NEURAL INFORMATION PROCESSING SYSTEMS

Position Paper

A Sustainable AI Economy Needs Data Deals That Work for Generators

Ruoxi Jia, Luis Oala, Wenjie Xiong, Suqin Ge, Jiachen Wang,
Feiyang Kang, Dawn Song

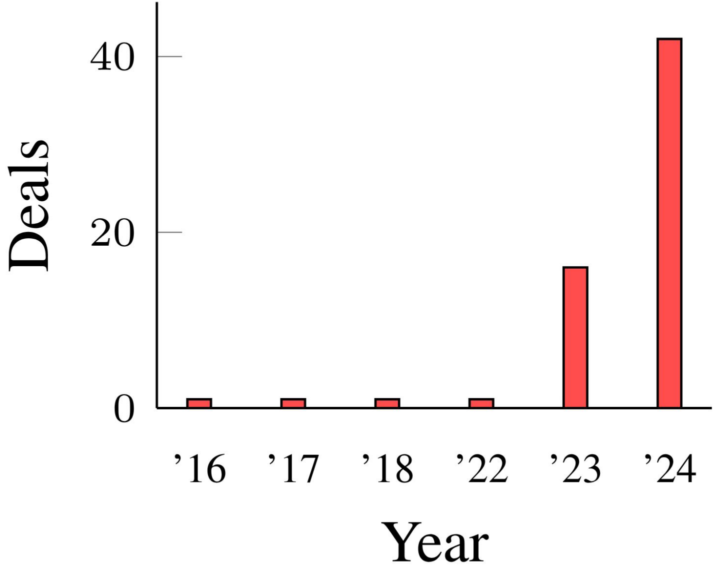
VIRGINIA TECH

Brickroad

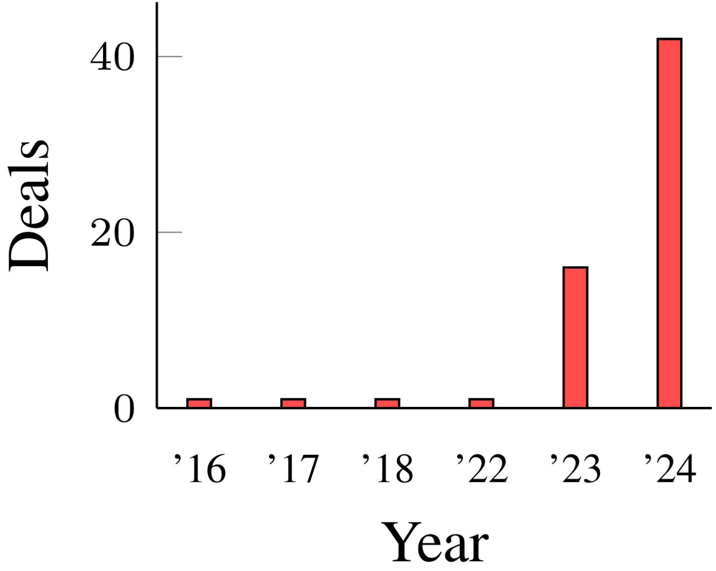
PRINCETON UNIVERSITY

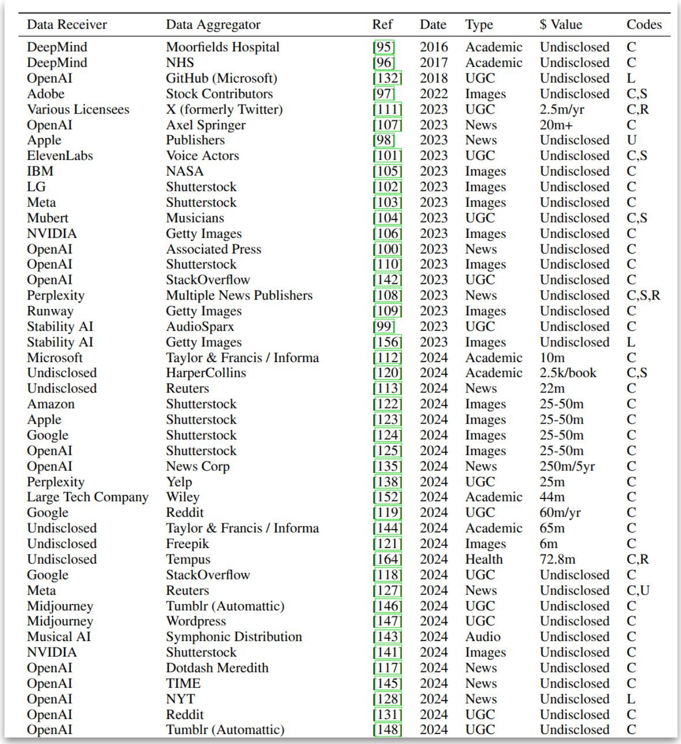
Berkeley UNIVERSITY OF CALIFORNIA

Context



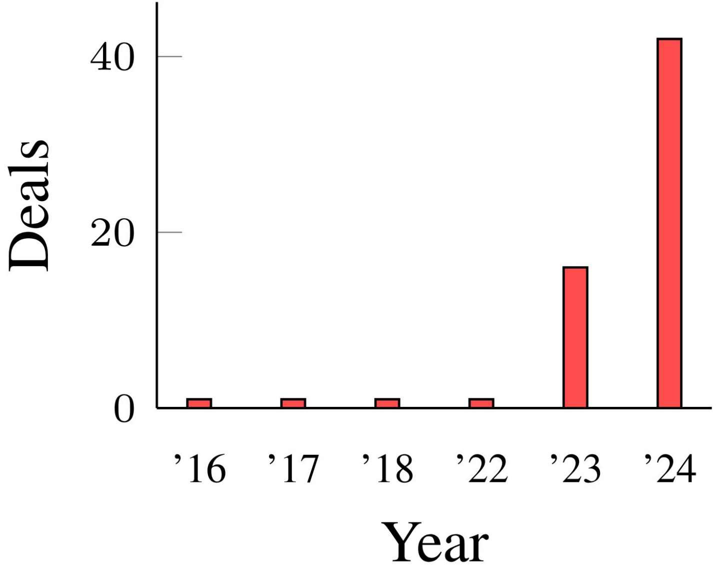
Context

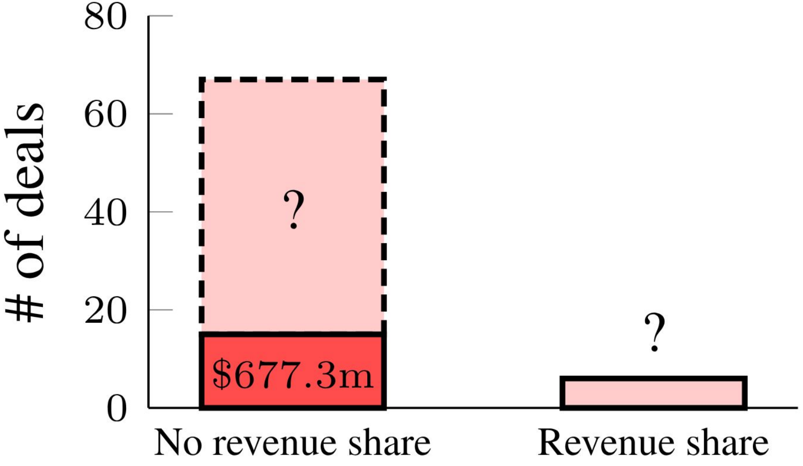




Data Receiver	Data Aggregator	Ref	Date	Type	$ Value	Codes
DeepMind	Moorfields Hospital	[95]	2016	Academic	Undisclosed	C
DeepMind	NHS	[96]	2017	Academic	Undisclosed	C
OpenAI	GitHub (Microsoft)	[132]	2018	UGC	Undisclosed	L
Adobe	Stock Contributors	[97]	2022	Images	Undisclosed	C,S
Various Licensees	X (formerly Twitter)	[111]	2023	UGC	2.5m/yr	C,R
OpenAI	Axel Springer	[107]	2023	News	20m+	C
Apple	Publishers	[98]	2023	News	Undisclosed	U
ElevenLabs	Voice Actors	[101]	2023	UGC	Undisclosed	C,S
IBM	NASA	[105]	2023	Images	Undisclosed	C
LG	Shutterstock	[102]	2023	Images	Undisclosed	C
Meta	Shutterstock	[103]	2023	Images	Undisclosed	C
Mubert	Musicians	[104]	2023	UGC	Undisclosed	C,S
NVIDIA	Getty Images	[106]	2023	Images	Undisclosed	C
OpenAI	Associated Press	[100]	2023	News	Undisclosed	C
OpenAI	Shutterstock	[110]	2023	Images	Undisclosed	C
OpenAI	StackOverflow	[142]	2023	UGC	Undisclosed	C
Perplexity	Multiple News Publishers	[108]	2023	News	Undisclosed	C,S,R
Runway	Getty Images	[109]	2023	Images	Undisclosed	C
Stability AI	AudioSparx	[99]	2023	UGC	Undisclosed	C
Stability AI	Getty Images	[156]	2023	Images	Undisclosed	L
Microsoft	Taylor & Francis / Informa	[112]	2024	Academic	10m	C
Undisclosed	HarperCollins	[120]	2024	Academic	2.5k/book	C,S
Undisclosed	Reuters	[113]	2024	News	22m	C
Amazon	Shutterstock	[122]	2024	Images	25-50m	C
Apple	Shutterstock	[123]	2024	Images	25-50m	C
Google	Shutterstock	[124]	2024	Images	25-50m	C
OpenAI	Shutterstock	[125]	2024	Images	25-50m	C
OpenAI	News Corp	[135]	2024	News	250m/5yr	C
Perplexity	Yelp	[138]	2024	UGC	25m	C
Large Tech Company	Wiley	[152]	2024	Academic	44m	C
Google	Reddit	[119]	2024	UGC	60m/yr	C
Undisclosed	Taylor & Francis / Informa	[144]	2024	Academic	65m	C
Undisclosed	Freepik	[121]	2024	Images	6m	C
Undisclosed	Tempus	[164]	2024	Health	72.8m	C,R
Google	StackOverflow	[118]	2024	UGC	Undisclosed	C
Meta	Reuters	[127]	2024	News	Undisclosed	C,U
Midjourney	Tumblr (Automattic)	[146]	2024	UGC	Undisclosed	C
Midjourney	Wordpress	[147]	2024	UGC	Undisclosed	C
Musical AI	Symphonic Distribution	[143]	2024	Audio	Undisclosed	C
NVIDIA	Shutterstock	[141]	2024	Images	Undisclosed	C
OpenAI	Dotdash Meredith	[117]	2024	News	Undisclosed	C
OpenAI	TIME	[145]	2024	News	Undisclosed	C
OpenAI	NYT	[128]	2024	News	Undisclosed	L
OpenAI	Reddit	[131]	2024	UGC	Undisclosed	C
OpenAI	Tumblr (Automattic)	[148]	2024	UGC	Undisclosed	C

Context

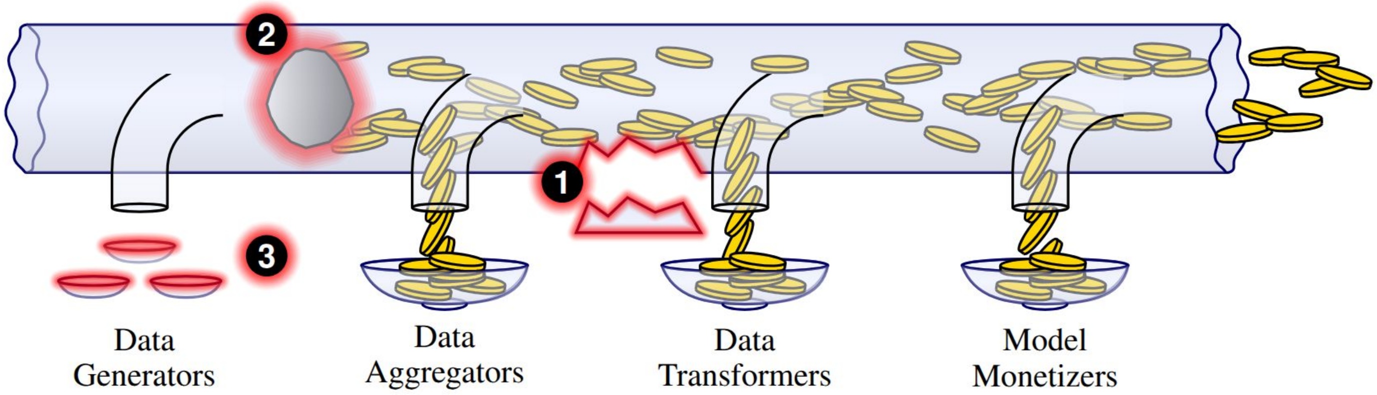

Deals
40
20
0
'16
'17
'18
'22
'23
'24
Year


# of deals
80
60
40
20
0
?
$677.3m
?
No revenue share
Revenue share

Economic Data Processing Inequality

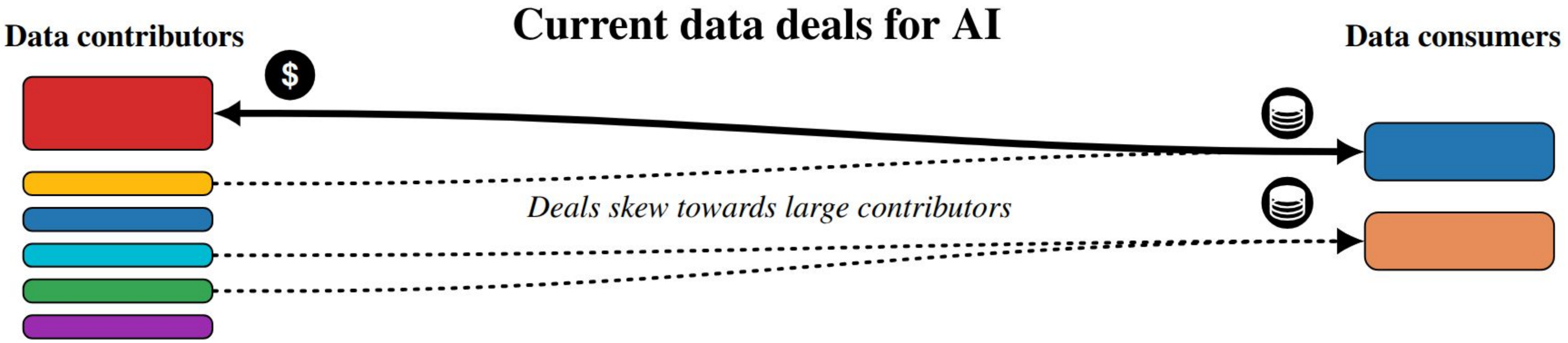
1 Invisible Provenance

2 Asymmetric Bargaining

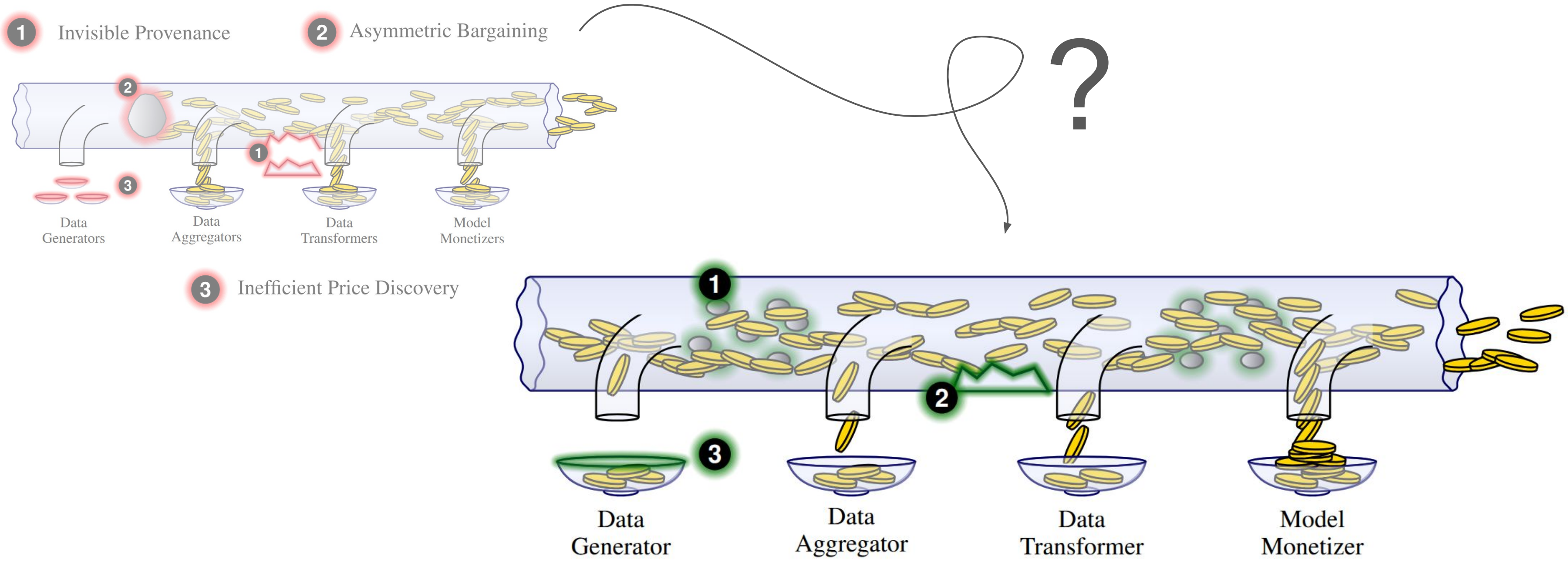


3 Inefficient Price Discovery

Economic Data Processing Inequality


Current data deals for AI
Data contributors
Data consumers
$
Deals skew towards large contributors

Position Paper



1 Invisible Provenance

2 Asymmetric Bargaining

3 Inefficient Price Discovery

Data Generators

Data Aggregators

Data Transformers

Model Monetizers

?

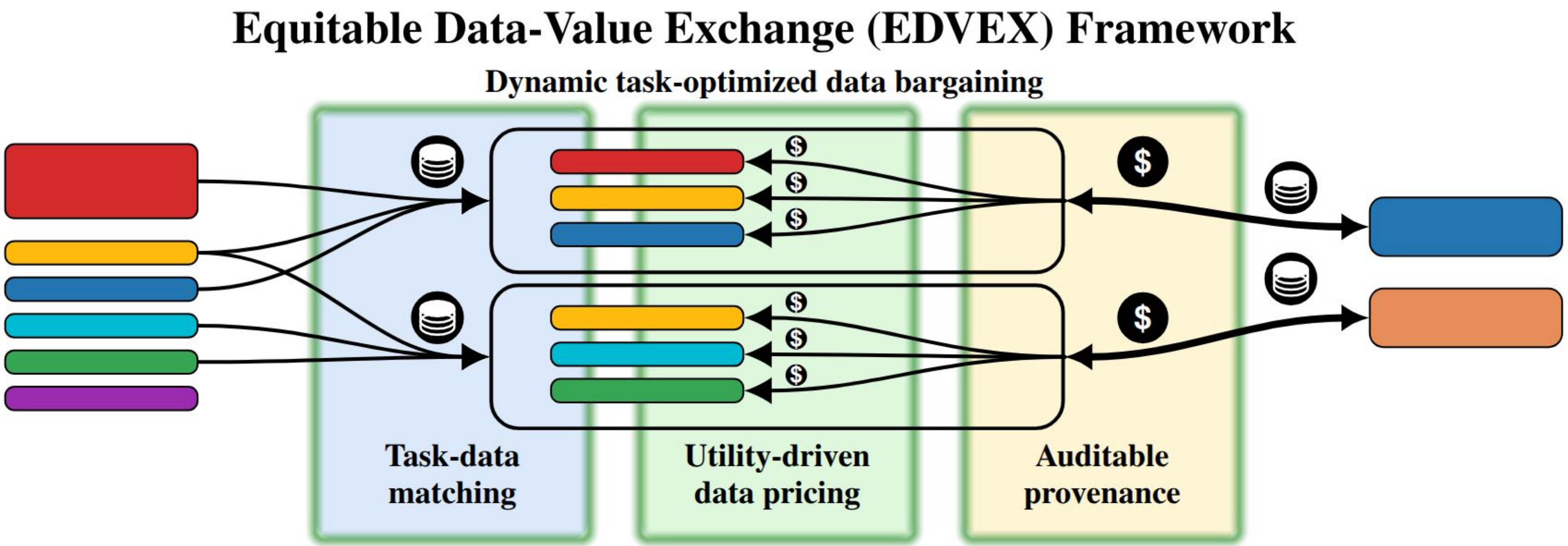
Data Generator

Data Aggregator

Data Transformer

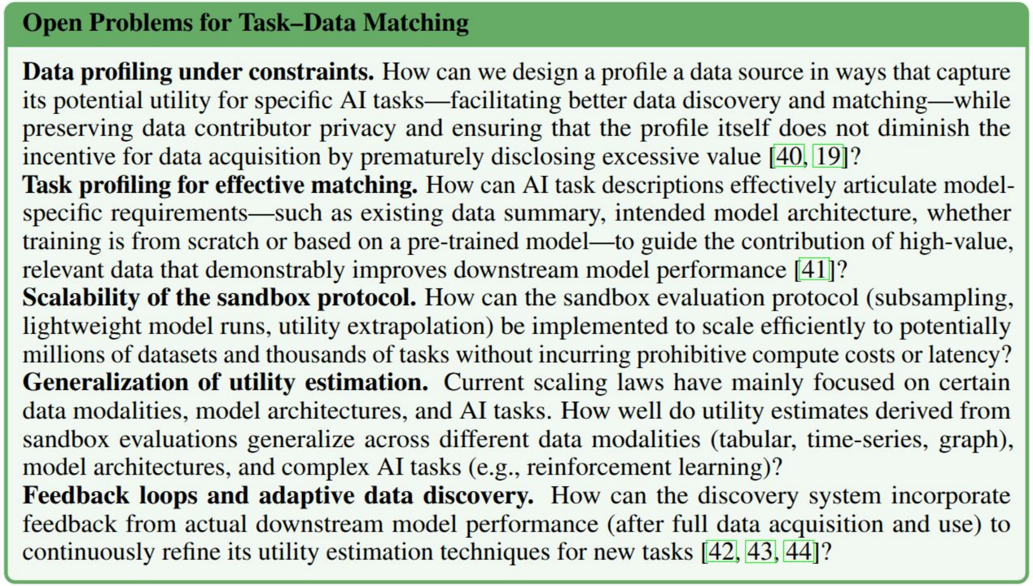
Model Monetizer

EDVEX: Primitives for an Efficient AI Data Economy


Equitable Data-Value Exchange (EDVEX) Framework
Dynamic task-optimized data bargaining
$
Task-data matching
Utility-driven data pricing
Auditable provenance

EDVEX: Open Problems

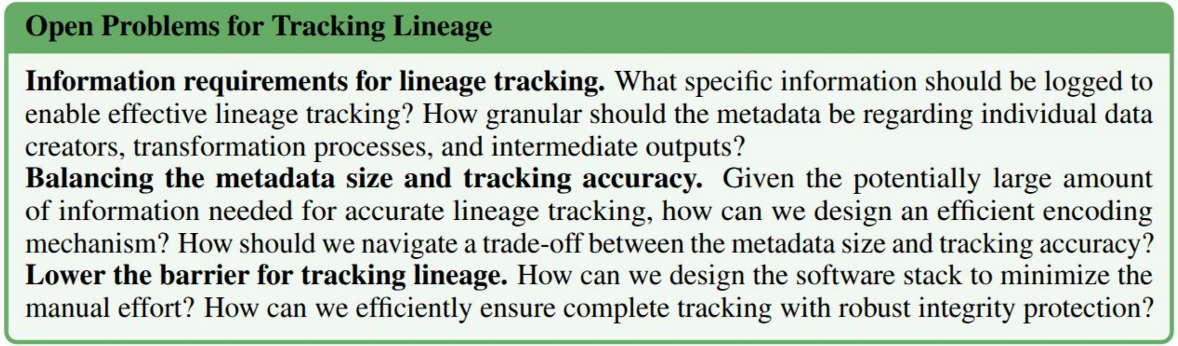
EDVEX: Open Problems - Task-Data Matching



Open Problems for Task–Data Matching

Data profiling under constraints. How can we design a profile a data source in ways that capture its potential utility for specific AI tasks—facilitating better data discovery and matching—while preserving data contributor privacy and ensuring that the profile itself does not diminish the incentive for data acquisition by prematurely disclosing excessive value [40, 19]?
Task profiling for effective matching. How can AI task descriptions effectively articulate model-specific requirements—such as existing data summary, intended model architecture, whether training is from scratch or based on a pre-trained model—to guide the contribution of high-value, relevant data that demonstrably improves downstream model performance [41]?
Scalability of the sandbox protocol. How can the sandbox evaluation protocol (subsampling, lightweight model runs, utility extrapolation) be implemented to scale efficiently to potentially millions of datasets and thousands of tasks without incurring prohibitive compute costs or latency?
Generalization of utility estimation. Current scaling laws have mainly focused on certain data modalities, model architectures, and AI tasks. How well do utility estimates derived from sandbox evaluations generalize across different data modalities (tabular, time-series, graph), model architectures, and complex AI tasks (e.g., reinforcement learning)?
Feedback loops and adaptive data discovery. How can the discovery system incorporate feedback from actual downstream model performance (after full data acquisition and use) to continuously refine its utility estimation techniques for new tasks [42, 43, 44]?

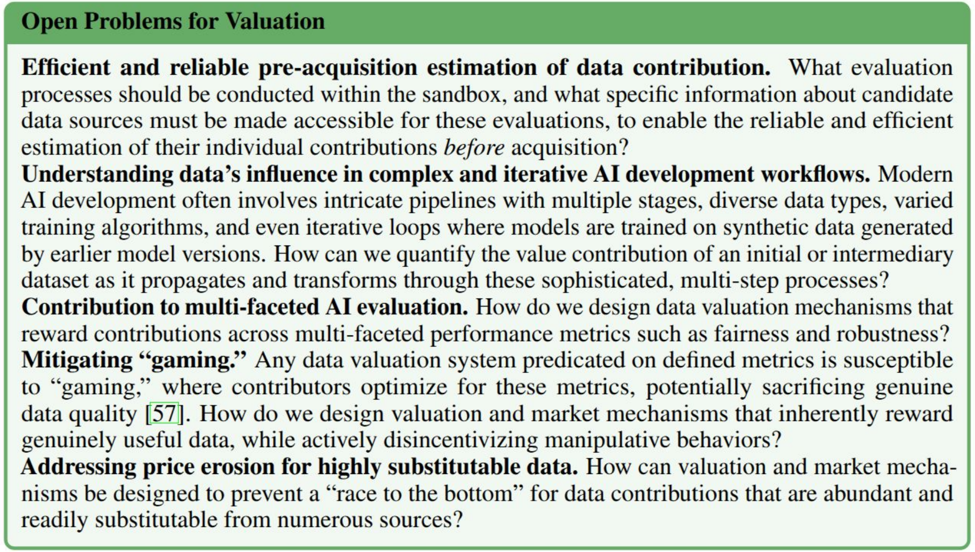
EDVEX: Open Problems - Lineage Tracking



Open Problems for Tracking Lineage

Information requirements for lineage tracking. What specific information should be logged to enable effective lineage tracking? How granular should the metadata be regarding individual data creators, transformation processes, and intermediate outputs?
Balancing the metadata size and tracking accuracy. Given the potentially large amount of information needed for accurate lineage tracking, how can we design an efficient encoding mechanism? How should we navigate a trade-off between the metadata size and tracking accuracy?
Lower the barrier for tracking lineage. How can we design the software stack to minimize the manual effort? How can we efficiently ensure complete tracking with robust integrity protection?

EDVEX: Open Problems - Data Valuation



Open Problems for Valuation

Efficient and reliable pre-acquisition estimation of data contribution. What evaluation processes should be conducted within the sandbox, and what specific information about candidate data sources must be made accessible for these evaluations, to enable the reliable and efficient estimation of their individual contributions *before* acquisition?

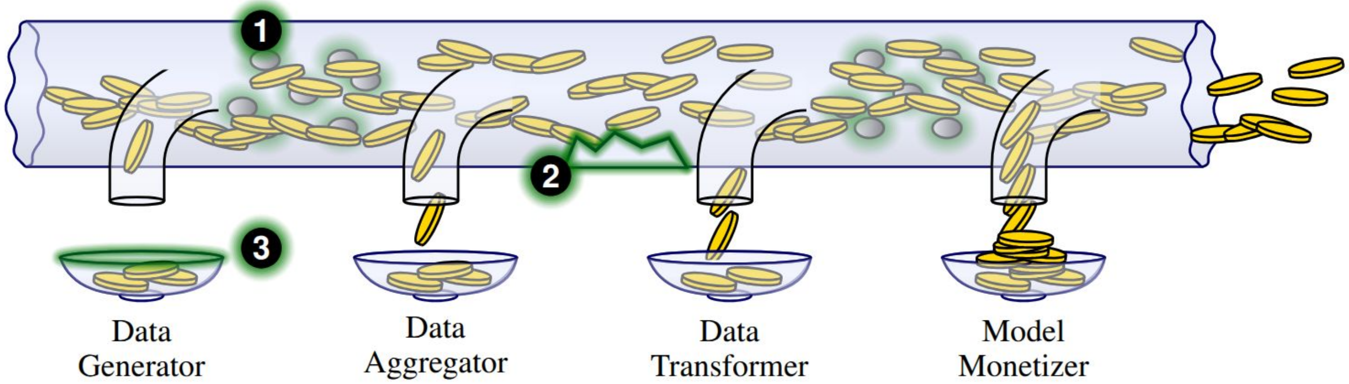
Understanding data's influence in complex and iterative AI development workflows. Modern AI development often involves intricate pipelines with multiple stages, diverse data types, varied training algorithms, and even iterative loops where models are trained on synthetic data generated by earlier model versions. How can we quantify the value contribution of an initial or intermediary dataset as it propagates and transforms through these sophisticated, multi-step processes?

Contribution to multi-faceted AI evaluation. How do we design data valuation mechanisms that reward contributions across multi-faceted performance metrics such as fairness and robustness?

Mitigating "gaming." Any data valuation system predicated on defined metrics is susceptible to "gaming," where contributors optimize for these metrics, potentially sacrificing genuine data quality [57]. How do we design valuation and market mechanisms that inherently reward genuinely useful data, while actively disincentivizing manipulative behaviors?

Addressing price erosion for highly substitutable data. How can valuation and market mechanisms be designed to prevent a "race to the bottom" for data contributions that are abundant and readily substitutable from numerous sources?

1 **Task-Data Matching** 2 **Lineage Tracking**



3 **Valuation**