Online Data Valuation and Pricing for Machine Learning Tasks in Mobile Health

Anran Xu, Zhenzhe Zheng, Fan Wu, and Guihai Chen

Shanghai Jiao Tong University, China

May 4, 2022

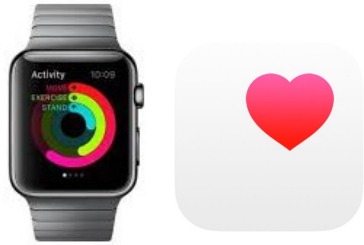
上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

1 Background

mHealth Applications

Google Fit

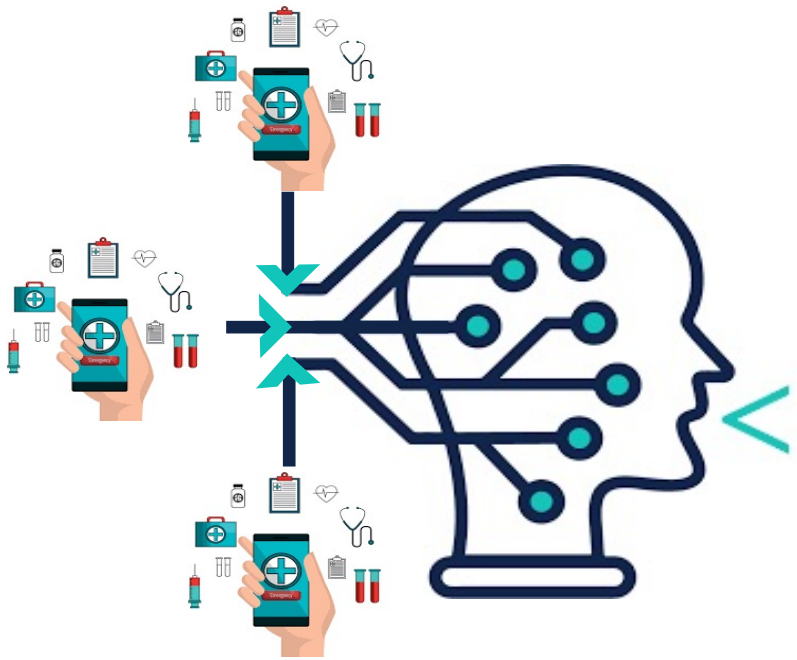


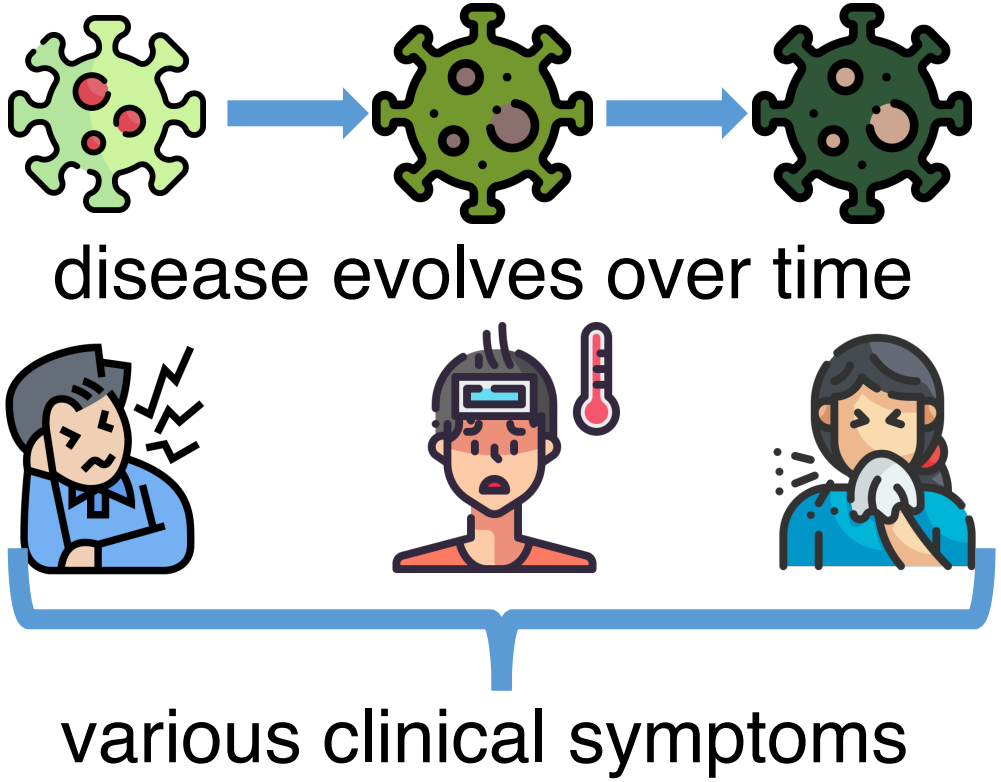
Apple Health

Microsoft Health

mHealth applications, benefiting from mobile computing, have emerged rapidly in recent years.

Machine Learning Tasks in mHealth

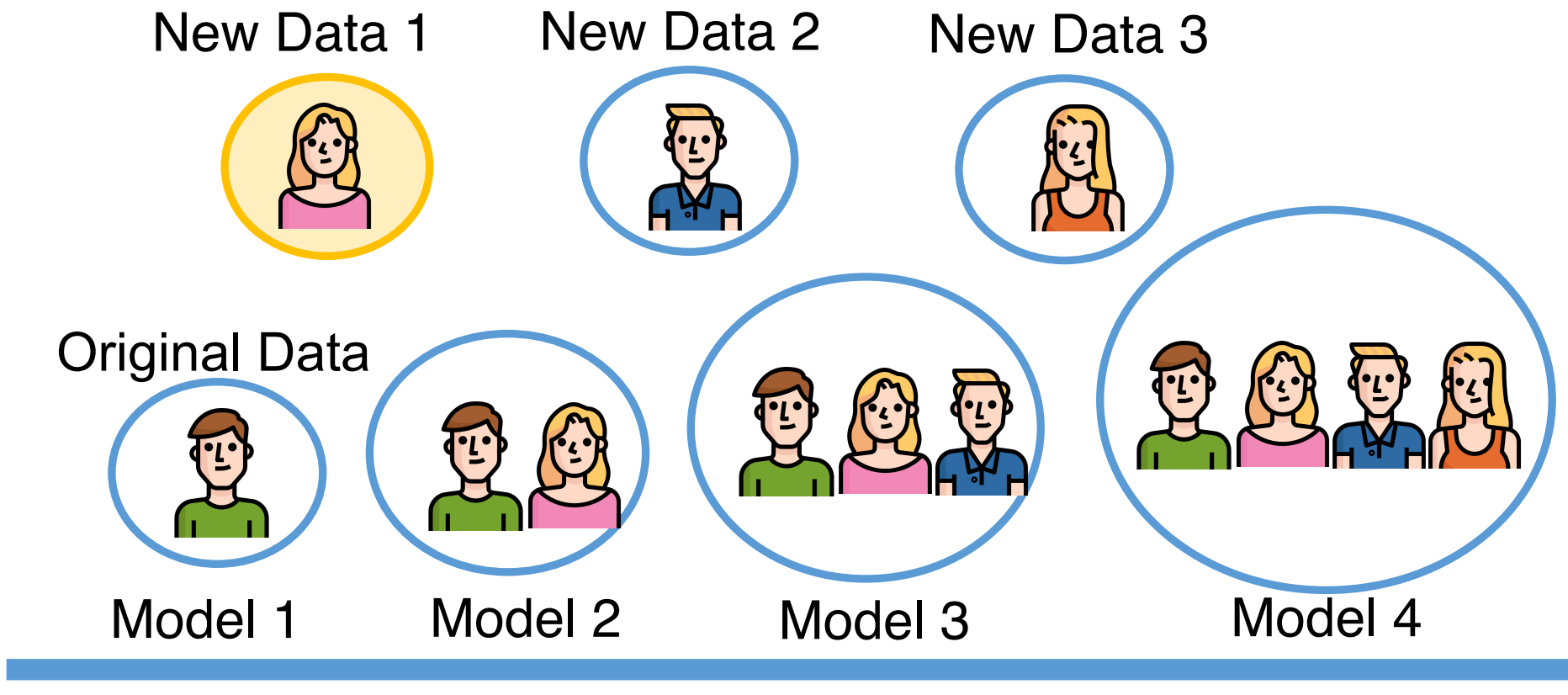




Stimulate users to contribute mHealth data for **ML models**

Online characteristics of the data acquisition

Online Learning in mHealth



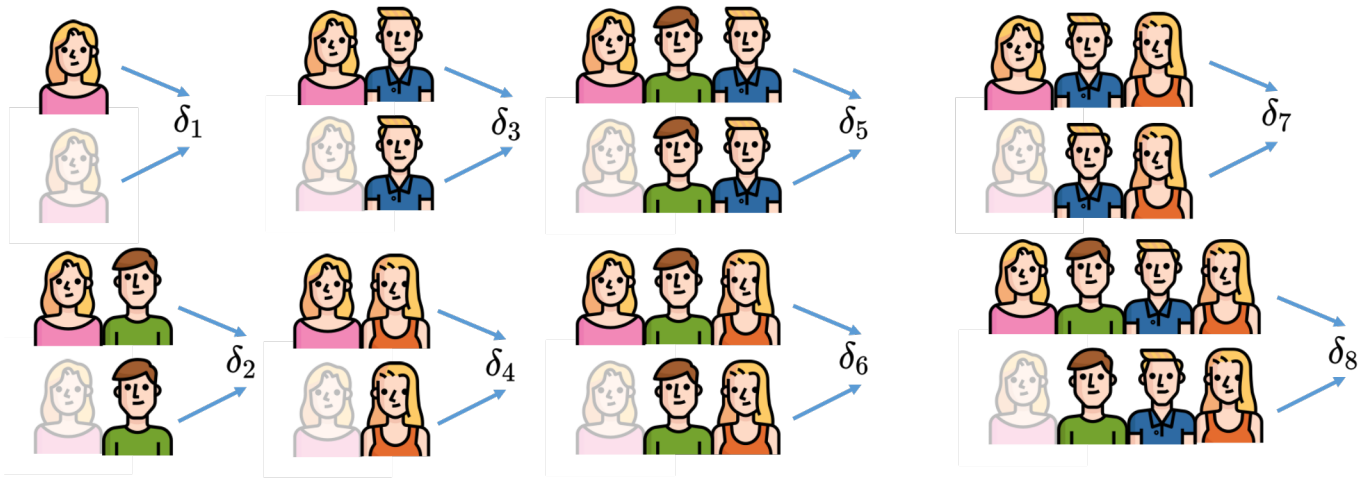
Online Learning models are proposed to address these dynamics.

Traditional Data Valuation Scheme

Shapley value $$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!}(v(S \cup \{i\}) - v(S))$$

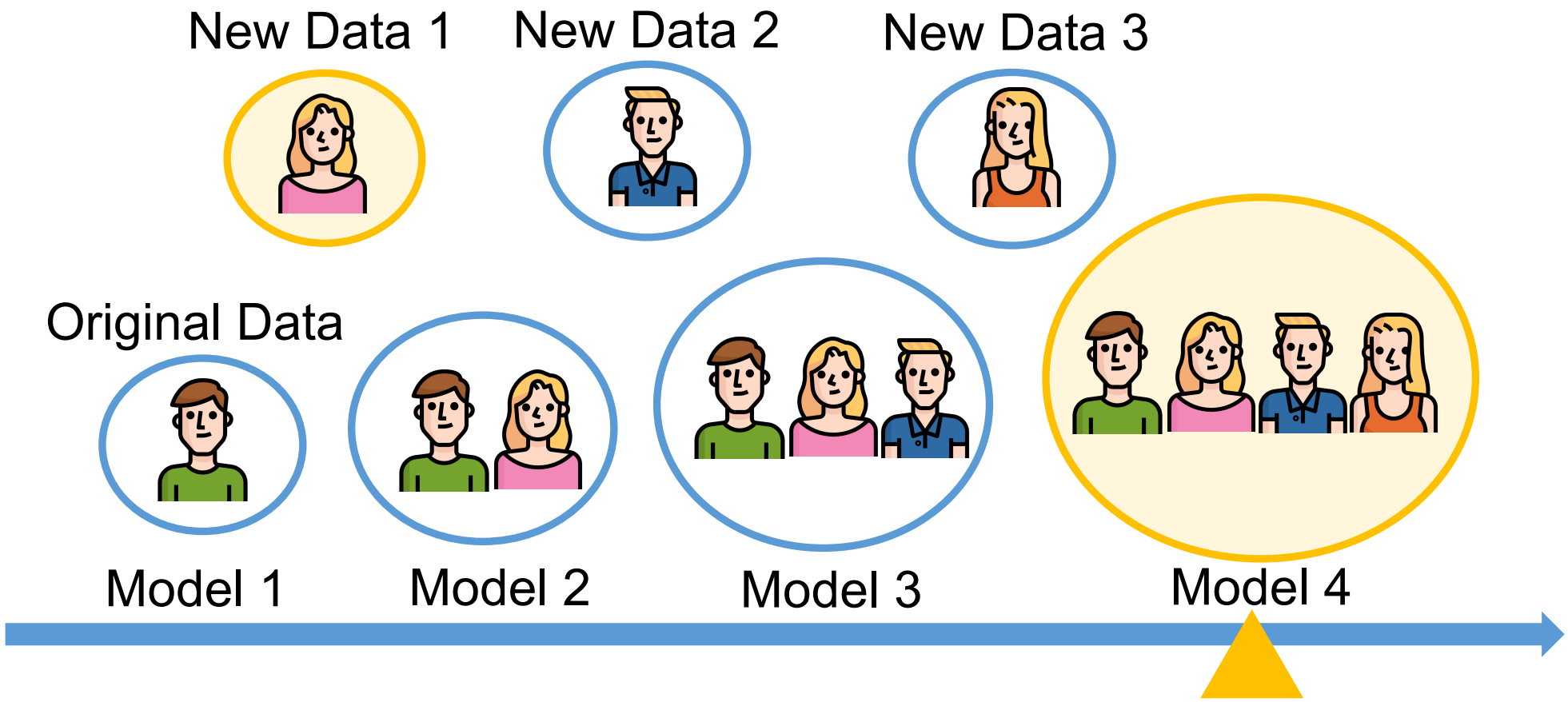
The Shapley value for user 

$$\phi_i = \frac{\delta_1 + \delta_3 + \delta_4 + \delta_5 + \delta_6 + \delta_7 + \delta_8}{8}$$



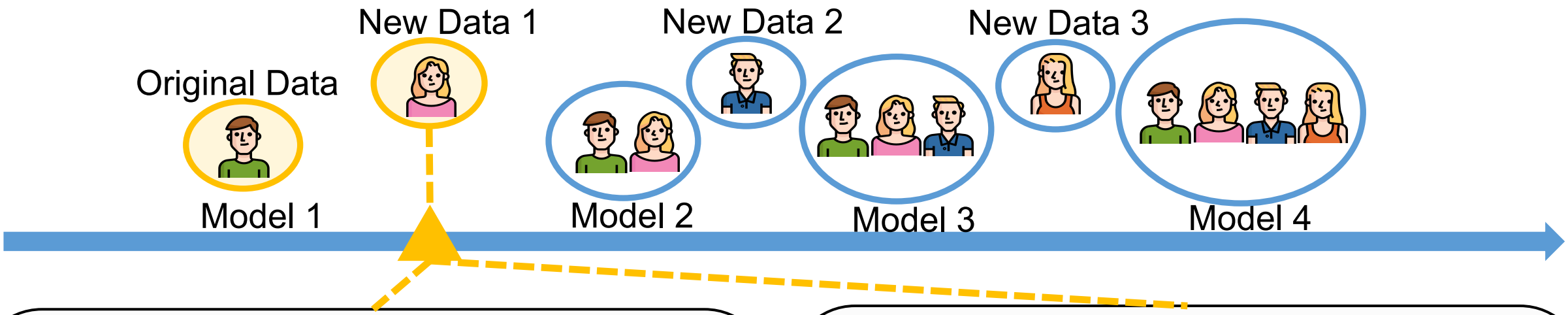
Shapley value is the average expected **marginal contribution** of one user after all possible combinations have been considered.

Traditional Data Valuation Scheme

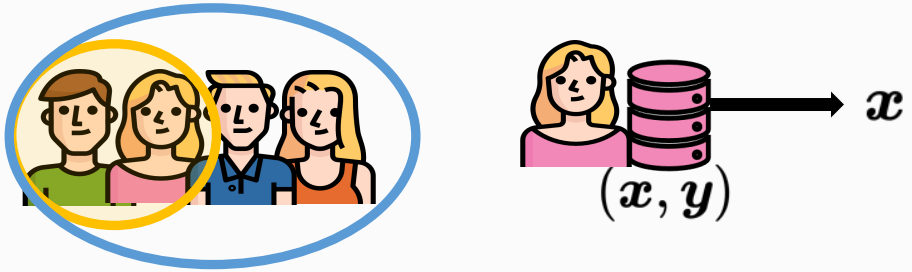


Evaluate data contribution at the **end** of model training

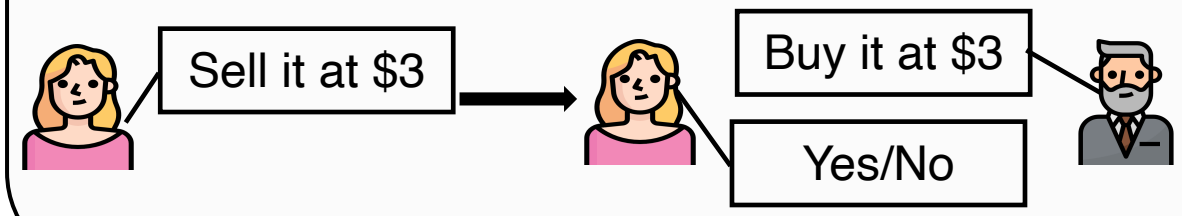
Online Valuation and Pricing



Measure the **data valuation in an online manner**, based on the currently collected data



Determine the **data pricing in an online manner**, based on the data valuation



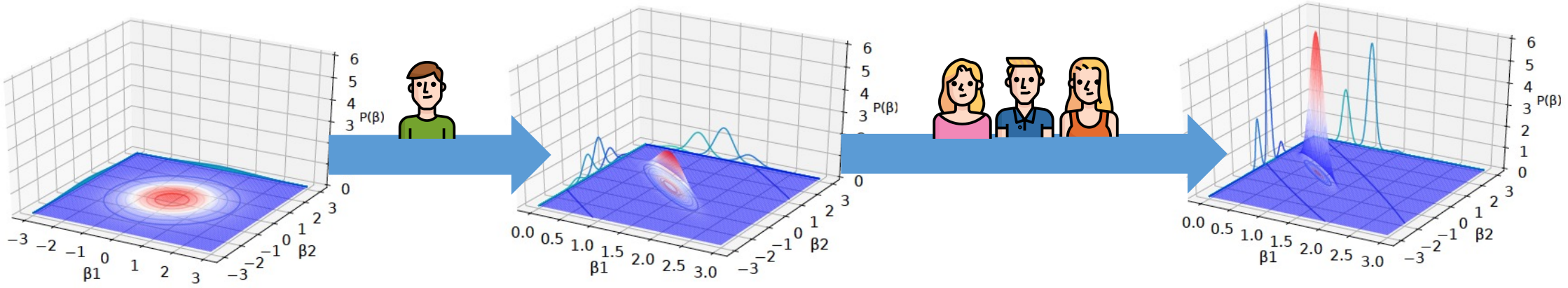
2 Data Valuation

Data Valuation: Bayesian Perspective

Bayesian theorem $$P(\boldsymbol{\beta}|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|\boldsymbol{\beta})P(\boldsymbol{\beta})}{P(\boldsymbol{Y})}$$

Prior

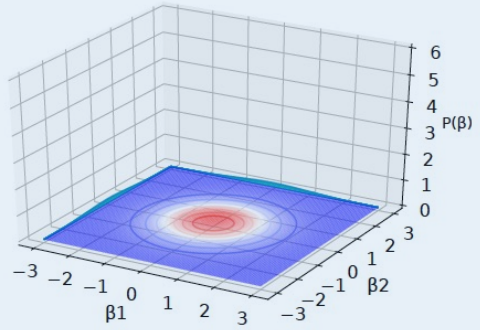
Posterior



Data Valuation can be measured by the change of the model **parameter distribution**, which can represent the evolution of the model training process.

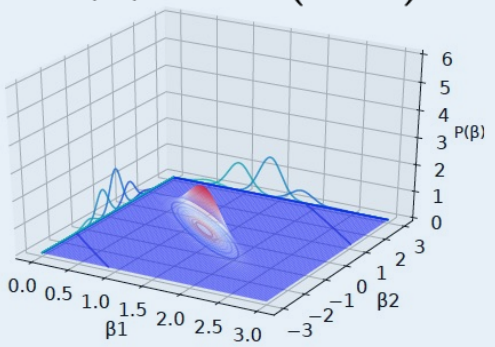
Example: Ridge Regression

$$P(\beta) \sim \mathcal{N}(0, \ell^2 \mathbf{I})$$



$(\boldsymbol{X}, \boldsymbol{Y})$

$$P(\beta) \sim \mathcal{N}(\bar{\beta}, \Sigma)$$



$$\Sigma = \left(\frac{1}{\delta^2} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} + \frac{1}{\ell^2} \mathbf{I} \right)^{-1}$$

$$\bar{\beta} = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} + \frac{\ell^2}{\delta^2} \mathbf{I} \right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{Y}$$

Ridge regression can be explained as a type of Bayesian Linear Regression, in which the **prior** probability of the parameters satisfy **Gaussian** distribution.

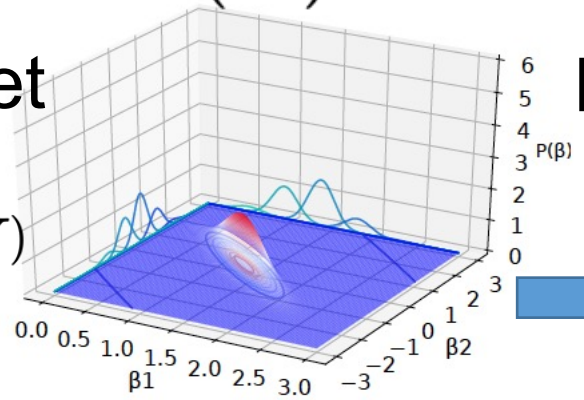
Example: Ridge Regression

Differential entropy $H(\boldsymbol{X}) = \frac{1}{2}\ln\left((2\pi e)^d \det(\Sigma_{\boldsymbol{X}})\right)$

Original Dataset

$(\boldsymbol{X}, \boldsymbol{Y})$

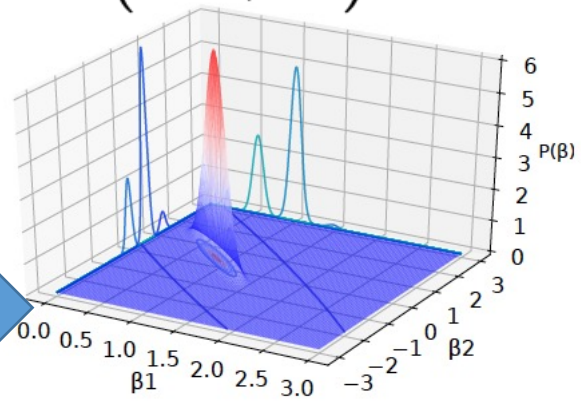
$H(\boldsymbol{X})$

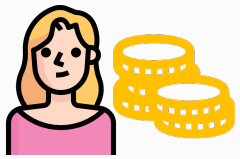


New Dataset

$(\boldsymbol{x}, \boldsymbol{y})$

$H(\boldsymbol{X}+\boldsymbol{x})$



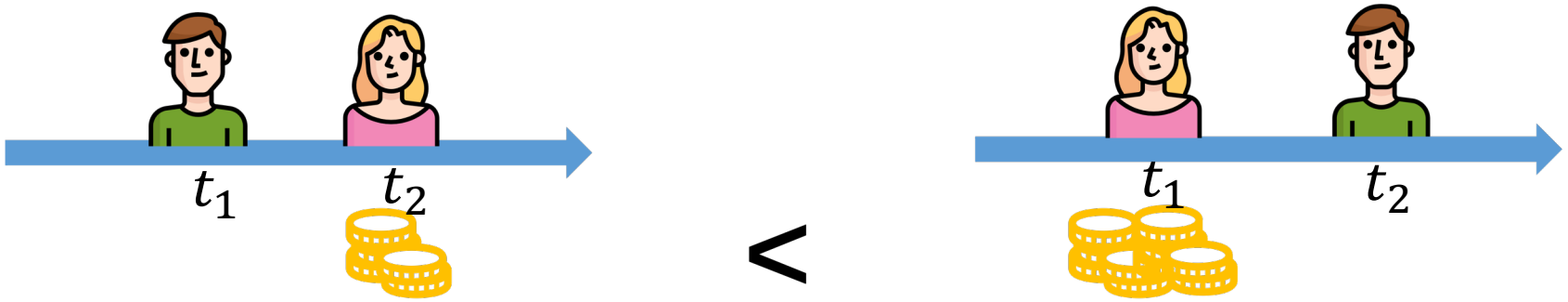


$$G_{\boldsymbol{X}}(\boldsymbol{x}) = H(\boldsymbol{X}) - H(\boldsymbol{X}+\boldsymbol{x}) = \frac{1}{2}\ln\left(1+\boldsymbol{x}^{\mathrm{T}}\Sigma_{\boldsymbol{X}}\boldsymbol{x}\right)$$

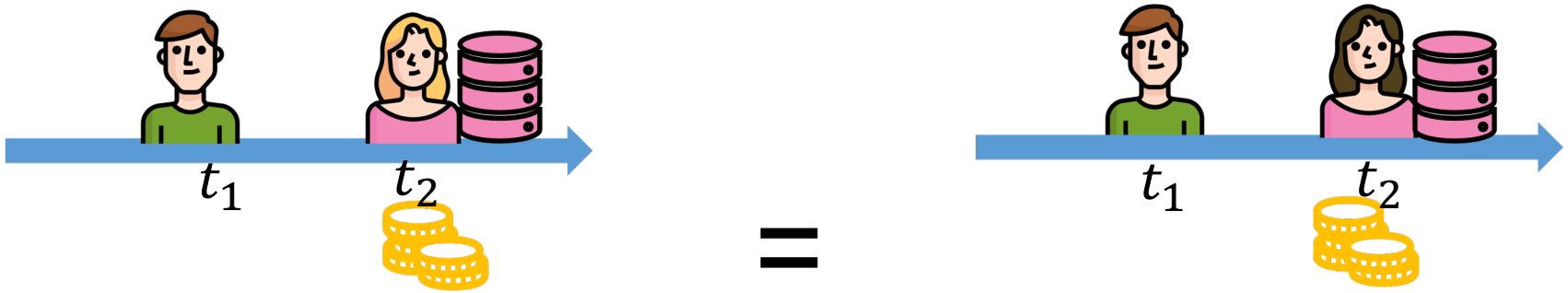
Quantify the data valuation by the change of the model parameter distribution's **entropy**

Data Valuation: Properties

Submodular: The contribution of data is **diminishing marginal**.

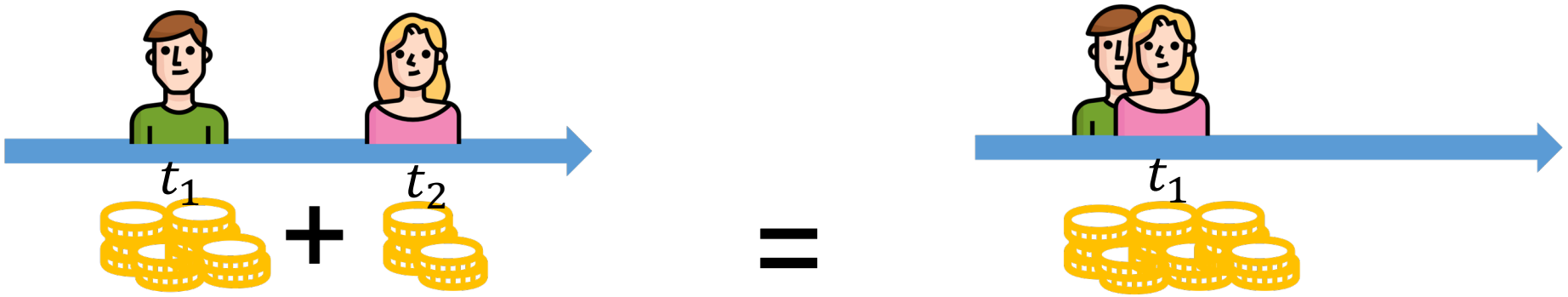


Online Fairness: The datasets which are identical in what they contribute to the model have the **same** valuation in an online manner.



Data Valuation: Properties

Additivity: The total contribution of all the datasets is the sum of the individual contribution of each dataset.

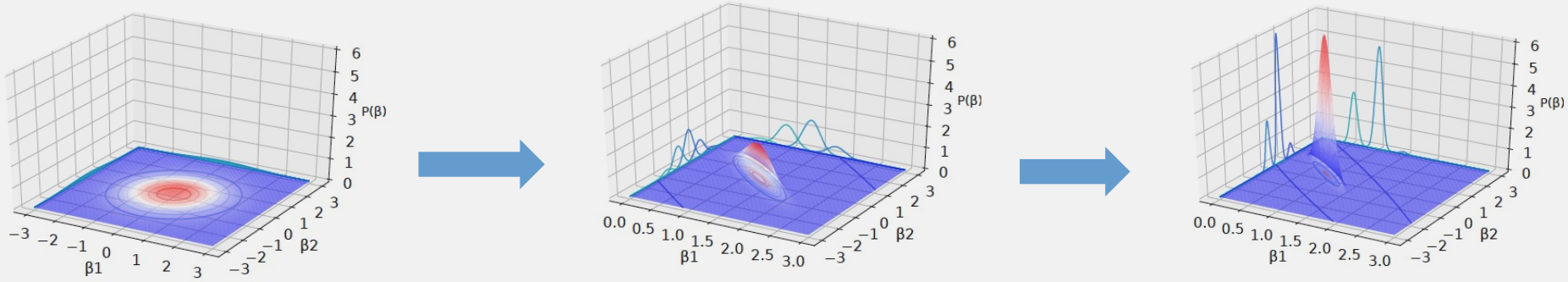


Group Rationality: The valuation of the entire dataset is **completely distributed** among all data.

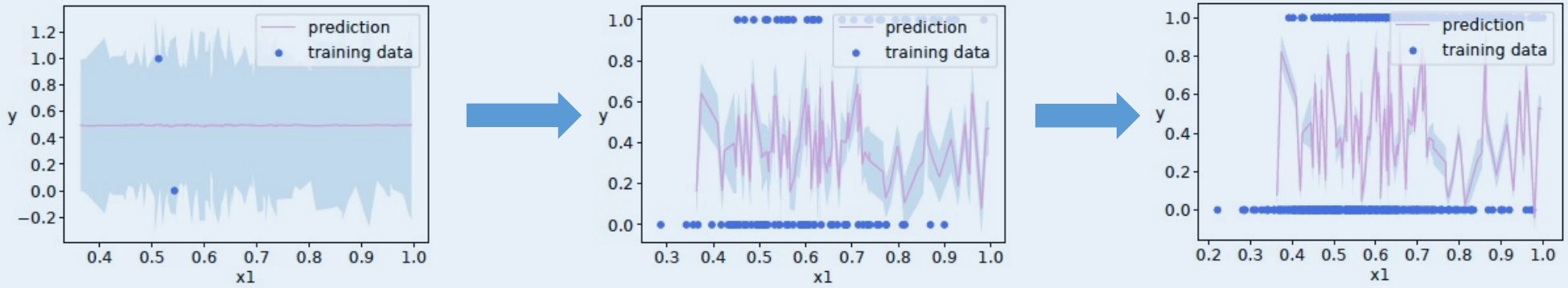
Inferrability: Each data's valuation only depends on the data features $\boldsymbol{x}$.

General Models: BNN/GP

Parameter Distribution



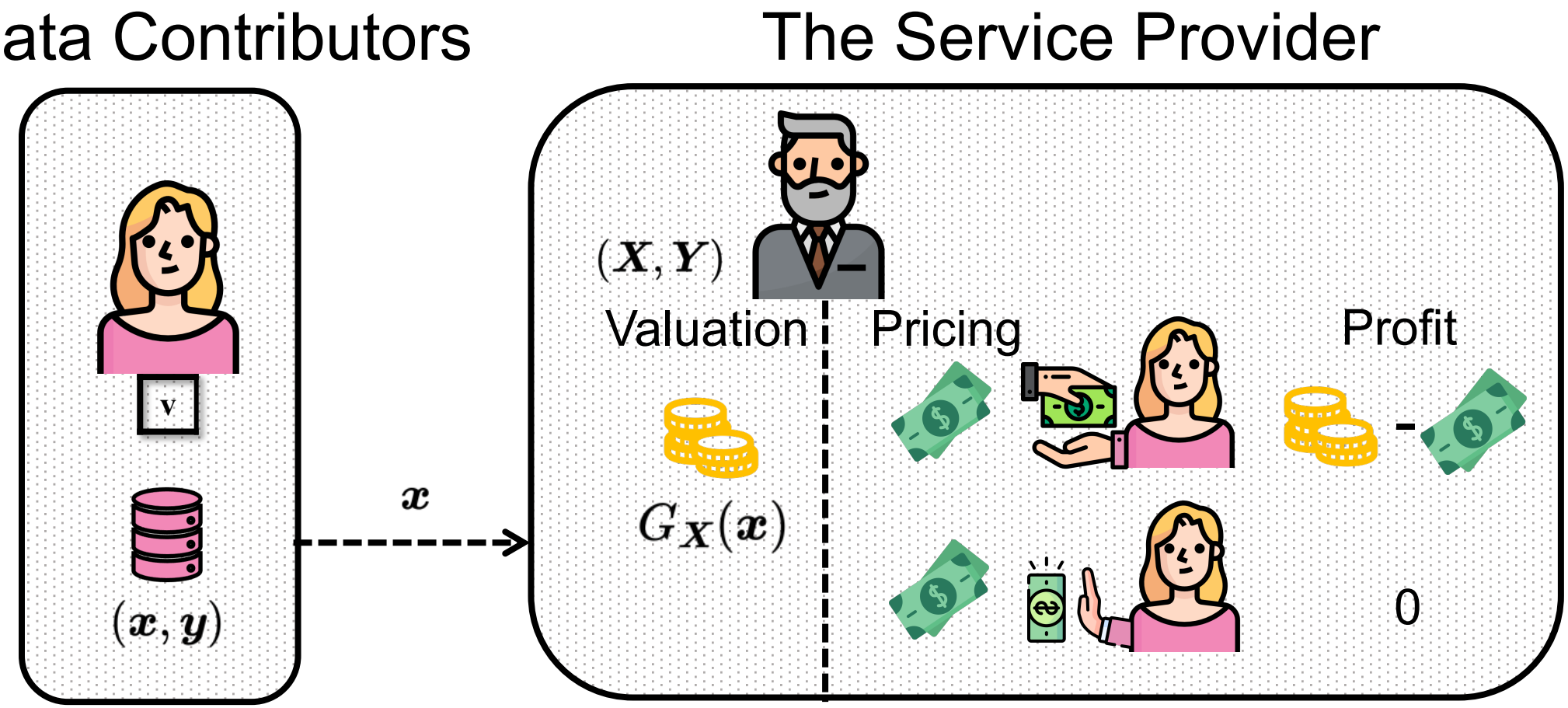
Predictive Distribution



Calculate from parameter space to prediction space.
Avoid exponentially hard with dimensionality.

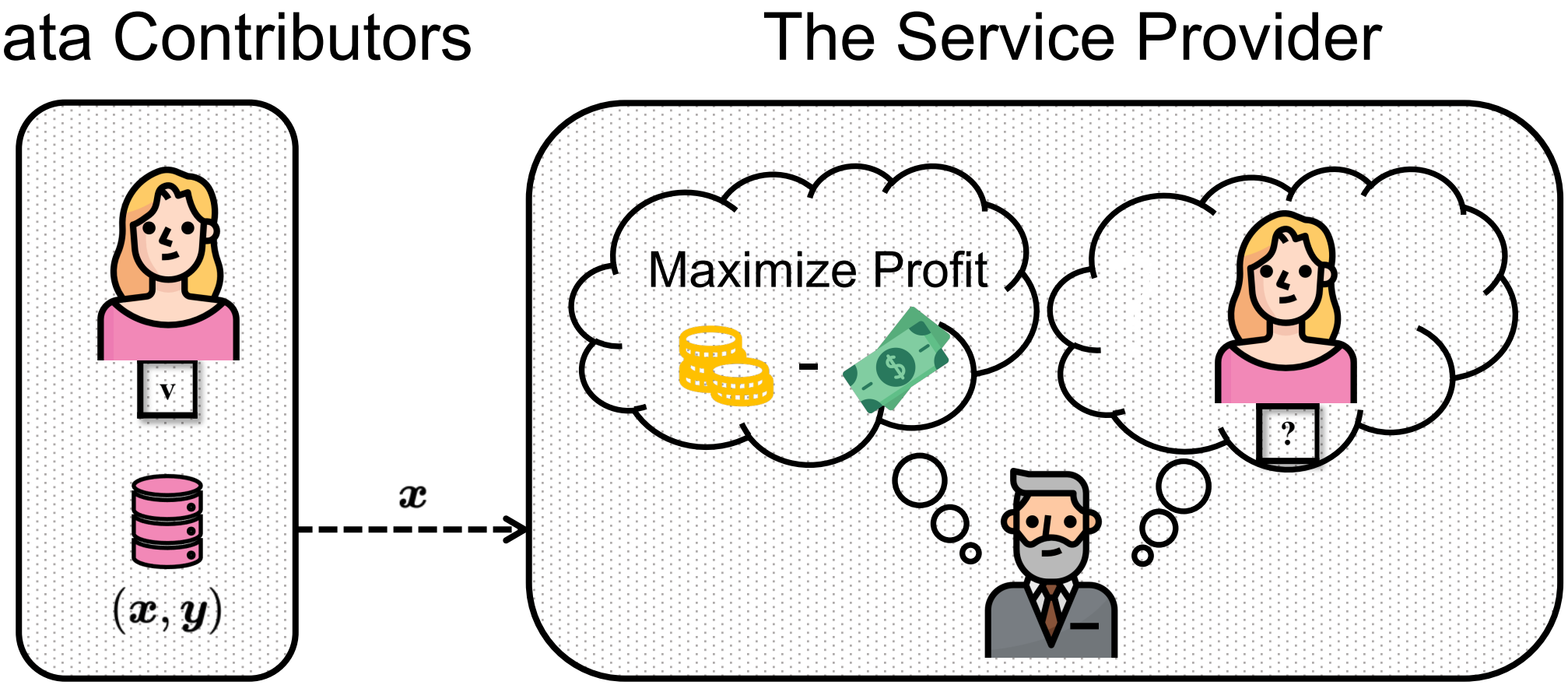
3 Data Pricing

Data Transaction Process in mHealth



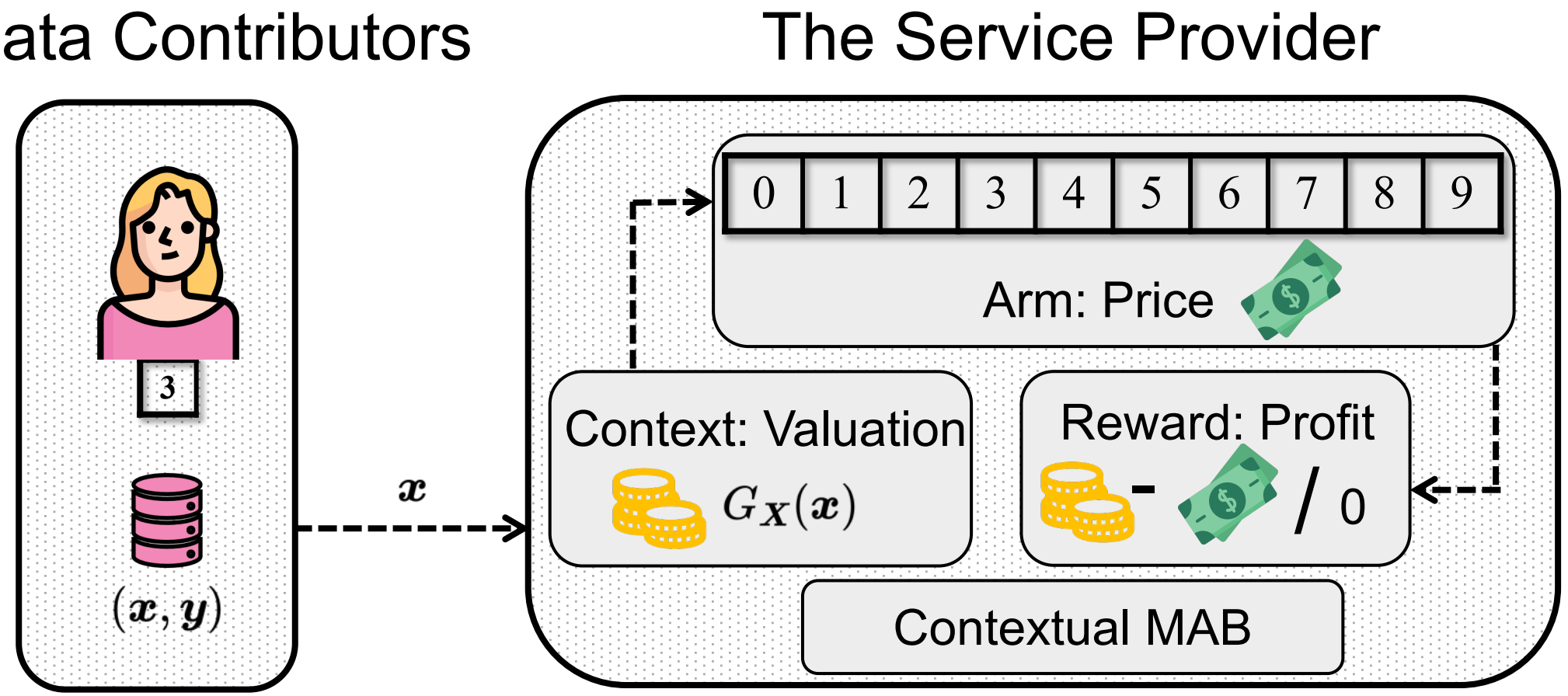
A **posted** pricing mechanism in an online manner

Data Pricing: Goal



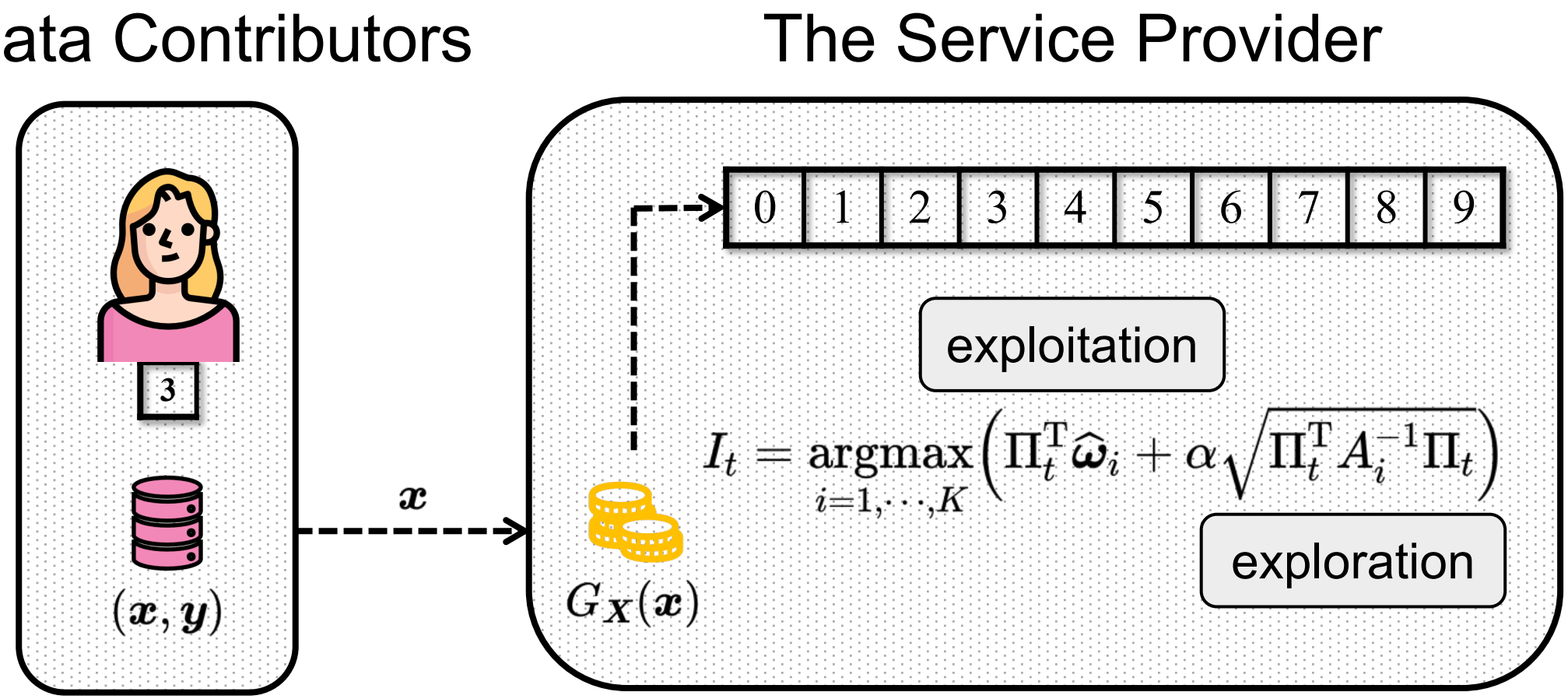
Design a **profit-maximizing** data pricing mechanism with **incomplete information**

Data Pricing: Problem Modeling



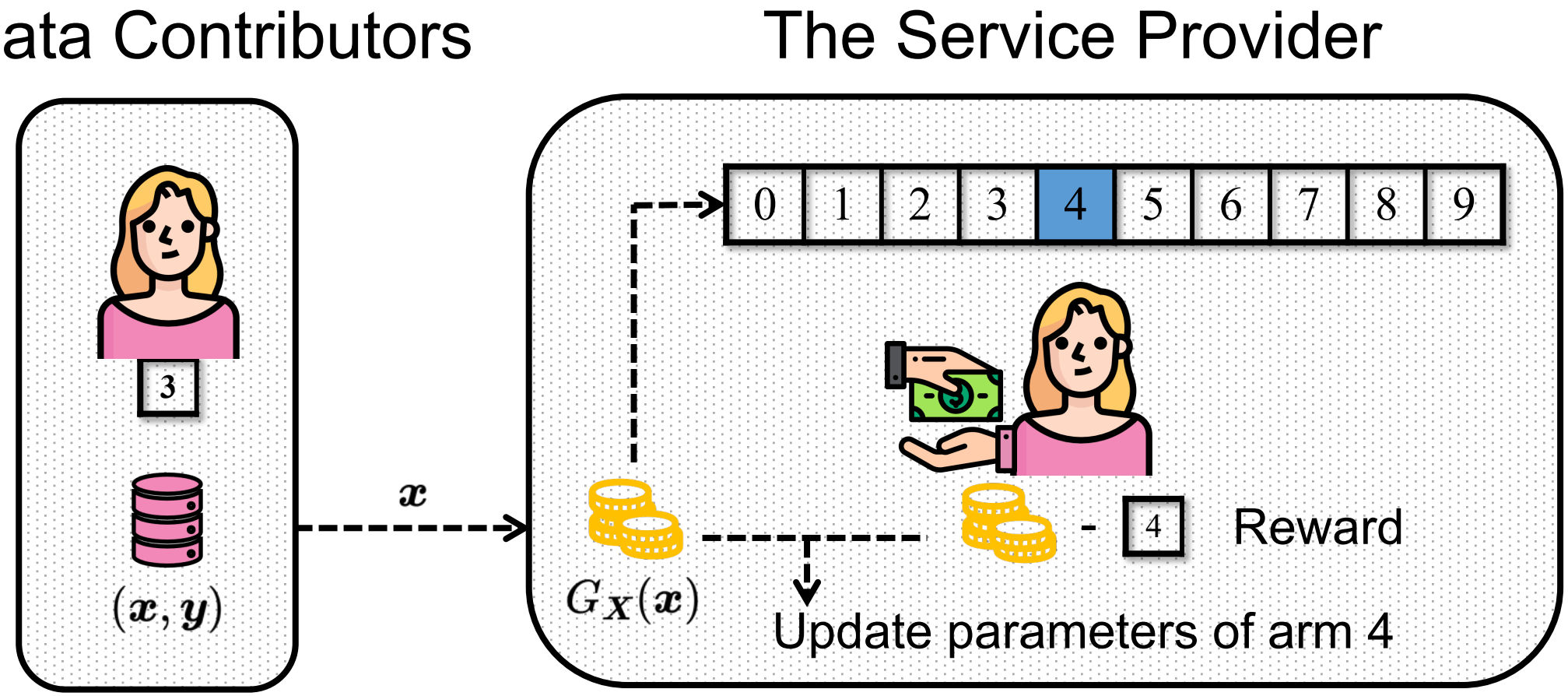
Model the payment determination process as a **contextual multi-armed bandit** with the goal of profit maximization

Data Pricing: Pricing Method



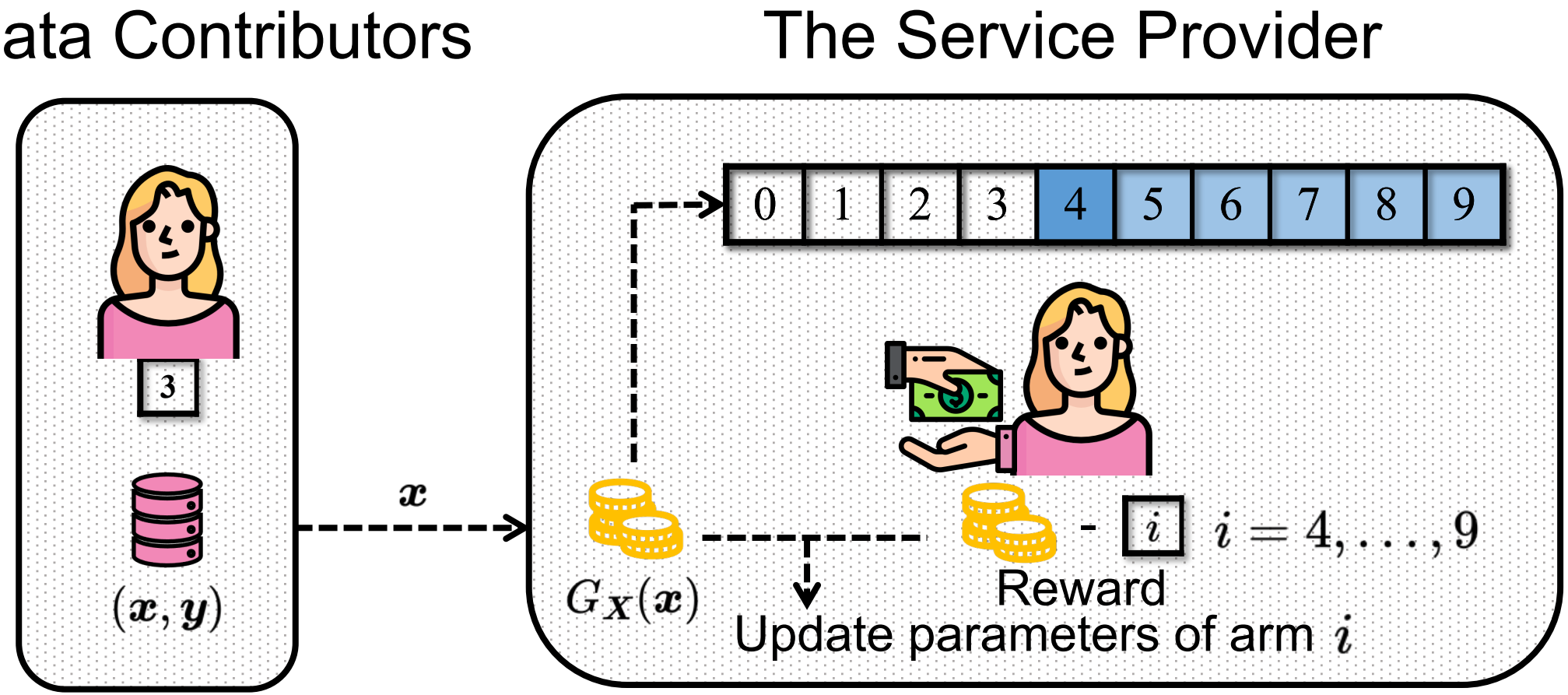
Conduct an **exploitation** and **exploration** process by choosing the arm with highest **upper confidence bound**

Data Pricing: Original Method



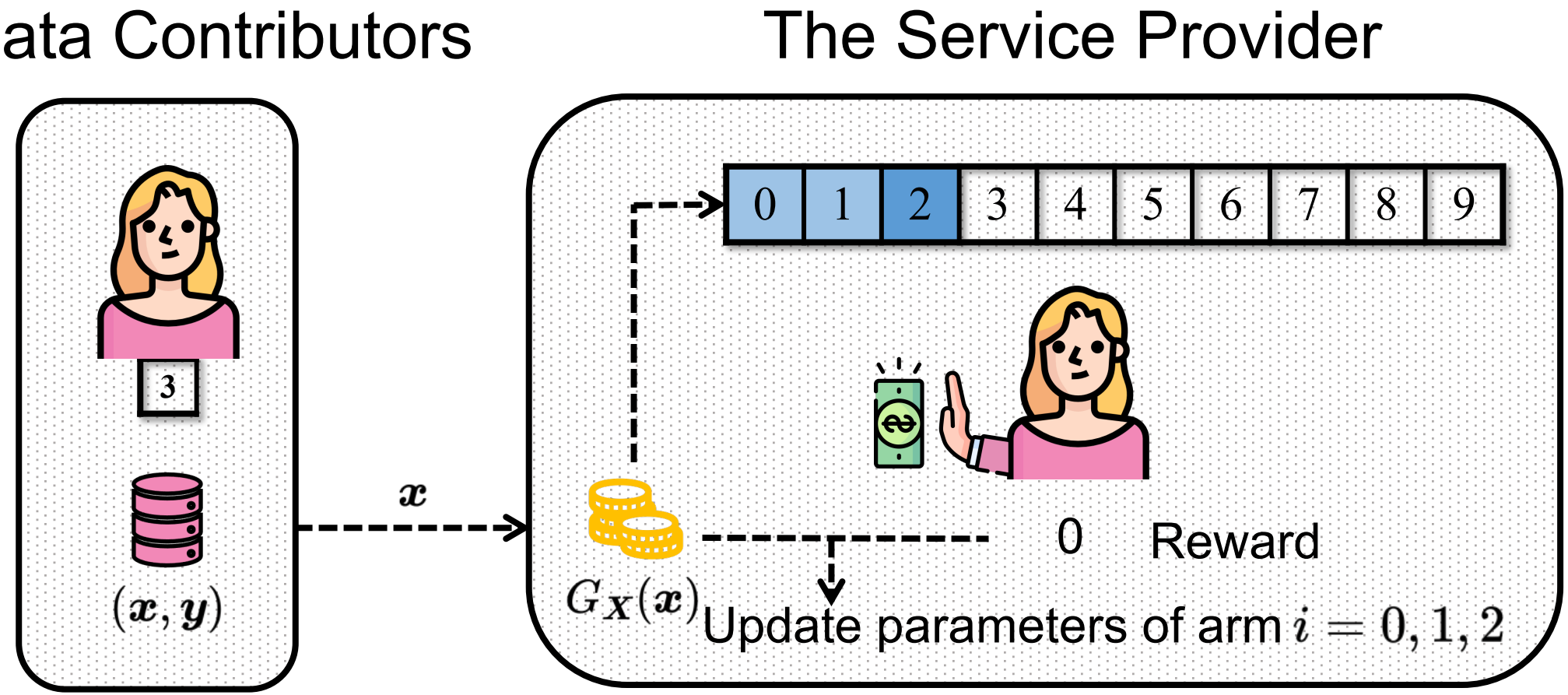
A data contributor is willing to sell her data at $4.

Data Pricing: Monotonic



If a data contributor is willing to sell her data at $4, she will also be willing to sell it at price higher than $4.

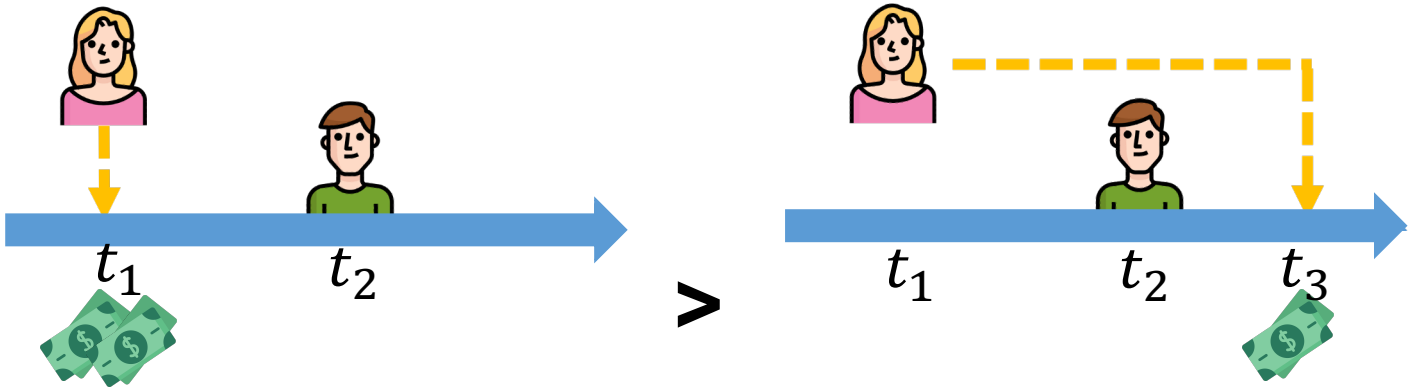
Data Pricing: Monotonic



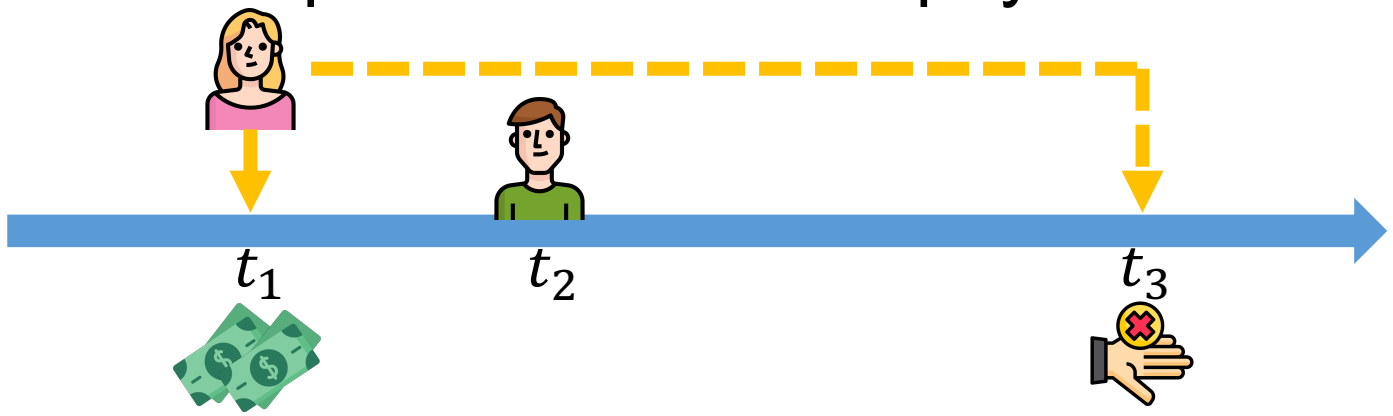
If a data contributor is unwilling to sell her data at $2, she will also be unwilling to sell it at price lower than $2.

Data Pricing: Properties

Incentive Mechanism: Earlier data contributors will get more payment.

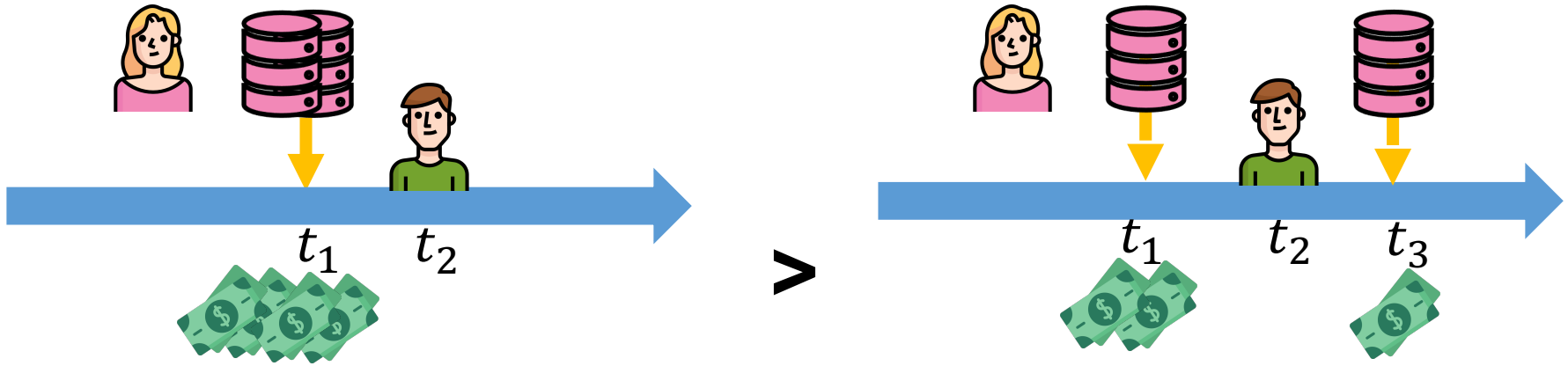


Robust to Strategic Behavior: VAP discounts similar data's valuation and subsequent contributor's payment.



Data Pricing: Properties

Arbitrage-Freeness: Splitting the data and uploading it in parts won't yield any higher payment than uploading the full dataset.



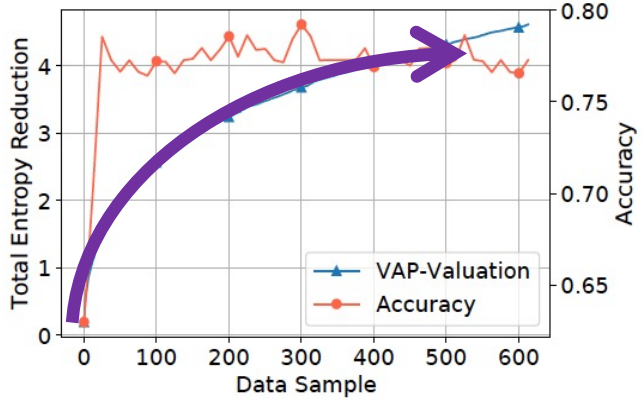
Data Privacy Preserving for mHealth: The label is not involved in the data valuation and pricing, reducing the risk of privacy leakage.

4 Evaluation

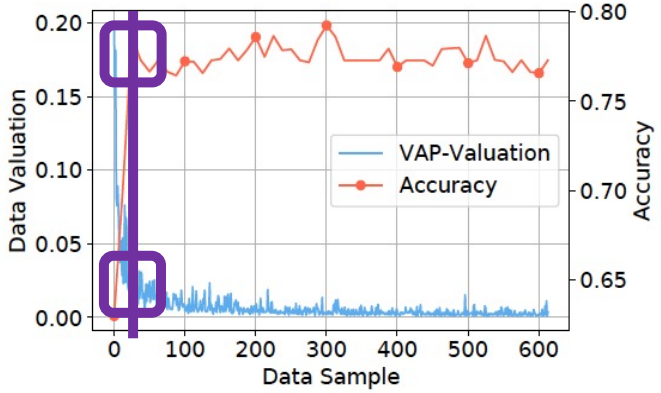
Setups

- Two real-world human behavior data sets:
  - Human Activity Recognition (HAR)
  - Pima Indians Diabetes (PID)
- Machine learning models:
  - Ridge Classification (RC)
  - Gaussian process classification (GPC)
  - Bayesian Neural Networks (BNN)
- Base Line:
  - TMC- Shapley; Gradient Shapley (G-Shapley); Random
  - UCB1; LinUCB; Random; Half-Fix; Half-Valuation

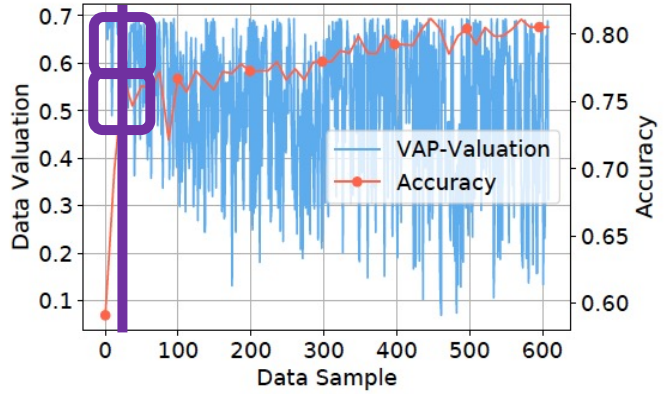
VAP-Valuation on Different Models



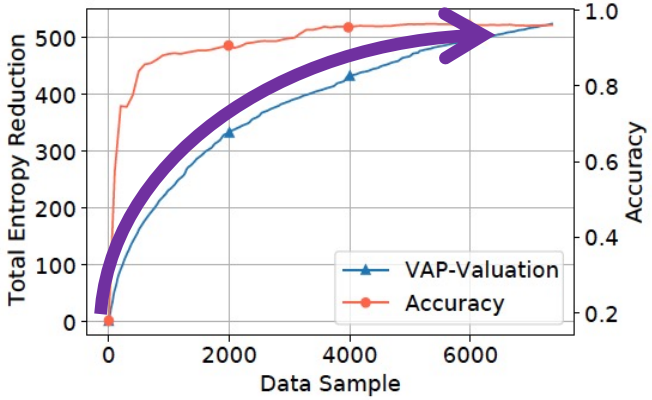
(a) RC on PID



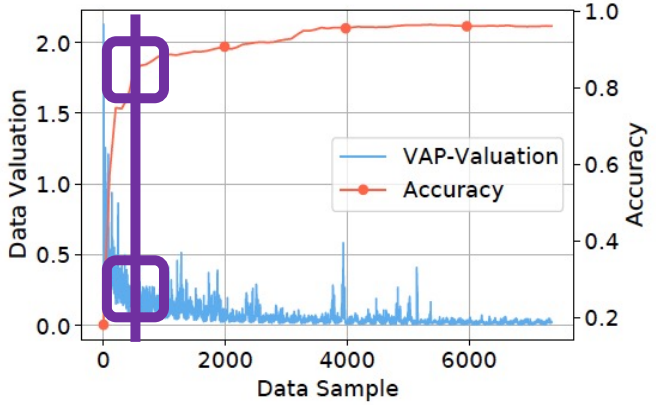
(b) RC on PID



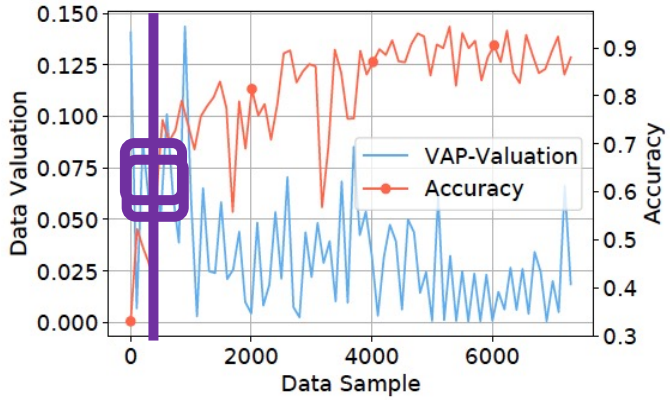
(c) GPC on PID



(d) RC on HAR



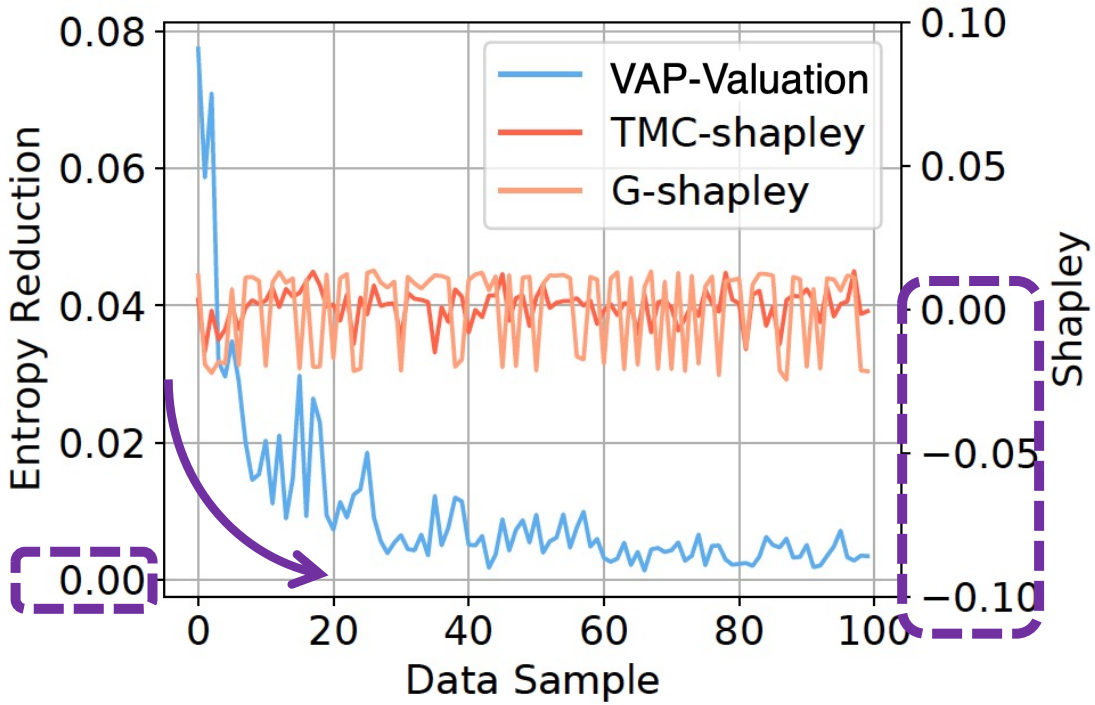
(e) RC on HAR

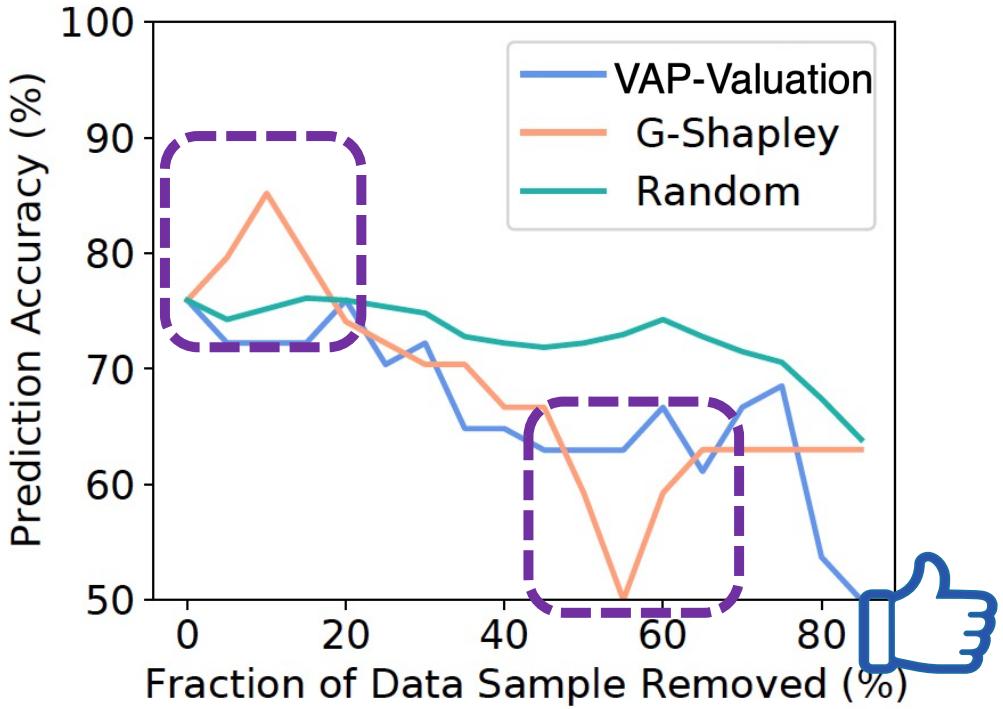


(f) BNN on HAR

Trends of total entropy reduction and accuracy boost are consistent.
The turning points of data valuation and accuracy are close.

Different Data Valuation Metrics

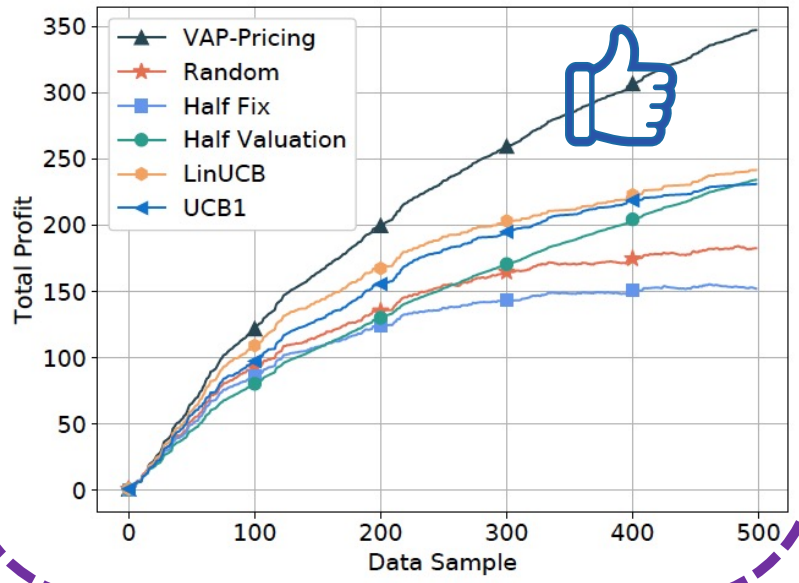




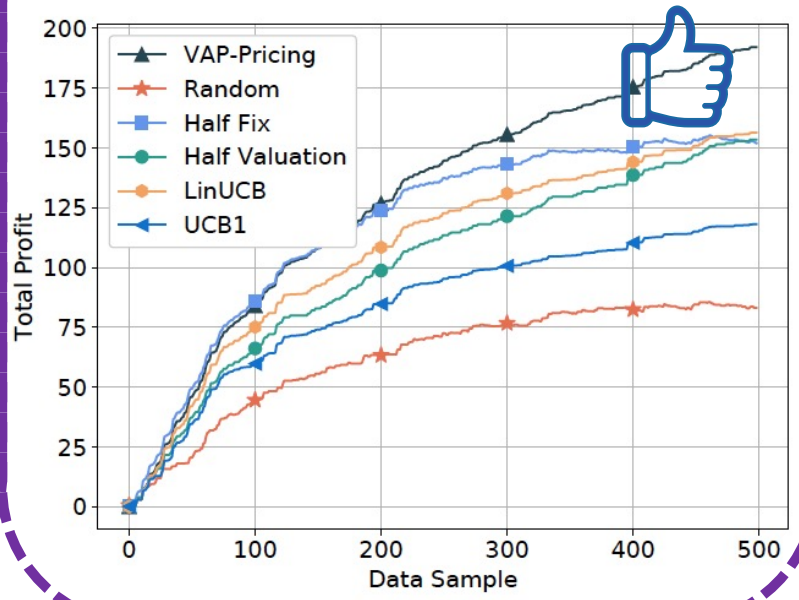
VAP-Valuation is diminishing marginal.
VAP-Valuation is always positive.
VAP is more stable for online learning than others.

Different Data Pricing Mechanisms: Profit

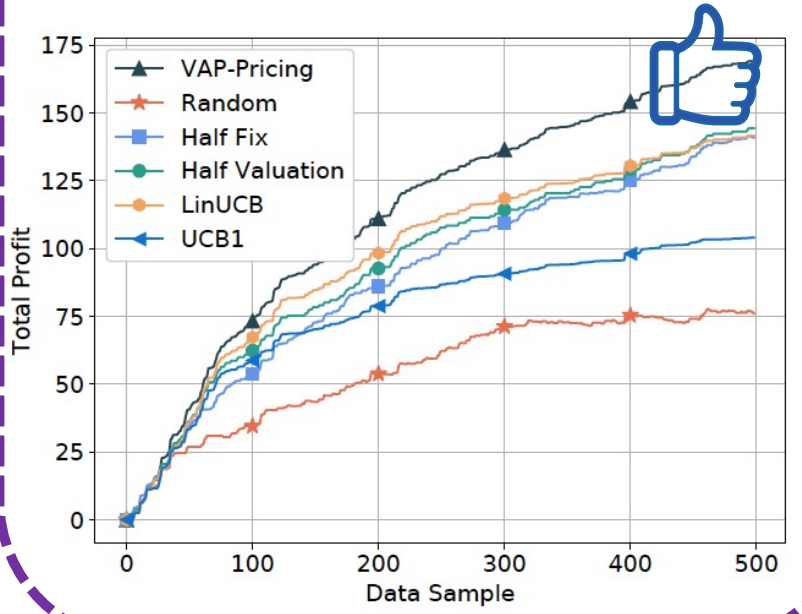
A uniform distribution within [0, 1]



A constant distribution as 0.5



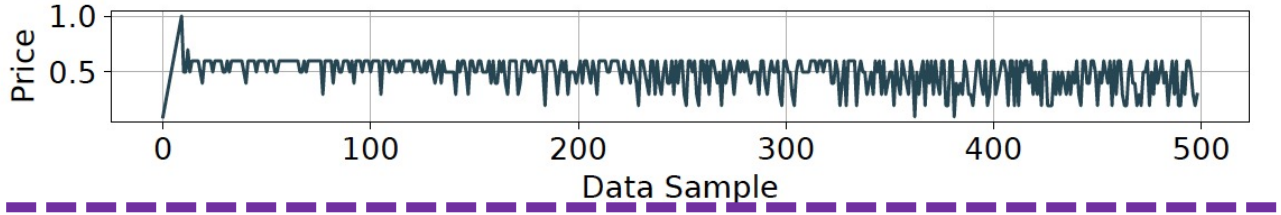
An approximately normal distribution within [0, 1], where the mean is 0.5, and variance is 0.1.



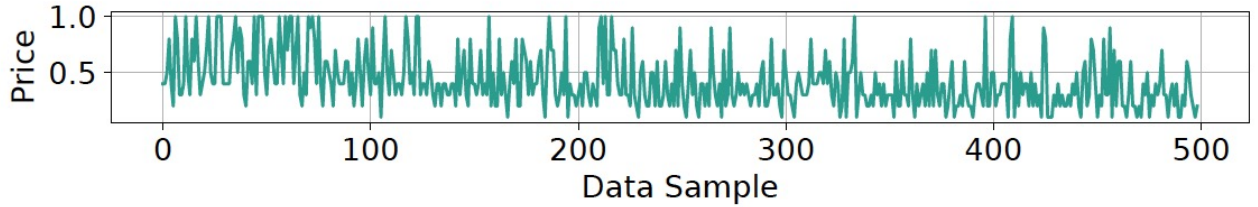
VAP outperforms other mechanisms in terms of extracted profit.

Different Data Pricing Mechanisms: Price

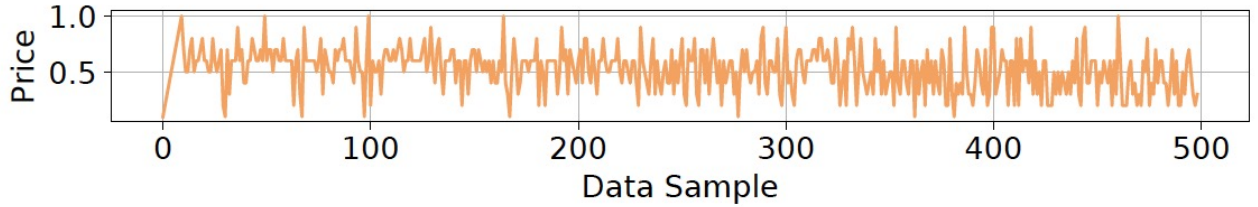
VAP



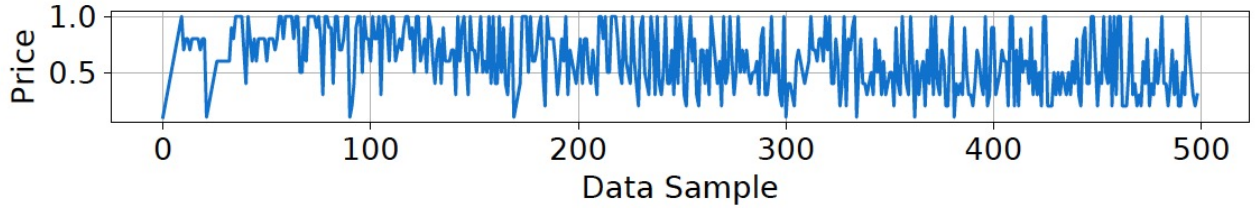
Half Valuation



LinUCB



UCB1



VAP-pricing maintains the downward trend of data valuation.

VAP-pricing has minimal price volatility.

5 Summary

Summary

- Proposed the first **online** metric of data valuation, which is quantified by the **entropy** of the distributions over model parameters.
- Proposed an **online** data pricing method under a **contextual multi-armed bandit** framework to maximize profit.
- Validated VAP over two real-world mHealth data sets: VAP outperforms better in **computational complexity** and **extracted profit**.

Thanks for watching!

Please refer to our paper for more details!