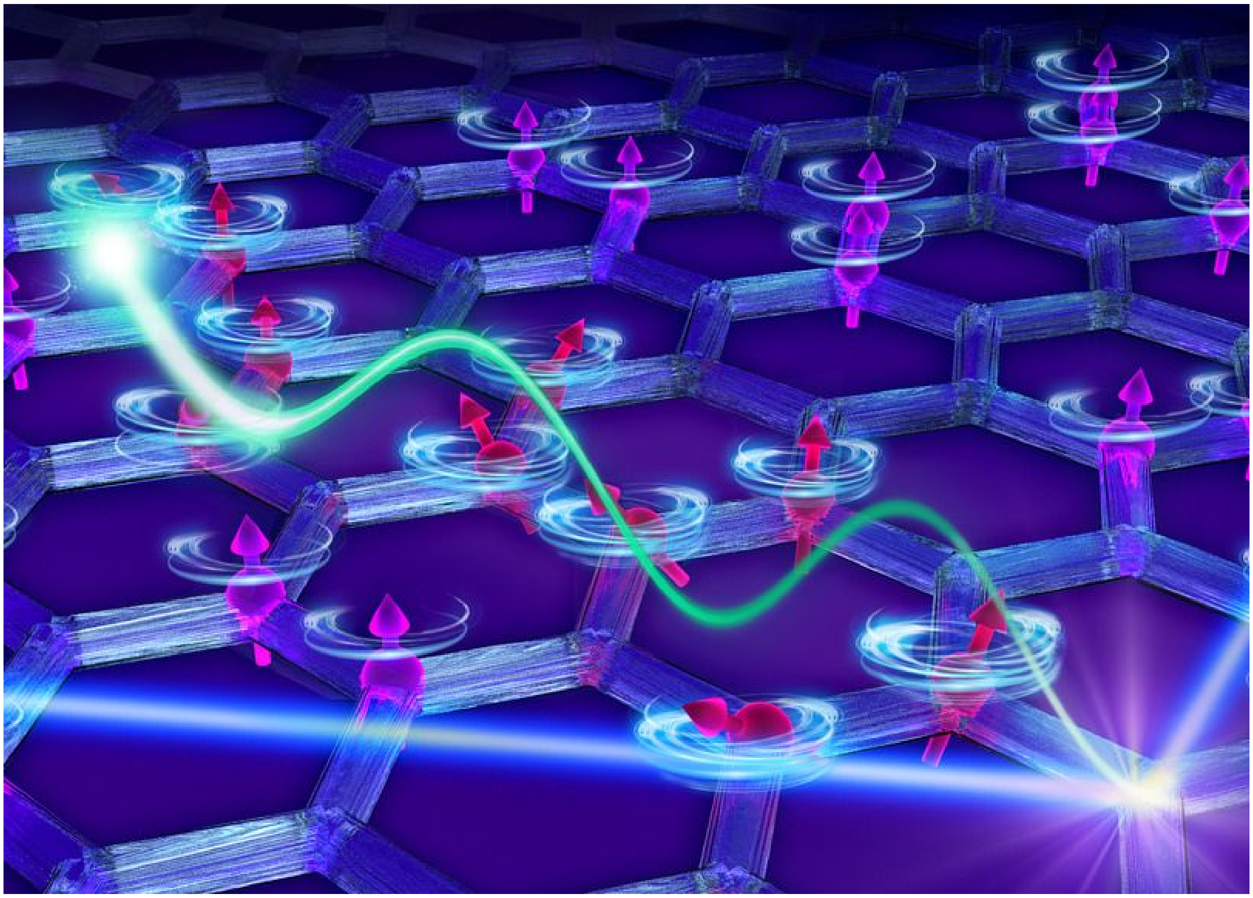
Python W200 - Fall 2018

What Scientists Have Wikipedia Pages:

A study of the missing scientists



Final Project

Prepared by: Lorrel Plimier, Sheila McGovern, and Daniel Lee

Introduction

Wikipedia has grown to represent a vast store of information on the world around us[1], but it's not comprehensive. Recognizing this, Primer.ai, a machine intelligence company, developed an AI tool to augment human-generated knowledge bases with machine-generated ones. They've released a public dataset of over 36,000 quantifiably notable[2] computer scientists discovered by analyzing 500 million news articles, 39 million scientific papers, and all of Wikipedia.

Incredibly, only 15% of these accomplished computer scientists had a Wikipedia article at the time of release, with others missing significant events, happenings, or context in their Wikipedia entries. Primer.ai refers to this phenomenon as Wikipedia's "recall problem...the articles that should be there but are entirely missing."[3] Other critics have acknowledged this problem as well across different dimensions, pointing to the inherent racial and gender biases present in Wikipedia's article corpus that mirror the mainstream underrepresentation of minority or disadvantaged groups across various topic areas (for instance, the fact that only 18% of all Wikipedia biographies are of women).[4]

Through their pioneering research, Primer.ai illustrates practical uses of machine learning as a means to fight missing or biased information, and their platform, Quicksilver, has shown successes in leveraging machines to augment human inputs. However, while it's clear Quicksilver can directly addresses the challenge of building a more complete article corpus, we believe it's equally important to understand whether these AI tools contribute to, or resolve some of the other biases.

This paper first seeks to understand the factors that may facilitate Wikipedia notoriety by comparing the characteristics of computer scientists that have and do not have Wikipedia pages. In doing so, we will also consider the effectiveness of Quicksilver in addressing these biases by determining the extent to which the AI corpus fits or disrupts the status quo.

[1] The english language edition of Wikipedia contains almost 6 million articles and adds 20,000 new articles per month. https://en.wikipedia.org/wiki/Wikipedia:Article_size
[2] Quantifiably notable refers to the requirements for scientists to be included, including at least 10 published papers with over 100 citations; https://github.com/PrimerAI/primer_quicksilver
[3] https://primer.ai/blog/quicksilver
[4] https://www.wired.com/story/using-artificial-intelligence-to-fix-wikipedias-gender-problem/

The Data

Primer.ai's public dataset is a .jsonl (JSON Lines) file with a single row for each of the 36,000 notable computer scientists identified by Quicksilver. This data was curated by linking and disambiguating across various large input sources, to include citations to news articles, professional affiliations, mappings to published works by that individual, and mappings to Wikipedia and Wikidata entries (if they exist). A full specification of variables in the raw Primer.ai data is seen in Figure 1 below.

Variable	Type	Description
name	string	name of the scientist
s2_affiliations	list	scientist affiliations from semantic scholar dataset
s2_paper_ids	list	semantic scholar paper ids attributed to the scientist
s2_id	string	id of the scientist in semantic scholar dataset
en_wikiUrn	list	wikipedia urn for the scientist
P21	list	gender fetched from wikidata if present or calculated using census data and news mentions
primer_id	string	primer generated unique id_
wikidata_id	string	id_ of the person in Wikidata
news_docs	list	disambiguated news docs mentioning the scientist (includes URL, publication date, article title, sentences where subject was quoted)

Figure 1: Specifications for Primer.ai's Computer Scientist dataset

Data Cleansing Approach

After raw data was loaded into a Pandas dataframe, additional preprocessing was conducted to:

1. Fill in missing values
2. Create categorical variables from strings
3. Process "lists of lists" and "lists of dictionaries" to support creation of new variables
4. Merge variables to create a final analysis dataset

The new variables created through Step 2 are summarized in the table below. A more comprehensive discussion on data cleansing efforts can be found in *Appendix A: Data Cleansing Overview.*

New Variable	Type	Source	Description
country	string	s2_affiliations	Probable country of scientist derived from text matching
continent	string	country	Probable continent of scientist derived from grouping countries
specialization	string	s2_affiliations	Probable secondary specialization of scientist derived from text matching
institution	string	s2_affiliations	Probable industry of scientist derived from text matching
num_pub	int	s2_paper_ids	Count of published papers
gender	string	P21	Enhanced gender column using python gender guesser package
num_news	int	news_docs	Count of news mentions
buzz_keyword	int	news_docs_title	Count of the number of times certain “buzz” words were found in the news article title. The buzz words were defined as AI, machine learning, deep learning, humanoid, robot, data science.
prize_keyword	int	news_docs_title	Count of the number of times words were found within the news article name that indicated an award of some kind. The prize words were defined as award, prize, medal
quote	int	news_doc_mentions_is_quoted	The number of times a scientist was quoted in news articles
news_before_2015	int	news_docs_publication_date	The number of articles where the scientist is mentioned published before 2015
news_2015	int	news_docs_publication_date	The number of articles where the scientist is mentioned published in 2015
news_2016	int	news_docs_publication_date	The number of articles where the scientist is mentioned published in 2016
news_2017_or_later	int	news_docs_publication_date	The number of articles where the scientist is mentioned published in 2017 or 2018

Figure 2: New variables created from Primer.ai’s raw data

Exploring Our Dataset

We started by analyzing the composition of the 36,000 computer scientists as a full population. We knew we ultimately wanted to look at how our variables were tied to whether or not the scientists had a Wikipedia page, but first we just wanted to get to know the data. We analyzed subsets of variables in order to discover what qualities we could attribute to our large group of scientists.

Geography

We analyzed the distribution of our scientists by country to understand where these notable individuals reside. The bulk of the scientists in our data (30.7%) were from the United States, but there was high representation from major players in Asia and Europe including China, Germany, the UK, Canada, Italy, France and Japan. This data is illustrated here in a choropleth map generated using the GeoPandas package.

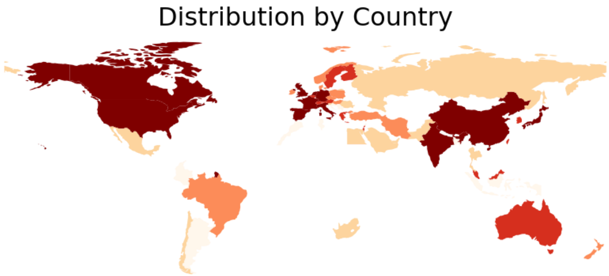


Figure 3

Gender

We also analyzed the distribution of our scientists by gender to understand how closely the AI-generated data aligned with the status quo.

Across the entire population, we see 66% males, 10% female, and 24% unknown. When we exclude the unknowns, we see the ratios of 87% male to 13% female. These ratios suggest that the AI-generated dataset still identifies females at a much lower rate than males. Current estimates by the National Science Foundation reveal 24% of the computer and information

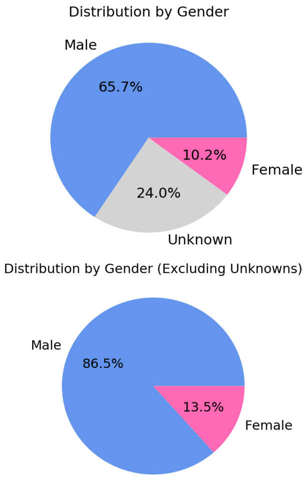


Figure 4

sciences workforce is comprised of women.[5] We discovered similar trends when analyzing the distribution of genders across geographic regions.

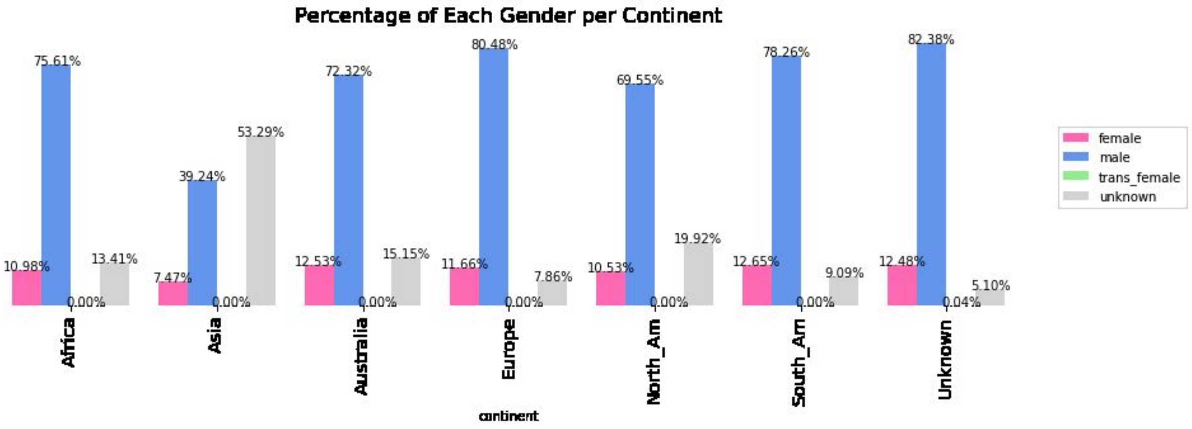


Figure 5

One interesting fact about our dataset was the large number of scientists with unknown gender (53%) in Asia. Some of our gender classification was done with first name recognition and we suspect that there were most likely some translation difficulties for Asian languages within the AI that generated the original database. Both of these factors could lead to an inability to classify the gender of a scientist.

We did some simple analyses of our count variables at this point, too. We have counts for the number of: published papers, affiliations, news articles, times quoted in the news, industry keywords included in news articles about the scientist, and award/prize specific keywords about the scientist.

Secondary Specialty

We used the affiliations variable from our raw data to determine what secondary specialties to assign to our computer scientists. Again, there were a fair number of unknowns in our results, but we were still able to gain insights into the composition of our group of scientists.

[5] https://www.nsf.gov/statistics/2018/nsb20181/report/sections/science-and-engineering-labor-force/women-and-minorities-in-the-s-e-workforce

To illustrate this, we can look at the number of published papers for each scientist. Because of the complexity of this plot type and the size of our database, we were forced to take a random sample of our data but the illustration still accurate depicts the anatomy of our scientists. In Figure 6, there is one dot per scientist, swarming around the relevant specialty and moving out to the right like lone bees as the number of published papers for an individual scientist reaches the outer limits of our dataset. From this illustration, not only can you get an insight into the gender distribution of our scientists across specialities, but you can also glean which specialties have scientists that are more prolific in publishing.

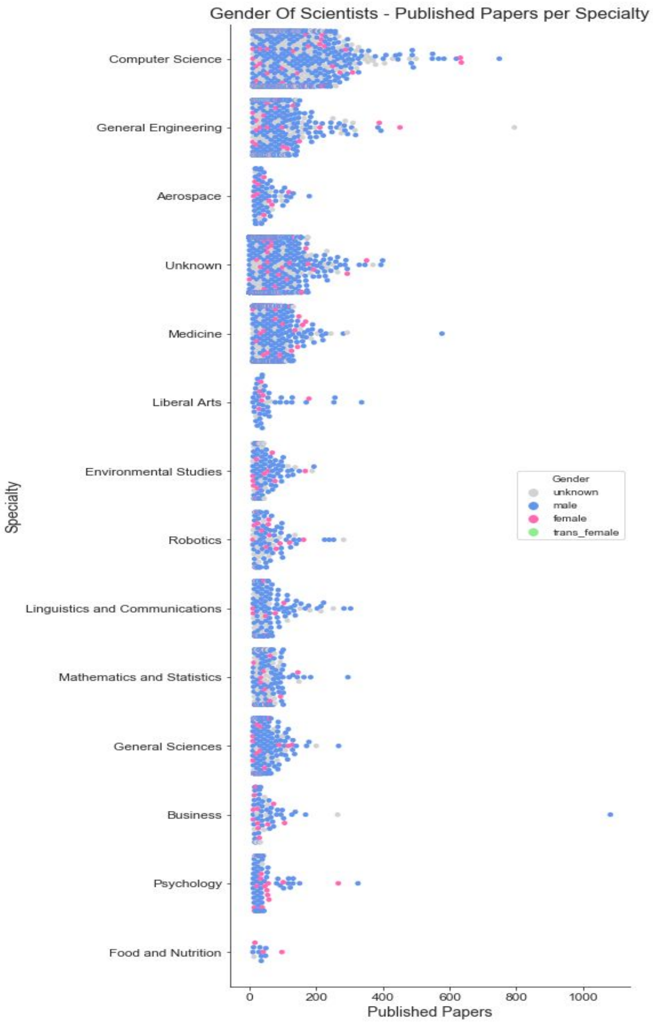


Figure 6

With our group of scientists, there appears to be a greater percentage of women in the fields of business, psychology, and food & nutrition. There is also an evident domination of computer scientists among our group. It is also clear that most of our scientists have published fewer than 200 papers and very few indeed publish greater that 400 papers (and who can blame them!)

Wikipedia - Are We There Yet?

Once we had a preliminary understanding of who our scientists were, we turned to discovering trends in the scientists who had Wikipedia pages. We had some hypotheses about how this might play out and we started by exploring these. We surmised that scientist with Wikipedia pages would be concentrated in certain regions, predominantly male, popular in the general public, and prolific publishers of scientific papers. We were able to confirm most of these theories, but also found some surprising results.

“Pop Factor”

We created a “Pop Factor” variable by combining the counts from news articles of the number of industry specific buzz words (e.g. AI, robot, humanoid, deep learning, machine learning, data science) and the counts of the number of keywords related to awards and recognition (e.g. prize, award, medal). Based on the Pop Factor, scientist who are more popular are much more likely to have a Wikipedia page. The interesting thing to note is that our database was unable to identify the geographic region for most of these popular scientists.

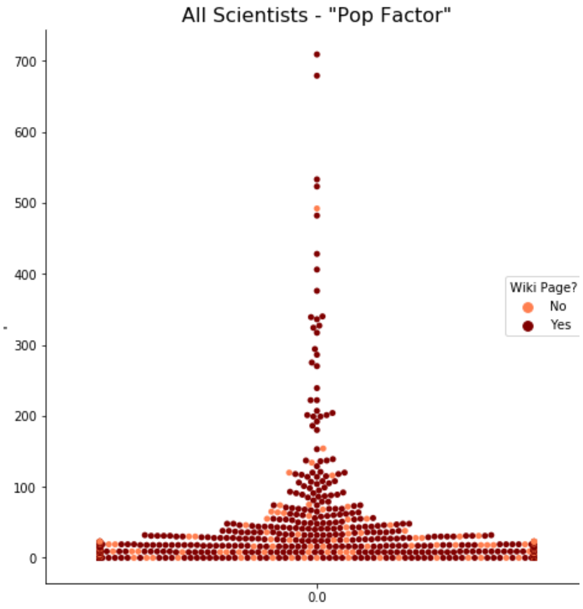


Figure 7

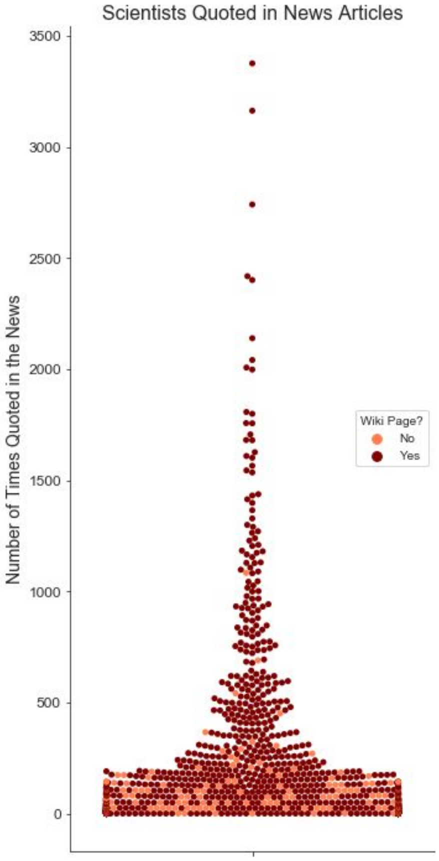


Figure 8

Being quoted

Looking at the number of times a scientist is quoted in news articles also seems to be a good indication of whether the scientist is likely to have a Wikipedia page. Again, this is related to the Pop Factor in that having lots of journalist quoting you tends be a good indicator of your popularity, or at least your infamy.

Gender and number of published papers per continent

Next we focused on women only as a subset of our data and confirmed for this subset that scientists with larger libraries of published papers are more likely to have a Wikipedia page. Of note is the line of Wikipedia pages with zero publications and unknown regional affiliation. This is undoubtedly a result of how the dataset was collected. The parameters for inclusion in the dataset included having a certain number and type of published papers or having a Wikipedia page. Clearly, the scientists with no published papers must have fulfilled the latter requirement.

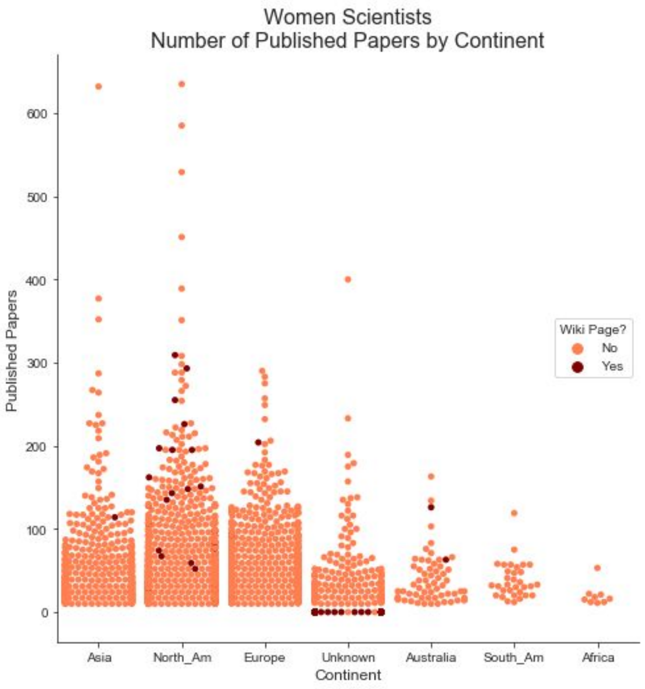


Figure 9

Gender Comparison of Scientists without a Wiki Page to All Scientists

Another interesting discovery with our data was that when looking at the entire population of scientists in our data, the percentage of women with a Wikipedia page is essentially the same as the percentage of men with a Wikipedia page. We had expected that more men than women would have Wikipedia pages and while this is true, it is not disproportionate to the gender breakout among the entire population.

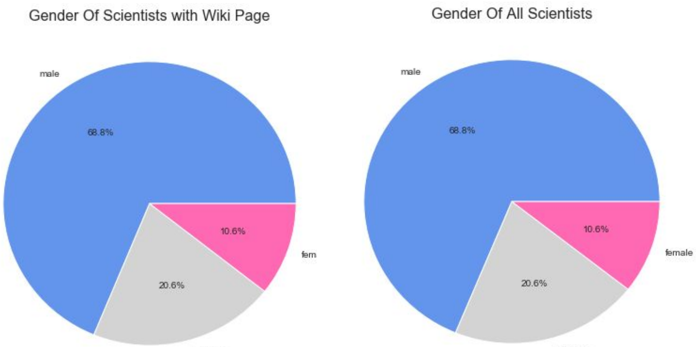


Figure 10

Further Analysis

We were able to draw preliminary conclusions on the data we had available to shed light on questions we had regarding which scientists have wikipedia pages versus those that don't. Our preliminary conclusions are as follows::

1) There doesn't appear to be gender bias in determining whether or not a scientist has a wikipedia page
2) There appears to be a correlation between the number of papers published and the existence of a page
3) There appears to be a correlation between the number of times a scientist is quoted and the existence of a wikipedia page
4) There appears to be a correlation in the types of mentions in news documents - the scientists who were mentioned in a news article pertaining to an award or a trending keyword such as AI or robot, were more likely to have a wikipedia page

There is additional analysis that could be performed on this data. This includes:

1) Analysis of the actual paper the scientist had published. We have the id number in the dataset, but would need to query an additional(very large) dataset (Semantic Scholar Open Research Corpus) to get the actual title. If we had that actual title, we could analyze for categories (such as buzz words, specialties, what publication it was for, etc.
2) We could analyze information on the actual wikipedia page to determine the relationship between the scientist and the person who contributed the entry. An example can be found in the edit history. Edit history for Paritosh Pandya
3) Analyze creators or editors of live Wikipedia pages of scientists to determine if there are additional conclusions to be drawn (i.e. is there a relationship between the page creator and the scientist?)

Appendix A: Data Cleansing Overview

Creating Categorical Variables from Strings

The raw data contained an alphanumeric codes for genders. We replaced these codes with gender names and assigned an "unknown" gender to missing values. We then used the **gender_guesser** Python package that uses first names and countries to "guess" the gender of those scientists with unknown gender. If the guess was "male" or "female," then we used this instead of the unknown classification. The guess could also be "androgynous," "mostly male," or "mostly female" in which case we left the classification as "unknown." The reliability of the gender_guesser package is not fully known, but it served as a mechanism to augment the existing data on genders.

Addressing Lists of Lists or Lists of Dictionaries

The s2_affiliations column held data in lists of lists, with one individual having the potential for multiple associations. A new dataframe was created for this variable before being flattened such that each record was expanded to n rows, with each row containing one string for each object in the original list of n objects.

At this point, each affiliation was generally represented as a string containing [Affiliation Department], [Affiliation Entity], and [Affiliation Location], but there was no standardized format or language for this information. To create the "Country", "Specialty", and "Industry" columns, the following approach:

1. Manually analyze observations to define a list of keywords that would drive the categorization of raw strings[6]
2. Translate observations into derived columns through simple text categorization

In some cases we elected to create higher-level groupings for these created categories as well. For example, because the data for the countries was so dispersed, we created a new

[6] For example, if the affiliations string contained strings like "New York", "NYC", or "Boston," we could categorize these into the country "USA". Similarly, if the affiliation string contained strings like "radiology", "neur", "cardio", "ortho", "anae", "gyn", or "neur" these were categorized into specialization "Medicine."

“continent” variable that grouped all of our identified countries into their respective continents.

Overall, there were a number of instances where we were not able to ascertain the country of the scientist either because of language barriers or because of missing information in the raw data. This is why the “unknown country” ends up being the second largest subset for the derived country variable.

Similarly, data in the news_docs column was contained in lists of nested dictionaries. The first level of the dictionary was pulled out into columns with the repeat and join commands on the primer_id(unique identifier for the scientist). This caused our approximately 39,000 record dataset to reach almost 400,000 records. There was a still a field - that was a nested dictionary in this resulting set. It contained the number of times the person’s name was referenced in a particular news article and whether they were quoted. To summarize, a particular scientist’s name could be mentioned 1 to many times in a news article. A scientist may have 1 or more articles where they are mentioned. When we tried to pull the data around number of mentions and whether or not it was a quote from the scientist out in a similar way as before, the program timed out. We pulled this information out when we analyze the individual rows in the dataframe. It was during this analysis, that we made the determination around how many buzz words, prize words were present in the mention, and what year the article was published.

Creating Summary Count Variables

We pulled out the useful content from the “affiliations,” “papers,” and “news” variables in the raw data, but knew that the number of entries in those variables might also be useful. If a scientist had numerous affiliations, dozens of published papers and was quoted in hundreds of news articles, was this scientist more likely to have a wiki page than a scientist who had only one of each?