University of Copenhagen

Department of economics

# The Gender Gap in Invention

August 2018

# Contents

# 1 Introduction

Institutions capable of securing private ownership of intellectual property has played a crucial role in the progress of economies throughout centuries. However, ensuring intellectual property rights, embodied by patents, is an undertaking dominated by men. Rosser [2009] points out that women are underrepresented in the field of commercial science measured by the number of patents obtained by women relative to the fraction of women in science. This paper investigates to what extend this is the case today as well as the development during the recent years. We investigate this across different nationalities as well as across different fields of science.

The analysis is based on data from the European Patent Office and is carried out by isolating characteristics of the inventor's names and nationality and guessing the gender of the inventors using different lookup- and machine-learning methods. The different fields of science will be classified using unsupervised clustering methods on the abstracts of the patent applications to see how the female inventors are distributed across different fields.

It is concluded that the fraction of women inventors has been increasing during the past 10 years. That is, the fraction of women inventors increased by 36 pct. but while this sounds like a large increase, the fraction only increased by roughly 3 pct. points. The analysis of the characteristics of the patents granted in 2017 indicates that women have the highest representation of patents in the field of chemistry (20,5 pct. of inventors), while they are heavily underrepresented in the field of mechanics (5,2 pct.).

# 2 Data

The data consists of granted patents from the European Patent Office (EPO)[1] from January in the period 2008 to 2018. EPO is an intergovernmental organization with 38 member states and 2 non-member states[2]. A granted patent from the EPO covers 40 states by default and is subject to local legislation in each state. The EPO granted 105,635 patents in 2017[3].

A granted patent contains a single assignee (the owner of the patent) and one or more inventors. The applicant and inventor is not necessarily the same person. An inventor is always a private

[1] https://www.epo.org/

[2] The member states are the 28 members of the European Union, Albania, the former Yugoslav Republic of Macedonia, Iceland, Liechtenstein, Monaco, Norway, San Marino, Serbia, Switzerland and Turkey. Montenegro and Bosnia and Herzegovina are non-member extension states.

[3] https://www.epo.org/about-us/annual-reports-statistics/annual-report/2017/statistics.html

individual, the assignee may also be a company or organization. If the invention is the result of a collaborative process, it may not be obvious who should be included as inventors in the patent. The EPO or the EU does not have a common definition of inventors but rely on legislation of the individual member states[4]. In the case of multiple inventors, the individual contributions may not be equal. However, the listed order of inventors does not carry any meaning.

## 2.1 Collecting Patent Data

*There's no such thing as a free lunch*, a term popularized by the well-known economist Milton Friedman, is highly applicable to patent data. Getting information about a specific patent is free and available from multiple sources such as *Google Patents* and the *World Intellectual Property Organization* (WIPO). However, when large amounts of data are needed it becomes much more of an obstacle to obtain patent data. In the case of *Google* the number of results shown are limited to 300 for a given search and sorting which makes it almost impossible to scrape any meaningful set of data given the scope of this paper. *WIPO* on the other hand limits the possibility of scraping by having *captchas* appear after just a few minutes of a user snooping around on the website.

### 2.1.1 Open Patent Services - API

With web-scraping exhausted as a possible way to gather large amounts of patent data, the only free way of obtaining a somewhat structured dataset is with the *Open Patent Services - API* (OPS). With a weekly free limit of 4 gigabytes of data this API allowed us to download a considerable amount of patents. To access the API an application needed to be filed in which we described our intention with the data. Each of the group members applied for access to the OPS, however, only two applications were granted.[5]

The API is provided together with a whopping 150 page documentation that covers the services offered by the API. The OPS provides access to all the patents available from the EPO's database, however, the API is far from flexible and several requests was needed to be made, to end up with the data we wanted. Figure 1 shows the three steps of requesting the API. To request the API, we imported a package called *epo-ops* that took care of the authorization needed to access the OPS. The package lacked the possibility of getting patent abstracts using the *equivalents* method, which we then added by editing a line of code in the package. We

[4]https://www.iprhelpdesk.eu/kb/2593-whats-difference-between-inventor-applicant-and-owner-patent

[5]Due to a maximum of one applicant per institution of education

ended up only extracting all abstracts for January 2017, since it was a very "expensive" request, where each response resulted in the same patent abstract appearing up to 10 times, in different languages and under different patent names.

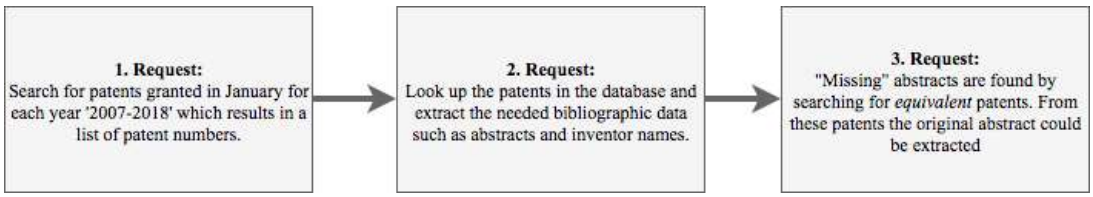



Figure 1: The requests made to collect bibliographic data and abstracts

The response from each request is received in `XML`-format - below is a snippet of what such a response looks like.

```
<publication-reference>
        <document-id document-id-type="docdb">
                <country>EP</country>
                <doc-number>1000000</doc-number>
                <kind>A1</kind>
                <date>20000517</date>
        </document-id>
        <document-id document-id-type="epodoc">
                <doc-number>EP1000000</doc-number>
                <date>20000517</date>
        </document-id>
</publication-reference>
```

We parsed this `XML`-format into a much more manageable format - `JSON`. To do this we converted the data into a dictionary by using the library called `xlmtodict`. With the dictionary we could load it as a `JSON`-format. Combining the responses provided us with the following variables of data for each patent:

- Document number
- Kind code [6]
- ID number (Document number + kind code)
- Published date

[6]For different types of kind codes: https://worldwide.espacenet.com/help?locale=en_EP&method=handleHelpTopic&topic=kindcodes

- IPC Code
- Applicant name
- Inventor(s) name
- Invention title
- Abstract
- Abstract language

Some of these variables were difficult to handle e.g. inventor's name. This was due to some patents having either none, one or multiple inventors. When converting these `XML` formats to `JSON` format the same variables could end up being different types i.e. list, string or `NaN`. Since we needed a massive amount of patents we had to come up with a generic solution to these different types of variables. With the generic solution we managed to create a `pandas.DataFrame` for each patent with identical structure. An example of a `DataFrame` is provided below:

| **Published date** | **IPC Code** | **Applicant name** | **Inventors name** | **Invention title** | **Abstract** | **Abstract language** |
|---|---|---|---|---|---|---|
| 19930908 | [G07C1/10, G07C9/00103, G07C9/00111] | BAUER KABA AG [CH] | ['HAECHLER CARLO\\2002[CH]', 'LOCHER JOHANN K\\2002[CH]'] | Individual identification system. | The basis of obtaining... | en |

Table 1: Database indexed on 'ID number', ID number = EP0559605A1

This dataset was then exported as a `CSV` in a tidy format, where each row contained a unique ID number.

## 2.2 Data parsing and restructuring of database

### 2.2.1 Data parsing

With inventor names appearing in the following manner - *'HAECHLER CARLO\\2002[CH]'*, we had to use `RegEx`[7] to extract usable information about the inventors. For this instance we employed two different `RegEx` expressions, one for isolating the inside of the square brackets (nationality) and one for isolating everything behind the double backslash. With the full names we could now split into first-, middle- and surname for later use on gender classification.

[7]Regular expression is a sequence of characters that define a search pattern

### 2.2.2 Restructuring the database

To make the database more manageable for gender classification we restructured it to be indexed on ID number and full name, such that each row contained a unique ID number *and* name combination. Table 1 is a snippet of the new `dataframe` for the same patent:

| **Document name** | **First name** | **Surname** | **Country** | **Days untill granted** | **Year granted** | **Abstract** | **Abstract language** |
|---|---|---|---|---|---|---|---|
| EP0559605 | CARLO | HAECHLER | CH | 5242 | 2018 | The basis of obtaining... | en |
| EP0559605 | JOHANN | LOCHER | CH | 5242 | 2018 | The basis of obtaining... | en |

Table 2: Database indexed on 'ID number' and 'Inventor name'

At this point we had 80,831 unique ID numbers, but as we did not want the same patent multiple times under different kind codes, we only kept the patent with the lowest kind code. This left us with 60,595 unique patents.

## 2.3 Descriptive statistics

As mentioned earlier we could not extract all abstracts via the standard bibliographic request. We therefore made a separate request for abstracts to extract all abstracts from January 2017, which we then merged on the existing dataset.

The final dataset therefore consists of 60,595 unique patents for January from 2008-2018. Abstracts for every patent in January 2017 is included. The dataset is indexed on ID number and inventor names and has 166,917 inventors, with an average of 2.75 inventors per patent. The dataset contains the following relevant variables:

- ID number
- IPC Code
- Inventor first name
- Inventor surname name
- Days until granted
- Abstract
- Abstract language

Figure 2 shows a histogram over the number of days from an application for a patent is filed until it is granted. We see the most common grant time is between 1 and 1.5 years. The average

application takes a mean of 3.9 years to be granted, which means our data on average shows a snapshot of how things were 3.9 years ago, which is worth considering when interpreting the results.

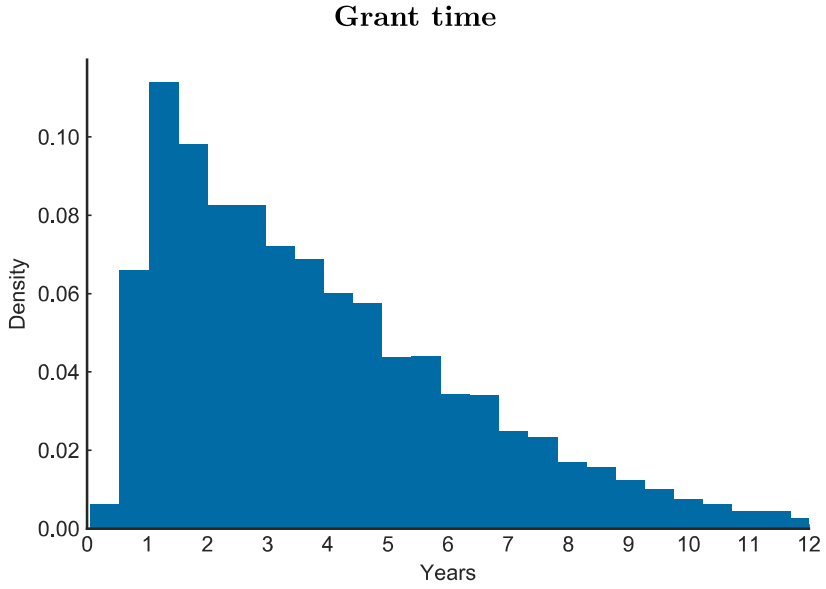



Figure 2: Distribution of time for a patent application to be granted.

### 2.4 Ethics

The main data collected was obtained using the EPO API. The European Patent Office granted the use of the API after we applied for permission to use it. Apart from the inventor names, we did not obtain any personal information on the applicants. Obtaining data classifying the most likely nationality of a person based on this person's last name was obtained by scraping http://www.forebears.co.uk/surnames. Doing this, we made sure to include a `time.sleep` function in order to avoid overwhelming their servers with traffic and thereby not make a DoS-attack. In sum, we did not face any challenges in regard to complying with the general code of conduct doing this assignment.

## 3 Classification of Gender

### 3.1 Classification using look-ups

Trying to classify the gender of the inventors at first seemed like a trivial problem, but the fact that some names are used for different genders in different countries made the matter quite

complex. Take for example the name Andrea, which is typically a female name in Denmark and in The United States, but typically a male name in Italy. The same is true for several other names. This was solved using a dataset containing 6.2 million records of first names and a classification of gender that depends on both the first name and the nationality of the person [Raffo, 2016][8]. Applying this method means we will classify gender in a binary way, leaving out the "unisex" classification. This is done due to the limited time scope of the paper and under the assumption that we will guess the wrong gender equally many times for both genders[9]. For rows where data on the nationality of the inventors were not present, we wrote a `request` that connects to a website capable of classifying the most likely country of origin based on last names[10]. The most likely country of origin based on the persons last-name is then assigned to the inventor and from that the most likely gender is assigned. An example of the resulting dataset[11] is shown in Figure 3.

| | Firstname | Surname | gender |
|---|---|---|---|
| 0 | ANDREA | KNUDSEN | F |
| 1 | ANDREA | PIRLO | M |
| 2 | ANDREAS | HANSEN | M |
| 3 | KIM | KRISTIANSEN | M |
| 4 | KIM | KARDASHIAN | F |

Figure 3: Example of gender-classification based on first- and surname

The genders not classified after this step were classified using a separate dataset from the same source that after filtering 'unisex' names contained about 172,000 records of names that 'unambiguously' can be classified as either a male or a female name[12]. 21,738 data-points could not be classified using the lookup method. We use a Random Forest model to classify the remaining names in the dataset which is covered in the next section.

[8]The dataset can be found on: https://econpapers.repec.org/software/wipeccode/10.htm

[9]This in turn means the distribution will be skewed if the distribution of men and women in the sample is not the same, but we will ignore this and continue

[10]Link: http://forebears.co.uk/surnames?q=Jensen

[11]We used dummy-data to make the point clear

[12]https://econpapers.repec.org/software/wipeccode/10.htm

## 3.2 Random Forest Classifier for unidentified names

### 3.2.1 The Random Forest in general

The Random Forest model combines the features of the decision tree, bootstrapping and feature bagging. A decision tree minimizes a cost function, e.g. *the sum of squared errors* by splitting the p-dimensional space one dimension at a time, where $X \in R^p$ is the variables of the model. Intuitively, the model finds a local minimum on each step of the tree by splitting a variable on a criterion in a given step. Regression trees can approximate discontinuous functions and are generally very flexible, which means they are prone to overfitting if many splits are included [Hastie et al., 2009]. In general terms, overfitting creates a high error from the variance in the data and arises when a model incorporates noise when modeling the data-generating process. This means it will perform well on training data, while it will perform poorly on the test data. In other words, the model is unable to generalize the data, so it only performs well on data it 'knows'. On the other hand, if the model is too simple, it will suffer from a high bias, which means the model will not be flexible enough to model the data-generating process. The designer of a model is thus faced with a variance-bias trade-off. By using bootstrap aggregation and feature bagging it is possible to reduce the error from the variance term without increasing the bias [Hastie et al., 2009]. Bootstrap aggregation (bagging) works by creating $B$ subsamples and calculate the estimate for each sample. Having calculated $B$ estimates, where a single estimate as a function of a subsample of data is $\hat{\theta}^B = f(x_i^B)$, the bagging estimate equal to the average of $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_B$ becomes:

$$\hat{\theta}_{bagging} = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j \tag{1}$$

Implementing bagging in a regression tree is, however, not unproblematic since each tree is not uncorrelated, but by estimating each tree from a random subsample of variables(features) - so called feature bagging - the trees are 'decorrelated'[Hastie et al., 2009]. The estimate of a random forest model as a function of the data, $x$ will thus become:

$$\hat{f}^b(x) = \frac{1}{B}\sum_{b=1}^{B}T(x;\Theta_b) \tag{2}$$

Where $\Theta_b$ is the parameters of a given tree and $B$ is the number of Bootstrap samples. Hastie et al. [2009] recommends using $\sqrt{N}$ features[13] for every bagging estimate, which is the default setting in the `SciKit-learn` module.

[13] Where $N$ is the total number of features

### 3.2.2 The Random Forest for classifying gender by name

We chose only to look at the first name of each data point and classify the names by extracting the features of these. That is - we extract the first and the second letter, the last and the second to last letter and the frequency distribution of each letter. Additionally, we extract the frequency distribution of each pair of letters appearing sequentially for each name. Each of these features are then vector-encoded using the One-Hot method and we are left with a total of 806 features. This way of extracting features from the names rely on the assumption that male and female names are characterized by distinct characteristics; if a name ends on the letter A for example, it is more likely a female name.

We use the dataset used for the look-up method containing 172,000 names that could unambiguously be classified as a female or a male name. Since the dataset is reasonably large, we refrain from doing a k-fold Crossvalidation and estimate the model using 120,000 names in the training set. The model is then validated on 30,000 names afterwards. All the 172,000 names in the dataset are shuffled before choosing the train- and the test set in order for the model to capture as many different patterns in the data as possible. We also specify the *n_estimators* which is the number of decision trees estimated and the *min_sample_split* which is the number of subsamples used to split a node in each decision tree. These hyperparameters were chosen based on 9 estimations of the model with every combination of 50, 100 and 150 for the parameter *n_estimators* and 10, 20 and 30 for the parameter *min_sample_split*. The model that performed best on the test data was chosen, which was *min_sample_split=150* and *min_sample_split=20*. The prediction accuracies for the different hyperparameters can be seen in Table 3.

| | | *n_estimators* | | |
|---|---|---|---|---|
| | | 50 | 100 | 150 |
| *min_* | 10 | 0.8534 | 0.8571 | 0.8546 |
| *samples_* | 20 | 0.8552 | 0.8549 | 0.8591 |
| *split* | 30 | 0.8489 | 0.8514 | 0.8534 |

Table 3: Accuracy scores for diffent hyperparameters

We get an accuracy score of 85.9 pct., which we consider reasonable, so we continue the classification of the remaining names in the data, keeping in mind the possible discrepancies for the final analysis. The most important features determining the gender can be seen in Figure 4.

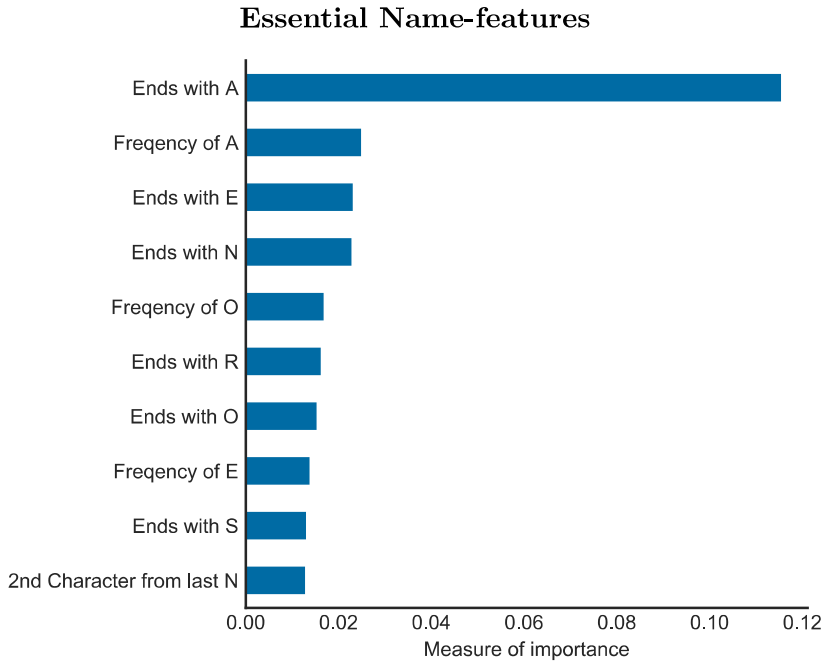



Figure 4: The top 10 most important features of a name determining the gender of a person

To sum up, we chose to classify the genders by running the data through the following steps:

1. We looked up the gender conditional on the nationality of the person using the dataset containing information on nationality (93,573 genders classified).
2. If the persons nationality did not appear on the application, we retrieved this information using a web-scraping technique and repeated step 1 (23,983 genders classified).
3. If the classification still was not successful, we looked up the name in a dataset containing data that was filtered to only contain names that unambiguously can be classified as either male or female (27,856 genders classified).
4. The datapoints still not classified were classified using the Random Forest classification. (21,505 genders classified)

# 4 Categorizing patents with topic modeling

The patent data is organized in nested categories based on the International Patent Classification (IPC). However, these categories may be too broad or narrow to get a sense of what the contents of the patent is. A new popular research area may not have a category yet. Furthermore, each patent may be assigned to several subcategories, which makes it difficult to separate the patents

into unique categories or fields. As an alternative classification scheme, we use an unsupervised clustering model on the patent abstracts.

The unsupervised clustering model explores existing patterns in the data and separates the abstract into different clusters. The exact separation is dependent on the chosen model and parameters. It is not possible to formally test the accuracy of the categorization, so multiple model specifications should be tested and the final categorization should be treated with caution. Nevertheless, topic modeling allows us to explore another dimension of the data, which would otherwise demand a cumbersome manual analysis.

As our dataset does not contain patent abstracts for the full time period, this part focuses on patents granted in January 2017.

### 4.1 Text vectorization

To perform the clustering analysis on the abstracts, the text must first be vectorized. Each abstract is tokenized into individual words (bag of words). These word tokens are lemmatized to remove inflectional endings and reduce each token to its dictionary form. This is done by looking up the tokens in the WordNet database. By default, the WordNet lemmatizer assumes that words are nouns. To also lemmatize verbs and adverbs, we run the lemmatizer again on tokens that fail to reduce with the proper verb or adverb setting. Another approach is stemming, which is algorithmically based, but the dictionary approach performed better in our case. Finally, common English words (*stopwords*) are filtered out. A word frequency matrix is generated, the document-term matrix, which contains a row for each document (abstract) and a column for each unique token.

A simple token count favors longer abstracts and common words. To remedy this, Term Frequency - Inverse Document Frequency (TF-IDF) weights are applied to reduce the influence of term count in each abstract and to decrease influence of common terms across all abstracts. The TF-IDF weights are defined as:

$$\text{TF-IDF}(t_d) = \frac{f_{t,d}}{f_d} \log \frac{N}{N_t}$$

$f_{t,d}$ : is the count of term $t$ in document $d$
$f_d$ : is the total term count in document $d$
$N$ : is the number of documents
$N_t$ : is the number of documents containing the term $t$

We use a custom word tokenizer and include all words with two characters or more and exclude numbers and words containing numbers. The TF-IDF weights are calculated using the `TfidfVectorizer` module from `SciKit-learn`.

### 4.2 Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is a technique that can be used for topic modeling. It factorizes an approximation of the document-term matrix $X$ into two matrices, a term-feature matrix $W$ and a feature-document matrix $H$, i.e. $X \approx WH$. The features are the topics (or clusters) into which the documents are sorted. Thus, the rows in the feature-document matrix are the topics and the columns are the documents. Each document can be related to several topics as indicated by the weights in the feature-document matrix. A factorization that approximates the document-term matrix well, must group terms together that coincide in several documents. This means that each feature can be interpreted as a topic, as the documents relating to the feature share common words not included in other documents.

Prior to the factorization the desired number of topics $k$ must be specified. We will later determine an optimal value for $k$. The factorization algorithm minimizes the distance between $X$ and $WH$. The distance measure used is the Frobenius norm. In practice, we use the `NKF` module from `SciKit-learn` for factorization. After factorization each document is assigned the topic with the highest weight.

To determine an optimal number of topics $k$ we factorize for values of $k$ between 4 and 18 and evaluate each model by comparing with a topic coherence measure based on a `word2vec` skip-gram model of the abstracts [O'Callaghan et al., 2015]. `word2vec` is a machine learning model that is trained on complete word sentences. Compared to a bag-of-words model, it also uses the position of a word within a sentence as information from the text data. Words that often appear in similar contexts (close to the same words in a sentence) are classified as being similar or related. `word2vec` creates a vocabulary of each unique word. It trains a neural network with word pairs of each unique word and the $z$ words that appear before and after in a sentence (we use $z = 5$). This allows the trained model to predict the probability of one word appearing near another word in a sentence. The trained model contains a weight matrix with a unique row of weights for each word. Similar words ought to give similar predictions of surrounding words, so similar words have similar rows of weights in the weight matrix. Thus, the model can infer similarity by comparing the weight rows for each word. Specifically, the cosine similarity between each weight vector is the measure of similarity between a word pair.

In our case the `word2vec` model is trained on individual abstracts as sentences. The abstracts are tokenized by the tokenizer used for the document-term matrix and the same stopwords are excluded. It is not TF-IDF weighted however. We used the module `word2vec` from `gensim`. The mean coherence score TC-W2V is calculated for each NMF model based on different values of $k$.

$$\text{TC-W2V} = \frac{1}{\binom{N}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} similarity(t_j, t_i)$$

It calculates the mean of the pairwise similarity scores between the $N$ highest ranked terms in each topic. In our case $N = 10$. The topic coherence score for the model is the mean across all topics. A model with a relatively high mean coherence score has greater coherence within its topics.

Figure 5 below shows the calculated mean coherence score for values of $k$ between 4 and 18. The highest score is for $k = 11$, so we choose a model with 11 topics. We note that an unsupervised model is very sensitive to the chosen parameter values. Different configurations resulted in slightly different topics and keywords, but the basic structure was similar.

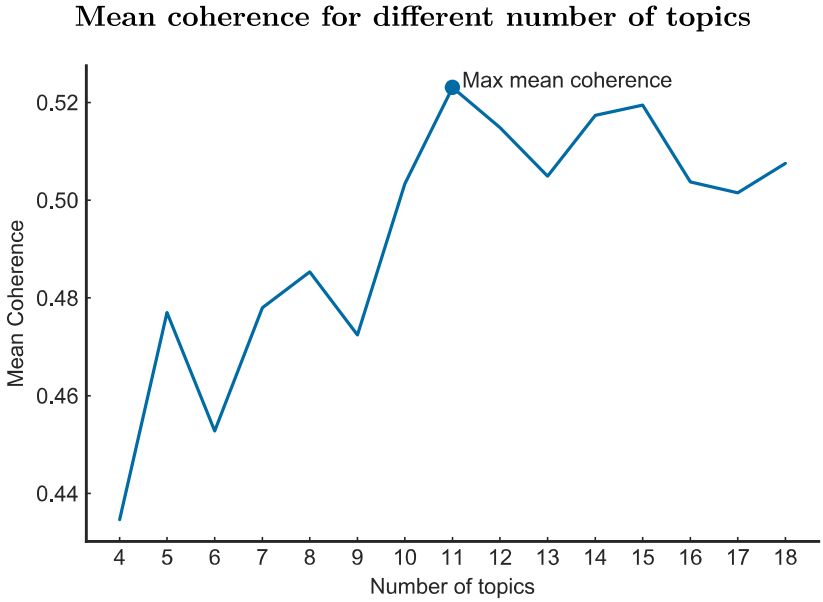
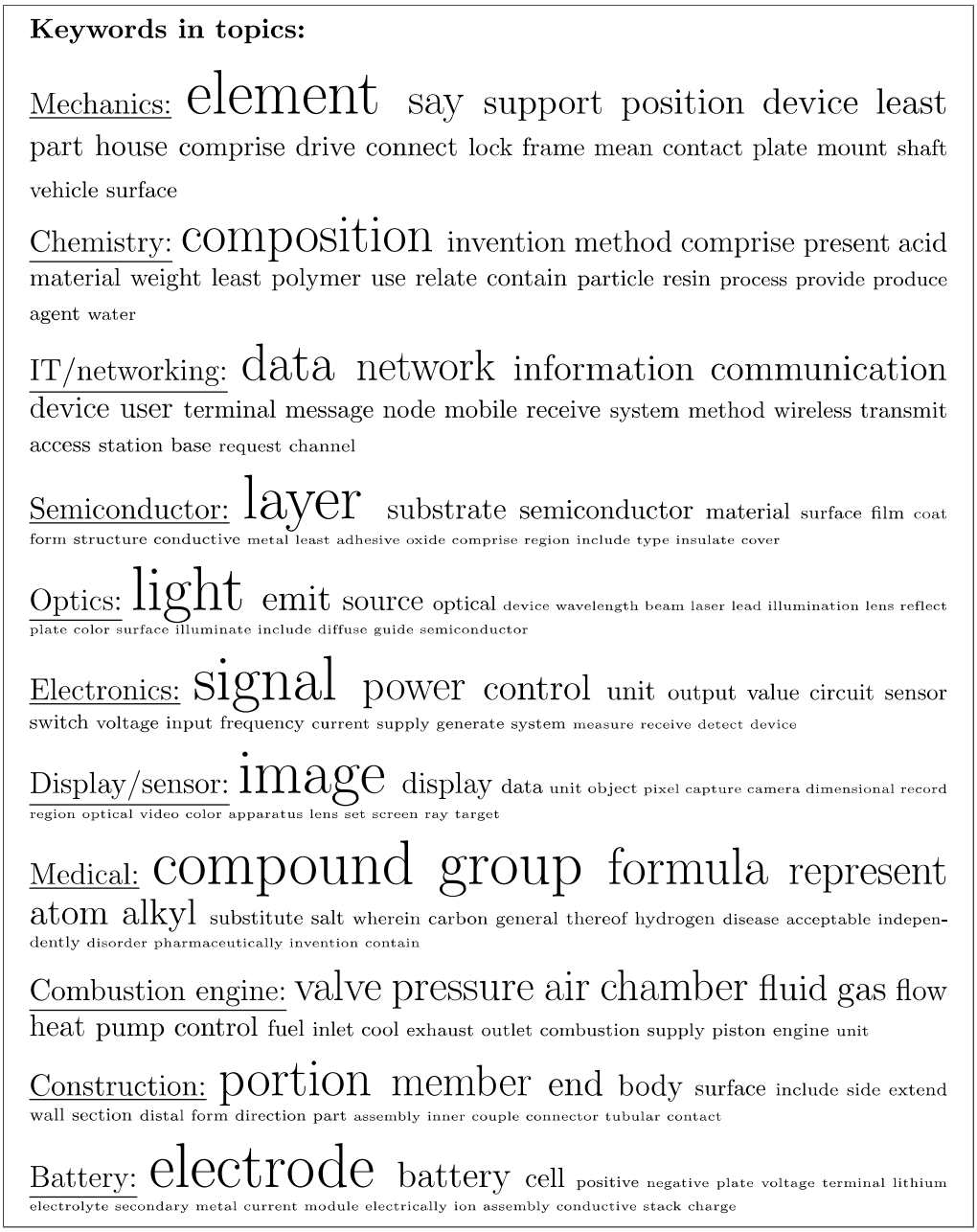



Figure 5: Mean coherence for different values of $k$

The resulting 11 topics are presented in the box below. Each topic shows the 20 highest scoring terms (keywords). The font size indicates the relative score within and between topics. We interpreted the keywords to give each topic a relevant name. Each abstract may belong to more than one topic. We assigned the abstract to the topic with the highest weight. The resulting

topics highlight the usefulness of topic modeling compared to the preassigned categories. For instance battery technology is a research area with many granted patents, which is not obvious from the IPC classifications (see Figure 12 in the Results section).




# 5 Results

## 5.1 Female inventors across countries

Figure 6 shows the development in the share of female inventors obtaining patents for a number of countries. Across all countries there has been an increase in the share of female inventors

from 9.6 pct. in 2008 to 13.1 pct. in 2018, an increase of 3.5 pct.-points. However small this may seem, it actually implies an increase of 36.2 pct.

For both Denmark and Sweden, we see a big spike in the share of female inventors in 2011. This volatility, especially in the case of Denmark, is likely due to the low number of patents granted each year, which makes the data more vulnerable to outliers.

It is however interesting that the share of female inventors in Sweden and Denmark is lower than the combined share for most years of our sample, considering that Sweden and Denmark are both on the forefront of gender equality. This may indicate that women select fields of study based on other parameters than wage, since there is a considerable wage premium in the fields of innovation.

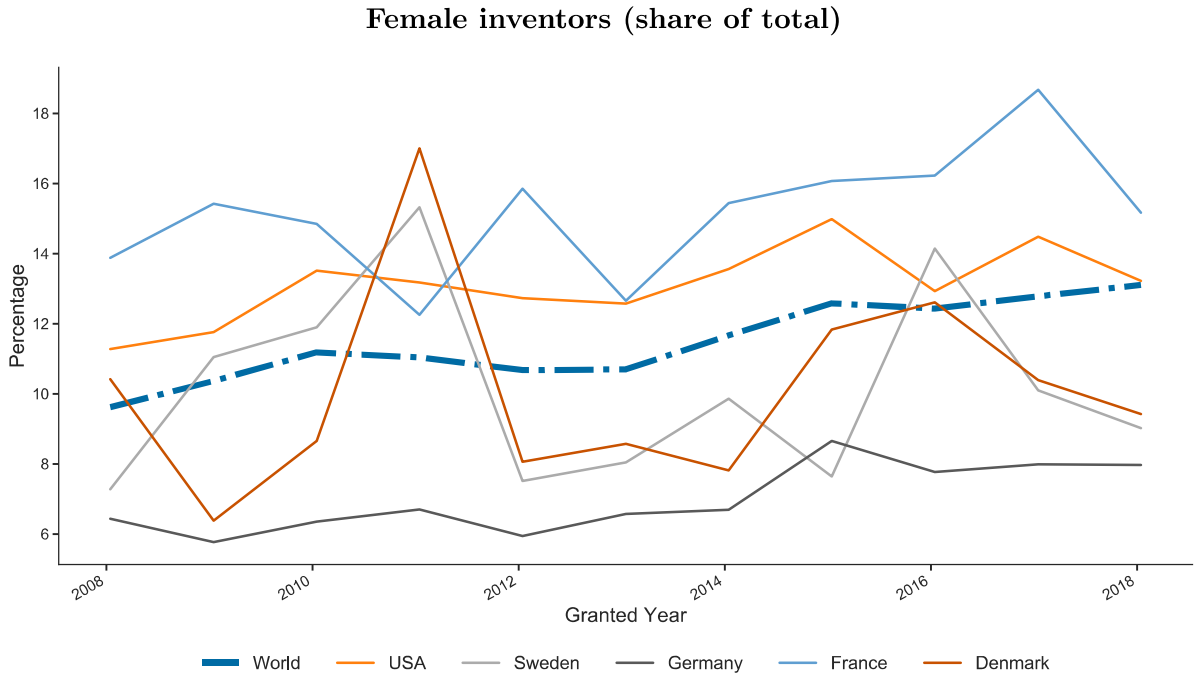



Figure 6: Share of female inventors of total inventors

Figure 7 shows the share of female inventors obtaining patents within countries in Europe and the total number of inventors in each country[14]. Countries with less than 50 inventors are excluded. The data covers inventors from patents from January 2008-2018. We see a relatively large difference in the share of female inventors throughout Europe ranging from 6 pct. to 27 pct.

Interestingly, when comparing the countries with the highest share of female inventors with

[14]The map was made using Tableau Software

the top 5 scoring countries on the Gender Equality Index of 2017[15], we do not find the same countries[16]. Except for France, all high-ranking countries are below the overall average share of 12 pct. in 2017. On the other end of the spectrum we have a country like Poland, which has a relatively large share of women in science, while being ranked 20th on the European Gender Equality-index. Beede et al. [2011] finds the same inverse relationship between countries scoring high on the Gender Equality Index and the share of women obtaining patents within these countries.

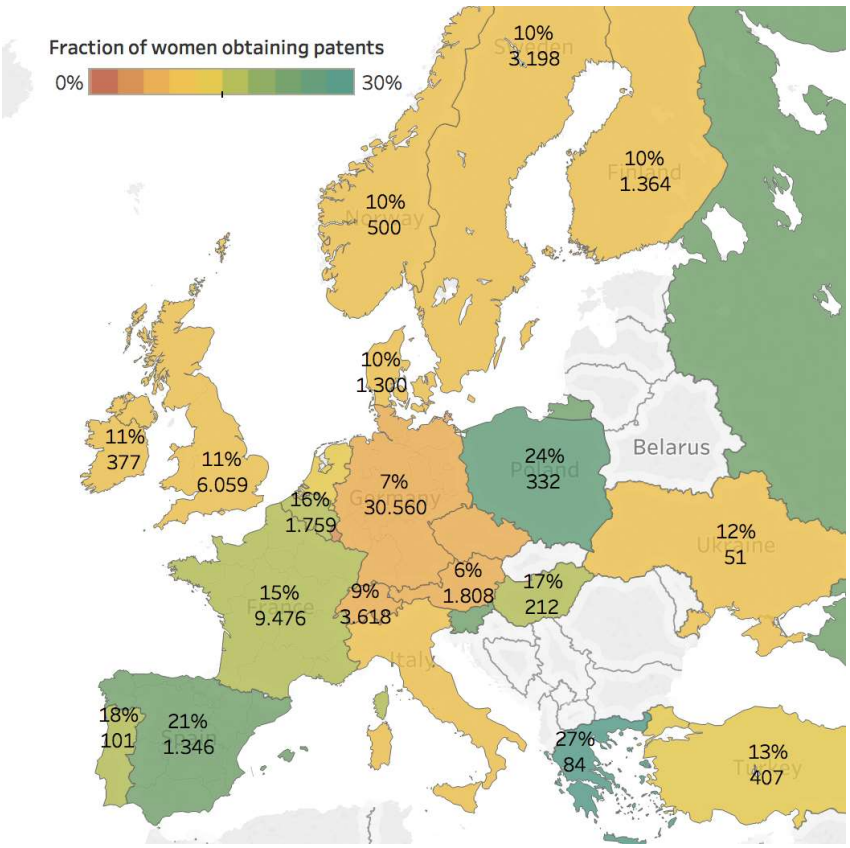



Figure 7: Map of the share of female inventors obtaining patents and the total inventors by country, January 2008-2018. Countries with less than 50 inventors are excluded.

## 5.2 Female inventors across patents

To dig a bit deeper into the inventor gap, we investigated the share of only male inventors, only female inventors and both genders on a single patent. The gender diversity within the group of inventors is not necessarily the same as gender diversity for a single patent.

[15]From the European Institute of Gender Equality, http://eige.europa.eu/rdc/eige-publications/gender-equality-index-2017-measuring-gender-equality-european-union-2005-2015-report

[16]The top scoring countries in descending order: Sweden, Denmark, Finland, The Netherlands and France

Figure 8 shows the share of patents with female inventors by patents with 1, 2 and 3+ inventors. As the separation by number of inventors reduces the sample size, the line plots become somewhat volatile. The increase in female inventors is predominantly on patents with 3 or more inventors.

**Share of granted patents with female inventors by number of inventors**

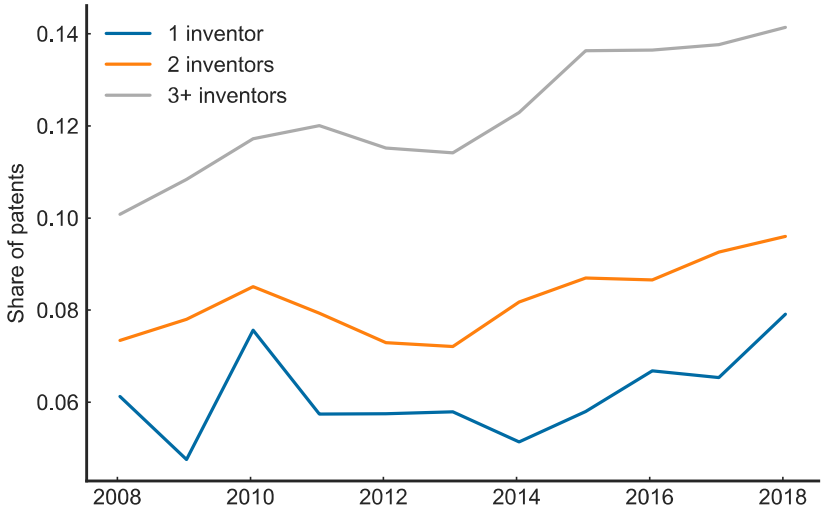



Figure 8: Share of granted patents with female inventors by number of inventors

As Figure 9 shows this is also the type of patent that exhibits the largest growth. This may indicate a shift in research environments or an increase in patent applications from specific research areas. Nonetheless, the growth of female inventors benefit from an increase in female inventors in collaborative patents.

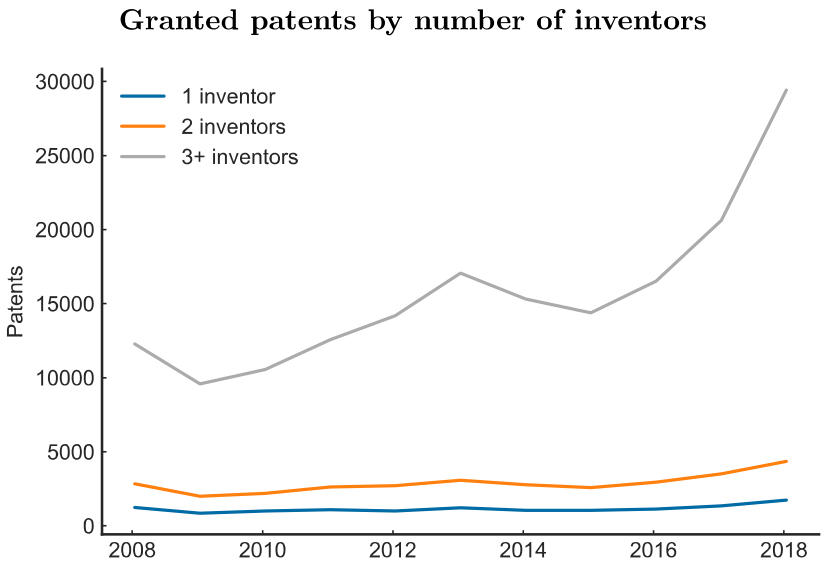



Figure 9: Granted patents by number of inventors

This effect is mirrored in Figure 10, which shows the share of patents with only male inventors, only female inventors and inventors of both genders. The share of patents with only female inventors (which includes solo inventors) is only 2 pct. and stable across the period. The share of patents with both genders is increasing in the period, while the share of male only patents is decreasing.

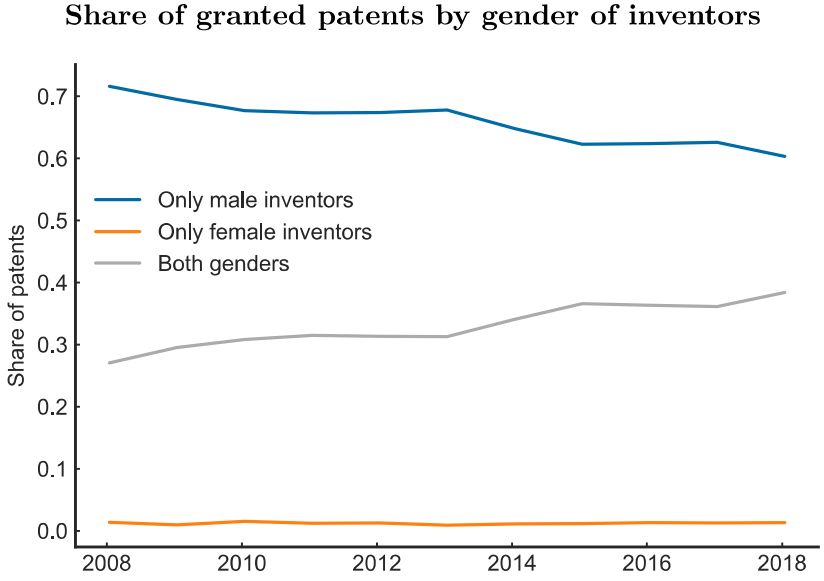



Figure 10: Share of granted patents by gender of inventors

### 5.3 Female inventors across sectors

Figure 11 shows the number of patents in a given *main IPC group*. These are sectors into which all patents are grouped, and each patent thus contain one or more IPC group classifications. The insight provided by Figure 11 is that these sectors are closely related to the fields of *Science, Technology, Engineering and Mathematics* (STEM) in which around 75 pct. of workers are men [Beede et al., 2011]. Based on this gender gap in fields of innovation, we would expect to find a gender gap in our data.

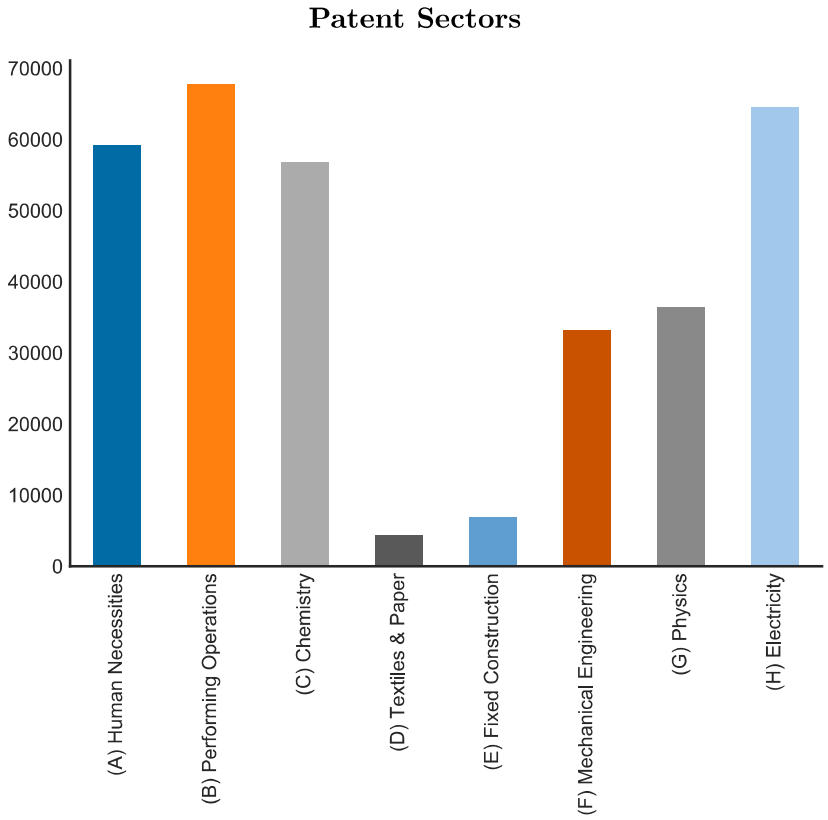



Figure 11: Patents granted by main IPC sectors, 2008-2018 (*Data for the month of January*)

This reflects on the discussion on *gender wage-gap*. There is no doubt that when it comes to wages, men earn more than women. This discrepancy has fallen throughout the past decades, yet the gap persists to this date. The most recent research within the field of gender inequality even shows that the gender inequality of wages in Denmark is mainly present due to childbearing rather than gender discrimination [Kleven et al., 2018]. Yet, simply looking at life-time income, men still earn substantially more than women. Even when correcting for childbearing. The main driver behind these wage differences is self-selection of women into fields that generally pay less than male-dominated fields.

Figure 12 shows how inventors are distributed across the topics inferred from the abstracts using topic modeling. The topics with the highest gender diversity are *chemistry* and *medical*. The topics with the lowest gender diversity are *mechanics* and *combustion engines*. There does not appear to be a relationship between gender diversity and number of inventors within each topic.

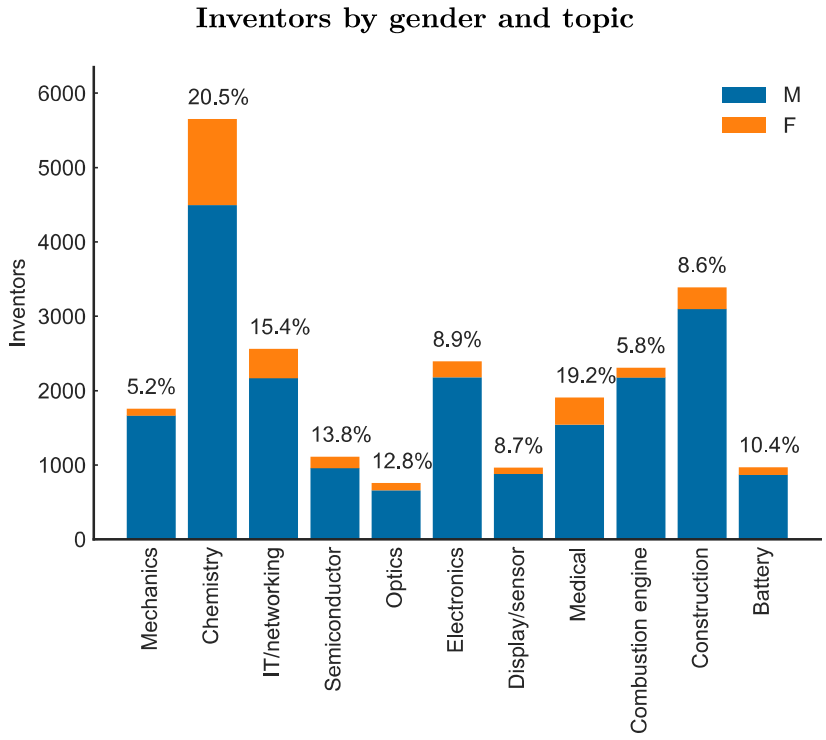



Figure 12: Inventors by gender and topic, granted January 2017. Percentages indicate the share of female inventors in the topic.

# 6 Discussion

Due to the time scope of this assignment and limitations on EPO's API, we were constrained in regard to how much data we could process. We chose to process patents granted in January for the past 10 years. In doing so we implicitly assume that January is representative for an entire year. This seems like a reasonable assumption given that there is no bias towards processing specific applications in each month. Further research could be based on the entire population of data within the period, as well as include data from other patent offices around the world. Additionally, we could include patent applications which has been denied or not been granted yet, and analyze if there is valuable information to gather, e.g. gender bias in denied patents. Furthermore, including the firms applying for each application could be useful, e.g. to gain

information on gender in relation to firm size.

In Figure 13 in the appendix we have included a world map of the share of female inventors. We see that on a global scale China tops the list with a share of 32 pct. female inventors followed by Malaysia (27 pct.), and Greece (27 pct.), Poland (24 pct.) and Russia (23 pct.). Later research could explore these differences further. The differences could however be due to a bias in our Random Forest model. The model trains by using male and female name characteristics. These characteristics might not be equivalent between continents or countries, e.g. what defines a female name in Asia is not the same characteristics that define a female name in Western countries. If this is the case, a way of solving it would be to train a model for each individual country or region.

We have relied on standard machine learning models and libraries, and due to time constraints and our relative inexperience, have not been able to do thorough robustness checks of the models. For instance, when classifying genders based on names there are many patterns of characters that might contain gender cues - especially when considering the differences between names across countries and regions. In this assignment we used predetermined features e.g. first letter in a name. This is one way to identify these patterns, but one might be able to explore these gender-revealing sequences without having to explicitly specify them by using a character-level Recurrent Neural Network[17].

# 7 Conclusion

This paper aimed at inferring some of the characteristics of the gender gap in invention. Unsurprisingly, the field is highly male-dominated, but women have been catching up in the last 10 years with an increase in the share of women obtaining patents of 36 pct. However, coming from a low level this only corresponds to an increase of 3 pct. points. The increase in the share is mainly driven by women working in gender-mixed teams, while the share of all-female teams remains constant during the period. Countries typically associated with having a high degree of gender-equality did not have have a higher share of female inventors. Within sectors we saw women having the highest representation in the fields of *chemistry* and *medical*. *Mechanics* is the most male-dominated field, based on the clustering of abstracts on the month of January 2017.

[17]link: https://towardsdatascience.com/name2gender-introduction-626d89378fb0

# 8 References

David N Beede, Tiffany A Julian, David Langdon, George McKittrick, Beethika Khan, and Mark E Doms. Women in stem: A gender gap to innovation. 2011.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009. URL http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

Henrik Kleven, Camille Landais, and Jakob Egholt Søgaard. Children and gender inequality: Evidence from denmark. Technical report, National Bureau of Economic Research, 2018.

Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645 – 5657, 2015. URL http://www.sciencedirect.com/science/article/pii/S0957417415001633.

Julio Raffo. Worldwide gender-name dictionary. 2016. URL https://EconPapers.repec.org/RePEc:wip:eccode:10.

Sue V. Rosser. The gender gap in patenting: Is technology transfer a feminist issue? *NWSA Journal*, Vol. 21, No. 2 (Summer, 2009), pp. 65-84, 2009.

# 9 Appendix

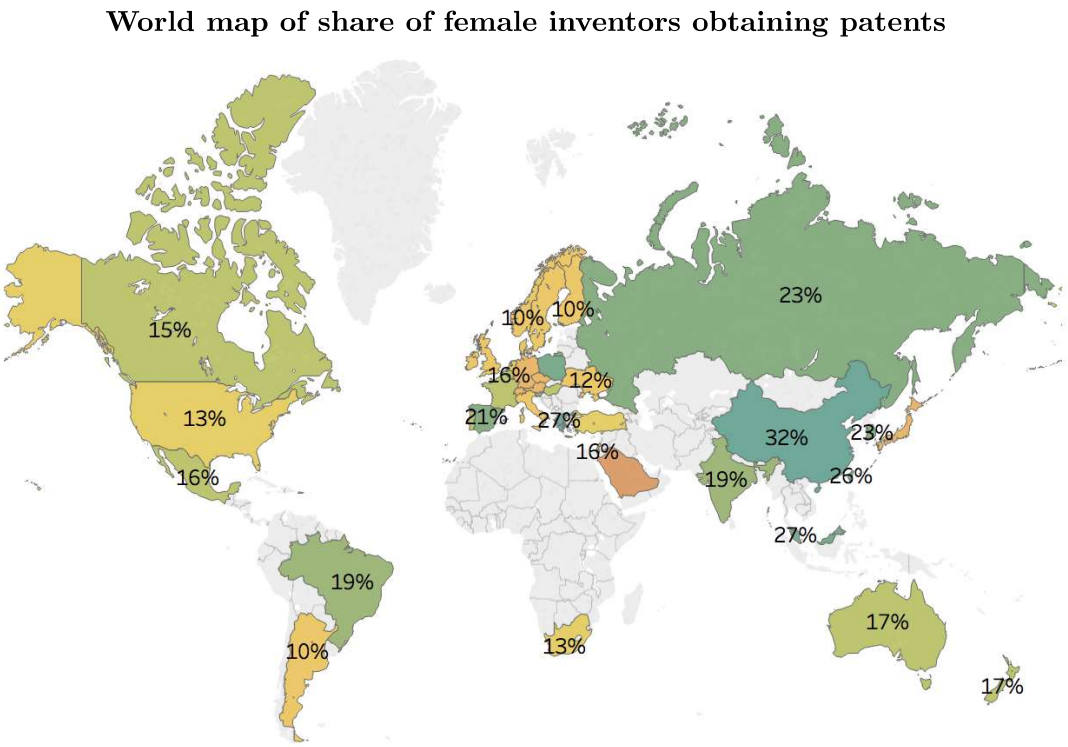



Figure 13: Map of the share of female inventors obtaining patents by country, January 2008-2018. Countries with less than 50 inventors are excluded.