Data-centric AI: Perspectives and Challenges

Daochen Zha†, Zaid Pervaiz Bhat‡, Kwei-Herng Lai†, Fan Yang†, Xia Hu†

†*Rice University*
‡*Texas A&M University*

What is Data-centric AI (DCAI)?

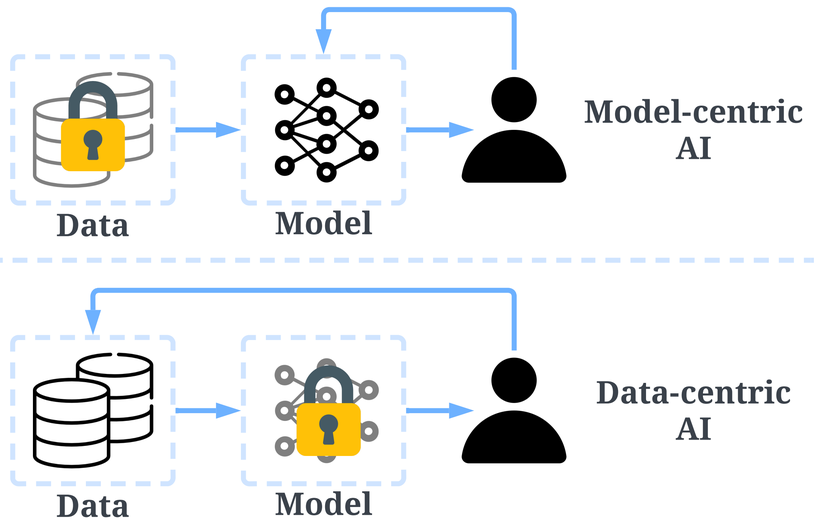

Figure: Model-centric AI versus Data-centric AI.

- **Andrew's definition:** DCAI is the discipline of systematically engineering the data used to build an AI system.
- **Our definition:** DCAI refers to a class of systematic techniques that develop, iterate, and maintain data for AI systems. DCAI involves three general goals: *training data development* (e.g., data collection, data labeling, data augmentation, etc.), *inference data development* (e.g., evaluation data, prompt engineering, etc.) and *data maintenance* (data valuation, data quality assessment, etc.).

Why DCAI?

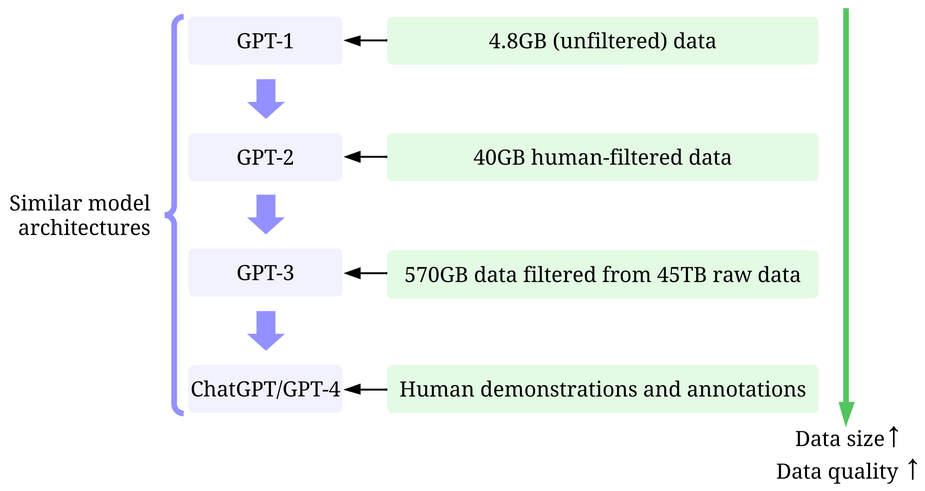

Figure: Motivating example 1: Large and high-quality training data are the driving force of recent successes of GPT models, while model architectures remain similar, except for more model weights.

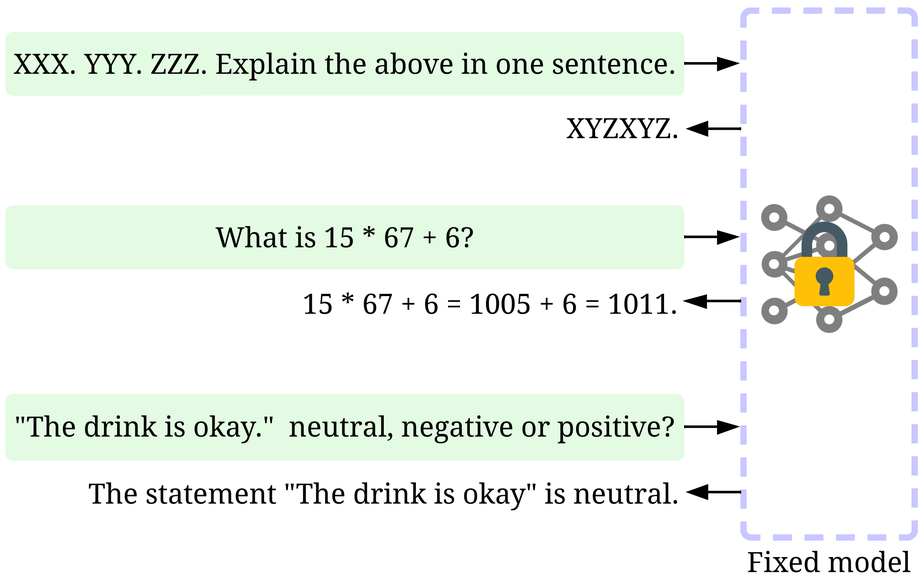

Figure: Motivating example 2: When the model becomes sufficiently powerful, we only need to engineer prompts (inference data) to accomplish our objectives, with the model being fixed.

Why DCAI (cont.)?

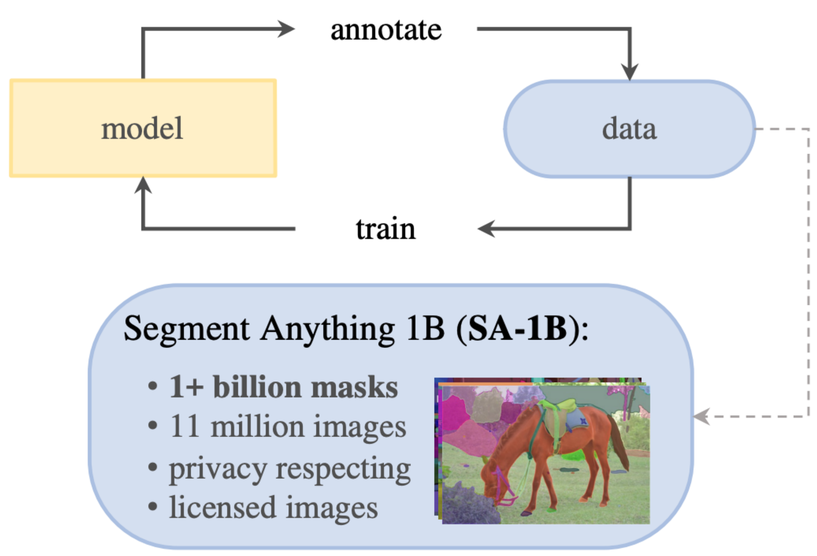

Figure: (Figure from the Segment Anything paper) Motivating example 3: The success of Segment Anything is largely attributed to a large annotated dataset with over 1 billion masks, which is 400x larger than the existing dataset.

FAQ

- **Is DCAI the same as "data-driven" AI?** "Data-centric" differs fundamentally from "data-driven", as the latter only emphasizes the use of data to guide AI development, which typically still centers on developing models rather than engineering data.
- **Is DCAI new?** While DCAI is a new concept, many relevant research topics are not new, such as feature selection and data augmentation. But some new tasks are emerging, such as data programming and data valuation. We need a top-level view of DCAI to motivate the collective initiative to push DCAI forward.

Outlooks and Challenges

- **Evaluation data & data maintenance:** Compared to training data developments, these two are less explored. They are important and require more research efforts.
- **Cross-task techniques:** How can we simultaneously optimize multiple tasks in DCAI?
- **Data-model co-design:** Can we achieve better performance by engineering the data and the model iteratively?
- **Data bias:** How can we remove the bias in data?
- **Data benchmark:** How can we develop *data* benchmark to benchmark data quality? Existing benchmarks only focuses on specific DCAI tasks but not the DCAI as a whole.

Perspective Paper | Survey Paper | GitHub Resources