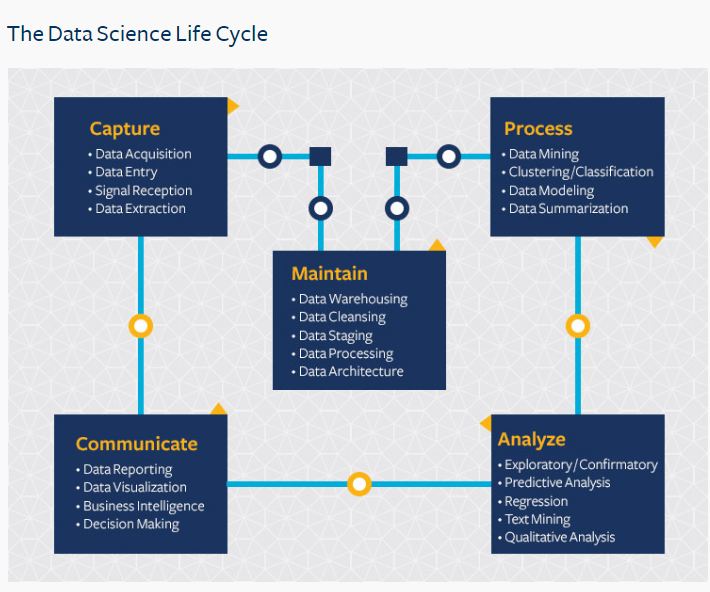


Reference : https://en.wikipedia.org/wiki/Exploratory\_data\_analysis



Reference :<https://ischoolonline.berkeley.edu/data-science/what-is-data-science/>

This shows the Typical Data Science Pathway.

Here we have 5 boxes, along which our DATA flows.

Let us take a step back to determine what Data is, and what Data isn’t.

<https://en.wikipedia.org/wiki/Data>

Data is a collection of qualitative and quantitative observations about one or more things.

Qualitative data refers to data that *describes* an object. This type of data is not necessarily quantifiable, I.E. Able to be reduced to numbers.

An example of qualitative data is eye color. We could say Brown, Blue, Black, Green, etcetera.

This is in contrast to Quantifiable Data. Data that *can* be reduced to numbers. Some examples include, height, weight, age.

A person could argue that Eye color can be measured quantitatively. We can use the RGB value of it and assign a numerical value to said color, and so we can quantitatively describe someone’s eye color. This is true. Why and how we choose to represent data is an important question, but it’s a question that we will think about later for now.

Where does the Data Science Life Cycle Begin?

It begins with us. I mean that not in the personal sense, but rather in a general us. Humans are obsessed, for better or worse, with the idea of quantifying and qualifying data. We like observing things and then trying to extrapolate from our observations. The first step begins with questioning.

It’s important to realize that we are not mathematicians, we are not computer scientists, we are not statisticians. We are a beautiful amalgamation of the three. Above all though, we are SCIENTISTS.

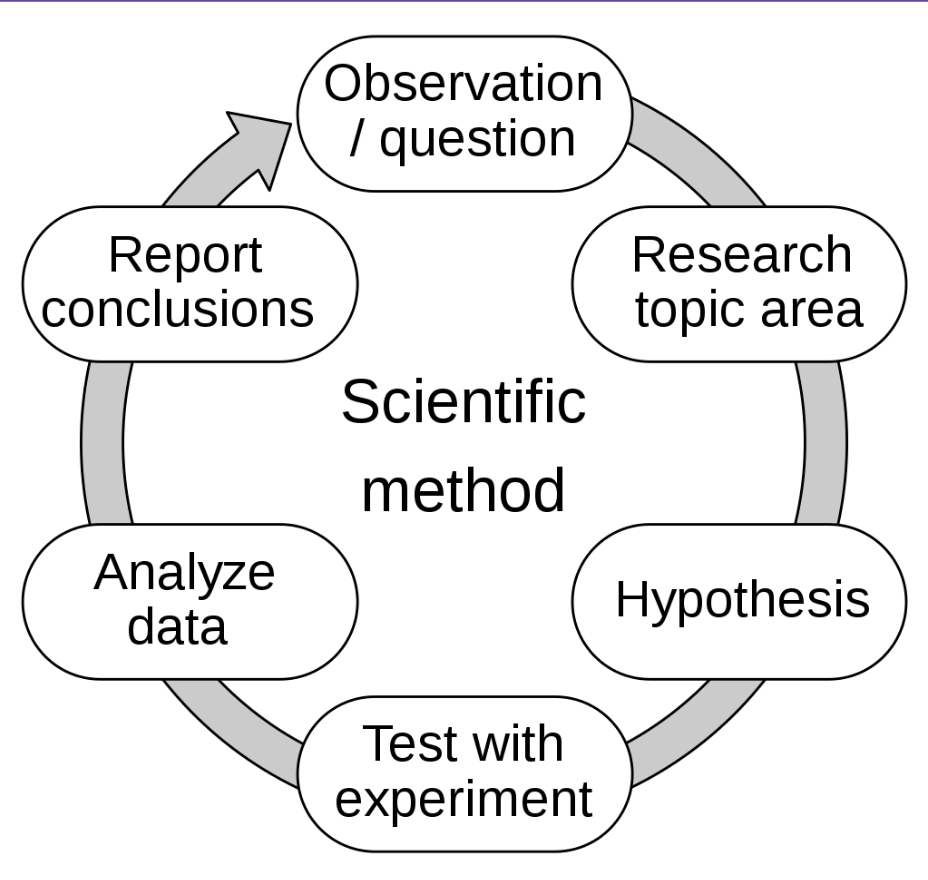
So what does it mean to BE a Scientist?

<https://en.wikipedia.org/wiki/Scientist>

A scientist is a person who conducts science (duh) on specific topics. Our specific topic is DATA. We are Scientists, for DATA. Before we unpack that loaded statement, let us talk openly about what it means to be a scientist. A scientist is a person who conducts scientific research, but what does that mean exactly?

<https://en.wikipedia.org/wiki/Scientific_method>

Being a scientist means that we use this. The Scientific Method. We apply this thought/idea pattern to objects in our lives, and then extrapolate from the data we collected.



This image goes through the “typical” Scientific Method. There is no definitive guide/answer to how to start and end science. There are general rules or guidelines however.

I will go through it using the starting point of Observation/Question. This is not the only starting point. Oftentimes, we collect data, analyze it, form an observation, and then test that observation. Don’t think of these as strict guidelines.

Observation/Question : This is the point where we as humans observe. We see things in the world. We try to make sense of things. We start most times with a question. Why? Why does this happen? Why does this person do this? How does this happen? This is the beauty of science. It’s inquiry, it’s understanding, it’s growth, it’s trying to explain the world as it happens.

Research Topic Area: We have our question. Now it is time to get down to work. We have to try to approach understanding. Are we asking a valid question? Are we asking something that hasn’t been answered? Are we trying to disprove something? Are we trying to affirm something? It’s important to gain some domain knowledge as to the question that we are asking. Do research, and have that research be honest.

Hypothesis:

https://en.wikipedia.org/wiki/Hypothesis

Now we are armed with our question, and with domain knowledge. Now it’s time to formulate a hypothesis. We are trying to explain some phenomenon that we have observed.

Let’s work through an example. Assume none of this has been proven yet.

I observe that it rains.

I cannot explain why it rains.

I observe that when it rains, there are clouds.

My hypothesis is that clouds cause the rain.

This is a good starting point, but I’m not finished. One might ask, why is that? It’s because I have only theorized at this point. I have to produce results, I have to produce quantifiable results and data that PROVES that the clouds cause the rain. I have to explain how I plan to test my hypothesis. My test has to produce accurate results and more importantly replicable results. What does that mean? That means that if someone was to mimic my tests, that they too would achieve the same results. I have to design a series of tests. An experiment.

I need more research, and I need to design a series of tests.

Clouds are water and ice particles suspended in the atmosphere. Evaporation brings water vapor up into the atmosphere, and they cool once they get high enough. The water is supercooled meaning that it remains in liquid form, even while being below the freezing point of water. The water and ice particles are too small to form rain or snow or hail, and instead float in the air. They bump into each other and sometimes nothing happens. They sometimes join together to make larger suspended particles in the atmosphere. This is called coalescence. Once enough particles join together, usually as a result of turbulence ( wind pushing and forcing the particles to collide) they form droplets large enough to break air resistance, and fall back to earth as droplets.

With this new domain knowledge, I can design an experiment to test that this is indeed the hypothesis that rain is generated by suspended particles of water in the air, cooling and condensing.

I can heat water to create vapor. Then, I need to cool the vapor. Then I need to give the cooled vapor a condensation point. ( in the case of the atmosphere dust, and debris and smoke that are naturally present serve as condensation points). Once this is true, I should see the vapor condense, and fall back into the earth because of gravity.

I design this test and then implement my tests.

<https://www.youtube.com/watch?v=x-r8jIBaIoo>

A test designed and implemented by another scientist ( A meteorologist!)

Analyze Data:

I write down my observations of the experiment. I see that this is indeed the case. Vapor from the hot water rises, it condenses on the top of the jar, and then it falls back into the jar. This proves that my hypothesis was correct.

Report Conclusions:

I can now share my results with other scientists, and share my methods of collecting data and of analyzing data as well. Good science is NOT done in a vacuum. Other scientists have to be able to replicate the results that either prove or disprove you. A scientist might ask the question. This is true in a jar, but how do we know this is true in the atmosphere? Can we prove that clouds are water? Now, we have to answer this question. We can start the process all over again.

Science is not a one and done thing. It is not solo work. It is collaborative and honest work. It is work that is done with precision and with care, to ensure that our data is accurate. We need clean, accurate data to analyze, and to report, and we need to do it in ways that are replicable.

We’ve now taken a better look at what it means to be a scientist, and how to “do science” so to speak. Now we have to cover the implications of being a scientist. There are moral and ethical obligations to being a scientist, and this is a conversation that must be had.

Class discussion time —- What can we think of as moral and ethical obligations —--

We as scientists have a duty to society. In calling ourselves scientists, (which after today you all should) you assert the idea that you have domain knowledge. That you are an expert in your field of study. Ours is going to be Data(huh?). More on that later.

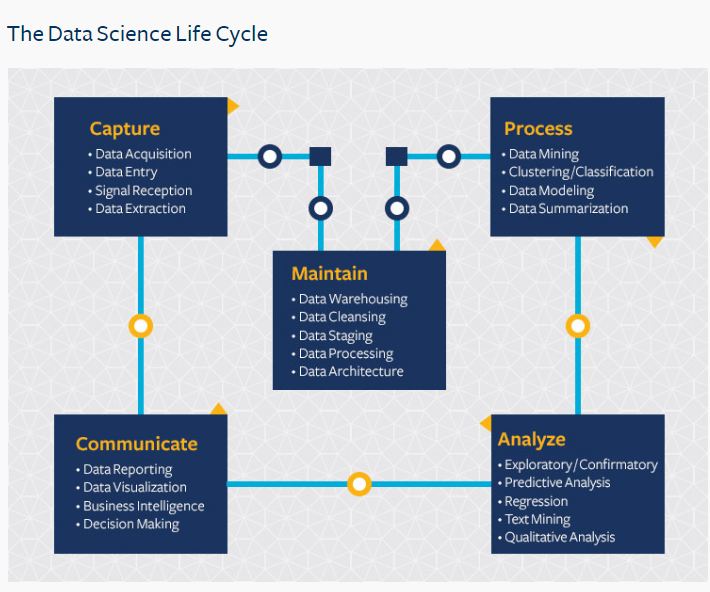
This means that we have a duty to accurately and truthfully hypothesize and synthesize. What we say and what we do is going to impact people. Decisions are going to be made by us, and by the analysis of our data. We are directly or indirectly impacting people’s lives, and people’s thought patterns. That has to be at the forefront of your mind and thinking. It is imperative that you remember that obligation. You present yourself as a scientist, and so it is your duty as one, to report things TRUTHFULLY and HONESTLY. Science has been used for Millennia to justify awful and despicable crimes. Racism, Sexism, Genocide has often used science to justify its existence and it perpuation. I will not teach people who do bad science. Willingly or unwillingly. There is weight now to what you say and do. I will not have you using Science as a method to spread misinformation, to spread dishonesty, and to justify crime. I am talking frankly because I am serious. Think very carefully about your science, and every aspect about it. Think about your own intrinsic bias, think about the bias of others. Think about the data you are collecting, and the data you aren’t collecting. Use reason, and logic to dictate your data collection and your analysis to the best of your ability. Question yourself, challenge yourself, do not take things for granted or for face value. Test everything you can, question and challenge everything you can, and ensure that what you analyze and report is done so to the best of your ability. Exhaust every single possible explanation and reasoning. Ensure that you are being as honest and as thorough as possible.

You have a moral and ethical obligation to say and do the right thing. Morality and ethics are personal for sure, but make sure you are challenging yourself. Your own morality and ethics. I will GLADLY be available to talk to you about these things. I will reiterate. I DO NOT want bad science being performed.

A baseline set of morals

1. Tell the truth
2. Keep your promises
3. Don’t Steal
4. Don’t Murder
5. Discourage Wrongdoing

Some of these may seem like they are out of your realm. This is not true. Data can be used to justify some of these heinous behaviors. Ensure that your data is not. Ensure that your analysis and interpretation of data isn’t breaking your ethical and moral code. Ensure that whoever you work for is not using your data you break your ethical and moral code. Be true to yourself, and what is right.



Now that we are clear about what being a Scientist is, and what our duties as a scientist are. Let’s see what Data Scientists in particular care about.

**Data science** is an [interdisciplinary](https://en.wikipedia.org/wiki/Interdisciplinarity) field that uses scientific methods, processes, algorithms and systems to extract [knowledge](https://en.wikipedia.org/wiki/Knowledge) and insights from noisy, structured and [unstructured data](https://en.wikipedia.org/wiki/Unstructured_data),[[1]](https://en.wikipedia.org/wiki/Data_science#cite_note-1)[[2]](https://en.wikipedia.org/wiki/Data_science#cite_note-2) and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to [data mining](https://en.wikipedia.org/wiki/Data_mining), [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and [big data](https://en.wikipedia.org/wiki/Big_data).

We want to capture data, we want to analyze data, and we use that analysis to inform decisions our companies make (sometimes we work with AI, and we use the analysis to inform our AI how to act).

We start with Data Capture. What are we getting data from? How are we getting that Data? What type of data are we getting? Why are we getting that data?

These are all the questions that you will have to answer. Justify your reasoning, and justify your choices. Here we use the scientific method. We use reasoning and logic as well.

Then we move onto data maintenance. How are we going to store that data? Why are we storing it like that? What part of our data is usable? What parts of our dataset are we going to need for analysis? Why?

This is one of the more technical parts of the job. Maintaining and building databases. Making methods that allow people to access the database and easily store and retrieve data from it. Cleaning up your data, and breaking up your data into more manageable chunks. Here we use a ton of technologies.

Python, R, SQL, NoSQL, PostGres, Flask, Django.

The possibilities are endless. We are going to focus on the Python Pipeline.

Next step is Data Processing. Taking out the data sets we need. Setting up Data into easy to read formats. Finding patterns in large data sets. Using our data to start developing groups and batches of specific (or unspecific data.) This is getting down and dirty with our data so that we can produce results.

Tools here are Pandas, Python, R.

Tools also include Math. We use math to start to process our data. We take our data and get data layered on top of that. Means, Medians, Standard Deviations.

Then we get to Data Analysis. This is one of the higher level operations in our pipeline. This is the question asking and or answering. What does our data show? Does our data prove or disprove our hypothesis? How can we prove that it proves it? What can we do with this data? Can we use this data to train algorithms?

Tools here include. Python, R, Pandas, TensorFlow, PyTorch.

We also employ: the Scientific Method, Math(Regression Analysis, T Testing)

“Lastly” we get to Communication. Here is where we put our analysis into words and images. We explain to people what our data looked like, what we were able to deduce, and how we were able to deduce it.

Tools include: Pandas, Python, R, Seaborn, MatplotLib, Scientific Method, Reasoning, Logic.

I say “lastly” because this can be done in any order. There is no order to these things. Don’t allow yourself to become limited in your thinking by assuming that this process is set in stone and always the same.