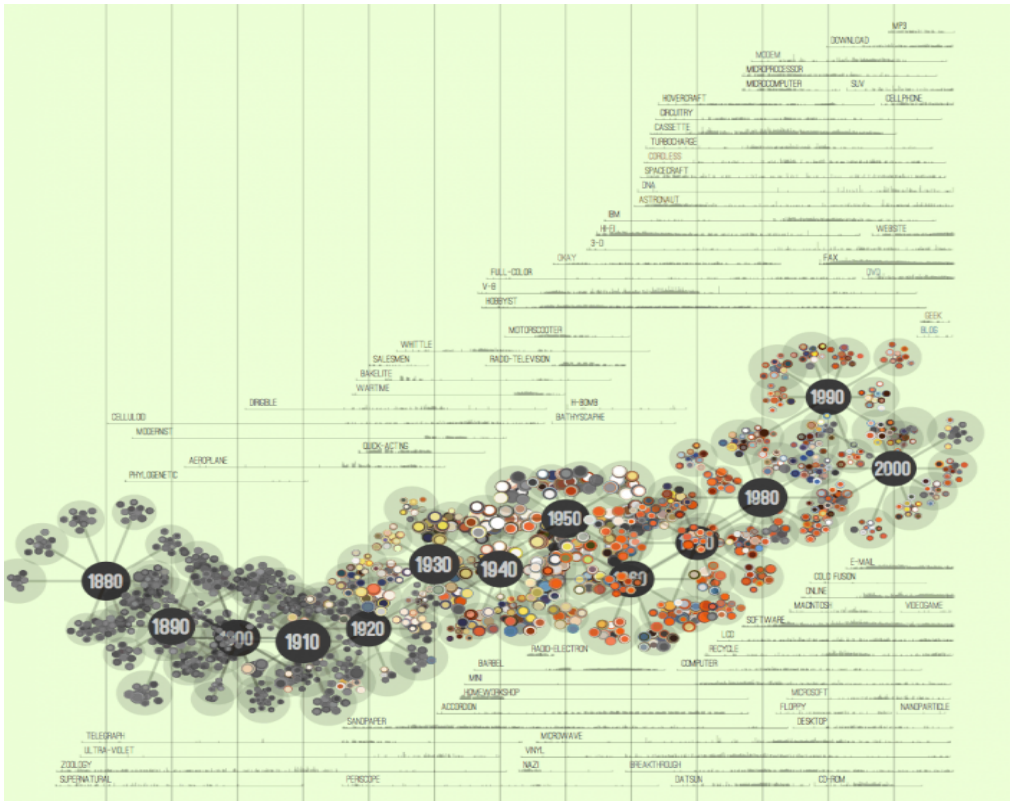
Miles Wilderman

Professor Wirfs-Brock

Introduction to Data Science

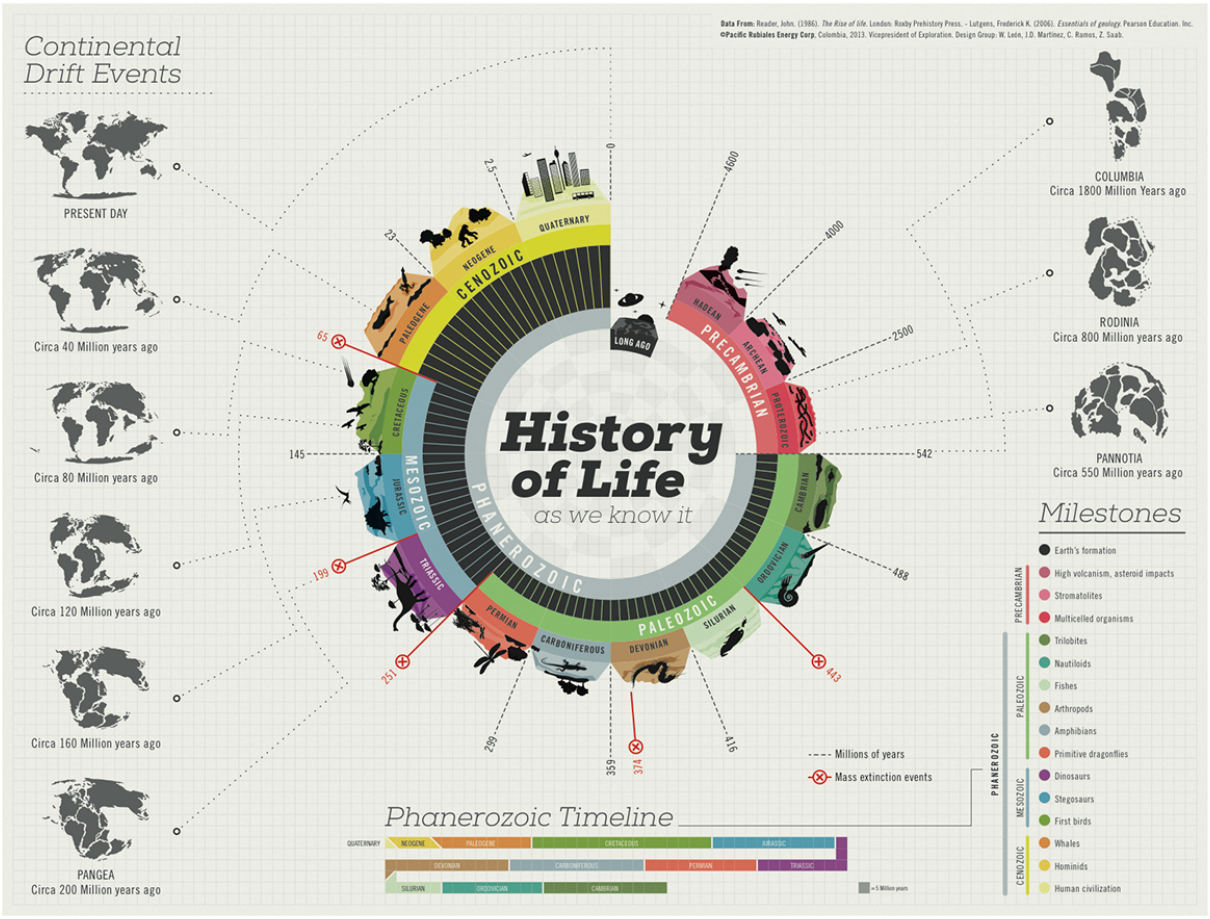
9 May, 2023

Data Manifesto to History



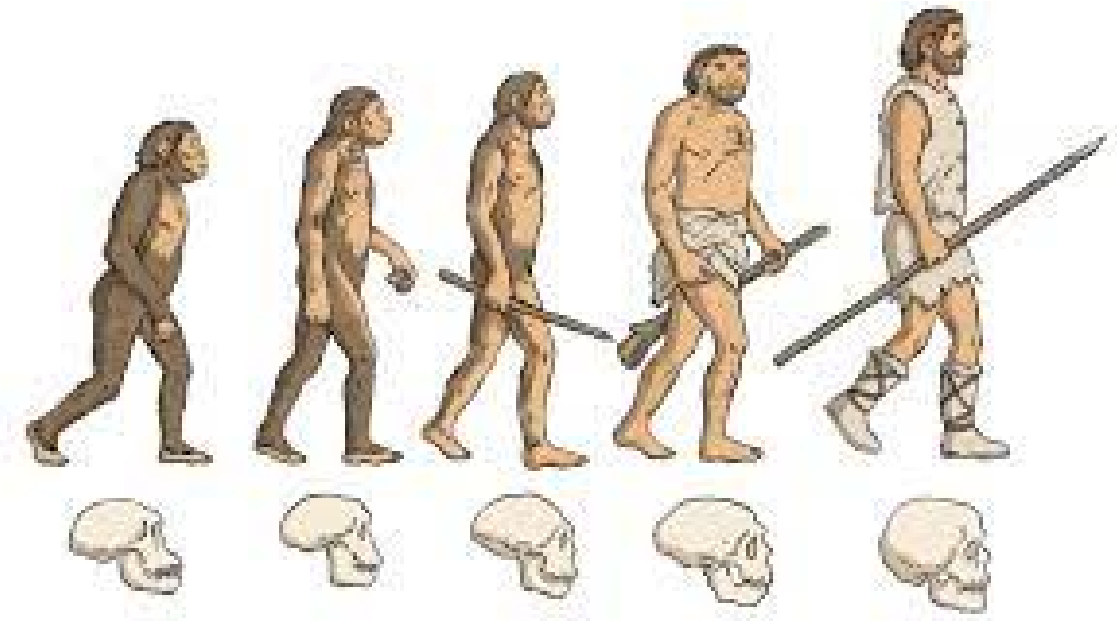
I think of data as a history record. Data does not have to be in a time format for it to be historical. What we then chose to do with this history, is up to us. Referring to the Data-Information-Knowledge-Wisdom pyramid by Russell Ackoff, I like to think of history as this bottom layer, and everything that we can gather above this is learning from this history. Then there is what we can do with this learning of information. Visualizations? Predicting new history with machine learning models? It is up to us to communicate what we gather from history as

well as we can.



Learning biology at the same time as data science and machine learning has been very interesting for me and made me think about a lot. Learning biology is kind of like getting a lot of data plugged into your brain and tested on how well you can remember it, which annoyed me a little. However it also opened my eyes up to some bigger questions. As I learned the “history” of

life on earth, through animals especially, I started analyzing evolution.



In such a short time period, compared to everything else in the world, apes evolved into humans. With thoughts of data science and machine learning in my head, this naturally made me think about evolution in the future. Could we predict future species that will be here in a million years? Will humans evolve further separating into two groups? Will humans go extinct as the Nianderthals did? I still don't know the answers to these questions but they definitely framed how I think about data and what I hope it can be used for.

To answer questions like this and what I have found learning even just intro data science and management, is that we must be organized. Organization was not my strong suit on computers before learning data science and taking some computer science courses. I would leave

things downloaded wherever never creating folders for anything.

I have also learned from biology that data must be organized very precisely for large powerful questions to be answered. I believe that data can be very powerful, but only if the proper care is taken throughout every step. Biology shows us how much diversity there is in the world, how much everything is constantly changing, and how the entropy of our universe will keep this continuing. In order for data to be able to overcome these problems, we need to continue improving our data management.

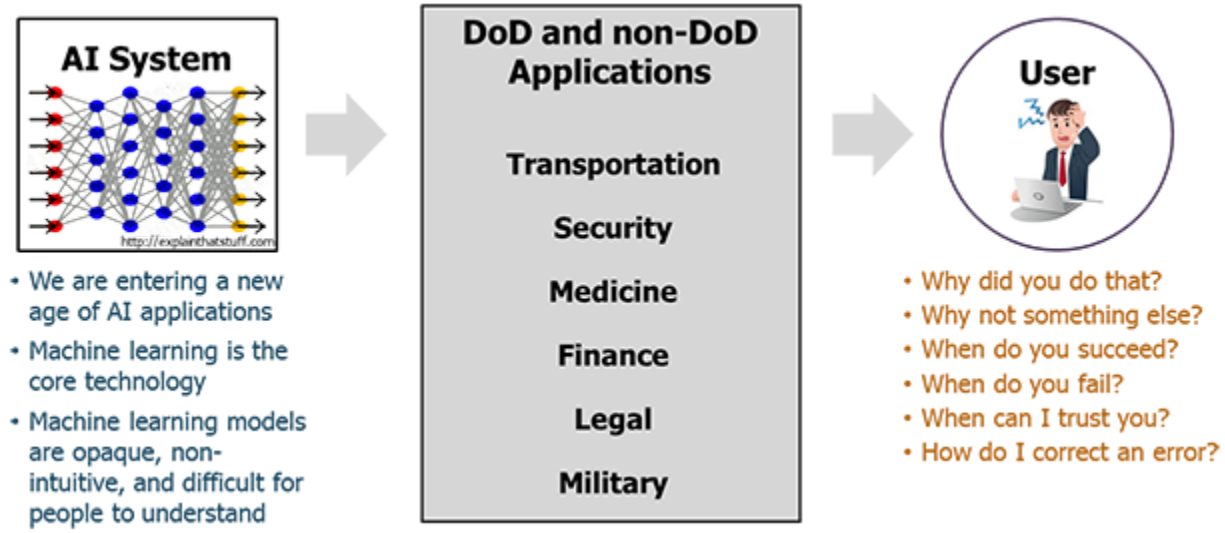
Some easy things that I have learned to do is firstly be very organized between different projects that I have that involve data. This includes the “No copying easy code” line that got drilled into my brain in intro CS. We also should scroll through multiple rows of our data to see

the raw stuff that we are working with.

Out[9]:

	Day	Max_Temp	Min_Temp	Avg_Temp	Precipitation
0	2022-09-01	96	69	82.5	0
1	2022-09-02	103	62	82.5	0
2	2022-09-03	86	66	76.0	T
3	2022-09-04	88	57	72.5	0
4	2022-09-05	84	59	71.5	0
5	2022-09-06	86	56	71.0	0
6	2022-09-07	89	62	75.5	0
7	2022-09-08	77	55	66.0	0
8	2022-09-09	81	51	66.0	0

Doing simple things like looking through my dataframe by hand really helps my awareness of a problem or situation. In the context of history as data we must look through events leading up to the larger events, because without causation we have no context. We can not begin to work with data unless we have fully explored the raw un hindered data.



When I mentioned predicting the future, I did not mention how scary this is. It is powerful but scary. We need to be able to explain what large future predicting models are doing.

Explainable AI is very important for these predictions to be accurate as well, because how can we even know how accurate it is if we can't explain it. For things like biology especially, we need to be able to explain why our model is saying that humans will regrow their tails again in a million years. Even for smaller problems we need to be able to explain our code and outputs better than most currently can in today's world. To do this we need to reference the history that we are basing our code off. What in our history of data shows us the reasoning for our code and our outputs?