one exists.

4. **Gender( $v$ )** - The gender of  $v$  converted to an integer, where male is set to 1, female is set to 2, and unknown gender is set to 0.
5. **Birth-Country( $v$ )** - The country in which  $v$  was born.
6. **Death-Country( $v$ )** - The country in which  $v$  died.
7. **Age-of-Death( $v$ )** - The age of death of  $v$  (also referred to as the lifespan of  $v$ ) which is calculated, if accurate dates are available, by subtracting the birth date of  $v$  from the death date of  $v$ .

### 3.2.2 Nuclear Family Features

8. **Children-Number( $v$ )** - the number of children which  $v$  had. The formal Children-Number( $v$ ) definition is:

$$\text{Children-Number}(v) := |\{u \in V | u \in V \wedge \exists(v, u, \text{Child}, t) \in E\}|.$$

9. **Spouse-Number( $v$ )** - the number of individuals to which  $v$  was married to. The formal Spouse-Number( $v$ ) definition is:

$$\text{Spouse-Number}(v) := |\{u \in V | (u \in V \wedge \exists(v, u, \text{Spouse}, t) \in E)\}|.$$

10. **Min-Spouse-Age-of-Death( $v$ )** - the minimum age of death of  $v$ 's spouses. The formal Min-Spouse-Age-of-Death( $v$ ) definition is:

$$\text{Min-Spouse-Age-of-Death}(v) :=$$

$$\min(\{\text{Age-of-Death}(u) \neq \emptyset | u \in V \wedge \exists(v, u, \text{Spouse}, t) \in E\}),$$

where the function  $\min(S)$  returns the minimum value among set  $S$  members, or 0 if  $S$  is empty.

11. **Max-Spouse-Age-of-Death( $v$ )** - the maximum age of death of  $v$ 's spouses. The formal Max-Spouse-Age-of-Death( $v$ ) definition is:

$$\text{Max-Spouse-Age-of-Death}(v) :=$$

$$\max(\{\text{Age-of-Death}(u) \neq \emptyset | u \in V \wedge \exists(v, u, \text{Spouse}, t) \in E\}),$$

where the function  $\max(S)$  returns the maximum value among set  $S$  members, or 0 if  $S$  is empty.

12. **Avg-Spouse-Age-of-Death( $v$ )** - the average age of death of  $v$ 's spouses. The formal Avg-Spouse-Age-of-Death( $v$ ) definition is:

$$\text{Avg-Spouse-Age-of-Death}(v) :=$$

$$\text{avg}(\{\text{Age-of-Death}(u) \neq \emptyset | u \in V \wedge \exists(v, u, \text{Spouse}, t) \in E\}),$$

where the function  $\text{avg}(S)$  returns the average value of set  $S$  members, or 0 if  $S$  is empty.

### 3.2.3 Extended Family Features

13. **Father-Age-of-Death( $v$ )** -  $v$ 's father age of death. The formal Father-Age-of-Death( $v$ ) definition is:

$$\text{Father-Age-of-Death}(v) := \text{Age-of-Death}(\text{Father}(v)),$$

where the function  $\text{Father}$  returns the father vertex of  $v$ , if one exists. Namely,  $\text{Father}(v) := u$ , where  $u \in V \wedge \text{gender}(u) = 1 \wedge \exists(v, u, \text{Parent}, t) \in E$ .

14. **Mother-Age-of-Death( $v$ )** -  $v$ 's mother age of death. The formal Mother-Age-of-Death( $v$ ) definition is:

$$\text{Mother-Age-of-Death}(v) := \text{Age-of-Death}(\text{Mother}(v)),$$

where the function  $\text{Mother}$  returns the mother vertex of  $v$ , if one exists. Namely,  $\text{Mother}(v) := u$ , where  $u \in V \wedge \text{gender}(u) = 2 \wedge \exists(v, u, \text{Parent}, t) \in E$ .

15. **Paternal-Grandfather-Age-of-Death( $v$ )** -  $v$ 's paternal grandfather's age of death, if one exists. The formal Paternal-Grandfather-Age-of-Death( $v$ ) definition is:

$$\text{Paternal-Grandfather-Age-of-Death}(v) := \text{Age-of-Death}(\text{Father}(\text{Father}(v))).$$

16. **Maternal-Grandfather-Age-of-Death( $v$ )** -  $v$ 's maternal grandfather's age of death, if one exists. The formal Maternal-Grandfather-Age-of-Death( $v$ ) definition is:

$$\text{Maternal-Grandfather-Age-of-Death}(v) := \text{Age-of-Death}(\text{Father}(\text{Mother}(v))).$$

17. **Paternal-Grandmother-Age-of-Death( $v$ )** -  $v$ 's paternal grandmother's age of death, if one exists. The formal Paternal-Grandmother-Age-of-Death( $v$ ) definition is:

$$\text{Paternal-Grandmother-Age-of-Death}(v) := \text{Age-of-Death}(\text{Mother}(\text{Father}(v))).$$

18. **Maternal-Grandmother-Age-of-Death( $v$ )** -  $v$ 's maternal grandmother's age of death, if one exists. The formal Maternal-Grandmother-Age-of-Death( $v$ ) definition is:

$$\text{Maternal-Grandmother-Age-of-Death}(v) := \text{Age-of-Death}(\text{Mother}(\text{Mother}(v))).$$

19. **Sibling-Number( $v$ )** - the number of brothers and sisters  $v$  had. The formal Sibling-Number( $v$ ) definition is:

$$\text{Sibling-Number}(v) := |\{u \in V | u \in V \wedge \exists(v, u, \text{Sibling}, t) \in E\}|.$$

20. **Max-Sibling-Age-of-Death( $v$ )** - the maximum age of death of  $v$ 's siblings. The formal Max-Sibling-Age-of-Death( $v$ ) definition is:

$$\begin{aligned} \text{Max-Sibling-Age-of-Death}(v) := \\ \max(\{\text{Age-of-Death}(u) \neq \emptyset | u \in V \wedge \exists(v, u, \text{Sibling}, t) \in E\}). \end{aligned}$$

21. **Avg-Sibling-Age-of-Death( $v$ )** - the average age of death of  $v$ 's siblings. The formal Avg-Sibling-Age-of-Death( $v$ ) definition is:

$$\begin{aligned} \text{Avg-Sibling-Age-of-Death}(v) := \\ \text{avg}(\{\text{Age-of-Death}(u) \neq \emptyset | u \in V \wedge \exists(v, u, \text{Sibling}, t) \in E\}). \end{aligned}$$

Using the features defined above, we specify the following feature sets, which will later be used to construct our multiple linear regression models and ML classifiers: (a) *All-Numeric-Features* - a set which contains all the defined-above features that return numeric values, except the Death-Year feature; (b) *Heritage-Features* - a set which includes all the extended family features, including the Birth-Year and Gender features; and (c) *Nuclear-Family-Features* - a set which includes all the nuclear family features, including Birth-Year and Gender.

### 3.3 Statistical and Predictive Analysis

In this study, we used various algorithms and methods to calculate the variations in human lifespan over the past centuries, to identify which features are correlated with human lifespan and longevity, and to create predictive models which can assist in predicting human lifespan.

In the remainder of this subsection, we describe in detail each one of our methods and algorithms.

#### 3.3.1 Lifespan Variations over Time

After we had extracted the features for each vertex in the graph, we could utilize these features to calculate the variations in human lifespan over an extended period of time. To perform these calculations, we created two vertices datasets. The first dataset was the *All-Dataset*, which included all the vertices with valid values of Age-of-Death, while the second dataset was the *<Country>-Dataset* which included only vertices with valid values of Age-of-Death of people who were born in a specific country - in this study, we chose to take a closer look at people born in the United States.

We utilized the *All-Dataset* and the *United-States-Dataset* to specifically look at the lifespan of people who were born in each quarter of a century between 1650 and 1900. For each quarter of a century on each dataset, we calculated the Age-of-Death distribution of those people born in the chosen quarter. For example, in the second dataset, we had a group of 22,021 people who were born in the United States and lived between 1700 and 1724; we then calculated the percent of the population that died at each age between 0 and 122.<sup>10</sup>

Additionally, for the *All-Dataset-10* and for the *United-States-Dataset-10*, and for each year from 1650 to 1900, we calculated both the average and median lifespans of the people who were born in each year and outlived the age of 10. We also repeated these average and median calculations for each gender, using the *Male-Dataset-10*, *Female-Dataset-10*, *Male-United-States-Dataset-10*, and *Female-United-States-Dataset-10* datasets.

#### 3.3.2 Linear Regression

One of the main goals of this study was to identify features which are correlated with lifespan and with longevity. To identify features correlated with an inheritance of human lifespan, we computed for each extended family feature, which was defined in Section 3.2.3, a simple linear regression  $Y = \alpha + \beta X$ , where  $Y$  was set to be the Age-of-Death vector, and  $X$  was set to be selected feature values. For each feature we chose only vertices from the *<Feature>-Dataset-10*, in which both the Age-of-Death value was greater or equal ten<sup>11</sup> and the selected feature value existed.<sup>12</sup> We then evaluated each simple linear regression by computing the regression's *P-value* and *R-squared* values.

To identify if an individual's lifespan was correlated with the lifespan of his or her spouse(s), we repeated the same process of constructing a simple linear regression between the Age-of-Death feature and the Avg-Spouse-Age-of-Death, Max-Spouse-Age-of-Death,

<sup>10</sup>122 is the maximum confirmed human lifespan [23].

<sup>11</sup>We chose to use a minimum lifespan of 10 to avoid adding infant and child mortality, which might be misreported.

<sup>12</sup>For features such as Max-Sibling-Age-of-Death and Min-Spouse-Age-of-Death, which involved calculation of minimum, maximum or average, we ignored vertices with missing values, although by definition these features returned a valid value of 0.

and Min-Spouse-Age-of-Death features. However, this time we used the *Married-Dataset* to include only individuals who were married at least once.

To identify if longevity is correlated with reproductive success, we repeated the same process of constructing a simple linear regression between the Age-of-Death feature and the Children-Number feature. However, with respect to Westendorp and Kirkwood's [22] results in mind, we used *Children-Number-Dataset-50*, *Male-Children-Number-Dataset-50*, and *Female-Children-Number-Dataset-50* datasets, which only contained vertices with age of death of at least 50, namely after menopause.

### 3.3.3 Multiple Linear Regression

In this study, we used backward stepwise multiple linear regression to create models for predicting the Age-of-Death of individuals who had been born by 1900. We constructed these regression models by using the *All-Numeric-Features*, the *Heritage-Features*, and the *Nuclear-Family-Features* sets, which were defined at the end of Section 3.2. For constructing our first two models, we only used vertices from the *No-Missing-Dataset-10* dataset with valid complete values, including defined gender values, for each selected features set of vertices who outlived the age of 10. Additionally, to prevent bias due to the tendency of people to get married and have children in later stages of life, for the Nuclear-Family-Features set we only used vertices from the *No-Missing-Dataset-50* dataset, i.e., those who outlived the age of fifty.

We evaluated these multiple linear regression models by calculating the *P-value*, as well as the *Multiple R-squared*, *Adjusted R-squared*, and *Residual Standard Error* (RSE) values.

### 3.3.4 Machine Learning Algorithms

One of the major drawbacks of using online genealogy datasets is the issue of missing values. In many genealogy datasets not all the profile data is complete; many profiles contain missing values due to nonexistent data or privacy considerations [28]. To overcome the issue of missing values and still gain predictive information from the profiles with nonexistent data, we chose to use Machine Learning algorithms, such as decision trees and Naive-Bayes algorithms, which can deal with missing values.

We evaluated various supervised learning algorithms in an attempt to predict which individuals who were born in the United States between 1650 and 1900, and outlived the age of fifty, will also outlive the age of 80. We constructed our classifiers using Weka [9], a popular suite of ML, and the features defined in the *United-States-Dataset-50* dataset. We used all numeric features in each dataset, except the Age-of-Death and Death-Year features. Additionally, we also treated unknown gender values as missing values, instead of replacing them with 0 values. Using these datasets as training sets, we used Weka's OneR, C4.5 (J48) decision tree, K-Nearest-Neighbors (IBk; with K=3,5), Naive-Bayes, RandomForest, and Bagging implementations of the corresponding algorithms. For each of these algorithms, most of the configurable parameters were set to their default values except for the J48 decision tree classifier, in which the pruning option was not enabled. We evaluated each classifier using the 10-folds cross validation method and calculated the True-Positive, False-Positive, F-Measure, and the Area-Under-Curve (AUC) measure. The AUC is a standard way to compare classifier performances [3], in which 0.5 a value represents a random classifier.

Additionally, to obtain an indication of the usefulness of the various features, we analyzed their importance using Weka's information gain attribute selection algorithm.

### 3.4 WikiTree Dataset

To test and evaluate our methods and algorithms, we chose to use information collected from the WikiTree website. This is a free and accessible collaborative family history website started by Chris Whitten [25], and it contains more than 5 million profiles [5] of individuals who primarily lived in the past. WikiTree contains many profile pages of people who lived in the previous centuries, and many of the profiles contain the following details about each individual: full name, gender, date of birth, date of death, location of birth, location of death, parents' profiles, children's profiles, spouses' profiles, and siblings' profiles. Often, in order to maintain the privacy of still-living people, the website limits access to their profile personal details [28]. In order to maintain the integrity of WikiTree profile data, many profiles give reference to the source of the data presented in the profile, and most profiles have a profile manager who has primary responsibility for WikiTree profiles [27]. In addition, to prevent editing of profiles by untrusted users, each WikiTree profile has an independent “Trusted List” of people who can edit and view the profile [26], making the data in many profiles only editable to a limited number of people.

To collect profile information from WikiTree, we developed a web crawler which crawled and parsed only public profiles from the website. Using our crawler, we have downloaded and parsed 1,070,189 public profile pages. Using these profiles, we were able to construct a directed multigraph with 9,192,212 links and 1,382,752 vertices, out of which 118,590 vertices represented at least distinct 28,011 private profiles. Moreover, the constructed multigraph contained at least 416,030 vertices represent individuals born in the United States, according to their profile pages. These vertices were connected by 5,168,275 links to other vertices in the multigraph (see Figure 1, and Tables 3 and 4).

Table 3: WikiTree Dataset Statistics

Graph Property	All Profiles	Vertices of People Born in the United States
<b>Downloaded Profile Number</b>	1,070,189	416,030
<b>Vertices Number</b>	1,382,752	416,030
<b>Links Number</b>	9,191,147	5,168,275
<b>Male Vertices Number</b>	553,411	217,387
<b>Female Vertices Number</b>	506,605	196,793
<b>Vertices with Valid Lifespan (Age-Of-Death) Value</b>	545,993	285,868

## 4 Results

In the following subsections, we present the results obtained using the algorithms and methods described in Section 3. The results consist of three parts: First, we present the results of calculating lifespan variations over time. Second, we present the results of the simple linear regression and multi-linear regression analysis techniques which were described in Section 3.3.2. Finally, we present the results of the ML algorithms mentioned in Section 3.3.4.

Table 4: *All-Dataset* Number of Profiles Born in Each Year

<b>Decade/Year</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
<b>1650</b>	1177	1243	1186	1213	1414	1218	1169	1311	1071	1814
<b>1660</b>	1192	1280	1330	1343	1645	1400	1368	1539	1372	2136
<b>1670</b>	1346	1626	1568	1552	1917	1469	1334	1660	1504	2324
<b>1680</b>	1500	1660	1489	1567	1840	1576	1573	1665	1431	2437
<b>1690</b>	1484	1516	1544	1624	2012	1658	1586	1734	1602	2919
<b>1700</b>	1657	1936	1804	1885	2173	1814	1765	1862	1630	2579
<b>1710</b>	1665	1916	1701	1753	2172	1876	1769	1854	1734	2739
<b>1720</b>	1797	1870	1921	1915	2378	1919	1853	1980	1800	2757
<b>1730</b>	1809	2013	1914	1987	2558	2037	2018	2188	1911	2997
<b>1740</b>	1939	2093	1970	2094	2643	2130	2006	2184	1960	3114
<b>1750</b>	1917	2195	2103	2190	2591	2411	2201	2264	2177	3118
<b>1760</b>	2303	2490	2349	2361	3019	2330	2345	2422	2275	3181
<b>1770</b>	2248	2399	2323	2471	2992	2500	2282	2428	2192	3297
<b>1780</b>	2358	2476	2448	2571	2923	2659	2617	2668	2629	3637
<b>1790</b>	2603	2901	2780	2901	3268	2831	3057	2990	2859	4007
<b>1800</b>	3086	3070	3068	3199	3653	3214	3159	3252	3085	3922
<b>1810</b>	3275	3489	3188	3091	3580	3359	3168	3302	3248	3903
<b>1820</b>	3399	3445	3459	3485	3987	3601	3630	3813	3651	4250
<b>1830</b>	3900	3920	3989	4051	4408	4034	4150	4168	4094	4511
<b>1840</b>	4156	4280	4516	4321	4657	4355	4256	4408	4469	4724
<b>1850</b>	4246	4117	4230	4197	4461	4417	4527	4578	4603	4868
<b>1860</b>	4418	4352	4072	3999	4183	4000	4286	4393	4320	4770
<b>1870</b>	4126	4164	4265	4191	4433	4333	4296	4114	4351	4387
<b>1880</b>	3874	3644	3852	3721	4020	3717	3741	3683	3697	3942
<b>1890</b>	3554	3577	3591	3479	3529	3330	3375	3249	3222	3357
<b>1900</b>	3060	2708	2786	2729	2713	2640	2530	2603	2579	2476

#### 4.1 Lifespan Variations over Time Results

As described in Section 3.3.1, we utilized the *All-Dataset* and the *United-States-Dataset* to compute the changes of lifespan over each quarter of a century between 1650 and 1900. Then, we used these same datasets to take a closer look at the people who had been born during this 250-year span. For each quarter of a century on each dataset, we calculated the Age-of-Death distribution of the people who were born in the chosen quarter. The results showing the lifespan variations over time for the *All-Dataset* are presented in Figure 2a, and for the *United-States-Dataset* in Figure 2b.

We also used the *All-Dataset-10* and the *United-States-All-Dataset-10* to calculate the average and the median lifespans for each gender, and for both genders, in each year between 1650 and 1900. The results of these calculations are presented in Figures 3 and 4.

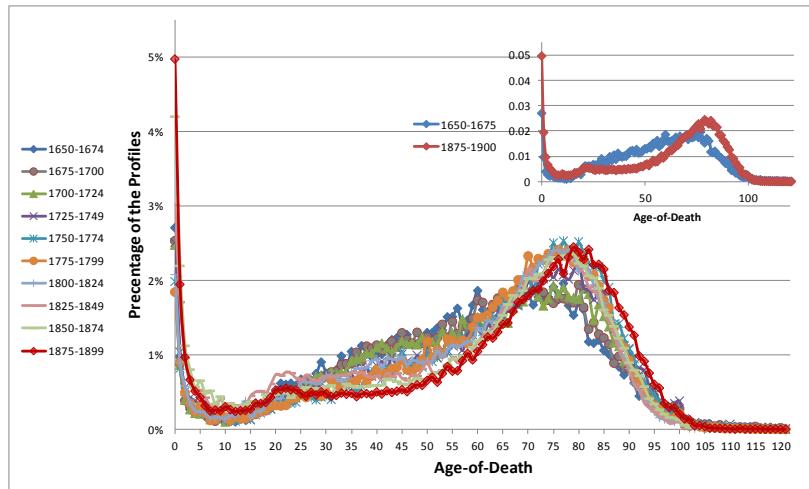
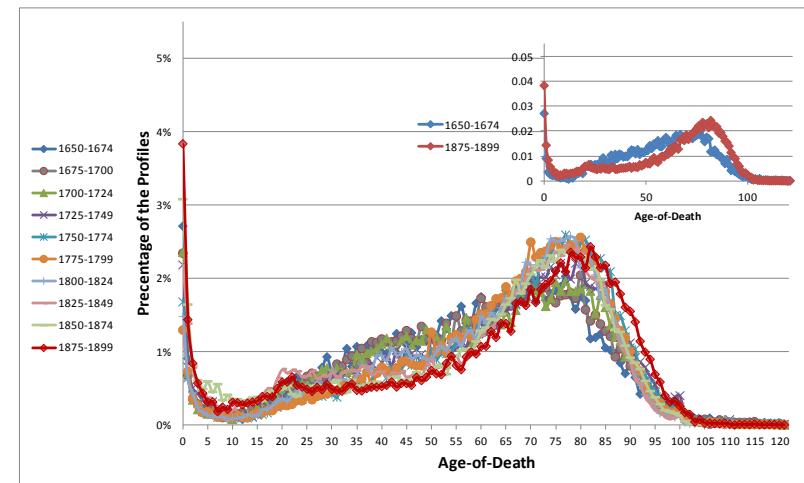
(a) *All-Dataset* Lifespan Variations.(b) *United-States-Dataset* Lifespan Variations.

Figure 2: Vertex Lifespan Variations, 1650-1900.

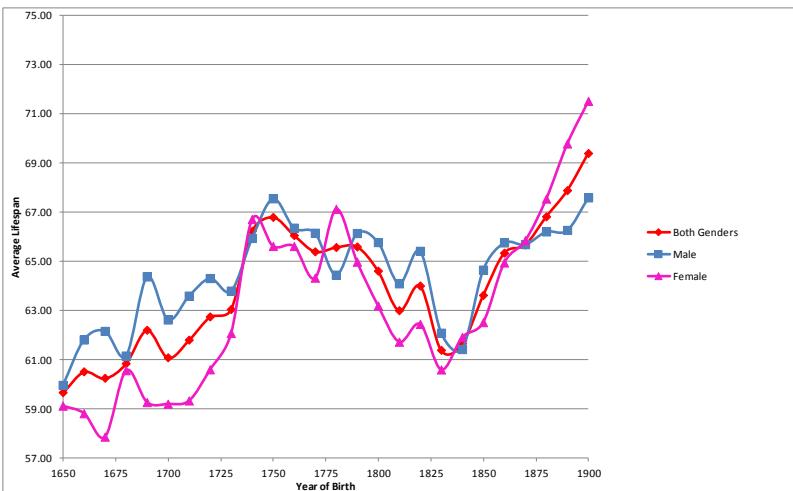
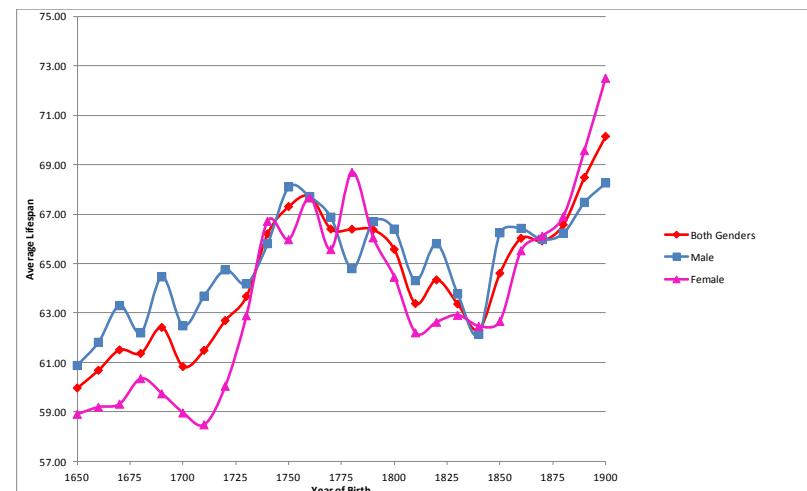
(a) *All-Dataset-10 - Average Lifespan.*(b) *United-States-Dataset-10 - Average Lifespan.*

Figure 3: Average Vertex Lifespans, 1650-1900.

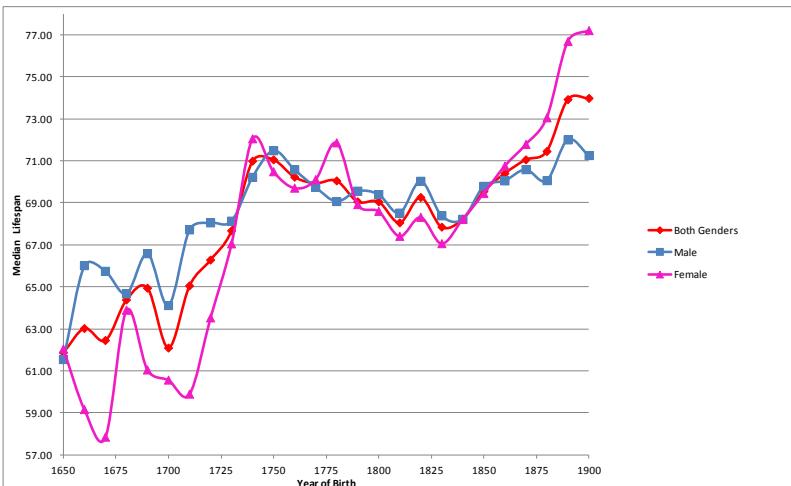
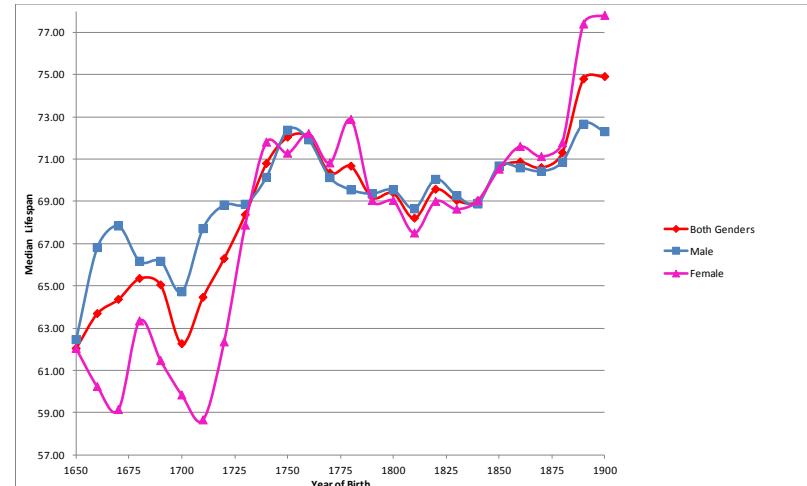
(c) *All-Dataset-10 - Median Lifespan.*(d) *United-States-Dataset-10 - Median Lifespan.*

Figure 4: Median Vertex Lifespans, 1650-1900.

## 4.2 Regression Analysis Results

Using R-project software [19], we ran several simple linear and multi-linear regression algorithms based on the features we defined in Section 3.2. From the regression algorithms, we generated and evaluated several prediction models in order to determine the vertices' Age-of-Death.

### 4.2.1 Simple Linear Regression Analysis Results

As described in Section 3.3.2, we computed several simple linear regression models in order to predict linear correlations between the Age-of-Death and other features. We first identified features correlated with the inheritance of human lifespan by computing a simple regression model for each feature in the *Extended Family Features* set, in order to predict the Age-of-Death feature. In these calculations, we used only the vertices who outlived the age of 10 and who also had valid existing information for each vertex's selected feature; i.e., we only used vertices which exist in the  $\langle Feature \rangle$ -Dataset-10 for each selected feature.

The simple regression results revealed that positive small but significant correlations exist between most of the *Extended Family Features* and the vertices' lifespans (see Table 5). These correlations have R-squared values ranging from 0.0015 to 0.05, with a very low P-value of  $2 \cdot 10^{-16}$  indicating that the correlation is highly significant, where the highest R-squared values were obtained for the Avg-Sibling-Age-of-Death (R-squared=0.05) and Max-Sibling-Age-of-Death (R-squared=0.0272) features, and the lowest R-squared values were obtained for the grandparents' lifespan features (R-squared ranging from 0.0015 to 0.0028). Additionally, we also discovered a small negative correlation between the Sibling-Number feature and the vertices' lifespan, with a slope of -0.155, R-squared of 0.0021, and P-value of  $2 \cdot 10^{-16}$ .

We then repeated the simple linear regression calculation to identify correlations between the vertices' lifespans and their spouses' lifespans by using the Avg-Spouse-Age-of-Death, Max-Spouse-Age-of-Death, and Min-Spouse-Age-of-Death features with the *Married-Dataset*. We discovered that each one of these features demonstrated a significant correlation, with a low P-value of  $2 \cdot 10^{-16}$  and a maximum R-squared value of 0.0564 (see Table 5).

Lastly, to identify if longevity is correlated with reproductive success, we computed simple linear regressions between the Age-of-Death feature and the Children-Number feature on the following datasets: *Children-Number-Dataset-50*, *Male-Children-Number-Dataset-50*, and *Female-Children-Number-Dataset-50*. Using the simple linear regression, we obtained the following correlation results: (a) on the *Children-Number-Dataset-50* dataset ( $n = 375, 938$ ) the regression returned a negative slope of -0.006, with a R-squared of  $4.1 \cdot 10^{-6}$  and a P-value of 0.2137; (b) on the *Male-Children-Number-Dataset-50* dataset ( $n = 214, 864$ ) the regression returned a positive slope of 0.044, with a R-squared of 0.0002 and a P-value of  $1.64 \cdot 10^{-11}$ ; and (c) on the *Female-Children-Number-Dataset-50* dataset ( $n = 160, 149$ ) the regression returned a negative slope of -0.079, with a R-squared of 0.0006 and a P-value of  $2.2 \cdot 10^{-16}$ .

### 4.2.2 Multi-Linear Regression Analysis Results

To create models which can estimate a vertex age of death based on the vertex's features, we chose to use the backward stepwise multiple linear regression technique. By combining this technique with the various predefined features sets, we created three multiple regression models which presented Multiple R-squared values of 0.085, 0.042, and 0.025, for

Table 5: Simple Linear Regression Results

Feature	Sample Size (n)	Y-Intercept ( $\alpha$ )	Slope ( $\beta$ )	R-Squared ( $R^2$ )	P-value
<b>Father-Age-of-Death</b>	339,039	53.69	0.129	0.0095	$2.2 \cdot 10^{-16}$
<b>Mother-Age-of-Death</b>	291,823	54.97	0.114	0.0098	$2.2 \cdot 10^{-16}$
<b>Paternal-Grandfather-Age-of-Death</b>	242,857	58.11	0.064	0.0026	$2.2 \cdot 10^{-16}$
<b>Maternal-Grandfather-Age-of-Death</b>	159,576	57.84	0.067	0.0028	$2.2 \cdot 10^{-16}$
<b>Paternal-Grandmother-Age-of-Death</b>	200,828	59.67	0.043	0.0015	$2.2 \cdot 10^{-16}$
<b>Maternal-Grandmother-Age-of-Death</b>	137,711	58.95	0.052	0.0021	$2.2 \cdot 10^{-16}$
<b>Sibling-Number</b>	508,520	63.87	-0.155	0.0021	$2.2 \cdot 10^{-16}$
<b>Max-Sibling-Age-of-Death</b>	309,282	46.40	0.201	0.0272	$2.2 \cdot 10^{-16}$
<b>Avg-Sibling-Age-of-Death</b>	309,282	45.64	0.287	0.0500	$2.2 \cdot 10^{-16}$
<b>Min-Spouse-Age-of-Death</b>	255,248	51.89	0.212	0.0453	$2.2 \cdot 10^{-16}$
<b>Max-Spouse-Age-of-Death</b>	255,248	49.26	0.247	0.0564	$2.2 \cdot 10^{-16}$
<b>Avg-Spouse-Age-of-Death</b>	255,248	50.02	0.238	0.0526	$2.2 \cdot 10^{-16}$

the *All-Numeric-Features* set, *Heritage-Features* set, and *Nuclear-Family-Features* sets respectively (see Table 6). We also obtained the following multiple linear regression models for each features set: For the *All-Numeric-Features* set, we computed the following model using data collected from 59,893 vertices who were born by 1900 and outlived the age of 10:

$$\begin{aligned}
 \text{Age-of-Death}(v) = & 9.813 - 2.194 \cdot \text{Gender}(v) + \\
 & 0.023 \cdot \text{Birth-Year}(v) + 0.261 \cdot \text{Children-Number}(v) + \\
 & 1.706 \cdot \text{Spouse-Number}(v) + \\
 & 0.084 \cdot \text{Max-Spouse-Age-of-Death}(v) + \\
 & 0.053 \cdot \text{Father-Age-of-Death} + \\
 & 0.039 \cdot \text{Mother-Age-of-Death}(v) + \\
 & 0.015 \cdot \text{Paternal-Grandfather-Age-of-Death}(v) + \\
 & 0.016 \cdot \text{Maternal-Grandfather-Age-of-Death}(v) + \\
 & 0.011 \cdot \text{Paternal-Grandmother-Age-of-Death}(v) + \\
 & 0.062 \cdot \text{Sibling-Number}(v) + \\
 & -0.1 \cdot \text{Max-Sibling-Age-of-Death}(v) + \\
 & 0.193 \cdot \text{Avg-Sibling-Age-of-Death}(v)
 \end{aligned}$$

For the *Heritage-Features* set, we computed the following model using data collected

from 59,893 vertices who were born by 1900 and outlived the age of 10:

$$\begin{aligned} \text{Age-of-Death}(v) = & 18.184 - 2.06 \cdot \text{Gender}(v) + \\ & 0.021 \cdot \text{Birth-Year}(v) + \\ & 0.057 \cdot \text{Father-Age-of-Death} + \\ & 0.041 \cdot \text{Mother-Age-of-Death}(v) + \\ & 0.013 \cdot \text{Paternal-Grandfather-Age-of-Death}(v) + \\ & 0.016 \cdot \text{Maternal-Grandfather-Age-of-Death}(v) + \\ & 0.016 \cdot \text{Paternal-Grandmother-Age-of-Death}(v) + \\ & -0.136 \cdot \text{Max-Sibling-Age-of-Death}(v) + \\ & 0.202 \cdot \text{Avg-Sibling-Age-of-Death}(v) \end{aligned}$$

For the *Nuclear-Family-Features* set, we computed the following model using data collected from 349,118 vertices who were born by 1900 and outlived the age of 50:

$$\begin{aligned} \text{Age-of-Death}(v) = & 47.73 + 1.063 \cdot \text{Gender}(v) + \\ & 0.013 \cdot \text{Birth-Year}(v) + \\ & -0.034 \cdot \text{Children-Number}(v) + \\ & -0.22 \cdot \text{Spouse-Number}(v) + \\ & -0.08 \cdot \text{Min-Spouse-Age-of-Death}(v) + \\ & 0.098 \cdot \text{Avg-Spouse-Age-of-Death}(v) \end{aligned}$$

Table 6: Multiple Linear Regression Results

Model Attributes \ Features Set	All-Numeric Features	Heritage Features	Nuclear-Family Features
<b>Sample Size (<i>n</i>)</b>	59,893	59,893	349,118
<b>Multiple R-Squared</b>	0.085	0.042	0.025
<b>Adjusted R-Squared</b>	0.085	0.042	0.025
<b>P-Value</b>	$2.2 \cdot 10^{-16}$	$2.2 \cdot 10^{-16}$	$2.2 \cdot 10^{-16}$
<b>Residual Standard Error</b>	20.33	20.80	11.95

### 4.3 Machine Learning Algorithms Results

We evaluated various supervised learning algorithms in an attempt to predict which individuals who were born in the United States between 1650 and 1900 and outlived the age of fifty, will also outlive the age of eighty. We constructed our classifiers using the *United-States-Dataset-50* dataset, which contained features of 183,494 vertices who outlived the age of fifty, out of which 58,975 vertices outlived the age of 80. To better understand which features were most useful to our classification algorithms, we analyzed the various features' importance using Weka's information gain features selection algorithm. For the *United-States-Dataset-50* dataset, the top eight features with the highest rank retrieved from Weka's information gain features selection algorithm were: (a) Birth-Year

(0.0058), (b) Max-Sibling-Age-of-Death (0.0047), (c) Avg-Sibling-Age-of-Death (0.0023), (d) Max-Spouse-Age-of-Death (0.0019), (e) Gender (0.0018), (f) Avg-Spouse-Age-of-Death (0.0016), (g) Min-Spouse-Age-of-Death (0.0016), (h) Father-Age-of-Death (0.0008), (i) Mother-Age-of-Death (0.0006), and (j) Parental-Grandfather-Age-of-Death (0.0002). At the end of the list, the Maternal-Grandfather-Age-of-Death, the Children-Number, and the Sibling-Number features received an information gain score of 0.

On this dataset, the RandomForest classifier received the maximum AUC of 0.632, better than a random classifier with AUC of 0.5, while the maximum True-Positive value of 0.976 was obtained by the Decision-Tree classifier, and the minimum False-Positive of 0.774 was obtained by K-Nearest-Neighbors (K=3) classifier (see Table 7). We used T-tests with a significance of 0.05 to compare the AUC results of the RandomForest and the naive OneR classifiers. According to the T-test result, RandomForest classifier performed better in terms of AUC, than the naive OneR classifier.

Table 7: Machine Learning Classifiers Results

<b>Measures</b>	<b>True Positive</b>	<b>False-Positive</b>	<b>F-Measure</b>	<b>AUC</b>
<b>Algorithms</b>				
<b>OneR</b>	0.908	0.895	0.779	0.507
<b>Decision-Tree (J48)</b>	<b>0.976</b>	0.946	0.805	0.587
<b>Naïve-Bayes</b>	0.973	0.949	0.803	0.571
<b>K-Nearest-Neighbors (K=3)</b>	0.797	<b>0.774</b>	0.737	0.518
<b>K-Nearest-Neighbors (K=5)</b>	0.846	0.820	0.757	0.522
<b>RandomForest</b>	0.947	0.858	0.805	<b>0.632</b>
<b>Bagging</b>	0.971	0.917	<b>0.807</b>	0.620

## 5 Discussion

To our knowledge, this study is the largest study to date which utilizes genealogical datasets to better understand factors that correlate with human lifespan. The algorithms and methods presented throughout this study, which were evaluated on the WikiTree dataset, reveal several interesting patterns and correlations.

Firstly, our results of lifespan variations over time, presented in Section 4.1 and in Figure 2, demonstrate how the lifespans of human population changed over the previous centuries. The lifespan graphs presented in Figure 2 show high infant and children death rates as well as local maximum values between the ages of 70 and 80; these results resemble the lifespan graphs presented in Mitchell et al. [14] and by the UK Office of National Statistics [17]. This resemblance supports our assumptions regarding the integrity of the WikiTree dataset, which indeed contains data on human population with largely accurate birth and death dates. However, the infant death rates presented in these graphs are not entirely accurate; according to Wegman [21], in 1900 the infant mortality rates in the United States were about 15%, which is higher than the values presented in our results. We assume that the main reason for this discrepancy was the lack of a uniform, formal

definition of “live births,” which was not standardized until 1951 [21]. Therefore, in most of this study we used as a sample set only people who outlived the age of ten. Nevertheless, by analyzing these graphs, we can observe that over time, lifespans increased and fewer people passed away at young ages. Another observation that can be concluded from these graphs is that even in the second half of the seventeenth century, people who outlived the age of ten would likely outlive the age of sixty. Indeed, according to our median lifespan analysis, presented in Figure 3d, the median age in 1650 for people who were born in the United States and outlived the age of ten was 62.46 for males and 62.04 for females.

Secondly, our median and average population lifespan calculations, presented in Figures 3 and 4, reveal some interesting patterns. By analyzing the graphs, we can locate several years in which the average lifespans sharply decreased for both males and females. For example, for people who were born in the United States in 1800 and outlived the age of 10, the average lifespans for males and females were 66.39 and 64.45, respectively. However, for people born in the United States ten years later, in 1810, the average lifespan was reduced by around 2 years: males’ lifespans decreased to 64.31, and females’ lifespans decreased to 62.20. An additional and even more interesting reoccurring pattern can be identified in Figure 3d in which, for a specific time period, the median lifespan for males increased while the median life span for females suddenly decreased, or vice versa. For example, from 1650 to 1660 the male median lifespan increased from 62.46, to 66.82 while in the same period of time the female median lifespan decreased from 62.04 to 60.24. Similar patterns reoccur between 1770 and 1780, only this time the female average lifespan increased from 65.57 to 68.69, while the male average lifespan decreased from 66.87 to 64.79 (see Figure 3b). Another interesting pattern can be found between 1850 and 1900 where in just a half a century the female average lifespan sharply increased from 62.66 to 72.5. We hope to discover underlying reasons for these patterns in our future research.

Thirdly, using simple linear regression algorithms, we uncovered small but significant correlations between various features and the Age-of-Death feature which are presented in Table 5. We found small positive significant correlations between the *Extended Family Features* and the Age-of-Death feature. For all these correlations, R-squared values were small and ranged from 0.0015 to 0.05 with a P-value of  $2.2 \cdot 10^{-16}$ , and these may indicate that lifespan can “run in the family.” However, due to the small R-squared values in our results, we can conclude that the influence of inherited lifespan is limited and, in fact, negligible after more than one generation. Alternately, the observed correlation could be explained due to socioeconomic reasons: ancestors with long lifespan might also indicate a higher socioeconomic status, which can be passed on to their offspring. We also found significant correlations between the Avg-Spouse-Age-of-Death, Max-Spouse-Age-of-Death, and Min-Spouse-Age-of-Death features and the Age-of-Death feature, with a low P-value and a maximum R-squared value 0.0564 (see Table 5). This indicates that correlations between the lifespans of spouses exist, supporting the claims for the existence of the “widow effect.” We hope to confirm this observation in a future study by taking a closer look at the time intervals between the deaths of married couples. Using simple linear regression models, we also identified small significant correlations between longevity and reproductive success. Namely, we discovered negligible negative correlation between females and their number of children ( $R^2 = 0.0006$ ), and negligible positive correlation between males and their number of children ( $R^2 = 0.0002$ ).

Fourthly, using multiple linear regression models, we were able to construct models which can predict a person’s age of death using various features that were extracted from the WikiTree social network directed multigraph. Our models presented a low P-value of  $2.2 \cdot 10^{-16}$  with Adjusted R-squared of up to 0.085 (see Table 6), indicating that the

extracted features can indeed assist in predicting a person Age-of-Death based on data which was extracted from the WikiTree dataset. However, the relative low Adjusted R-squared values indicate that other external factors are also responsible for influencing an individual's lifespan. We hope to test these assumptions in future studies by merging WikiTree genealogy datasets with other datasets that contain additional information about individuals' habits and lifestyles.

Fifthly, our machine learning classifiers presented better than random performances, with AUCs up to 0.632 (see Table 7), in identifying which people who were born in the United States and outlived the age of 50 would also outlive the age of 80. These results support our previous claims that the data collected from genealogy datasets can be utilized to predict a person's lifespan. Additionally, the information gain algorithm results revealed that the Max-Sibling-Age-of-Death, Avg-Sibling-Age-of-Death, and Max-Spouse-Age-of-Death (see Section 4.3) were among the most useful features. These results also indicate that a correlation exists both between spouses' lifespans and between siblings' lifespans. In our future studies, we hope to use similar techniques to predict other personal attributes based on data collected from online genealogy datasets.

The study presented here is among the first of its kind and offers many future research directions to pursue. One possible research direction is to analyze not only the structured data which appear in the WikiTree profile pages, but also to use Natural Language Processing (NLP) algorithms to analyze content data which appear in these pages. Another possible research direction is to compare the results presented in this study from the WikiTree dataset to other online genealogy datasets, such as FamiLinx,<sup>13</sup> which is publicly available and contains information from Geni.com,<sup>14</sup> a genealogy-driven social network. Additionally, in future work, we plan to utilize the results obtained through this study on the population of United States and test them on various populations in other countries.

## Acknowledgment

The authors would like to thank Carol Teegarden for her editing expertise and helpful advice.

## References

- [1] Y. Altshuler, N. Aharony, M. Fire, Y. Elovici, and A. S. Pentland. Incremental learning with accuracy prediction of social and individual properties from mobile-phone data. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pages 969–974. IEEE, 2012.
- [2] Ancestry.com. Ancestry.com inc. reports q3 2012 financial results. 2013. (last accessed on November 2th, 2013).
- [3] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.
- [4] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

<sup>13</sup><http://erlichlab.wi.mit.edu/familinx/data.html>

<sup>14</sup><http://www.geni.com/>

- [5] D. Eastman. Wikitree reaches five million profiles. [http://blog.eogn.com/eastmans\\_online\\_genealogy/2013/04/wikitree-reaches-five-million-profiles.html](http://blog.eogn.com/eastmans_online_genealogy/2013/04/wikitree-reaches-five-million-profiles.html), 2013. (last accessed on November 2th, 2013).
- [6] F. Elwert and N. A. Christakis. The effect of widowhood on mortality by the causes of death of both spouses. *Journal Information*, 98(11), 2008.
- [7] M. Fire, G. Katz, Y. Elovici, B. Shapira, and L. Rokach. Predicting student exam's scores by analyzing social network data. In *Active Media Technology*, pages 584–595. Springer, 2012.
- [8] M. Gögele, C. Pattaro, C. Fuchsberger, C. Minelli, P. P. Pramstaller, and M. Wjst. Heritability analysis of life span in a semi-isolated population followed across four centuries reveals the presence of pleiotropy between life span and reproduction. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 66(1):26–37, 2011.
- [9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November 2009.
- [10] J. Knowles. Myheritage hits 1 billion profiles and announces new features for historical research. 2012. (last accessed on November 2th, 2013).
- [11] É. Le Bourg. Does reproduction decrease longevity in human beings? *Ageing research reviews*, 6(2):141–149, 2007.
- [12] P. Martikainen and T. Valkonen. Mortality after the death of a spouse: rates and causes of death in a large finnish cohort. *American Journal of Public Health*, 86(8 Pt 1):1087–1093, 1996.
- [13] P. F. McArdle, T. I. Pollin, J. R. O'Connell, J. D. Sorkin, R. Agarwala, A. A. Schäffer, E. A. Streeten, T. M. King, A. R. Shuldiner, and B. D. Mitchell. Does having children extend life span? a genealogical study of parity and longevity in the amish. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 61(2):190–195, 2006.
- [14] B. D. Mitchell, W.-C. Hsueh, T. M. King, T. I. Pollin, J. Sorkin, R. Agarwala, A. A. SchaÈffer, and A. R. Shuldiner. Heritability of life span in the old order amish. *American journal of medical genetics*, 102(4):346–352, 2001.
- [15] N. H. Murdock. Teenage pregnancy. *Journal of the National Medical Association*, 90(3):135, 1998.
- [16] MyHeritage. Myheritage members map. <http://www.myheritage.com/myheritage-member-map>, 2013. (last accessed on November 2th, 2013).
- [17] O. of National Statistics. Mortality in england and wales: Average life span. [http://www.ons.gov.uk/ons/dcp171776\\_292196.pdf](http://www.ons.gov.uk/ons/dcp171776_292196.pdf), 2012. (last accessed on Nov. 13th, 2013).
- [18] C. M. Parkes, B. Benjamin, and R. G. Fitzgerald. Broken heart: a statistical study of increased mortality among widowers. *British Medical Journal*, 1(5646):740, 1969.

- [19] R. D. C. Team et al. R: A language and environment for statistical computing, 2005.
- [20] F. Thomas, A. Teriokhin, F. Renaud, T. De Meeûs, and J.-F. Guégan. Human longevity at the cost of reproductive success: evidence from global data. *Journal of Evolutionary Biology*, 13:409–414, 2000.
- [21] M. E. Wegman. Infant mortality in the 20th century, dramatic but uneven progress. *The Journal of Nutrition*, 131(2):401S–408S, 2001.
- [22] R. G. Westendorp and T. B. Kirkwood. Human longevity at the cost of reproductive success. *Nature*, 396(6713):743–746, 1998.
- [23] C. R. WHITNEY. Jeanne calment, world's elder, dies at 122. *New York Times*, 1997.
- [24] Wikipedia. Wikipedia:statistics. <http://en.wikipedia.org/wiki/Wikipedia:Statistics>, 2013. (last accessed on November 2th, 2013).
- [25] WikiTree. About wikitree. [http://www.wikitree.com/wiki/About\\_WikiTree](http://www.wikitree.com/wiki/About_WikiTree), 2013. (last accessed on November 2th, 2013).
- [26] WikiTree. Trusted list. [http://www.wikitree.com/wiki/Trusted\\_List](http://www.wikitree.com/wiki/Trusted_List), 2013. (last accessed on November 2th, 2013).
- [27] WikiTree. Wikitree manager. [http://www.wikitree.com/wiki/Profile\\_Manager](http://www.wikitree.com/wiki/Profile_Manager), 2013. (last accessed on November 2th, 2013).
- [28] WikiTree. Wikitree privacy. <http://www.wikitree.com/wiki/Privacy>, 2013. (last accessed on November 2th, 2013).