Virtual Conference
KDD2021
August 14th – 18th

Multi-facet Contextual Bandits: A Neural Network Perspective

Yikun Ban
UIUC
yikunb2@illinois.edu

Jingrui He
UIUC
jingrui@illinois.edu

Curtiss B. Cook
Mayo Clinic Arizona
cook.curtiss@mayo.edu

ILLINOIS

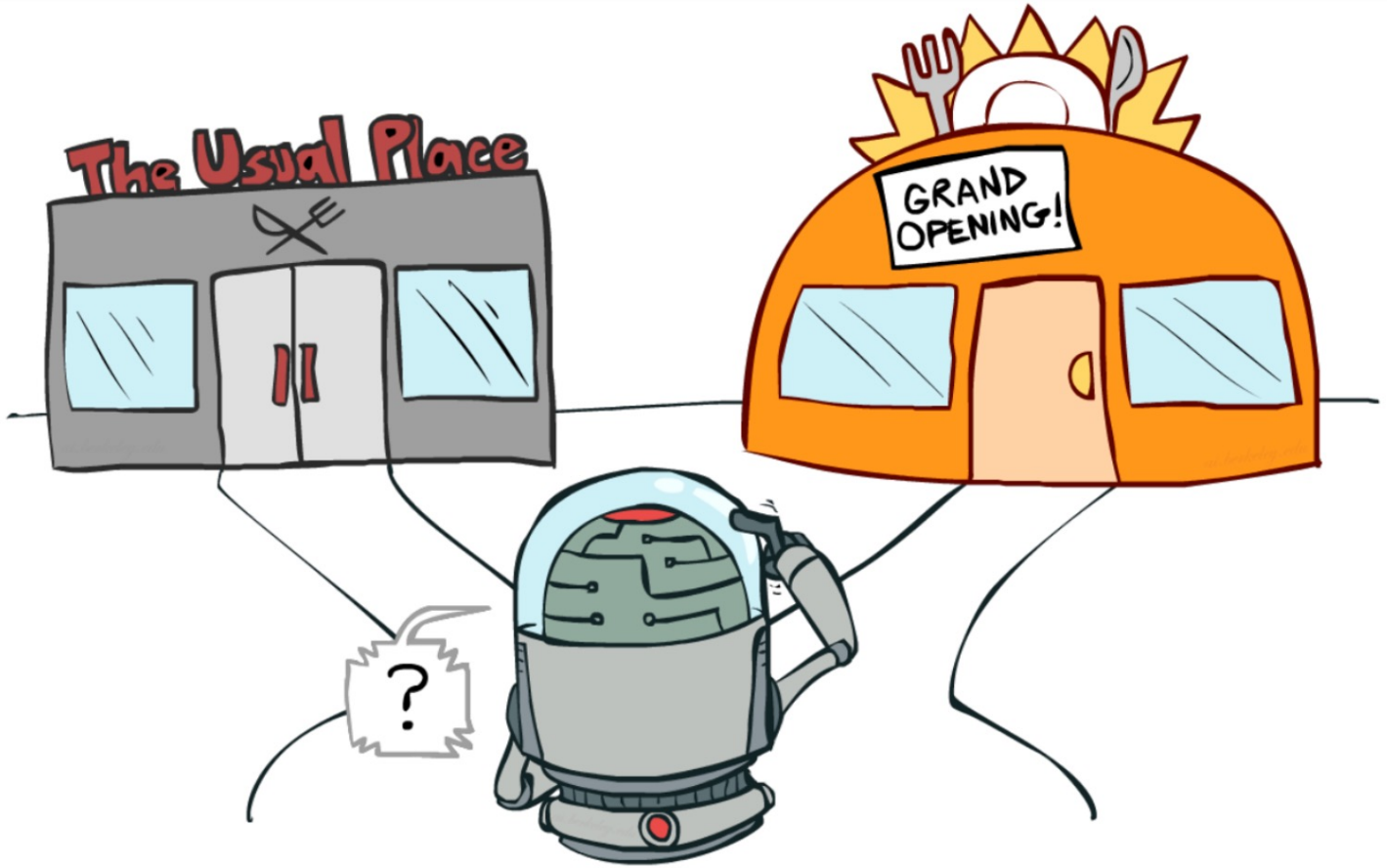
Roadmap

- Background
- Motivation and Problem Definition
- Proposed Algorithm
- Theoretical Analysis
- Experiments and Conclusion

Exploitation VS Exploration

- The dilemma of *Exploitation and Exploration* is everywhere.

Where to eat?

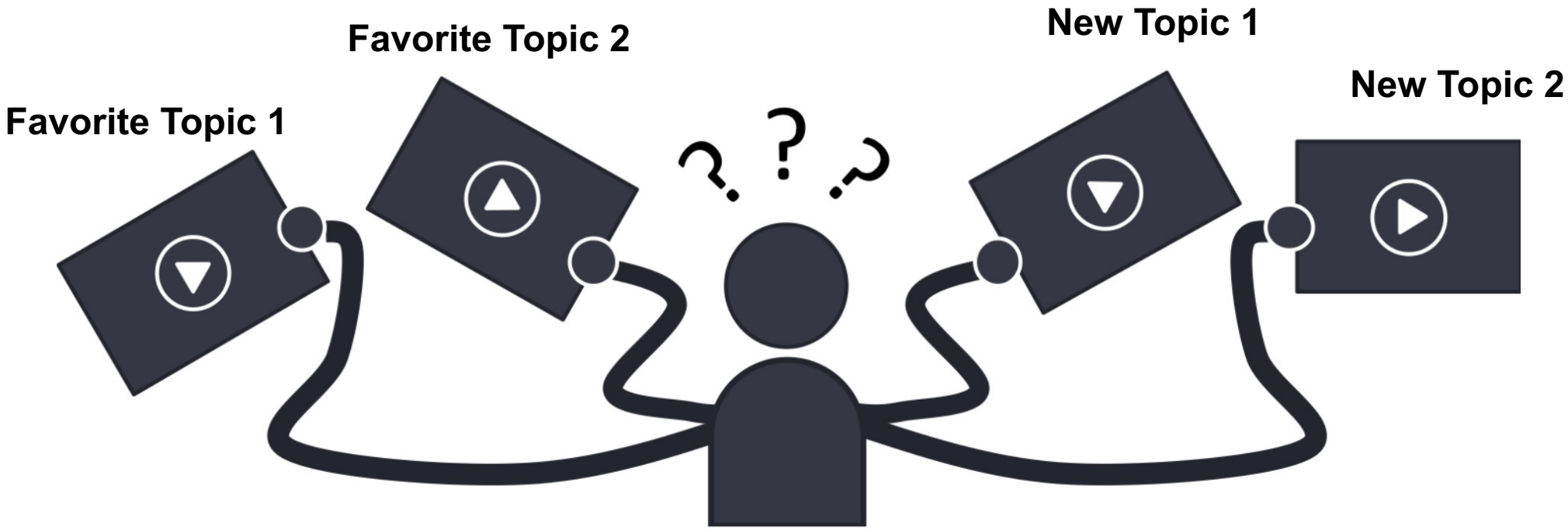


UC Berkeley CS 188: http://ai.berkeley.edu/lecture_slides.html

Exploitation VS Exploration

- **The dilemma of *Exploitation and Exploration* is everywhere.**

Personalized Recommendation:



SpotX: Introduction to Multi-Armed Bandits with Applications in Digital Advertising

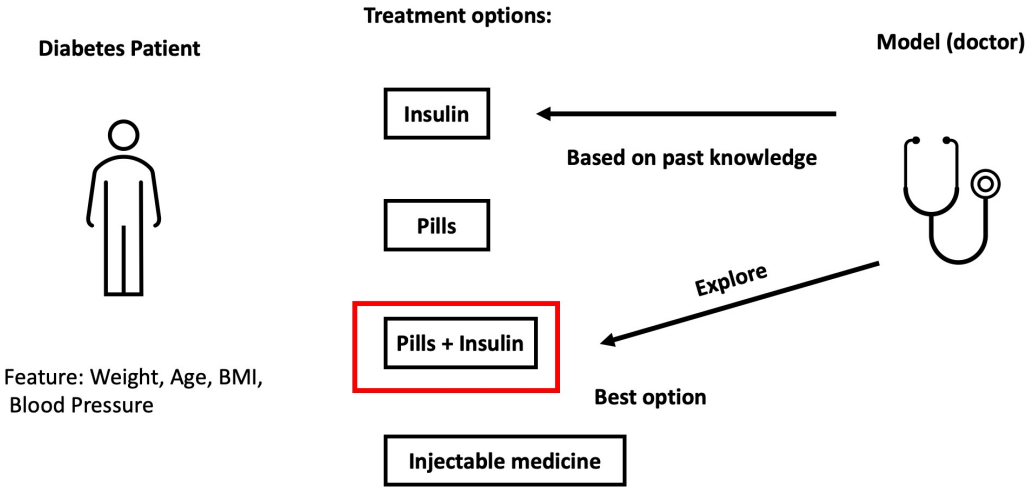
Exploitation VS Exploration

➢ **The dilemma of *Exploitation and Exploration* is everywhere.**

Online Advertising:

[Li et al. 2010]

Clinic Trial:



Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010, April). A contextual-bandit approach to personalized news article recommendation. WWW'10.

Multi-armed Bandit

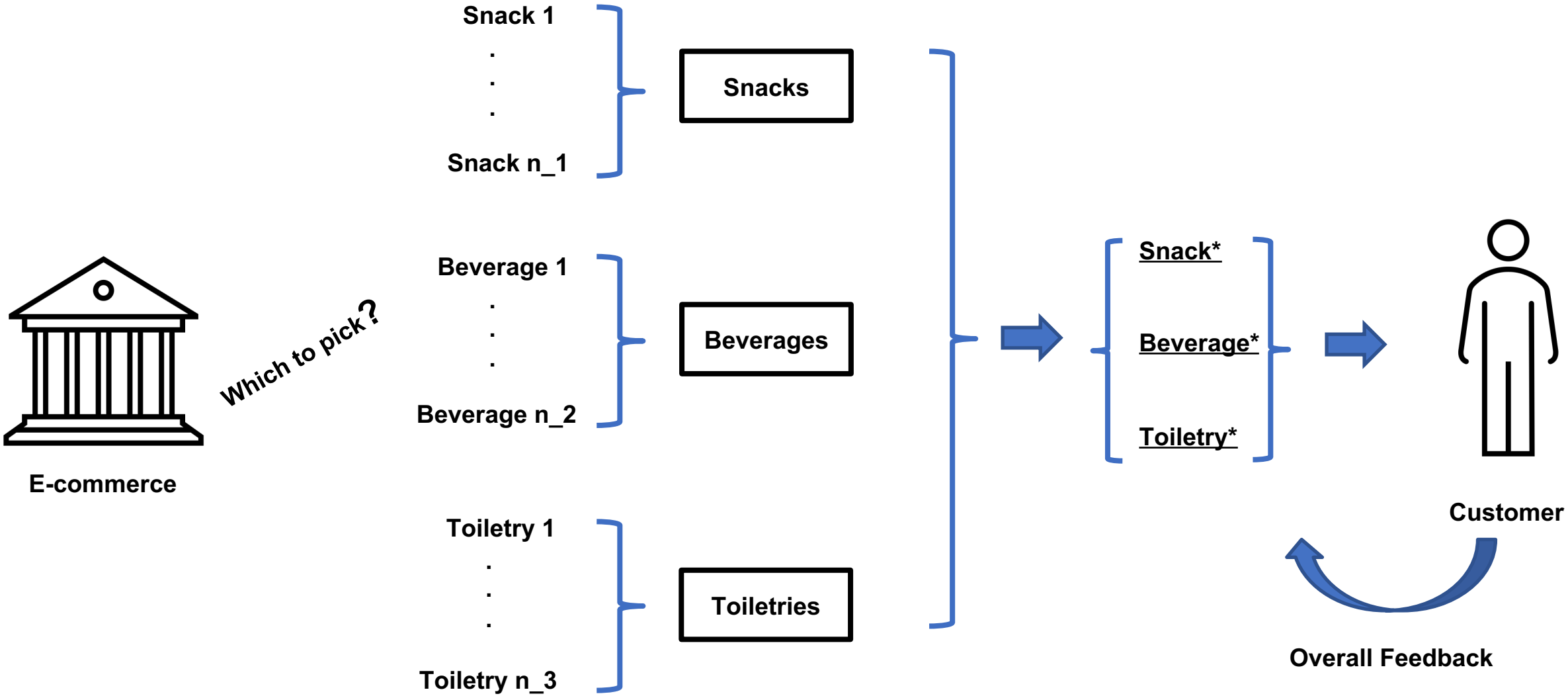
- **Multi-armed bandits provide powerful tools for solving the dilemma of exploitation and exploration.**

  - ✓ Can adapt to personalized recommendation, online advertising, etc.
  - ✓ Existing algorithms:
    - ○ Epsilon – Greedy
    - ○ **Upper Confidence Bound** ← This paper
    - ○ Thompson Sampling

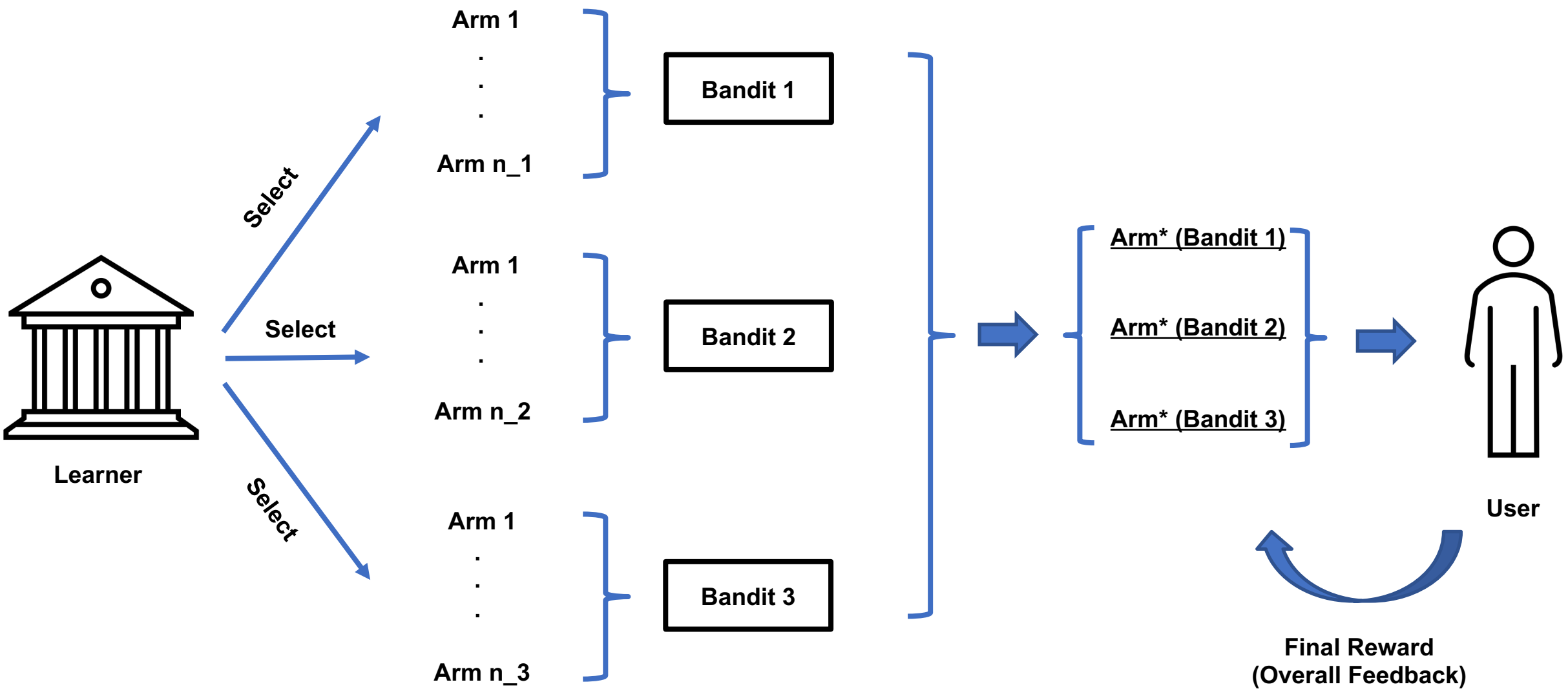
Roadmap

- Background
- Motivation and Problem Definition
- Proposed Algorithm
- Theoretical Analysis
- Experiments and Conclusion

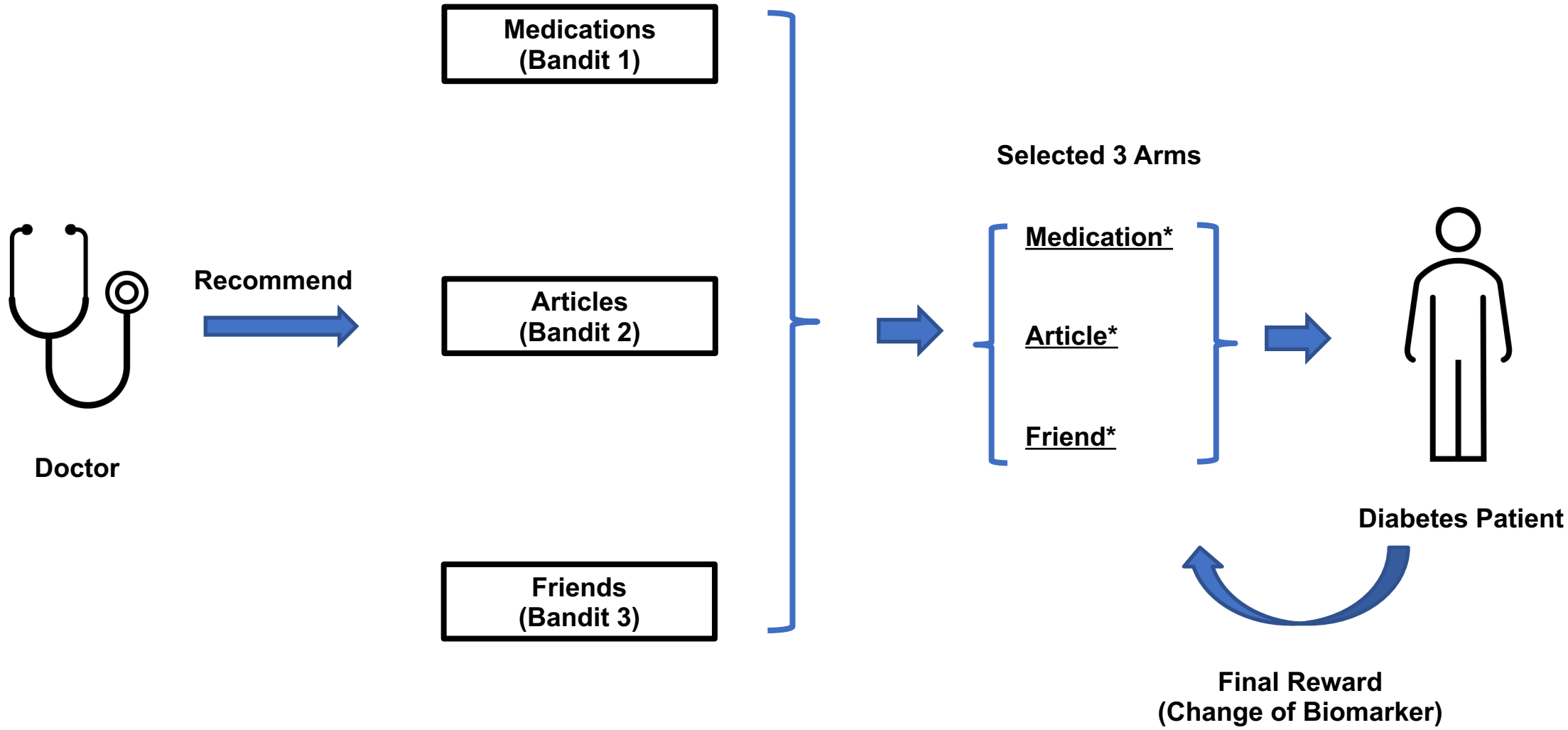
Motivated Case: Promotion Campaign



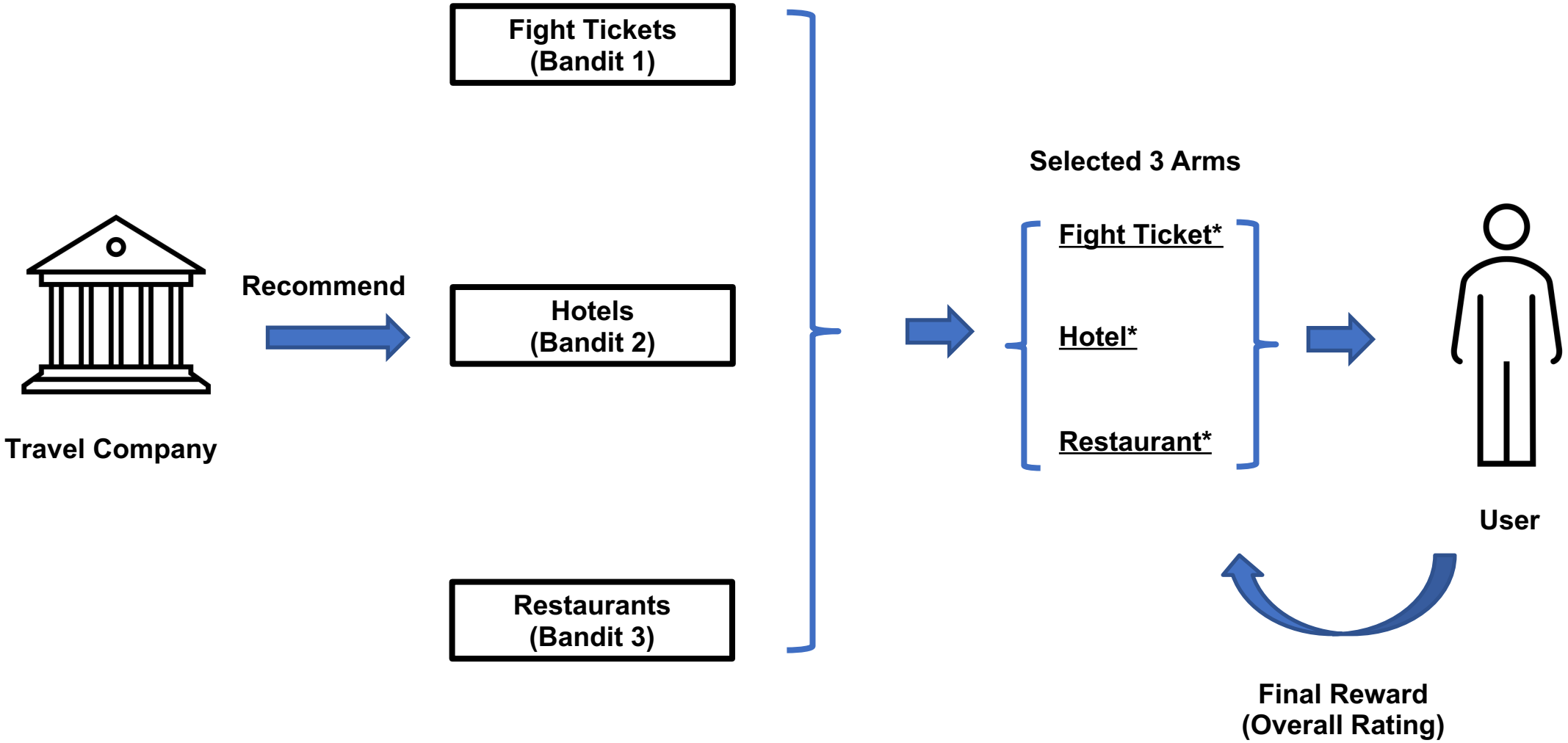
Novel Problem: Multi-facet Contextual Bandits



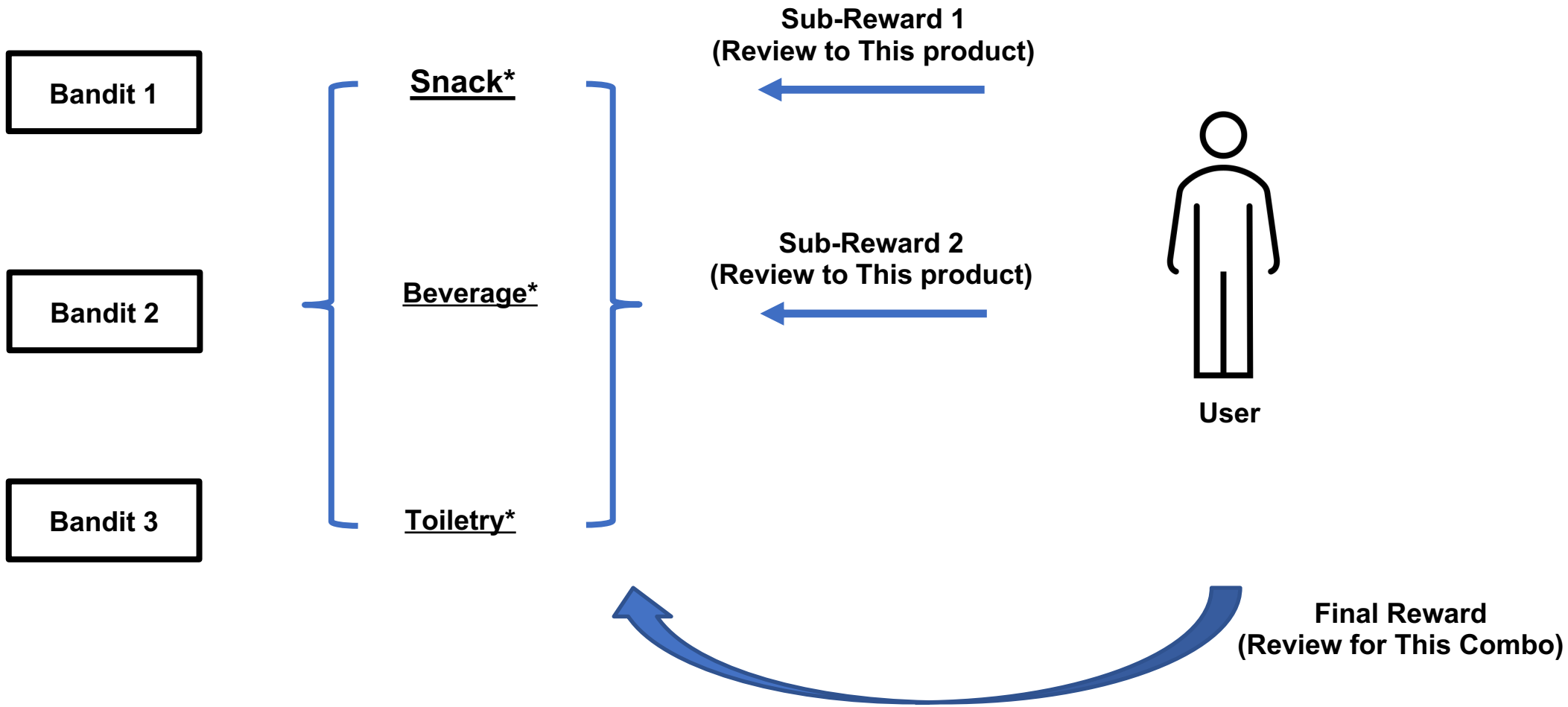
Application in Clinic Trials



Application in Travel Business

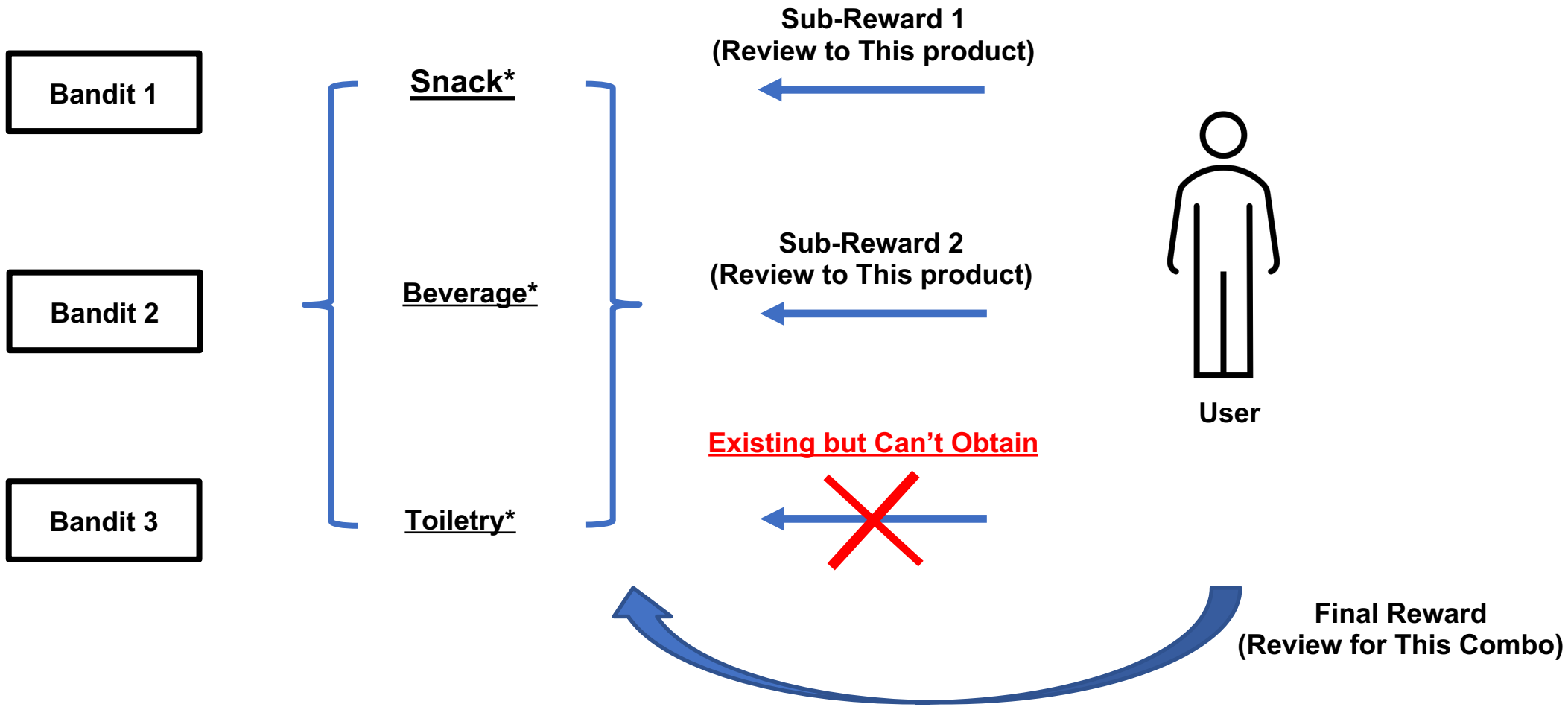


Main Challenge in Multi-facet Contextual Bandit



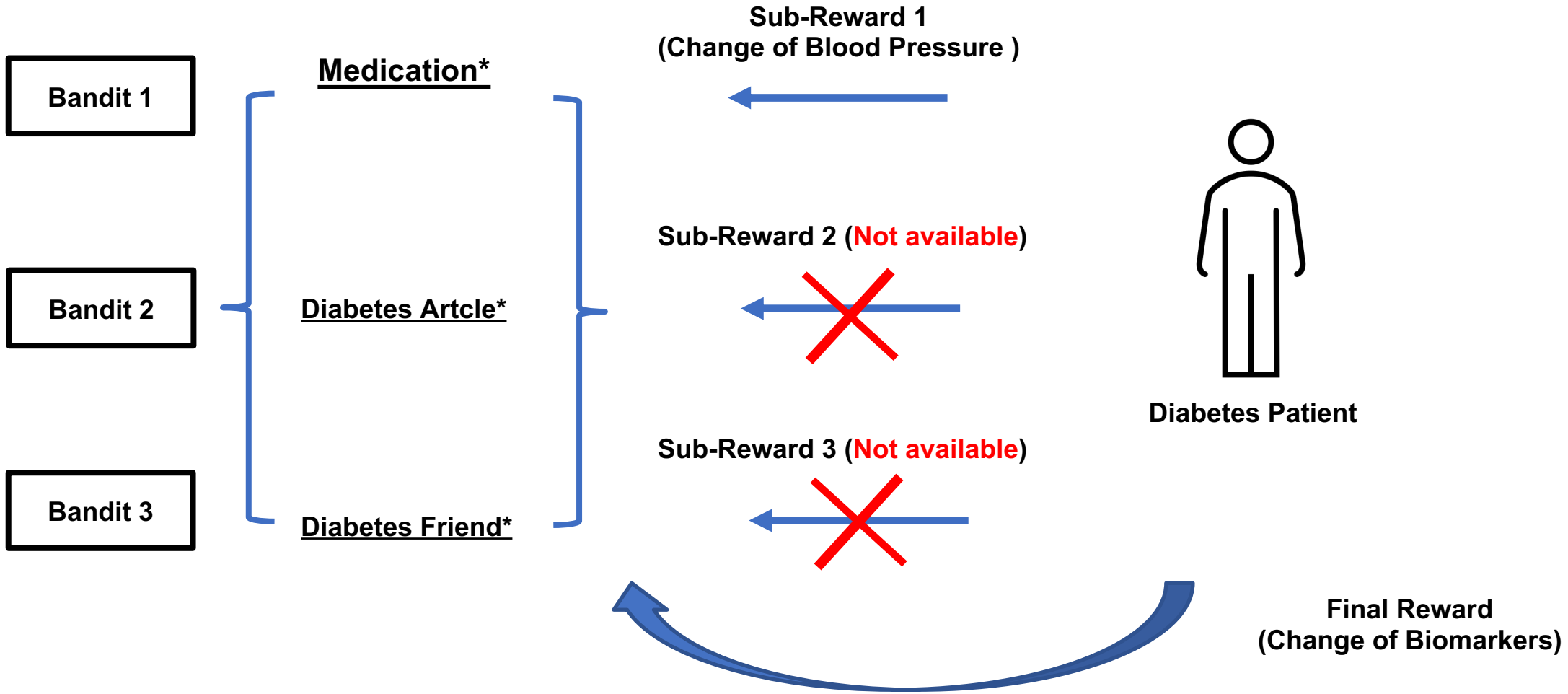
Main Challenge in Multi-facet Contextual Bandit

- **Partial availability of sub-rewards (E-Commerce).**



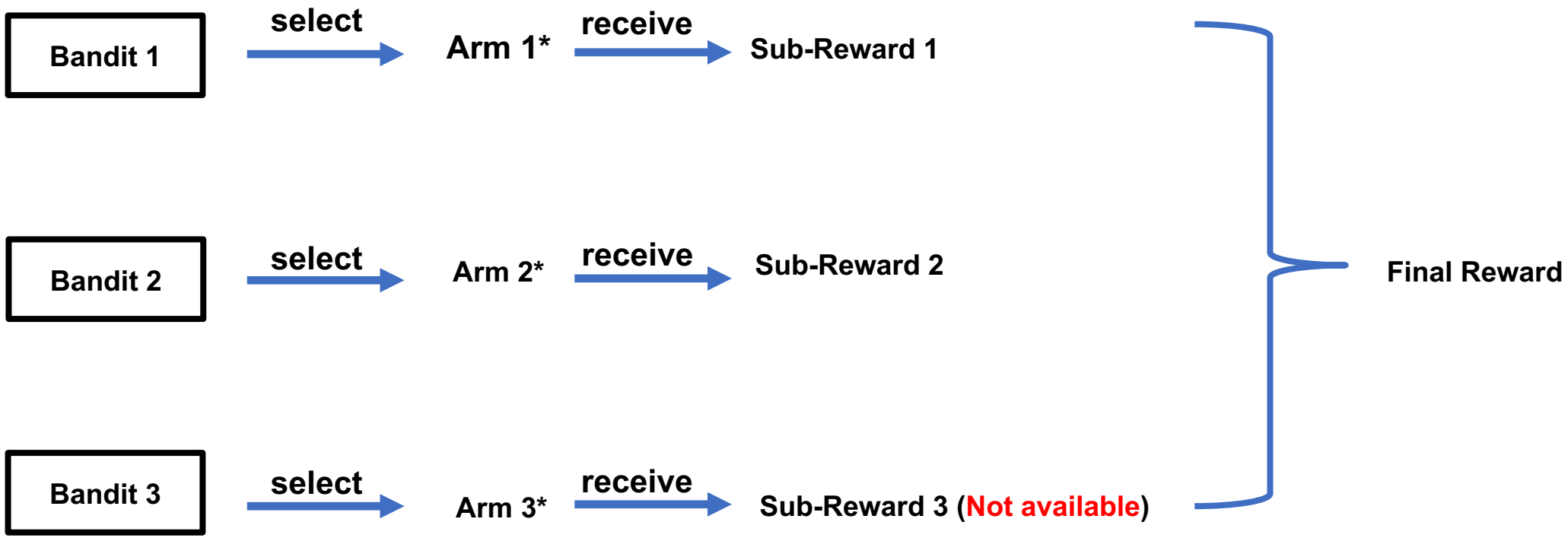
Main Challenge in Multi-facet Contextual Bandit

- Partial availability of sub-rewards (Clinic Trials).



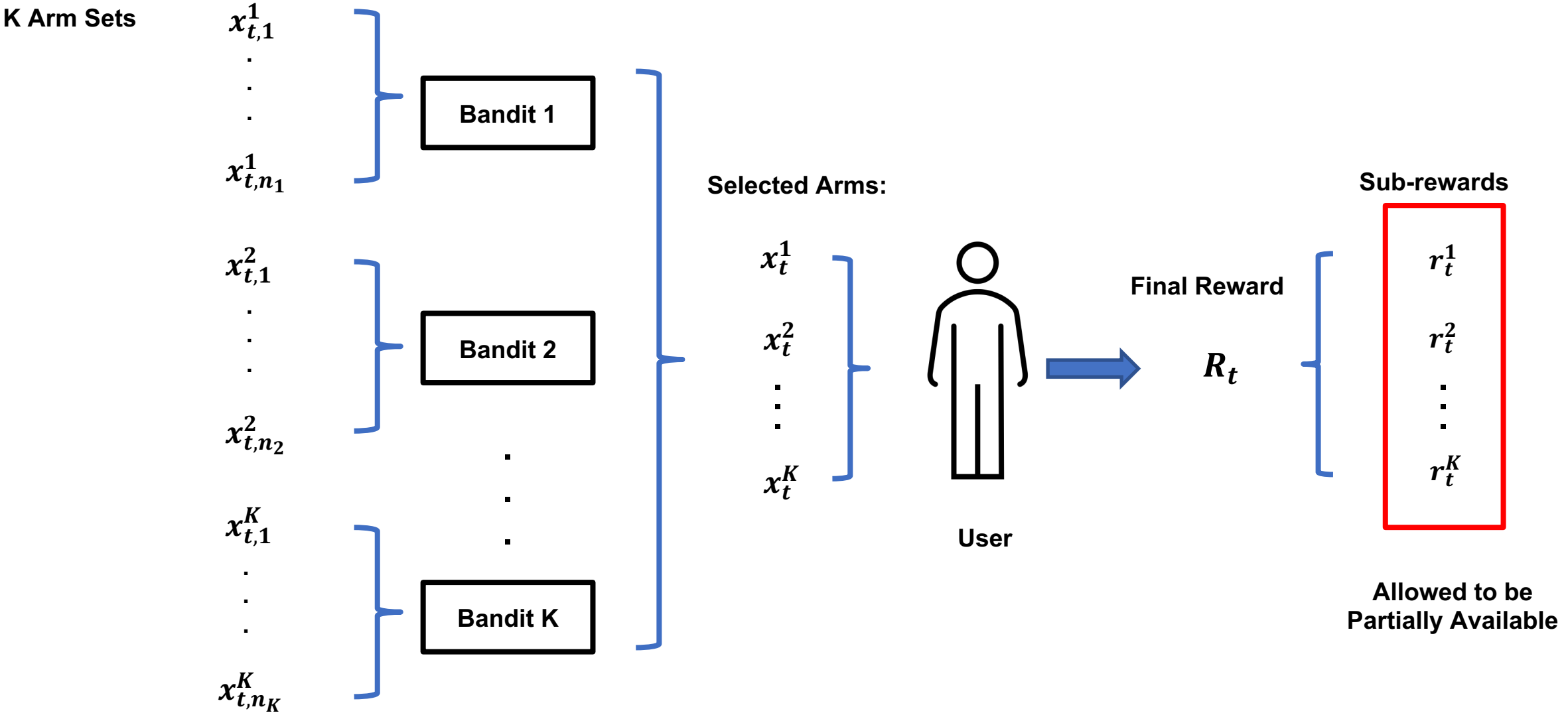
Main Challenge in Multi-facet Contextual Bandit

How to utilize partial availability of sub-rewards?



Sub-rewards determine final reward

Formal Definition of Multi-facet Contextual Bandit : In Round *t*



Formal Definition of Multi-facet Contextual Bandit : In Round *t*

- **Sub-reward Functions (unknown):**

$$r_t^1 = h_1(x_t^1) \quad \text{(Linear or Non-linear)}$$

$$r_t^2 = h_2(x_t^2)$$

$$\vdots$$

$$r_t^K = h_K(x_t^K)$$

Assumtion1: $h_k(0) = 0, \ \forall\, k$

Formal Definition of Multi-facet Contextual Bandit : In Round *t*

- **Sub-reward Functions (unknown):**

$$r_t^1 = h_1(x_t^1) \quad \text{(Linear or Non-linear)}$$

$$r_t^2 = h_2(x_t^2)$$

$$\vdots$$

$$r_t^K = h_K(x_t^K)$$

Assumtion1: $h_k(0) = 0, \ \forall\, k$

- **Final Reward Function (unknown):**

$$R_t = H\left(r_t^1, r_t^2, \dots, r_t^K\right) + \boxed{\epsilon_t}$$ ← Noise

Expectation: $H(X_t) = E[R_t|X_t] = H\left(r_t^1, r_t^2, \dots, r_t^K\right)$

Assumtion2: *H* **is** $\bar{C}$***- Lipschitz continuous.***

Formal Definition of Multi-facet Contextual Bandit : In Round *t*

- **Sub-reward Functions (unknown):**

$$r_t^1 = h_1(x_t^1) \quad \text{(Linear or Non-linear)}$$

$$r_t^2 = h_2(x_t^2)$$

$$\vdots$$

$$r_t^K = h_K(x_t^K)$$

Assumtion1: $h_k(0) = 0, \ \forall\, k$

- **Final Reward Function (unknown):**

$$R_t = H\left(r_t^1, r_t^2, \dots, r_t^K\right) + \epsilon_t$$ ← Noise

Expectation: $H(X_t) = E[R_t|X_t] = H\left(r_t^1, r_t^2, \dots, r_t^K\right)$

Assumtion2: *H* **is** $\bar{C}$***- Lipschitz continuous.***

- **Evaluation Measure: Regret**

$$Reg = E\left[\sum_t (R_t^* - R_t)\right]$$

$$= \sum_t [H(X_t^*) - H(X_t)]$$

Optimal Final Reward **Received Final Reward**

- **Goal: Minimize the regret of T rounds.**

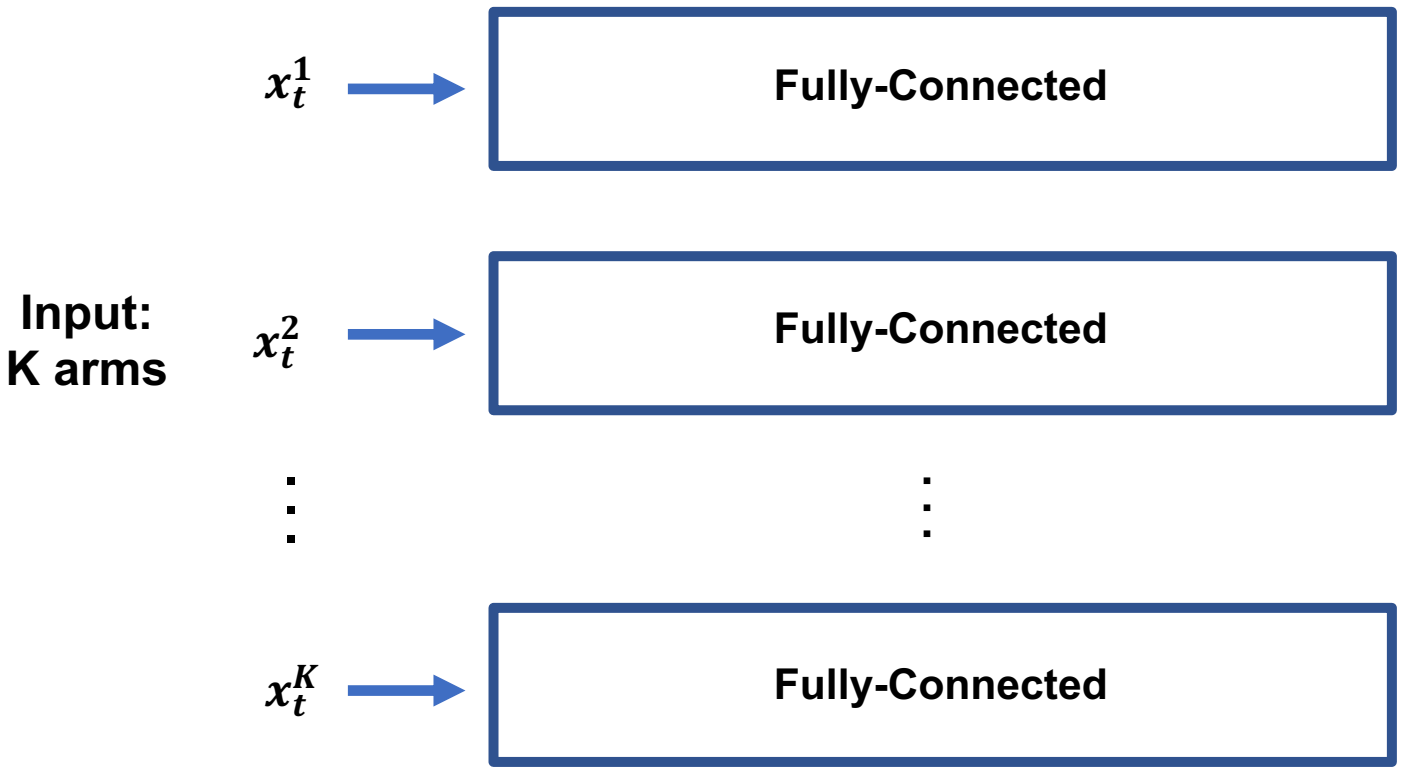
Roadmap

- Background
- Motivation and Problem Definition
- Proposed Algorithm
- Theoretical Analysis
- Experiments and Conclusion

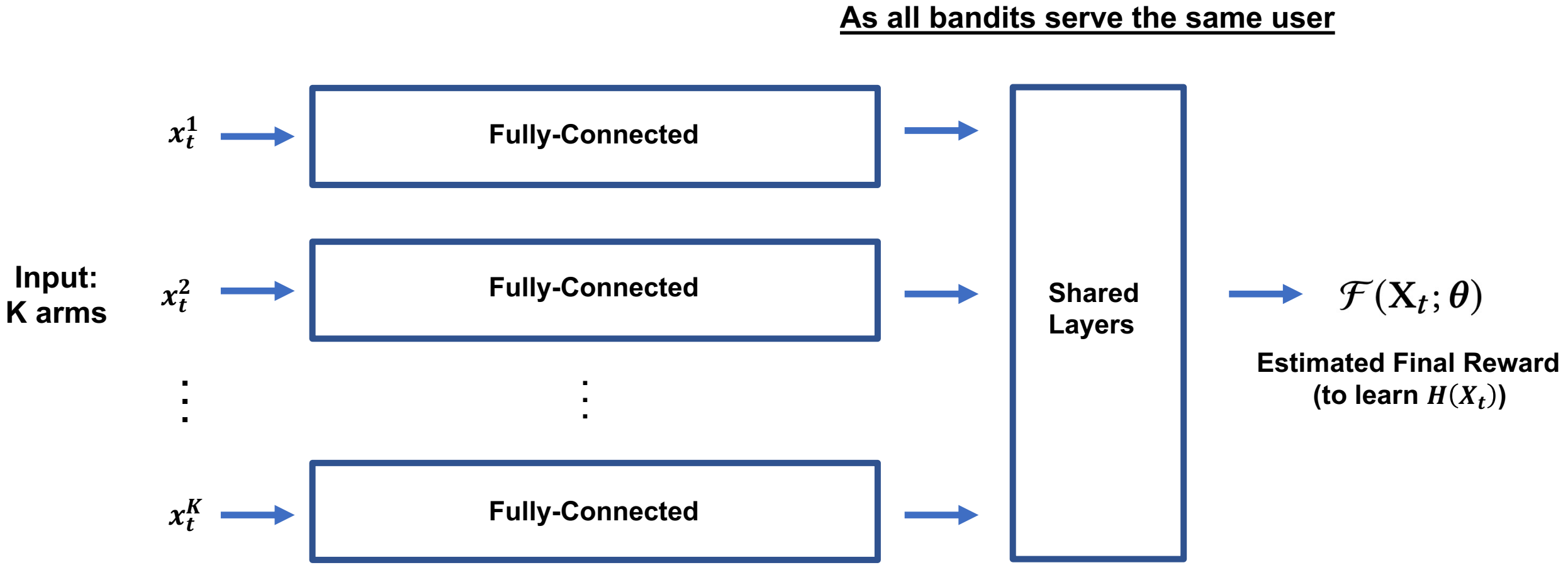
Proposed Algorithm: MuFasa

- **Exploitation**
  - **Neural Network Model**
  - **Training Process**
- **Exploration**
  - **Upper Confidence Bound**

MuFasa: Exploitation (Neural Network Model)



MuFasa: Exploitation (Neural Network Model)



MuFasa: Exploitation (Training Process)

- **Minimization Problem:** **In round t:**
  1) **Observe K arm sets;**
  2) **Select and play K arms;**
  3) **Observe rewards;**
  4) **Gradient descent and update parameters.**

- **Loss function:**

Estimated Final Reward

Regularization

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \left(\mathcal{F}(\mathbf{X}_t; \boldsymbol{\theta}) - R_t\right)^2 /2 + m_2 \lambda \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2 /2.$$

Received Final reward

MuFasa: Exploitation (Training Process)

- **Partially Available Sub-rewards!** $\{r_t^1, \ldots, r_t^K\}$
  - Consider a pair: $\mathbf{x}_t^k, \quad r_t^k$

Selected Arm for bandit k

Sub-reward of bandit k

MuFasa: Exploitation (Training Process)

- **Partially Available Sub-rewards!** $\{r_t^1, \ldots, r_t^K\}$
  - **Consider a pair:** $\mathbf{x}_t^k, \quad r_t^k$

    Selected Arm for bandit k — **Sub-reward of bandit k**
  - **Construct a new pair:** $\widetilde{\mathbf{x}}_{t,k} = \{0, \ldots, \mathbf{x}_t^k, \ldots, 0\}$, $\widetilde{\mathbf{r}}_{t,k} = \{0, \ldots, r_t^k, \ldots, 0\}.$

    Unfortunately, $H(\mathrm{vec}(\widetilde{\mathbf{r}}_{t,k}))$ **is unknown.**

MuFasa: Exploitation (Training Process)

- **Partially Available Sub-rewards!** $\{r_t^1, \ldots, r_t^K\}$
  - **Consider a pair:** $\mathbf{x}_t^k, \quad r_t^k$

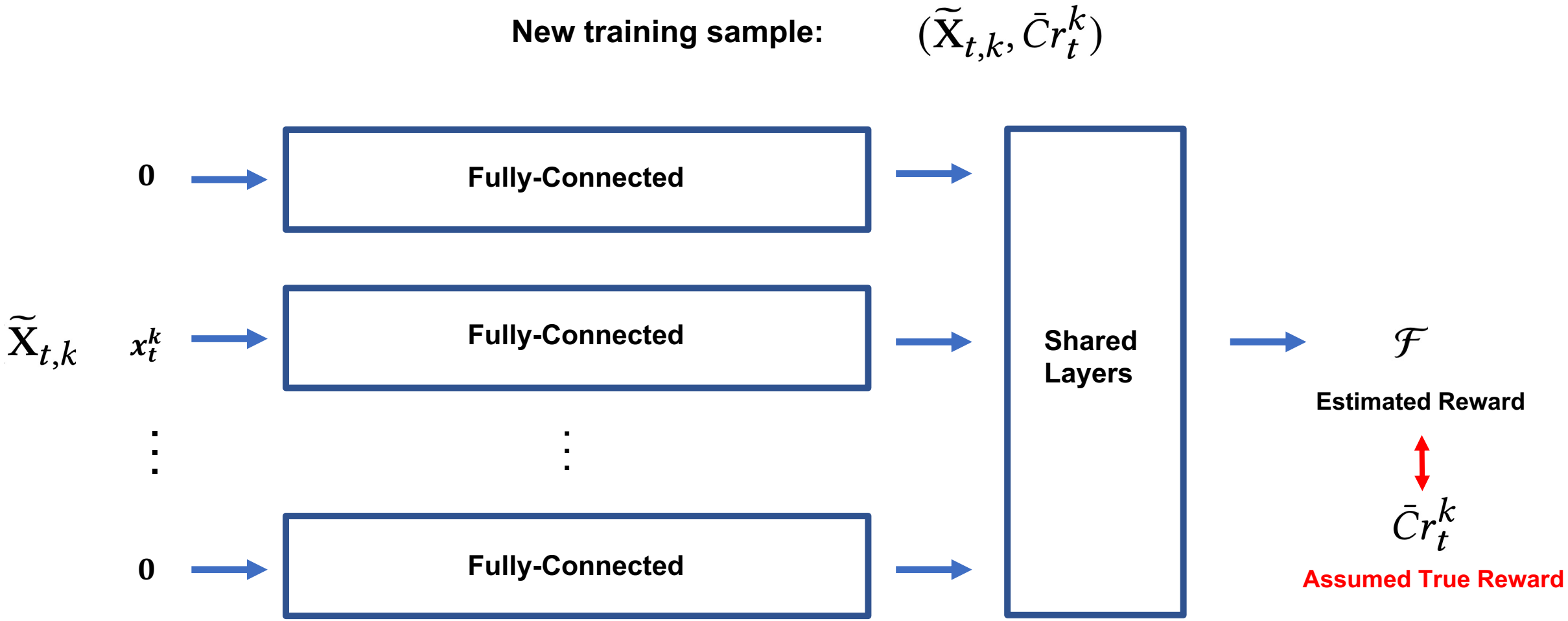
    Selected Arm for bandit k — **Sub-reward of bandit k**
  - **Construct a new pair:** $\widetilde{\mathbf{X}}_{t,k} = \{\mathbf{0}, \ldots, \mathbf{x}_t^k, \ldots, \mathbf{0}\}$, $\widetilde{\mathbf{r}}_{t,k} = \{0, \ldots, r_t^k, \ldots, 0\}$.

    Unfortunately, $H(\mathrm{vec}(\widetilde{\mathbf{r}}_{t,k}))$ **is unknown.**
  - **Based on** $\bar{C}$**-** ***Lipschitz continuity,***

$$H(\mathrm{vec}(\widetilde{\mathbf{r}}_{t,k})) \leq \bar{C} r_t^k$$

New training sample: $(\widetilde{\mathbf{X}}_{t,k}, \bar{C} r_t^k)$

MuFasa: Exploitation (Training Process)

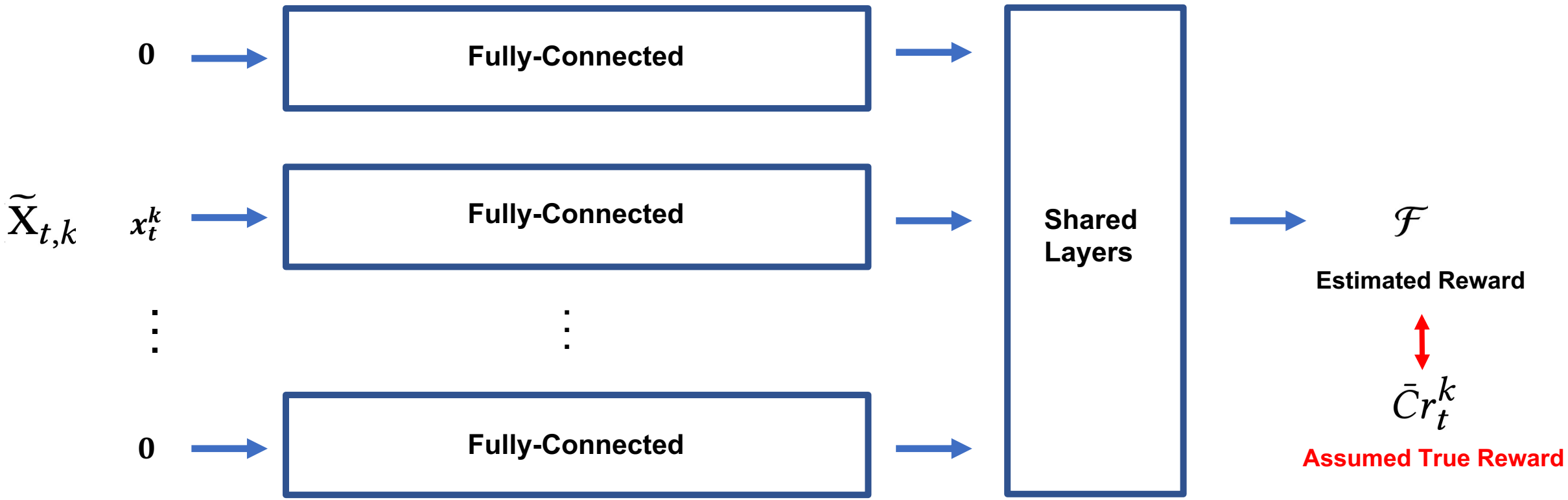


MuFasa: Exploitation (Training Process)

Suppose we have Q $(Q \leq K)$ available sub-rewards,

we have $\underline{Q+1}$ training samples in a round totally

$$\Omega_t = \{(\widetilde{\mathbf{X}}_{t,k}, \bar{C}r_t^k)\}_{k\in[\mathcal{K}]} \bigcup \{(\mathbf{X}_t, R_t)\}.$$



MuFasa: Exploration (Upper Confidence Bound)

- **UCB:** $$\mathbb{P}\left(|\mathcal{F}(\mathbf{X}_t;\boldsymbol{\theta}_t) - \mathcal{H}(\mathbf{X}_t)| > \mathrm{UCB}(\mathbf{X}_t)\right) \leq \delta,$$

- **K selected arms are determined by:**

$$\mathbf{X}_t = \arg\max_{\mathbf{X}'_t \in \mathbf{S}_t} \left(\mathcal{F}(\mathbf{X}';\boldsymbol{\theta}_t) + \mathrm{UCB}(\mathbf{X}'_t)\right).$$

Where

$$\mathbf{S}_t = \{(\mathbf{x}_t^1, \ldots, \mathbf{x}_t^k, \ldots, \mathbf{x}_t^K) \mid \mathbf{x}_t^k \in \mathbf{X}_t^k, k \in [K]\},$$

(all possible combinations of K arms)

MuFasa: Workflow

➢ Workflow

In round t:

1) **Observe K arm sets;**
2) **Select and play K arms using UCB strategy;**
3) **Observe final reward and sub-rewards;**
4) **Construct new training samples;**
5) **Gradient descent and update parameters.**

Roadmap

- Background
- Motivation and Problem Definition
- Proposed Algorithm
- Theoretical Analysis
- Experiments and Conclusion

New Upper Confidence Bound 1

- Consider a single bandit where reward function is $h(\mathbf{x}_t)$

  learned by a fully-connected neural network $f(\mathbf{x}_t; \boldsymbol{\theta})$

- With probability at least $1 - \delta$,

$$\left|h(\mathbf{x}_t) - f(\mathbf{x}_t; \boldsymbol{\theta}_t)\right| \leq \gamma_1 \|g(\mathbf{x}_t; \boldsymbol{\theta}_t)/\sqrt{m}\|_{\mathbf{A}_t^{-1}} + \gamma_2 \|g(\mathbf{x}_t; \boldsymbol{\theta}_0)/\sqrt{m}\|_{\mathbf{A}_t'^{-1}} + \gamma_1\gamma_3 + \gamma_4,$$

$$\gamma_1(m, L) = (\lambda + tO(L)) \cdot ((1 - \eta m\lambda)^{J/2}\sqrt{t/\lambda}) + 1$$

$$\gamma_2(m, L, \delta) = \sqrt{\log\left(\frac{\det(\mathbf{A}_t')}{\det(\lambda\mathbf{I})}\right) - 2\log\delta} + \lambda^{1/2}S$$

$$\gamma_3(m, L) = C_2 m^{-1/6}\sqrt{\log mt}^{1/6}\lambda^{-7/6}L^{7/2}$$

$$\gamma_4(m, L) = C_1 m^{-1/6}\sqrt{\log mt}^{2/3}\lambda^{-2/3}L^3$$

$$\mathbf{A}_t = \lambda\mathbf{I} + \sum_{i=1}^{t} g(\mathbf{x}_t; \boldsymbol{\theta}_t)g(\mathbf{x}_t; \boldsymbol{\theta}_t)^\top/m$$

$$\mathbf{A}_t' = \lambda\mathbf{I} + \sum_{i=1}^{t} g(\mathbf{x}_t; \boldsymbol{\theta}_0)g(\mathbf{x}_t; \boldsymbol{\theta}_0)^\top/m.$$

New Upper Confidence Bound 1

- Consider a single bandit where reward function is $h(\mathbf{x}_t)$

  learned by a fully-connected neural network $f(\mathbf{x}_t; \boldsymbol{\theta})$

- With probability at least $1 - \delta$,

$$\left|h(\mathbf{x}_t) - f(\mathbf{x}_t; \boldsymbol{\theta}_t)\right| \leq \gamma_1 \|g(\mathbf{x}_t; \boldsymbol{\theta}_t)/\sqrt{m}\|_{\mathbf{A}_t^{-1}} + \gamma_2 \|g(\mathbf{x}_t; \boldsymbol{\theta}_0)/\sqrt{m}\|_{\mathbf{A}_t'^{-1}} + \gamma_1\gamma_3 + \gamma_4,$$

Gradient after J iterations of gradient descent (higher variance)

Gradient at initialization (higher bias)

$$\gamma_1(m, L) = (\lambda + tO(L)) \cdot ((1 - \eta m\lambda)^{J/2}\sqrt{t/\lambda}) + 1$$

$$\gamma_2(m, L, \delta) = \sqrt{\log\left(\frac{\det(\mathbf{A}_t')}{\det(\lambda\mathbf{I})}\right) - 2\log\delta + \lambda^{1/2}S}$$

$$\gamma_3(m, L) = C_2 m^{-1/6}\sqrt{\log mt}^{1/6}\lambda^{-7/6}L^{7/2}$$

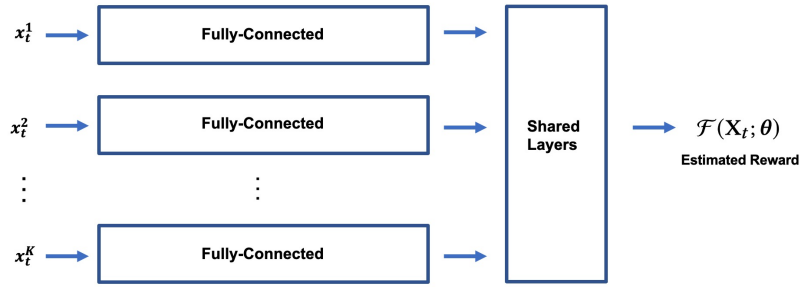
$$\gamma_4(m, L) = C_1 m^{-1/6}\sqrt{\log mt}^{2/3}\lambda^{-2/3}L^3$$

$$\mathbf{A}_t = \lambda\mathbf{I} + \sum_{i=1}^{t} g(\mathbf{x}_t; \boldsymbol{\theta}_t)g(\mathbf{x}_t; \boldsymbol{\theta}_t)^\top/m$$

$$\mathbf{A}_t' = \lambda\mathbf{I} + \sum_{i=1}^{t} g(\mathbf{x}_t; \boldsymbol{\theta}_0)g(\mathbf{x}_t; \boldsymbol{\theta}_0)^\top/m.$$

New Upper Confidence Bound 2

- **Consider the using assembled neural network in MuFasa**



- **With probability at least $1-\delta$,**

$$|\mathcal{F}(\mathbf{X}_t;\boldsymbol{\theta}_t) - \mathcal{H}(\mathbf{X}_t)| \leq \bar{C}\sum_{k=1}^{K}\mathcal{B}^k + \mathcal{B}^F = UCB(\mathbf{X}_t),\ \textit{where}$$

$$\mathcal{B}^k = \gamma_1 \|g_k(\mathbf{x}_t^k;\boldsymbol{\theta}_t^k)/\sqrt{m_1}\|_{\mathbf{A}_t^{k-1}} + \gamma_2\left(\frac{\delta}{k+1}\right)\|g_k(\mathbf{x}_t^k;\boldsymbol{\theta}_0^k)/\sqrt{m_1}\|_{\mathbf{A}_t^{k\prime-1}} + \gamma_1\gamma_3 + \gamma_4$$

$$\mathcal{B}^F = \gamma_1 \|G(\mathbf{f}_t;\boldsymbol{\theta}_t^\Sigma)/\sqrt{m_2}\|_{\mathbf{A}_t^{F-1}} + \gamma_2\left(\frac{\delta}{k+1}\right)\|G(\mathbf{f}_t;\boldsymbol{\theta}_0^\Sigma)/\sqrt{m_2}\|_{\mathbf{A}_t^{F\prime-1}} + \gamma_1\gamma_3 + \gamma_4$$

$$\mathbf{A}_t^k = \lambda\mathbf{I} + \sum_{i=1}^{t} g_k(\mathbf{x}_i^k;\boldsymbol{\theta}_t^k)g_k(\mathbf{x}_i^k;\boldsymbol{\theta}_t^k)^\intercal/m_1$$

$$\mathbf{A}_t^{k\prime} = \lambda\mathbf{I} + \sum_{i=1}^{t} g_k(\mathbf{x}_i^k;\boldsymbol{\theta}_0^k)g_k(\mathbf{x}_i^k;\boldsymbol{\theta}_0^k)^\intercal/m_1$$

$$\mathbf{A}_t^F = \lambda\mathbf{I} + \sum_{i=1}^{t} G(\mathbf{f}_i;\boldsymbol{\theta}_t^\Sigma)G(\mathbf{f}_i;\boldsymbol{\theta}_t^\Sigma)^\intercal/m_2$$

$$\mathbf{A}_t^{F\prime} = \lambda\mathbf{I} + \sum_{i=1}^{t} G(\mathbf{f}_i;\boldsymbol{\theta}_0^\Sigma)G(\mathbf{f}_i;\boldsymbol{\theta}_0^\Sigma)^\intercal/m_2$$

Regret Analysis

$$\boldsymbol{Reg} = \boldsymbol{E}\left[\sum_{t}(\boldsymbol{R}_t^* - \boldsymbol{R}_t)\right]$$
$$= \sum_{t}[\boldsymbol{H}(\boldsymbol{X}_t^*) - \boldsymbol{H}(\boldsymbol{X}_t)]$$

- After T rounds, with probability at least $1-\delta$,

$$\boldsymbol{Reg} \leq (\bar{C}K+1)\sqrt{T}2\sqrt{\widetilde{P}\log(1+T/\lambda)+1/\lambda+1}$$
$$\cdot\left(\sqrt{(\widetilde{P}-2)\log\left(\frac{(\lambda+T)(1+K)}{\lambda\delta}\right)+1/\lambda+\lambda^{1/2}S+2}\right)+2(\bar{C}K+1),$$

- Achieve near-optimal regret bound $\widetilde{O}\left((K+1)\sqrt{T}\right)$, same as a single linear bandit $\widetilde{O}(\sqrt{T})$

Roadmap

- Background
- Motivation and Problem Definition
- Proposed Algorithm
- Theoretical Analysis
- Experiments and Conclusion

Experiments Setting

- **Data sets: Yelp (Personalized recommendation)**
- **Two bandits: Restaurants and Friends**

- **Data sets: Mnist + NotMnist (Classification)**
- **Two bandits: Number and Letter**

- **Ground-truth reward functions:** $H_1(\mathrm{vec}(\mathbf{r}_t)) = r_t^1 + r_t^2; \;\; H_2(\mathrm{vec}(\mathbf{r}_t)) = 2r_t^1 + r_t^2.$

- **Baselines: (1) (K-) LinUCB [1]; (2) (K-) KerUCB [2]; (3) (K-) NeuUCB [3]**
- **Run them on 2 bandits respectively**

[1] Li et all. 2010. A contextual- bandit approach to personalized news article recommendation. WWW'21
[2] Valko et all. 2013. Finite-time analysis of kernelised contextual bandits. arXiv preprint arXiv:1309.6869 (2013).
[3] Zhou et all. 2020. Neural contextual bandits with UCB-based exploration. ICML'20.

All Sub-rewards Available (Same Final Reward Function)

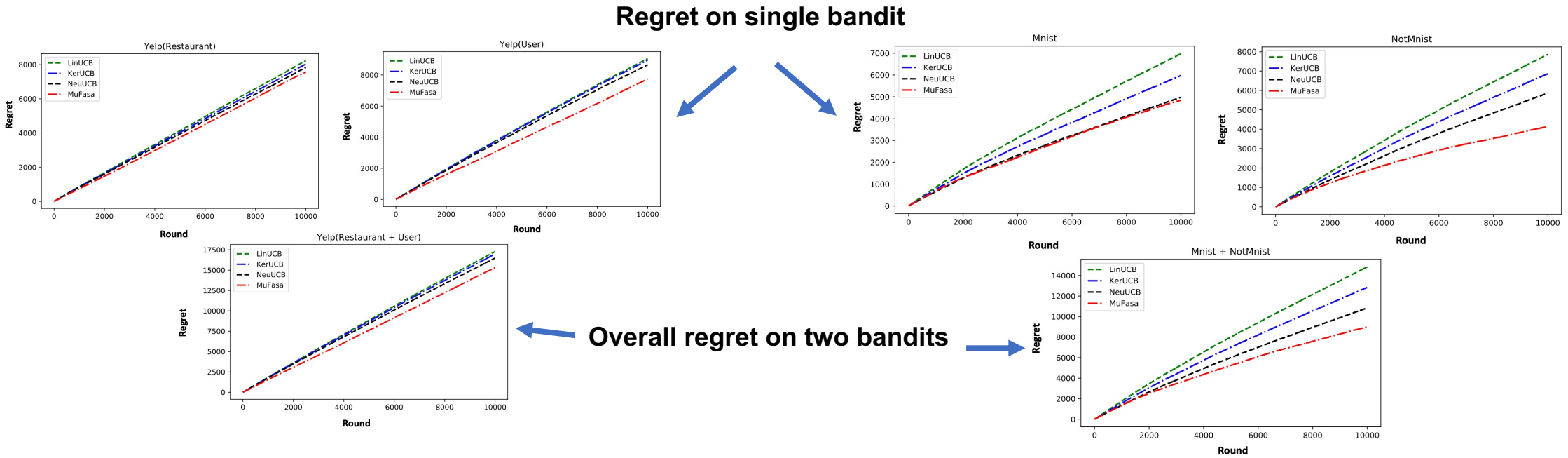


Figure 1: Regret comparison on Yelp with H_1 final reward function.

Figure 2: Regret comparison on Mnist+NotMnist with H_1 final reward function.

Observation:

- **With same final reward function,**

  MuFasa outperforms all baselines.

Insights:

- **Leveraging mutual influence among bandits can improve the overall performance**

All Sub-rewards Available (Different Final Reward Function)

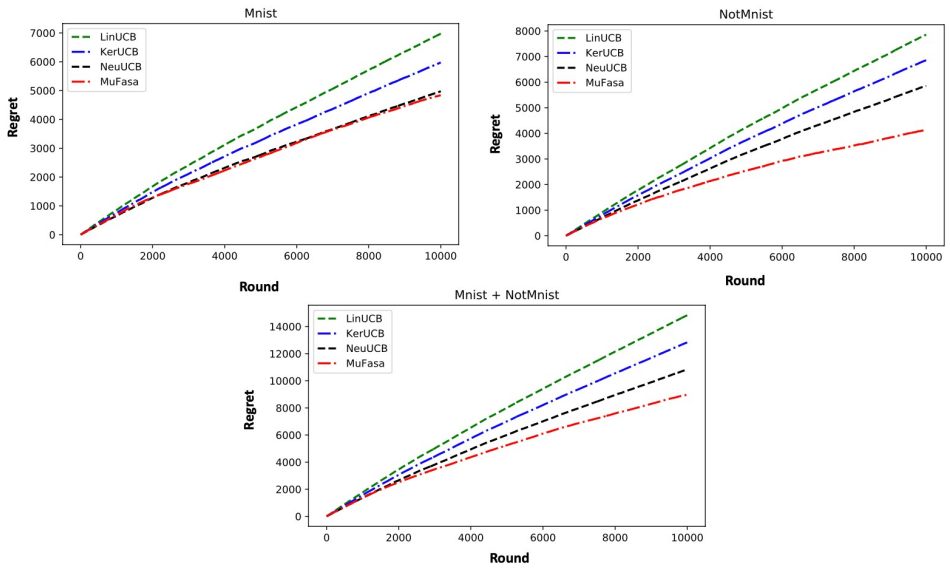


Figure 3: Regret comparison on Mnist+NotMnist with H_1.

$$H_1(\mathrm{vec}(\mathbf{r}_t)) = r_t^1 + r_t^2$$

Observation:

- **Superiority of MuFasa is slightly higher on H_2, compared to H_1.**

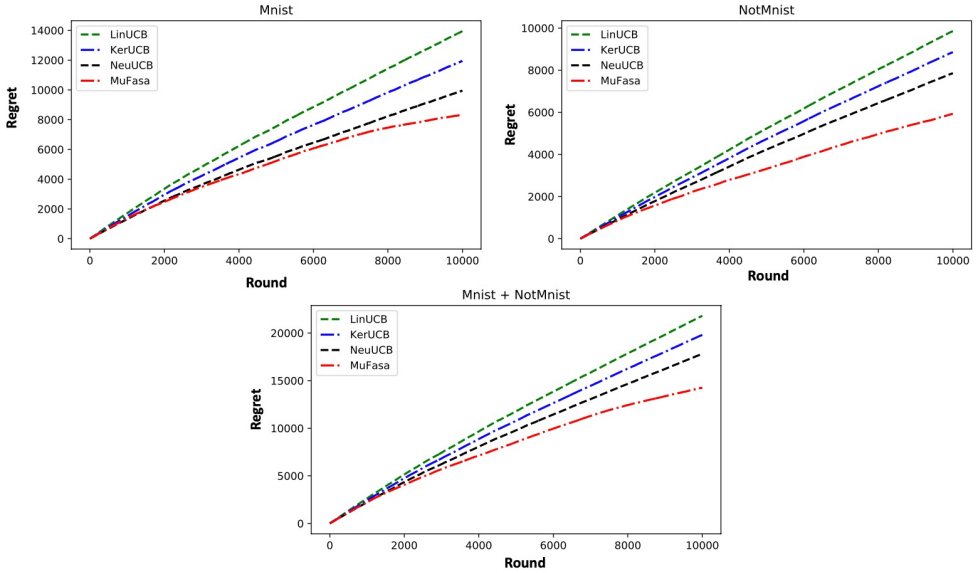


Figure 4: Regret comparison on Mnist+NotMnist with H_2.

$$H_2(\mathrm{vec}(\mathbf{r}_t)) = 2r_t^1 + r_t^2.$$

Insights:

- **MuFasa can select arms according to different weights of bandits (Bandit 1 has higher weight in H_2).**

Partial Sub-rewards Available

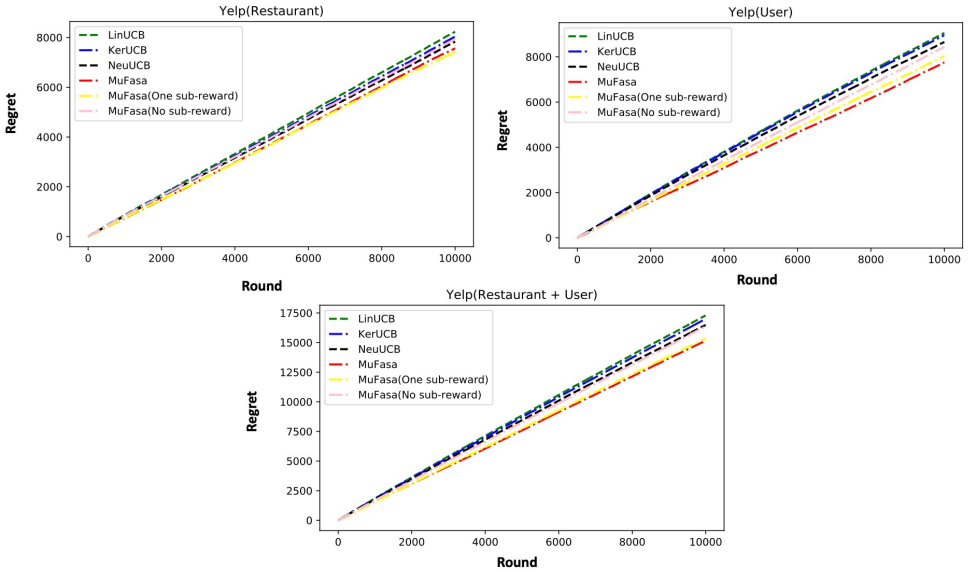


Figure 5: Regret comparison on Yelp with different reward availability.

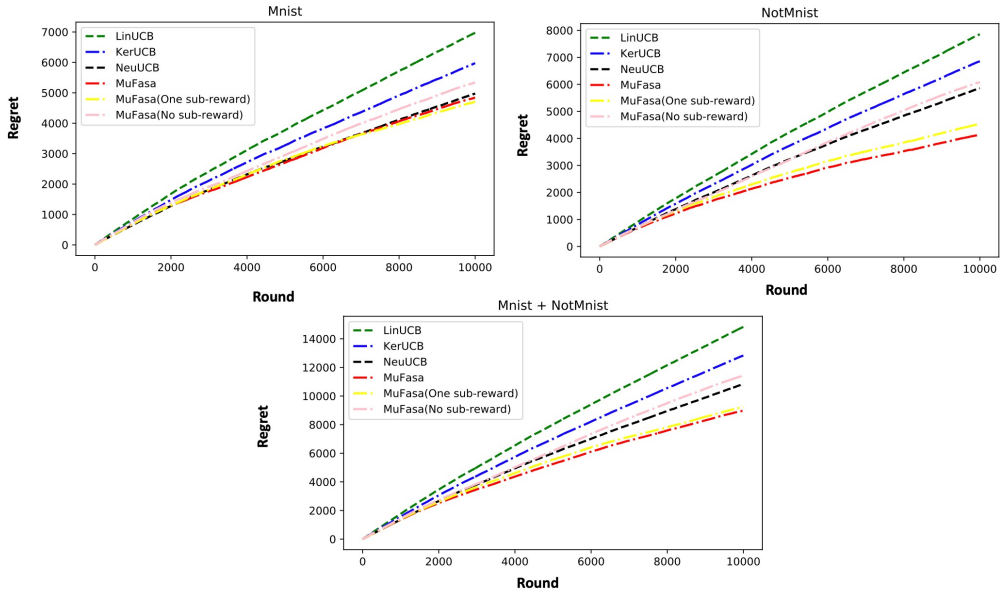


Figure 6: Regret comparison on Mnist+NotMnist with different reward availability.

Observation:

- **With one sub-reward, MuFasa still outperforms all baselines.**
- **Without any sub-reward, MuFasa's performance is close to the best baseline.**

Conclusion:

- ✓ [Problem]. We introduce a novel bandit problem, Multi-facet Contextual Bandit, which has direct applications in personalized recommendation, online advertising, travel business, etc.
- ✓ [Algorithm]. We propose a novel algorithm, MuFasa, which exploits the final reward and (partial) sub-rewards by an assembled neural network with UCB-based exploration.
- ✓ [Theory]. We provide a new Upper Confidence Bound and achieve near-optimal regret bound for MuFasa.
- ✓ [Experiments]. We conduct extensive experiments to show: (1) MuFasa outperforms baselines; (2) Mutual influence, weighting, and sub-rewards of bandits play important roles in multi-facet contextual bandit problem.

THANKS

Yikun Ban
UIUC
yikunb2@illinois.edu
www.banyikun.com

Jingrui He
UIUC
jingrui@Illinois.edu
www.hejingrui.org

Curtiss B. Cook
Mayo Clinic Arizona
cook.curtiss@mayo.edu

ILLINOIS