Hayden Garner
5/10/2023
Introduction to Data Science
Professor Wirfs-Brock

Data Manifesto

Data differs from information, knowledge, facts and other terms used to describe the understanding of something by eliminating anything but the bones of the information. While data is similar to other methods of knowledge by being inherently descriptive in nature, it leaves full space for analysis by omitting its own interpretation. In this way data is generally much more useful than knowledge and facts because although they have their own merits, they occupy a space of more informal, less descriptive and ultimately less useful support for a given subject. On the other hand data is malleable and essentially formless and has the capacity to provide support to a much wider range of projects that may have otherwise been inaccessible.

For example, in a data analysis about crime rates in Los Angeles, knowing the facts about each crime would be helpful, but ultimately overwhelming due to the sheer amount of extraneous data that accompanies the relevant. If you wanted to make a map with a dot representing every instance of crime committed, knowing things like the alleged perpetrator and the type of crime is just another obstacle needing to be overcome. Facts, knowledge and other terms to describe information carry this same burden. The resources pictured below are more classically descriptive, but lack a clear and concise description that is easily usable on a larger scale. That being said, due to the

fact that all things are inherently data, these resources are descriptive of many things, but their main focus is lost due to their formatting.

Driver charged with manslaughter after 8 killed outside Texas migrant shelter

Authorities are investigating to determine whether the deadly incident was intentional.

21H AGO

Neighbors says Allen outlet shooter Mauricio Garcia kept to himself

Neighbors would see Garcia come and go to work, dressed in a uniform that looked similar to a security guard's. But for the most part, they said he kept to himself.

MAY 7

https://www.cbsnews.com/losangeles/crime/

That being said, another addition to the usefulness of data and data collection is its inherent standardization. The organization of data in tables may initially be less descriptive than well strung sentences, but the true reason for such standardization is revealed quickly.

DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAM	Rpt Dist Nc	Part 1-2	Crm Cd	Crm Cd Des	Mocodes	Vict Age	Vict Sex
10304468	########	########	2230	3	Southwest	377	2	624	BATTERY -	0444 0913	36	F
1.9E+08	########	########	330	1	Central	163	2	624	BATTERY -	0416 1822	25	M
2E+08	########	########	1200	1	Central	155	2	845	SEX OFFEN	1501	0	X
1.92E+08	########	########	1730	15	N Hollywoc	1543	2	745	VANDALISM	0329 1402	76	F
1.92E+08	########	########	415	19	Mission	1998	2	740	VANDALISM	329	31	X
2E+08	########	########	30	1	Central	163	1	121	RAPE, FOR	0413 1822	25	F
2E+08	########	########	1315	1	Central	161	1	442	SHOPLIFTIN	1402 2004	23	M
2E+08	########	########	40	1	Central	155	2	946	OTHER MIS	1402 0392	0	X
2E+08	########	########	200	1	Central	101	1	341	THEFT-GRA	1822 0344	23	M
2.02E+08	########	########	1925	17	Devonshire	1708	1	341	THEFT-GRA	1300 0202	0	X
2E+08	########	########	2200	1	Central	192	1	330	BURGLARY	1822 1414	29	M
2E+08	########	########	955	1	Central	111	2	930	CRIMINAL	0421 0906	35	M
2E+08	########	########	1355	1	Central	162	1	341	THEFT-GRA	1822 0344	41	M
2E+08	########	########	1638	1	Central	162	1	648	ARSON	1402 1501	0	X
2E+08	########	########	1805	1	Central	128	1	442	SHOPLIFTIN	0325 1402	24	F
2.12E+08	########	########	730	19	Mission	1916	2	626	INTIMATE I	2000 1814	24	F
2.01E+08	########	########	2018	11	Northeast	1124	2	626	INTIMATE I	0400 0416	34	F
2.01E+08	########	########	1900	5	Harbor	511	1	440	THEFT PLAI	0319 0344	29	F
2.11E+08	########	########	1200	9	Van Nuys	932	2	354	THEFT OF I	1501 1822	46	M
2E+08	########	########	1330	1	Central	152	1	210	ROBBERY	0416 0411	66	M
2.01E+08	########	########	1735	9	Van Nuys	909	2	354	THEFT OF I	0377 1822	40	M
2E+08	########	########	1730	1	Central	162	1	341	THEFT-GRA	0344 1822	31	M
2E+08	########	########	1445	1	Central	162	1	442	SHOPLIFTIN	0325 1402	27	M
2.11E+08	########	########	1	10	West Valley	1045	2	354	THEFT OF I	1822 0930	46	F
2E+08	########	########	700	1	Central	166	1	230	ASSAULT W	0416 0913	62	M
2.11E+08	########	########	1200	10	West Valley	1043	2	354	THEFT OF I	1822	34	F
2.01E+08	########	########	1830	11	Northeast	1101	1	330	BURGLARY	0344 0352	43	M
2E+08	########	########	2000	1	Central	111	1	230	ASSAULT W	2004 0305	71	M

Pictured above are some crime statistics from Los Angeles from the year 2020 onwards. It's jarring at first, every word and description is minimalized and standardized. But when we manipulate this data we are able to uncover powerful things and utilize previously clunky data to make streamlined visualizations. Data in this context refers not to the method of collection or the information itself necessarily. All it means is a standardized set of information. By utilizing these standardizations then only are we able to efficiently create interpretations.

In the same vein, a data scientist is someone who can interpret this more abstract data in more easily representable formats, someone who can properly use the ingredients to make a full meal, so to speak. Though that being said, there are many things that contribute to naming someone as a data scientist. It could really be anything from creating advanced visualizations by using machine learning on thousands of sets of data to train a nigh perfect algorithm, to realizing and documenting the correlations between your target data and the data in question. These definitions are as malleable as the data itself and as such are subject to that degree of change between individuals. Strictly speaking, the only requirement in being a data scientist is intentionally working with data.

There's a few requirements for being able to work with data, though these aren't necessarily strict guidelines. Generally speaking, having an understanding of basic programming and the ability to use a computer are some hypothetical requirements but

in reality, due in part to the massive amount of accumulated knowledge that is derivable from the internet, there is no real barrier to entry.

That being said, although there may not be a concrete technical barrier with regards to working with data, the data scientist in question has to understand or at least acknowledge the aforementioned ideas surrounding data malleability. That data within a set is standardized but otherwise, data comes in all shapes and sizes. The transition point from observations to real concrete data is simply the process of observing and recording.

For example, in our Linear Motion project we did exactly that. We understood that our movement must have some data correlation, that there must be some way to represent such an abstract thing like movement into a more workable form. And we did!

	time	ax (m/s^2)	ay (m/s^2)	az (m/s^2)	aT (m/s^2)
0	0.002174	-0.0659	-0.0317	-0.0036	0.073
1	0.004631	-0.0229	-0.0074	-0.0730	0.077
2	0.013049	0.0263	-0.0378	0.0036	0.046
3	0.020658	0.0305	-0.0292	-0.0515	0.067
4	0.028815	0.0130	0.0322	-0.0755	0.083

```
height2.head()
```

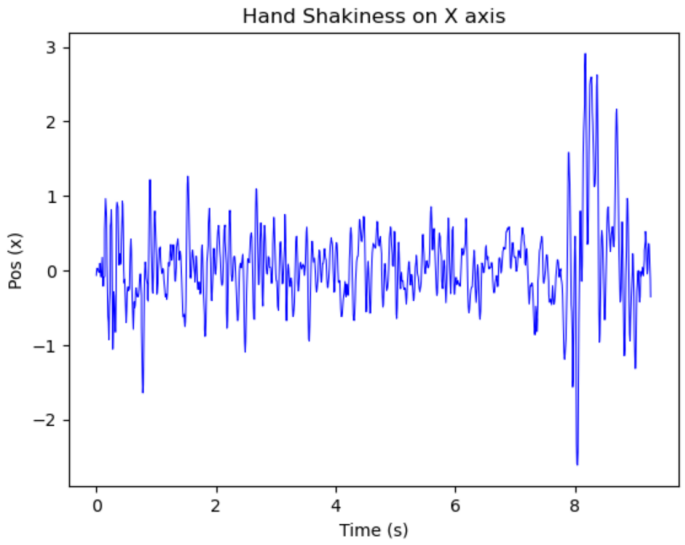
	Time (s)	Linear Acceleration x (m/s^2)	Linear Acceleration y (m/s^2)	Linear Acceleration z (m/s^2)	Absolute acceleration (m/s^2)
0	0.006469	-0.053707	-0.023127	-0.041767	0.071859
1	0.016526	-0.063557	-0.025577	-0.057485	0.089433
2	0.026584	-0.061033	-0.011483	-0.041920	0.074927
3	0.036641	-0.055774	-0.009183	-0.037879	0.068043
4	0.046698	-0.071056	-0.014492	-0.046113	0.085938

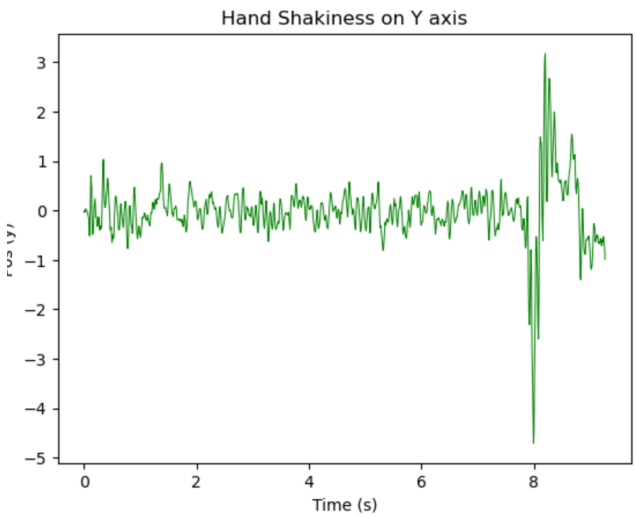
```
height3.head()
```

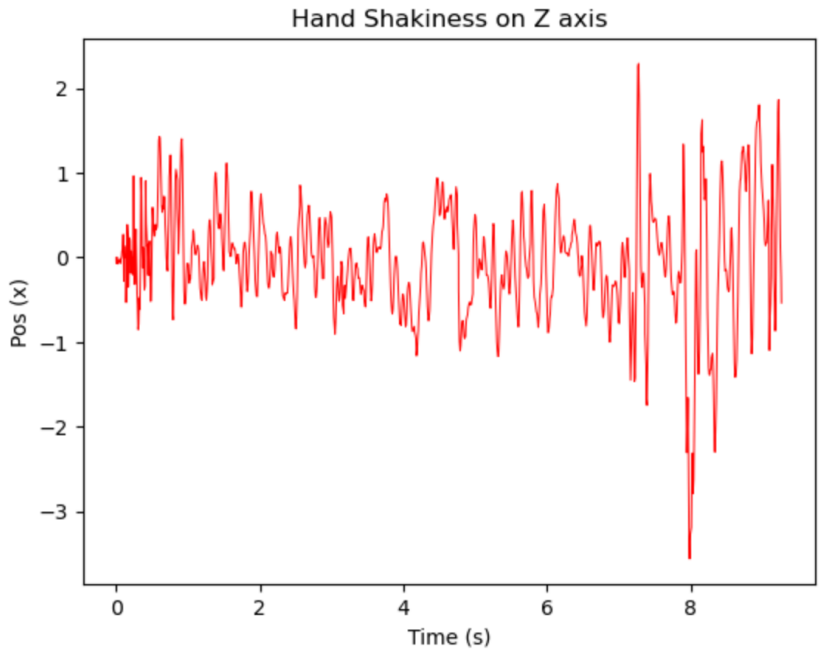
	Time (s)	Linear Acceleration x (m/s^2)	Linear Acceleration y (m/s^2)	Linear Acceleration z (m/s^2)	Absolute acceleration (m/s^2)
0	0.035940	-0.011031	0.025299	0.138819	0.141536
1	0.045947	-0.010959	0.025551	0.128186	0.131166
2	0.055954	-0.013522	0.026771	0.128182	0.131644
3	0.065961	-0.006221	0.025461	0.138060	0.140526
4	0.075968	-0.003480	0.025348	0.123538	0.126160

Then from this our previously (rather disconnected data) we were able to find out how shaky each of our hands were! Instances like this were the most eye opening part of this experience. These times when we were able to use data to find out something that seemed previously disconnected was definitely the most pivotal part.

Pictured below are a few visualizations describing how shaky our hands were in each direction on the X, Y and Z axis respectively.







Despite the encouragement that really anyone can be a data scientist, some may still fear that it's beyond them due to their lack of programming skills. If that is the case specifically then there are a multitude of free internet resources that cater to such a need for any desirable level.

Realistically there is an inherent knowledge barrier with regards to visualizing data electronically, if that is something that is most desirable, learning a degree of a programming language like Python is invaluable and has the potential to open up a huge amount of resources to be freely used.

With electronic visualization, learning the implementation of the Pandas and Matplotlib libraries is also recommended. They provide invaluable tools for quick, high quality data visualizations which would take an individual who is processing the same data manually significantly longer to complete.

For reference, the visualizations depicting hand shakiness above were all made with matplotlib through a few lines of code.

```
plt.plot(concatenated['time'], concatenated['ay (m/s^2)'], color='green',linewidth = .7, linestyle='-')
plt.xlabel('Time (s)')
plt.ylabel('Pos (y)')
plt.title('Hand Shakiness on Y axis')
```

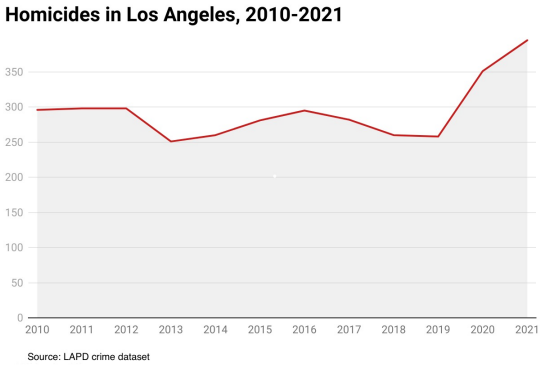
(Pictured above is the code to represent 'Y axis hand shakiness')

The interesting thing about data and data visualization is that you can really solve any question by using it. Even in some cases that require much more descriptive answers than a single output from a program.

Anything from “What is the current crime rate of Los Angeles by county?” to “what flavors of icecream are universally popular?” can be solved effectively by data. Because of this I believe that it is more appropriate to ask which problems can’t be solved by data science and visualization. To which I say: “anything that doesn’t have enough data collected to make an argument.”

Data can solve literally any question, the only barrier is that of whether the data has already been collected. Then, if it hasn’t been collected, find a way to collect it!

The principles that define my own data analysis process which are accentuated due to the properties of data malleability are as follows: data is formless, data is a word representative of an infinite number of observable points, the idea of forming your question around your data and finally, visualization is key.



https://xtown.la/2022/01/26/crime-rate-los-angeles2021/

In order to complete a problem regardless of its topic I use these principles. Starting with analyzing the data manually and searching for a question that it can be asked. The next two parts of my process are more internal, understanding that this data may need to be supplemented with other relevant data to reach a conclusion that makes sense, expected or otherwise. Finally, visualization. Which is in my opinion one of the most important parts of the process because it provides an easily accessible understanding of the data for those with less technical knowledge. My own personal process differs greatly from others. While data in its interpretation is infinite, data analysis itself has no standardizations, therefore the analysis is as unique as the person performing it. Data's nature allows for complex problems to be solved in completely unique ways and is more simple to work with than one may initially think!

That all being said, one of the most important parts of data science and analysis is that central theme of data malleability. That data can be stretched and twisted into unique forms. The process allows you to mold the perfect key for an impossible lock, so to speak.