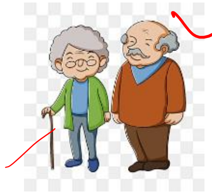
Contextual Bandits

Subrahmanya Swamy Peruru

Photos provided by Pexels

Contextual Bandits – Multiple states

- News article Recommendation systems



Retirement Benifit

90%

10%

90%

- Articles – arms
- +1 −1
  Like / Dislike – Reward
- User – State

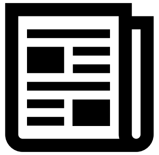
Different users have different preferences for articles

Photos provided by Pexels

Multi-arm Bandits – One state

Arms / Articles

a



Expected

$\mu(a) = 0.2$

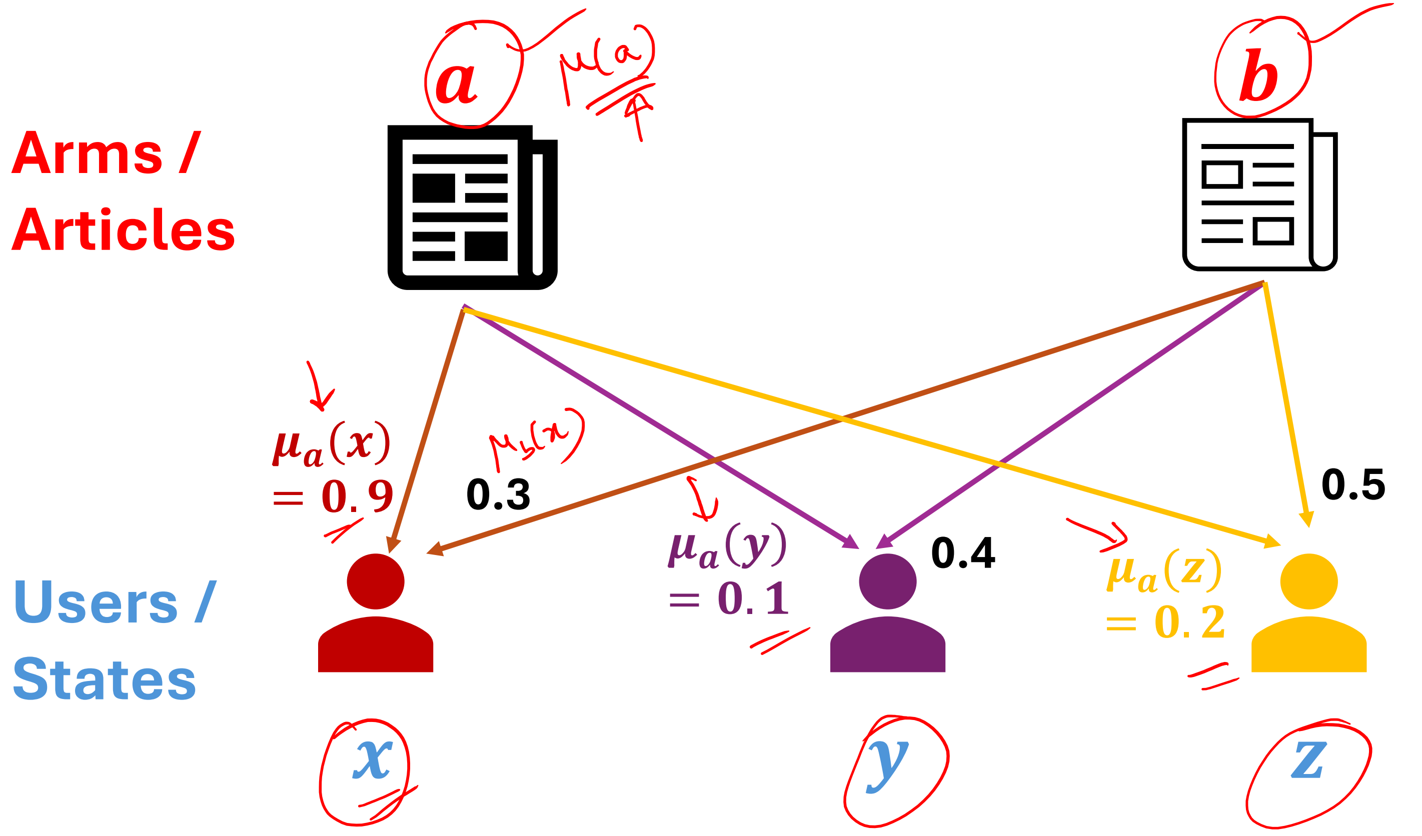
b

$\mu(b) = 0.9$

- Each arm has only one expected reward associated with it

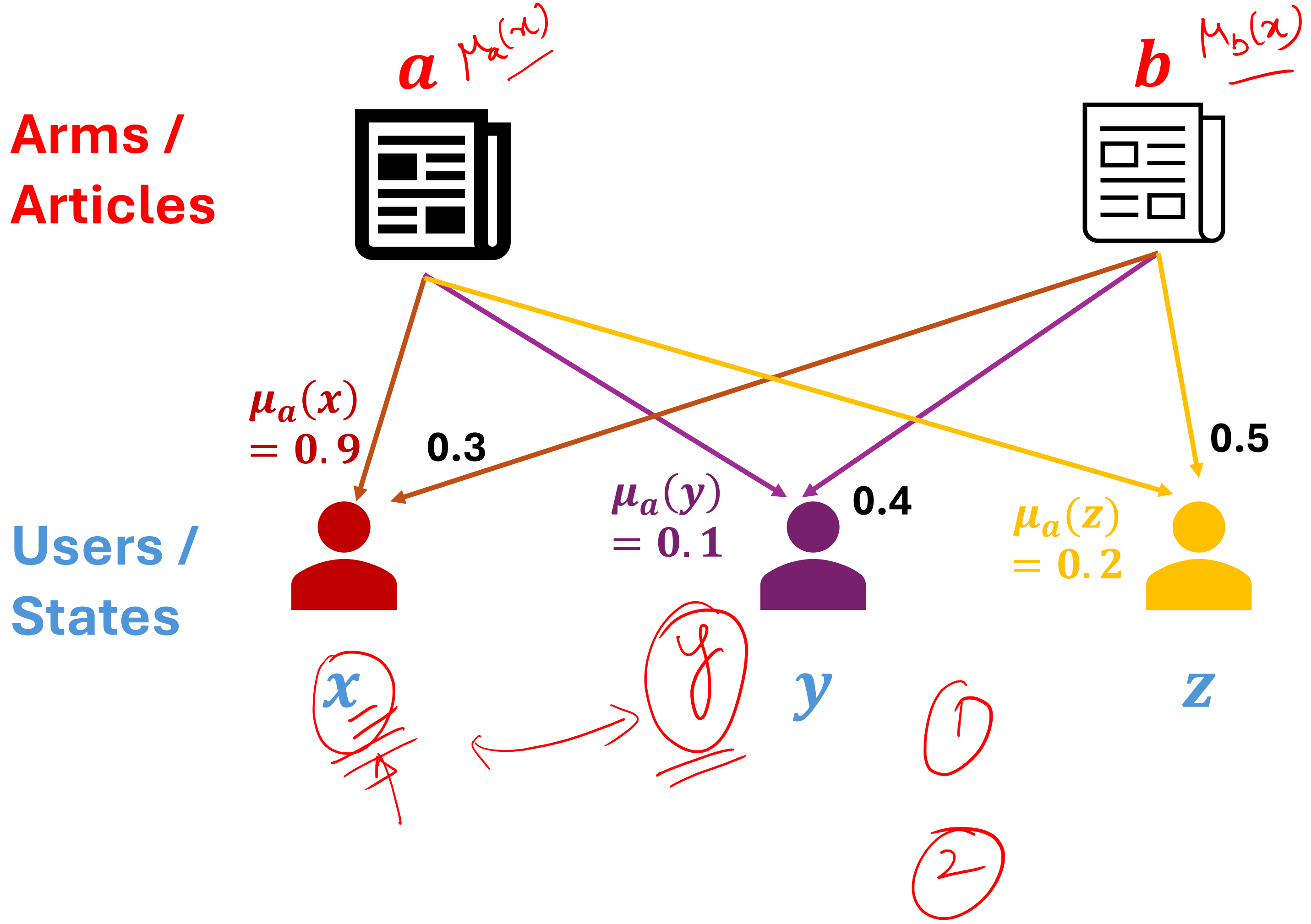
Contextual Bandits – Multiple States

ETC
ε-greedy
UCB



- Expected reward of an arm changes with user $\mu_a(x)$
- **How to deal with it?**

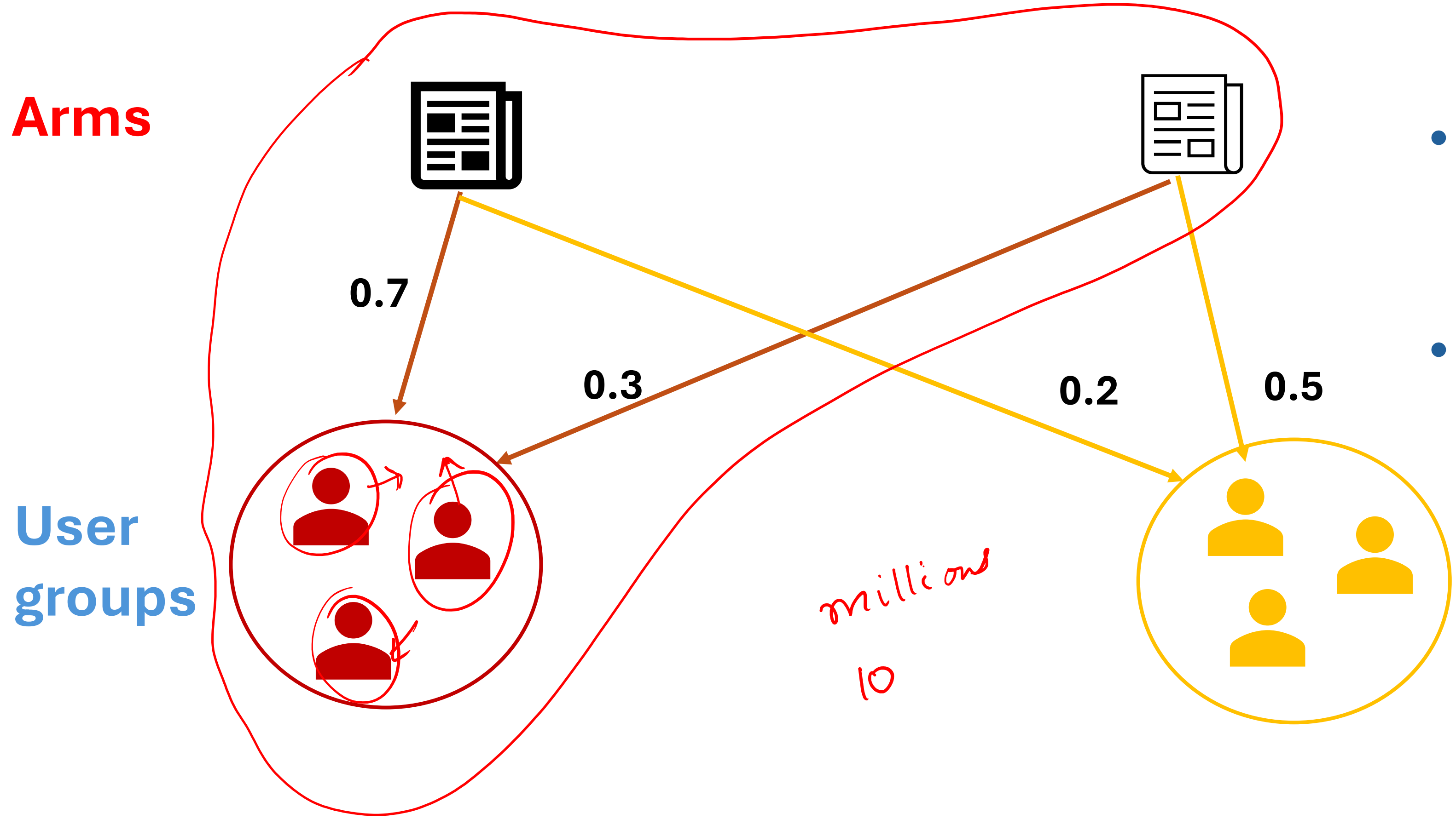
Contextual Bandits – Multiple States



unsup
sup

- Expected reward of an arm changes with user $\mu_a(x)$
- Treat each user as a separate bandit problem
- Practically infeasible with millions of users!

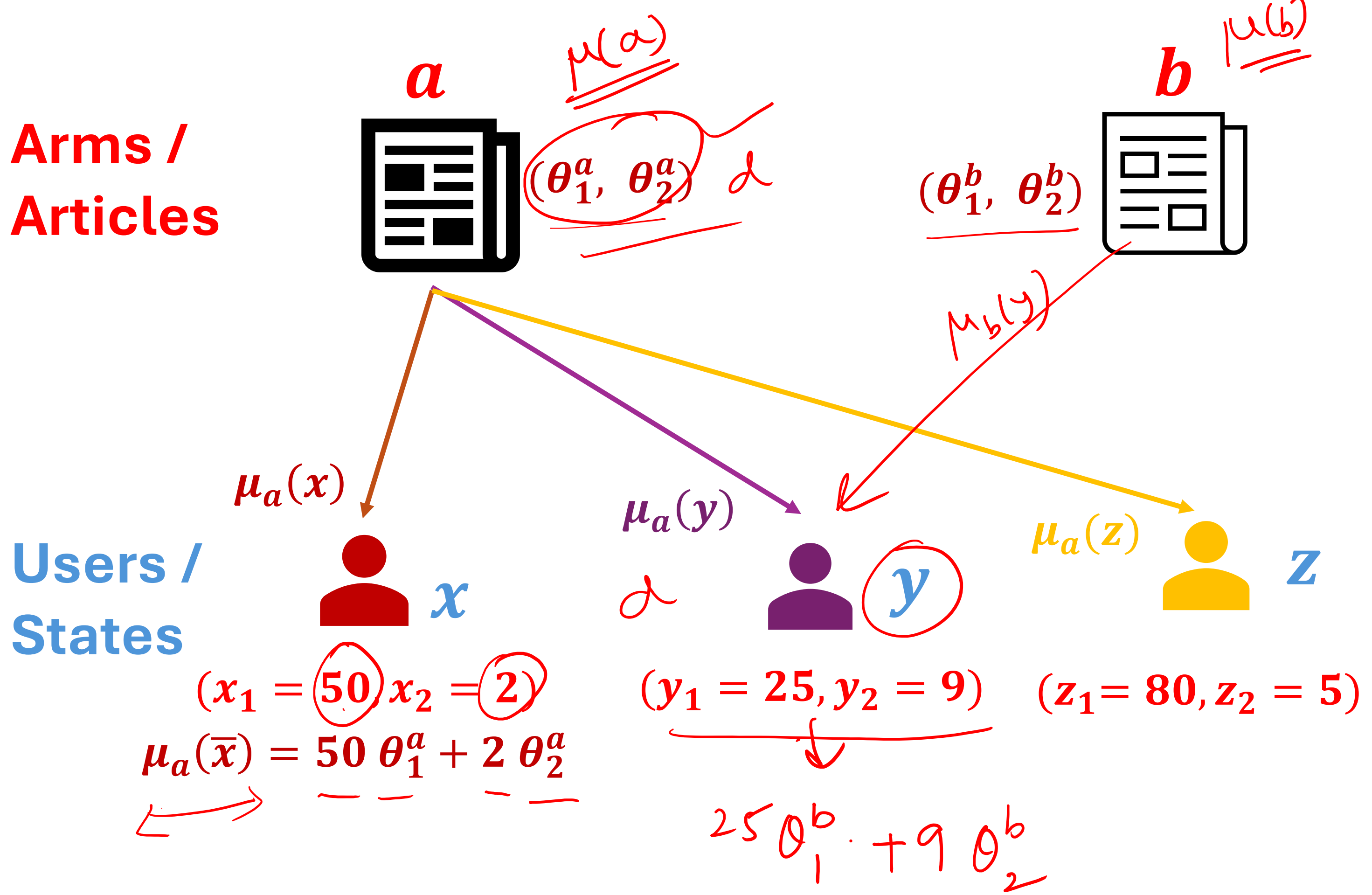
Bandits + Unsupervised Learning



- Form user groups by clustering similar users together
- Solve a separate bandit problem for each cluster

Bandits + Supervised Learning

ETC



- User (state) represented by features such as age, income
  $\bar{x} = (x_1, x_2)$
- Model the expected reward for user $\bar{x}$ for pulling arm a as
  $\mu_a(\bar{x}) = \theta_1^a x_1 + \theta_2^a x_2$

Contextual (Linear) Bandits

$\mu(a) = 2$

$R_1 = 2.1$, 2.7, 1.8 } $\frac{2.1 + 2.7 + 1.8}{3}$

- Users (state) represented by features such as age, gender
  - State feature vector: $\bar{x} = (x_1, x_2)^T$ Eg: (50, 1), (25, 0), (80, 1)
- Each article (arm) has a different expected reward associated with each user (state)
  - The expected reward of an arm is characterized by **unknown** $\bar{\theta^a} = (\theta_1^a, \theta_2^a)^T$
  - State-specific expected reward: $\mu_a(\bar{x}) = \boxed{\theta_1^a x_1 + \theta_2^a x_2}$ (Linear Bandits)
- The reward for playing arm a_t under state $\bar{x_t}$ is $R_t = \mu_{a_t}(\bar{x_t}) + \epsilon_t$, where ϵ_t is independent mean-zero noise.
- If "T arbitrary users" visit our website sequentially, How to maximize the total reward $\sum_{t=1}^{T} R_t$?

Solution Approach

- Explore each arm N times and
- Estimate the mean parameters based on sample rewards

One-state Multi-arm Bandits

$\mu(a) = 2$ Unknown parameter

Sample reward: $R_t = \mu(a) + noise$

Rewards from 3 rounds of exploration

$$R_1 = \mu(a) \approx 2.7$$
$$R_2 = \mu(a) \approx 1.6$$
$$R_3 = \mu(a) \approx 2.1$$

Best estimate: $\hat{\mu}(a) = \arg\min_x (R_1 - x)^2 + (R_2 - x)^2 + (R_3 - x)^2$

Linear Bandits

ETC

$\downarrow a \to 3$ times

Unknown parameter: $(\theta_1^a, \theta_2^a) = \theta_a$

Sample reward: $R_t = \mu_a(x) + noise$

$= \theta_1^a x_1 + \theta_2^a x_2 + noise$

Exploration:

3 users with features: (50, 1), (25, 4), (80, 7)

Observed rewards: 0.7, 0.4, 0.5

$$\left.\begin{aligned} R_1 &= 0.7 \approx \theta_a^1\, 50 + \theta_a^2\, 1 \\ R_2 &= 0.4 \approx \theta_a^1\, 25 + \theta_a^2\, 4 \\ R_3 &= 0.5 \approx \theta_a^1\, 80 + \theta_a^2\, 7 \end{aligned}\right\}$$

$$\underset{D_a}{\begin{bmatrix} 50 & 1 \\ 25 & 4 \\ 80 & 7 \end{bmatrix}} \underset{\theta_a}{\begin{bmatrix} \theta_a^1 \\ \theta_a^2 \end{bmatrix}} = \underset{b_a}{\begin{bmatrix} 0.7 \\ 0.4 \\ 0.5 \end{bmatrix}}$$

Best estimate: $\widehat{\theta^a} = \arg \min_{(\theta_1^a, \theta_2^a)} \left\{ \begin{aligned} &(0.7 - \theta_a^1\, 50 + \theta_a^2\, 1)^2 + \\ &(0.4 - \theta_a^1\, 25 + \theta_a^2\, 4)^2 + \\ &(0.5 - \theta_a^1\, 80 + \theta_a^2\, 7)^2 \end{aligned} \right\}$

$\min \quad f(\theta_1^a, \theta_2^a)$

$\to$ Linear Regression

Photos provided by Pexels

ETC algorithm for Contextual Bandits

$\hat{\mu}(a) = \bar{\mu}(a)$, $\bar{\mu}(b)$ $\quad \frac{N}{K} \quad$ $\max\{\bar{\mu}(a), \bar{\mu}(b)\}$

- Explore each arm "N" times
- Based on the history $\{(x_t, a_t, R_t)\}_{t=1}^{NK}$, estimate the parameters for all $a \in \mathcal{A}$ using Ridge regression.

  ε-greedy

  "t"

  - $\widehat{\theta^a} = \left(D_a^T D_a + I_d\right)^{-1} D_a^T b_a$
  - D_a is $N \times d$ context matrix whose rows represent user feature vectors
  - b_a is $N \times 1$ reward vector with rewards obtained during N exploration rounds

  0.2

  $x_t^1 \hat{\theta}_b^1 + x_t^2 \hat{\theta}_b^2$

  - $D_a = \begin{bmatrix} x_1^1 & x_2^1 \\ x_1^2 & x_2^2 \\ x_1^3 & x_2^3 \end{bmatrix}$ feature dimension $d = 2$, arm a played N = 3 times

  0.9

  $x_t^1 \hat{\theta}_a^1 + x_t^2 \hat{\theta}_a^2$

  NK

- At any round $t > NK$, play the arm $a_t = \arg\max_a \; x_t^T \widehat{\theta^a}$

$a \rightarrow (\hat{\theta}_a^1, \hat{\theta}_a^2)$

x_t

(x_t^1, x_t^2)

$b \rightarrow (\hat{\theta}_b^1, \hat{\theta}_b^2)$

Photos provided by Pexels

Exploit

UCB

$$\text{UCB score} = \bar{\mu}_t(a) + \sqrt{\frac{2\log T}{n_t(a)}}$$

Exploration

$\hat{\mu}_a(x_t) +$

$(1,2)$
$(2,3)$

$(1,2)$
$(1,2)$
$(1,2)$

ε-greedy

$$\left.\begin{aligned} \hat{\mu}_t(a) &\approx \bar{\mu}(a) \\ \hat{\mu}_t(b) &\approx \bar{\mu}(b) \end{aligned}\right\}$$

t :

Prob: 'ε' $\rightarrow$ heads $\rightarrow$ Explore

$1-\varepsilon \rightarrow$ Tail $\rightarrow$ Exploit

$(\hat{\theta}_a^1(t), \hat{\theta}_a^2(t))$

$(\hat{\theta}_b^1(t), \hat{\theta}_b^2(t))$

random

t

prob ε $\rightarrow$ Heads $\rightarrow$ a, b

$1-\varepsilon \rightarrow$ Exploit

$x_t \rightarrow \hat{\mu}_a(x_t) = \hat{\theta}_a^1(t)x_1^t + \hat{\theta}_a^2(t)x_2^t$

$\rightarrow \hat{\mu}_b(x_t)$

Photos provided by Pexels

LinUCB (UCB for Linear Bandits)

- At time t based on the history $\{(x_s, a_s, R_s)\}_{s=1}^{t-1}$, estimate the parameters for all $a \in \mathcal{A}$ using Ridge regression.
  - $\widehat{\theta^a} = (\mathrm{D_a^T\, D_a + I_d})^{-1}\, \mathrm{D_a^T b_a}$
  - If arm a is played m times till round t,
  - D_a is $m \times d$ context matrix whose rows correspond to the respective user feature vectors
  - b_a is $m \times 1$ reward vector corresponding to the rewards obtained during those rounds
  - $D_a = \begin{bmatrix} x_1^1 & x_2^1 \\ x_1^2 & x_2^2 \\ x_1^3 & x_2^3 \end{bmatrix}$ feature dimension $d = 2$, arm a played $m = 3$ times

Exploit **Explore**

- Pick the arm $a_t = \arg\max_a \; x_t^T \widehat{\theta^a} + \sqrt{x_t^T (D_a^T D_a + I_d)^{-1} x_t}$

Photos provided by Pexels