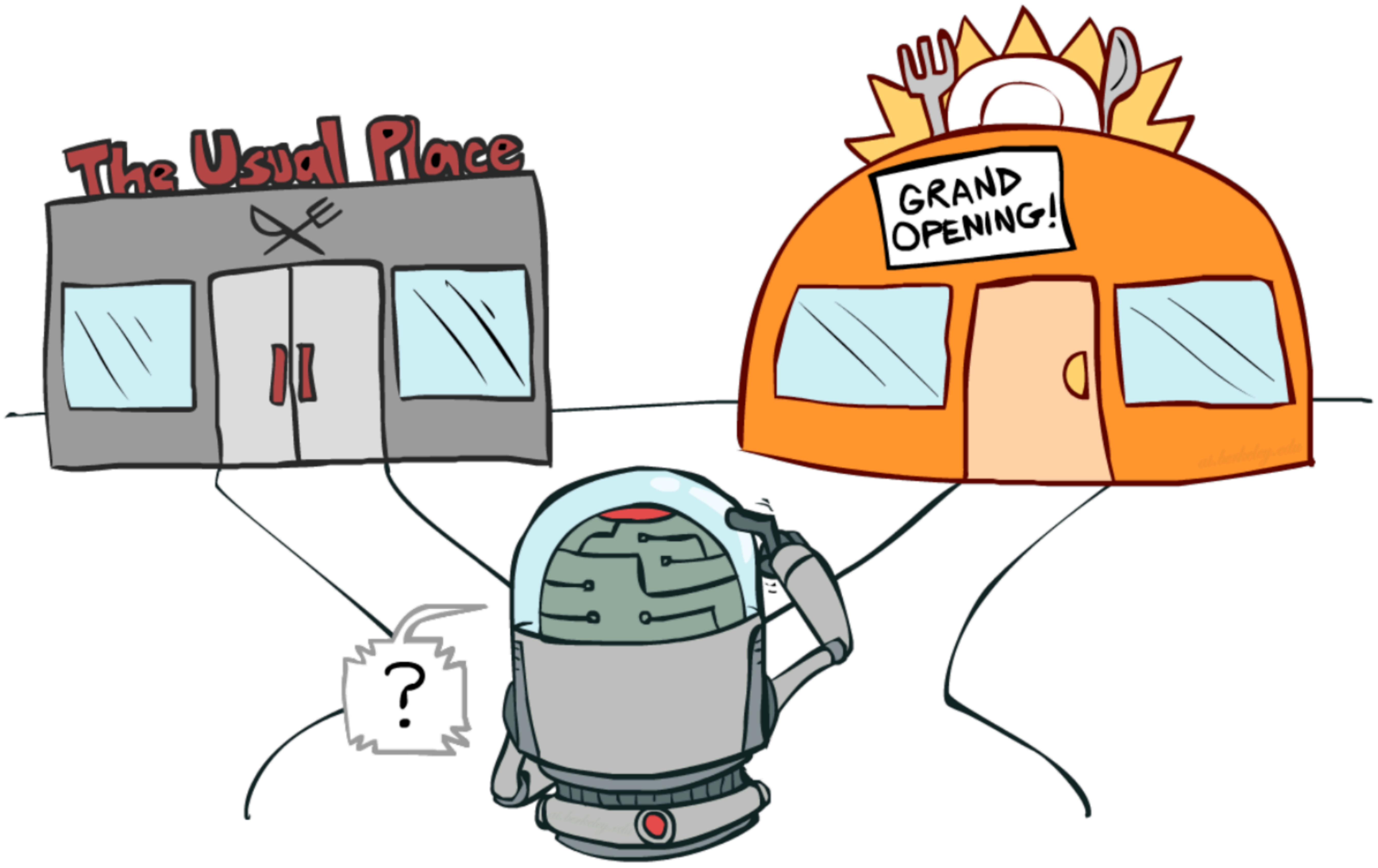
Exploration-Exploitation Dilemma in Multi-Armed Bandit

Kimang KHUN, Ph.D.
Ministry of Industry, Science, Technology & Innovation

Seminar of AMS Department, Institute of Technology of Cambodia, Phnom Penh

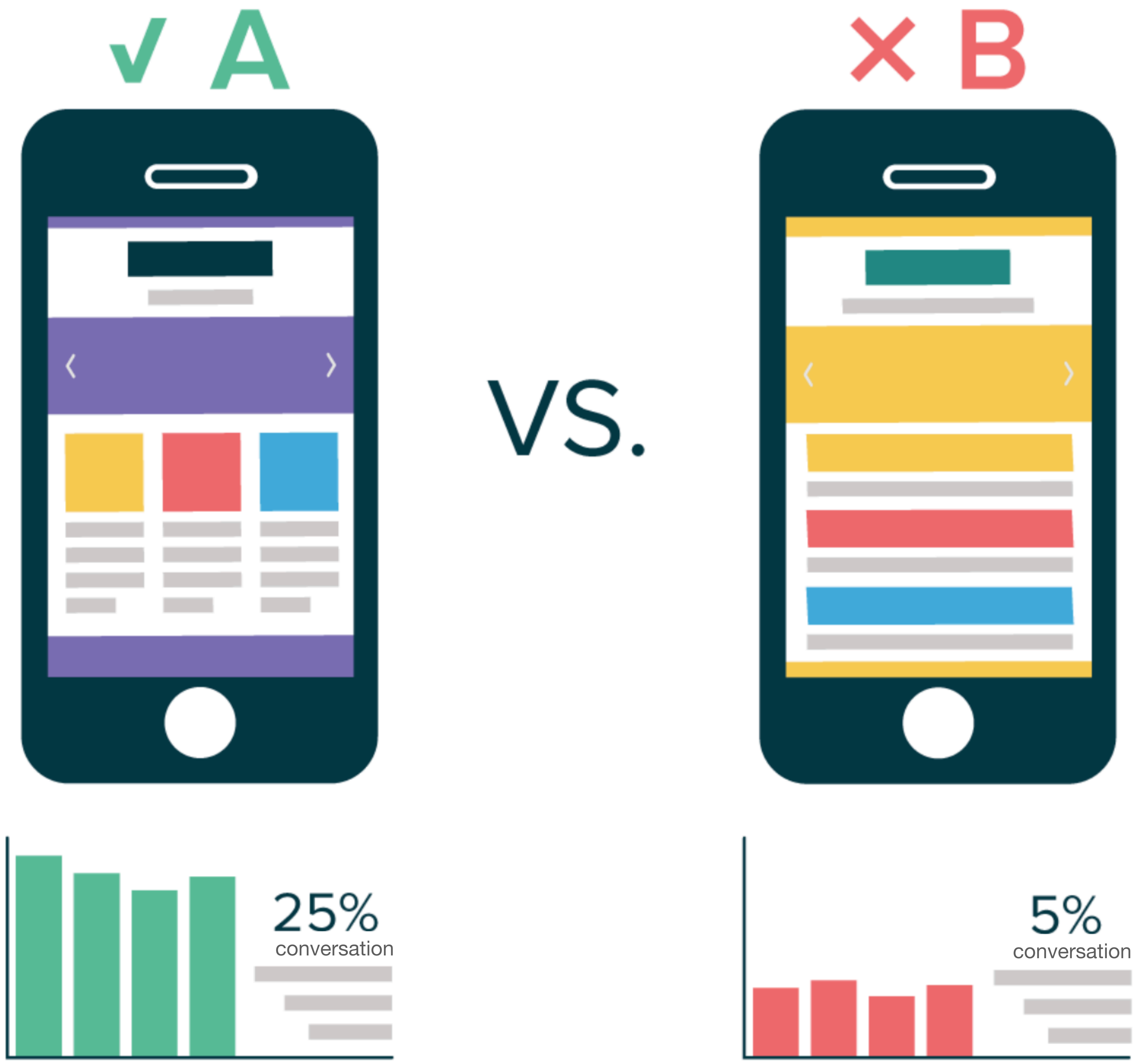
12th July 2023

Exploration and Exploitation Dilemma



Source: Nicolas Gast's slides

Example: A/B Testing

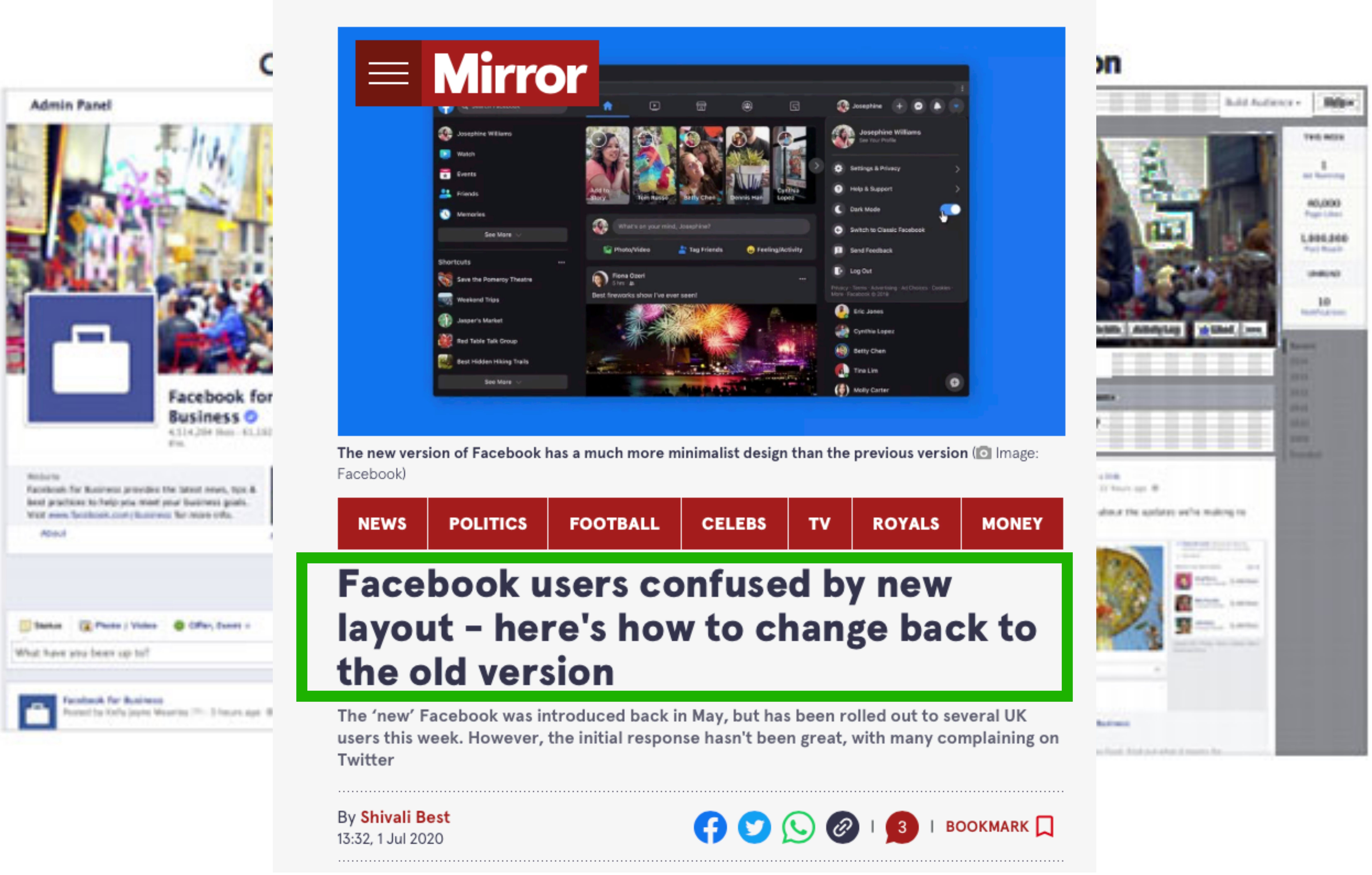


Example: A/B Testing

Old version

New version

Example: A/B Testing

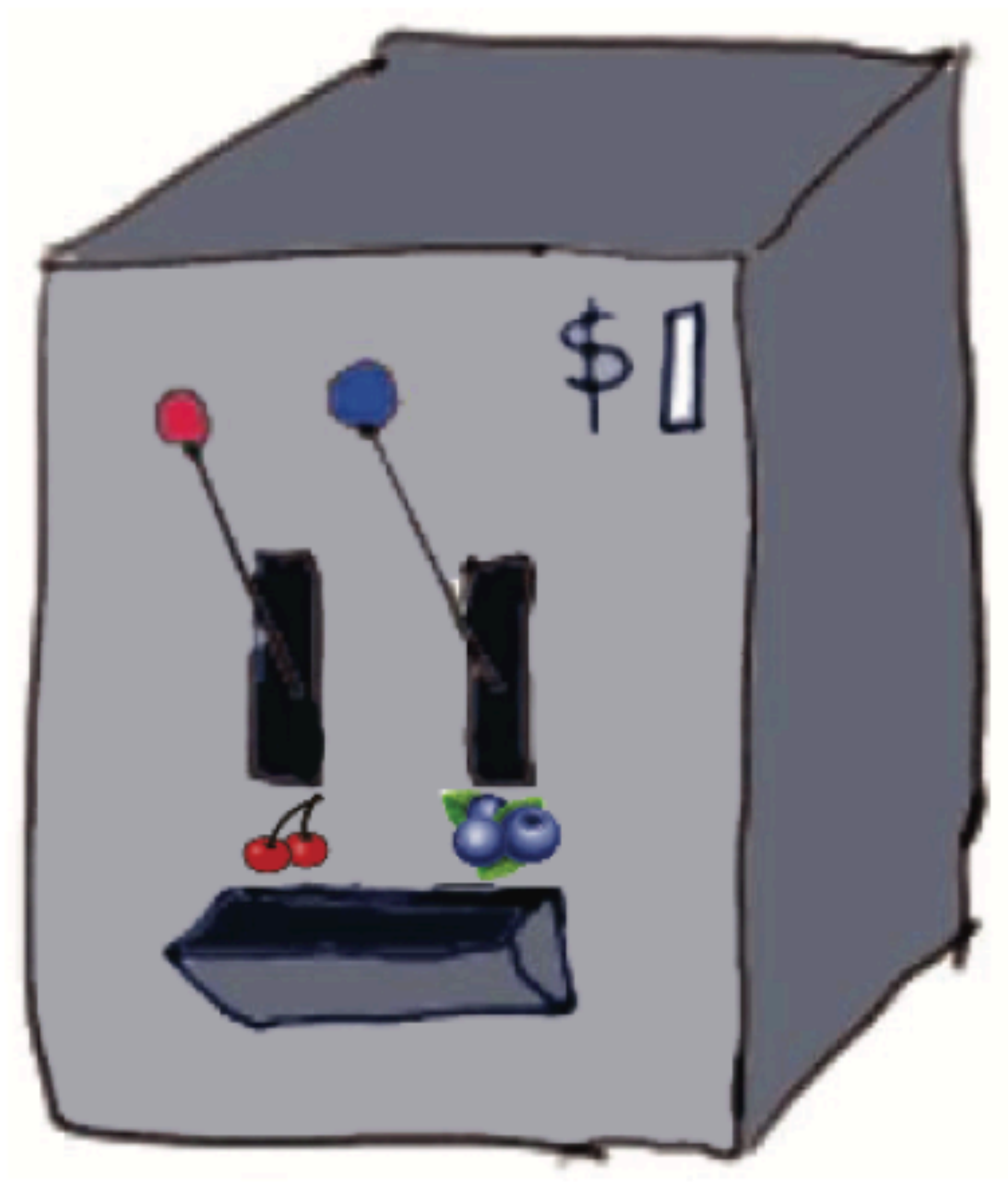


What can a scientist do?

We need mathematical formulation:

1. to design new models (algorithms)
2. to quantify the trade-off

Two-Armed Bandit Problem



Arm 1 Arm 2

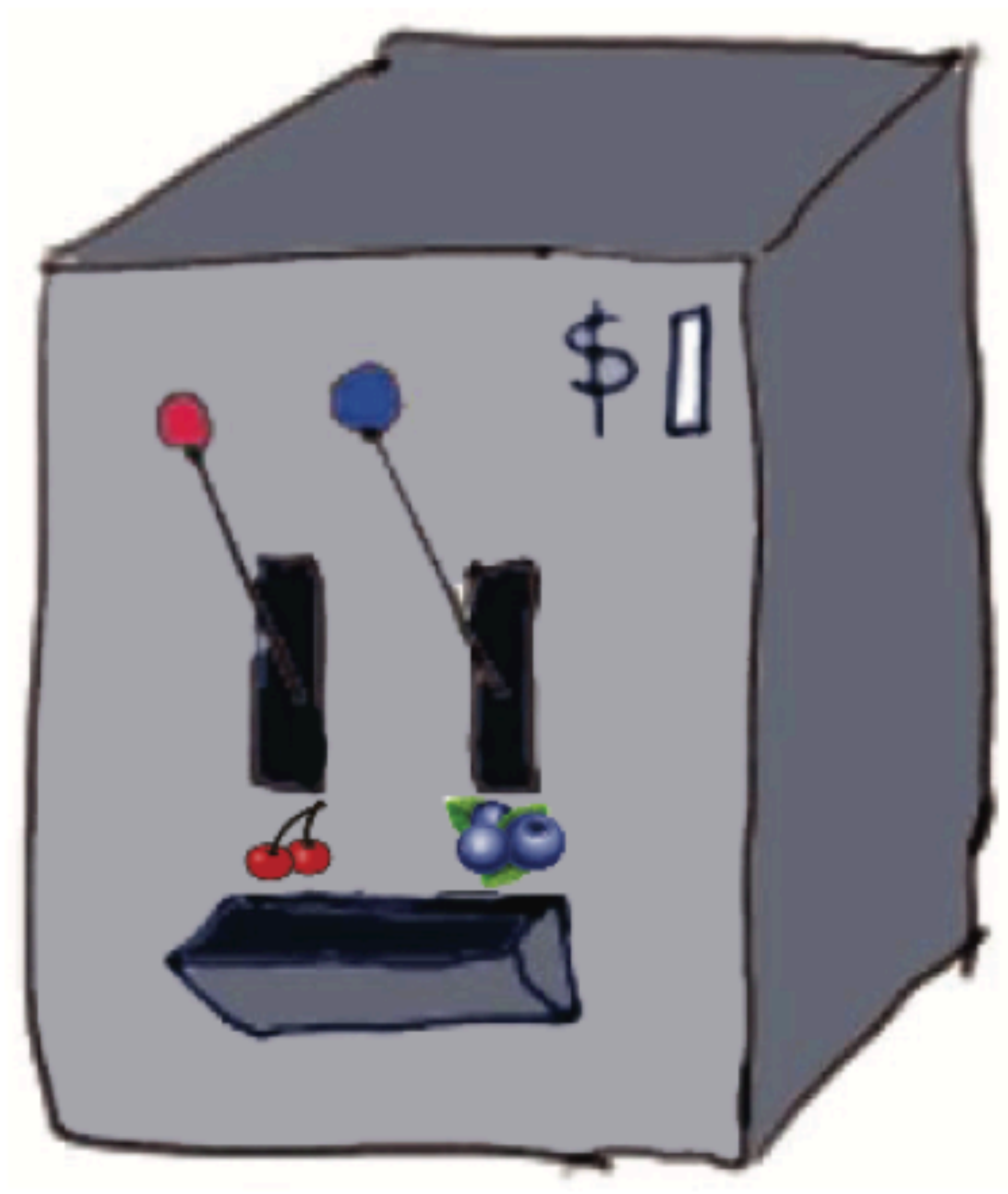
μ_1

Pulling Arm 1 gives 1 w.p. μ_1 or 0 w.p. $1 - \mu_1$.

A possible sequence of outcomes: $\underbrace{1,0,0,1,1,\ldots}_{T \text{ terms}}$ s.t.

$$\mu_1 = \lim_{T\to\infty} \frac{1}{T}\left(\underbrace{1+0+0+1+1+\ldots}_{T \text{ terms}}\right)$$

Two-Armed Bandit Problem



Arm 1 μ_1

Arm 2 μ_2

Pulling Arm 2 gives 1 w.p. μ_2 or 0 w.p. $1 - \mu_2$.

A possible sequence of outcomes: $\underbrace{0,0,0,0,1,\ldots}_{T \text{ terms}}$ s.t.

$$\mu_2 = \lim_{T \to \infty} \frac{1}{T} \left(\underbrace{0 + 0 + 0 + 0 + 1 + \ldots}_{T \text{ terms}} \right)$$

Two-Armed Bandit Problem

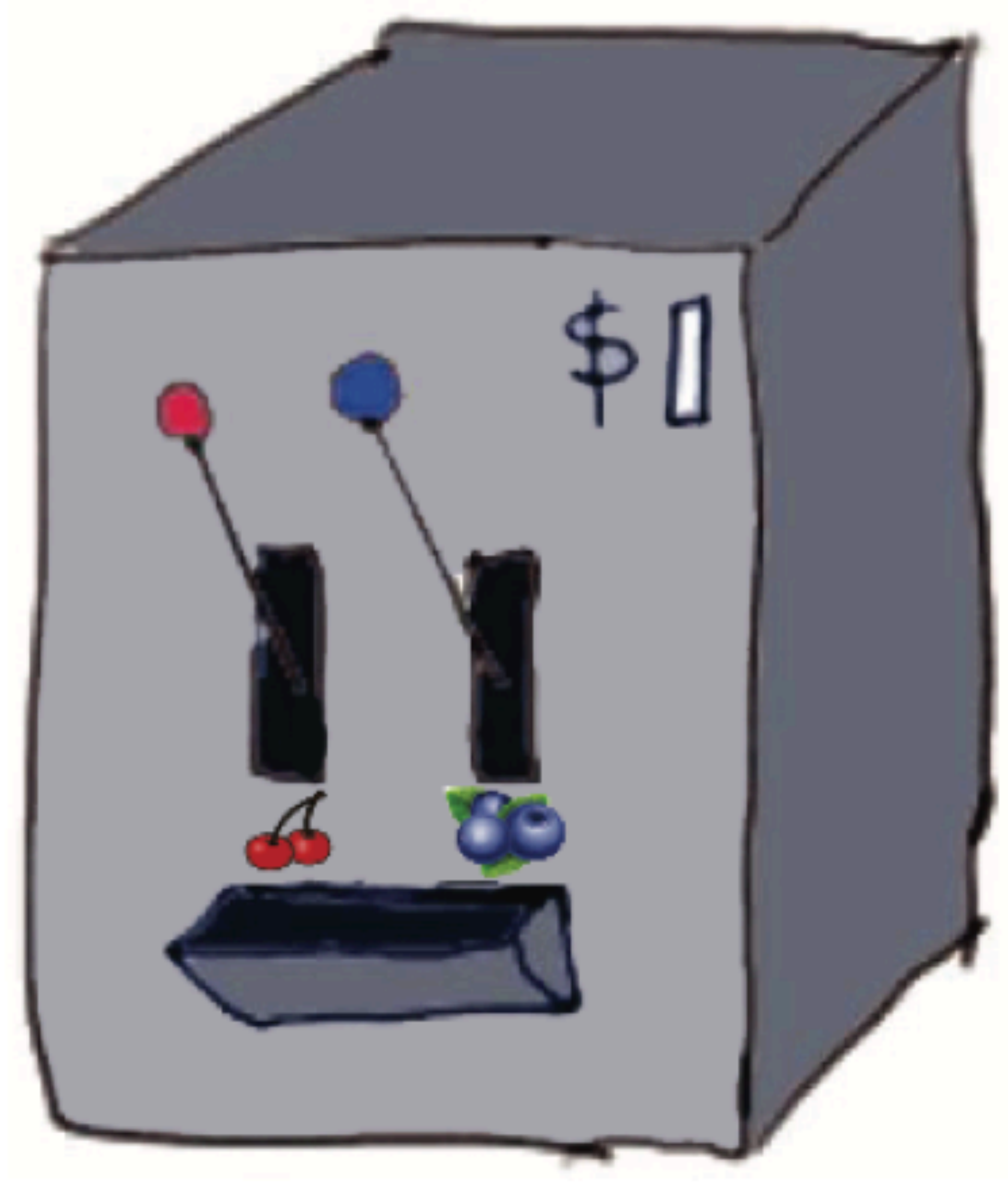
Arm 1 μ_1 Arm 2 μ_2

1, 0, 0, 1, 1, 0, 0, 0, 0, 1, …

Cumulative reward := 1+0+0+1+1+0+0+0+0+1+ …

Question: Which arm to pull so that the expected cumulative reward is **maximized**?

Two-Armed Bandit Problem



Arm 1 μ_1

Arm 2 μ_2

<u>**Question**</u>: Which arm to pull so that the expected cumulative reward is **maximized**?

If μ_1 and μ_2 are **KNOWN**, then
- always pull Arm 1 if $\mu_1 > \mu_2$
- always pull Arm 2 otherwise.

Challenge: μ_1 and μ_2 are **UNKNOWN**.

This is called "**Stochastic bandit**".

Motivation

Maximize clicks

Title	Click probability
"Murder Victim found in an Adult Entertainment Venue"	μ_1
"Headless body found in Topless bar"	μ_2

Choose which title to display. Observe "Click/Not Click".

Clinical trials

μ_1 μ_2

Choose treatment for patient.

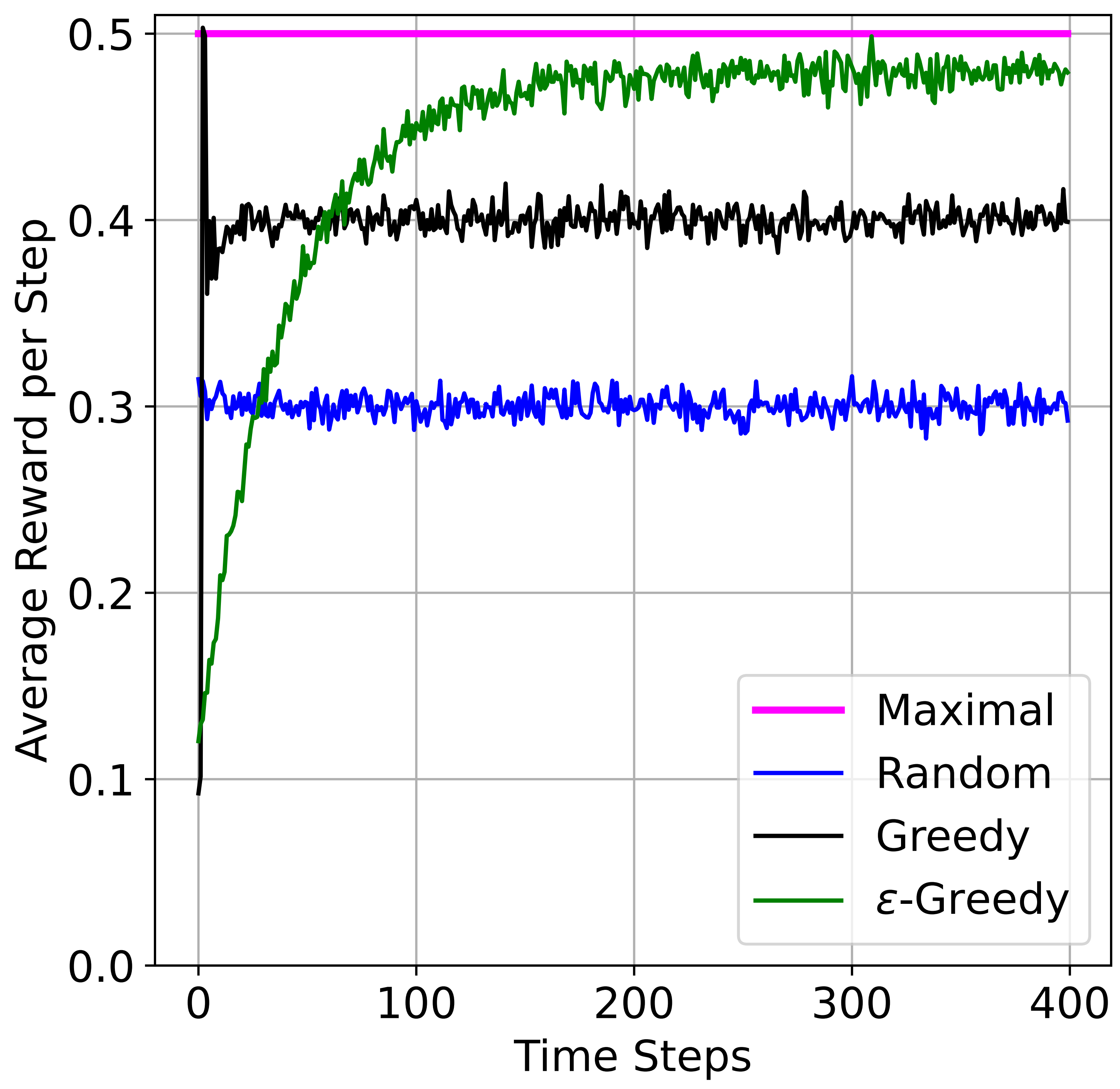
Observe "Heal/Not Heal".

Algorithm Design

- **Random**: at each decision time, **uniformly randomly** pull one arm.
- **Greedy**: initially try each arm the same number of pulls, then commit on the best arm.
- ε**-Greedy**: w.p. ε, **uniformly randomly** pull one arm (<u>Exploration</u>), and w.p. $1 - \varepsilon$, pull the best arm so far (<u>Exploitation</u>).

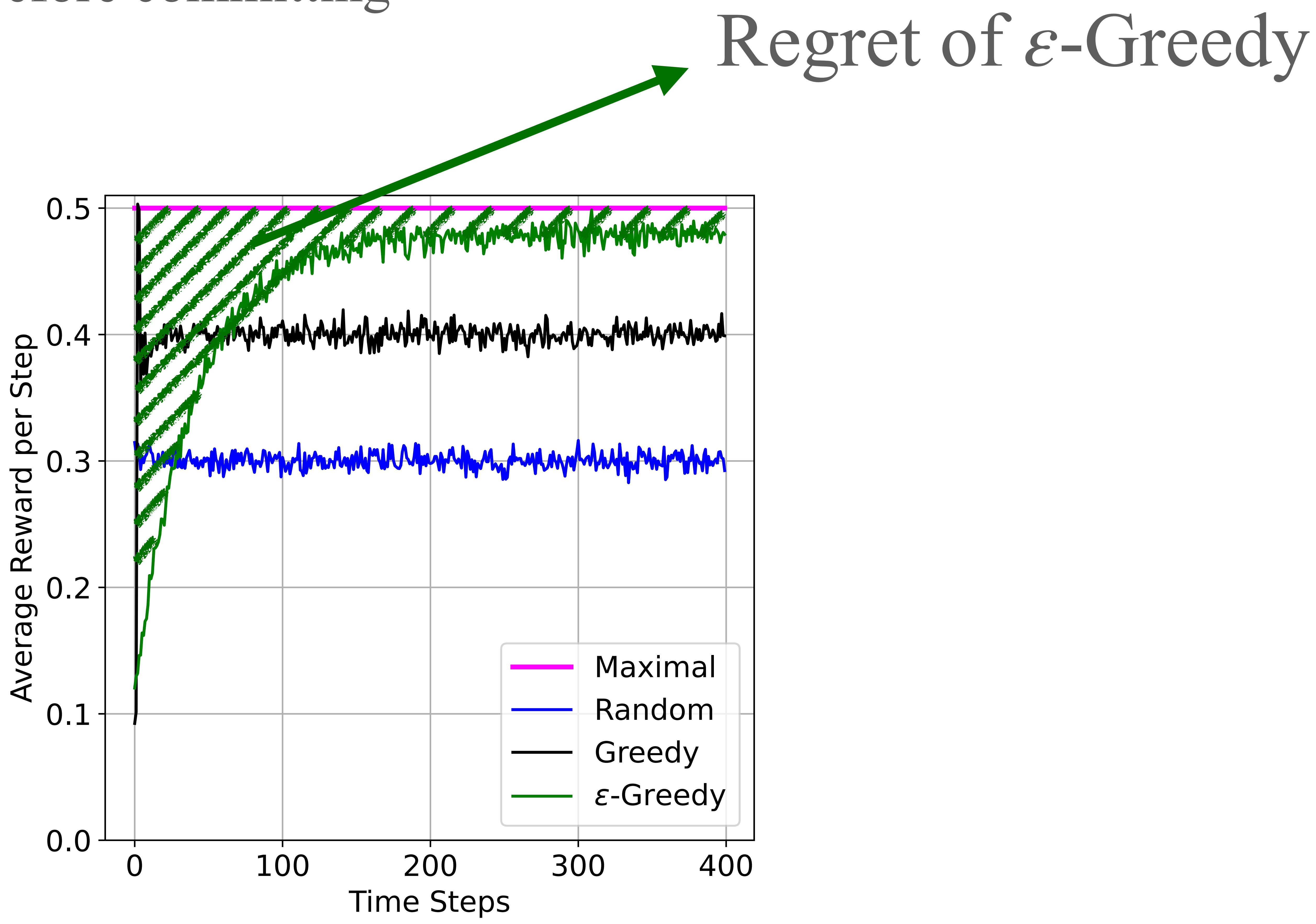
Algorithm Design

- Setup: $\mu_1 = 0.1$ and $\mu_2 = 0.5$
- Greedy: try each arm 2 pulls before committing
- ε-Greedy: $\varepsilon = 0.1$



Algorithm Design

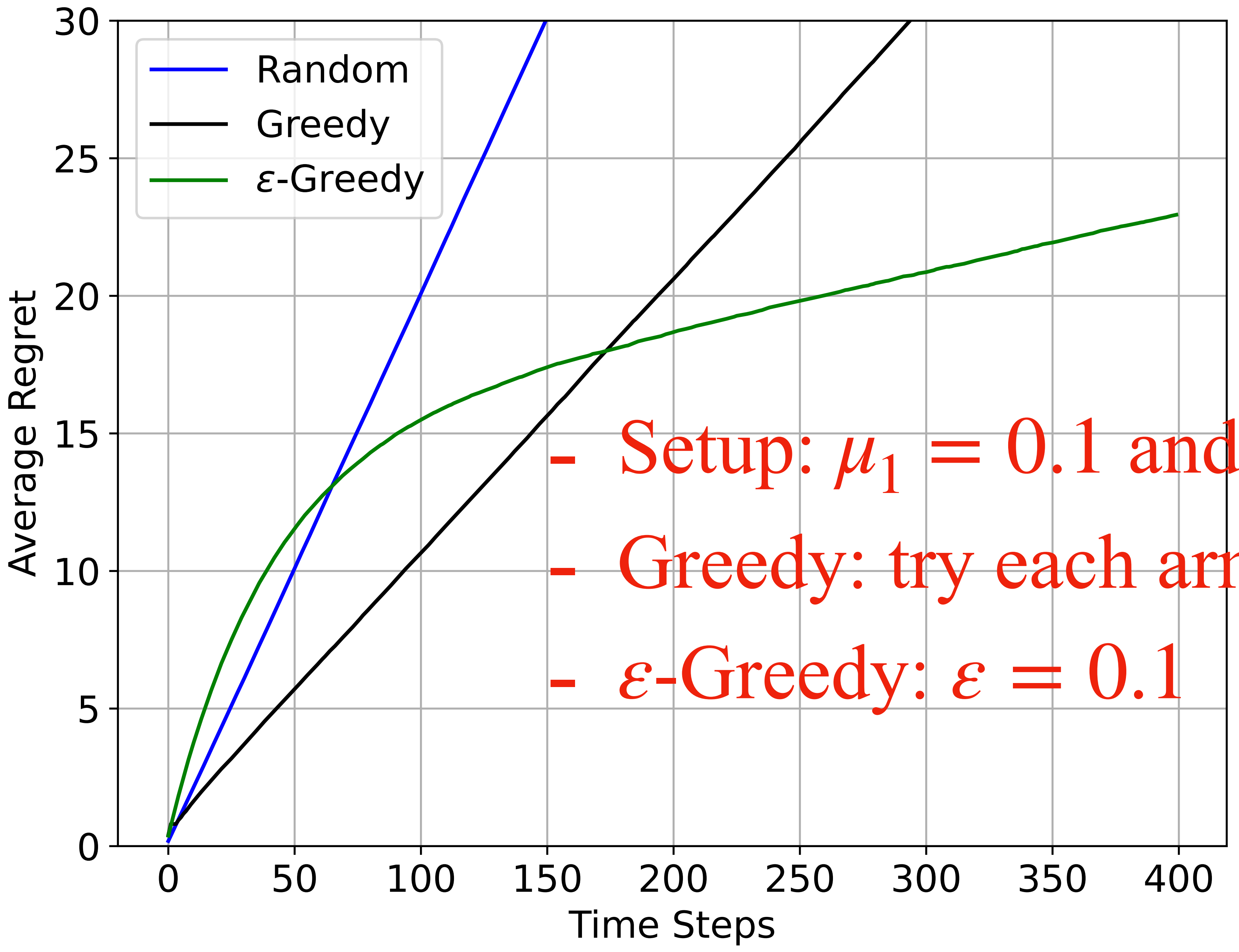
- Setup: $\mu_1 = 0.1$ and $\mu_2 = 0.5$
- Greedy: try each arm 2 pulls before committing
- ε-Greedy: $\varepsilon = 0.1$



Performance metric: Regret

Regret of $\mathscr{A}$:= (**maximal** cumulative reward) - (cumulative reward of $\mathscr{A}$).

The **smaller** the **regret** is, the **better** $\mathscr{A}$ **performs**.



- Setup: $\mu_1 = 0.1$ and $\mu_2 = 0.5$
- Greedy: try each arm 2 pulls before committing
- ε-Greedy: $\varepsilon = 0.1$

Performance metric: Regret

Regret of $\mathscr{A}$:= (**<u>maximal</u>** cumulative reward) - (cumulative reward of $\mathscr{A}$).

The **smaller** the **regret** is, the **better** $\mathscr{A}$ **performs**.

Let T be total steps.

The **regret** of ε-Greedy is $O(T)$ (this is called **linear regret**).

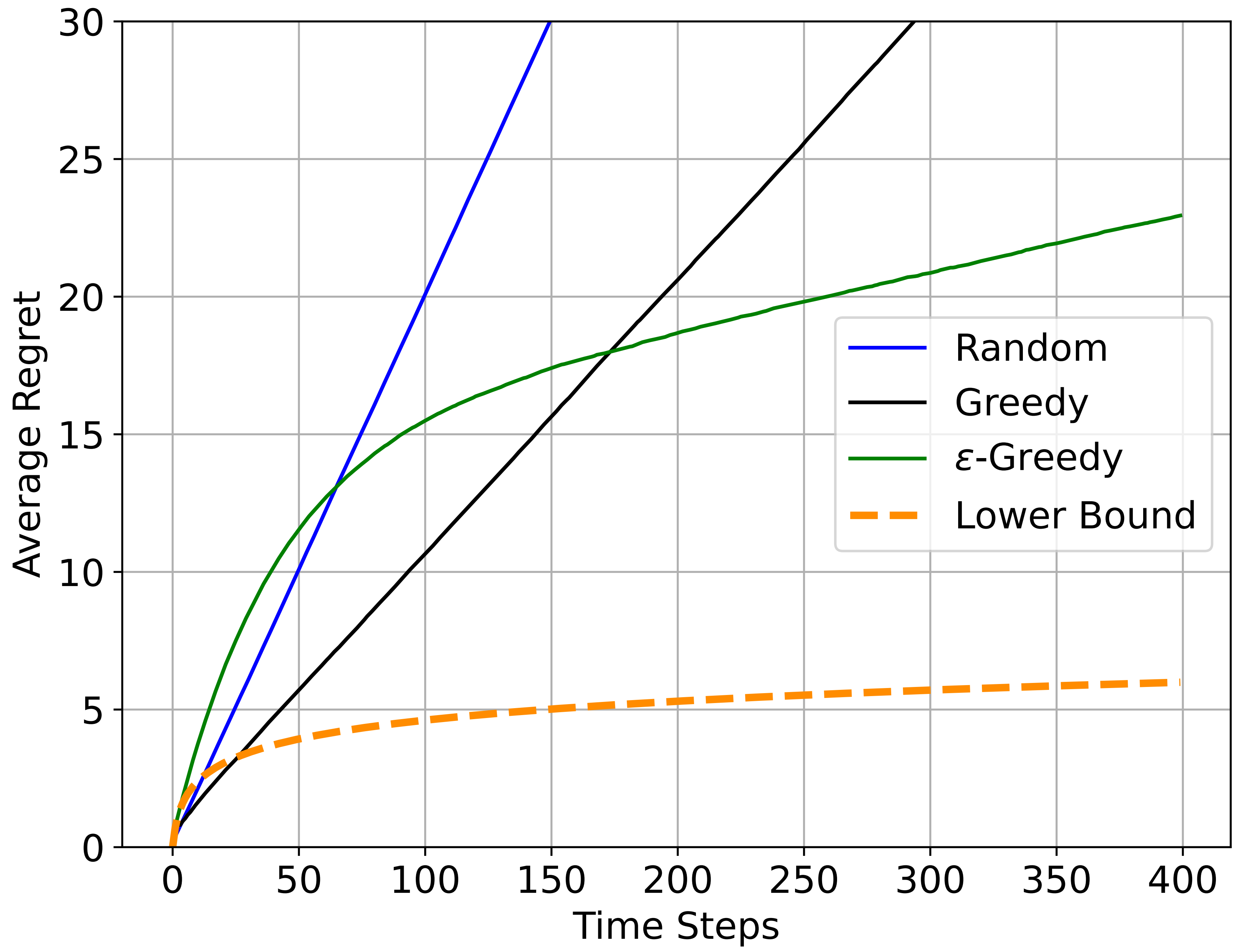
Can we do better?

Lower bound on Regret

Theorem 1 (Lai & Robbins, 1985)

There exists a constant c (that depends on μ) s.t. any uniformly efficient[1] algorithm $\mathscr{A}$ satisfies:

$$\text{Regret of } \mathscr{A} \geq c \ln T.$$



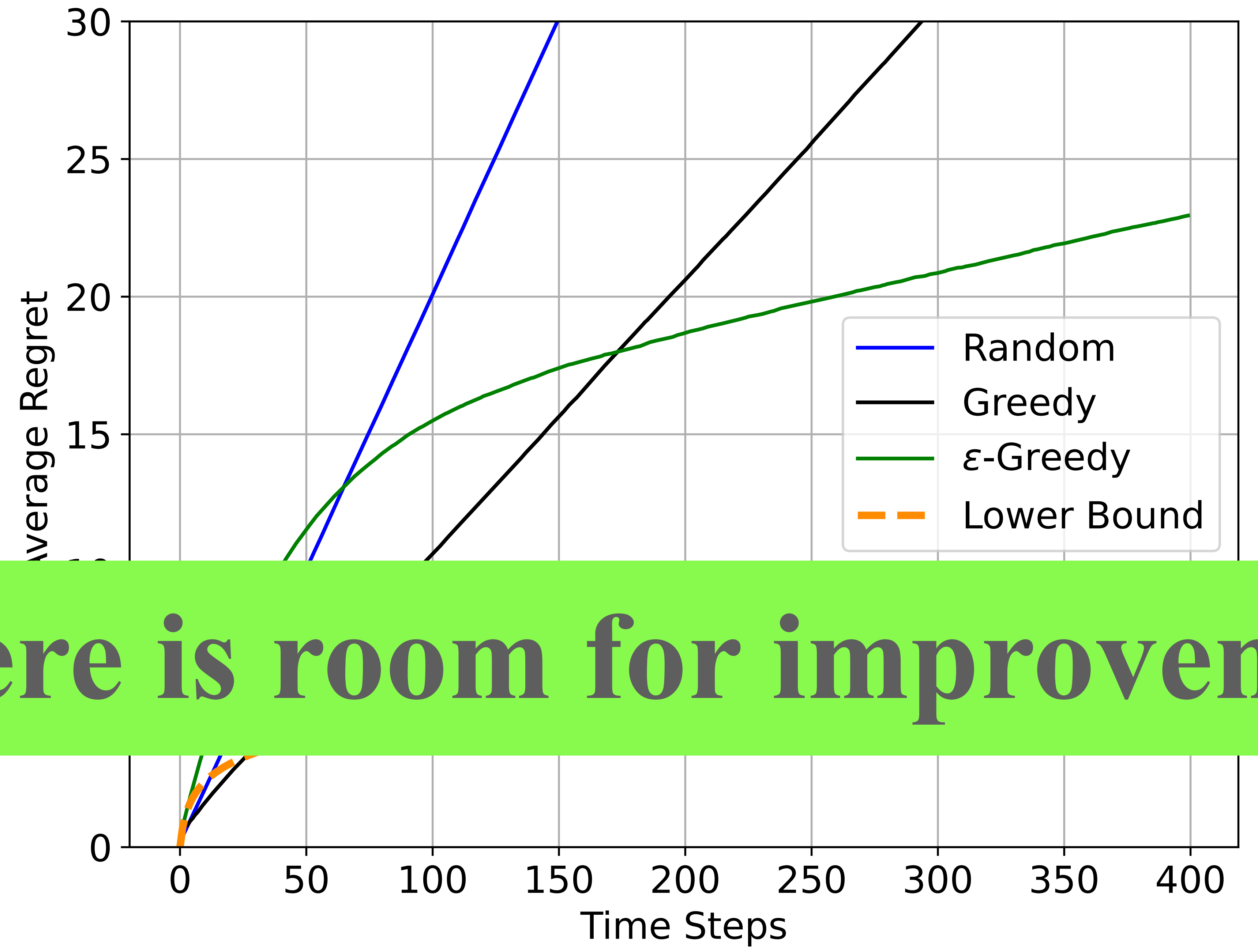
[1] Meaning that Regret of $\mathscr{A}$ is $o(T^\alpha)$ for all μ and α.

Lower bound on Regret

Theorem 1 (Lai & Robbins, 1985)

There exists a constant c (that depends on μ) s.t. any uniformly efficient[1] *algorithm $\mathscr{A}$ satisfies:*

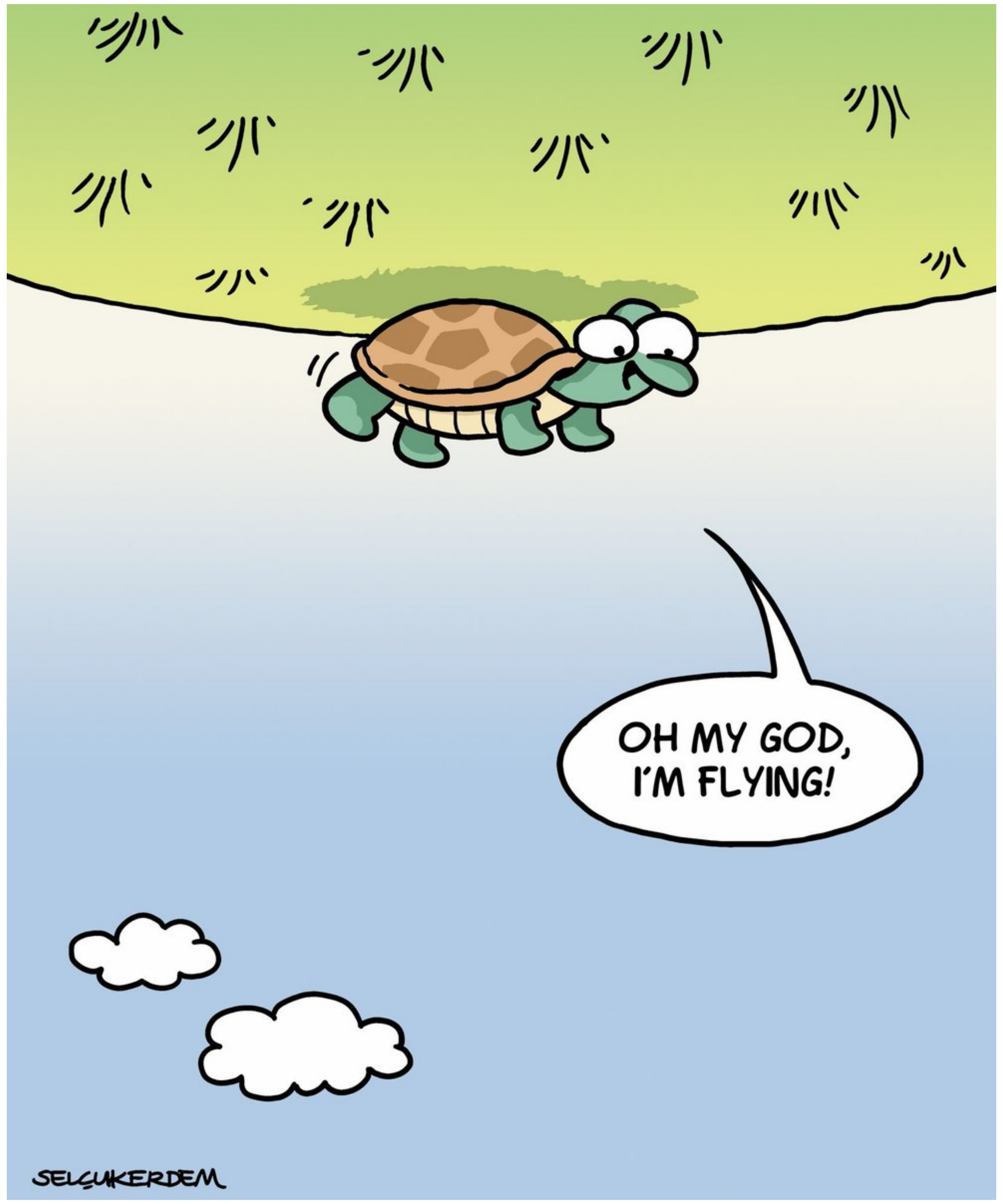
$$\text{Regret of } \mathscr{A} \geq c \ln T.$$



There is room for improvement!

[1] Meaning that Regret of $\mathscr{A}$ is $o(T^\alpha)$ for all μ and α.

Optimism in Face of Uncertainty (OFU)



When you are uncertain, consider the **best possible environment**.

If the **best possible environment** is **correct**
⇒ No reward lost
Exploitation

If the **best possible environment** is **wrong**
⇒ Gather useful info.
Exploration

Upper Confidence Bound (UCB)

Consider a coin that gives "Head" w.p. μ.

Suppose that you toss the coin N times and observe "Head" n times.

The natural estimator of μ is:

$$\hat{\mu} := \frac{n}{N}.$$

By Hoeffding's inequality, we have that[2] for $x > 0$,

$$\mathbb{P}\left\{-\sqrt{\frac{x}{2N}} + \hat{\mu} \leq \mu \leq \hat{\mu} + \sqrt{\frac{x}{2N}}\right\} \geq 1 - 2e^{-x}.$$

[2] under the assumption that all the observations are i.i.d.

Upper Confidence Bound (UCB)

Consider a coin that gives "Head" w.p. μ.

Suppose that you toss the coin N times and observe "Head" n times.

The natural estimator of μ is:

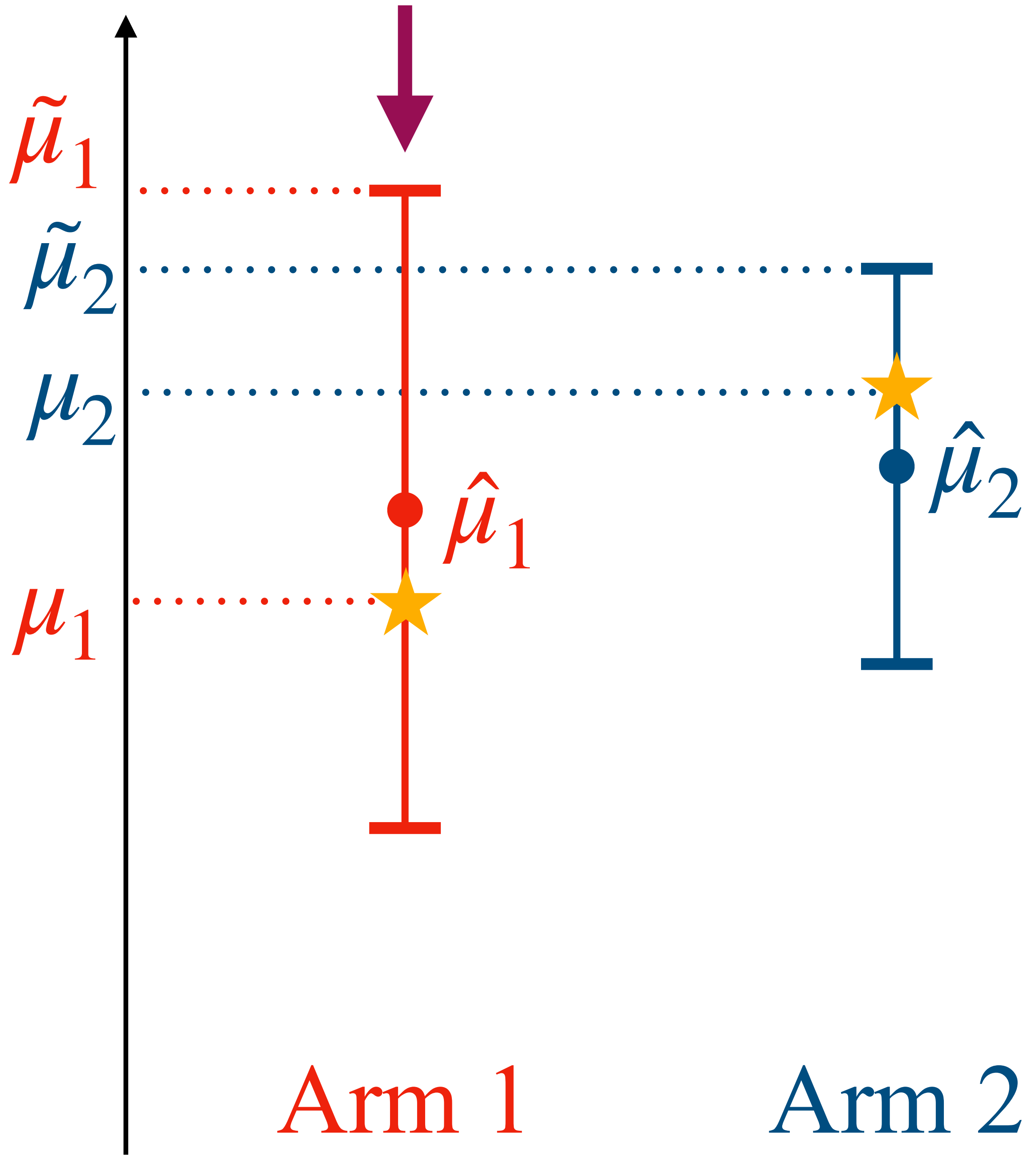
$$\hat{\mu} := \frac{n}{N}.$$

By Hoeffding's inequality, we have that[2] for $x > 0$,

$$\mathbb{P}\left\{-\sqrt{\frac{x}{2N}} + \hat{\mu} \leq \mu \leq \underbrace{\hat{\mu} + \sqrt{\frac{x}{2N}}}_{\tilde{\mu}}\right\} \geq 1 - 2e^{-x}.$$

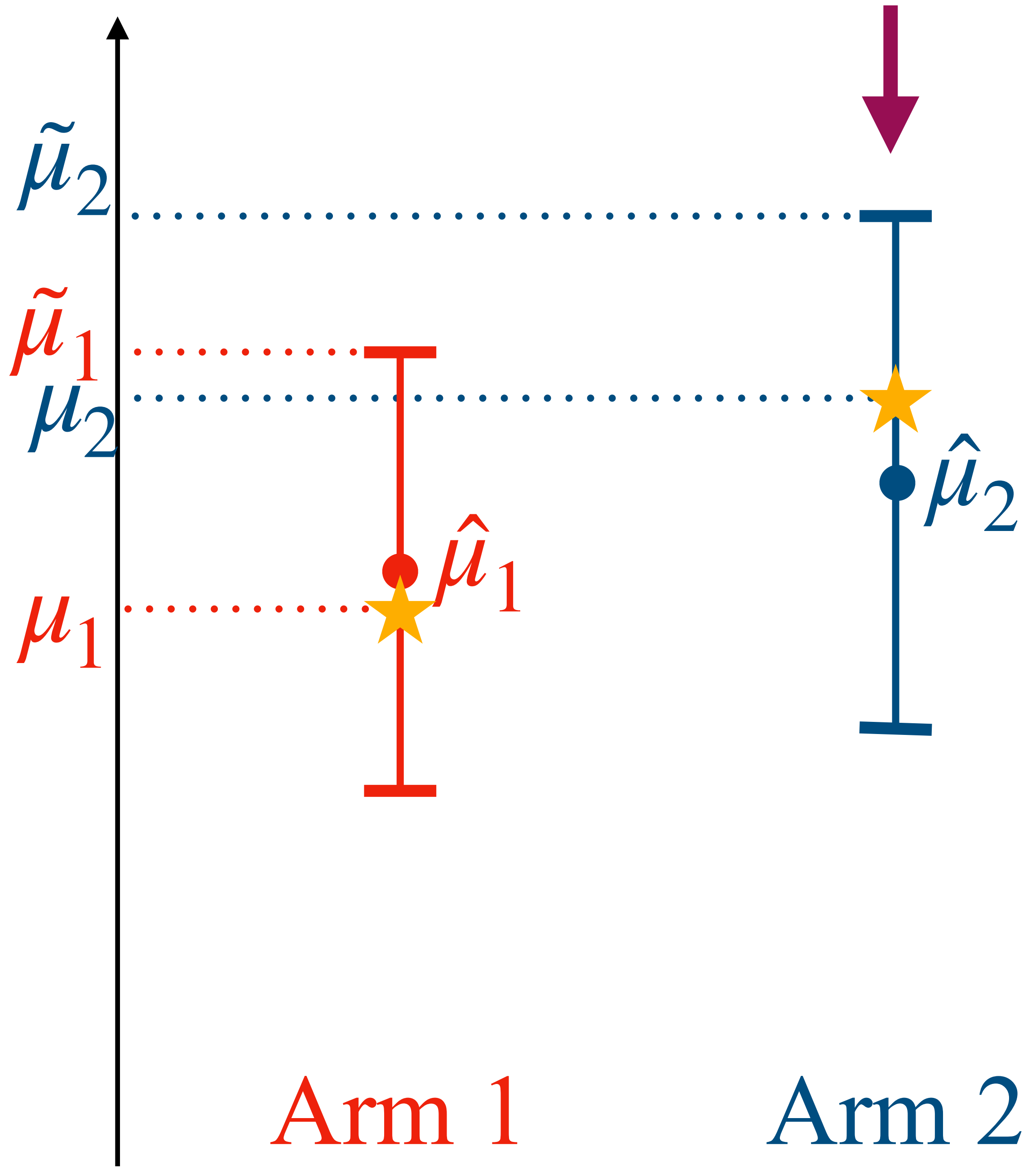
[2] under the assumption that all the observations are i.i.d.

Upper Confidence Bound (UCB)



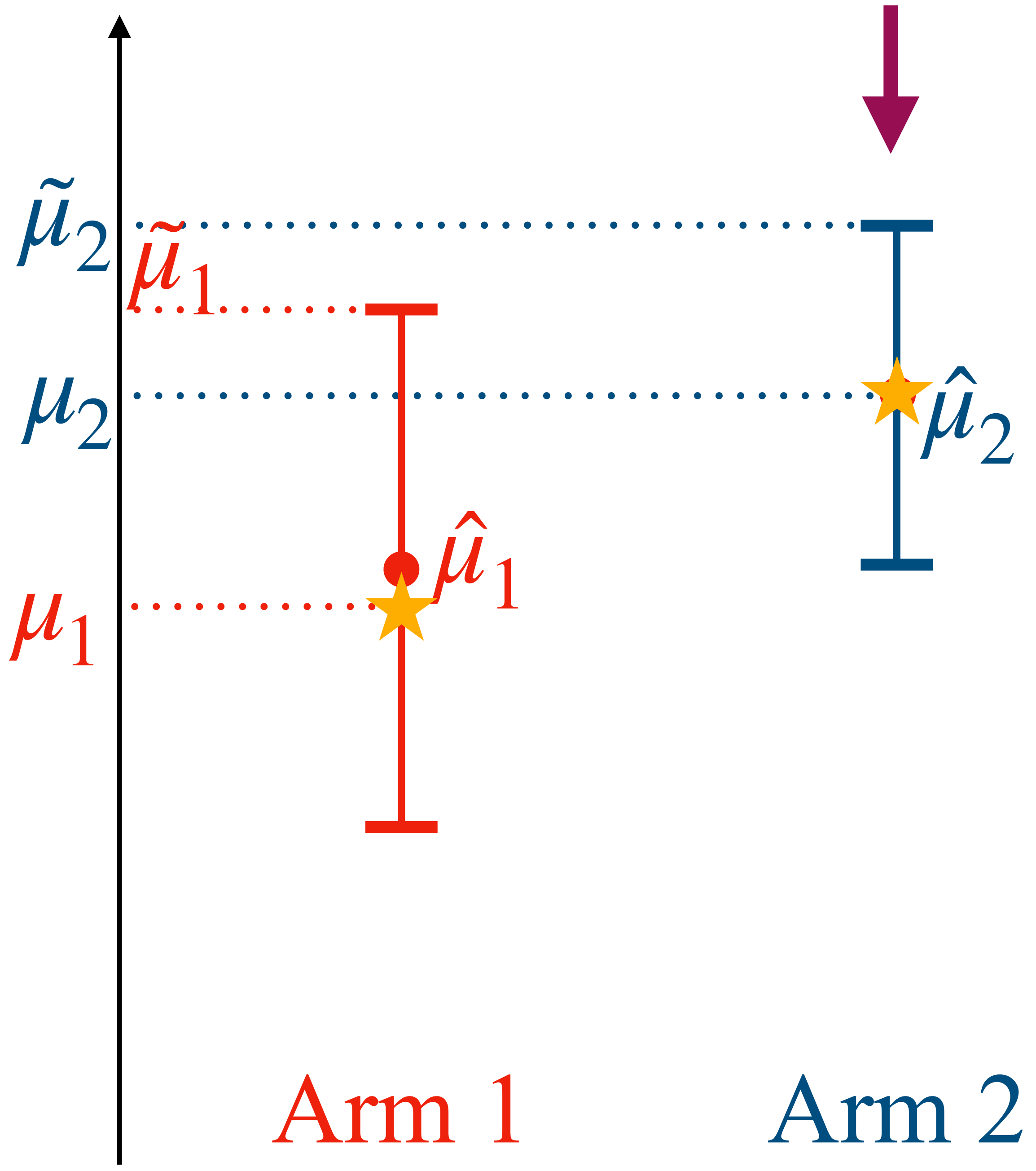
At time step t

Upper Confidence Bound (UCB)



At time step $t + 1$

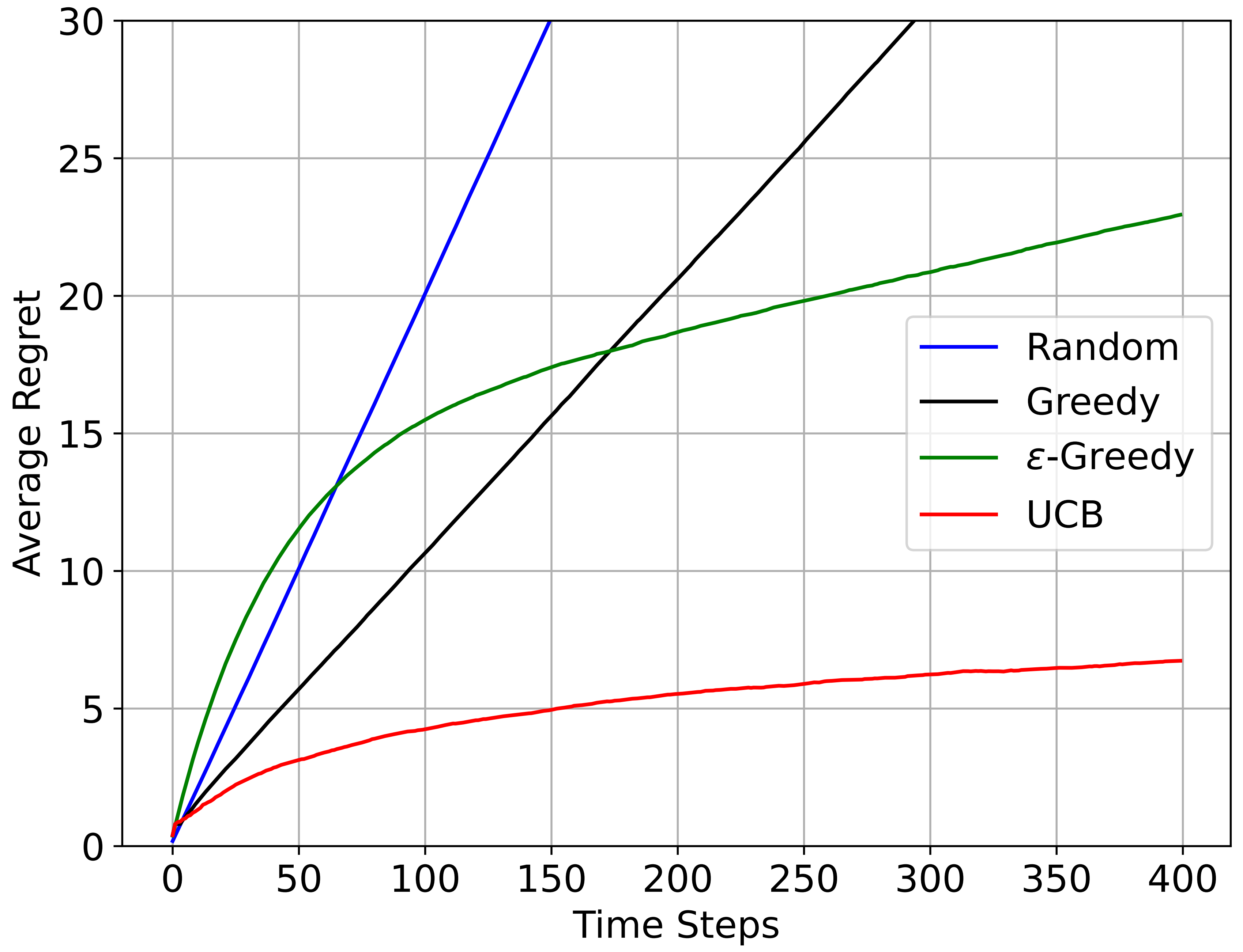
Upper Confidence Bound (UCB)



At time step $t + 2$

Upper Confidence Bound (UCB)

- Setup: $\mu_1 = 0.1$ and $\mu_2 = 0.5$
- Greedy: try each arm 2 pulls before committing
- ε-Greedy: $\varepsilon = 0.1$



Upper Confidence Bound (UCB)

Theorem 2 (Auer et al., 2002):

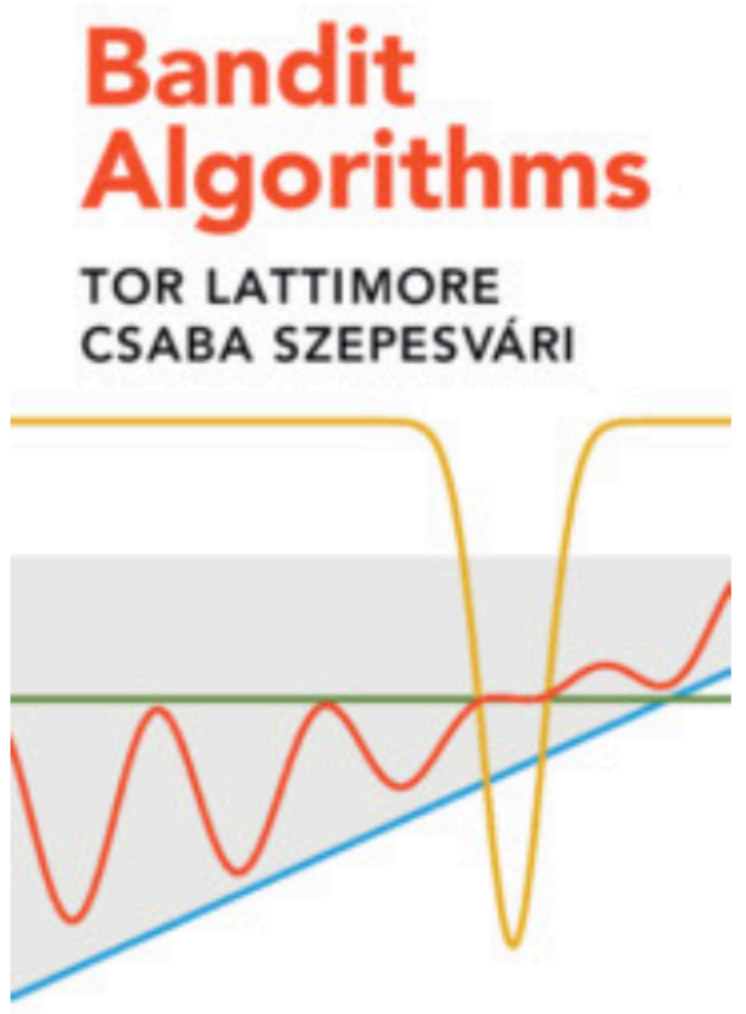
$$\textit{Regret of UCB} \leq c' \ln T.$$

We say that UCB is **asymptotically optimal**.

Conclusion

- **Exploration vs. Exploitation (EE) dilemma** happens in the world of decision-making under uncertainty.
- **Multi-armed bandit** (MAB) problem is a mathematical formulation allowing us to consider the EE trade-off and design new algorithms.
- We use **Regret** to measure the algo.'s performance.
- **No algorithm** has a regret smaller than $O(\ln T)$ uniformly over all MAB problems.
- **UCB** algorithm from **OFU** approach has a regret bounded by $O(\ln T)$ (it is **asymptotically optimal**).

For more on bandit, check out this book



source: https://twitter.com/parveenkaswan/status/1364791588442890240?lang=zh-Hant

https://kimang18.github.io or khun.kimang@misti.gov.kh

Questions?