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.

1er paragraphe : La façon dont on utilise les données pour l'intelligence artificielle (IA) a changé récemment. On se concentre désormais plus sur la qualité des données que sur la création de nouveaux modèles. C'est ce qu'on appelle l'approche centrée sur les données pour l'IA, ou DCAI. Plutôt que de se focaliser uniquement sur l'amélioration des modèles, on met l'accent sur la qualité et la fiabilité des données utilisées.

Dans le passé, on se concentrait principalement sur l'amélioration des modèles sans trop se soucier des données utilisées. Mais maintenant, on réalise que des données de bonne qualité sont essentielles pour obtenir des résultats fiables avec l'IA. La DCAI nous

encourage à repenser la manière dont nous utilisons les données et à nous assurer qu'elles sont fiables et représentatives.

La DCAI englobe différentes techniques pour développer, évaluer et maintenir les données utilisées dans les systèmes d'IA. Cela inclut la création de données d'entraînement, l'évaluation de la qualité des données et la maintenance continue pour s'assurer qu'elles restent pertinentes. La communauté s'efforce donc de trouver des moyens pour améliorer ces différents aspects des données.

Pour faire progresser la DCAI, il est important de reconnaître l'importance des données de qualité dans la construction de systèmes d'IA performants. En mettant l'accent sur la qualité des données, nous pouvons réaliser le potentiel de l'IA pour résoudre des problèmes complexes dans de nombreux domaines.

2eme paragraphe :

Pour créer des systèmes d'intelligence artificielle efficaces, il est essentiel de disposer de données d'entraînement de qualité. Cela implique plusieurs étapes : la collecte de données, l'ajout d'étiquettes informatives, la préparation des données pour les rendre adaptées à l'apprentissage, la réduction de la taille des données pour les rendre plus simples et plus compréhensibles, et enfin l'augmentation de la diversité des données pour améliorer la performance du modèle. Bien que nous ayons investi beaucoup d'efforts dans le traitement des données, il reste des défis à relever pour garantir la qualité des données utilisées dans les modèles d'intelligence artificielle.

3eme :

Avant de mettre en service un modèle d'intelligence artificielle (IA), il est important de le tester pour savoir s'il fonctionne bien. Pour cela, on utilise des données spéciales appelées données d'évaluation. Ces données permettent de vérifier si le modèle fait un bon travail. Il y a deux types de données d'évaluation : celles qui ressemblent aux données utilisées pour entraîner le modèle, et celles qui sont différentes. Pour les premières, on peut créer de nouveaux ensembles de test en fusionnant des données similaires à celles d'entraînement. Pour les deuxièmes, on peut utiliser des techniques comme la perturbation adversaire pour tester la capacité du modèle à reconnaître les informations correctes même lorsqu'elles sont légèrement modifiées. La recherche dans ce domaine est en cours et va probablement se développer davantage à l'avenir.

4eme :

Une façon importante de simplifier les données complexes est de les réduire à deux dimensions pour mieux les comprendre. On peut aussi évaluer chaque donnée pour savoir laquelle est la plus utile. Cela aide à garder les données de qualité en les vérifiant régulièrement pour éviter les erreurs. Pour améliorer les données, on peut les trier ou les corriger, soit manuellement soit avec des outils automatiques. Accélérer l'obtention des données est essentiel pour travailler plus rapidement. Cela peut se faire en organisant mieux les ressources et en rendant les requêtes plus rapides. La

maintenance des données est importante pour soutenir la création constante des données d'apprentissage et d'évaluation. Dans l'avenir, on va probablement rendre cette maintenance plus dépendante de la façon dont on construit les données.

5eme et dernier:

Le paragraphe aborde les défis à surmonter dans le domaine de l'intelligence artificielle centrée sur les données (DCAI). Il met en lumière l'importance d'évaluer soigneusement les modèles d'IA avant de les déployer pour garantir leur fiabilité. De plus, maintenir la qualité des données tout au long du processus est crucial pour assurer des performances optimales des modèles. Le paragraphe souligne également l'importance de comprendre comment différentes tâches de DCAI interagissent entre elles, ce qui peut être complexe. Une approche intégrée de la conception des pipelines de données et des modèles d'IA est recommandée pour maximiser l'efficacité des systèmes d'IA. De plus, il est essentiel de prendre des mesures proactives pour atténuer les biais potentiels dans les données, car cela peut affecter les performances et l'équité des modèles. Enfin, le paragraphe met en évidence le besoin de développer des références et des benchmarks solides pour évaluer les progrès réalisés dans le domaine de la DCAI, ce qui nécessitera des efforts de recherche concertés et une collaboration entre les acteurs de l'industrie et de la recherche.