The main base for developing an ML model is the dataset that it works on. The entire performance of the model depends upon the dataset only. If the dataset is improper, the model will not work properly. So, the dataset that the model works upon shuld be developed and curated with carefulness.

The objections around data raise a serious challenge to the justifiability of the datasets used by many ML applications. How can AI systems be trusted if they are built on the basis of false datasets?

If ML is to survive this crisis in a responsible fashion, it must adopt visibility practices that enable accountability and responsibility throughout the data development lifecycle. The adaptions can be taken from the traditional Software Engineering practices like the Software Development Life cycle.

This paper explores in detail a number of the practices that need to be adopted to mitigate dataset risks. These practices in turn require adopting a deliberative and intentional methodology, rather than the justifications that are sometimes observed when datasets are developed hurriedly without proper research.

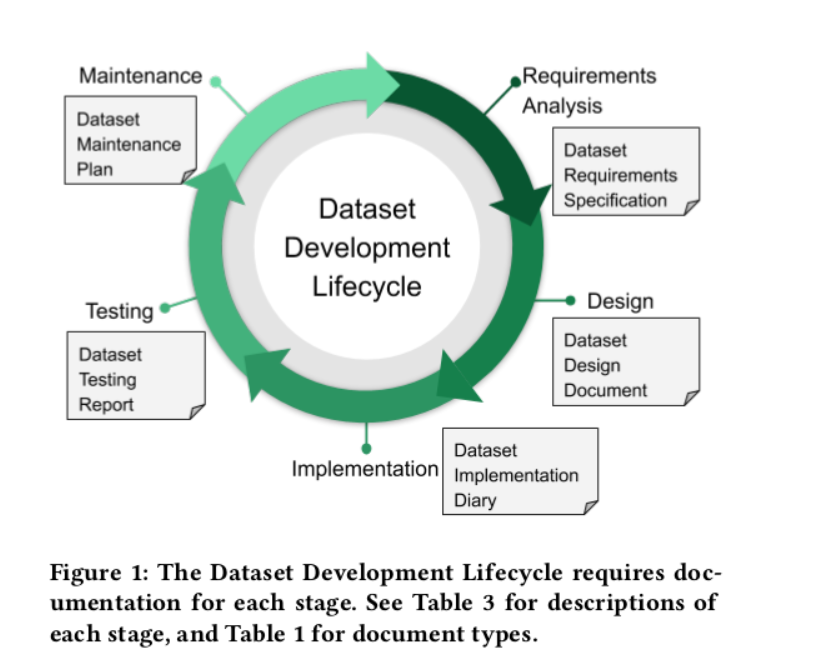
**Non-linear cycle of Dataset development:**

• **Documentation**: a model of documentation practices throughout the dataset development lifecycle, drawing on software lifecycle practices. The documentation helps for the next daraset developers about the problems faced and rules to be followed while gathering the data.

• **Oversight**:

• **Maintenance**

**Dataset Development Lifecycle:**

****

The Dataset development model is insipred from the Software Development Lifecycle model to develop the datasets more reliable and error free.

SDLC promises that the software developed provides accountability to the client. Similarly, the dataset developed by using the DDLC model will also be accountable to the ML model developers.

The main reason behind developing DDLC model is, the development teams involve in lots of work while developing the model by referring the dataset. Atlast, if they found that dataset has lots of errors and this dataset doesn’t suit for the model development, all the work done and money invested will go for a toss. So, to avoid all the uncertainities during the development, we need to ensure that the dataset is perfect and ready for the work.

**Real life Challenges:**

There might be politics invloved in dataset development. Suppose, if a company is developing dataset, then it will not include the data which causes harm for its reputation. So, it requires **accountability** for enabling objectives, even if those objectives are not one’s own.

For example, amazon removes the bad reviews for the products in their datasets. So, while working on the model for recommendation of products, the model may predict that the product is very good. Blindly floowing the model, will result in loss of money for the customer.

**References:**

*Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure*

*Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson,*

*Parker Barnes, Margaret Mitchell*

.