Improving data quality through high precision gender categorization

Daniel Müller, Yiea-Funk Te, Pratiksha Jain
Department of Management, Technology and Economics
ETH Zurich
Zurich, Switzerland
danielmueller@ethz.ch, fte@ethz.ch, 25pratiksha@gmail.com

Abstract— **First name to gender mappings have been widely recognized as a critical tool to complete, study and validate data records in a range of different areas. In this study, we investigate how organizations with large databases of existing entities can create their own mappings between first names and gender and how these mappings can be improved and utilized. Therefore, we first explore a dataset with demographic information on more than 6 million people, provided by a car insurance. We then study how naming conventions have changed over time and how they differ by nationality. Second, we build a probabilistic first name to gender mapping and augment the mapping by adding nationality and decade of birth to improve the mapping's performance. We test our mapping in a two label and three label setting and further validate our mapping by categorizing patent filings by gender of the inventor. We compare the results with previous studies' outcomes and find that our mapping produces high precision results. We validate that the additional information of nationality and year of birth improve the recall scores of name to gender mappings. Therefore, it constitutes an efficient process to improve data quality of organizations' records, whenever the attribute gender is missing or unreliable.**

Keywords-data quality improvement, record completion, gender name mapping, patenting

I. INTRODUCTION

High quality data is a key asset to improve efficiency in competitive business environments [1]. Practitioners and academics consider quality data a key organizational resource and agree that it should be managed accordingly [2]. Therefore, extracting valid information to support decision making is a critical success factor for an organization in today's society [3]. As organizations began to investigate their own records to build decision support systems, some of the resulting data quality problems became observable for the first time [4]. A data quality problem caused by incorrect or incomplete attributes such as gender is commonly-cited by practitioners integrating two data sources. They report that linking two data sets through common demographic fields often fails since records have missing or flawed attributes, which are required for the matching [5]. Besides record linkage problems caused by missing or wrong gender attributes, the unvalidated application of erroneous gender derived insights can have devastating effects on a firms' reputation, efficacy of marketing budgets and its strategic decisions [6], [7]. Organizations aim to avoid such data quality related problems by finding ways to improve and validate their own records in systematic ways [8]. Those organizations with data privacy concerns eventually avoid buying data for record completion from unknown and untested sources [7], [9]. Instead, they aim to leverage their internal data records to gain insights, build capabilities, comply with data laws and avoid reputational damages [7]. Building an algorithmic systematic categorization of gender, which is based on organizations' existing records, helps organizations achieve these goals. Such a mappings allows the validation and completion of the attribute gender, however can only be performed, if a sufficiently large and representative amount of records are available.

Existing research of the systematic categorization of gender can be found in the domains of humanities, marketing and information systems [10]–[12], [8], [13]. This is due to new forms of the widespread digitization, which opened up new opportunities to apply computational methods to archival databases. As predominantly corporate and federal digital records were made available for the public or for smaller circles of researchers, academia has been able to process this information and derive new knowledge. Alongside with the proliferation of machine-readable datasets and digital texts in general, a vast trove of material became available to study and estimate gender [10] inferred from various attributes. Those studies partially resulted in the creation of name to gender mapping tables. Intended as the outcome of an interim step to perform the studies' goal, these mappings of first names and gender were created and applied to categorize user records at large scale. Such mapping algorithms can infer the respective genders by looking up for example "Anna" and "Bob" in a dataset that matches first names to genders. Such n-to-gender rule based categorizations have been applied in humanities to categorize records in order to uncover the gender differences in a range of domains [10]. Further, such mappings were applied and discussed in the context of information science, marketing and census. Some author looked at gender differences over longer periods of time and have addressed the concern that many other studies' findings may not generalizable or stable over time. This concern is particularly stressed in studies evaluating the publication performance of female and male authors within the last decades [14], [15]. Generalization of

findings is questionable due to the fact, that naming conventions not only differ by nationality, but also change over time. This problem is also known as the "Leslie problem" and has been researched in Anglo-Saxon countries [10]. A study on gender and naming practices found that conventions, like language itself, is not static [16]. Instead, we need an improved approach to infer gender from digital data sources, dynamically capturing other user attributes to improve categorization precision. Also, depending on the application type of the gender mapping, a user might be interested only in the high precision outcomes of the mapping or a "best guess", providing a binary gender categorization with a high recall score for a complete lists of names. Therefore, the quality of a mapping can only be evaluated in context of the application.

How to effectively build such a method with the goal of improving the data quality of a company's data records and how precision and recall affect each other will be discussed in the remainder of this study. In section two we review the current literature on how naming gender mappings were created and where they have been applied with which success. Further we illustrate in Section 3, how our own mapping is created and how we can improve recall and precision results, which we state in Section 4. Finally, we discuss our findings in Section 5 and offer implication for practitioners while also stating the limitation of our study.

II. RELATED WORK

The categorization of gender is a problem that has been approached in different research domains. Studies and applications of available gender categorization tools can be found in the domains of Marketing, Humanities, Information Science and Census literature. Many researchers have performed gender categorization to uncover instances of gender inequity in a range of different areas, ranging from authorship of French Literature, to the disparities between attendees of the annual Digital Humanities Conference [17]. Others have contributed to identify attributes which allow gender categorization and can be applied in reversal.

In Marketing, gender is one of the most common forms of segmentation used by marketers in general and advertisers in particular [18]. In general, males and females are likely to differ in information processes and decision making. Therefore marketing researchers have for example tried to identify gender based on a web users' perception of Web advertising [19] or browsing use pattern [20]. They found that the genders make use of the web differently [21], hence a user's gender can be identified, which presents opportunities for advertisers such as ad placement targeting. A prerequisite of any gender segmentation is the representation of real world entities to which they refer in a consistent, accurate, complete, timely and unique way [8]. The quality of the input data strongly influences the quality of the results [22] ("garbage in, garbage out" principle) and is essentially studied in two research communities: databases and management [23]. The first one studies data quality from a technical point of view (e.g., [24]), while the second one is also concerned with other aspects or dimensions (e.g., accessibility, believability, relevancy, interpretability, objectivity) involved in data quality (e.g., [25], [26]). The completion and the data quality improvement of a firms records, hence creates an opportunity for those enterprises that engage in efforts to take advantage of best practices in data quality management from a technical and dimensional point of view [11].

In the domain of Information Science, scholars predicted for example gender based on people's internet browsing history [27]. The authors' experimental results, based on click-through log, showed that they were able to achieve 79.7% precision in gender categorization. Similarly, authors of [28] have found significant linguistic differences between men and women, which can be identified in written or spoken form. The author determined multiple linguistic features such as character usage, writing syntax, functional words, and word frequency, which can be mapped to gender. Those and other features have been examined and are contained in the Media Research Center' (MRC) Psycholinguistic database, however have not yet been used in reversal to categorize gender [28]. Authors of [29] applied a simple Neural Network to categorize gender on a sample that was extracted from the Enron email dataset, provided by the Carnegie Mellon University. The emails were labeled according to gender and the authors algorithm were able to achieve a 95% precision using word based features [29].

In the domain of Humanities, initiatives like the Orlando Project, the Poetess Archive and the Women Writers Project have evaluated the share of female authors and writers. Further, authors such as [17] have studied the linguistic styles of male and female playwrights or representations of gendered bodies in European fairy tales. Authors of [30] uncovered the underlying trends across a century of academic articles in literary studies [30], also mapping first name to gender. Further, authors of [31] have performed a mapping between names and gender, based on historical African American naming practices and have identified a set of "distinctively African American names". This mapping of name and gender has been used in a number of papers evaluating the share of African Americans in a list of names attributed to patenting activity [4,5] gender categorization in general [33], age [34] and income estimation [35]. Most recently, the author of [10] inferred gender from one of the most common features of humanities datasets, the personal names of authors. Mapping first names and gender over time, they investigated changes in naming conventions in the United States.

In census, scholars have examined naming conventions and changes to those. They found that for example the conventional gender for names switched over the course of a few decades. In 1900 some 92% of the babies born were named Leslie were male, while in 2000 about 96% of the Leslies born in that year in the United States were female [10]. Changes in naming practices create a problem, especially for databases with records of people born in the

"transition phase" of naming conventions. As the average lifespan of humans is increasing, this problem is further intensifying. Today, the average European citizen reaches an age of more than 80 years [36], which is sufficient time for naming practices to change [10]. Further, applying a name to gender mapping allowed to examine the role of patents attributed to female inventors to be discussed in research about gender disparity in patenting [37]. In their research authors [37] created their mapping based on a universal and country-specific name lists of unknown origin, which they did not further elaborate upon and applied it on a sample of 4.6 million utility patents granted by the United States Patent and Trade Office (USPTO). The authors of [14] concluded that the lack of academic and female innovators, which they identified from their mapping of name and gender, is a suitable metric for female innovation. Patent activity has also been studied in terms of gender distribution in a recent study by the Elsevier Analytical Services [14]. In this study, the authors have mapped research performance, including patents granted with inventors' gender. The authors have relied on social networking service data to calculate the probabilities, hence naming conventions from the 1930 - 1950 may not have been entirely representative for the current share of people alive. In the study using social media profiles, the authors have limited their mapping to names, which appeared at least 5 times in their data set and had a probability of male or female of at least 85%.

Many studies we found discuss differences in genders' behavior or segment based on gender. In order to label by gender, some authors found creative ways to create name gender mappings, whereas some relied on list created by other researchers. Some identified the need to consider other variables in order to improve the mapping quality and only very few [10] have quantified their findings, which carries importance depending on the usage goal of the name mapping, either preferring a high precision or a high recall.

According to our systematic literature review, there is no research investigating the change in naming conventions and nationality together, nor has anybody evaluated or compared their mappings on recall and precision together, which carries importance in many data quality problems. Our literature review identified studies addressing how additional attributes can enhance the mapping quality ("Leslie problem"), however only for the United States. By composing an algorithm sensitive to nationality and year of birth, we aim to overcome the above mentioned research gaps and answer the following research question:

RQ: How can we improve the categorization of gender based on first names?

In order to answer this research questions, we collect a dataset from a Swiss car insurer and extract the first names and gender of each available policy. Further, we extract nationality and date of birth from the same vehicle insurance policies. In the process, we evaluate if the inclusion of those demographic characteristics hold discriminative power, allows to improve the categorization precision over basic first name gender mappings. We then apply our mapping to 20,000 names, which the mapping algorithm has not seen and evaluate the mapping performance in two scenarios. First, by evaluating precision of the mapping outcome, categorizing into male and female names. In a second step, we evaluate the performance of the mapping in a three class scenario which contains female, male and unisex labels. Finally, we apply our mapping to a dataset provided by the Swiss Patent office and compare the results to data provided by Statista Data Services, elaborating on the share of patent applicants that were female in Switzerland from 1980 to 2013 [38].

With our study, we contribute to the research stream on data quality improvement by presenting new evidence how organizations can leverage their own data records. Through the combination of demographic information with available name gender mappings, incomplete data records can be completed with the attribute gender in an efficient and accurate manner. Further, we contribute to the research area of data quality analytics within experimental algorithms in Information Science by following the recent call of [10] to provide a method for gender categorization that takes demographic changes over time into account.

III. METHODOLOGY

A. Dataset

We create our name-gender tables from policy data of Swiss car insurer. This data is comprised of 6,872,617 names and their corresponding gender, nationality and date of birth. Out of 6,872,617 names, 45.5% (3,128,688) are labeled as females and 54.5% (3,743,929) as males. Our data is comprised of 217,360 unique first names of which 17,079 unique names are occurring more than ten times in our data. Out of the 217,360 unique names, a total of 99,243 are uniquely labeled as females, whereas 110,905 are uniquely labeled as males. The remaining 7,212 names have certain frequency of being male and female. Of the 17,079 names occurring more than ten times in our dataset, a total of 7,807 are uniquely male names, whereas 6,450 are female names and the remaining 2,822 names of our samples have instances of male as well as female first names in our data. We computed the probabilities of such first names being male or being female. Figure 1. illustrates the percentage share of females between 1900 and 2016. Our data shows a rising number of females from period 1900 to 1908 with female population reaching 70% in 1907.

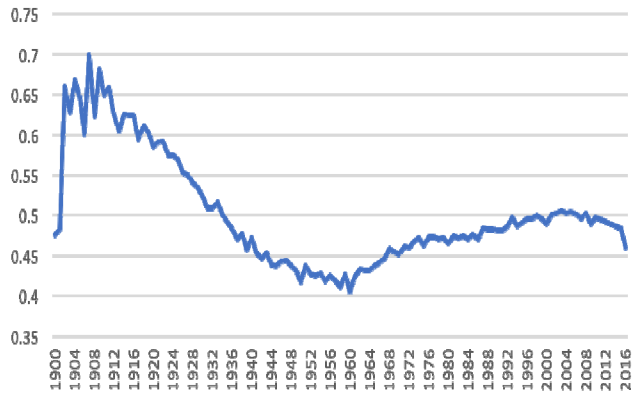

Figure 1. Percentage of females over the years in the data sample

B. *Sample distribution*

Further exploring our data set, we test whether naming conventions have changed over time. Naming practices mostly show consistency, with the exception of a few first names such as Gabriele, Michele, Dominique, Deniz, Isa or Kim. These names occur frequently as male or female names. We find that some of these naming conventions for gender have changed over the years, partially even within one decade. Figure 2. shows the share of females named "Gabriele" in our database, relative to total number of people named "Gabriele" born in the respective year. As indicated by Figure 2., the name "Gabriele" used to be carried predominantly by females in early 1900s.

In 1920-1940 it was a unisex name, while from 1960s onwards, "Gabriele" became a predominantly male name. Figure 3. shows a sharp decline of "Gabriele" as a female name from 1961 onwards and from 1985 on, we found only rare instances of "Gabriele" as female name in our data.

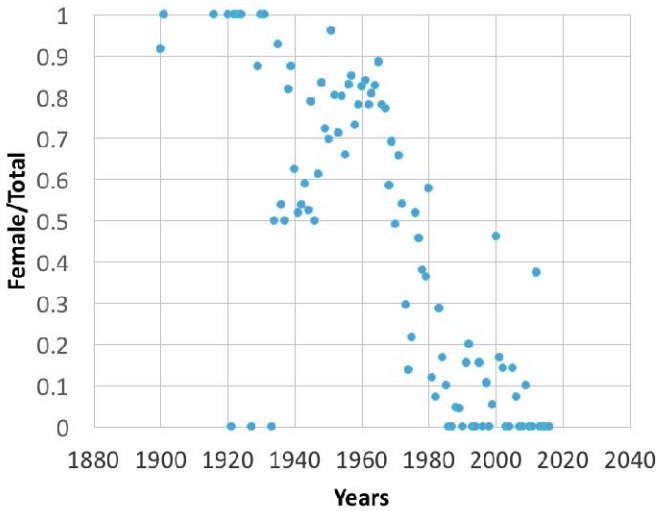

Figure 2. Female share of the name Gabriele

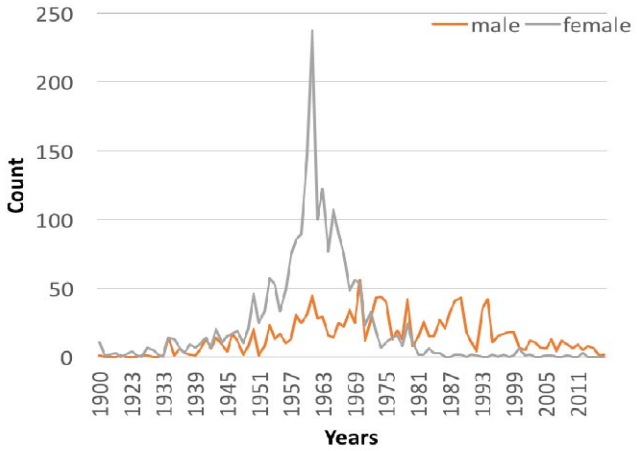

Figure 3. Change in frequency of the name Gabriele

Finally, we explored differences in naming convention across 37 different countries. We found that a few names such as Andrea, Luca, Nicola showed particular differences in conventions and had sufficiently large counts (>100) for evaluation. For illustration, Figure 4. shows the percentage male and female share of the name "Andrea" in countries Switzerland (CH), Italy (IT), Germany (GE), with counts of 29,448, 1,171, 790, 190. In Italy, people named "Andrea" are predominantly males, whereas in Switzerland, the name is male.

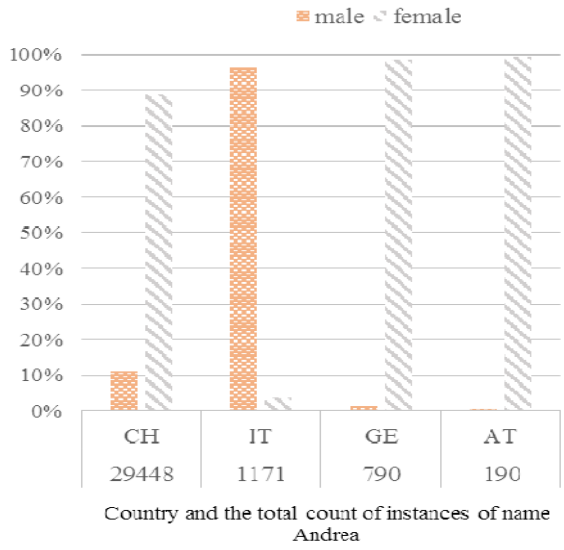

Figure 4. Nationality difference for the name Andrea

C. Test sample 1: First names of vehicle insurees

Out of the 6,872,617 observations, we randomly selected 20,000 instances having name, decade of birth, nationality and gender. We excluded those observations before proceeding in calculating probabilities of our mapping algorithm, allowing to evaluate our model performance with a test set of observations, which the model has not seen.

We take this random set of 20,000 rows to test our gender prediction algorithm. This set has 2,634 unique names, 8,512 unique name, decade combinations and 9,800 unique name, decade, nationality combinations. Out of these 20,000 instances 9,731 are female instances and 10,261 are male instances.

D. Test sample 2: First names of Swiss patent inventors

To test and evaluate our model's performance in a second real life scenario, we downloaded all patents which were filed in Switzerland and available in the database provided by the Lens, an independent nonprofit institute of Cambia [39]. The patent dataset is comprises of detailed information of over 15,000 Swiss patents from the years 2009-2017. Each patent has detailed information such as the applicant name, type of the patent, inventor names, title, number of citations, application date and publication date. We have taken a random sample of 1,500 institutions, which filed a total of 1,543 patents. We apply our gender mappings to the names of the inventors, which for many patents are teams of people and compare the results with the numbers provided by Statista Switzerland [38].

E. Probabilistic frequency calculation

Our model is based on the probabilistic occurrence of names and gender (see Table I, II and III). Therefore we count the occurrence of first name labeled as male and female. From the occurrence in each category, relative to the overall counts, we then compute the male and female probability of the first names being categorized male (P(m)) and female (P(f)), as illustrated in Table I. for the name 'Peter'.

TABLE I. DISTRIBUTION OF THE NAME PETER

Name	***Male (m)***	***Female (f)***	***Total (t=m+f)***	***P(m)***	***P(f)***
Peter	67,759	52	67,811	.999	.001

We then further subgroup the two categories by their birth decade. From the occurrence in each subgroup, relative to the overall counts, we then compute the male and female probability of the first name classified male and female, as illustrated in Table II. for the name 'Gabriele'.

TABLE II. DISTRIBUTION OF GABRIELE BY DECADE

Name = *Gabriele*	***Male (m)***	***Female (f)***	***Total (t=m+f)***	***P(m)***	***P(f)***
Decade:					
1900	0	9	9	0	1
1910	3	12	20	.15	.85
….	…	…	…	…	…
1990	180	6	186	.96	0.04

Finally, we distinguished nationalities for all the subgroups. From the occurrence in each subgroups, relative to the overall counts, we then computed the male and female probability of the first name classified male and female, as illustrated in Table III for name 'Andrea'. We assume that adding decade and nationality will improve the mapping of name and gender for some of the naming exceptions, which we identified previously (see section *A. Dataset*).

TABLE III. DISTRIBUTION OF ANDREA BY NATIONALITY

Name = *Andrea* Decade = *1970*	***Male (m)***	***Female (f)***	***Total (t=m+f)***	***P(m)***	***P(f)***
Nationality					
CH	718	8,313	9,036	.07	.93
AT	1	67	68	.01	.99
…	…	…	…	…	…
IT	386	3	389	.99	.01

F. Mapping of Name to Gender

We mapped the names to gender based on the male or female probability in Table I, II and III. First, we label our naming dataset binary as male and female, depending on a probability of more than 50%. Second, we add another label to those which showed dispersion of gender for each names. We choose to label all names, with a probability of less than 0.95 of being either male or female as unisex names. For example, in the case of a binary categorization, "Gabriele" in row 2 of Table II. is labeled as a female name as the probability of "Gabriele" being female is greater than that of being male according to our dataset. But in case of three category labeling, we assigned a unisex label because the probability of "Gabriele" being male and female is less than 0.95 for some of the decades. We performed this labeling procedure for all entries of Table I, II and III.

G. Testing of categorization methodology

We estimate the gender of our test set of 20,000 observations (see Test sample 1) by mapping first names from these rows to first names in predicted labels for the two and three categorization. We compare the results against the

ground truth for the binary and 3 labeled categorization for Table I, II and III. For binary categorization, we compute true positives (tp) as the number of males and females which are actually identified as male and female respectively by our predicted labels [40]. For the 3-label categorization, we omit the names which are being labeled as unisex from the random sample, and then compute true positives (tp) as males and females which are actually identified as male and female respectively by the predicted labels. We assume that names with unisex labels cannot be predicted with the information of name, age or nationality. Therefore, omitting unisex labels will reduce the number of samples from the random set but precision of model should increase as we try to predict the genders of only those names whose probability associated with gender is greater than 0.95. We further calculate precision, recall and F1-Score for all three tables in both the categorizations which allows the comparison of the results of the algorithms according to Formula 1-3.

$$Precision = tp/(tp+fp) \quad (1)$$

$$Recall = tp/(tp+fn) \quad (2)$$

$$F1\text{-}Score = 2* (precision * recall) / (precision + recall) \quad (3)$$

Using the binary categorization approach, we also apply our mapping procedure to categorize test sample 2, estimating the share of inventors by gender. We use the calculated probabilities from Table I, as attributes such as nationality or year born is not published with patent data. As we do not know the actual gender of the inventors, we can only compare our mapping results with other studies using aggregated numbers, hence continue without calculating the evaluation metrics precision, recall and the F1-Score.

IV. RESULTS

We apply our mapping procedure in a 2-label, also known as binary categorization and 3-label categorization setting. We further calculate the recall, precision and F1-Score from both kinds of categorization. Then, we apply our mapping to identify the gender of patent applicants.

A. Results of 2-label categorization

In the binary categorization setting, we matched 19,917 names from Table I. We matched 19,910 name, decade combinations from Table II and 19,920 name, decade, nationality combinations of our test set from Table III. Out of the 20,000 instances in our test set 1, 10,543 were males and 9,374 were females. The results of our binary categorization can be found in Table IV. With only first names, a total of 10,063 out of 10,543 (95.44%) males were identified correctly. This percentage of true positive matches increases for both male and female by adding decade and nationality attributes to the name. Thus, we see precision increases from 96.30% to 96.98% by adding decade to first name, and to 97.02% by adding nationality to name and decade. The F1-Score also increases with the addition of decade and nationality to name, which indicates that the name to gender mapping augmented by decade and nationality is more accurate than the model with just name or name and decade.

TABLE IV. MODEL PERFORMANCE 2-LABEL

Label	***Name***	***Name, Decade***	***Name, Decade, Nationality***
Males (tp)	10,063	10,131	10,184
Females (tp)	9,119	9,179	9,221
Total (tp)	19,182	19,310	19,405
Recall (%)	95.91	96.55	97.02
Precision (%)	96.30	96.98	97.41
F1-score (%)	96.10	96.76	97.21

B. Results of 3-label categorization

In the 3-label categorization, we were able to match 19,992 names; 19,979 name, decade combinations and 19,987 name, decade, nationality combinations out of 20,000 instances of the random sample with our datasets from Table I, II and III. respectively. For evaluation purposes, we only considered the male and female labels which reduced our sample to 18,354 names; 18,759 name, decade combinations and 19,050 name, decade, nationality combinations. With just first names 8,410 females and 9,547 males out of 8,506 and 9,848 respectively were identified correctly. The precision of the model increased from 97.83% to 99.72% by adding decade and nationality to predict the first names. The F1-score also increases from left to right, which we observed in binary categorization also.

TABLE V. MODEL PERFORMANCE 3-LABEL

Label	***Name***	***Name, Decade***	***Name, Decade, Nationality***
Males (tp)	9,547	9,785	10,155
Females (tp)	8,410	8,642	8,842
Total (tp)	17,957	18,427	18,997
Recall (%)	89.78	92.13	64.98
Precision (%)	97.83	98.23	99.72
F1-score (%)	93.63	95.08	97.29

The precision of the 3-label categorization is higher than the precision of the binary categorization in all three cases. However, the recall is lower in the 3-label categorization due to elimination of unisex labels. The F1-Score also is lower for 3-label categorization due to lower recall values.

C. Results of Gender identification of patent inventors

We evaluated a total 1,543 patents, identifying the gender of 2,058 inventors. We successfully matched 2,058 first names from our database and using the name matching binary categorization, we identified 1,829 (89%) males and 229 (1%) female inventors.

V. Discussion

In this research study we try to accurately identify the gender attributed to the instances of a data set collected by an insurance company and to a data set we downloaded from a patent database. We do this by creating a mapping table which consists of more than 1 million unique user records, which is representative to the population of Switzerland (8.4 million in 2015, [41]). Our goal is to show that adding information such as nationality and year of birth to the mapping table can improve the precision of a categorization algorithm. This is due to the fact that the conventional gender of some names has switched over time as well as due to differences in naming practices by nationality. In our two class categorization study, we find that adding age of birth as well as nationality to our mapping table improves the recall score of the mapping from 95.91% to 97.02% (1.1%), a relative decrease in the error rate by 27.1%. In our three class categorization, adding decade of birth and nationality improved our recall scores from 89.78% to 94.98% (5.2%). This is a relative decrease in the error rate by 49.6%. Applying our two class categorization to our patent set for Switzerland, we find that 11 % of all inventors are woman and 89% are men, which is identical to the findings published by Statista Switzerland [38], further validating our mapping results. Hence, we find that the mapping methodology of name to gender lookup tables are a suitable way to categorize gender by name, however adding nationality and age, can improve the recall and precision scores in a binary categorization as well as 3 class categorization setting compared to simple mappings by probabilistic occurrence of first names.

We further show how the results of a two-class (male, female) categorizations compared to a three-class (male, female, unisex name) categorization differ by evaluating the precision and F1-score. We find that in a setting where the first name is available and a high categorization precision is desired, a three-class categorization score improves from 97.83% to 99.72% , an increase of 1.89% by augmenting decade and nationality to first names. An example of setting in which a high precision is desirable may be illustrated by an organization which plans to approach users by “Mr” or “Mrs” (i.e. in letters, E-mail or chat-bots) and may suffer from reputational damages, if the salutation does not correspond with the actual gender of the user. On the contrary, in a setting where one strives for the highest recall value, a two class categorization enriched by nationality and year of birth seems to be the most promising approach, achieving a F1-Score of 97.21%, the highest outcome of all our evaluations scores. A setting where a high recall is desirable could be found in an online marketing campaign with limited amount of participants, where an advertisement banner needs to be displayed to users identified as male and another one to users identified as female. In the latter scenario, the damage of a wrongly categorized user is much smaller and may not even be recognized by the user.

In this study, we apply a humanistic starting point, the historical relationship between naming practices and gender. This way, we aim contribute to advances in the methodology for data analysis. Our method uses an abundant existing database from a Swiss car insurer, adding additional information generally found in a vehicle insurance policy to compose a first name gender mapping. Our contribution is to incorporate a historical method, and take into account how gender naming practices have changed over time, and how some first names are attributed to both sexes depending on nationality. When high recall is required, our temporal, nationality specific approach provided much improved precision results over the simpler, anachronistic, ahistorical lookup method. As such, our method illustrates the pragmatic trade-off between complexity and discovery. Its reliance on a male vs. female binary categorization dampens the complexity of gender identity, while the three label categorization benefits most when adding additional information. In this more complex case, incorporating additional information in the mapping also yields the best F1-Score, however increases the complexity of the model (compare [8]).

We encourage other researchers to study how changes in naming conventions in other countries affect the performance of first name to gender mappings and if the addition of nationality, year of birth or other attributes can improve mapping results. A joint effort of researchers around the world, merging their name to gender mappings would be highly beneficial for all studies of mixed nationalities aiming for high precision. Especially for the years between 1900-1935, we encourage researchers to replicate our study with other mappings, as we see a skew towards female instances in our data set for these years. Further, our results only discuss gender as stated in legal documents and do not represents an individual's self-perceived endorsement of masculine and feminine personality traits, and as such, may or may not be congruent with a self-assigned gender of an individual or the biological sex unrecognized in legal documents. We encourage fellow researchers to evaluate self-perceived gender in ways, which are less reliant on historical naming practices and would further complete gender identification methodologies through data analysis.

REFERENCES

[1] P. Oliveira, F. Henriques, and P. Henriques, "A formal definition of data quality problems," *2005 Int. Conf. Inf. Qual.*, pp. 1–14, 2005.

[2] G. K. Tayi and D. P. Ballou, "Examining data quality," *Commun. ACM*, vol. 41, no. 2, pp. 54–57, 1998.

[3] Y. Wang, V. C. Storey, and P. Firth, "A framework for analysis of data quality research," *IEEE Trans. Knowl. Data Eng.*, vol. 7, no. 4, pp. 623–640, 1995.

[4] K. Orr, "Data Quality and," no. 1, 2000.

[5] L. Gu, R. Baxter, D. ickers, C. ainsford, and C. M. and I. Sciences, "Record Linkage: Current Practice and Future Directions," *Tech. Rep.*, no. June 2003, 2003.

[6] Y. Qin, S. Zhang, X. Zhu, J. Zhang, and C. Zhang, "Semi-parametric optimization for missing data imputation," *Appl. Intell.*, vol. 27, no. 1, pp. 79–88, 2007.

[7] C. Duhigg, "How Companies Learn Your Secrets," *N. Y. Times Mag.*, pp. 1–16, 2012.

[8] B. Saha and D. Srivastava, "Data quality: The other face of Big Data," *Proc. - Int. Conf. Data Eng.*, pp. 1294–1297, 2014.

[9] A. Alharthi, V. Krotov, and M. Bowman, "Addressing barriers to big data," *Bus. Horiz.*, vol. 60, no. 3, pp. 285–292, 2017.

[10] C. Blevins and L. Mullen, "Jane, John … Leslie? A Historical Method for Algorithmic Gender Prediction," *Digit. Humanit. Q.*, vol. 9, no. 3, pp. 1–19, 2015.

[11] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Sci. J.*, vol. 14, no. 0, p. 2, 2015.

[12] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data quality in context," vol. 40, no. 5, pp. 103–110.

[13] L. Zhou and T. Wang, "Socialmedia: a New Vehicle for Citymarketing in China," *Cities*, vol. 37, pp. 27–32, 2014.

[14] J. Hunt, J. Garant, H. Herman, and D. J. Munroe, "Why are women underrepresented amongst patentees ?," *Res. Policy*, vol. 42, no. 4, pp. 831–843, 2013.

[15] J. Milli, B. Gault, E. Williams-Barron, J. Xia, and M. Berlan, "The gender patenting gap," no. July, pp. 1–10, 2016.

[16] J. Simkins-Bullock and B. Wildman, "An Investigation into the Relationships Between Gender and Language," vol. 24, pp. 149–160, 1991.

[17] S. Argamon and C. Science, "DHQ : Digital Humanities Quarterly Vive la Différence ! Text Mining Gender Difference in French Literature," vol. 3, no. 2, pp. 1–11, 2017.

[18] W. K. Darley, R. E. Smith, W. K. Darley, and R. E. Smith, "Gender Differences in Information Processing Strategies : An Empirical Test of the Selectivity Model in Advertising Response Gender," vol. 3367, no. October, 2017.

[19] O. Shopping, "Consumers' Attitude towards Online Shopping."

[20] S. Selfie-promotion, "Gender Differences in Internet Use Patterns and Internet Application Preferences: A Two-Sample Comparison," no. September, 2015.

[21] S. Xue-wui, N. I. E. Gui-hua, and S. Ling, "Gender-Based Differences in the Effect of Web Advertising in E-business," no. 70572079, 2000.

[22] K. Sattler and P. O. Box, "based on a Multidatabase Language," pp. 219–228, 2001.

[23] I. Chen and K. Popovich, "Understanding customer relationship management (CRM): People, process and technology," *Bus. Process Manag. J.*, vol. 21, no. 2, pp. 191–206, 2017.

[24] D. Shasha, H. Galhardas, C. Saita, E. Simon, and I. Rocquencourt, "Improving Data Cleaning Quality using a Data Lineage Facility," *Informatica*, pp. 1–13, 1996.

[25] L. L. Pipino, Y. W. Lee, R. Y. Wang, M. W. Lowell Yang Lee, and R. Y. Yang, "Data Quality Assessment," *Commun. ACM*, vol. 45, no. 4, p. 211, 2002.

[26] O. O. R. Data, Q. Can, H. A. Severe, I. On, and T. H. E. Overall, "Wand, Yair Wang, Richard Y.," 1996.

[27] J. Hu, H. Zeng, H. Li, C. Niu, and Z. Chen, "Demographic Prediction Based on User ' s Browsing Behavior," pp. 151–160, 2007.

[28] D. U. Interface, U. W. A. Psychology, W. Number, C. N. Meaningfulness, P. N. Age, W. Type, C. Irregular, N. Nphon, N. K. K. Number, and T. B.-F. Thorndike-lorge, "MRC Psycholinguistic Database," pp. 1–4, 2017.

[29] W. Deitrick, Z. Miller, B. Valyou, B. Dickinson, T. Munson, and W. Hu, "Author Gender Prediction in an Email Stream Using Neural Networks," *J. Intell. Learn. Syst. Appl.*, vol. 4, no. August, pp. 169–175, 2012.

[30] A. Goldstone and T. Underwood, "The Quiet Transformations of Literary Studies : What Thirteen Thousand Scholars Could Tell Us Welcome to Project MUSE Connect with Project MUSE," no. 0, pp. 9–10, 2017.

[31] L. D. Cook, "Explorations in Economic History Inventing social capital : Evidence from African American inventors , 1843 – 1930 ☆," *YEXEH*, vol. 48, no. 4, pp. 507–518, 2011.

[32] D. Tannen, “Gender differences in topical coherence: Creating involvement in best friends’ talk,” vol. 13, no. 1, pp. 9–13, 2017.

[33] T. Jung and O. Ejermo, “Technological Forecasting & Social Change Demographic patterns and trends in patenting : Gender , age , and education of inventors,” *Technol. Forecast. Soc. Chang.*, vol. 86, pp. 110–124, 2014.

[34] B. F. Jones, “The Burden of Knowledge and the ‘ Death of the Renaissance Man ’: Is Innovation Getting Harder ?,” no. October, pp. 283–317, 2017.

[35] M. A. Celik, “Does the Cream Always Rise to the Top ? The Misallocation of Talent in Innovation,” 2015.

[36] C. Ludwig, S. Cavalli, and M. Oris, “‘“ Vivre / Leben / Vivere ”’: An interdisciplinary survey addressing progress and inequalities of aging over the past 30 years in Switzerland,” *Arch. Gerontol. Geriatr.*, vol. 59, no. 2, pp. 240–248, 2014.

[37] C. R. Sugimoto, C. Ni, J. D. West, and V. Larivi, “The Academic Advantage : Gender Disparities in Patenting,” pp. 1–10, 2015.

[38] Statista, “Share of patent applicants that were female in Switzerland from 1980 to 2013 Statista Accounts,” 2017.

[39] Cambia, “Lens patent database,” 2017. [Online]. Available: https://www.lens.org/lens/.

[40] D. Colquhoun, “An investigation of the false discovery rate and the misinterpretation of P values,” *R. Soc. Open Sci.*, pp. 1–15, 2014.

[41] S. Federal Statistical Office, “Federal Statistical Office - Look for Statistics,” 2017. [Online]. Available: https://www.bfs.admin.ch/bfs/en/home/statistics.html. [Accessed: 02-Mar-2017].