Document 4 :

https://epubs.siam.org/doi/epdf/10.1137/1.9781611977653.ch106

EN ANGLAIS :

The way data is used for AI has recently changed. There's now a greater focus on data quality than on creating new models. This is known as the DCAI approach. Instead of just improving models, the focus is on using high-quality data.

In the past, the focus was on improving models, with less concern about the data used. But now, there's recognition that good data quality is essential for reliable AI results. DCAI encourages us to rethink how we use data and ensure its reliable and representative.

DCAI covers different techniques for developing, evaluating, and maintaining the data used in AI systems. This includes creating training data, evaluating data quality, and ongoing maintenance to ensure it stays relevant. So, the community is working on ways to improve these different aspects of data.

To advance DCAI, it's important to recognize the significance of high quality data in building effective AI systems. By focusing on data quality, we can unlock AI's potential to solve complex problems in many fields.

Having high quality data involves several steps: data collection, adding informative labels, preparing the data to make it suitable for learning, reducing data size to make it simpler and more understandable, and increasing data diversity to enhance model performance. Although we have made great strides in processing data, there are still challenges to overcome to ensure the quality of data used in AI models.

Before deploying an AI model, it's essential to test it to see if it performs well. For this, we use special data called evaluation data. This data allows us to check if the model does a good job. There are two types of evaluation data: those that look like the data used for training the model, and those that are different. For the former, we can create new test sets by merging data like that used for training. For the latter, we can use techniques to test the model's ability to recognize correct information even when it's slightly altered. Research in this area is ongoing and likely to expand further in the future.

An important way to simplify complex data is to reduce it to two dimensions for better understanding. We can also evaluate each piece of data to see which is most useful. This helps maintain data quality by regularly checking it to avoid errors. To improve data, we can sort or correct it, either manually or with automatic tools. Speeding up data acquisition is essential for working faster. This can be done by better organizing resources. Data maintenance is important to support the ongoing creation of training and evaluation data. In the future, they will likely make this maintenance more dependent on how we construct the data.

And finally, the challenges to overcome in the field of DCAI (Data Centric AI). It highlights the importance of carefully evaluating AI models before deploying them to ensure their reliability. Moreover, maintaining data quality throughout the process is crucial to ensure optimal model performance. The article also shows the importance of understanding how different DCAI tasks interact with each other, which can be complex. An integrated approach to designing data pipelines and AI models is recommended to maximize the efficiency of AI systems. It's crucial to take steps to reduce possible unfairness in the data, as this can impact how well the model works. Lastly, the paragraph points out the importance of creating solid standards to measure progress in the DCAI field. This will need cooperation between businesses and researchers, and a lot of research.