Exploration of a Scientific Events Dataset

Charles Murray's book Human Accomplishment surveys "outstanding contributions to the arts and sciences from ancient times to the mid-twentieth century" (quote from Wikipedia). The data gathering for this work took years, and the raw dataset was made available as an Excel file. Each of the 7131 records a major event in human history such as an invention or scientific accomplishment.

Dataset Basics

The dataset was read into a Pandas DataFrame for initial inspection.

The raw field names and data types are the following:

```
id_event              int64
year                  int64
where                object
category             object
subcategory          object
description          object
source_pct            int64
significantevent      int64
centralevent        float64
id_person           float64
person               object
dtype: object
```

Several of these fields were then renamed according to personal preference (e.g. **where** changed to **location**).

The following data dictionary shows relevant information for the fields that were deemed useful for analysis. (The renamed column designators are used here instead of the original ones.)

Field	Type	Description	Values
id	int	Unique identifier of the event	unique
year	int	Year the event occurred	-10000 to 1950
location	string	Location of event (typically a country or region)	137
category	string	Major category (e.g. Technology	9
subcategory	string	Subcategory with main category	68
description	string	Text description of the event	unique
significant	boolean	True if event is considered significant	true/false
person	string	Person to whom the event is attributed	3648

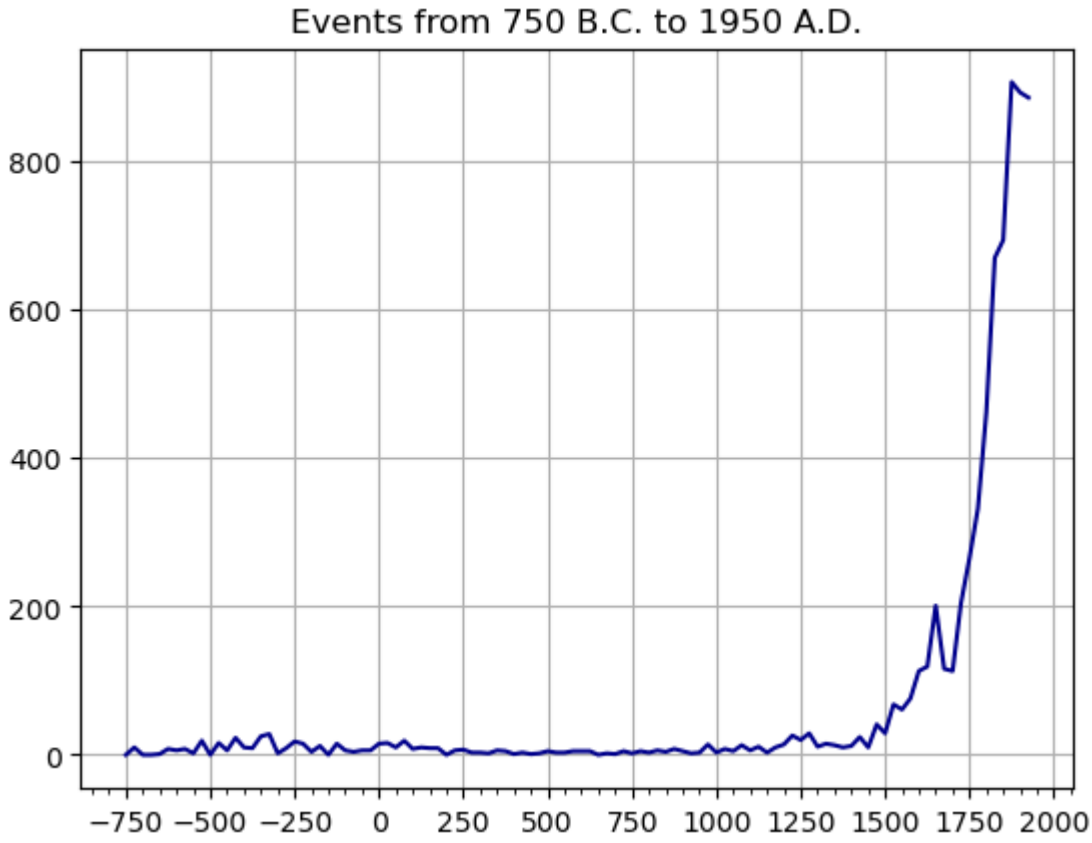
The columns are mostly text strings with year encoded as an int and significant as boolean (technically 0 or 1 in the original dataset). The above **Values** field describes if the values of the variable are unique, what the range is for year, and the number of unique values for **location**, **category**, **subcategory** and **person**.

Time Span and Counts

The dataset covers a large span of time from the dawn of human history into the modern era. The range of years was the following, generated by calculating minimum and maximum values of the **year** field:

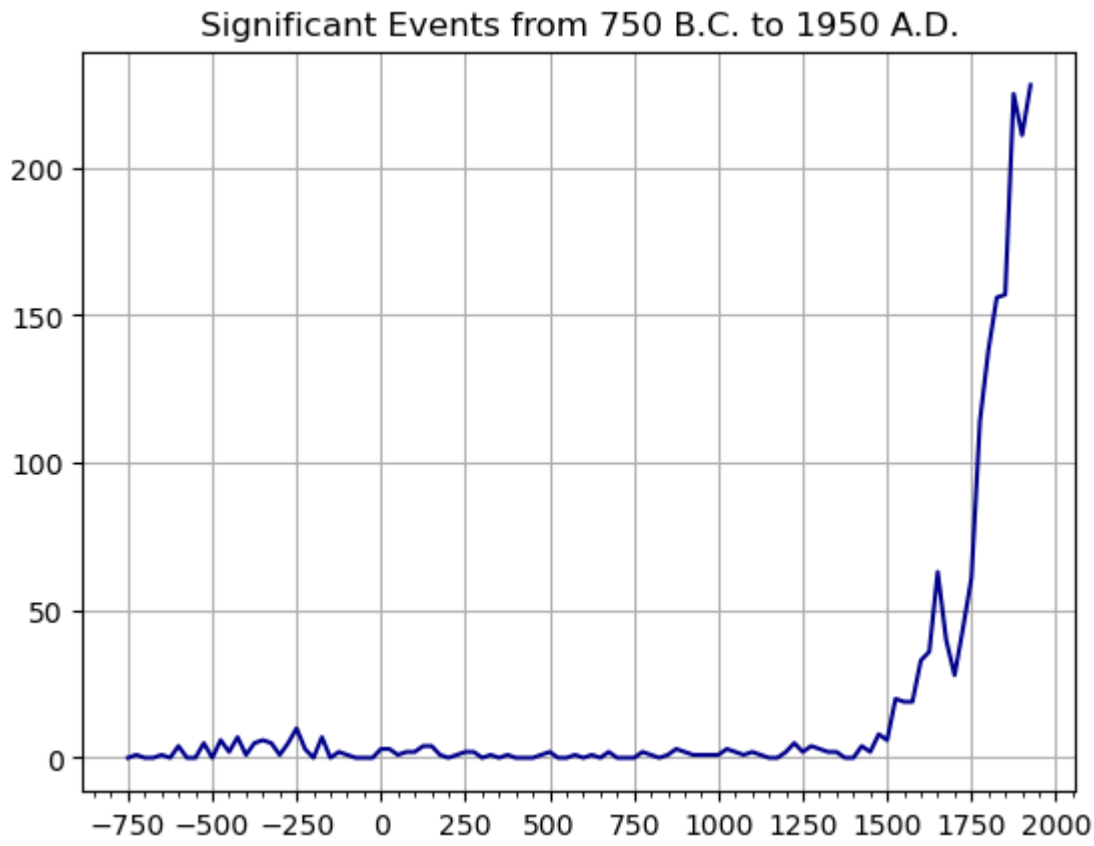
```
-10000 BC to 1950 AD
```

There are only a sparse number of events in the ancient era until a certain year, so 750 B.C. (-750 in the data) was picked as a cutoff for graphing the data.



The graph shows events counts by year binned into 25 year periods. As might be expected, exponential growth occurs somewhere around 1500 (perhaps corresponding roughly to widespread adoption of the printing press and dissemination of knowledge). This would also align with the start of the Renaissance when many notable achievements occured. The curve becomes particularly steep in the latter half of the 18th century until the last date of 1950, which is also unsurprising, given the number of advancements during this time period.

The data also includes an indicator as to whether the researcher felt an event was very significant. These significant events can be plotted similarly to those from the entire dataset.



This looks roughly the same as the plot of all events, showing that the ones marked as significant follow a roughly similar distribution.

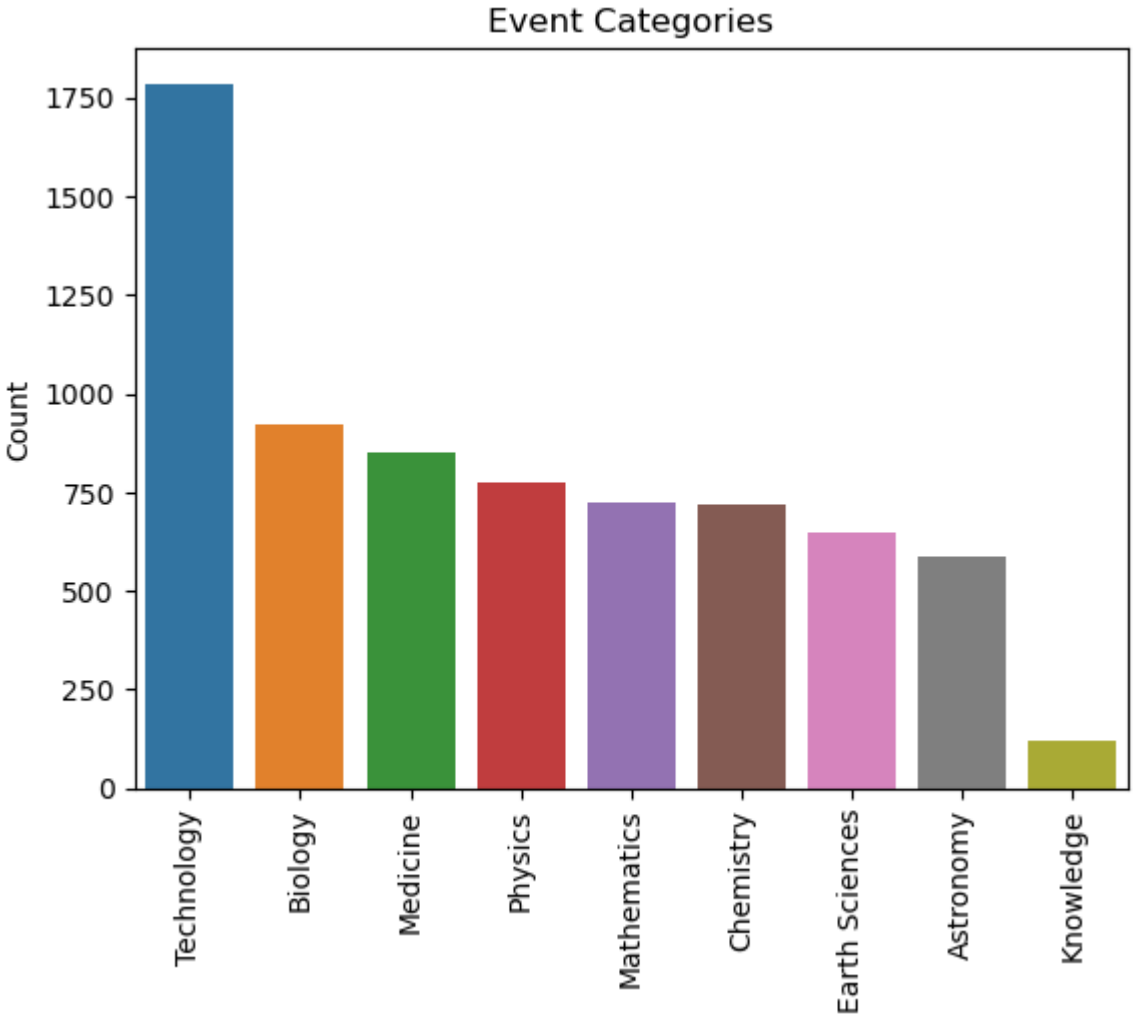
Locations

The location is the first text-based field that will be analyzed. It designates a country or region where the event took place. For the ancient era, these are broad designators such as "Ancient China," and country names such as "USA" for the modern era. One way to visualize the importance of different locations is a word cloud, scaled by frequency of a country in the entire dataset. This will become somewhat unreadable if all the values are included, and many of the locations only have one or a few events associated to them. Some of these locations also list multiple countries ("Germany UK," for instance). So the top 20 locations will be used.

Here it is clear from the graphic that the top locations across the huge period of time represented in the dataset are Britain, Germany, France, USA, etc. Perhaps an element of Western ethnocentrism becomes evident, given that the largest countries representing the most counts are all European (or the USA). Other societies or civilizations are present as well, though, including Ancient Egypt, Ancient China and Ancient Mideast (China appears to be the only non-Western country represented in the top 20 aside from ancient civilizations). Given the nature of the dataset, focusing primarily on Western-style scientific discovery and theory, the emphasis on a certain set of countries shown here is perhaps not surprising.

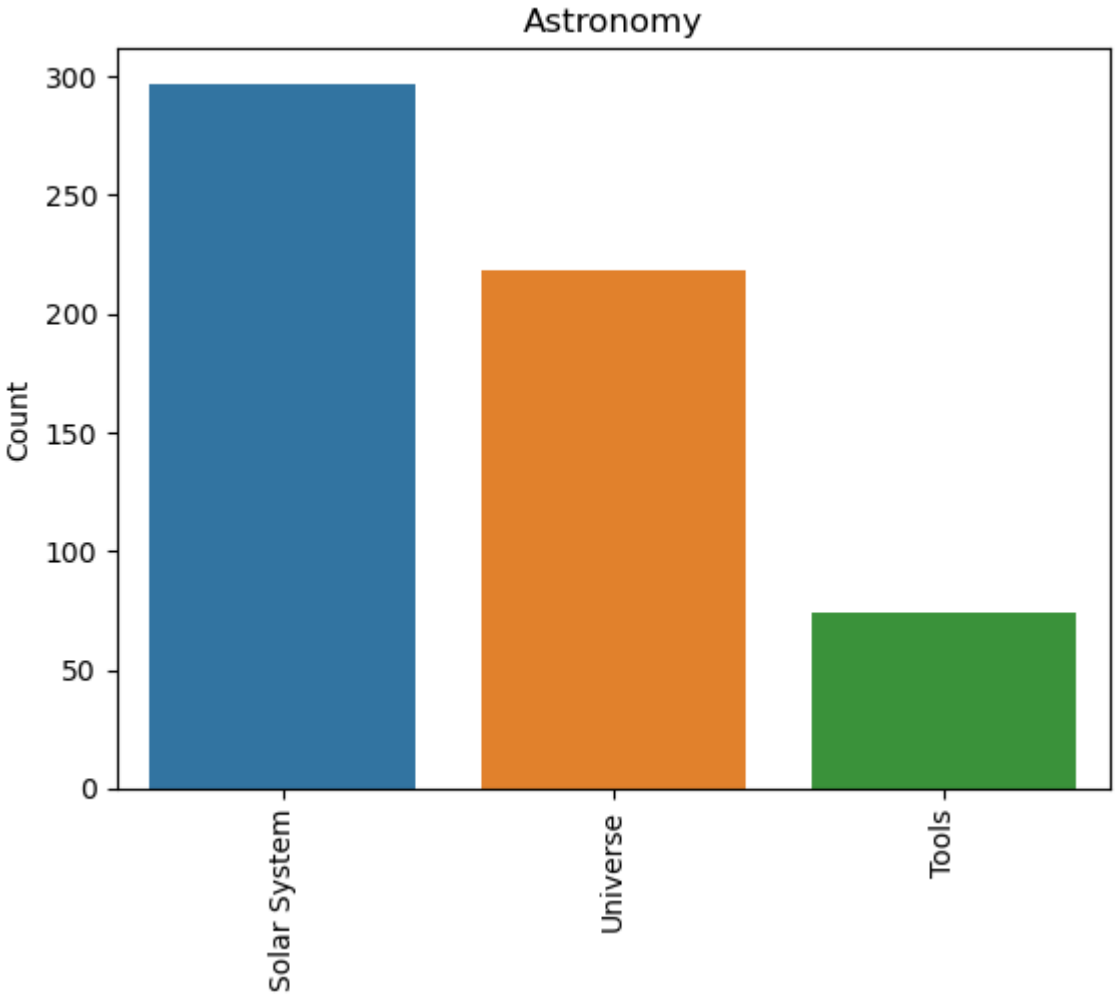
Categories and Subcategories

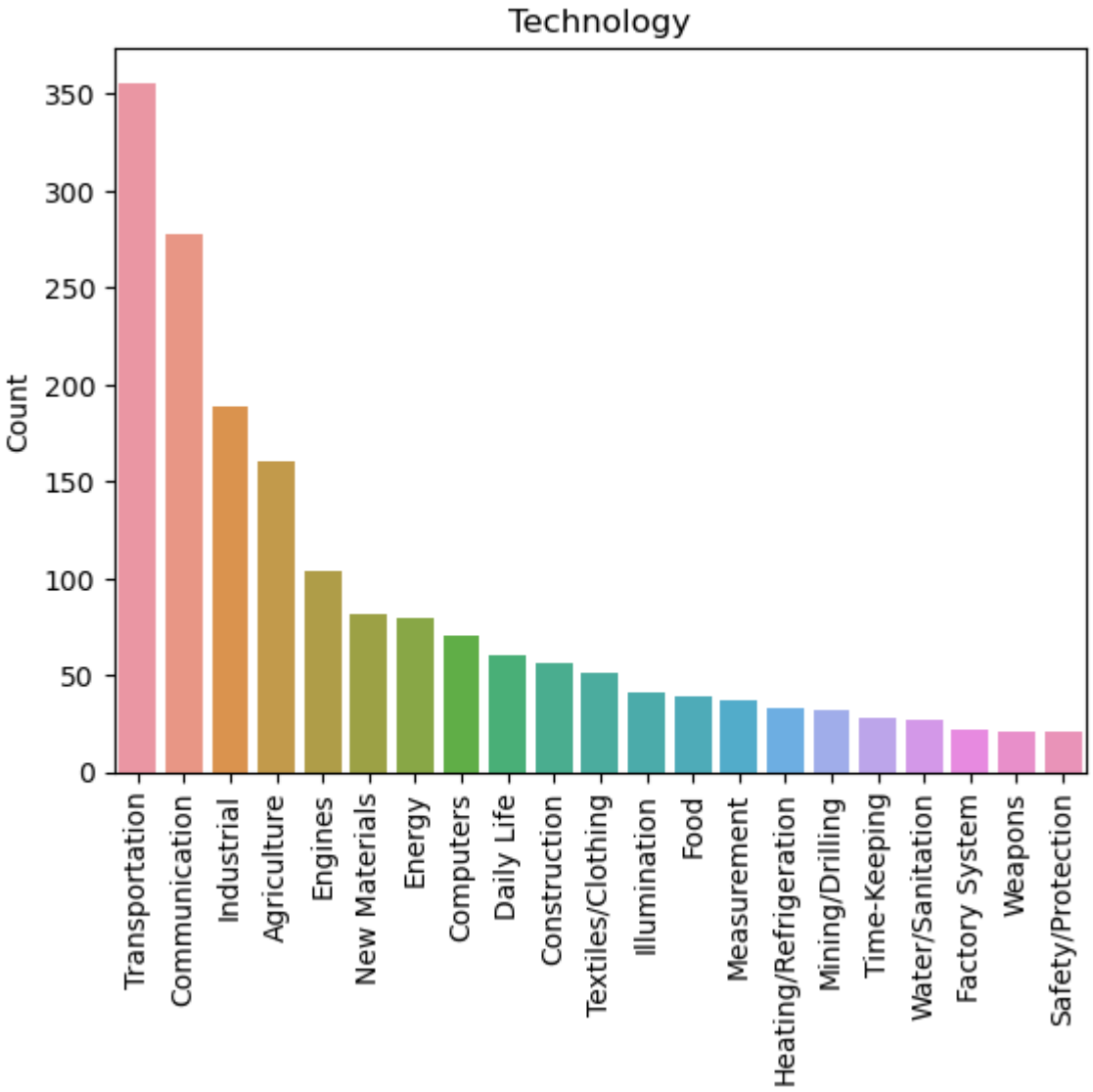
The next set of text fields covered in the analysis are categories and subcategories. The categories represent broad designators of the event type such as "Technology" or "Astronomy." The subcategories further divide these into more specific topics. Bar charts are used to show event counts for these categories, and subcategories are displayed within each category.

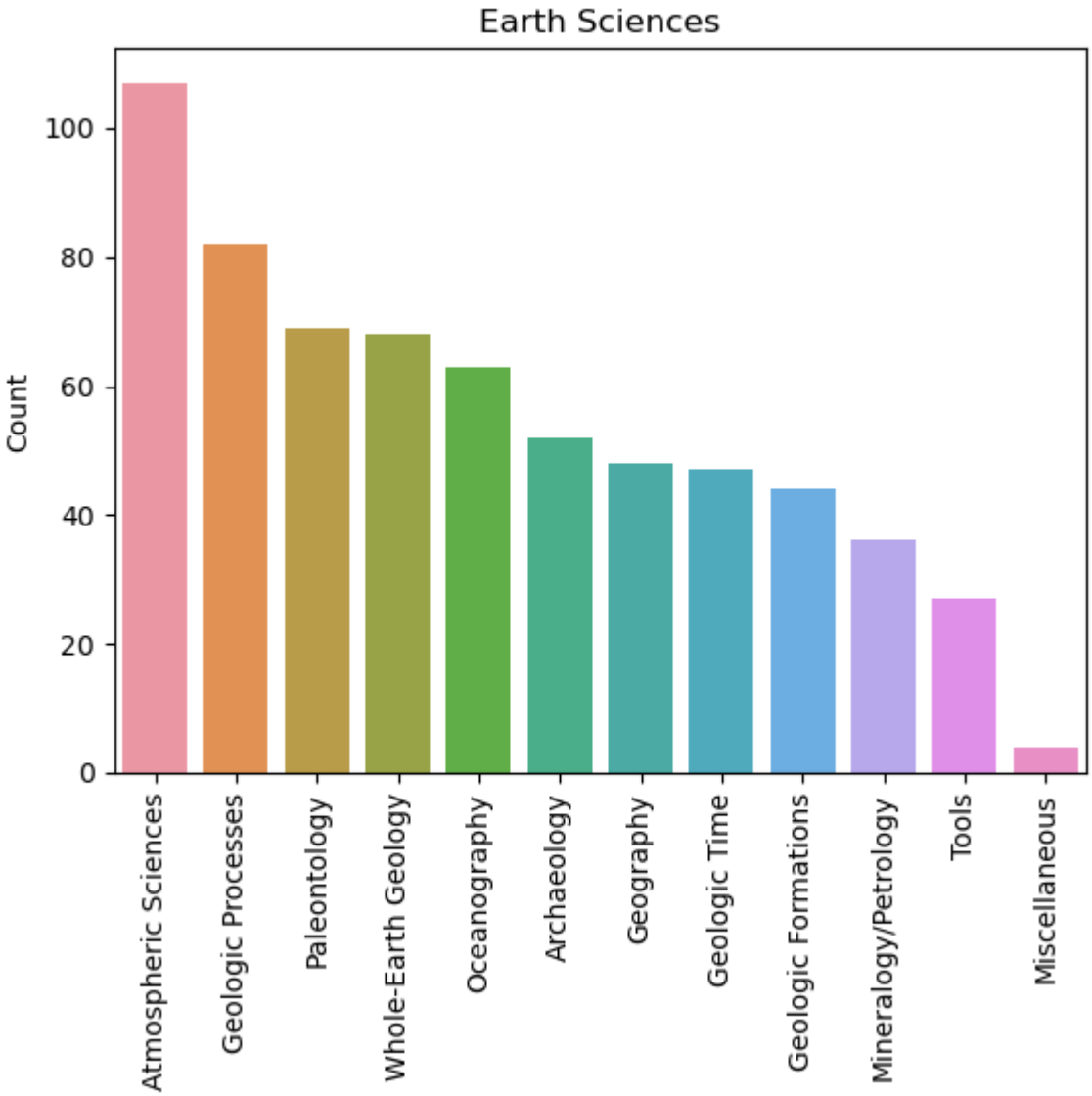


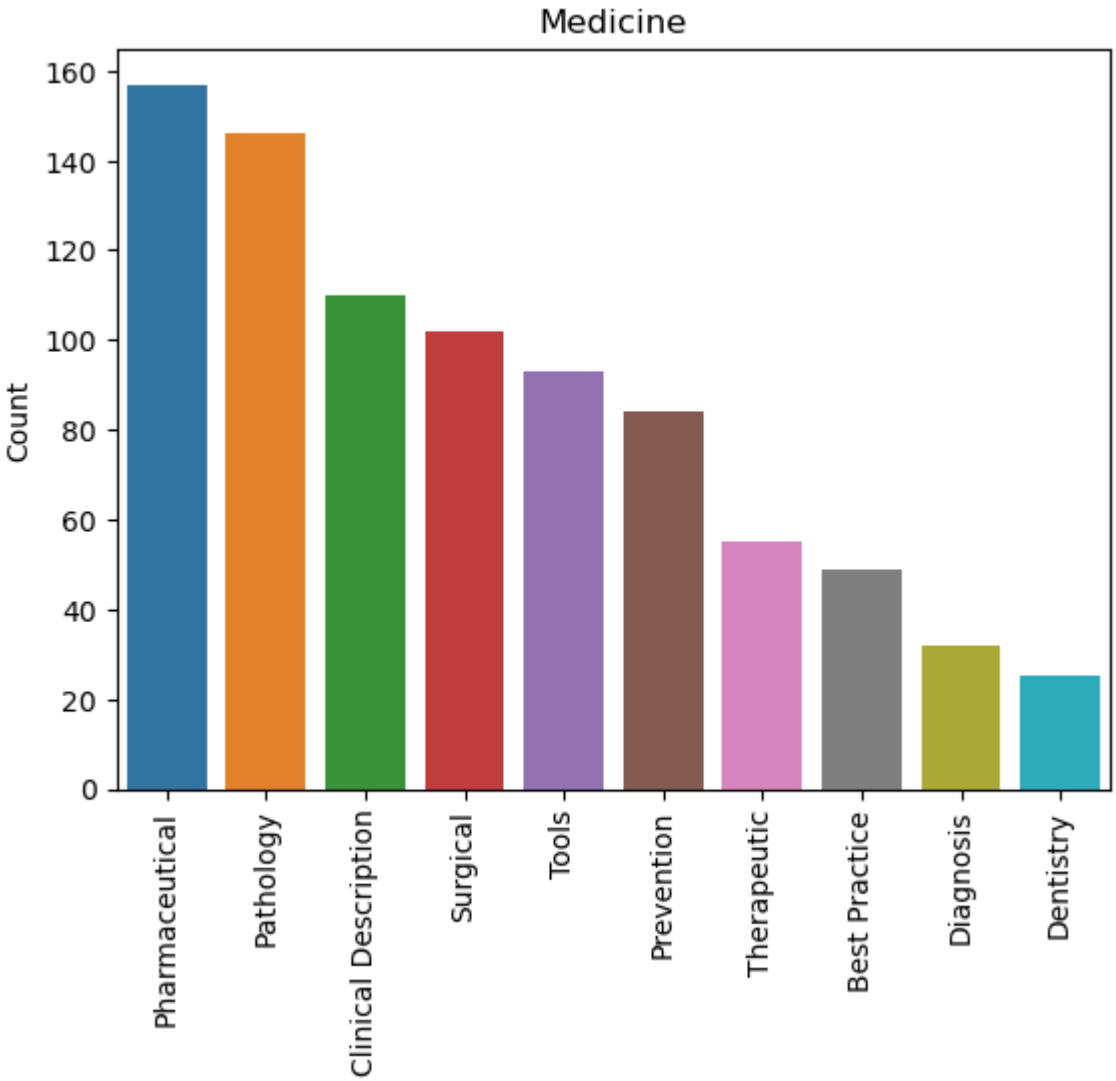
Technology is the most heavily represented category, which, again, makes sense given the nature of the data. The relative counts of the other categories are not too different from each other, indicating that the selection of events, while derived from a qualitative process, is fairly evenly distributed across domains.

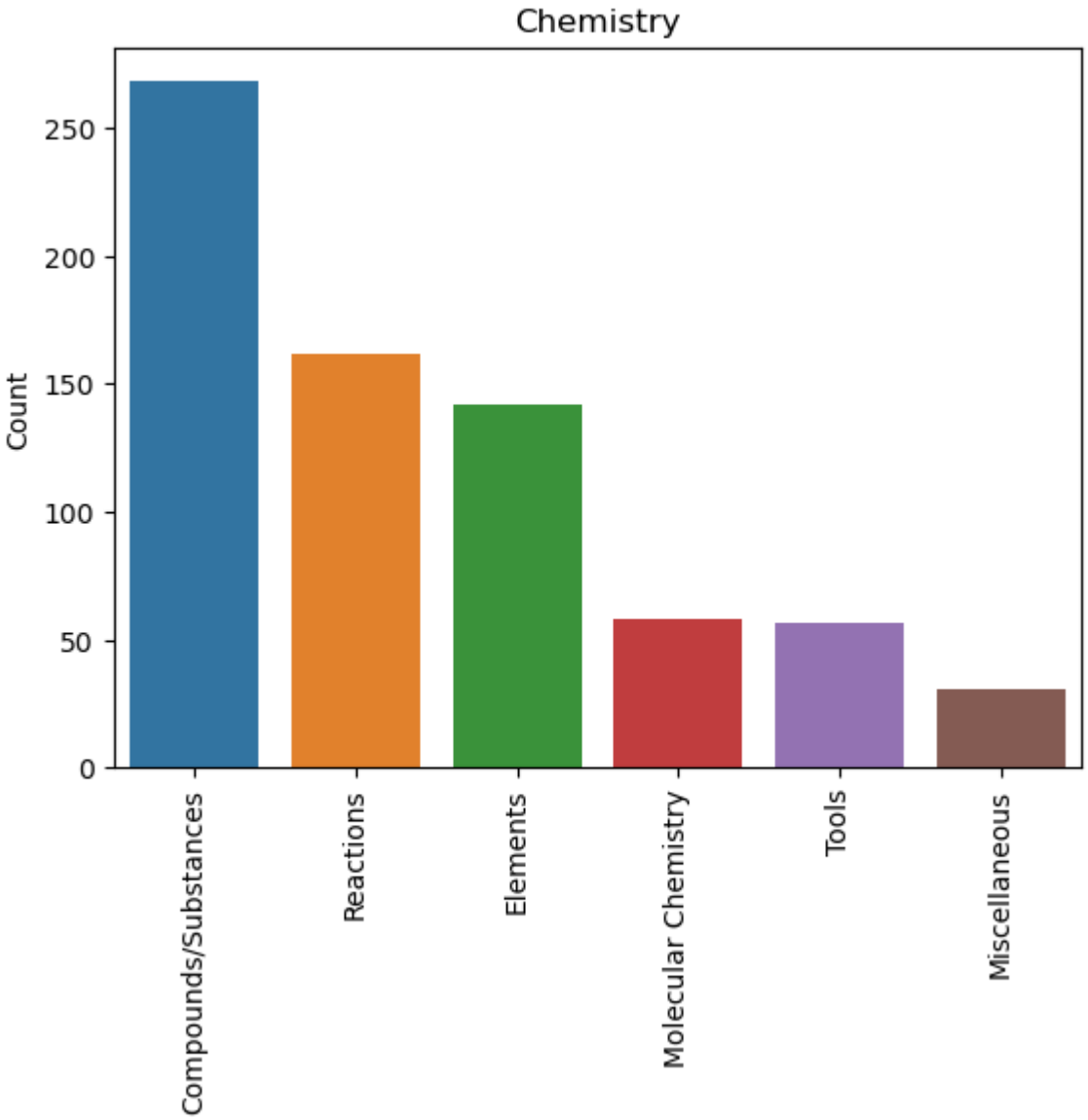
Next, the subcategories are charted similarly and displayed for each category, with the events being first filtered to a specific category. The subcategories within that category are then counted.

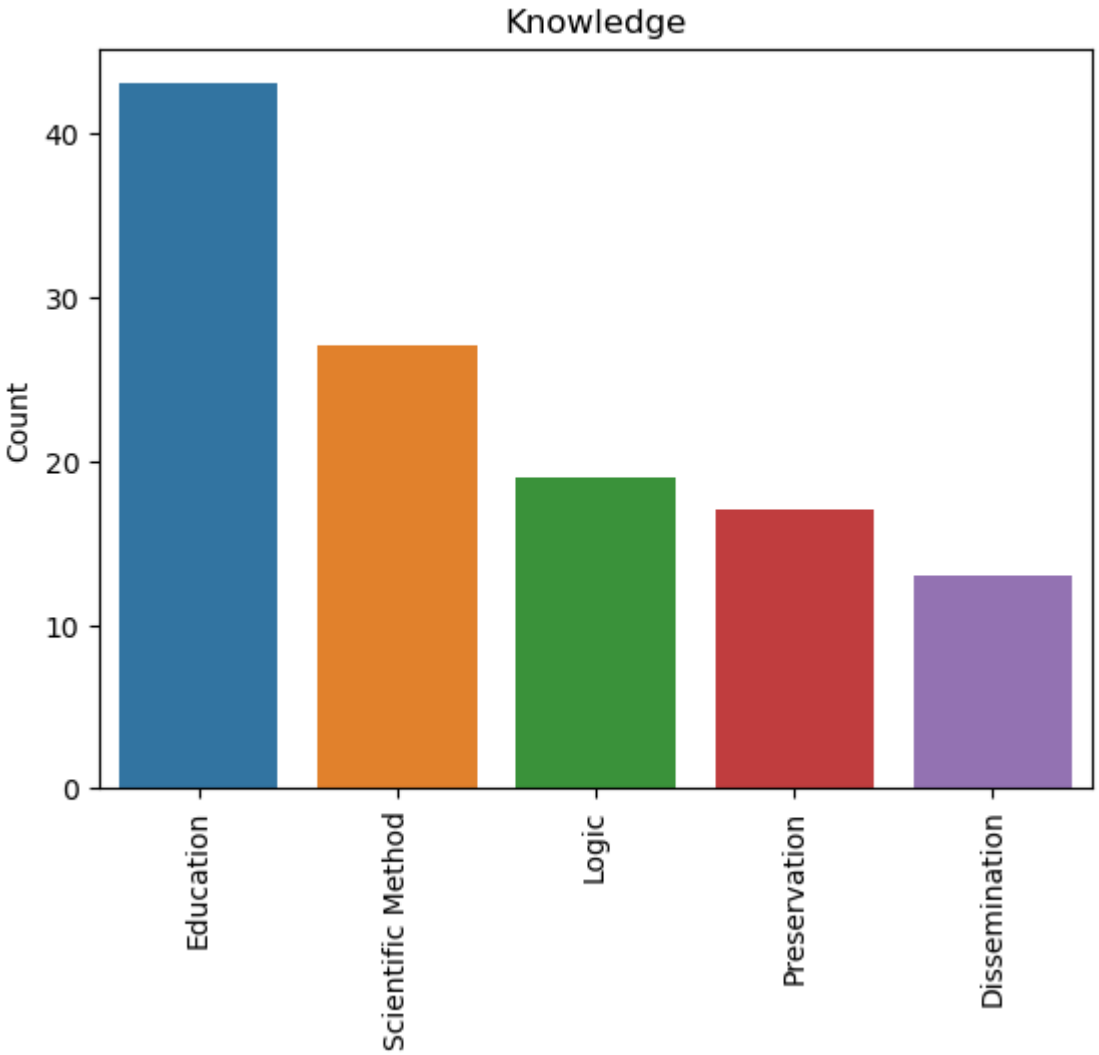

Astronomy
Count
300
250
200
150
100
50
0
Solar System
Universe
Tools

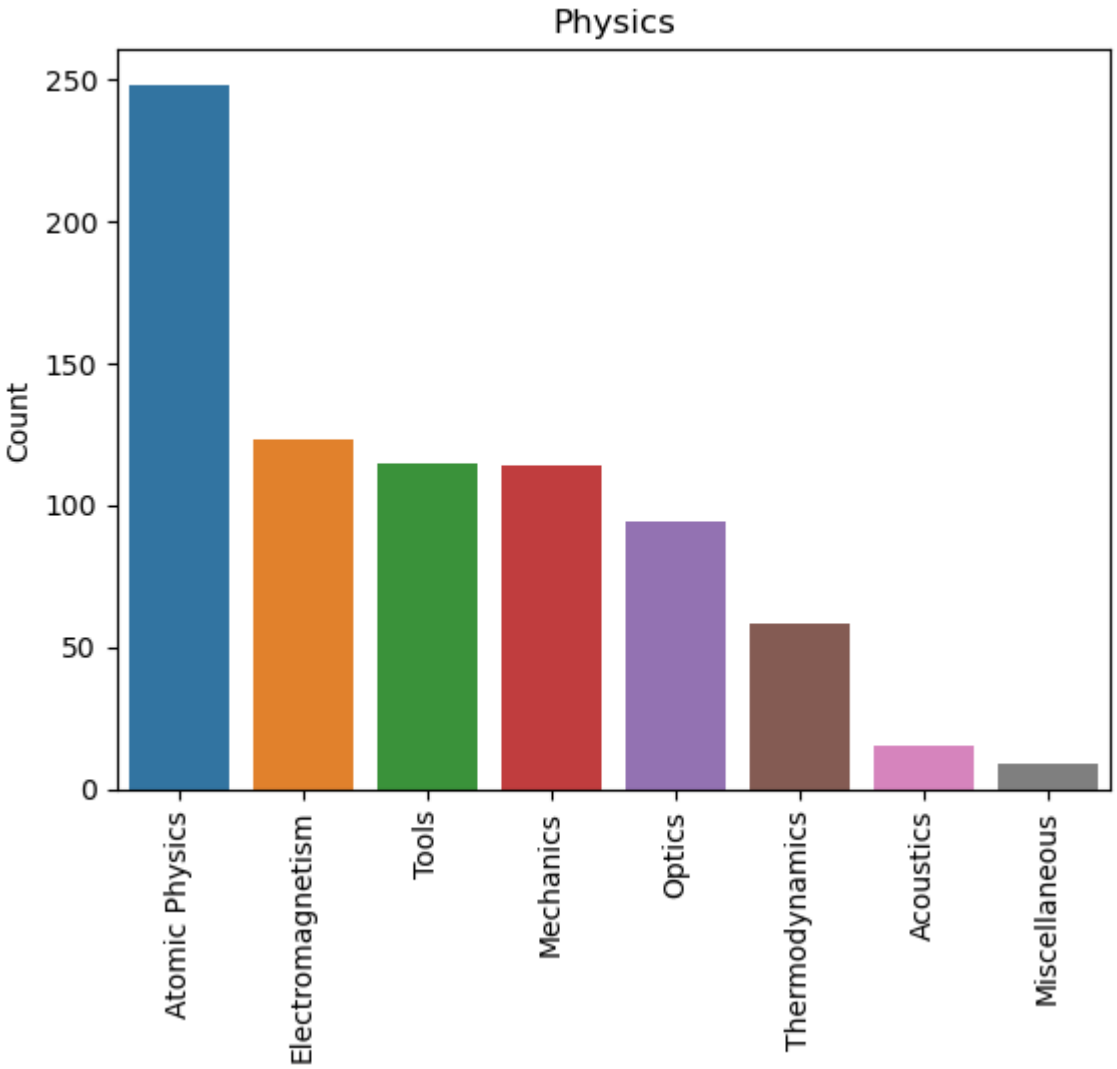

Technology
Count
0
50
100
150
200
250
300
350
Transportation
Communication
Industrial
Agriculture
Engines
New Materials
Energy
Computers
Daily Life
Construction
Textiles/Clothing
Illumination
Food
Measurement
Heating/Refrigeration
Mining/Drilling
Time-Keeping
Water/Sanitation
Factory System
Weapons
Safety/Protection

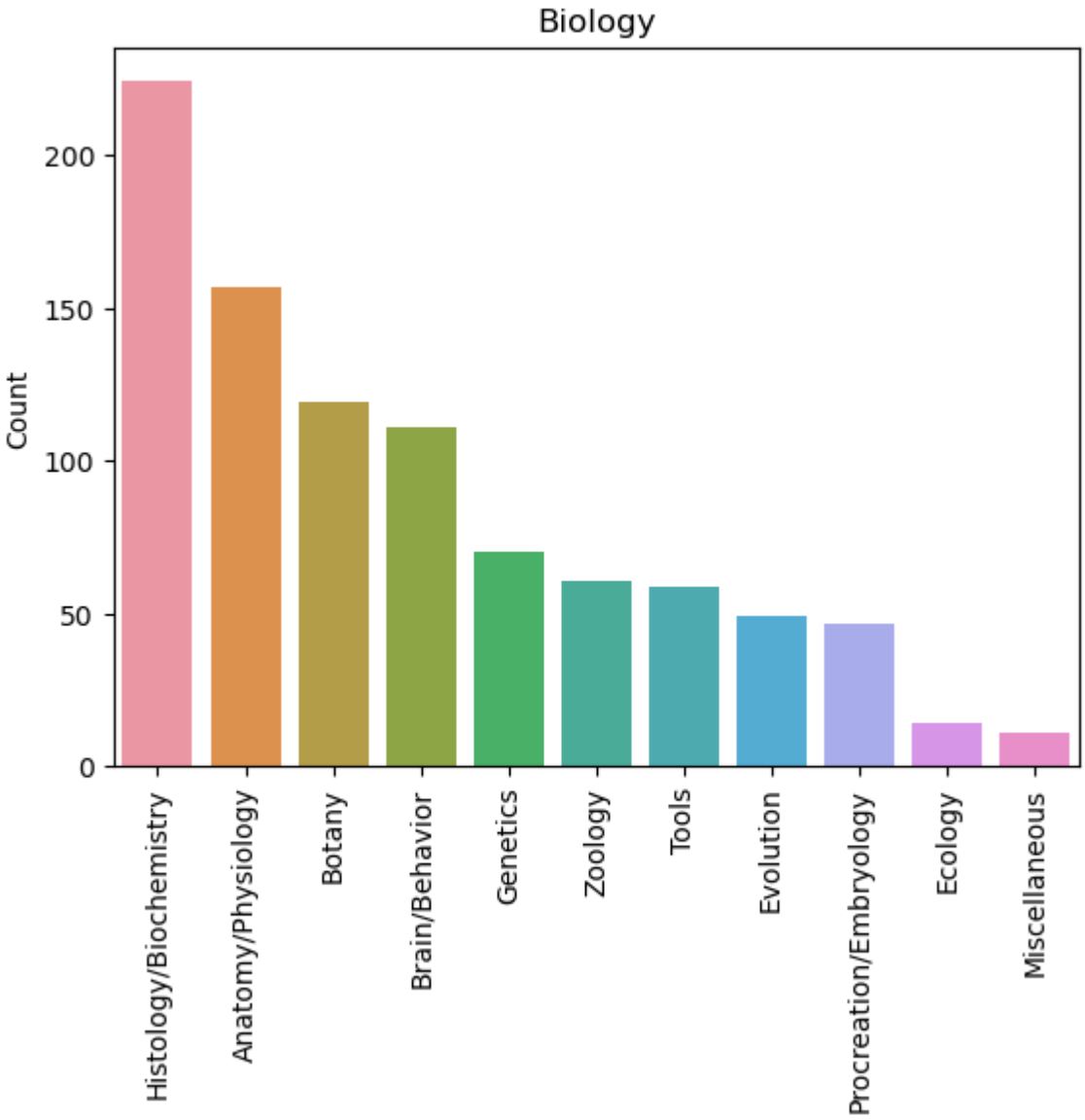

Earth Sciences
Count
0
20
40
60
80
100
Atmospheric Sciences
Geologic Processes
Paleontology
Whole-Earth Geology
Oceanography
Archaeology
Geography
Geologic Time
Geologic Formations
Mineralogy/Petrology
Tools
Miscellaneous


Medicine
Count
160
140
120
100
80
60
40
20
0
Pharmaceutical
Pathology
Clinical Description
Surgical
Tools
Prevention
Therapeutic
Best Practice
Diagnosis
Dentistry


Chemistry
Count
0
50
100
150
200
250
Compounds/Substances
Reactions
Elements
Molecular Chemistry
Tools
Miscellaneous


Knowledge
Count
40
30
20
10
0
Education
Scientific Method
Logic
Preservation
Dissemination

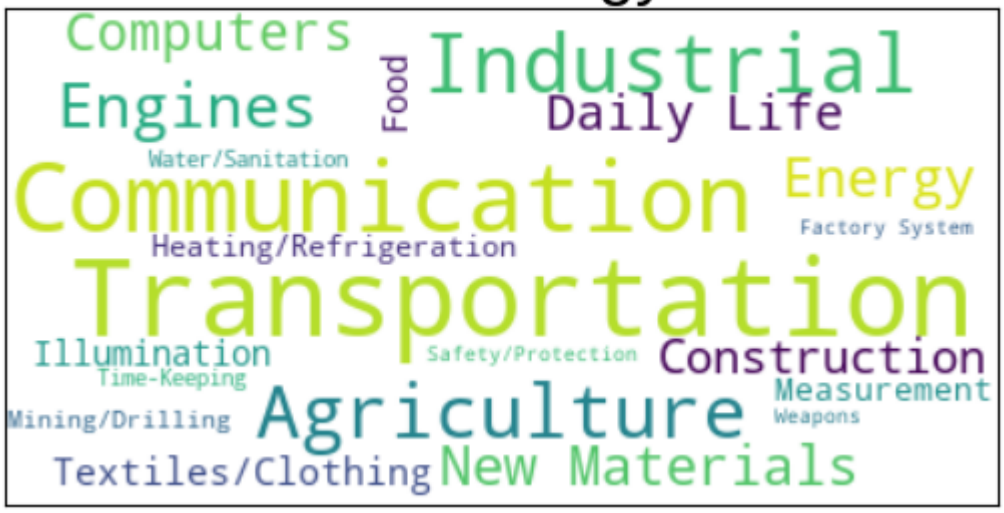

Physics
Count
250
200
150
100
50
0
Atomic Physics
Electromagnetism
Tools
Mechanics
Optics
Thermodynamics
Acoustics
Miscellaneous



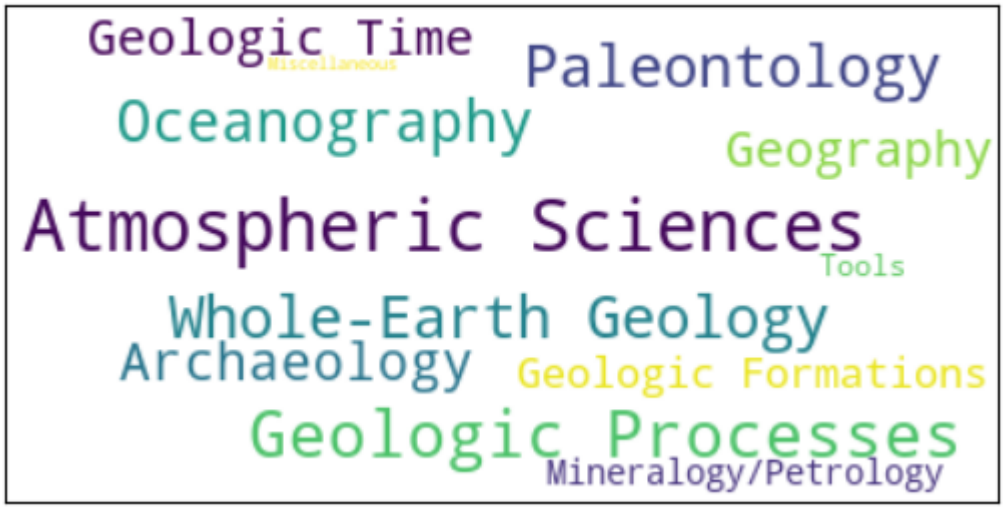
Here there is a wide variety in the number of subcategories within each category, as well as interesting distributions of the counts. This illustrates the heterogeous nature of the data, where some categories have many sub-domains, like "Technology," whereas others like "Astronomy" have only a few. Some categories also seem to have one or only a few dominant subcategories, as in "Atomic Physics" within "Physics," whereas others like "Earth Sciences" or "Medicine" are somewhat more evenly distributed. The relative complexity of each category can be discerned here. For instance, "Technology," being a very broad topic, has the largest number of subcategories. On the other hand, the "Mathematics" category has no subcategories! Since this category does, in fact, have a large number of valid subcategories (geometry, calculus, and so on), the qualitative nature of the data gathering process and the knowledge or biases of the researcher become apparent. (A mathematician would surely have categorized the events differently.)

An alternate method of presenting the information on subcategory frequencies is a word cloud. A few of these are shown below.

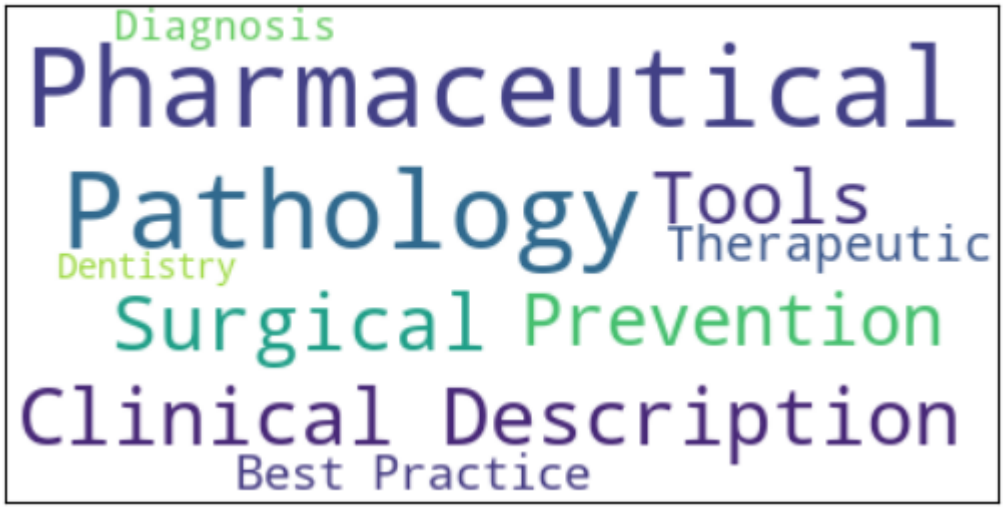
Technology

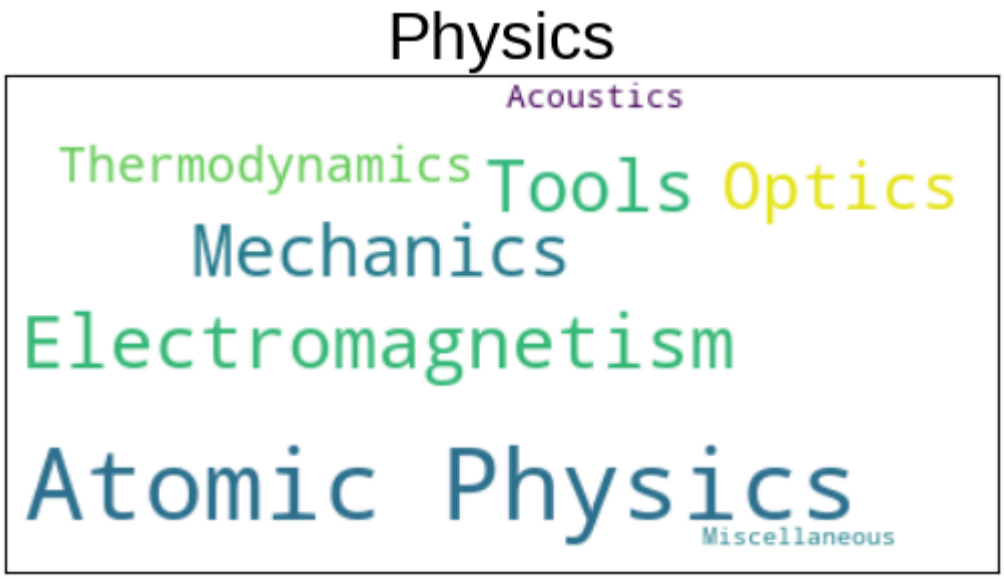


Earth Sciences



Medicine



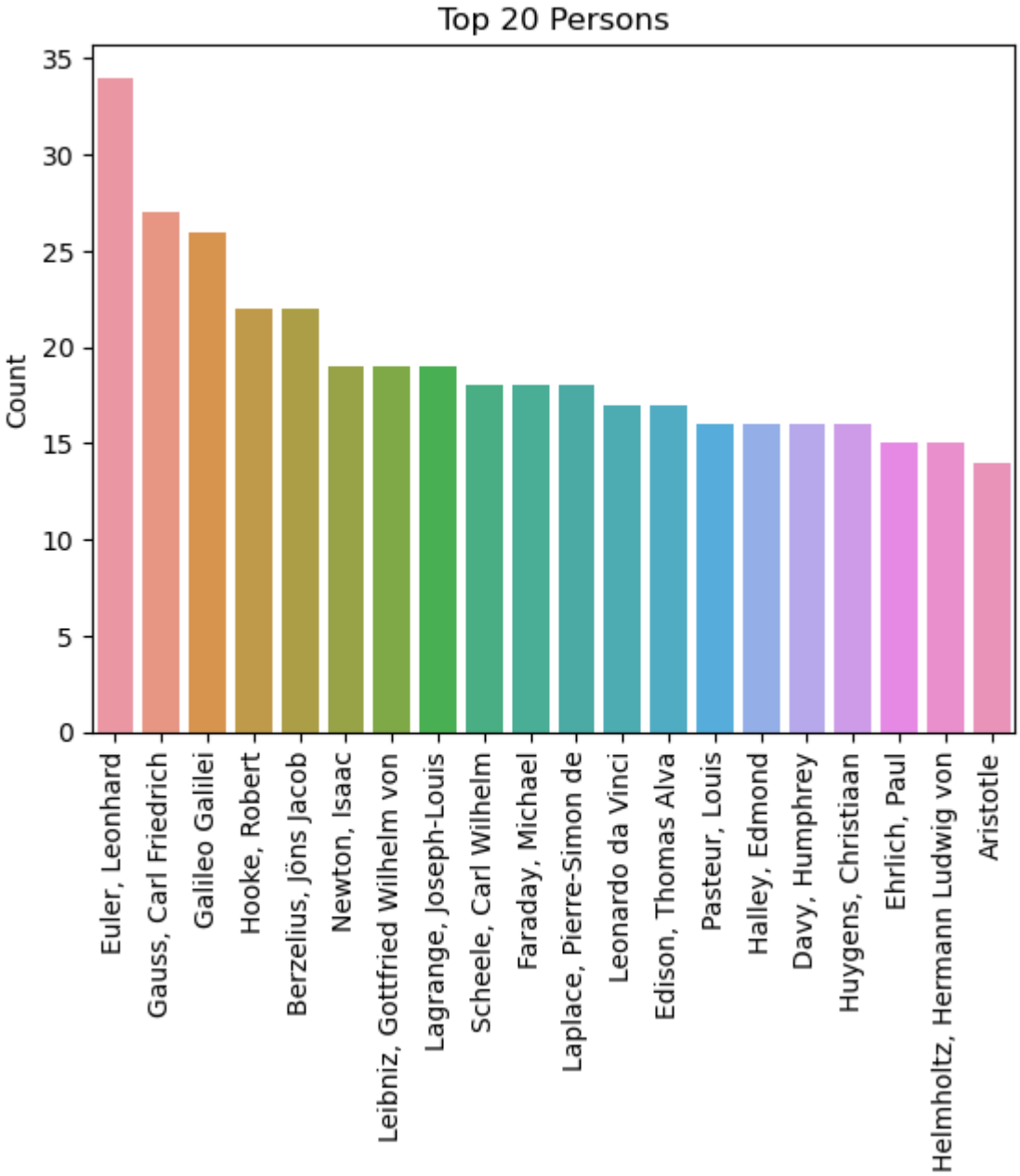


It is interesting to note the different effect in the viewer that the word clouds induce compared with the bar charts. The bar charts, sorted by count, seem to emphasize the relative importance of each subcategory by rank. On the other hand, the word clouds do not show these differences as much, suggesting more equality between subcategories. This is partially due to the interpretation of written words as uniformly relational, regardless of their size in the visualization. Small differences in text sizes are also more difficult to discern than relative bar heights. While an interesting device, word clouds are not ideal for presentation of ranked, qualitative data.

People

Thousands of different individuals are represented as the person associated with an achievement, ranging from ancient philosophers to modern scientists. These are denoted using a text field in the dataset. The top 20 individuals with the most achievements are shown in the chart below.

person	Event Count
Euler, Leonhard	34
Gauss, Carl Friedrich	27
Galileo Galilei	26
Berzelius, Jöns Jacob	22
Hooke, Robert	22
Leibniz, Gottfried Wilhelm von	19
Newton, Isaac	19
Lagrange, Joseph-Louis	19
Faraday, Michael	18
Laplace, Pierre-Simon de	18
Scheele, Carl Wilhelm	18
Leonardo da Vinci	17
Edison, Thomas Alva	17
Halley, Edmond	16
Davy, Humphrey	16
Huygens, Christiaan	16
Pasteur, Louis	16
Helmholtz, Hermann Ludwig von	15
Ehrlich, Paul	15
Koch, Heinrich Hermann Robert	14



All of these are men, which is also a consequence of the qualitative reasoning used to gather the data. However, there are some women represented, just not associated with as many events.

Similarity Measurement

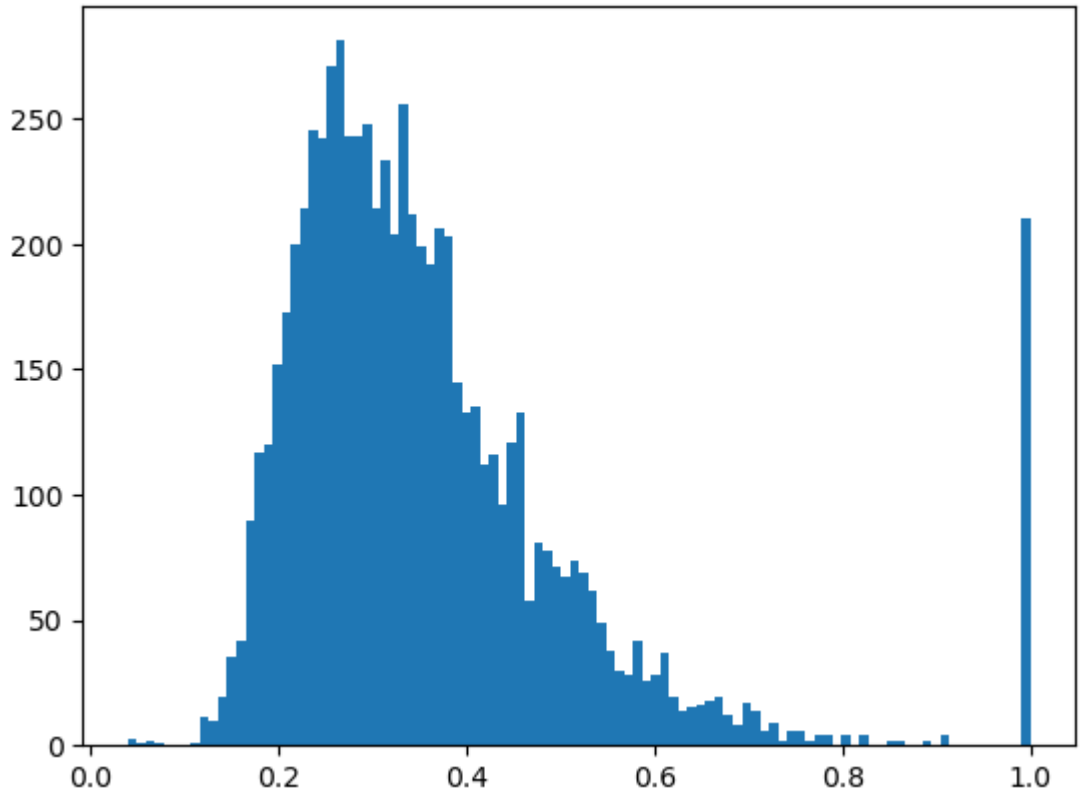
Each event has a unique description with it. This makes the content difficult to analyze with traditional methods. Techniques used in recommendation engines can be used to determine which of these descriptions are similar to each other. The raw text can be transformed into a feature vector and then the cosine similarity calculated to determine which descriptions are similar to others based on this scoring.

Some example rows from this process are shown below.

```
[[1.         0.01371014 0.         ... 0.00960077 0.02976811 0.00630068]
 [0.01371014 1.         0.         ... 0.01637584 0.01082964 0.        ]
 [0.         0.         1.         ... 0.06802356 0.         0.        ]
 ...
 [0.00960077 0.01637584 0.06802356 ... 1.         0.00758365 0.        ]
 [0.02976811 0.01082964 0.         ... 0.00758365 1.         0.02087683]
 [0.00630068 0.         0.         ... 0.         0.02087683 1.        ]]
```

```
7131
```

The scores of the most similar descriptions for each record are shown below with a high granularity (each bin is 0.01 out of 1). Just over 200 scores seem to be very similar, though this is a relatively small proportion out of the more than 7000 records. Perhaps these are close to being duplicates. The rest of the scores show a slightly right-tailed distribution with the peak around 0.3, indicating that most of the descriptions do not have very similar records in the dataset. As the text is supposed to represent "unique" human achievements, this could be taken as a good sign that the researcher achieved their goal of representing events in this fashion.



Similar to a recommendation engine, an event can be picked and other similar records located in the dataset based on the description. The description below was picked somewhat arbitrarily for this.

```
Claude Shannon outlines the nature of a program for a chess playing computer.
```

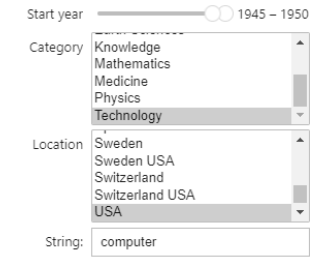
The most similar records to this description are listed below.

```
['John Eckert designs the first stored computer program, for EDVAC.',
 'Maurice Wilkes designs EDSAC, the second computer built with a program.',
 'Leonardo Torres y Quevedo designs a chess-playing machine, an early attempt to imitate complex
human thought processes with Boolean logic.',
 'Georges Claude discovers the Claude process for the bulk liquefaction of air.',
 'Jöns Berzelius begins a program for the chemical analysis of meteorites.',
 'Claude Berthollet publishes a major work on the nature of chemical affinity publishes.',
 'Peter Goldmark invents the long-playing record.',
 'Claude Shannon conducts the first systematic work on information theory, a general approach to
all kinds of information handled electronically, including symbolic logic.',
 'Gottfried von Leibniz outlines the principles of the aneroid barometer.',
 'Claude Shannon\'s "A Symbolic Analysis of Relay and Switching Circuits" proves that it is poss
ible to use on-off relays (later transistors) to solve Boolean algebra problems, a founding docu
ment of the mathematical theory of information.']
```

It is interesting to note that none of these are really all that similar in terms of the text. There are common terms such as "computer," "chess," and the person "Claude Shannon" present amongst them. This indicates that what a mathematical procedure considers "similar" may be different to what we might actually categorize as such based on our own knowledge. This may depend on the fidelity of the information, itself, or how it has been expressed in the text.

Search Interface

Given the nature of the data in terms of being interesting to filter and explore, a simple user interface was developed using ipywidgets. This included a range slider for year, multiple select boxes for category and location, and a search box for the description. An example of using this interface is presented below. The query interface below allows searching by year range, categories, locations or a phrase in the description.



	year	location	category	description	person
6960	1945	USA	Technology	John von Neumann's "First draft on a report on the EDVAC" proposes novel elements, including a memory for storing programs & recommendation that codes be treated as numerals, a foundational text for the computer revolution.	Neumann, John von
6979	1946	USA	Technology	ENIAC, the first entirely electronic computer, developed by John Eckert, John Mauchly, Arthur Burks & John von Neumann, becomes fully operational.	Eckert, John Presper, Jr.
6980	1946	USA	Technology	Arthur Burks, Herman Goldstine & John von Neumann publish "Preliminary Discussion of the Logical Design of an Electronic Computing Instrument," providing the conceptual foundation for computer development.	Burks, Arthur Walter
7050	1948	USA	Technology	John Eckert designs the first stored computer program, for EDVAC.	Eckert, John Presper, Jr.
7088	1949	USA	Technology	John Eckert oversees the construction of BINAC, the first US computer using programs stored electronically.	Eckert, John Presper, Jr.
7126	1950	USA	Technology	Claude Shannon outlines the nature of a program for a chess playing computer.	Shannon, Claude Elwood

Reflection

This was not a purely textual dataset, but the additional fields provided an interesting context to the analysis. The category and subcategory were text, though treated here as categorical variables. This seems like a powerful approach when the number of unique values is manageable. There were a large number of subcategories, but these could be broken down by category in order to make the presentation more managable. The usage of similarity algorithms was similar to a recommendation engine, though generalized to find events here. Given more time, this type of analysis could have been performed on interesting subsets of the data or filtered versions of the description. For instance, the names of the associated person were included in many of the descriptions. Removing these may have lead to more interesting results in the "recommendations" of similar events. Additionally, similarity measurements could have been done only within certain categories. Finally, this was my first time really experimenting with the particular GUI toolkit (ipywidgets) that I used to mockup a search interface, which was enjoyable to use for data exploration. This level of interactivity is particularly powerful when dealing with textual data where it can be enlightening to narrow down results by words or phrases, and I plan to include this type of functionality more in the future in my notebooks.