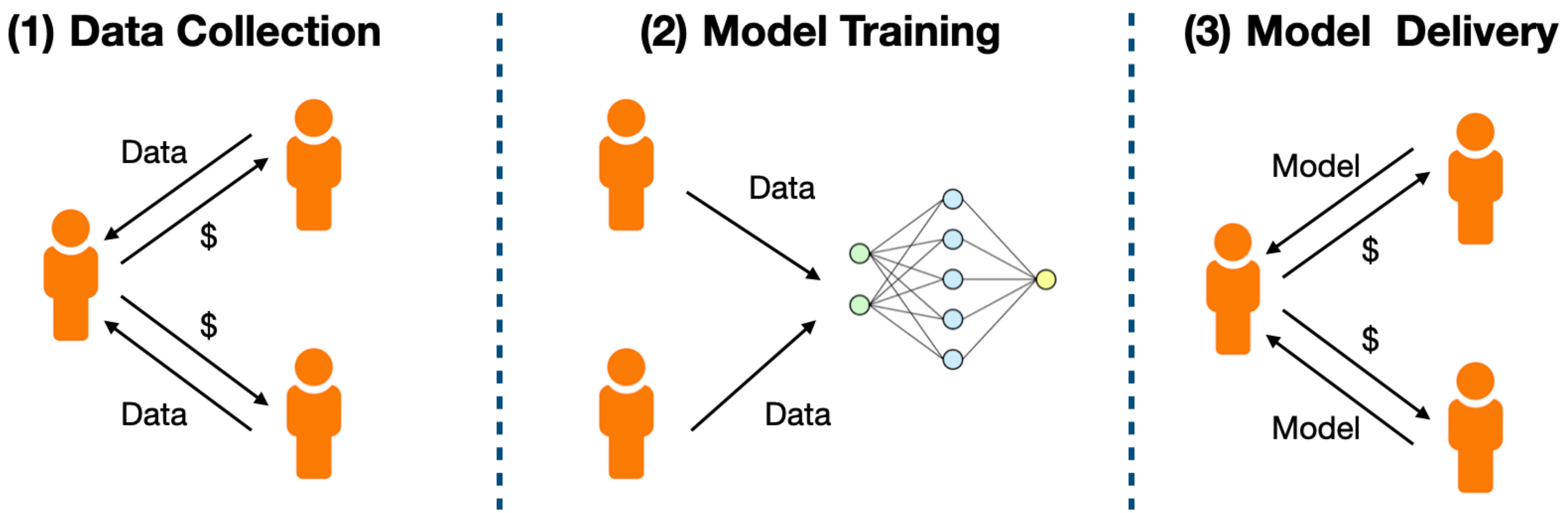
Data Pricing in Machine Learning Pipelines

Zicun Cong, Xuan Luo, Jian Pei (Simon Fraser University)
Feida Zhu (Singapore Management University)

Part I: Introduction

Collaborations in Machine Learning (ML) Pipeline

- The disruptive success of ML in many applications has led to an explosion in demand
- Many parties need to collaborate to build a powerful machine learning application



Example: collaboration scenarios in ML pipelines

- Machine learning applications are indeed pipelines connecting many parties

Data and Model Exchange in ML Pipelines

- Data is critical for machine learning and penetrates the whole ML pipelines
- Obtaining data for machine learning is far from easy
- Data exchange becomes a fundamental interaction among different parties
    - Share, exchange, and reuse data sets and ML models
- What is a principled mechanism to connect many parties in ML pipelines in scale?

Data Products as Economic Goods

- **Data products** refer to data sets as products and information services derived from data sets
- Advantages of data marketplaces
  - Data owners can monetize their data and intelligent properties
  - Data buyers can access data products of high quality and large quantities
- To enable tradings, data has to be priced

Pei, Jian. "A survey on data pricing: from economics to data science." *IEEE Transactions on Knowledge and Data Engineering* (2020).

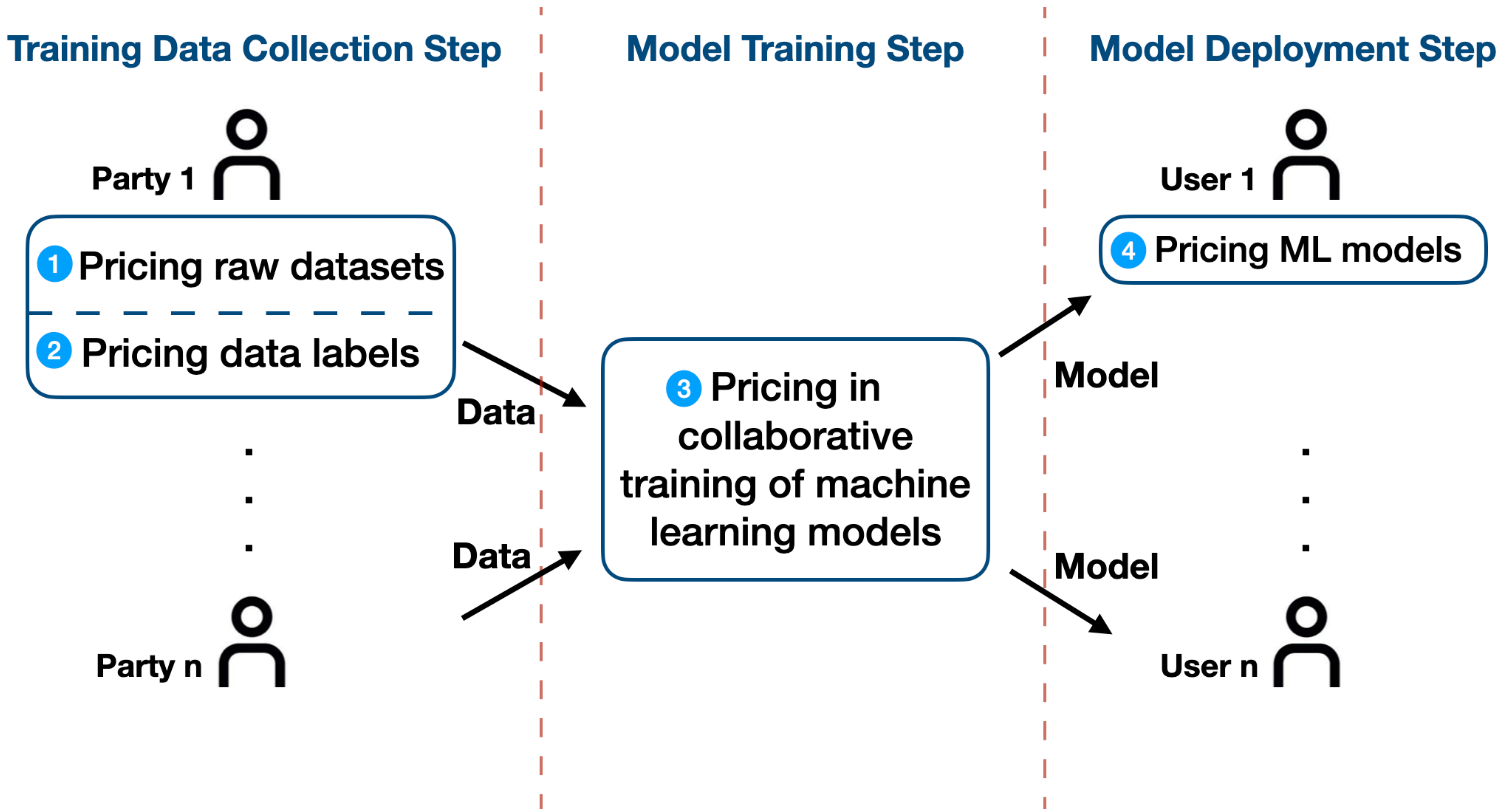
Pricing of Data Products

- What is pricing?
  - The practice that a business sets a price at which a product or a service can be sold
  - 3c's of pricing strategies: cost, consumer, and competitors
- Four challenges
  - Data can be replicated at zero marginal cost
  - The value of data is inherently combinatorial
  - The value of data varies widely among different buyers
  - The usefulness of data lies in the value of information derived from it, which is difficult to verify a priori

De Toni, Deonir, et al. "Pricing strategies and levels and their impact on corporate profitability." *Revista de Administração (São Paulo)* 52 (2017): 120-133.

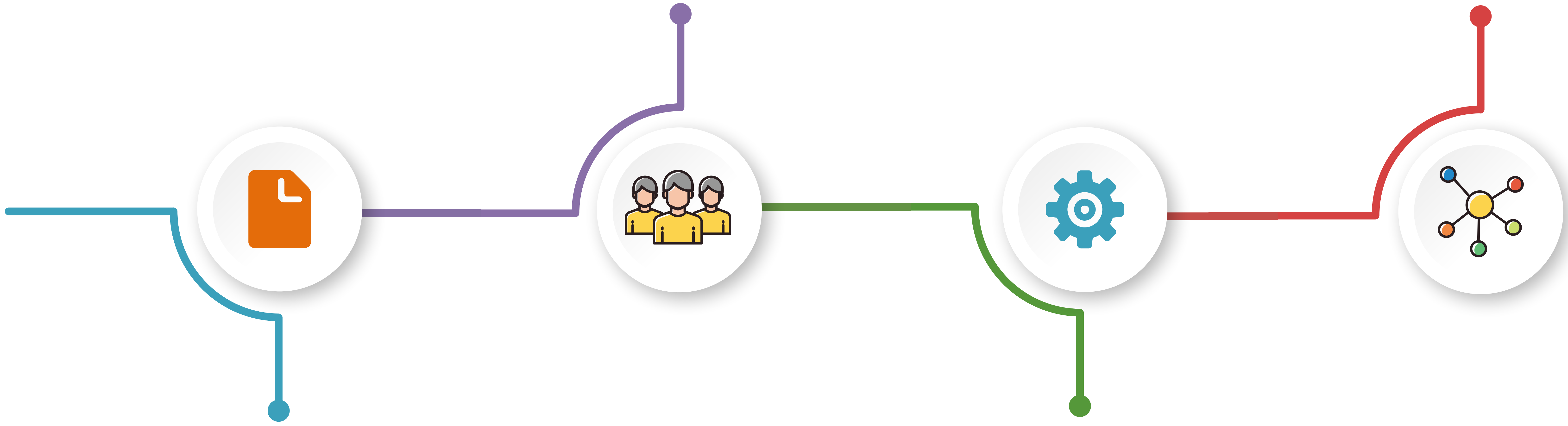
Agarwal, Anish, Munther Dahleh, and Tuhin Sarkar. "A marketplace for data: An algorithmic solution." *Proceedings of the 2019 ACM Conference on Economics and Computation*. 2019.

Four Major Data Pricing Tasks in ML Pipelines



Steps and pricing tasks in machine learning pipelines

Key Challenges in the Four Data Pricing Tasks



Pricing Raw Data Set
Set the price reflecting the usefulness of a data set

Pricing Data Labels
Motivate crowd-workers to invest high efforts and report accurate data labels

Revenue Allocation in Collaborative ML
Fairly reward data owners' contributions

Pricing ML Models
Version machine learning models and avoid arbitrage among multiple versions

A Principle in Data Pricing

- Common core idea: linking prices of data products to their utilities to customers
- Two types of utility functions
    - Absolute utility function
    - Relative utility function

Roadmap

- Introduction
- Essentials of pricing data and machine learning models
- Pricing in data collection - pricing raw data sets
- Pricing in data collection - pricing data labels
- Pricing in collaborative training of machine learning models
- Pricing machine learning models
- Summary and future directions

Part II:
Essentials of Pricing Data and Machine Learning Models

What Is a Data Marketplace?

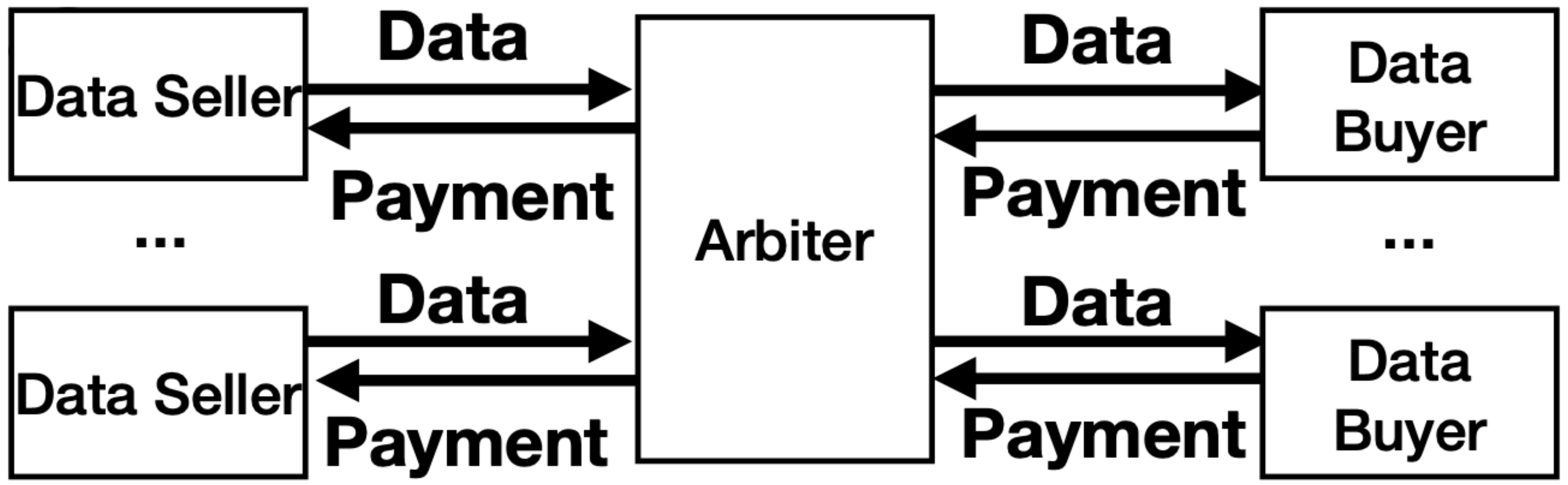
- A platform that allows people to buy and sell data products
- Seven categories of participants
  - Data providers, analysts, application vendors, data processing algorithm developers, consultants, licensing and certification entities, and data market owners
- Four types of market structures
  - Monopoly, oligopoly, strong competition markets, and monopsony
- Examples of data marketplaces
  - Personal data marketplaces, crowd-sensing data marketplaces, and ML model marketplaces, etc.

Muschalle, Alexander, et al. "Pricing approaches for data markets." *International workshop on business intelligence for the real-time enterprise*. Springer, Berlin, Heidelberg, 2012.

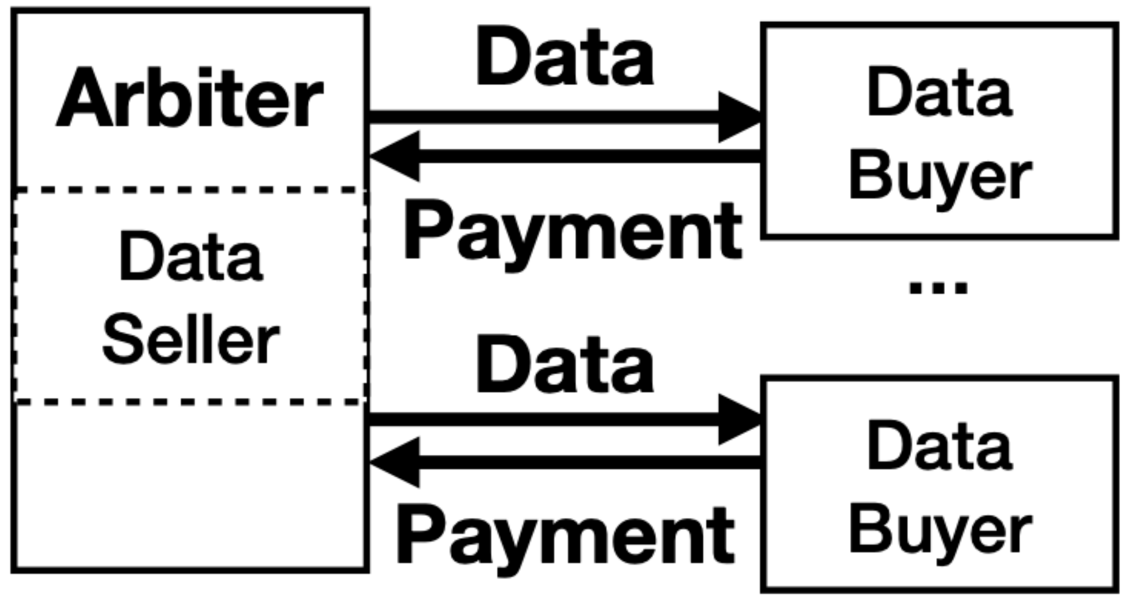
Fricker, Samuel A., and Yuliyan V. Maksimov. "Pricing of data products in data marketplaces." *International Conference of Software Business*. Springer, Cham, 2017.

Fernandez, Raul Castro, Pranav Subramaniam, and Michael J. Franklin. "Data market platforms: trading data assets to solve data problems." *Proceedings of the VLDB Endowment* 13.12 (2020): 1933-1947.

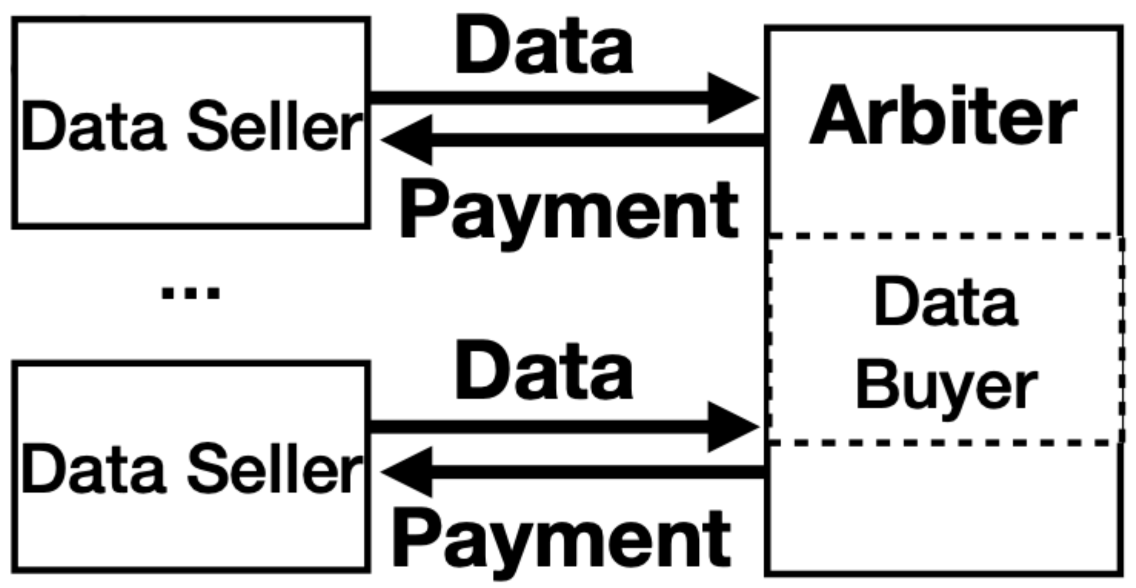
Data Marketplace Architectures



(a) General data marketplace



(b) Sell-side marketplace



(c) Buy-side marketplace

Zhang, Mengxiao, and Fernando Beltrán. "A Survey of Data Pricing Methods." *Available at SSRN 3609120* (2020).

Major Data Pricing Strategies

- Three major data pricing strategies
  - Cost-based pricing
  - Customer value-based pricing
  - Competition-based pricing
- Other pricing strategies
  - Operation-oriented pricing, revenue-oriented pricing, and relationship-oriented pricing

Nagle, Thomas T., and Georg Müller. *The strategy and tactics of pricing: A guide to growing more profitably*. Routledge, 2017.

De Toni, Deonir, et al. "Pricing strategies and levels and their impact on corporate profitability." *Revista de Administração (São Paulo)* 52 (2017): 120-133.

Desiderata of Data Pricing

- Truthfulness
- Revenue Maximization
- Fairness
- Arbitrage-free Pricing
- Privacy-preservation
- Computational Efficiency
- Effort Elicitation

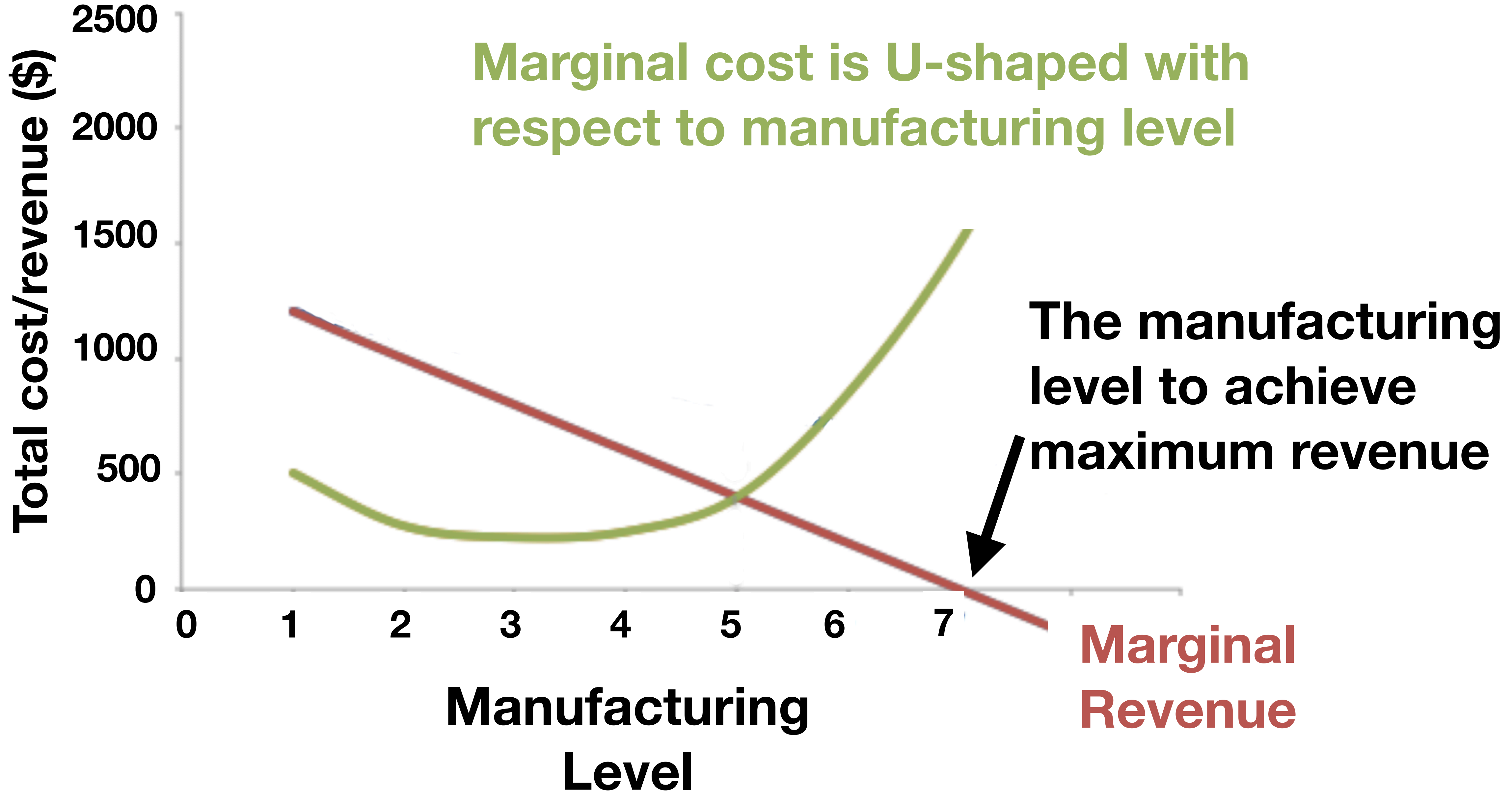
Truthfulness

- In a truthful market, all participants are selfish and only offer prices that maximize their utility values
    - Offering the real values of products is a participant's best strategy
- Reverse auction is a common tool to implement truthful data markets
    - Reverse auction: one buyer and many potential sellers (Forward auction: one seller and multiple competing buyers)
- Myerson's lemma of truthful sealed-bid reverse auction
    - The selection rule of auction winners is monotone
        - If a seller s_i wins the auction by bidding b_i, the seller also wins by bidding $b_i' \leq b_i$
    - Each selected seller s_i is paid the critical value p_i
        - Critical value p_i : seller s_i would not win the auction if s_i bids higher than p_i

Myerson, Roger B. "Optimal auction design." *Mathematics of operations research* 6.1 (1981): 58-73.

Revenue Maximization

- Increase a seller's customer base by having low prices
- Revenue maximization for physical goods is achieved when the marginal revenue is zero
- Data products can be re-produced at almost zero costs
    - The revenue maximization techniques for data products are quite different



Burkett, John P. *Microeconomics: optimization, experiments, and behavior*. Oxford University Press, 2006.

Fairness

- A market is fair if a seller gets a fair allocation for the seller's contribution in a coalition
- Shapley fairness
  - *Balance*: the payment should be fully distributed to the sellers
  - *Symmetry*: the same contribution to the payment should be paid the same
  - *Zero element*: no contribution means no payments
  - *Additivity*: if the data sets can be used for two tasks t_1 and t_2 with payments v_1 and v_2, respectively, then the payment to solve both tasks $t_1 + t_2$ should be $v_1 + v_2$

Shapley, Lloyd S. *17. A value for n-person games*. Princeton University Press, 2016.

Shapley Value

- Shapley value is the unique allocation solution that satisfies Shapley fairness

$$\psi(s) = \frac{1}{N} \sum_{S \subseteq D \setminus \{s\}} \frac{\mathcal{U}(S \cup \{s\}) - \mathcal{U}(S)}{\binom{N-1}{|S|}}$$

Equivalently,

$$\psi(s) = \frac{1}{N!} \sum_{\pi \in \prod(D)} (\mathcal{U}(P_s^\pi \cup \{s\} - \mathcal{U}(P_s^\pi)))$$

- Exponential computational cost with respect to the number of sellers
  - Can be estimated by sampling methods

Shapley, Lloyd S. *17. A value for n-person games*. Princeton University Press, 2016.

Arbitrage-free Pricing

- Arbitrage is the activities that take advantage of price differences between multiple markets
- A data buyer may circumvent the advertised price of a product through buying a bundle of cheaper ones
  - Example: an answer with a variance of 5 is sold at $5 and with a variance of 1 is sold at $50. A data buyer wants to obtain an answer of variance 1. The buyer can purchase the cheaper answer 5 times and compute their average. The total cost is only $25

Li, Chao, et al. "A theory of pricing private data." *ACM Transactions on Database Systems (TODS)* 39.4 (2014): 1-28.

Privacy Preservation

- In data marketplaces, the privacy of buyers, sellers, and involved third parties are highly vulnerable
- Our tutorial focuses on compensations for the privacy disclosure of data owners
- Data owners' data sets are protected by differential privacy
- Data owners are paid according to how much their privacy is disclosed

Dandekar, Pranav, Nadia Fawaz, and Stratis Ioannidis. "Privacy auctions for recommender systems." *ACM Transactions on Economics and Computation (TEAC)* 2.3 (2014): 1-22.

Computational Efficiency

- Prices should be computed in polynomial time with respect to the number of participants or the number of data products
- It takes exponential time to compute the pricing functions with some desirable properties, such as Shapley fairness, arbitrage-freeness, and revenue maximization

Chen, Lingjiao, Paraschos Koutris, and Arun Kumar. "Towards model-based pricing for machine learning in a data marketplace." *Proceedings of the 2019 International Conference on Management of Data*. 2019.

Koutris, Paraschos, et al. "Query-based data pricing." *Journal of the ACM (JACM)* 62.5 (2015): 1-44.

Effort Elicitation

- A data buyer may purchase training data labels via crowdsourcing
- Control the quality of collected label is challenging
  - Spammers may provide random labels without solving the tasks
- Design rigorous incentives to guide worker behaviors
- Motivate workers to invest efforts and report accurate labels

Dasgupta, Anirban, and Arpita Ghosh. "Crowdsourced judgement elicitation with endogenous proficiency." *Proceedings of the 22nd international conference on World Wide Web*. 2013.

Summary: Introduction and Essentials of Data Pricing

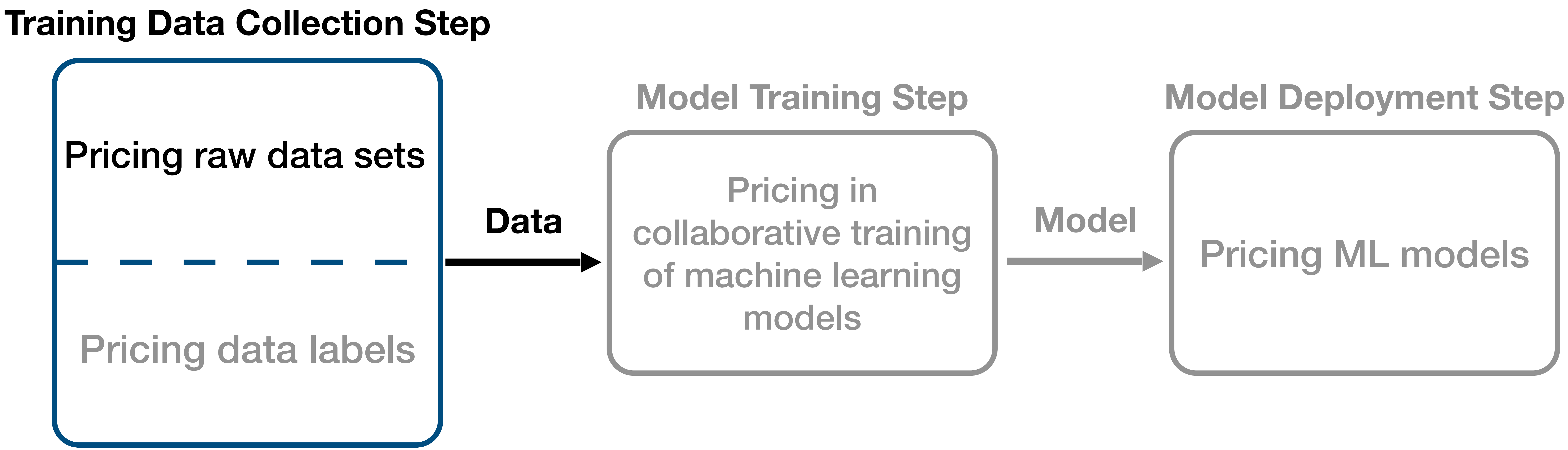
- Data and ML models as economic goods
- Four major pricing tasks in ML pipelines
- Architectures and players in data marketplaces
- Core idea and seven desiderata of data pricing

Part III:
Pricing Raw Data Sets

Outline: Pricing Raw Data Sets

- Introduction
- Pricing General Data Sets
- Pricing Crowd-sensing Data
- Pricing Data Queries
- Compensating Privacy Loss
- Summary

Pricing Raw Data Sets in Machine Learning Pipelines



Major Factors in Data Pricing Models of Raw Data Sets

- Intrinsic factors
  - Data quality: accuracy, volume, freshness, completeness, …
  - Consumption units: whole datasets and subsets
- Extrinsic factors
  - Market supply and demand: participants’ competitions and customers’ valuations

Typical Pricing Scenarios in Literature

1

Pricing General Data Sets

Pricing data sets as *indivisible* units in a *monopoly* market

2

Pricing Crowd-sensing Data

Pricing *indivisible* data sets in a *competitive* market

3

Pricing Data Queries

Data consumers can purchase just a *subset* of an entire data set

4

Compensating Privacy Loss

Pricing *personal data* by privacy compensation

Outline: Pricing Raw Data Sets

- Introduction
- Pricing General Data Sets
- Pricing Crowd-sensing Data
- Pricing Data Queries
- Compensating Privacy Loss
- Summary

Pricing General Data Sets

- Linear model: price $= \text{Fixed cost} + \sum_i w_i \cdot \text{factor}_i$
- Two level optimization model for revenue maximization
  - Different versions are constructed by different data quality factors
  - Customers' demands for different versions are public
  - Both the data seller and customers want to maximize their utility
  - The problem is a bi-level programming model, which is NP-hard

Heckman, Judd Randolph, et al. "A pricing model for data markets." *iConference 2015 Proceedings* (2015).

Yu, Haifei, and Mengxiao Zhang. "Data pricing strategy based on data quality." *Computers & Industrial Engineering* 112 (2017): 1-10.

Outline: Pricing Raw Data Sets

- Introduction
- Pricing General Data Sets
- Pricing Crowd-sensing Data
- Pricing Data Queries
- Compensating Privacy Loss
- Summary

Crowd-sensing Systems

- A task requester initiates a data collection task and compensates participating workers according to their reported costs
- Workers may exaggerate their costs to manipulate the market
- A truthful market is assumed

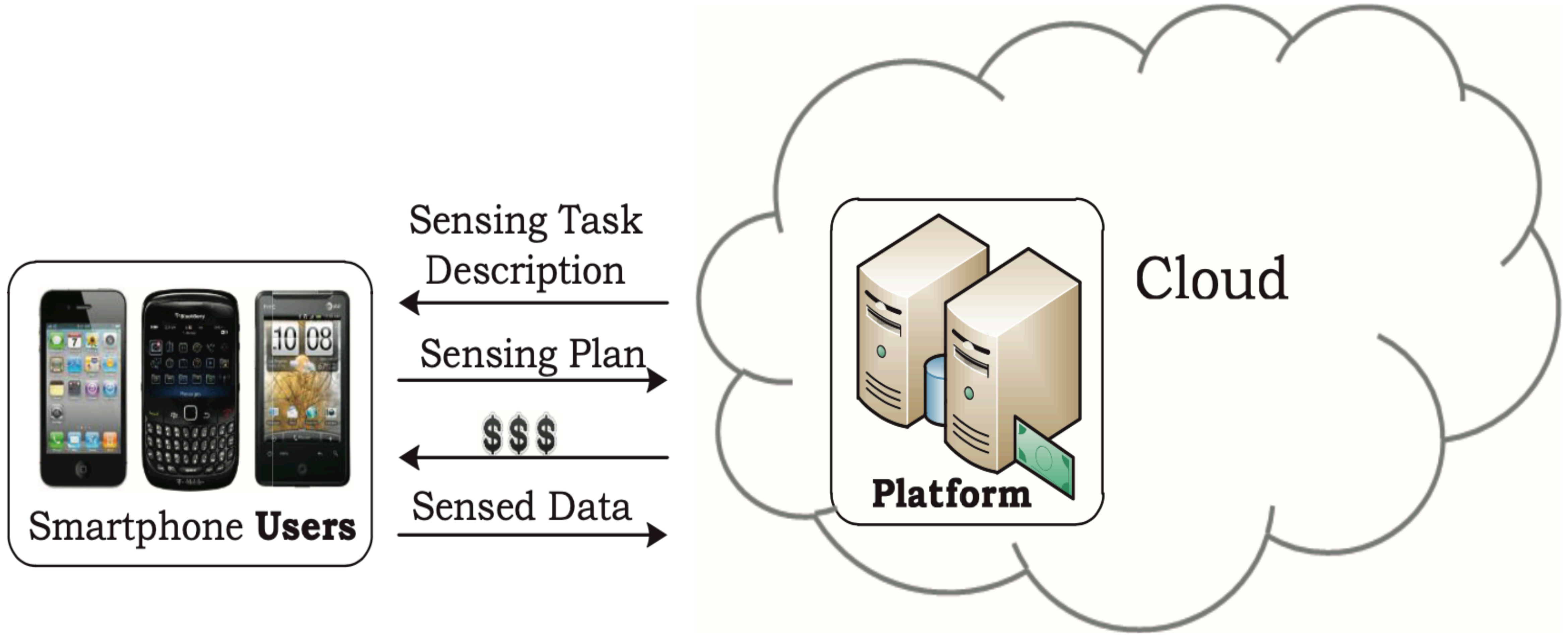
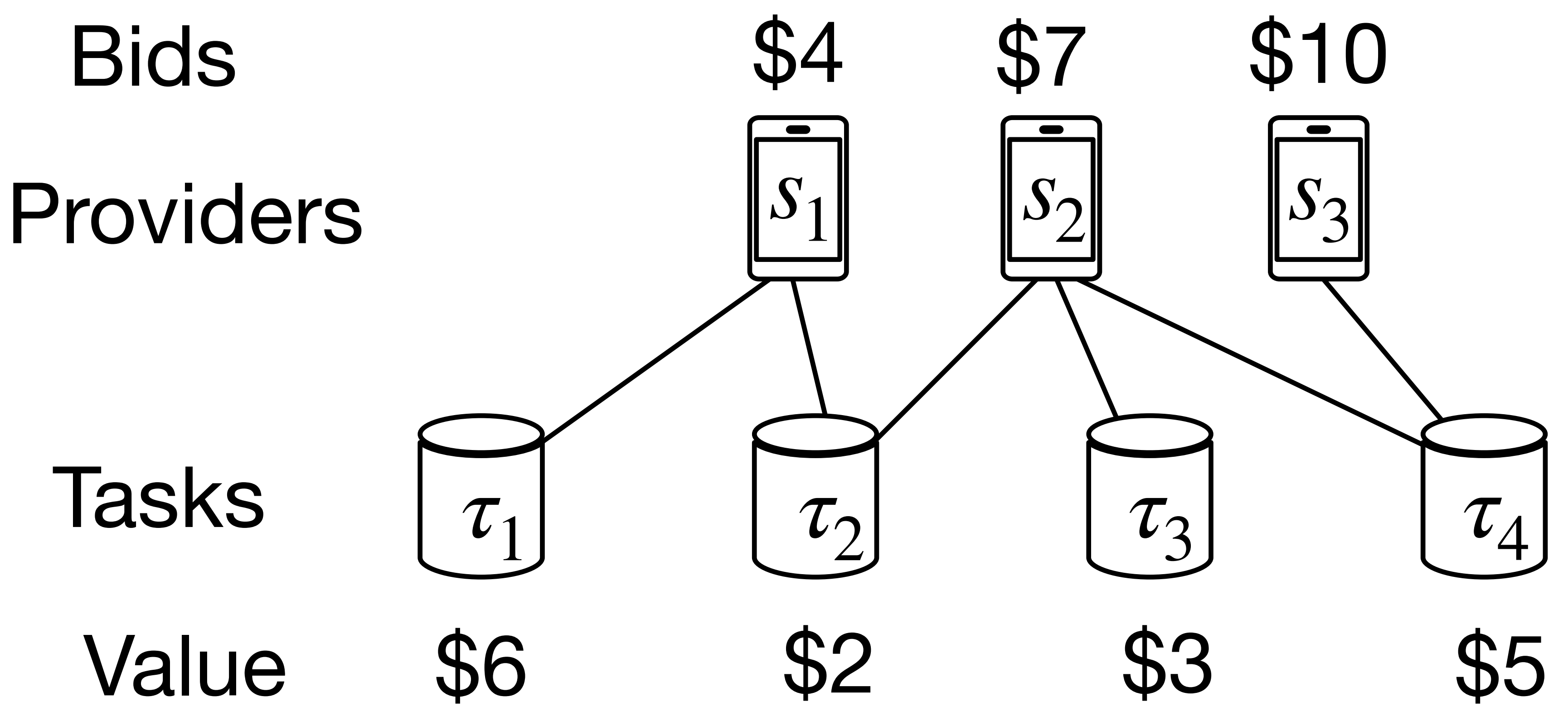


Figure from [Yang, Dejun, et al., 2012]
Example: a crowd-sensing system

Yang, Dejun, et al. "Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing." *Proceedings of the 18th annual international conference on Mobile computing and networking*. 2012.

Truthful Crowd-sensing Marketplaces

- A buyer has a set $\Gamma = \{\tau_1, \ldots, \tau_n\}$ of sensing tasks, where a task τ_i has a value v_i
- Each seller s_i chooses to perform a subset of tasks $\Gamma_i \subseteq \Gamma$ and has a private cost c_i
- The task-bid pair (Γ_i, b_i) is submitted to the buyer, where b_i is s_i's asking price for performing the tasks Γ_i
  - The asking price b_i can be greater than the true cost c_i
- Design an auction that is truthful and all participants have non-negative utilities



Yang, Dejun, et al. "Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing." *Proceedings of the 18th annual international conference on Mobile computing and networking*. 2012.

Truthful Crowd-sensing Marketplaces

- Determine auction winners
  - Select winners in a greedy way by iteratively choosing the seller that brings the largest non-negative net marginal profit to the buyer
- Determine payments to winners
  - Each winner s_i is paid his/her critical value (seller s_i would not win the auction if s_i bids higher than his/her critical value)
- Sellers achieve highest expected profits by bidding truthfully
  - Truthful bidding: for each seller s_i, $b_i = c_i$

Yang, Dejun, et al. "Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing." *Proceedings of the 18th annual international conference on Mobile computing and networking*. 2012.

Data Quality Aware Truthful Crowd-sensing Marketplaces

- Assume data quality q_i of each seller s_i is public and q_i is the same for all sensing tasks
- Each sensing task t_j has a data quality requirement Q_j, that is, $\sum_{s_i \in S, \text{ if } s_i \text{ performs } t_j} q_i \geq Q_j$
- Select sellers to maximize total utility of all participants (social welfare) under data quality constraints
- A greedy algorithm with a guaranteed approximation ratio is proposed
  - First, all sellers with positive social welfare contributions are selected
  - Then, select sellers with negative social welfare contributions greedily to fulfill data quality constraints
  - Critical payment is made to each winner

Jin, Haiming, et al. "Quality of information aware incentive mechanisms for mobile crowd sensing systems." *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 2015.

Outline: Pricing Raw Data Sets

- Introduction
- Pricing General Data Sets
- Pricing Crowd-sensing Data
- Pricing Data Queries
- Compensating Privacy Loss
- Summary

Charging Customers by Data Queries

- Customers can purchase their interested parts of a data set through data queries
- Arbitrage allows buyers to obtain a query result in a cost less than the advertised price

Name	Gender	Age
John	M	25
Alice	F	13
Bob	M	45
Anna	F	19

- Q_1 = SELECT count(*) FROM User WHERE Gender='F'
- Q_2 = SELECT Gender, count(*) FROM User
- Q_1 is sold for $7 and Q_2 is sold for $5

Arbitrage-free Pricing of Data Queries

- Given a database D, a multi-set of query bundles $\mathbf{S} = \{\mathbf{Q}_1, \ldots, \mathbf{Q}_m\}$ is said to determine a query bundle $\mathbf{Q}$, if the answer to $\mathbf{Q}$ can be computed only from the answers to the query bundles in $\mathbf{S}$

- A pricing function is arbitrage-free if the advertised price satisfies

$$\pi(\mathbf{Q}) \leq \sum_{i=1}^{m} \pi(\mathbf{Q_i})$$

View-Based Pricing

- The seller determines the price of a few views $\mathbf{V}$ over a database, then the price of a query bundle $\mathbf{Q}$ is decided algorithmically
- The query price $\pi(\mathbf{Q})$ is the total price of the cheapest subset of $\mathbf{V}$ that determines $\mathbf{Q}$
- Computing the price function is NP-hard for general conjunctive queries
  - Polynomial time algorithms for chain queries and cyclic queries are proposed
  - Example chain query $Q(x, y) = R(x) \bowtie S(x, y) \bowtie T(y)$
  - Example cyclic query $Q(x, y, z) = S(x, y) \bowtie B(y, z) \bowtie C(z, x)$

Koutris, Paraschos, et al. "Query-based data pricing." *Journal of the ACM (JACM)* 62.5 (2015): 1-44.

QueryMarket: Prototype of View-based Pricing

- Formulate the pricing model as an integer linear program (ILP) with the objective to minimize the total cost of purchased views $\mathbf{V}_p$
- A large class of queries can be priced efficiently in practice
- Constraints of the ILP
  - For a tuple $t \in Q(D)$
    - For each relation R in Q, at lease one view on R should be purchased
    - $\exists V' \subseteq V_p$ that can produce t
  - For a tuple $t \notin Q(D)$, $\exists V' \subseteq V_p$ that can indicate $t \notin Q(D)$

Koutris, Paraschos, et al. "Toward practical query pricing with querymarket." *proceedings of the 2013 ACM SIGMOD international conference on management of data*. 2013.

QueryMarket: An Example

Query: $Q(x, y) = R(x) \bowtie S(x, y)$

Table R

A
a_1

Table S

A	B
a_1	b
a_2	b

- Objective: Minimize the total price of purchased views
- subject to

$$
\begin{array}{llll}
(a_1, b) \in Q & x[R.A = a_1] & & \geq 1 \\
 & x[S.A = a_1] & + \; x[S.B = b] & \geq 1 \\
(a_2, b) \notin Q & & x[R.A = a_2] & \geq 1
\end{array}
$$

Binary variable indicating whether a view is purchased

One view from each relation in $\mathcal{Q}$

This view indicates $(a_2, b) \notin Q$

Koutris, Paraschos, et al. "Toward practical query pricing with querymarket." *proceedings of the 2013 ACM SIGMOD international conference on management of data*. 2013.

Arbitrage-free Pricing of Linear Aggregate Queries

- A linear query over real-valued data set $\mathbf{x} = \langle x_1, \ldots, x_n \rangle$ is a real-valued vector $\mathbf{q} = \langle w_1, \ldots, w_n \rangle$, and the answer is $\mathbf{q}(\mathbf{x}) = \sum_{i=1} w_i x_i$

- Unbiased estimator of $\mathbf{q}(\mathbf{x})$ is traded and priced based on variance v
  - v trades off between data accuracy and query price

- Arbitrage example
  - Q_1 and Q_2 are sold for \$5 and \$20, respectively
  - $Q_1 = (\mathbf{q}, v), Q_1 = (\mathbf{q}, v) \rightarrow Q_2 = (\mathbf{q}, v/2)$

Li, Chao, et al. "A theory of pricing private data." *ACM Transactions on Database Systems (TODS)* 39.4 (2014): 1-28.

Arbitrage-free Pricing of Linear Aggregate Queries

- An arbitrage-free pricing function must satisfy $\pi(\mathbf{q}, v) = \Omega(\frac{1}{v})$
- Basic arbitrage-free function: $\pi(\mathbf{q}, v) = \frac{f^2(\mathbf{q})}{v}$, where the function $f(\cdot)$ is semi-norm
  - E.g. $\pi(\mathbf{q}, v) = \frac{|\mathbf{q}|_\infty^2}{v} = \frac{max_i q_i^2}{v}$
- Composition $\pi(\mathbf{q}, v) = f(\pi_1(\mathbf{q}, v), \ldots, \pi_k(\mathbf{q}, v))$ of arbitrage-free functions $\pi_1, \ldots, \pi_k$ is still arbitrage-free if $f(\cdot)$ is subadditive and nondecreasing
  - E.g. $f(\pi_1, \pi_2) = \sqrt{\pi_1 * \pi_2}$

Li, Chao, et al. "A theory of pricing private data." *ACM Transactions on Database Systems (TODS)* 39.4 (2014): 1-28.

Arbitrage-free Pricing for General Queries

- Three types of pricing models for query bundles
  - Instance-independent pricing: the price depends only on the query
  - Answer-dependent pricing: the price depends on the query and the query output
  - Data-dependent pricing: the price depends on the query and the database instance

Lin, Bing-Rong, and Daniel Kifer. "On arbitrage-free pricing for general data queries." *Proceedings of the VLDB Endowment* 7.9 (2014): 757-768.

Five Types of Arbitrage for General Queries (1)

- Price-based arbitrage: if prices are quoted by queries, a buyer may deduce answers to queries from prices along

  Q = SELECT a, T.b, c FROM T, R WHERE T.b=R.b AND a=1 AND b=2 AND c=3

  - Let $\pi(T)$ and $\pi(R)$ be the price of the whole tables T and R, respectively
  - In view-based pricing, $\pi(Q) = \pi(T) + \pi(R)$ if and only if the answer to Q is not empty
  - Customer can infer that the tuple $(1,2,3)$ is in the join of T and R by checking the price

Lin, Bing-Rong, and Daniel Kifer. "On arbitrage-free pricing for general data queries." *Proceedings of the VLDB Endowment* 7.9 (2014): 757-768.

Five Types of Arbitrage for General Queries (2)

- Separate-account arbitrage: a buyer may use multiple accounts to derive answers to a query bundle
  - Recall the arbitrage example in linear aggregate query
- Almost-certain arbitrage: two queries have almost identical answers but their prices are dramatically different
  - Consider a query asking the population of Canada
  - π(an answer of a variance 1)=$10,000
  - π(an answer of a variance 1.1)=$1

Lin, Bing-Rong, and Daniel Kifer. "On arbitrage-free pricing for general data queries." *Proceedings of the VLDB Endowment* 7.9 (2014): 757-768.

Five Types of Arbitrage for General Queries (3)

- Post-processing arbitrage: if the answers to a query bundle $\mathbf{Q}$ can always be deduced from the answers to another query bundle $\mathbf{Q}'$, the price of $\mathbf{Q}'$ should be no cheaper than that of $\mathbf{Q}$
    - Q_1 = SELECT * FROM T WHERE g="F" $\rightarrow$ Q_2 = SELECT count(*) FROM T WHERE g="F"
- Serendipitous arbitrage: for a specific database instance, the answers to $\mathbf{Q}$ may be derived from the answers to $\mathbf{Q}'$
    - Assume that table T does not have any records with g="F"
    - $Q_2 \rightarrow Q_1$

Lin, Bing-Rong, and Daniel Kifer. "On arbitrage-free pricing for general data queries." *Proceedings of the VLDB Endowment* 7.9 (2014): 757-768.

Qirana: Efficient and Scalable Pricing

- Compute the price of a query bundle $\mathbf{Q}$ from the view of uncertainty reduction
- Denote by S a set of all possible database instances with the same schema as the true database instance D
- The buyer can rule out database instances $D_i \in S$ that cannot be D by checking whether $\mathbf{Q}(D_i) = E$
- Arbitrage-free pricing function should be monotone and subadditive with respect to how much S shrinks

Deep, Shaleen, and Paraschos Koutris. "QIRANA: A framework for scalable query pricing." *Proceedings of the 2017 ACM International Conference on Management of Data*. 2017.

Qirana: Efficient and Scalable Pricing

- Denote by $C_{\mathbf{Q}} = \{D_i \in S \,|\, \mathbf{Q}(D) \neq \mathbf{Q}(D_i)\}$ the set of ruled out database instance

$$\pi(\mathbf{Q}) = \sum_{D_i \in C_{\mathbf{Q}}} w_i, \text{ where } w_i \text{ is the weight of } D_i$$

- The weights could be manually set by the buyer or learned from exemplar queries and their prices

---

- $\mathbf{Q}$ = SELECT count(*) FROM User WHERE Gender = “F”
- Table 1 and Table 2 are ruled out, thus $\pi(\mathbf{Q}) = w_1 + w_2$

Name	Gender
John	M
Alice	F
Bob	M
Anna	F

True Table

Name	Gender
John	F
Alice	F
Bob	M
Anna	F

Table 1

Name	Gender
John	M
Alice	M
Bob	M
Anna	F

Table 2

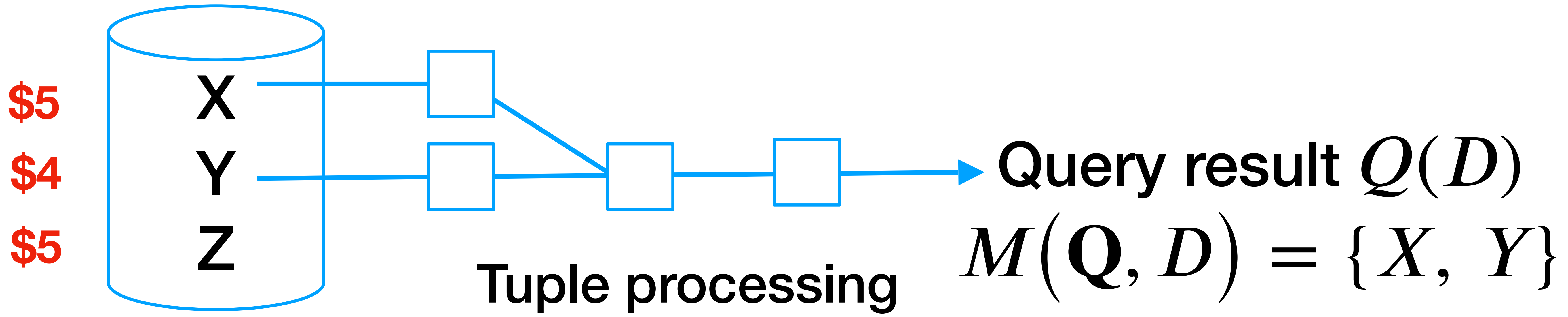
Name	Gender
John	M
Alice	F
James	M
Anna	F

Table 3

Deep, Shaleen, and Paraschos Koutris. "QIRANA: A framework for scalable query pricing." *Proceedings of the 2017 ACM International Conference on Management of Data*. 2017.

Query Pricing based on Query Lineage

- Price selection-projection-natural join queries over incomplete databases
- The lineage tuples $M(\mathbf{Q}, D)$ is the set of tuples in the database D that contribute to $\mathbf{Q}(D)$
- Each tuple in D has a price, which is proportional to the completeness of the tuple
- Query price $\pi(\mathbf{Q})$ is the total price of the tuples in $M(\mathbf{Q}, D)$



Miao, Xiaoye, et al. "Towards Query Pricing on Incomplete Data." *IEEE Transactions on Knowledge and Data Engineering* (2020).

Revenue Maximization in Query-based Pricing

- A buyer is single-minded if the buyer wants to purchase the answer to a single set of queries
- A buyer purchases $\mathbf{Q}$ if the advertised price $\pi(\mathbf{Q})$ is smaller than or equal to the buyer's valuation
- Follow the idea in Qirana, which prices $\mathbf{Q}$ as a bundle of items
- Uniform bundle pricing: set the same price for all queries
- The additive/item pricing: set a weight for each item and $\pi(\mathbf{Q})$ is the total weights of the items in the bundle
- XOS pricing: set k weights $w_i^1, \ldots, w_i^k$ for each item D_i
    - The price of $\mathbf{Q}$ is $\pi(\mathbf{Q}) = \max_{j=1}^{k} \sum_{D_i \in S, \mathbf{Q}(D) \neq \mathbf{Q}(D_i)} w_i^j$

Chawla, Shuchi, et al. "Revenue maximization for query pricing." *Proceedings of the VLDB Endowment* 13.1 (2019): 1-14.

Bounds on Revenue Maximization

•Cheung, Maurice, and Chaitanya Swamy. "Approximation algorithms for single-minded envy-free profit-maximization problems with limited supply." *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, 2008.

•Chawla, Shuchi, et al. "Revenue maximization for query pricing." *Proceedings of the VLDB Endowment* 13.1 (2019): 1-14.

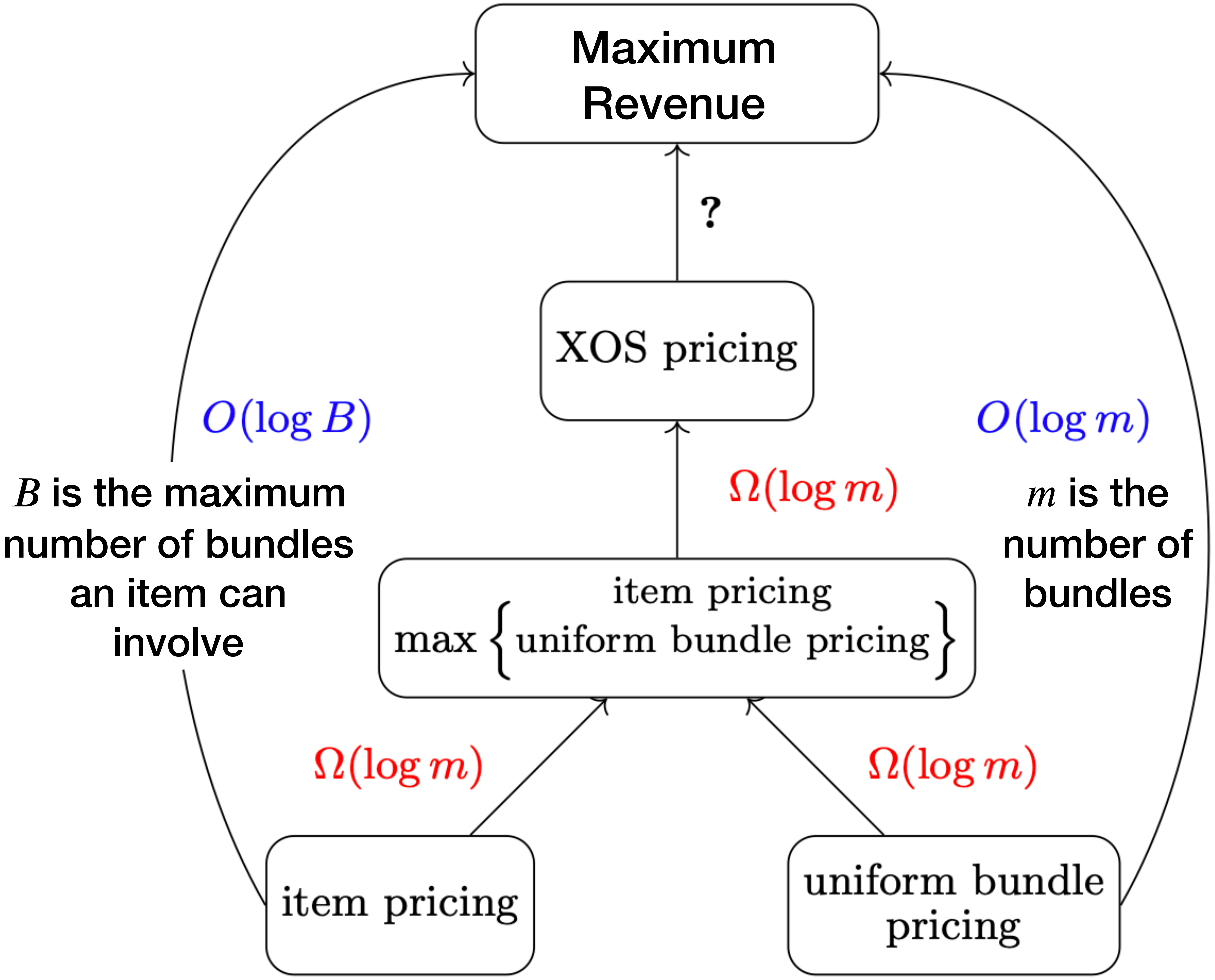


Figure from [Chawia, Shuchi, et al., 2018]

History Aware Pricing

- A history-aware pricing function does not charge the customer twice for already purchased information
- QueryMarket tracks the purchased views of a customer and avoids charging those views when pricing future queries of the customer
- Allow buyers to ask for refunds of already purchased data
  - An identifier (coupon) for each tuple in the query answer $\mathbf{Q}(D)$, which records the identity information of a tuple
  - If the buyer receives the same tuple t from two queries, the buyer can ask for a refund of t by presenting the two coupons associated with t in the two corresponding queries
  - No arbitrage-free guarantee

Koutris, Paraschos, et al. "Toward practical query pricing with querymarket." *proceedings of the 2013 ACM SIGMOD international conference on management of data*. 2013.

Upadhyaya, Prasang, Magdalena Balazinska, and Dan Suciu. "Price-optimal querying with data apis." *Proceedings of the VLDB Endowment* 9.14 (2016): 1695-1706.

Outline: Pricing Raw Data Sets

- Introduction
- Pricing General Data Sets
- Pricing Crowd-sensing Data
- Pricing Data Queries
- Compensating Privacy Loss
- Summary

Differential Privacy

- Differential privacy provides privacy protection by injecting controlled random noise into a data set

- Two data sets D and D' are neighboring datasets if they differ in one element

- A is an algorithm that returns noisy query answers over a data set

- A is ϵ-differential private if and only if for any two neighbouring data sets D and D'

$$\exp(-\epsilon) \leq \Pr(\frac{A(D) = y}{A(D') = y}) \leq \exp(\epsilon)$$

- An adversary cannot distinguish between D and D' only from the query answers

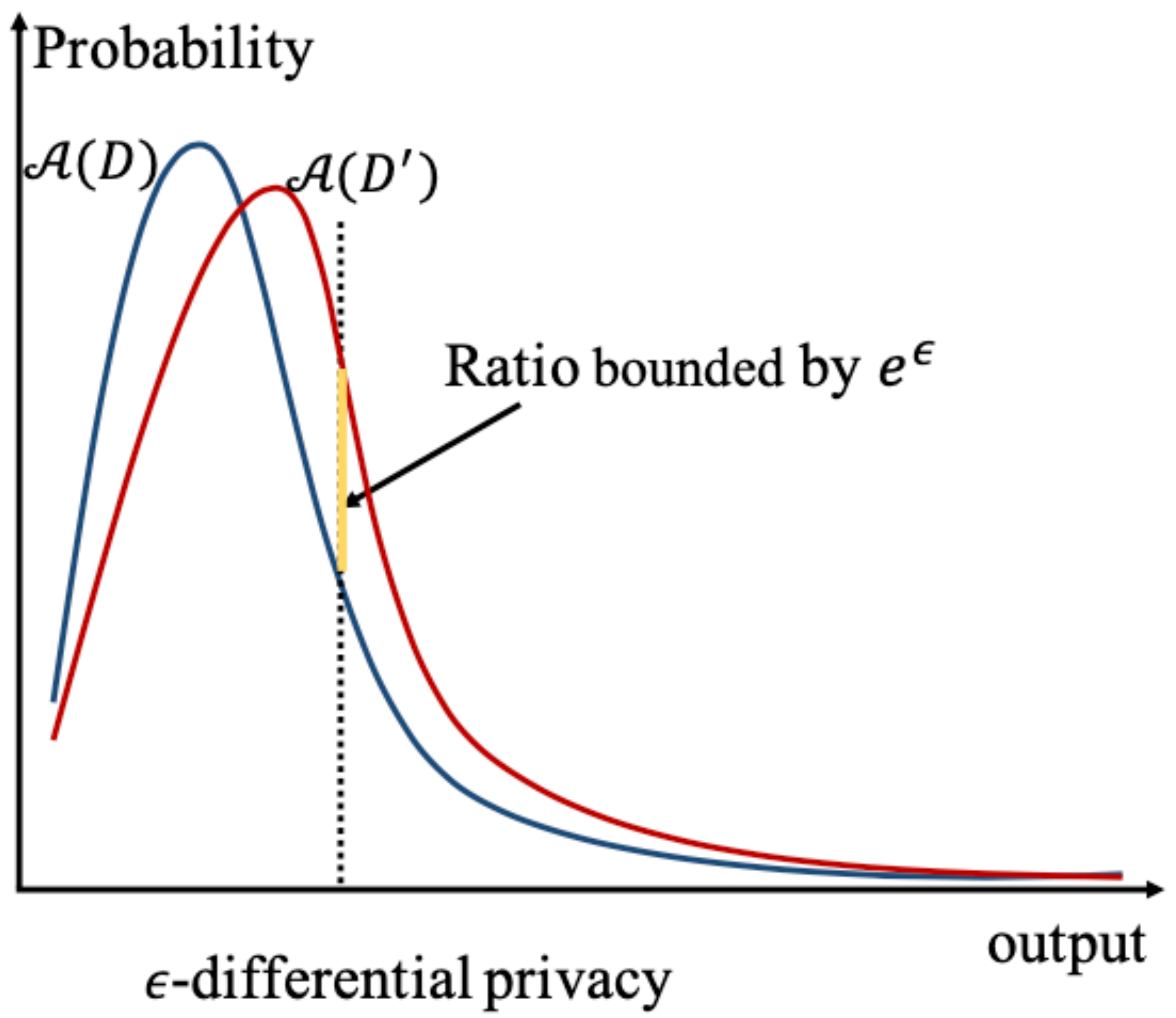


Figure from [Jiang, Honglu, et al., 2020]

Jiang, Honglu, et al. "Differential privacy and its applications in social network analysis: A survey." *arXiv e-prints* (2020): arXiv-2010.

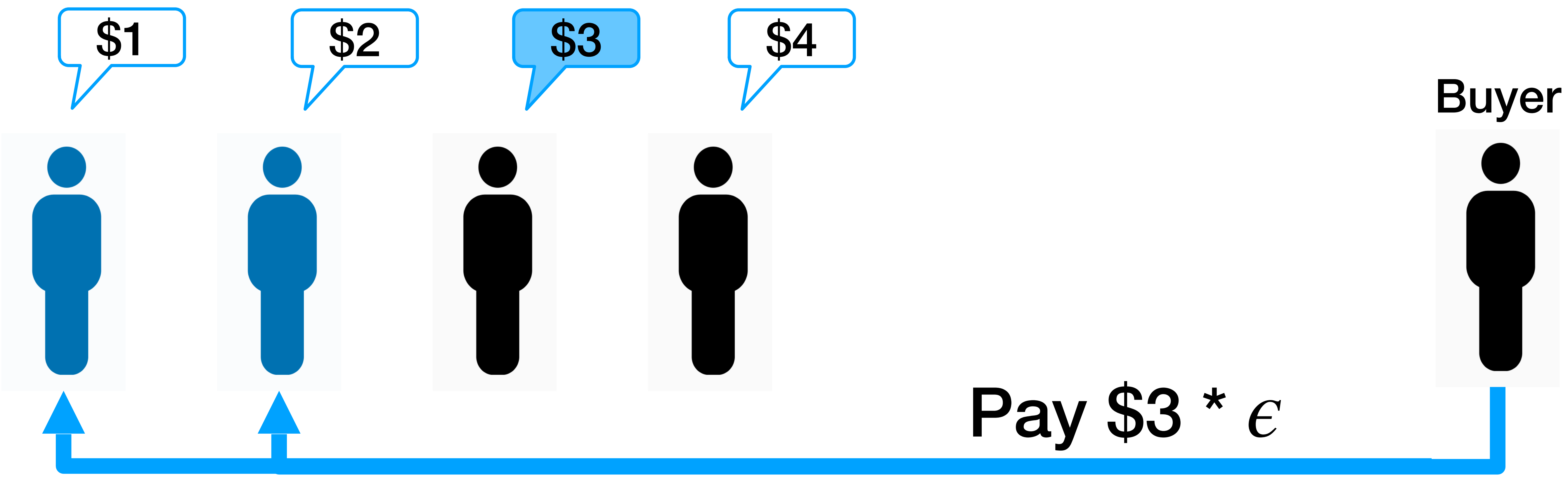
Privacy Compensation in Data Market

- Private data has values
  - A unique user values \$4 to Facebook and \$24 to Google
- Differential privacy plays an essential role in personal data pricing
- The magnitude of injected random noise impacts data providers' privacy loss ϵ, data set usability, and the data price

Li, Chao, et al. "A theory of pricing private data." *ACM Transactions on Database Systems (TODS)* 39.4 (2014): 1-28.

A Truthful and Privacy Preserving Marketplace

- Only need to purchase data from m individuals and use them in an ϵ-differential privacy manner, where m and ϵ only depend on the accuracy goal
- Transform the problem into variants of multi-unit procurement auction. The classic Vickrey auction minimizes the buyer's payment and guarantees the accuracy goal
    - Vickrey auction (second-price sealed-bid auction): every bidder submits a bid without knowing others' bids. The bidder making the highest bid wins and pays only the second highest bid
- Negative result: may not work well if the value of personal data and privacy valuation may be correlated



Top-m smallest bids win the auction and the winners are paid by the $m+1$ smallest bid

Ghosh, Arpita, and Aaron Roth. "Selling privacy at auction." *Proceedings of the 12th ACM conference on Electronic commerce*. 2011.

Pricing Linear Aggregate Queries by Auction

- A data buyer wants to purchase an estimator of a linear aggregate queries $\mathbf{q} = \langle w_1, \ldots, w_n \rangle$ over real-valued personal data

- Minimize the expected squared error of the returned estimator with respect to the buyer's budget

  - Need to maximize $\sum |w_i| x_i$, where $x_i \in \{0,1\}$ indicates whether provider i is used

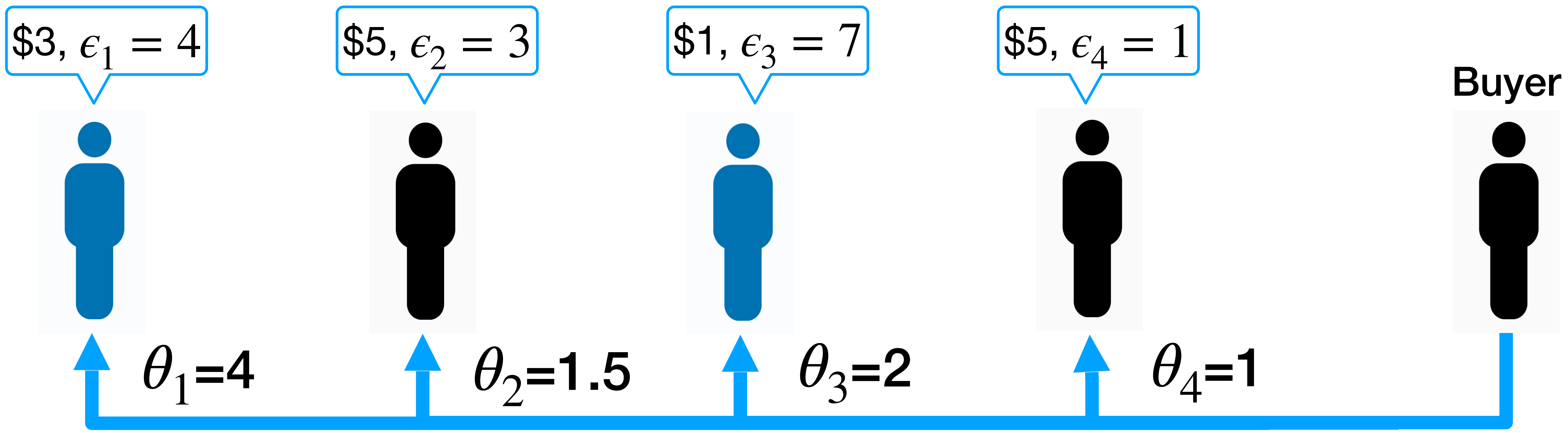
- Transform to a variant of knapsack reverse auction

  - Budget, compensation to a provider p_i, and w_i are regarded as knapsack capacity, item value, and item weight, respectively

Dandekar, Pranav, Nadia Fawaz, and Stratis Ioannidis. "Privacy auctions for recommender systems." *ACM Transactions on Economics and Computation (TEAC)* 2.3 (2014): 1-22.

Pricing Under the Maximal Privacy Loss Constraint

- Estimators of linear aggregate queries over real-valued personal data are traded
- Each data provider i can specify the personal maximum tolerable privacy loss ϵ_i
- Assume that the distribution of privacy cost of each provider is public
- Transform to Bayesian optimal knapsack procurement
- The data of each selected provider i is used in ϵ_i-differential privacy manner



A take-it-or-leave-it price θ_i is computed for each provider i

Zhang, Mengxiao, Fernando Beltran, and Jiamou Liu. "Selling Data at an Auction under Privacy Constraints." *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020.

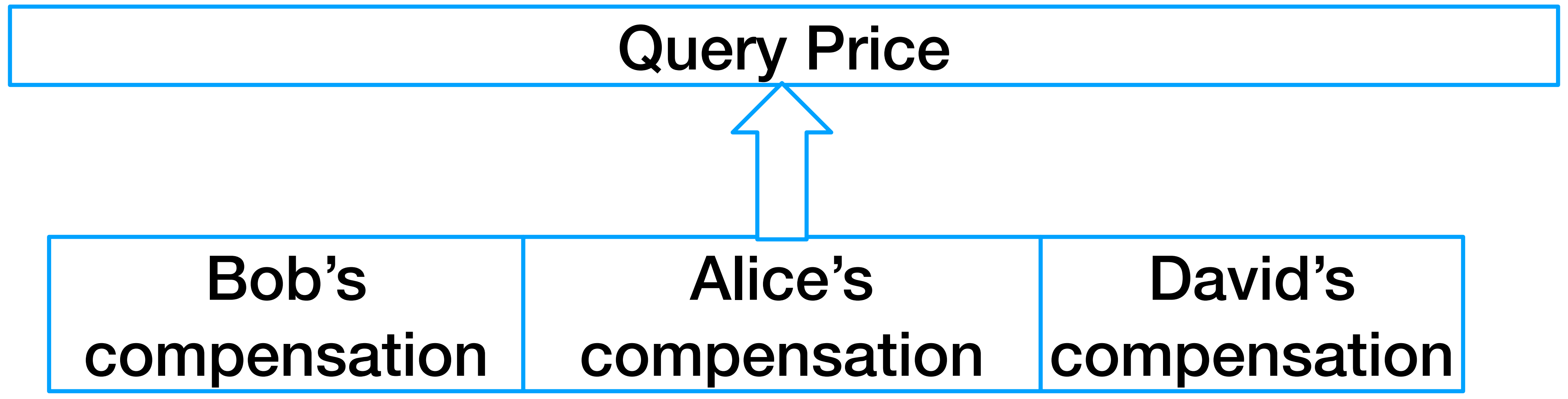
Privacy Compensation in Arbitrage-free Pricing

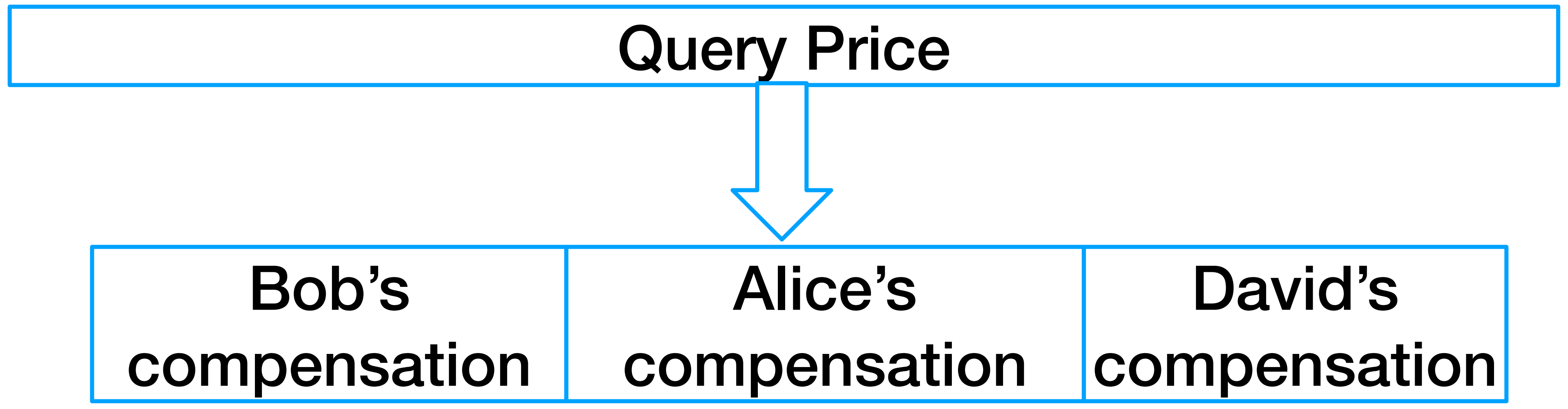
- A linear aggregate query $\mathbf{Q} = (\mathbf{q}, v)$ is traded under differential privacy, where v is defined by the buyer
- Laplace noise with variance $\sqrt{\frac{v}{2}}$ is added for privacy protection
- The privacy loss of an individual s_i is upper-bounded by $\epsilon = \frac{\text{max}_i \mathbf{q}_i}{\sqrt{\frac{v}{2}}}$
- Provider s_i receives a compensation $p_i(\epsilon) = c_i \epsilon$, where c_i is the unit privacy cost of s_i
- The price of a query is the sum of the privacy compensations, which is arbitrage-free

Li, Chao, et al. "A theory of pricing private data." *ACM Transactions on Database Systems (TODS)* 39.4 (2014): 1-28.

Compensating Correlated Private Data

- Two individuals' data may be correlated, the privacy of a not-involved individual may be leaked due to the revelation of the other individual's data
- The privacy loss of a data provider s_i caused by a query is upper-bounded by $\epsilon_i = \frac{ds_i}{\sqrt{\frac{v}{2}}}$, where ds_i is the dependent sensitivity of the query at provider i's data
- Propose bottom-up mechanism and a top-down mechanism to determine privacy compensations and query prices





Niu, Chaoyue, et al. "Unlocking the value of privacy: Trading aggregate statistics over private correlated data." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018.

Outline: Pricing Raw Data Sets

- Introduction
- Pricing General Data Sets
- Pricing Crowd-sensing Data
- Pricing Data Queries
- Compensating Privacy Loss
- Summary

Summary: Pricing Raw Data Sets

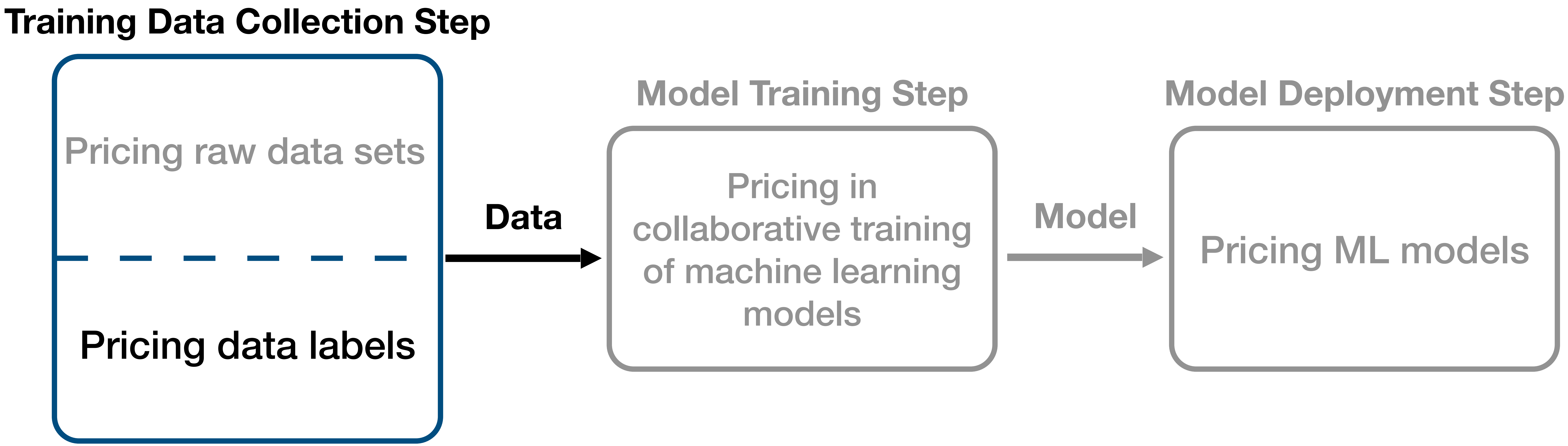
- The price of a data set is determined by both intrinsic and extrinsic factors
- Four typical pricing scenarios in existing studies
- Pricing models with different desiderata, namely revenue maximization, truthfulness, arbitrage-free, and privacy preservation
- **Limitation:** price of a data set is determined without considering the down-stream applications of the data set

Part IV:
Pricing Data Labels

Outline: Pricing Data Labels

- Introduction
- Gold Task Based Methods
- Peer Prediction Based Methods
- Summary

Pricing Data Labels in Machine Learning Pipelines



Crowdsourcing Data Labeling Tasks

- Crowdsourcing is a popular way for label collection
    - Tasks solved by workers recruited through the internet
- Quality control methods for collected labels
    - Filter, reputation, incentives, etc
    - Incentives: encourage participation and effort of good data providers by rigorously designed rewards
- How do we pay workers in proportion to their efforts?

Vaughan, Jennifer Wortman. "Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research." *J. Mach. Learn. Res.* 18.1 (2017): 7026-7071.

Faltings, Boi, and Goran Radanovic. "Game theory for data science: Eliciting truthful information." *Synthesis Lectures on Artificial Intelligence and Machine Learning* 11.2 (2017): 1-151.

Model of Workers

- Workers can have different behaviors
  - Heuristic behaviors: report a random label or a constant label
  - Truthful behaviors: perform accurate measurement and report truthfully
- Assume rational workers choose behaviors with the highest payoff
- Motivate workers to behave truthfully through payments

Faltings, Boi, and Goran Radanovic. "Game theory for data science: Eliciting truthful information." *Synthesis Lectures on Artificial Intelligence and Machine Learning* 11.2 (2017): 1-151.

Principle of Pricing Data Labels

- Reward workers based on consistency with a reference
  - Gold task-based methods: some tasks with ground-truth answers are used as reference
  - Peer prediction-based methods: the answers from peer workers are used as reference

Outline: Pricing Data Labels

- Introduction
- Gold Task Based Methods
- Peer Prediction Based Methods
- Summary

Pricing Binary Labels

- A worker has a private belief Pr$(y_t = l)$ about how likely the true label y_t of a task t is l
- Motivate workers to skip questions for which his/her confidence is lower than T
- No free lunch axiom
  - If all the answers attempted by the worker in the gold standard are wrong, then the payment is zero

$$\pi(u) = \beta \cdot \frac{1}{T^c} \cdot 1(r = 0)$$

Number of correct answers **Number of wrong answers**

Figure from [Shah, Nihar Bhadresh, et al., 2015]

Shah, Nihar Bhadresh, and Dengyong Zhou. "Double or nothing: Multiplicative incentive mechanisms for crowdsourcing." *Advances in neural information processing systems* 28 (2015): 1-9.

Pricing Multiple Labels

- Workers can select multiple answers $\hat{Y}$ to a question
- Assume a worker's beliefs for any label being the true label for a task lie in the set $\{0\} \cup (p, 1]$ for some (fixed and known) p
- Motivate workers to report all labels with positive confidences
- A worker u receives $\pi(u, t)$ for his answers $\hat{Y}$ to a question t

$$\pi(u, t) = (1 - p)^{|\hat{Y}|} \cdot 1(r = 0)$$

- Total payment to a worker is $\prod_{t \in G} \pi(u, t)$, where G is the set of gold tasks

Figure from [Shah, Nihar, et al., 2015]

Shah, Nihar, Dengyong Zhou, and Yuval Peres. "Approval voting and incentives in crowdsourcing." *International conference on machine learning*. PMLR, 2015.

Reduce the Number of Gold Tasks

- Gold task-based methods require a sufficient number of tasks to achieve good performance
  - Gold tasks are expensive to obtain
- Arrange the workers in a hierarchy
  - Every worker shares one common task with each of its children
- The answers from workers are used as pseudo gold tasks for workers in the next layer

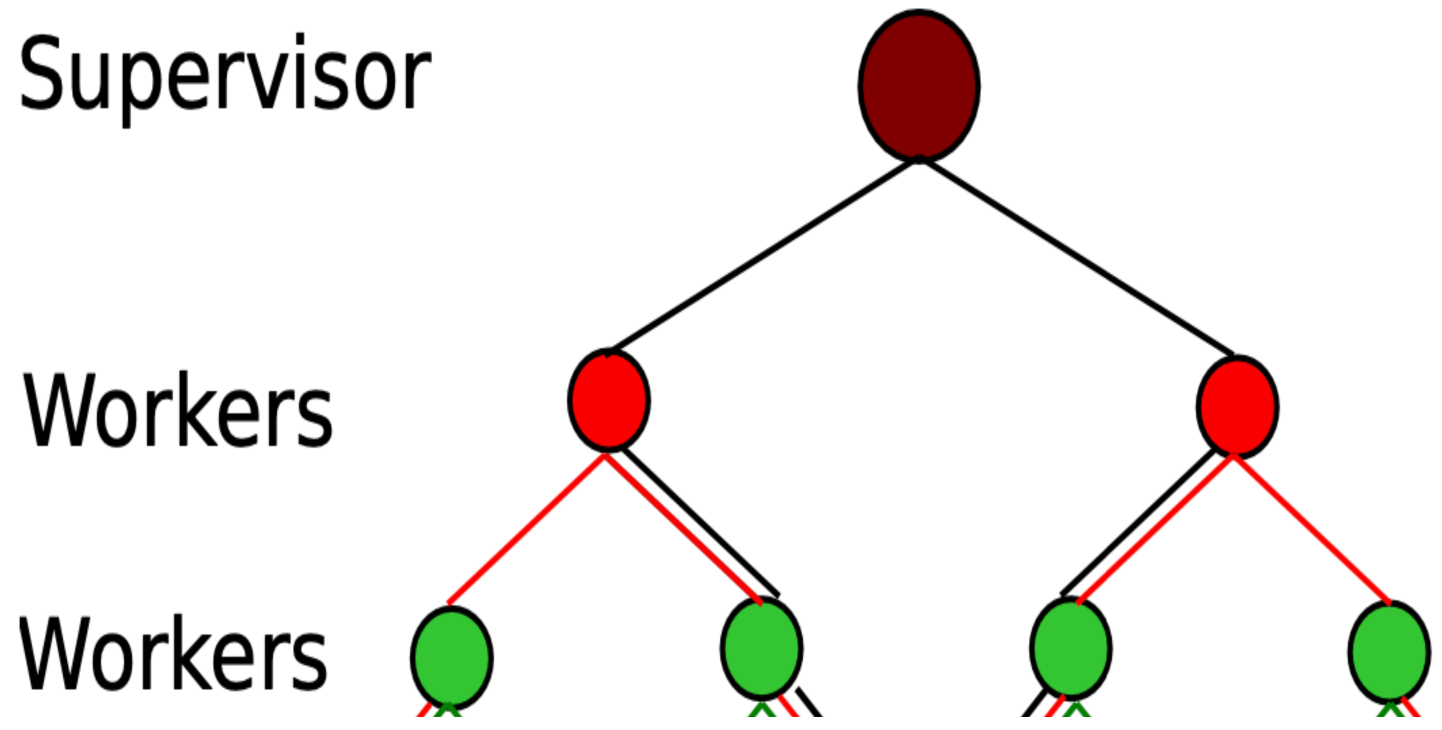
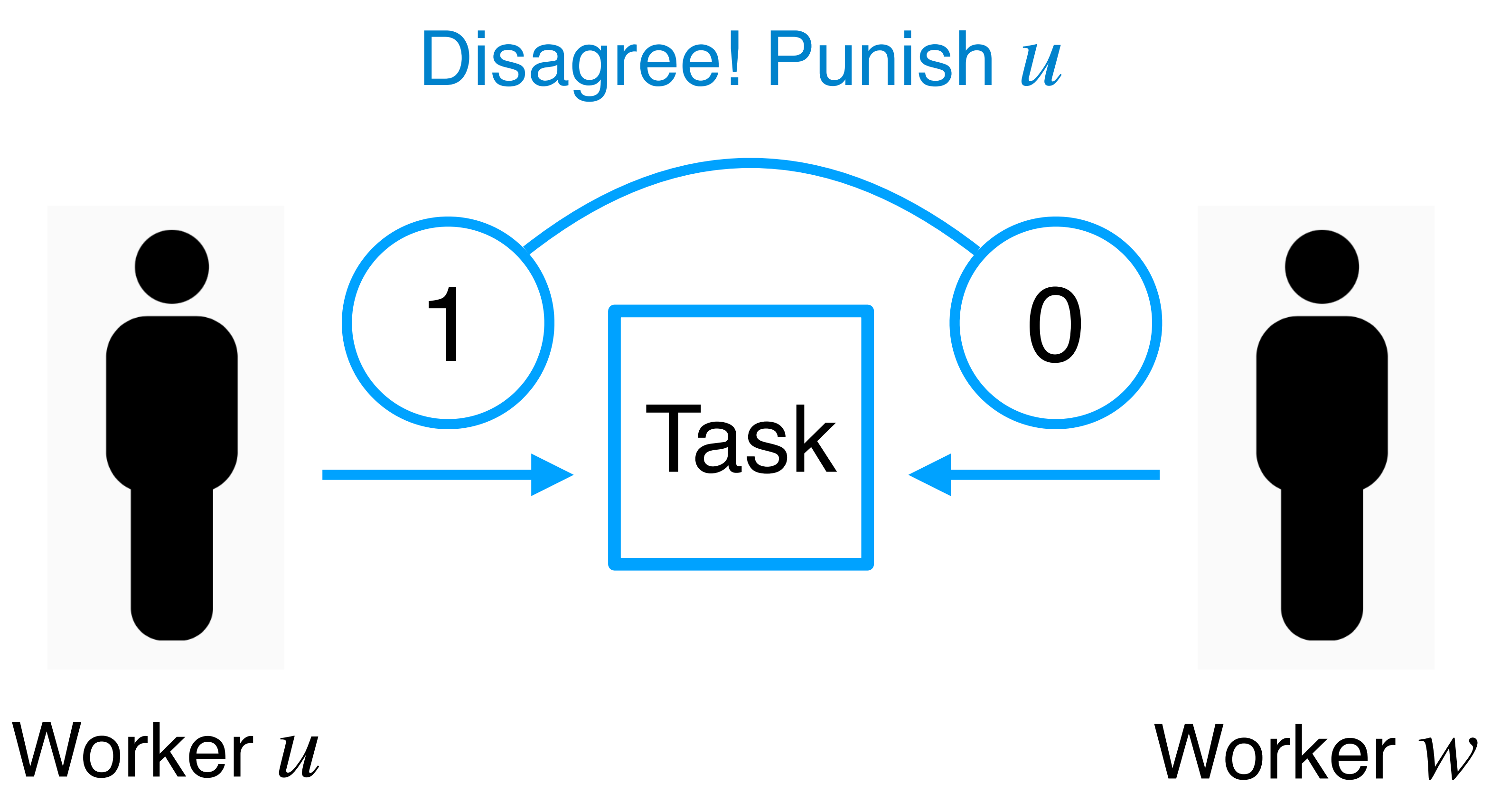


Figure from [de Alfaro, Luca, et al, 2016]

de Alfaro, Luca, et al. "Incentives for truthful evaluations." *arXiv preprint arXiv:1608.07886* (2016).

Reduce the Number of Gold Tasks

- A worker u needs to exert $f(e_u)$ efforts if u wants to achieve error rate e_u
- Motivate each worker u to exert enough efforts, such that $e_u \leq \epsilon$
- Worker u receives a penalty if u does not agree with the parent w on their shared task
- Worker u chooses error rates e_u^* to minimize his/her expected loss $e_u^* = \text{argmin } L(e_u, e_w)$
- If $e_w \leq \epsilon$, the optimization problem has a unique solution $e_u^* \leq \epsilon$
- All Nash equilibria guarantee that all workers exert enough efforts



de Alfaro, Luca, et al. "Incentives for truthful evaluations." *arXiv preprint arXiv:1608.07886* (2016).

Fair Performance Evaluation

- Fair evaluation: expected reward of a worker is directly proportional to the worker's proficiency
    - Proficiency matrix $T_p \in R^{K \times K}$ of a worker p: $T_p[l_k, l_j] = P(p \text{ report } l_j \mid \text{ ground−truth is } l_k)$
- Gold tasks are used to estimate the proficiency of a small group G of workers
- The answers by the small group of workers to non-gold tasks are used as contributed gold tasks, which are used to estimate the proficiency of more workers
- Payment is based on a worker's proficiency: $\pi(p) = \beta * (\sum_{g \in [K]} T_p[g, g] - 1)$
- Workers reporting random labels get zero payments in expectation

Goel, Naman, and Boi Faltings. "Deep bayesian trust: A dominant and fair incentive mechanism for crowd." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 2019.

Outline: Pricing Data Labels

- Introduction
- Gold Task Based Methods
- Peer Prediction Based Methods
- Summary

When No Ground-Truth Labels Are Available

- Peer prediction: evaluate consistency with peer reports
- Formulate a game among workers: reward of a worker depends on the reports of the worker and other workers
- Design the game such that
  - Exerting effort in solving the tasks can achieve high expected rewards
  - Spammers providing random answers on average receive no payments

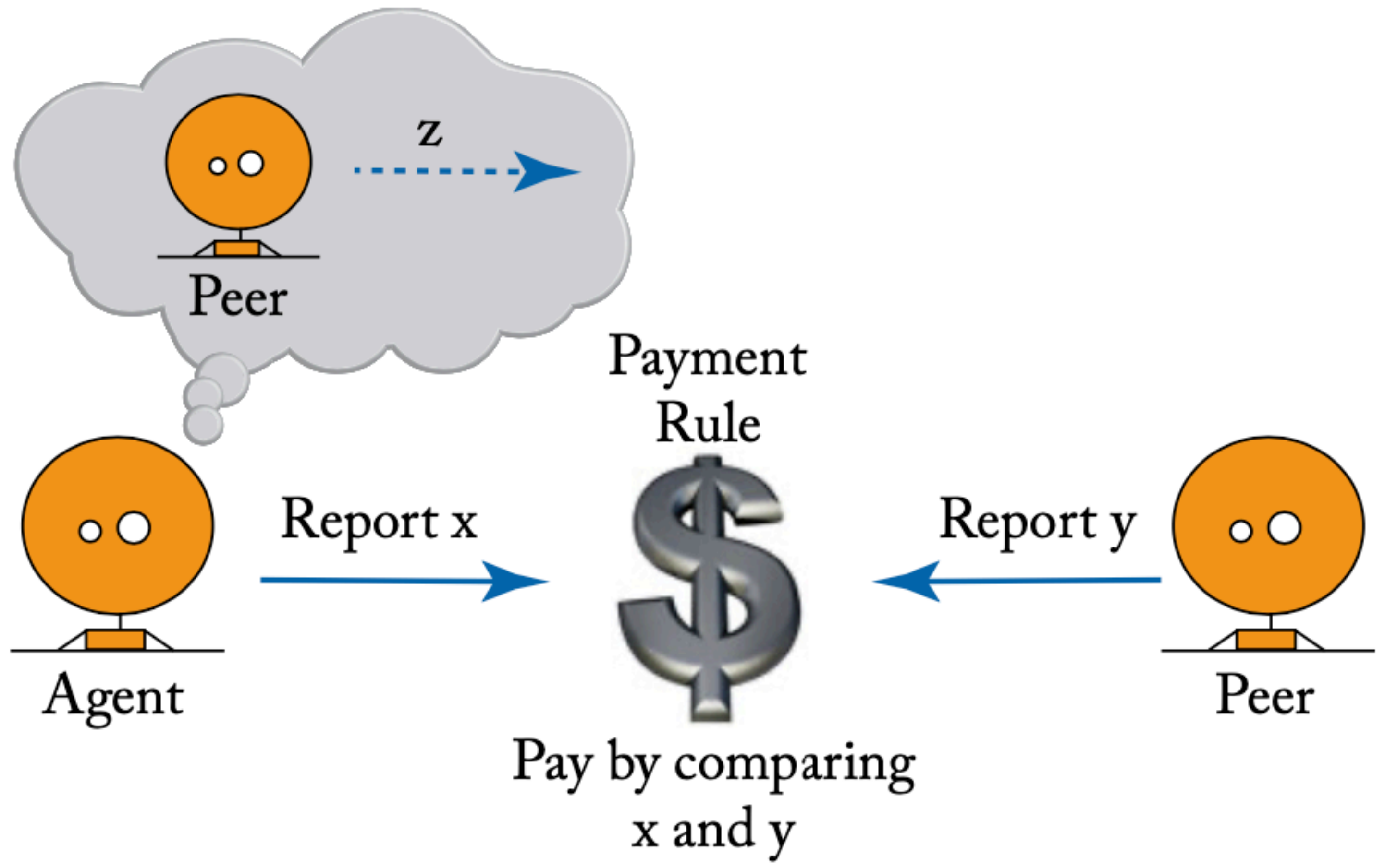
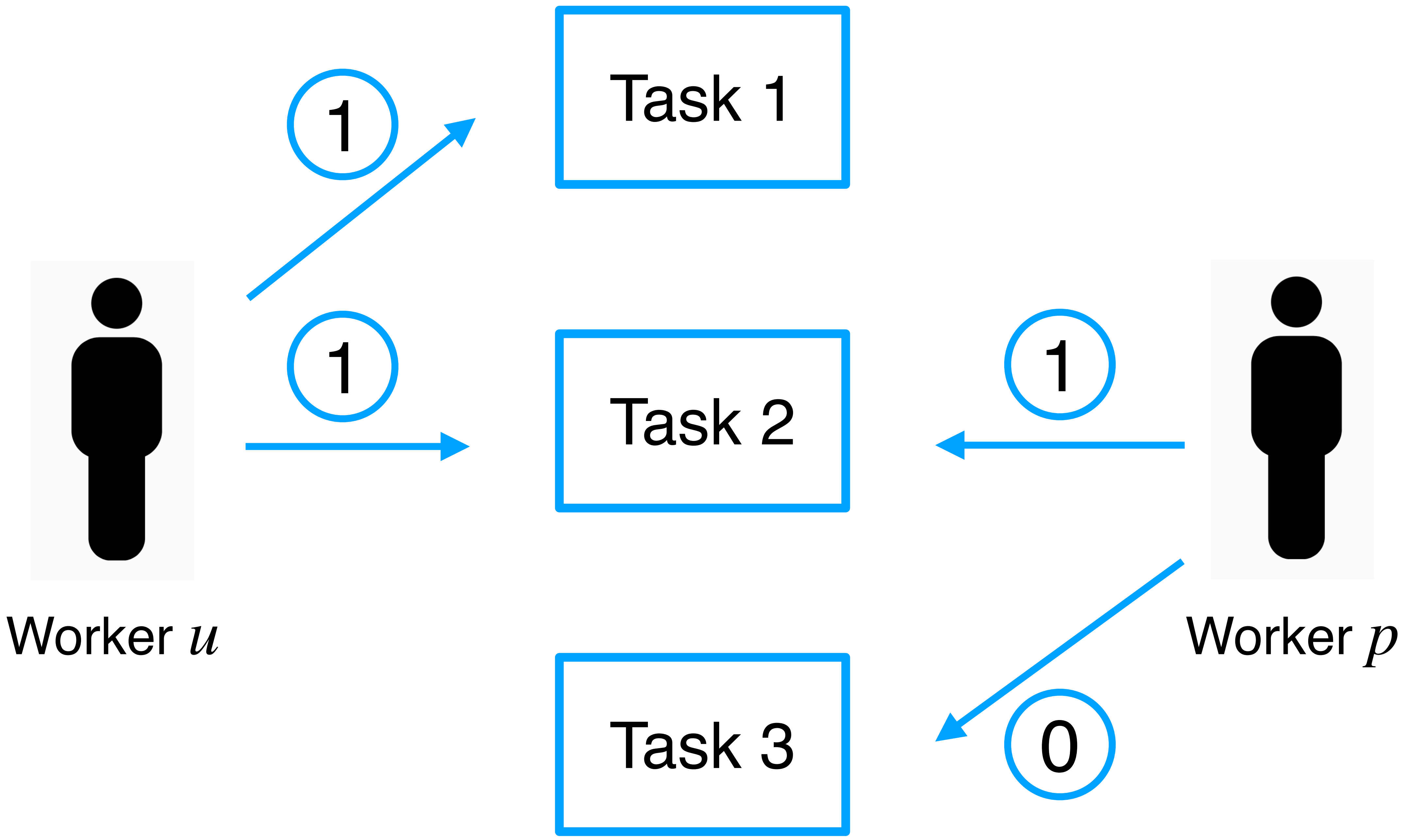


Figure from [Faltings, Boi, et al, 2017]

Faltings, Boi, and Goran Radanovic. "Game theory for data science: Eliciting truthful information." *Synthesis Lectures on Artificial Intelligence and Machine Learning* 11.2 (2017): 1-151.

Pricing Binary Labels (1)

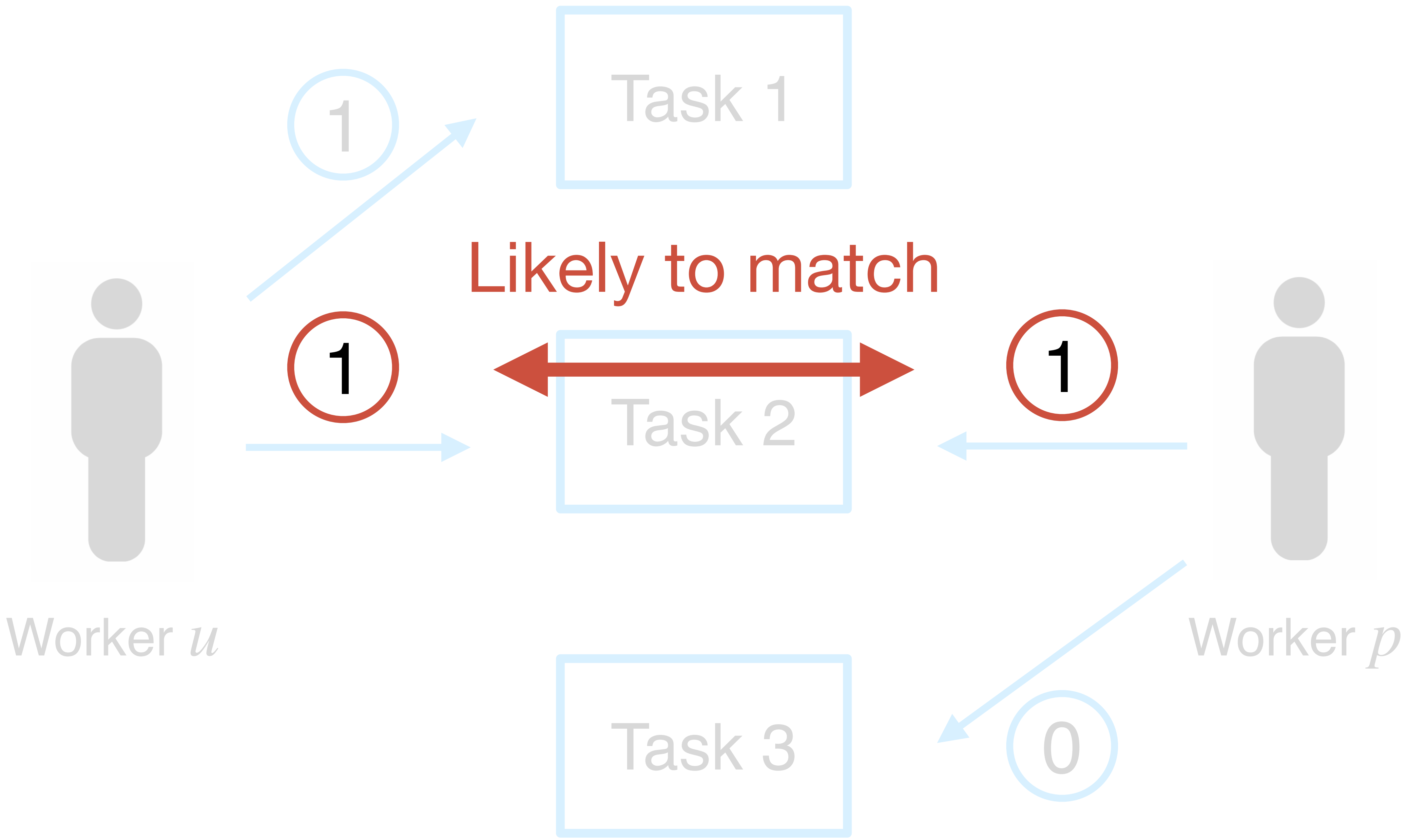
- Each task is labeled by multiple workers and each worker labels multiple tasks
- A worker u_i's behaviors
  - Invest no effort and thus provide a random label
  - Invest full effort with a cost and provide a true label with probability $p_i > 1/2$



Dasgupta, Anirban, and Arpita Ghosh. "Crowdsourced judgement elicitation with endogenous proficiency." *Proceedings of the 22nd international conference on World Wide Web*. 2013.

Pricing Binary Labels (2)

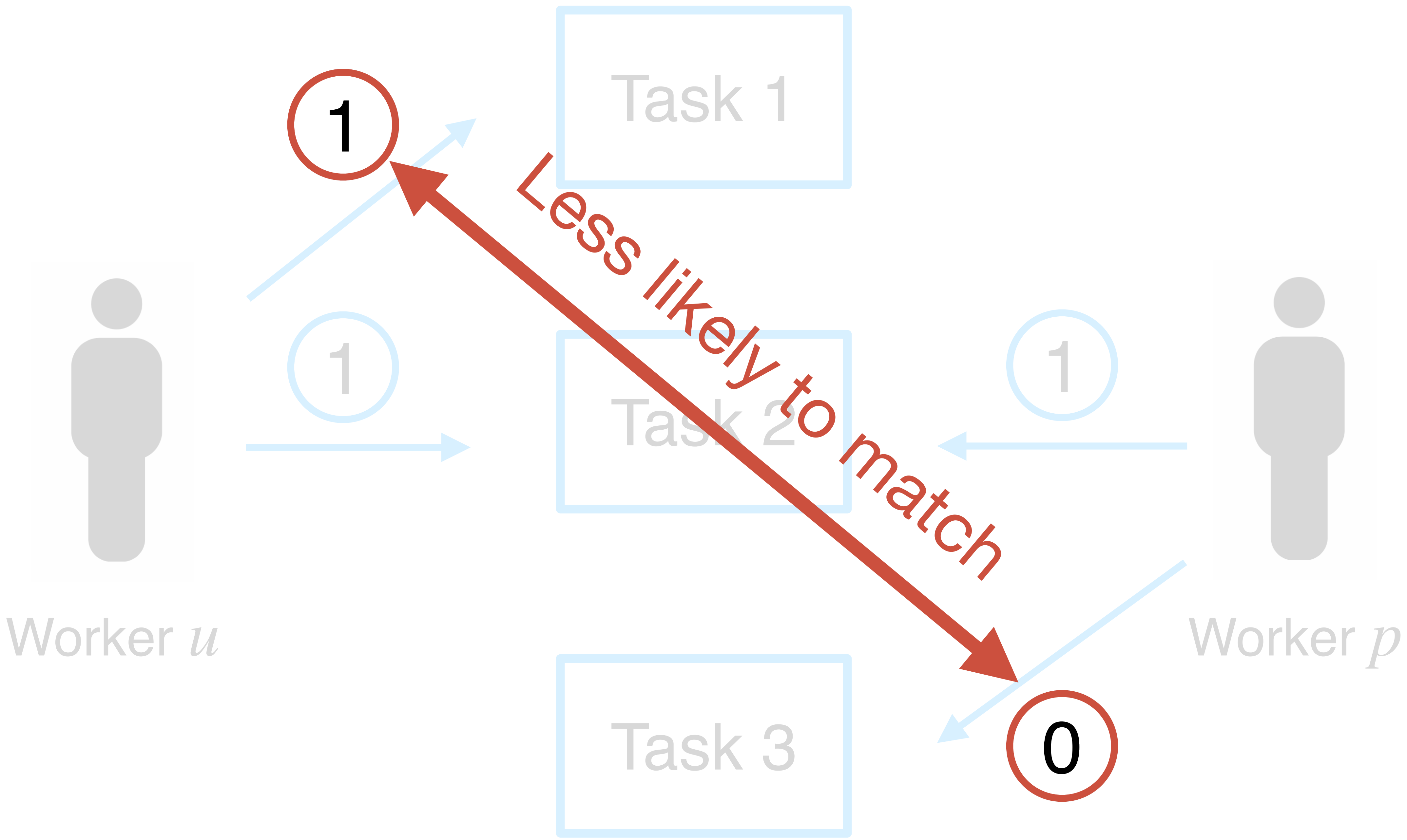
- Each task is labeled by multiple workers and each worker labels multiple tasks
- A worker u_i's behaviors
    - Invest no effort and thus provide a random label
    - Invest full effort with a cost and provide a true label with probability $p_i > 1/2$



Dasgupta, Anirban, and Arpita Ghosh. "Crowdsourced judgement elicitation with endogenous proficiency." *Proceedings of the 22nd international conference on World Wide Web*. 2013.

Pricing Binary Labels (3)

- Each task is labeled by multiple workers and each worker labels multiple tasks
- A worker u_i's behaviors
  - Invest no effort and thus provide a random label
  - Invest full effort with a cost and provide a true label with probability $p_i > 1/2$



Dasgupta, Anirban, and Arpita Ghosh. "Crowdsourced judgement elicitation with endogenous proficiency." *Proceedings of the 22nd international conference on World Wide Web*. 2013.

Pricing Binary Labels (4)

- Reward a worker u_i on a task t based on how surprisingly u_i's report is consistent with that of the peer worker u_p

$$\pi(u_i, t) = \beta \cdot (\mathbf{1}(\hat{y} = \hat{y}_p) - \boxed{\Pr(u_i, u_p)})$$

The probability that u_i and u_p agree on a random task

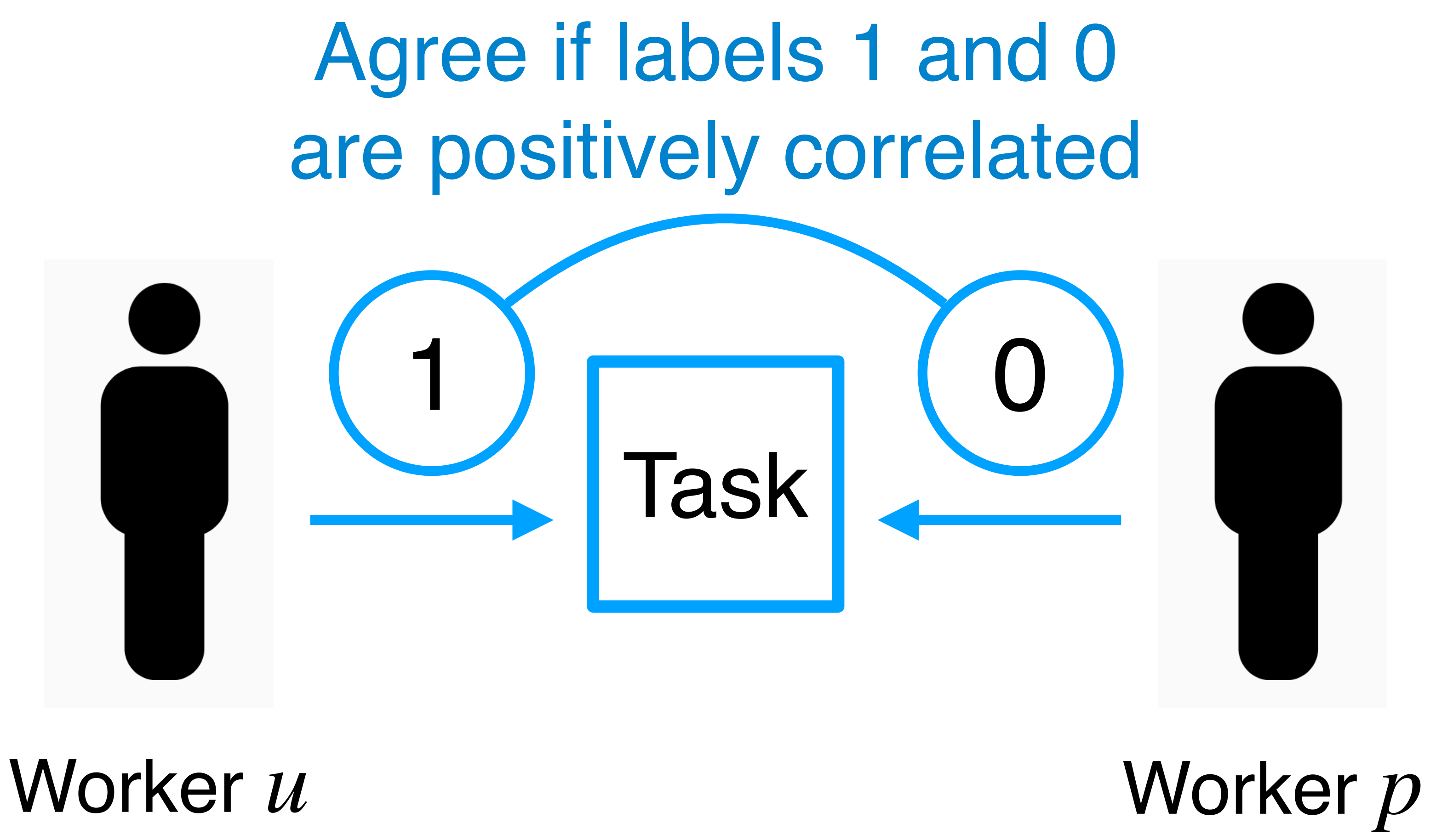
- All workers exerting full efforts and reporting truthfully is an equilibrium
- Exists non-informative equilibrium, that is, all workers report constant labels
    - Workers receive zero rewards in expectation

Dasgupta, Anirban, and Arpita Ghosh. "Crowdsourced judgement elicitation with endogenous proficiency." *Proceedings of the 22nd international conference on World Wide Web*. 2013.

Correlated Agreement (CA) Mechanism

- Consider pricing multi-labels tasks, where two labels l_i and l_j may be positive correlated
  - Workers can misreport l_j by l_i to receive more rewards
- CA mechanism rewards worker u if u's report is positively correlated with that of peer p
- $S(l_i, l_j) = 1$ if the two labels are positively correlated and -1 otherwise

$$\pi(u, t) = \beta \cdot (S(\hat{y}, \hat{y}_p) - S(\hat{y}_a, \hat{y}_b))$$



Shnayder, Victor, et al. "Informed truthfulness in multi-task peer prediction." *Proceedings of the 2016 ACM Conference on Economics and Computation*. 2016.

Correlated Agreement (CA) Mechanism

- Expected payment of exerting efforts and truthful reporting
  - Δ is label correlation matrix

$$\text{E[pay]} = \beta * \sum_{l_i, l_j} \Delta[l_i, l_j] S(l_i, l_j) = \beta * \sum_{\boxed{l_i, l_j, \Delta[l_i, l_j] > 0}} \Delta[l_i, l_j]$$

- Reporting random labels could bring negative elements in Δ into the expected payments
- CA mechanism fails if two labels l_1 and l_2 are not distinguishable with respect to $S(\cdot)$

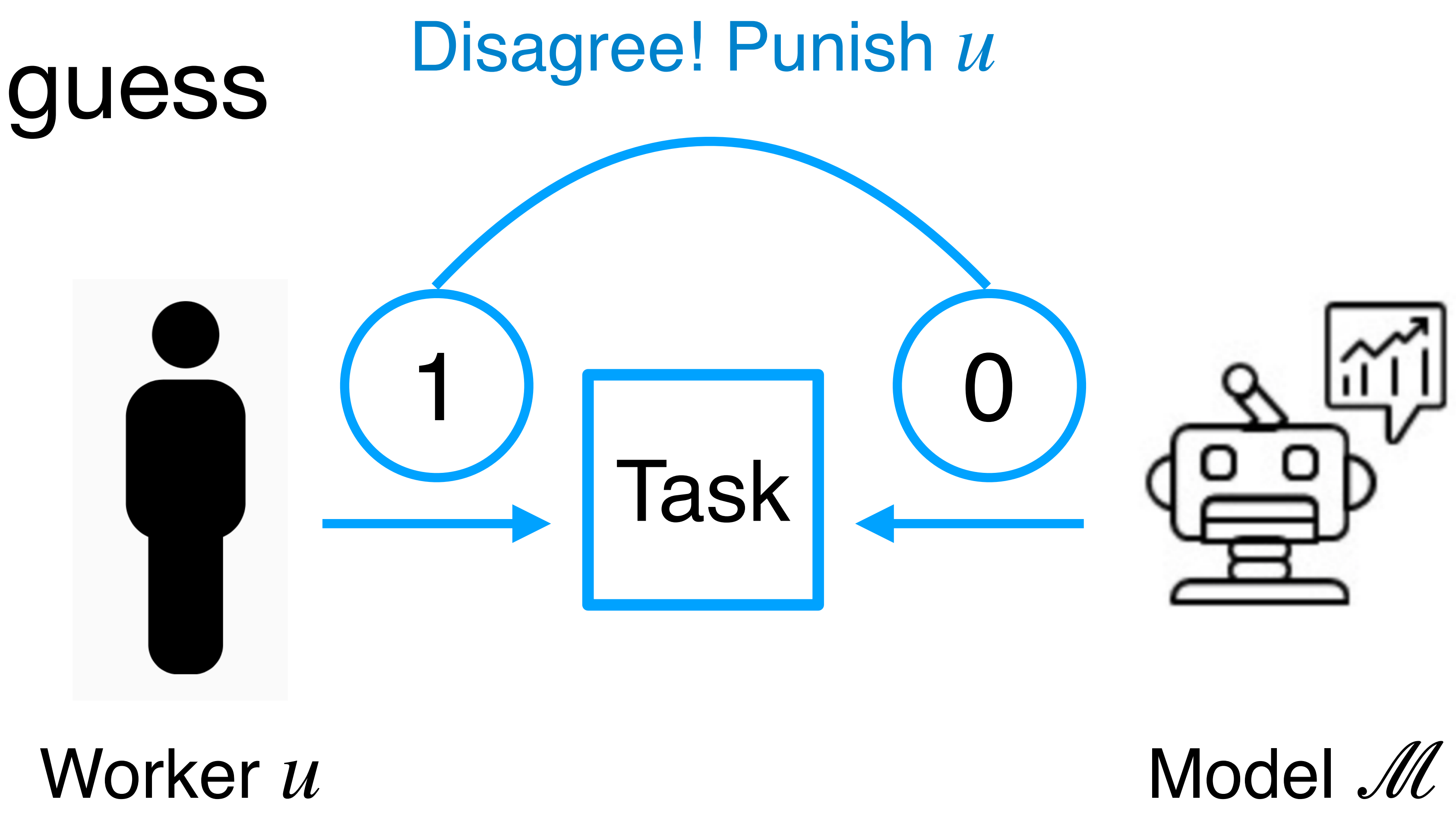
Shnayder, Victor, et al. "Informed truthfulness in multi-task peer prediction." *Proceedings of the 2016 ACM Conference on Economics and Computation*. 2016.

Machine Learning Models as Peer Workers

- Each task must be completed by at least two workers, which leads to duplicate answers

- Learn a classifier $\mathscr{M}$ from workers' reports, and use the classifier's predictions as peer reports

- Assume workers' proficiency is better than random guess



Liu, Yang, and Yiling Chen. "Machine-learning aided peer prediction." *Proceedings of the 2017 ACM Conference on Economics and Computation*. 2017.

Machine Learning Models as Peer Workers

- Learn $\mathscr{M}$ with an error rate calibrated loss function $\varphi(\cdot)$
  - The model is as if evaluated using the ground-truth labels in expectation
- Error calibrated loss function as a payment function

$$\pi(\hat{y}_i) = -\beta * \varphi(\mathscr{M}(t), \hat{y})$$

- Since label noises are removed by the calibrated loss function, reporting true labels can minimize loss
- Exerting efforts and truthful reporting is the most profitable Bayesian Nash Equilibrium

Liu, Yang, and Yiling Chen. "Machine-learning aided peer prediction." *Proceedings of the 2017 ACM Conference on Economics and Computation*. 2017.

Scale Payments by Reputation (1)

- Scale payments such that:
  - Avoid negative payments
  - Increase the difference between the payments to good workers and the payments to spammers

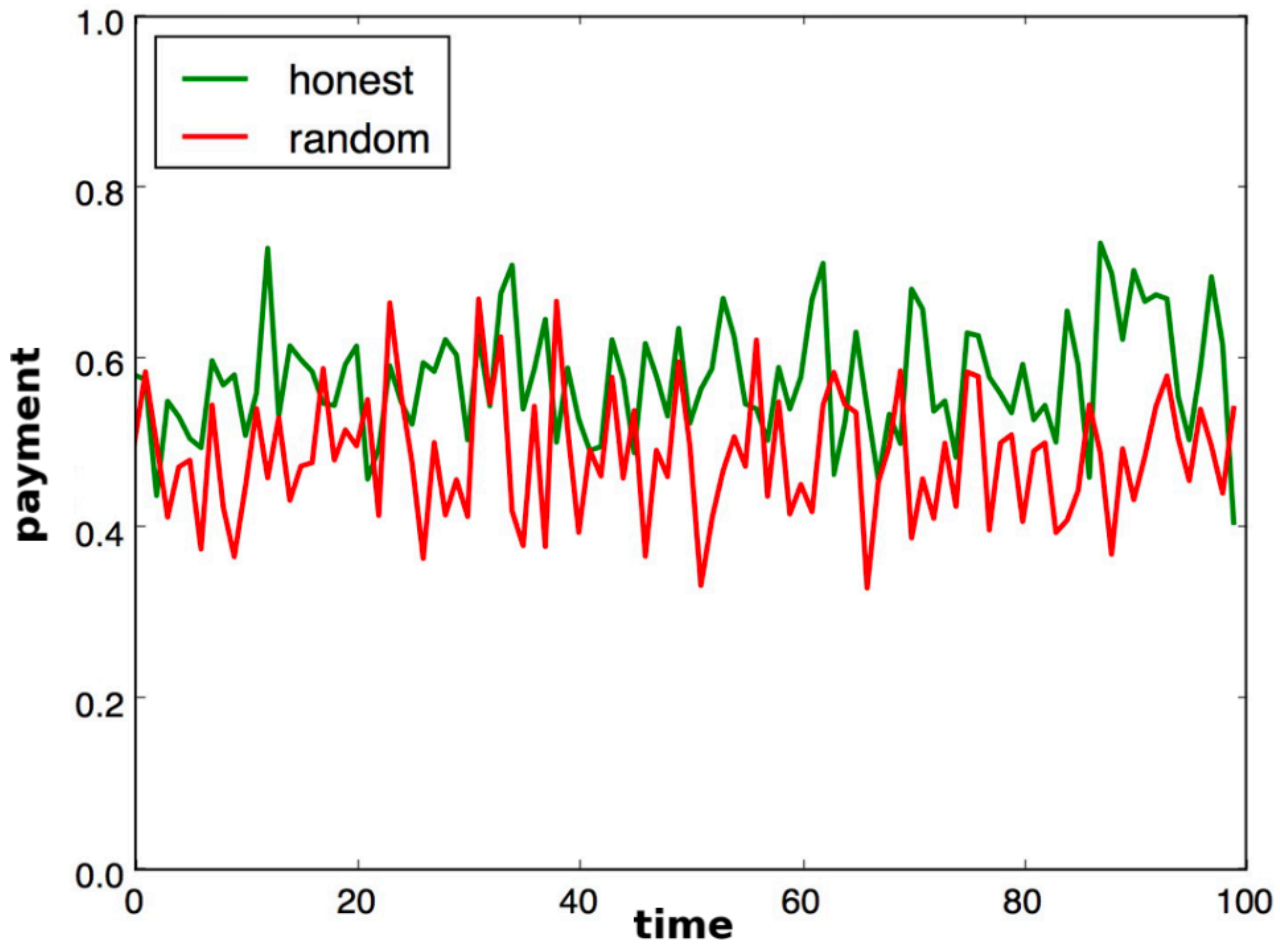


Figure from [Radanovic, Goran, et al., 2016]

Payments of honest workers and spammers

Radanovic, Goran, and Boi Faltings. "Learning to scale payments in crowdsourcing with properboost." *Fourth AAAI Conference on Human Computation and Crowdsourcing*. 2016.

Scale Payments by Reputation (2)

- Publish tasks to workers in multiple rounds T
- Reputation score for each worker is updated based on the worker's report in each round
  - Quality of a report $\hat{y}$ is evaluated by comparing $\hat{y}$ with the estimated true label $\hat{y}_t$
  - Update reputation r_i of worker u_i by

$$r_i = r_i * (1 + \text{constant} * \text{score}(i, t))$$

- Final payment = payment * r_i
- Average payment of a spammer converges to 0 as T approaches infinity

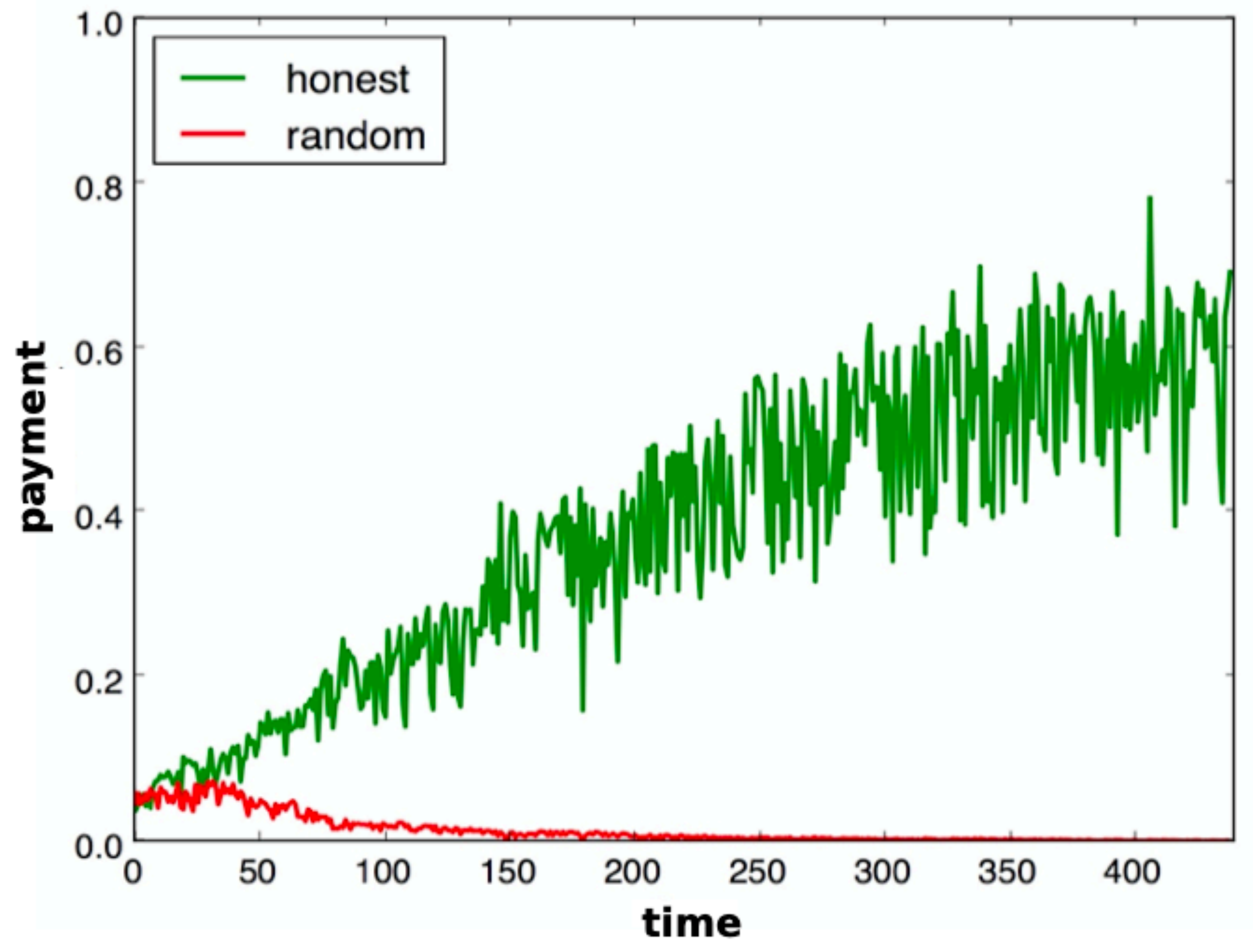


Figure from [Radanovic, Goran, et al., 2016]

Radanovic, Goran, and Boi Faltings. "Learning to scale payments in crowdsourcing with properboost." *Fourth AAAI Conference on Human Computation and Crowdsourcing*. 2016.

Outline: Pricing Data Labels

- Introduction
- Gold Task Based Methods
- Peer Prediction Based Methods
- Summary

Summary: Pricing Data Labels

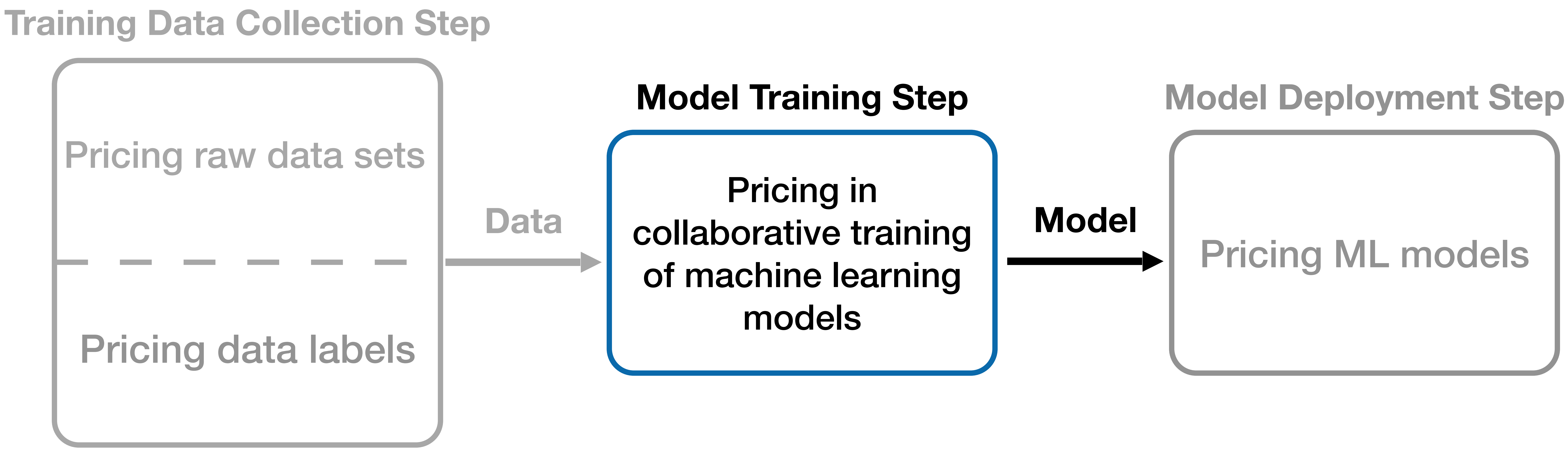
	Gold Task	Peer Prediction
Idea	Uniformly mix gold tasks at random within the tasks for workers to evaluate workers' performance	Form a game among workers such that exerting efforts is the most profitable equilibrium
Pro	Exerting effort is worker's dominant strategy	Do not rely on ground-truth tasks
Con	Gold tasks may be hard to collect	Existence of non-informative equilibria

Part V:
Pricing in Collaborative Training of Machine Learning Models

Outline: Pricing in Collaborative Training of Machine Learning Models

- Introduction
- Revenue Allocation by Shapley value
- Revenue Allocation by Other Fairness Models
    - Leave-one-out
    - Core Based Algorithms
    - Reinforcement Learning Based Algorithm
- Summary

Pricing in Collaborative Training of Machine Learning Models



Introduction

- Collaborative Machine Learning
  - Multiple data owners collaboratively build high quality machine learning models by contributing their data
- Revenue allocation measures
  - Cost-based measure:
    - privacy cost, energy cost, etc.
  - Performance-based measure
    - Make sure data owners who contribute more valuable data achieve more rewards
    - Our tutorial focuses on this measure

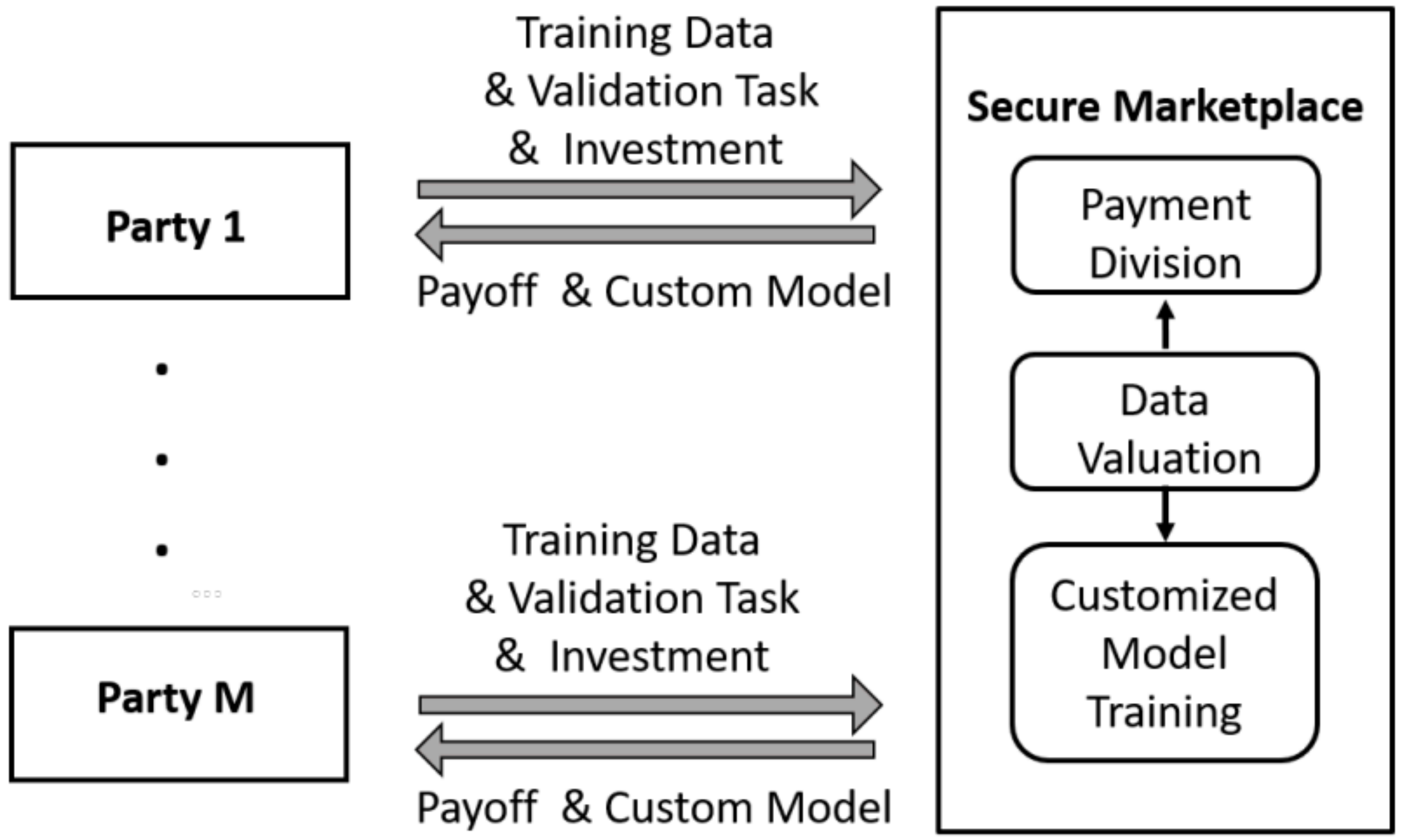


Figure from [Ohrimenko et al., 2019]

Example: Collaborative marketplace setup

Ohrimenko, Olga, Shruti Tople, and Sebastian Tschiatschek. "Collaborative machine learning markets with data-replication-robust payments." *arXiv preprint arXiv:1911.09052* (2019).

Desirable Properties of Revenue Allocation

- Balance
- Symmetry
- Zero Element
- Additivity
- Adversarial Robustness
- Collaboration Stability
- Efficiency

Shapley Fairness

Outline: Pricing in Collaborative Training of Machine Learning Models

- Introduction
- Revenue Allocation by Shapley value
- Revenue Allocation by Other Fairness Models
  - Leave-one-out
  - Core Based Algorithms
  - Reinforcement Learning Based Algorithm
- Summary

Shapley Value

- Defintion $\psi(s) = \dfrac{1}{N!} \sum_{\pi \in \prod(D)} (\mathcal{U}(P_s^\pi \cup \{s\} - \mathcal{U}(P_s^\pi)))$
  - Example:
    - 123, 132, 213, 231, 312, 321
  - Assume the utility function $\mathcal{U}$ is non-decreasing
- $\psi(\cdot)$ is the unique allocation method that possesses Shapley fairness
- Flexibility to support different utility function
  - E.g., performance of trained model in collaborative machine learning
- Challenge: exponential computational cost

Shapley, Lloyd S. 17. A value for n-person games. Princeton University Press, 2016.

Permutation Sampling Algorithm for Bounded Utility Function

- Core idea: get an unbiased estimator of Shapley value via uniform sampling
- Approximate Shapley value by sample mean
  - Simple random sampling
- How to bound estimate error:
  - Chebyshev's inequality
    - $Pr(|\hat{\phi} - \phi| >= \epsilon) <= \frac{\sigma^2}{m\epsilon^2} <= \delta$
  - Hoeffding's inequality
    - $Pr(|\hat{\phi} - \phi| >= \epsilon) <= 2\exp(-\frac{2m^2\epsilon^2}{mr^2}) <= \delta$
- Cons:
  - Evaluating the utility function is computationally expensive, as it requires training a machine learning model

Maleki, Sasan, et al. "Bounding the estimation error of sampling-based Shapley value approximation." arXiv preprint arXiv:1306.4265 (2013).

Truncate-based and Gradient-based Approximation Methods

- Truncated Monte Carlo Shapley
  - Reduce the number of utility evaluations
  - In a sampled permutation, set the marginal contribution to be zero for some S whenever $V(D) - V(S) < a$ predefined threshold
- Gradient-based method
  - Speed up the evaluation of utility functions by reducing training time
  - In a sampled permutation, update the model by performing gradient descent on one data point at a time
  - The marginal contribution is the change in model's performance

Ghorbani, Amirata, and James Zou. "Data shapley: Equitable valuation of data for machine learning." International Conference on Machine Learning. PMLR, 2019.

Truncate-based and Gradient-based Approximation Methods

- Pros
  - Empirically speed up computation
- Cons
  - Introduce estimation bias into the approximated Shapley values
  - Have no guarantee on the approximation error

Ghorbani, Amirata, and James Zou. "Data shapley: Equitable valuation of data for machine learning." International Conference on Machine Learning. PMLR, 2019.

Reduce the Number of Utility Evaluations with Provable Error Bounds

- Two approximation algorithms to reduce the number of utility evaluations
- Algorithm 1: group testing-based approximation algorithm
  - Group testing

$$\psi(i) - \psi(j) = \frac{1}{N-1} \sum_{S \subseteq D \setminus \{i,j\}} \frac{\mathcal{U}(S \cup \{i\} - \mathcal{U}(S \cup \{j\})}{\binom{N-2}{|S|}} = E[(\beta_i - \beta_j)\mathcal{U}(\beta_1, \ldots, \beta_N)]$$

  - O($N(logN)^2$) utility evaluations

Jia, Ruoxi, et al. "Towards efficient data valuation based on the shapley value." The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019.

Reduce the Number of Utility Evaluations with Provable Error Bounds

- Algorithm 2: sparse signal recovering-based approximation algorithm
  - Based on observation that Shapley values are approximately sparse
    - Most of values are concentrated around its mean and only a few data have significant values
  - Sparse signal recovering idea
  - $O(Nlog(logN))$ utility evaluations

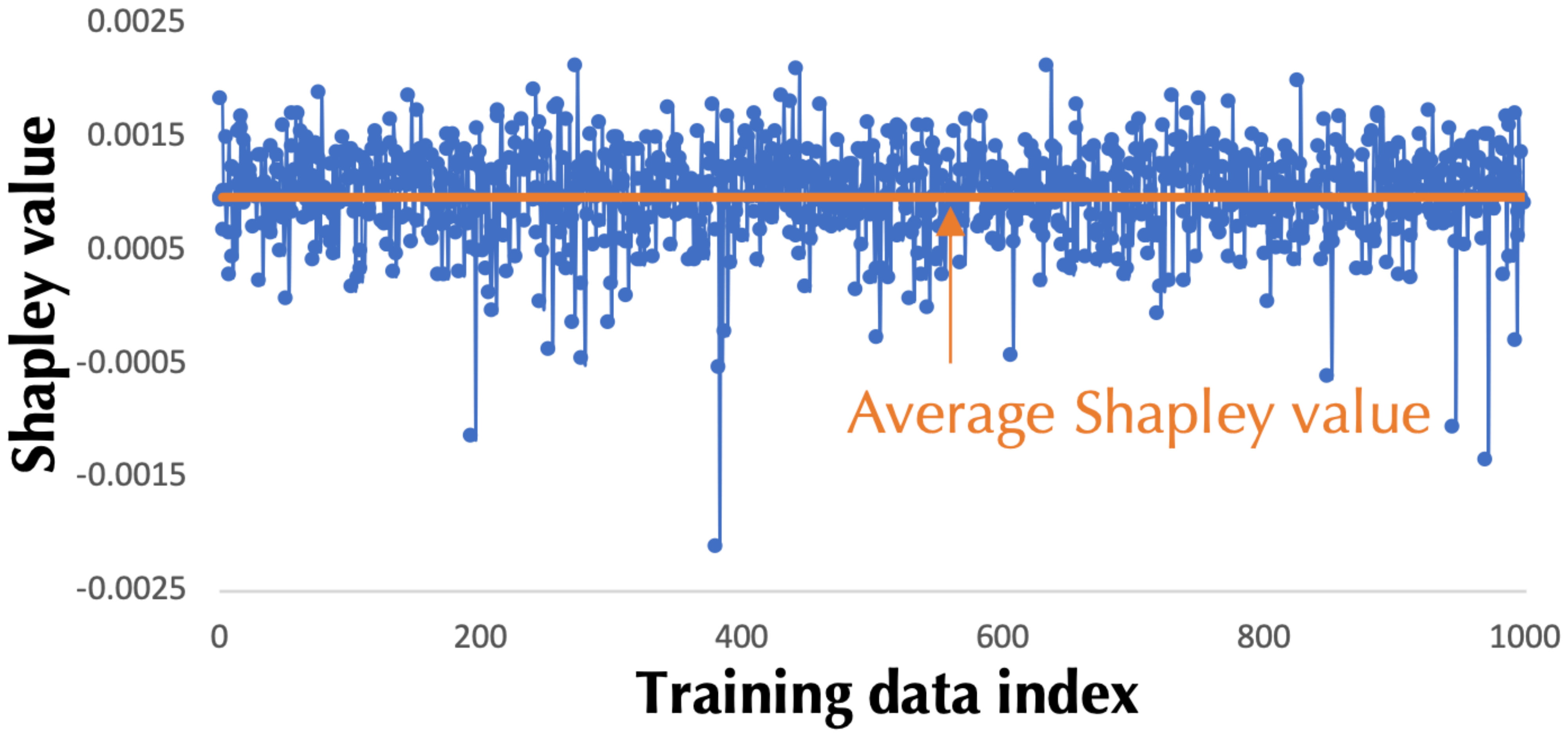


Figure from [Jia, Ruoxi, et al., 2019]

Jia, Ruoxi, et al. "Towards efficient data valuation based on the shapley value." The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019.

Shapley Value in Unweighted KNN Classifiers

- Define a special utility function to enable efficient computation of Shapley differences between two data points
  - For a single testing point, define the utility of KNN classifiers by the likelihood of the right label

$$\nu(S) = \frac{1}{K} \sum_{k=1}^{\min\{K,|S|\}} \mathbb{1}[y_{\alpha_k(S)} = y_{\text{test}}]$$

  - Based on above utility function, the Shapley value of each training points can be calculated recursively as

$$s_{\alpha_N} = \frac{\mathbb{1}[y_{\alpha_N} = y_{test}]}{N}$$

$$s_{\alpha_i} = s_{\alpha_{i+1}} + \frac{\mathbb{1}[y_{\alpha_i} = y_{test}] - \mathbb{1}[y_{\alpha_{i+1}} = y_{test}]}{K} \frac{\min\{K, i\}}{i}$$

  - Generalize above utility function to the case with multiple testing points
    - The Shapley value computation cost complexity is $O(NlogNN_{test})$

Jia, Ruoxi, et al. "Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms." *Proceedings of the VLDB Endowment* 12.11.

Shapley Value in Unweighted KNN Classifiers

- Develop an algorithm to only compute Shapley values for the retrieved k nearest neighbours
  - Reduce the computational cost to $O(NlogN)$ time
- The idea can be adapted to any “local” models
  - Models which only use a subset of the entire data set for data prediction

Jia, Ruoxi, et al. "Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms." *Proceedings of the VLDB Endowment* 12.11.

Information Gain based Algorithm

- This algorithm considers the situation where no validation data sets are available
- Use information gain on model parameters as the utility function

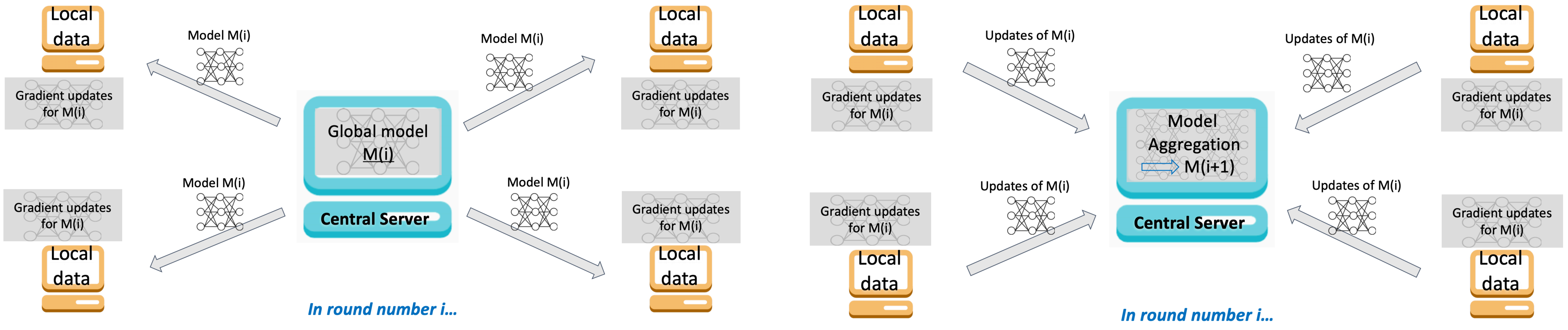
$$IG(\theta) = H(\theta) - H(\theta | D)$$

- Three additional incentive conditions are proposed
    - Individual rationality
    - Stability of the grand coalition
    - Group welfare
- Machine learning models as rewards over money incentives

Sim, Rachael Hwee Ling, et al. "Collaborative machine learning with incentive-aware model rewards." International Conference on Machine Learning. PMLR, 2020.

Federated Learning

- Federated Learning
  - Collaborative machine learning without centralized training data



Example: Federated Learning

https://inst.eecs.berkeley.edu/~cs294-163/fa19/slides/federated-learning.pdf

Federated Shapley Value

- Definition:
  - Federated Shapley value of participant i at round t is defined as
    - $\phi_t(s_i) = \frac{1}{|I_t|} \sum_{S \subseteq I_t \setminus \{i\}} \frac{1}{\binom{|I_t|-1}{|S|}} [\mathcal{U}(I_{1:t-1} + (S \cup \{i\})) - \mathcal{U}(I_{1:t-1} + S)]$
  - Federated Shapley value of participant i:
    - $\phi(s_i) = \sum_{t=1}^{T} \phi_t(s_i)$
- Advantages
  - Satisfy the balance and additivity axioms of Shapley fairness
  - Symmetry and zero element are satisfied in each round
- Extend the permutation sampling and group testing approximation methods to compute federated Shapley value

Wang, Tianhao, et al. "A principled approach to data valuation for federated learning." Federated Learning. Springer, Cham, 2020. 153-167.

Replication-robust Shapley Value

- Shapley value is vulnerable to data-replication attacks

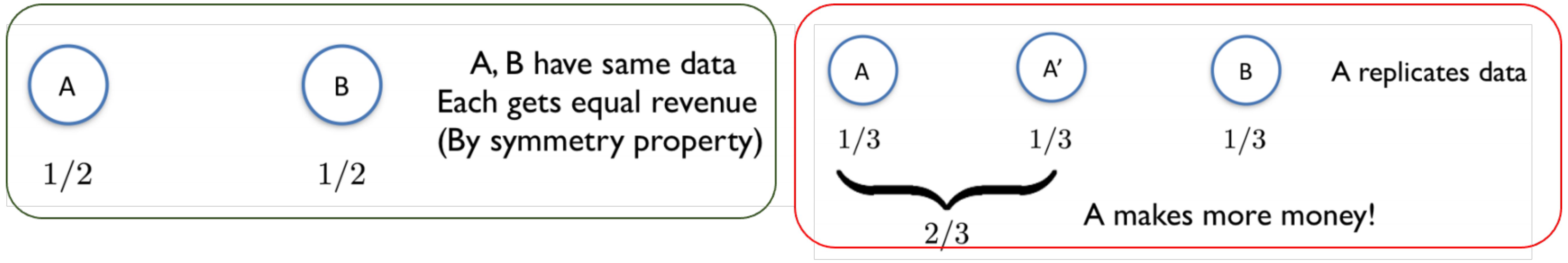


Figure from [Agarwal et al., 2019]

- Replication-robust Shapley value
  - Robust to data replication-attacks by penalizing similar data sets to disincentive replication
  - No longer satisfies the balance axiom in Shapley fairness

Agarwal, Anish, Munther Dahleh, and Tuhin Sarkar. "A marketplace for data: An algorithmic solution." Proceedings of the 2019 ACM Conference on Economics and Computation. 2019.

Outline: Pricing in Collaborative Training of Machine Learning Models

- Introduction
- Revenue Allocation by Shapley value
- Revenue Allocation by Other Fairness Models
    - Leave-one-out
    - Core Based Algorithms
    - Reinforcement Learning Based Algorithm
- Summary

Leave-one-out Methods

- To formalize the impact of a training point on a prediction
- Evaluating data importance by comparing the performance of a model trained on the full data set with that trained on the full set minus one point
- Challenge
  - Perturbing the data and retraining the model can be expensive
- We have influence functions!!!
  - A classic technique from robust statistics that tells us how the model parameters change as we upweight a training point by an infinitesimal amount

Influence Function When Upweighting Training Point

- How do we know the change in model parameters due to removing a training point z?
  - There exists a close-form influence function to approximate parameter change when upweighting z
  - Removing a training point z is the same as up-weighting it to a degree
- How do we know the change in model's predictions due to removing a training point z?
  - Similar to above!

Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions." International Conference on Machine Learning. PMLR, 2017.

Influence Function for Federated Learning

- Reward participants in federated learning for their contributed data points
- Two improvements compared to the previous influence function when upweighting training points
  - Batch processing to handle sequential data
  - Resolve the issue that mean influence is zero

Richardson, Adam, Aris Filos-Ratsikas, and Boi Faltings. "Rewarding high-quality data via influence functions." arXiv preprint arXiv:1908.11598 (2019).

Leave-one-out vs Shapley Value

- Leave-one-out vs Shapley value
  - Leave-one-out models are more efficient as they do not require model retraining
  - Leave-one-out models may not accurately assess the values of data points
    - E.g., it may assign a low value to one the two exactly equivalent data points, as high performance may still be achieved by including the other datum

Core Based Data Pricing Model

- Core
  - Revenue allocation solutions that satisfy the following
    - Constraint: the total reward of each coalition should be at least equal to its utility
  - E.g. an cooperative game including three players A, B, C
    - $u(A, B, C) = 1000, \quad u(A, B) = 500, \quad u(B, C) = 500, \quad u(A, C) = 500$
    - Two solutions belong to core
      - $\varphi(A) = 0, \quad \varphi(B) = 500, \quad \varphi(C) = 500$
      - $\varphi(A) = 100, \quad \varphi(B) = 400, \quad \varphi(C) = 500$
    - Choose the solution with the smallest l_2-norm
  - Pros: achieve maximum stability of how participants team up with each other
  - Cons: only satisfies the balance, symmetry, and zero element axioms of Shapley fairness

Gillies, Donald B. "3. Solutions to general non-zero-sum games." Contributions to the Theory of Games (AM-40), Volume IV. Princeton University Press, 2016. 47-86.

Core Based Data Pricing Model

- Least Core
  - Relax the constraint by allowing a minimum difference between the utility and the total reward for a given coalition
- The number of constraints grows exponentially with the number of participants!
- Monte Carlo algorithm
  - Reduce computational cost by allowing a relaxed version of the least core

Yan, Tom, and Ariel D. Procaccia. "If you like shapley then you'll love the core." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 6. 2021.

Reinforcement Learning Algorithm

- Intuition: integrate data valuation into the training procedure of the predictor model

- Mechanism

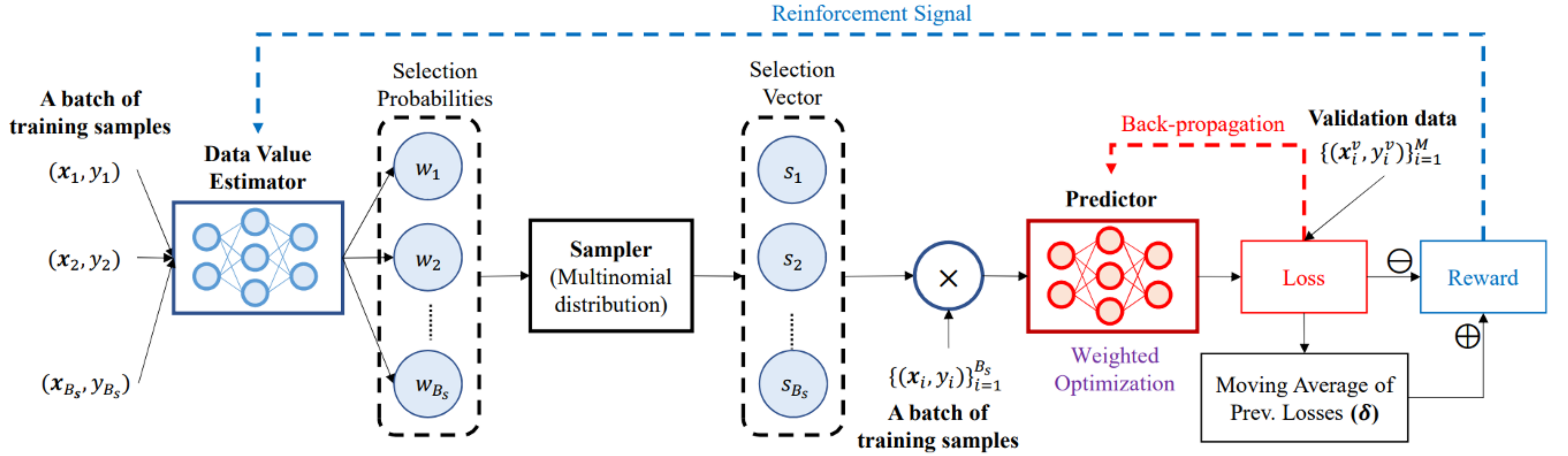


Figure from [Yoon et al., 2020]

- Pros
  - Scalable to large datasets
  - Integrate data valuation into the training procedure of the predictor model, allowing the predictor and data value estimator to improve each other's performance.

Yoon, Jinsung, Sercan Arik, and Tomas Pfister. "Data valuation using reinforcement learning." *International Conference on Machine Learning*. PMLR, 2020.

Outline: Pricing in Collaborative Training of Machine Learning Models

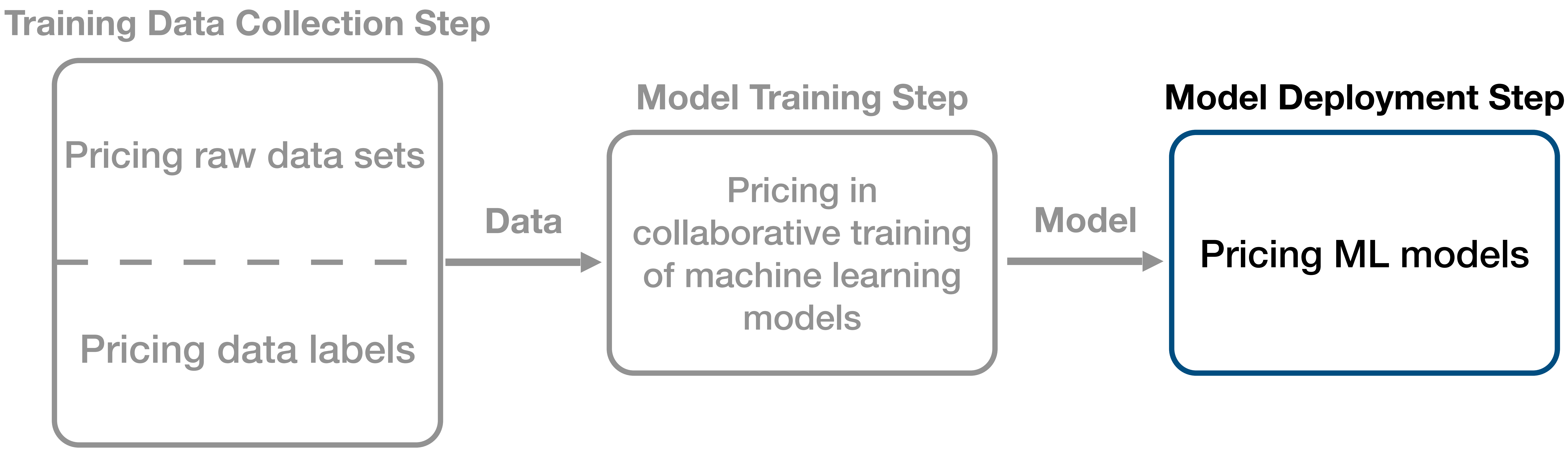
- Introduction
- Revenue Allocation by Shapley value
- Revenue Allocation by Other Fairness Models
  - Leave-one-out
  - Core Based Algorithms
  - Reinforcement Learning Based Algorithm
- Summary

Summary: Pricing in Collaborative Training of Machine Learning Models

	Types	Limitations
Shapley value based	Shapley value equation	Exponential computational cost Non-decreasing utility function
	Sampling based method	Achieve partial Shapley fairness
	Utilize properties of machine learning model to reduce computational cost	Limited application
Non-Shapley-value based	Estimate data importance by comparing model performance with and without a training point	Achieve partial Shapley fairness
	Revenue allocation by resolving mathematical equations with predefined constraints	Achieve partial Shapley fairness
	Estimate importance of training examples via reinforcement learning process	Achieve partial Shapley fairness

Part VI:
Pricing Machine Learning Models

Pricing ML Models in Machine Learning Pipelines



Machine Learning Model as a Service

- Machine learning as a service (MLaaS) is a rapidly growing industry
- Customers may purchase well-trained machine learning models or build models on top of those well-trained rather than building models from scratch by themselves
    - Example: one may use Google prediction API to classify an image for only $0.0015

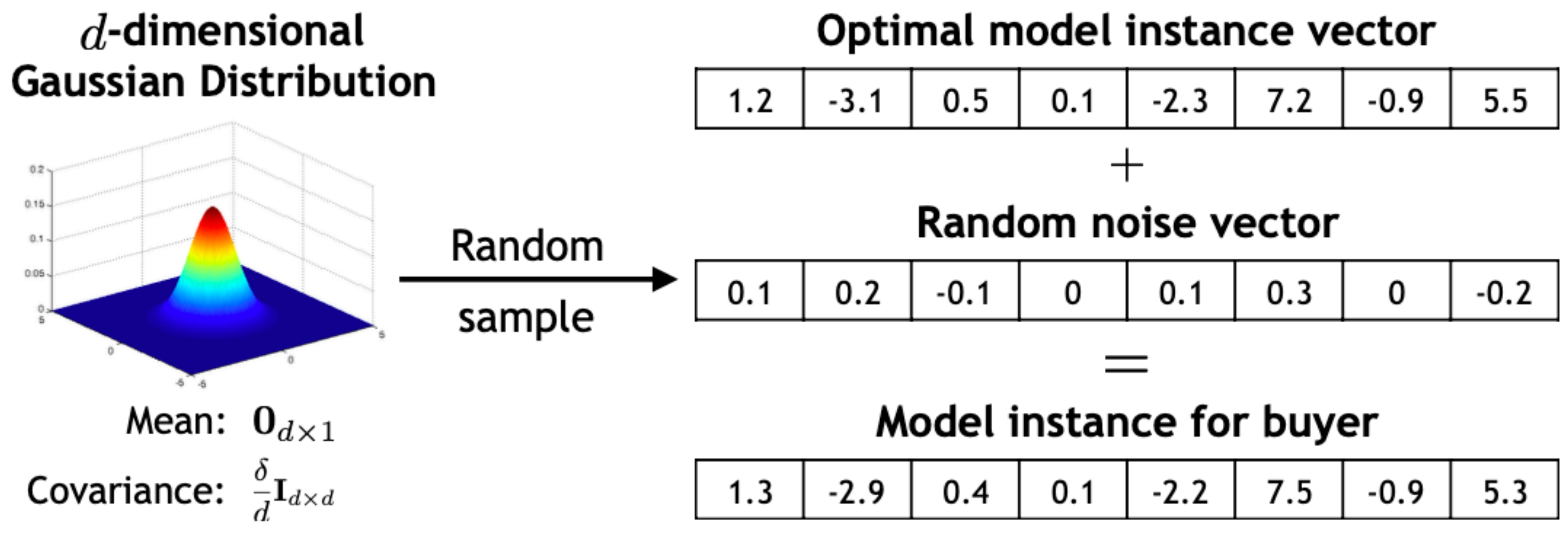
Chen, Lingjiao, et al. “FrugalML: How to Use ML Prediction APIs More Accurately and Cheaply.” Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 10685–10696.

Two Challenges in Pricing ML Models

- Model versioning
  - Perturb model parameters
  - Perturb training data
- Model pricing
  - Arbitrage-free
  - Revenue maximization

Model Versioning and Arbitrage-free Pricing

- Produce model instances with different performances to target customers with different demands
- Assume ML models are trained by strictly convex loss functions
- Train an optimal classifier $\mathscr{M}$ on training data
  - Add Gaussian random noise $w \sim \mathcal{N}(0, \delta * I_d)$ to the parameters of $\mathscr{M}$
  - The expected error $\mathbb{E}[\epsilon(\mathscr{M} + \mathbf{w}, D)]$ is monotonic with respective to δ
- Arbitrage-free pricing function π
  - A buyer cannot derive a high performance model by paying less
  - If and only if π is sub-additive and monotone over $\dfrac{1}{\delta}$



Chen, Lingjiao, Paraschos Koutris, and Arun Kumar. "Towards model-based pricing for machine learning in a data marketplace." *Proceedings of the 2019 International Conference on Management of Data*. 2019.

Revenue Maximization in Model Pricing

- Set prices to different versions to maximize the revenue of model sellers
- A customer purchases a model if the price is lower than his valuation
  - Customers demands are public information
  - Total revenue: $\sum_{i} \pi(\frac{1}{\epsilon}) * \text{purchase}(\pi(\frac{1}{\epsilon}))$
  - Constraint: $\pi(\cdot)$ is arbitrage-free

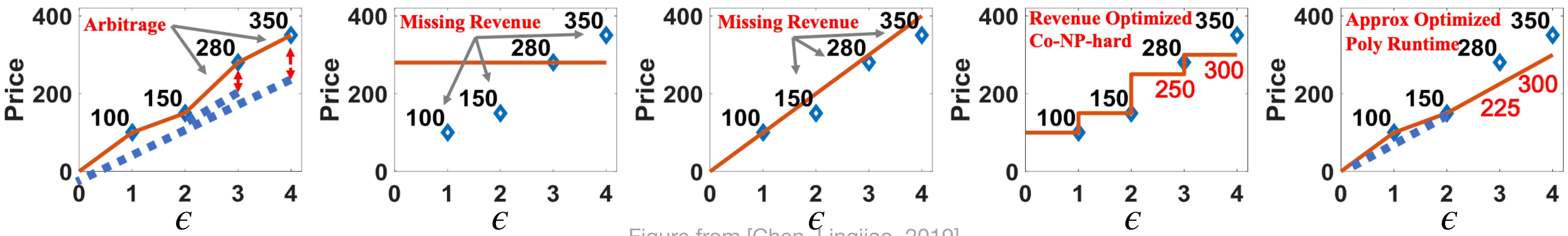


Figure from [Chen, Lingjiao, 2019]

Example pricing functions

Chen, Lingjiao, Paraschos Koutris, and Arun Kumar. "Towards model-based pricing for machine learning in a data marketplace." *Proceedings of the 2019 International Conference on Management of Data*. 2019.

Revenue Maximization in Model Pricing

- Determining the revenue maximization price is co-NP hard
- Relax the subadditive constraints $\pi(x+y) \leq \pi(x) + \pi(y)$ by $\frac{\widehat{\pi}(x)}{x} \leq \frac{\widehat{\pi}(y)}{y}$, where $y \geq x \geq 0$
  - Bounded approximation error $\pi(x)/2 \leq \hat{\pi}(x) \leq \pi(x)$
  - Price $\hat{\pi}(x)$ can be computed by dynamic programming in $O(n^2)$

Chen, Lingjiao, Paraschos Koutris, and Arun Kumar. "Towards model-based pricing for machine learning in a data marketplace." *Proceedings of the 2019 International Conference on Management of Data*. 2019.

Model Market with Differential Privacy

- ML models with different differential privacy levels ϵ are traded
  - Constructed by objective perturbation
- A model may have multiple data contributors and the revenue should be distributed to contributors
- Cost for using data owner s_i's data in ϵ-differential privacy manner:

$$\pi(s_i, \epsilon) = b_i \cdot c_i(\epsilon)$$

Data quality

Privacy cost

- Fairness: A data owner contributing to a model receives a reward promotional to $\pi(s_i, \epsilon)$
- Each model has multiple survey prices

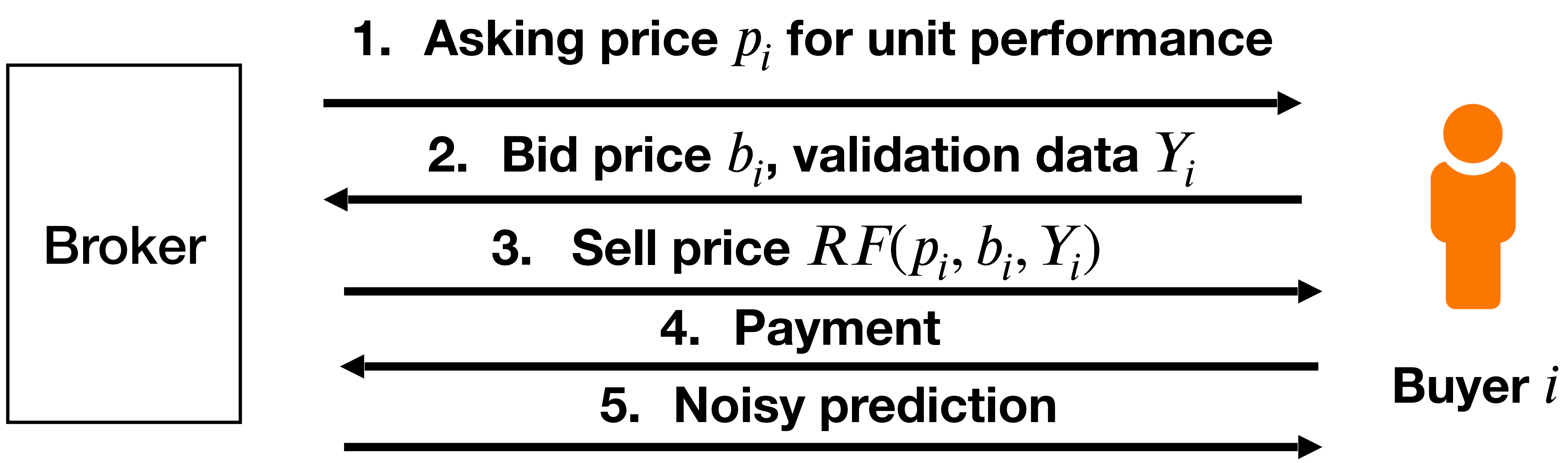
Liu, Jinfei, et al. "Dealer: an end-to-end model marketplace with differential privacy." *Proceedings of the VLDB Endowment* 14.6 (2021): 957-969.

Model Market with Differential Privacy

- Properties of pricing function $p(\epsilon)$
  - Arbitrage-free with respect to ϵ: sub-additive and monotone over ϵ
  - Maximizing revenue: co-NP hard to optimize
  - Cover data owners' costs: NP hard
- Two optimization problems
  - Determine revenue maximization price $p(\epsilon)$ with respect to customer's demands and valuations
  - Given manufacturing cost $p(\epsilon)$, select a subset of data providers, such that the total data quality is maximized

Liu, Jinfei, et al. "Dealer: an end-to-end model marketplace with differential privacy." *Proceedings of the VLDB Endowment* 14.6 (2021): 957-969.

Online Auction for ML Models (1)



- Online auction: different customers bid at different times
- The broker sets the asking price to maximize cumulative revenue by learning from historical transactions
- Noisy models are generated for the buyer by adding calibrated noise w into training data

$$w \sim (p_i - b_i) * \mathcal{N}(0,\sigma^2)$$

- The model's performance $G(\hat{Y}_i, Y_i)$ is inversely proportional to $p_i - b_i$

Agarwal, Anish, Munther Dahleh, and Tuhin Sarkar. "A marketplace for data: An algorithmic solution." *Proceedings of the 2019 ACM Conference on Economics and Computation*. 2019.

Online Auction for ML Models (2)

- Buyers are selfish and wants to maximize their utility by choosing b_i

$$\underbrace{\mathcal{U}(b_i) = \mu_i}_{\text{Buyer } i\text{'s valuation on unit performance}} \cdot G(\hat{Y}_i, Y_i) - \underbrace{RF(p_i, b_i, Y_i)}_{\text{Determined following Myerson's payment function to motivate truthful bids}}$$

- The seller determines asking price p_i by multiplicative weights algorithm
  - Price p_i is sampled from a list of pre-defined prices
  - Prices that bring larger historical revenues are more likely to be sampled
  - Average regret goes to zero as $i \to \infty$

Agarwal, Anish, Munther Dahleh, and Tuhin Sarkar. "A marketplace for data: An algorithmic solution." *Proceedings of the 2019 ACM Conference on Economics and Computation*. 2019.

Pricing Raw Data Products Versus Machine Learning Models

- The pricing units of machine learning models are often well defined and fixed
- Versioning ML models is harder than versioning raw data sets
- The value of raw data sets to customers is generally harder to measure than that of machine learning models
- Preventing arbitrage is usually harder in model market than in raw data market

Summary: Pricing Machine Learning Models

- Versioning techniques for ML models
- Arbitrage-free and revenue maximization pricing models of ML models
- Major differences between machine learning model products and raw data set products, including pricing units, versioning, arbitrage prevention, and customer valuation

Part VII: Conclusion

What Did We Discuss?

What is data pricing	• Machine learning pipeline • Data and ML models as economic goods
Essentials of pricing data and ML models	• Data markets • Pricing strategies • Data and model pricing desiderata
Pricing raw data sets	• Pricing general data sets • Pricing crowdsensing data • Pricing data queries • Compensating privacy loss
Pricing data labels	• Gold task-based methods • Peer prediction-based methods
Pricing in collaborative training of ML models	• Revenue allocation by Shapley Value • Pricing by other fairness models
Pricing in ML models	• Pricing ML models • Pricing raw data products versus ML models

The Principle and Seven Desiderata of Data Pricing

Effort elicitation - 7

- Reward workers based on consistency with a reference

Computational efficiency - 6

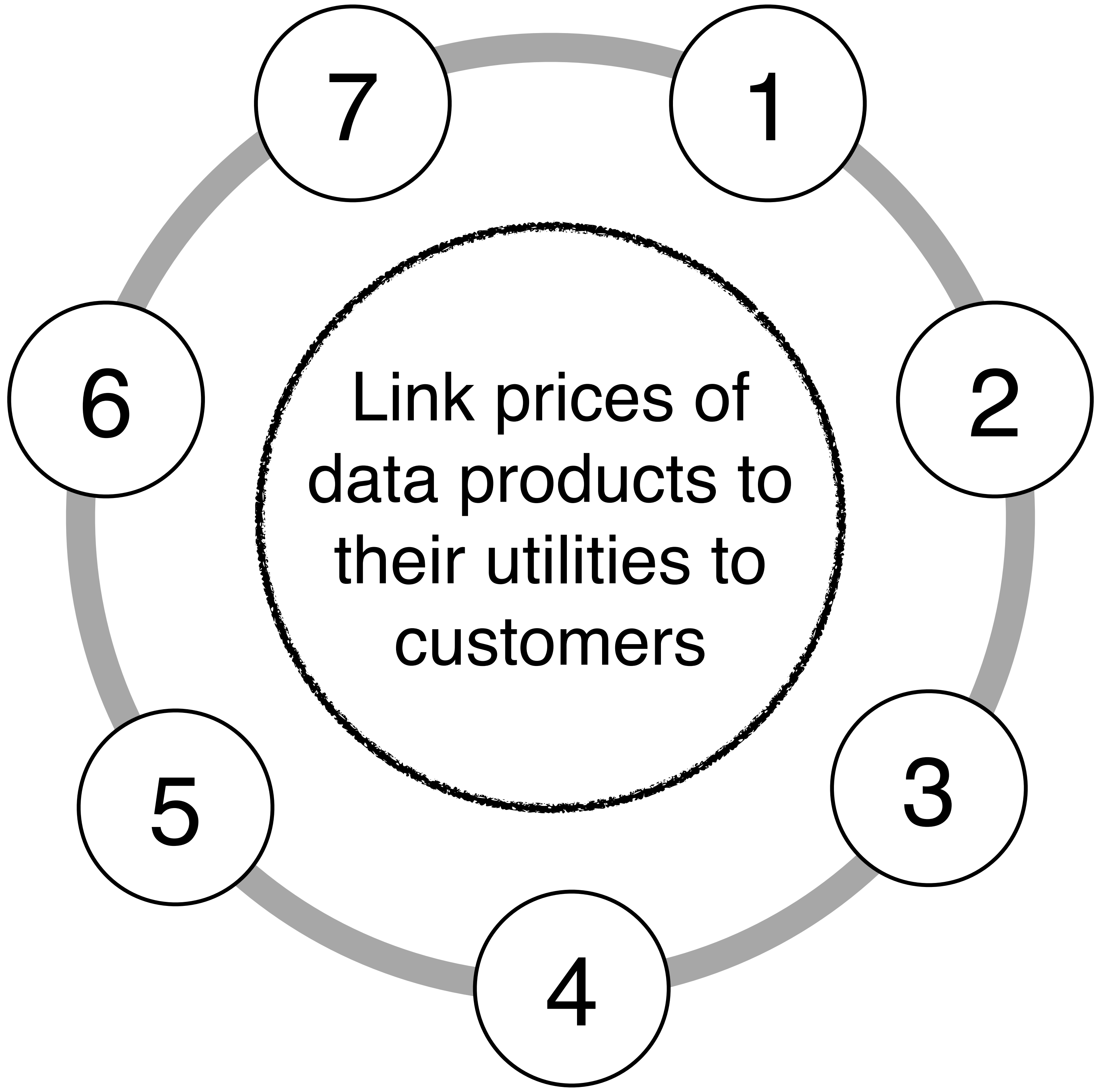
- Develop approximation algorithms by leveraging properties of ML tasks

Privacy preservation - 5

- Protect data owners' privacy by differential privacy and compensate data owners by their privacy loss

Arbitrage-free - 4

- In general, the pricing function is subadditive and monotone with respect to the utility of a data product



1- Truthfulness

- Adapt well-developed truthful auction mechanisms to develop truthful marketplaces

2 - Revenue maximization

- Determine revenue maximization prices by solving an optimization problem with respect to public demands and valuations of customers

3 - Fairness

- Adapt revenue allocation solutions developed in cooperative game theory to reward participants

Future Directions (1)

Task Complexity

- Price data products in more complicated and realistic environments
  - E.g., studying fine-grained data procurement in competitive markets
  - Data sellers need to assign prices to different parts of their data sets based on supply and demand
  - Data buyers need to explore how to distribute their budgets among data sellers to maximize the utility of purchased data sets

Model Axioms

- Understand the necessary axioms for data pricing in different scenarios
  - E.g., Shapley value vs non-Shapley value based methods
  - Some axioms of revenue allocation methods
    - Balance, symmetry, zero element, additivity, adversarial robustness, collaboration stability, and computational efficiency
  - Shapley value can only satisfy the first four axioms

Future Directions (2)

End-to-End Pricing Models

- Systematic study of an end-to-end pricing model in ML pipelines
  - Develop a mechanism that can measure and compare the contributions of different parties in different stages
  - Develop a system that can dynamically adjust the budget allocations in response to the changes in supply and demand

Model Evaluation

- Rigorous evaluation methods for data pricing models
  - Develop a platform that can simulate complicated behaviors of market participants
  - Test the robustness of designed data markets against adversarial participants

Fernandez, Raul Castro, Pranav Subramaniam, and Michael J. Franklin. "Data market platforms: trading data assets to solve data problems." *Proceedings of the VLDB Endowment* 13.12 (2020): 1933-1947.

Thank You!