Data Manifesto

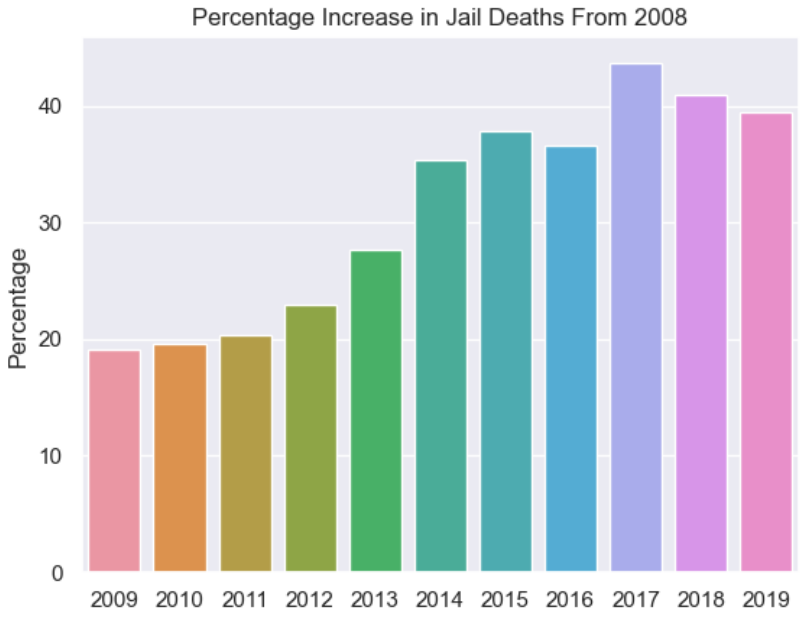
What is data and how does it acquire meaning?

I see ***data*** as gathered attributes for a "to-be-determined". The source of individual instances of data is often clear, for example, when asking the favorite color of *n* people and placing the results into a data frame it would be clear that each color is an attribute of one of *n* people. Although, this is not why data is gathered. The "to-be-determined" is not the parent of an attribute of a specific instance, but instead a product of a question asked by data scientists that connects many data instances. The "to-be-determined" regarding the data frame of *n* people's favorite colors may stem from the question of, "What is the most common favorite color among *n* people?" Or, "How many of *n* people favorite the color blue?"

id	state	county	jail	year	date_of_death	full_name	last_name	first_name	mid_name
13.0	AL	Lee	W.S. Buck Jones Detention Center	2017	12/8/2017	NaN	Winningham	Danny	NaN
8.0	AL	Marshall	Marshall County Jail	2017	4/15/2017	Jemal Seid Mohammad	Mohammad	Jemal	Seid
9.0	AL	Mobile	Mobile County Metro Jail	2017	7/14/2017	Ryan Scott Burkhardt	Burkhardt	Ryan	Scott

Example of data, attributes that are not given meaning until analyzed together

Information makes a more concise name for "to-be-determined." Information is the parent to the many attributes, or data, that allow for meaning to be conveyed in a structured and useful manner. If there is a conclusion that blue is the most common favorite color among *n* people, this is information which was derived from a collection of attributes. Information illustrates trends which are attributes for another "to-be-determined."



An example of Information, a structured and organized trend derived from data

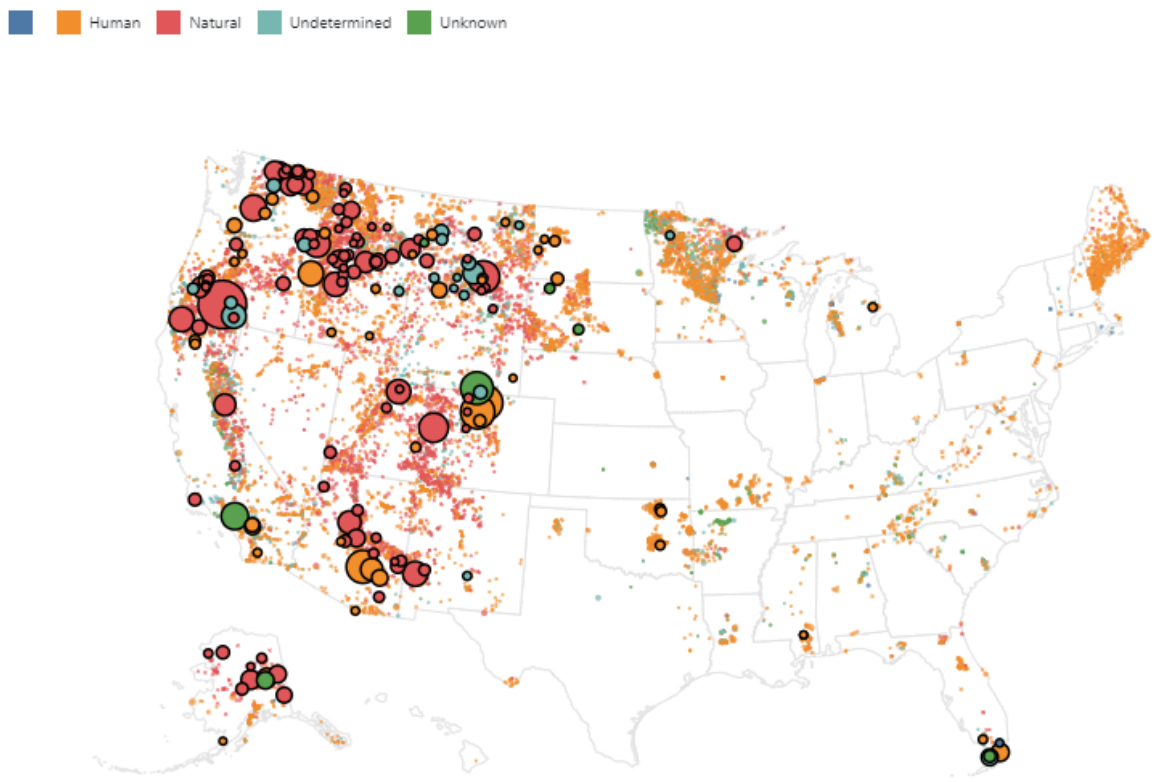
Whatever data scientists seek to use information for will be reflected in this next "to-be-determined" as ***knowledge***. Knowledge is the synthesis of new, larger, and more complex attributes (information). Take a new collection of data consisting of *n* people's one-word description of how the sky or the ocean makes them feel. Ask a question to derive information from this data, such as "What is the general emotional tone of the people's descriptions?" Assuming that most people described the sky as something along the lines of beautiful, and the ocean as calming or peaceful, using sentiment analysis the emotional tone would be positive. Where knowledge is achieved is using the information from both datasets to answer the question of, say, "Why do people like the color blue?" While this is hypothetical, the information showing the positive emotional response to things that are blue (sky, ocean) synthesized with the information showing blue as the most common favorite color forms into knowledge. This knowledge being: blue is the most common favorite color because positive emotional responses are associated with things that are blue.

The acquisition of ***wisdom*** occurs as a product of reflection over oneself and the methods used in gathering knowledge. Wisdom offers insight into better decision making and the recognition of patterns which, in turn, make gathering knowledge a more efficient process. Most importantly,

wisdom is the opportunity of integrating knowledge for the purpose of improvement. I say most importantly because the nature of this world is to strive for efficiency. I find this dangerous as it encourages an abundance of consumption that creates the idea of more consumption equating to more value. Hopefully, the wisdom gained from striving for efficiency will one day become wisdom expressing the danger of it.

What makes a data scientist?

Through my experience, data science is a field requiring an open mind. While there are tasks that can be completed routinely, intensive tasks require a great deal of data wrangling that require different approaches. A personal experience I had with this was first being introduced to geospatial visualization using JavaScript. Firstly, I had never worked with JavaScript before. Secondly, I had never worked with geospatial visualization before. Because of this I was lost and overwhelmed, a position that I often found myself in. Despite not knowing how to approach the problem, it is important that a data scientist remembers their problem is not the first and that there are always tools and instructions available.



In my case with creating this map of wildfires, I was instructed via an observable notebook and able to create a map relatively easily without understanding the code. While it is ideal that a data scientist understands their code, I believe that being open to the resources available is a defining characteristic of one. This may sound contradictory to my earlier comments on the dangers of efficiency though that is not my intent. The reality is, a data scientist will find themselves in circumstances where they must work with a language they are not proficient in, or are given specific tasks that will not translate into their greater knowledge. What I mean by this is that there are too many tasks for a data scientist to be proficient in all of them. The decision to remain open to resources and the new ways of solving problems presented in them will keep a data scientist from stubbornly brute forcing their existing knowledge into a solution, and provide new ways of approaching problems upon reflection.

What skills do you need to do data work?

While you can reach many solutions by using outside resources, to do data work you do need to know how to manipulate data proficiently. A data scientist cannot afford to reference something for every line of code. Not only is it defeating if you do not know how anything is working, it will limit how creative and personalized your data analysis can be. To do valuable data work, a data scientist needs to be creative. A data scientist needs to ask ***their*** question in which they can turn information from the data into knowledge that answers that question. If a data scientist has a very specialized question, it will likely require the joining of a great deal of information. The data scientist must know how to make these work together efficiently, as there may be bigger problems for which resources can be used. It should go without saying that proficiency in Python, SQL, R, or any frequently used data analysis language is necessary for a data scientist. I use Python as it is most familiar to me and straightforward. Accessing columns and rows within a data frame is one of the first things I learned when working with the pandas module in python, and it continues to be something I do in almost every data science task.

```
most_common_age = df[(df['gender'] == 'M') & (df['age'] == 51) | (df['gender'] == 'F') & (df['age'] == 28)]
```

```
cause_short_common_age_df = most_common_age[most_common_age['cause_short'].isin(['M', 'S','DA','AC','H'])]
```

The code above is from my most recently completed data science project. In the first line, I am accessing the ‘gender’ and ‘age’ columns to create a new data frame with only rows containing the specified values. The second line is similar, the ‘cause short’ column is being accessed and only the rows with 'M', 'S','DA','AC','H' I want to keep. Knowing how to access individual columns and their contents is essential in being able to manipulate data.

Advice I would give to new data scientists.

Most of the data a data scientist will work with will not be in a generous form. A great amount of effort and time frequently needs to be put into parsing data from the web/APIs so that it is in a format that you can begin to work with. In my experience, data analysis has proven to be the easier and fun part. Scraping data from the web is something I am not proficient in one bit, I would recommend that new data scientists begin working with web scraping as it unlocks so many possibilities. I also recommend that new data scientists remain patient and calm with it because it can be frustrating. Web data can be in formats that require extensive coding to actually make it readable and manipulatable, because of this I have become discouraged with it many times, so try to keep your spirits up.

What problems can and can't be solved?

When beginning a data analysis it is important to have either a goal or a question. Maybe that goal is to find and answer questions upon viewing a dataset. Regardless of what your goal or question may be, a data scientist must keep in mind that answers to their questions could be presented in numerous ways. Not finding a correlation between data is just as significant as finding one, though it is easy to feel defeated if there are no trends jumping out from your analysis. One might try to manipulate their data so that the information derived from it is more interesting, this would of course result in ethical issues and data biases. Data may offer a solution to a problem or suggest ways of approaching another problem, it is strictly a tool that should not be weaponized by those forming their desired solution. Ideally, a data scientist should seek to be taught by the data when seeking an answer to a problem instead of manipulating data for the purpose of drawing preconceived conclusions.