RICE

ATM

siam 2023 International Conference on Data Mining

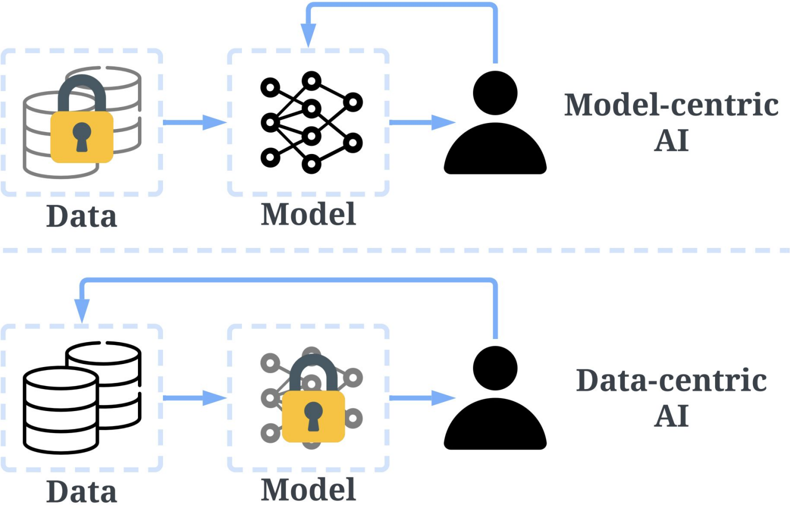
Data-centric AI: Perspective and Challenges

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

Rice University
Texas A&M University

What is data-centric AI?

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.

Need for data-centric AI

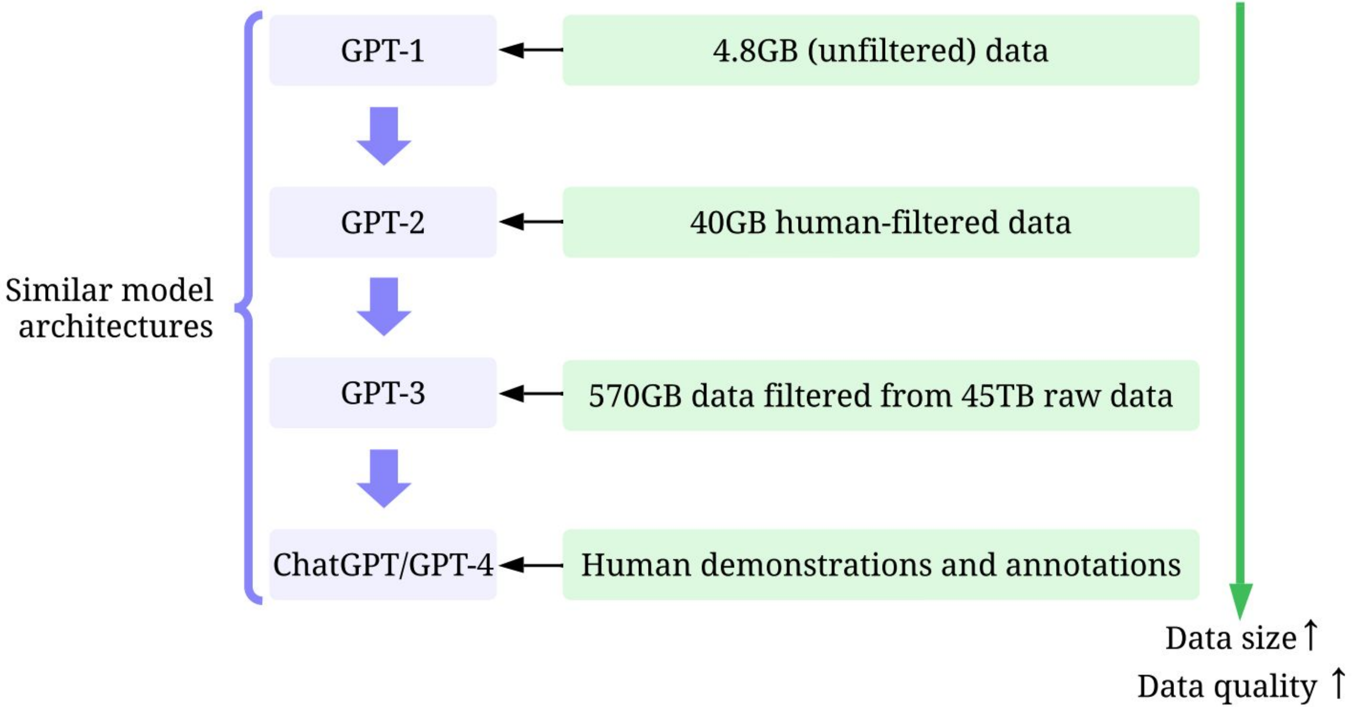
Many major AI breakthroughs occur only after we have the access to the right training data.

Year	AI Breakthrough	Dataset
1994	Human-level spontaneous speech recognition	Spoken Wall Street Journal articles and other texts (1991)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games (1991)
2012	AlexNet, one of the first successful CNNs	ImageNet corpus of 1.5 million labeled images (2010)
2021	AlphaFold, AI for science	Annotated protein sequence (2017)
Now	Large language models	Large text data

[1] http://www.spacemachine.net/views/2016/3/datasets-over-algorithms

Need for data-centric AI

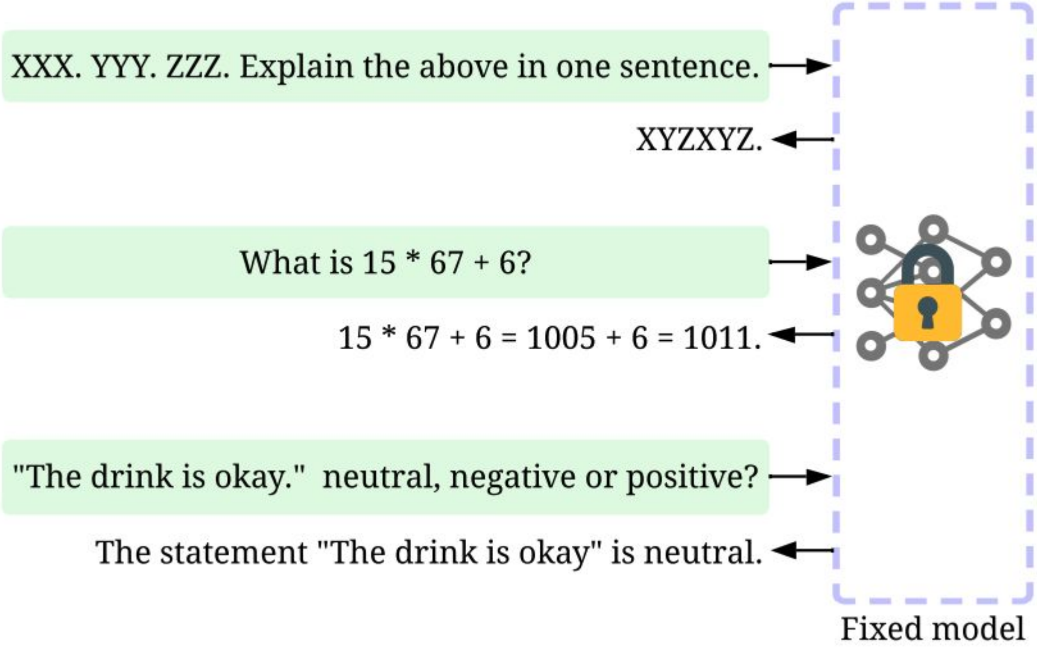
Data is the driving force when model design becomes mature.



Similar model architectures
GPT-1
GPT-2
GPT-3
ChatGPT/GPT-4
4.8GB (unfiltered) data
40GB human-filtered data
570GB data filtered from 45TB raw data
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.

XYZXYZ.

What is 15 * 67 + 6?

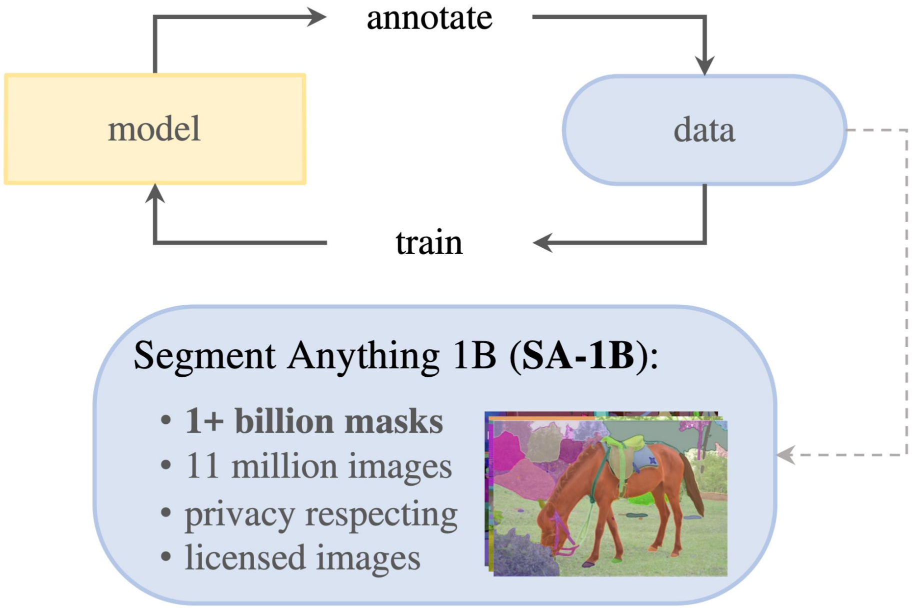
15 * 67 + 6 = 1005 + 6 = 1011.

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

The statement "The drink is okay" is neutral.

Need for data-centric AI

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**.



annotate

model

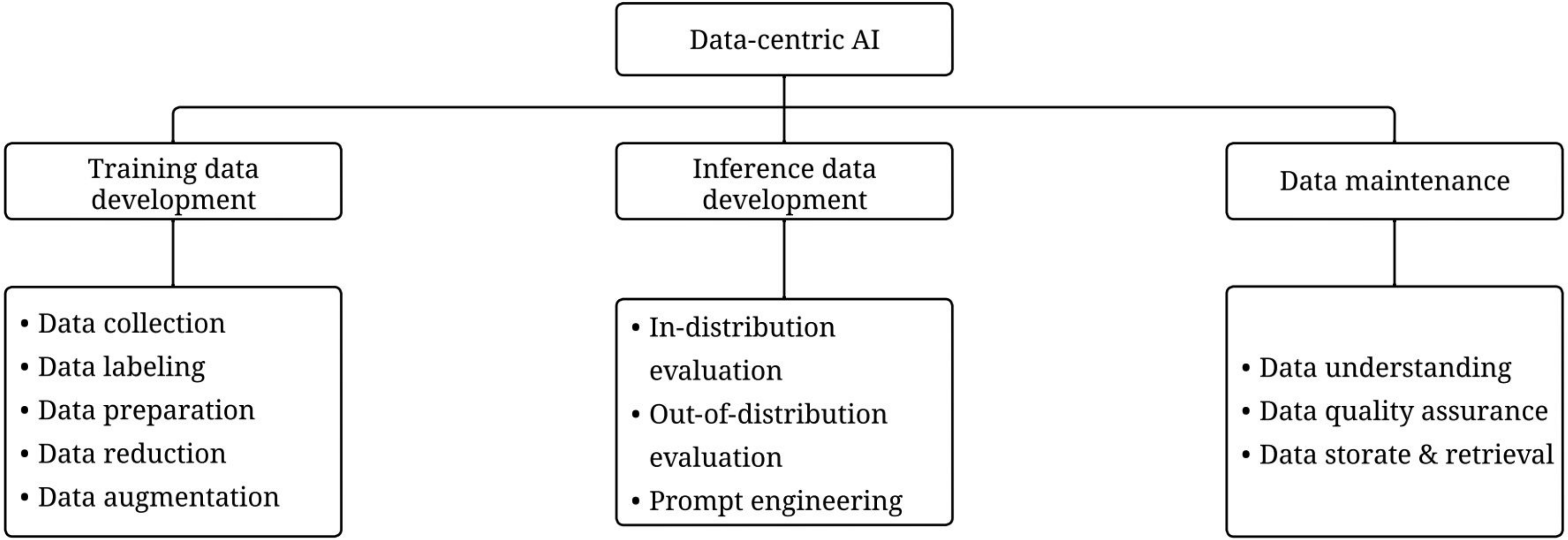
data

train

Segment Anything 1B (**SA-1B**):

- **1+ billion masks**
- 11 million images
- privacy respecting
- licensed images

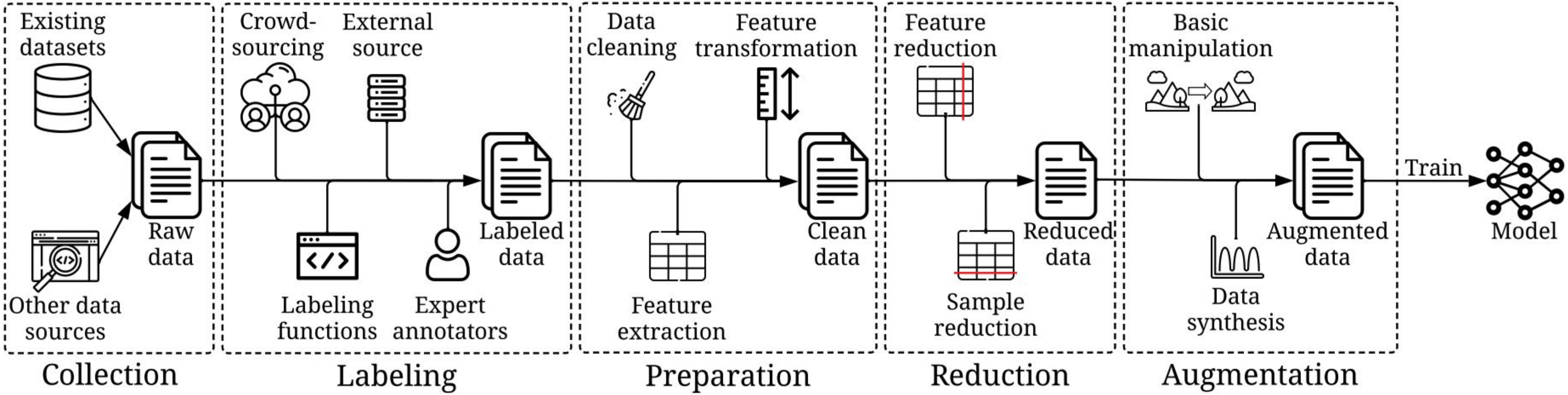
A data-centric AI framework



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.

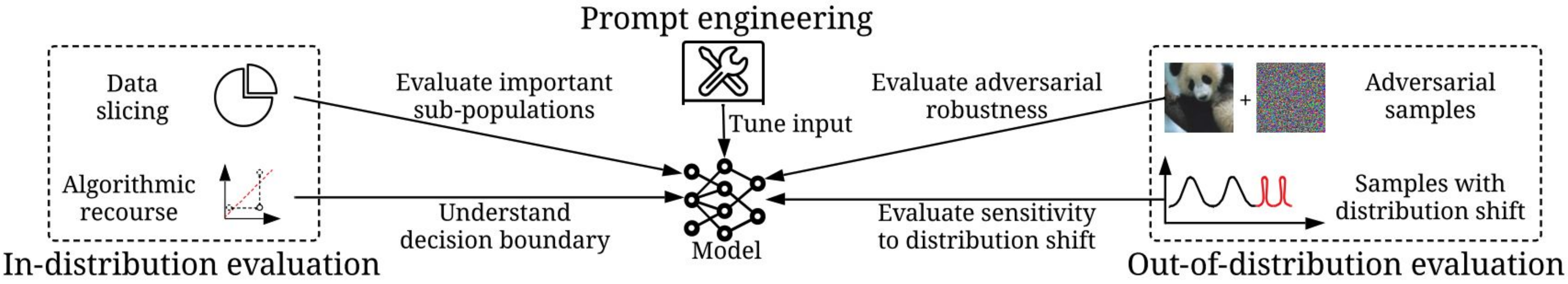
Training data development

Research question 1: How can we construct the right training data to improve the performance?



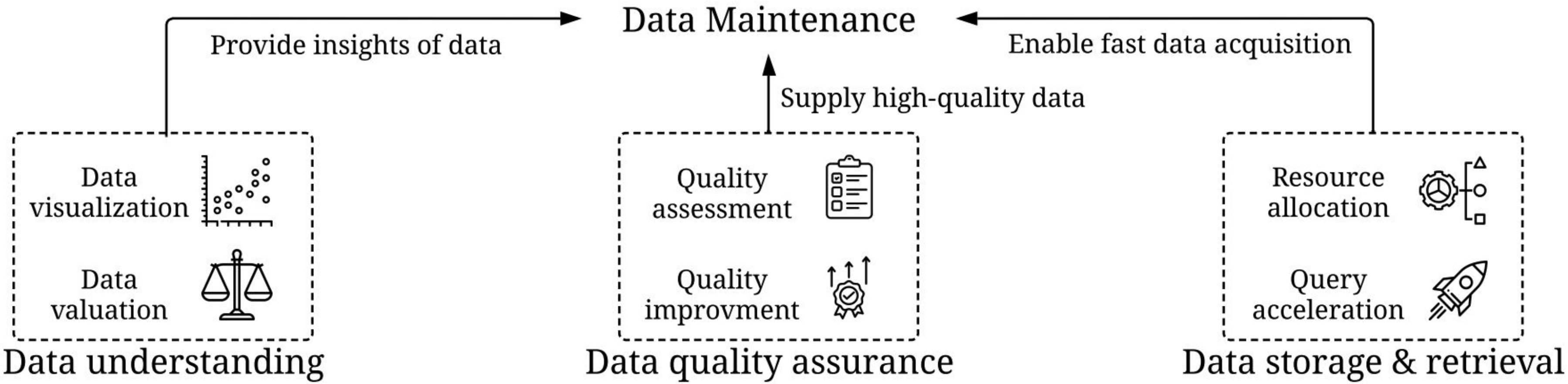
Inference data development

Research question 2: How can we construct the right inference data to evaluate the model or probe knowledge from the model?


Prompt engineering
Tune input
Model
Data slicing
Algorithmic recourse
In-distribution evaluation
Evaluate important sub-populations
Understand decision boundary
Evaluate adversarial robustness
Evaluate sensitivity to distribution shift
+
Adversarial samples
Samples with distribution shift
Out-of-distribution evaluation

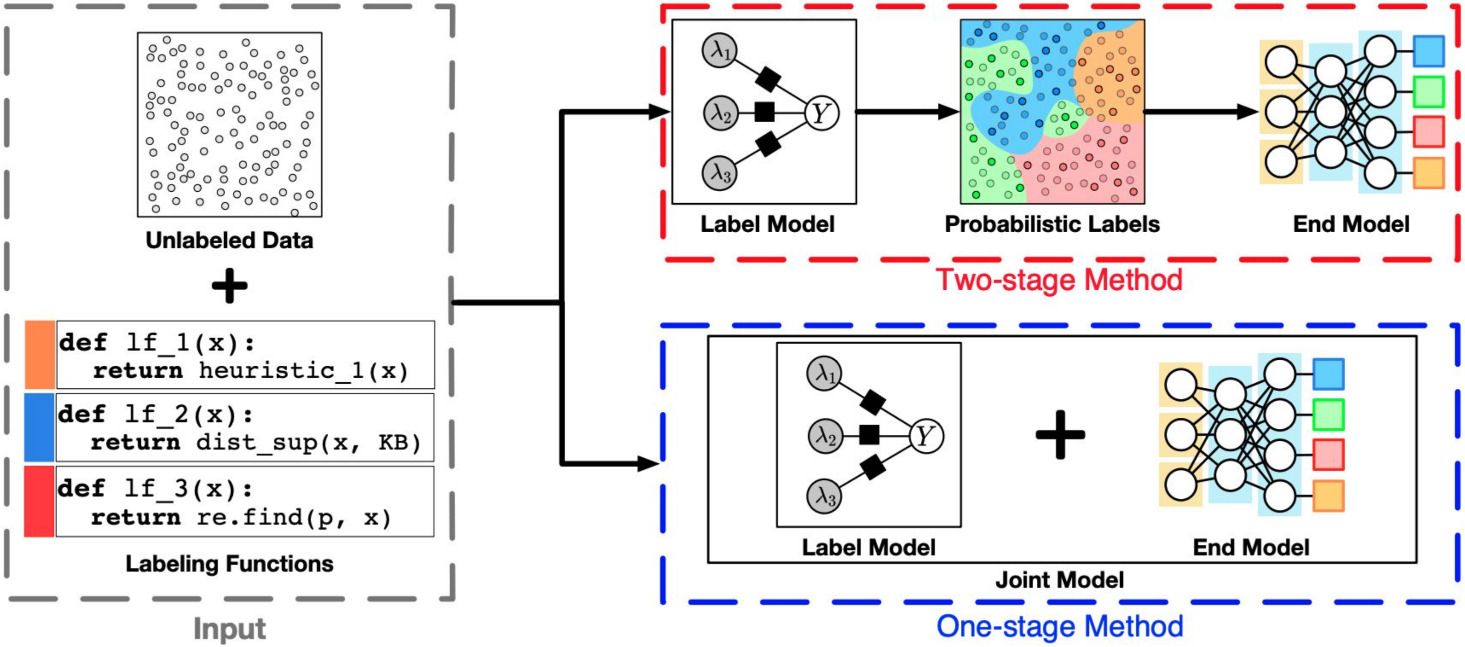
Data maintenance

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.

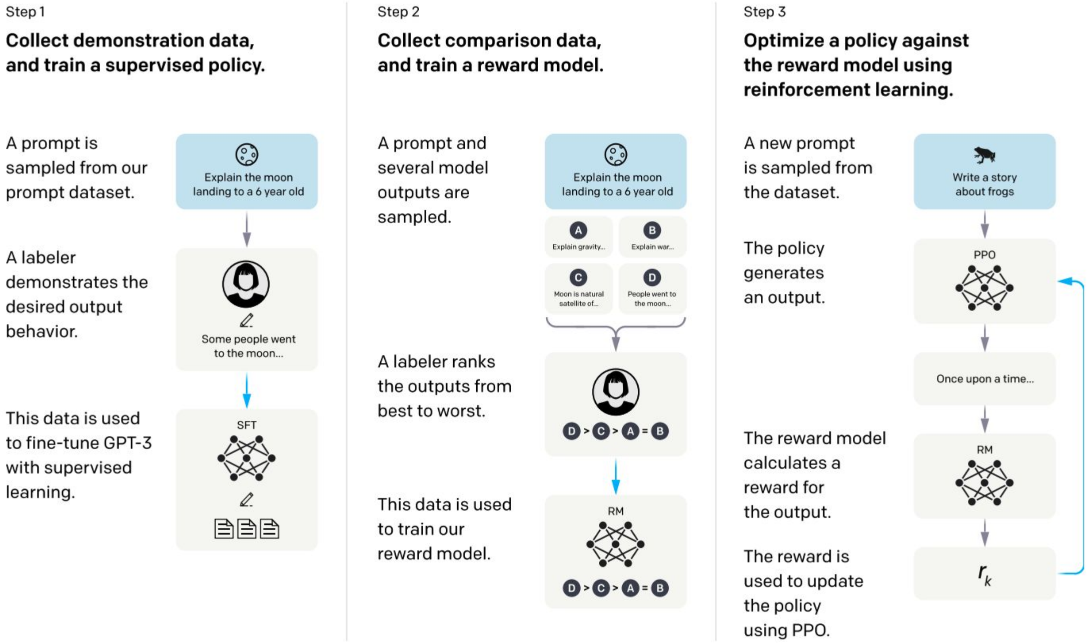


```
def lf_1(x):
  return heuristic_1(x)
def lf_2(x):
  return dist_sup(x, KB)
def lf_3(x):
  return re.find(p, x)
```

[1] Zhang, Jieyu, et al. Wrench: A comprehensive benchmark for weak supervision. NeurIPS, 2021.

Representative data-centric AI techniques

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



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

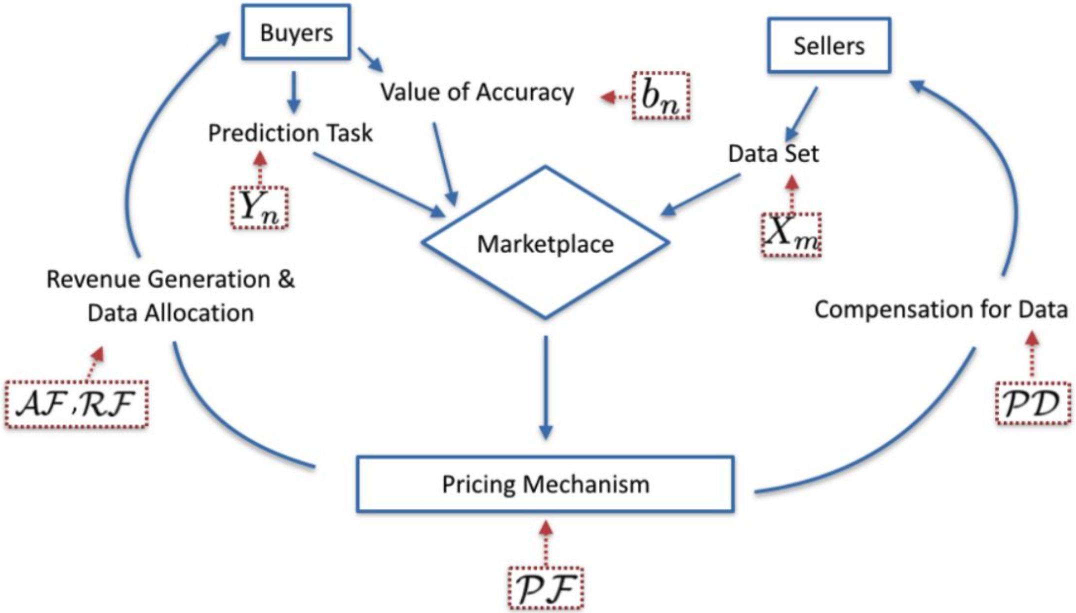
The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

[1] Ouyang, Long, et al. Training language models to follow instructions with human feedback. NeurIPS 2022.

Representative data-centric AI techniques

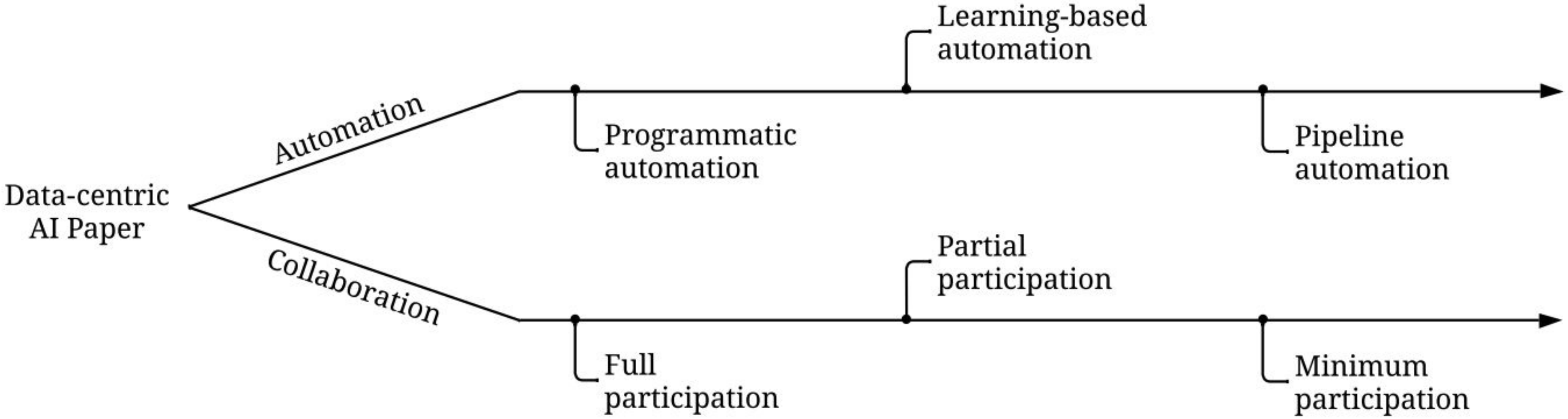
Data valuation: How valuable is the data in the marketplace?



[1] Agarwal, Anish, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. EC, 2019.

Trend 1: Automation and collaboration

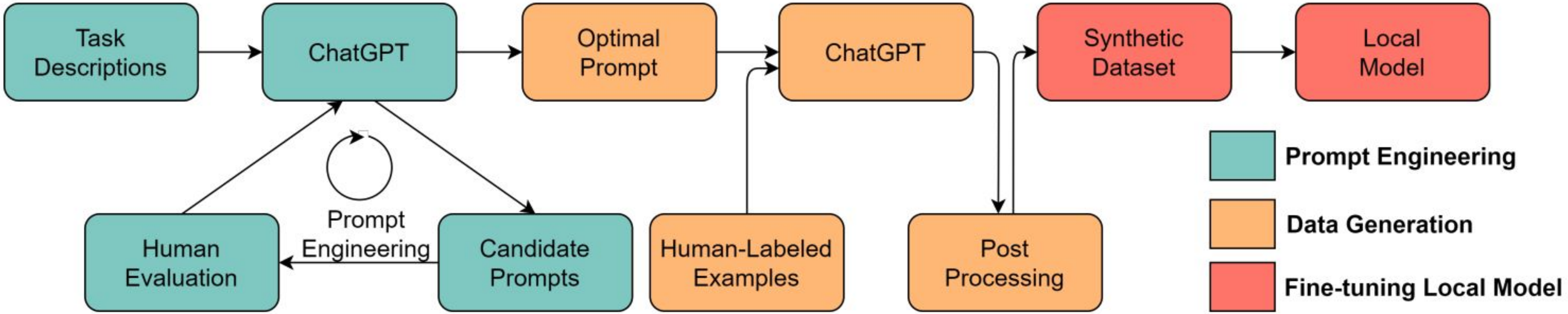
Automation & Collaboration: To keep pace with the ever-growing size of the available data, we need more efficient algorithms to incorporate human knowledge or automate the process.



[1] Zha, Daochen, et al. Data-centric Artificial Intelligence: A Survey. arXiv, 2023.

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.



[1] Tang, Ruixiang, et al. "Does Synthetic Data Generation of LLMs Help Clinical Text Mining?." arXiv preprint arXiv:2303.04360 (2023).

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?

Data-centric AI Perspectives

Data-centric AI Survey

GitHub Resources