Data-centric AI: Perspective and Challenges

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

Rice University
Texas A&M University
Data-centric AI is the discipline of systematically engineering the data used to build an AI system. – Andrew Ng

**Pitfall:** The concept “data-driven” differs fundamentally from “data-centric”. “Data-driven” only emphasizes the use of data to guide AI development, which typically still centers on developing models rather than engineering data.
Many major AI breakthroughs occur only after we have the access to the right training data.

<table>
<thead>
<tr>
<th>Year</th>
<th>AI Breakthrough</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>IBM Deep Blue defeated Garry Kasparov</td>
<td>700,000 Grandmaster chess games (1991)</td>
</tr>
<tr>
<td>2012</td>
<td>AlexNet, one of the first successful CNNs</td>
<td>ImageNet corpus of 1.5 million labeled images (2010)</td>
</tr>
<tr>
<td>2021</td>
<td>AlphaFold, AI for science</td>
<td>Annotated protein sequence (2017)</td>
</tr>
<tr>
<td>Now</td>
<td>Large language models</td>
<td>Large text data</td>
</tr>
</tbody>
</table>

Data is the driving force when model design becomes mature.

- **GPT-1**: 4.8GB (unfiltered) data
- **GPT-2**: 40GB human-filtered data
- **GPT-3**: 570GB data filtered from 45TB raw data
- **ChatGPT/GPT-4**: Human demonstrations and annotations

Data size ↑
Data quality ↑
Need for data-centric AI

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

XXX. YYY. ZZZ. Explain the above in one sentence.

What is 15 * 67 + 6?

15 * 67 + 6 = 1005 + 6 = 1011.

"The drink is okay." neutral, negative or positive?

The statement "The drink is okay" is neutral.
The success of Segment Anything is largely attributed to a annotated dataset with over 1 billion masks, which is 400x larger than the existing one. Segment Anything has three stages of labeling: **assisted-manual stage**, **semi-automatic stage**, and **fully automatic stage**.
Pitfall: While “data-centric AI” is a new concept, it is not completely new. Many tasks (e.g., data augmentation and data labeling) have been studied since decades ago. At the same time, many new tasks and ideas are also emerging, such as data programming.
Research question 1: How can we construct the right training data to improve the performance?
Research question 2: How can we construct the right inference data to evaluate the model or probe knowledge from the model?
Data maintenance

Research question 3: How can we ensure the data is right in a dynamic production environment?
**Representative data-centric AI techniques**

**Data programming (labeling):** We infer labels based on human-designed labeling functions.

---

Representative data-centric AI techniques

**RLHF (labeling):** Reinforcement learning from human feedback, a key technique behind ChatGPT and GPT-4.

Representative data-centric AI techniques

Data valuation: How valuable is the data in the marketplace?

**Trend 1: Automation and collaboration**

**Automation & Collaboration:** To keep pace with the ever-growing size of the available data, we need more efficient algorithms to incorporate human knowledge or automate the process.

![Diagram showing automation and collaboration categories](image)

**Trend 2: blurred data-model boundary**

**Foundation models become a form of data or a “container” of data:** When model becomes sufficiently powerful, we can use models to generate data.

---

Moving towards data-centric AI

Cross-task automation: Can we jointly optimize tasks aimed at different goals, ranging from training data development to inference data development and data maintenance.

Data-model co-design: Can we co-design data and models towards better performance?

Debiasing data: How can we mitigate bias for the tasks under the three data-centric AI goals?

Tackling data in various modalities: How can we effectively deal with data in other formats, such as graph and time-series?

Data benchmarks development: Can we develop a more unified data benchmark?