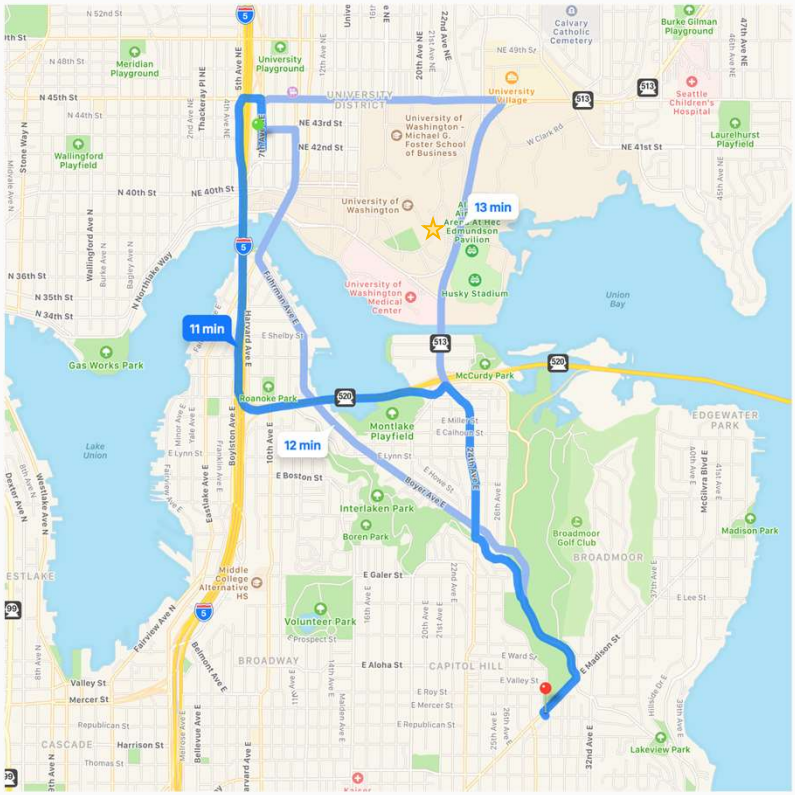
Multi-Armed Bandits

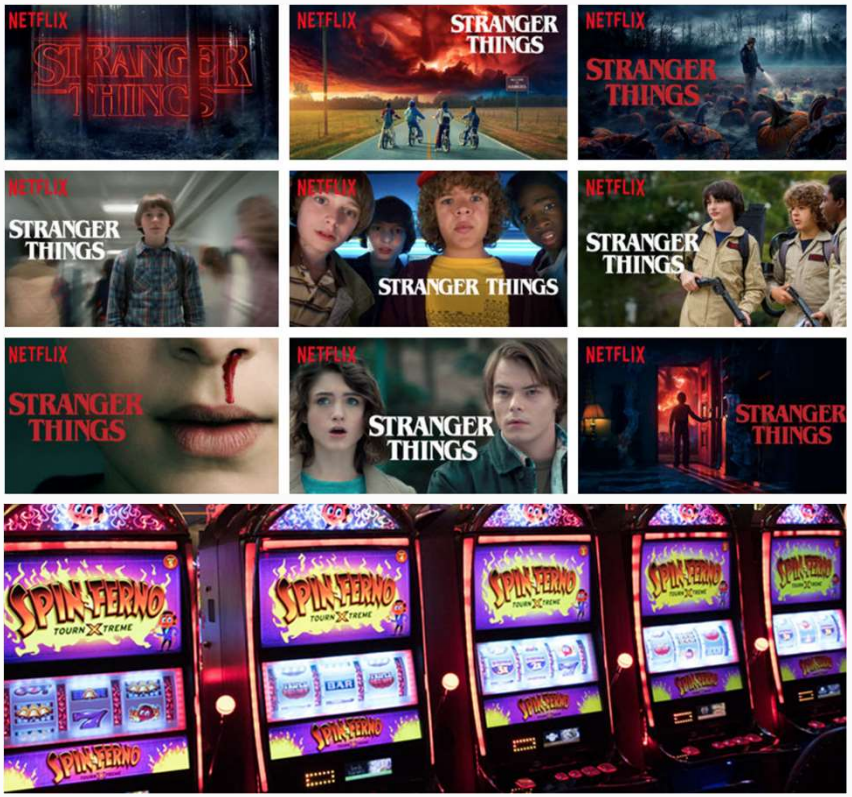
A talk on the benefits of being optimistic in the face of uncertainty

Hunter Schafer, Paul G. Allen School of Computer Science & Engineering

BE BOUNDLESS

Some not so similar problems





Let's go gambling

- We walk into a casino and are offered 100 free pulls on a slot machine.
- How much money will we make?
- Expected Payout: $\mu = \$2$


```
Rewards: [2,   2, 2, 2, 2, ...]

Rewards: [0,   4, 4, 0, 0, ...]

Rewards: [0, 100, 0, 0, 0, ...]
```

A bunch of bandits

$\mu_1 = \$?.4$

$\mu_2 = \$?.5$

$\mu_3 = \$?.1$

The multi-armed bandit game

- N slot machines (arms) each with **unknown** mean μ_i
- The rewards from an arm are **normally distributed** with mean μ_i
- T rounds of the game

for $t = 1 \ldots T$:

1) Choose arm A_t out of all the possible slot machines
2) Receive reward X_t with $E[X_t] = \mu_{A_t}$

- Goal: Minimize notion of "regret"
    - Compare against expected winnings of a player who knew the best arm in hindsight

How good is a slot machine?

- Simplify problem: Estimate μ of just one slot machine after receiving rewards: $[4,\ 1,\ 2,\ 3,\ 3]$

$\hat{\mu}(5)$

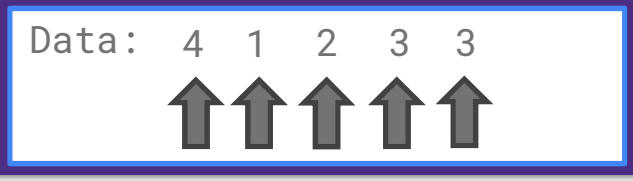
↳ $\hat{\mu} = \frac{1}{5}(4 + 1 + 2 + 3 + 3) = 2.6$

In general with n pulls

$$\hat{\mu}(n) = \frac{1}{n}(X_1 + X_2 + \ ... + Xn)$$

- How confident are you in your prediction?
  - If given 2 estimates $\hat{\mu}(2)$ and $\hat{\mu}(500)$, which one is better? What does "better" mean?

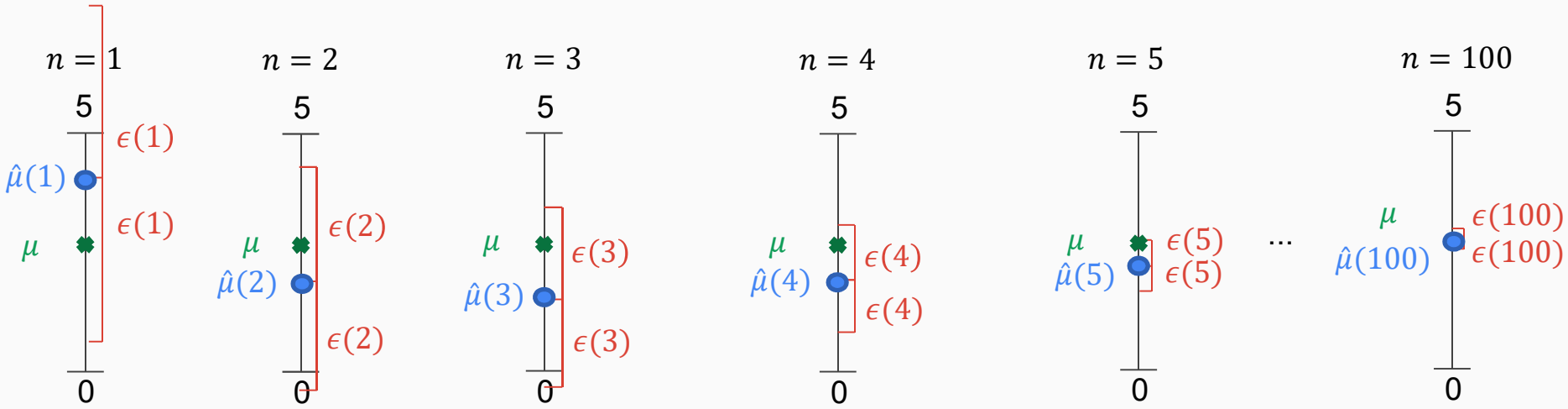
Quantifying Uncertainty



Reads: Low probability ($\leq 5\%$) that estimate $\hat{\mu}(n)$ being more than $\epsilon(n)$ far away from μ

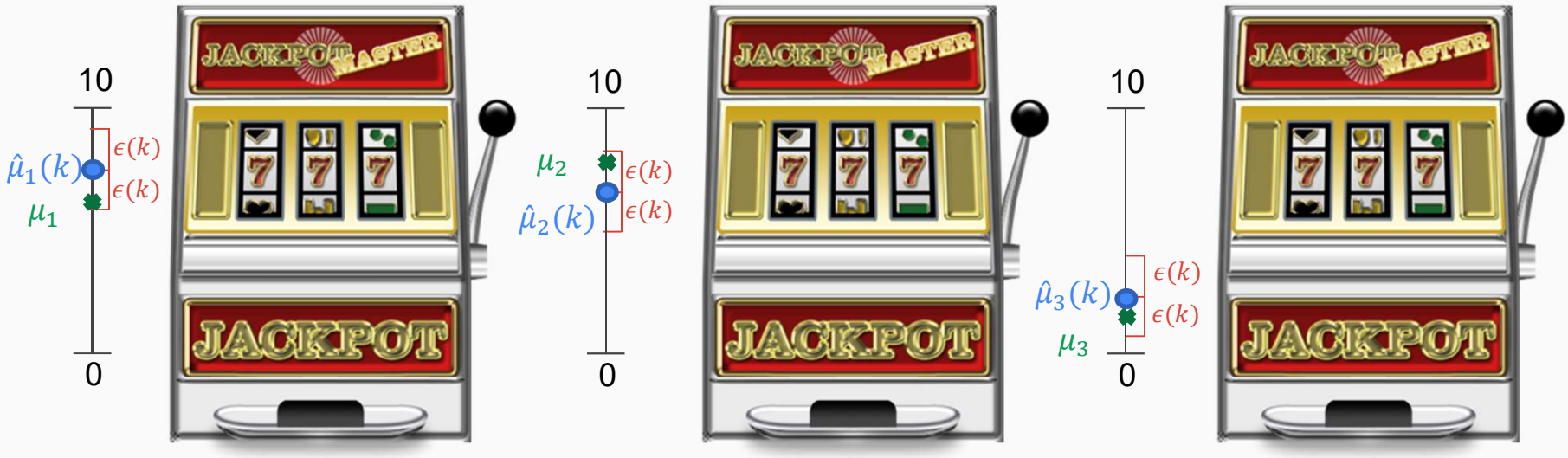
- Estimate is good if it's unlikely to be too far away from true value

$$\Pr(|\hat{\mu}(n) - \mu| \geq \epsilon(n)) \leq 0.05$$



Strategy: Explore then commit

- *Explore* each arm k times to estimate means
- *Commit* to the one that looks best



How to choose k

k should be large!

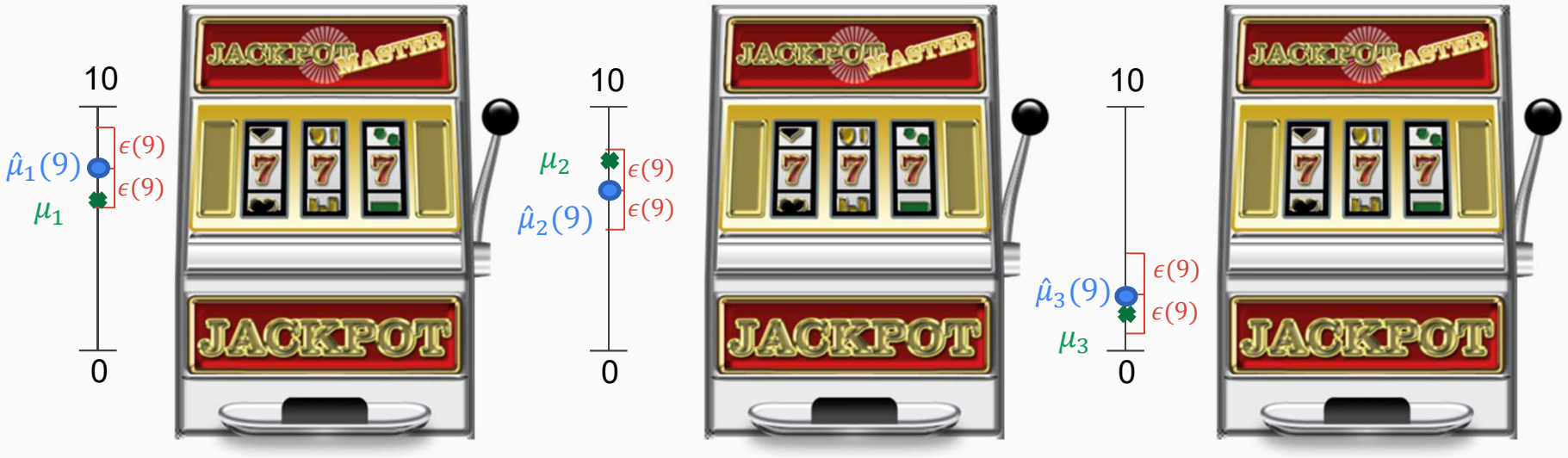
- Allows for good approximation of means
- Unlikely to commit to the incorrect arm

k should be small!

- We only get T turns, this strategy requires Nk exploratory pulls
- *Might be wasting pulls on obviously bad arms!*

Strategy: Explore then commit

- Assume *k = 20*
- Say we have pulled each arm 9 times already
- Is it safe to disregard arm 3?



How to choose k

k should be large!

- Allows for good approximation of means
- Unlikely to commit to the incorrect arm

k should be small!

- We only get T turns, this requires Nk exploratory pulls
- *Wasted pulls on obviously bad arms*

The right k?

- Optimal choice depends on gaps of true means, which we don't know 😢

A better idea: Adapt to the situation

- Use a strategy that figures out which arm to pull on each round based on past information (don't use a fixed strategy)
- Must strike a balance between

Exploration

- Want to explore arms that we are unsure about

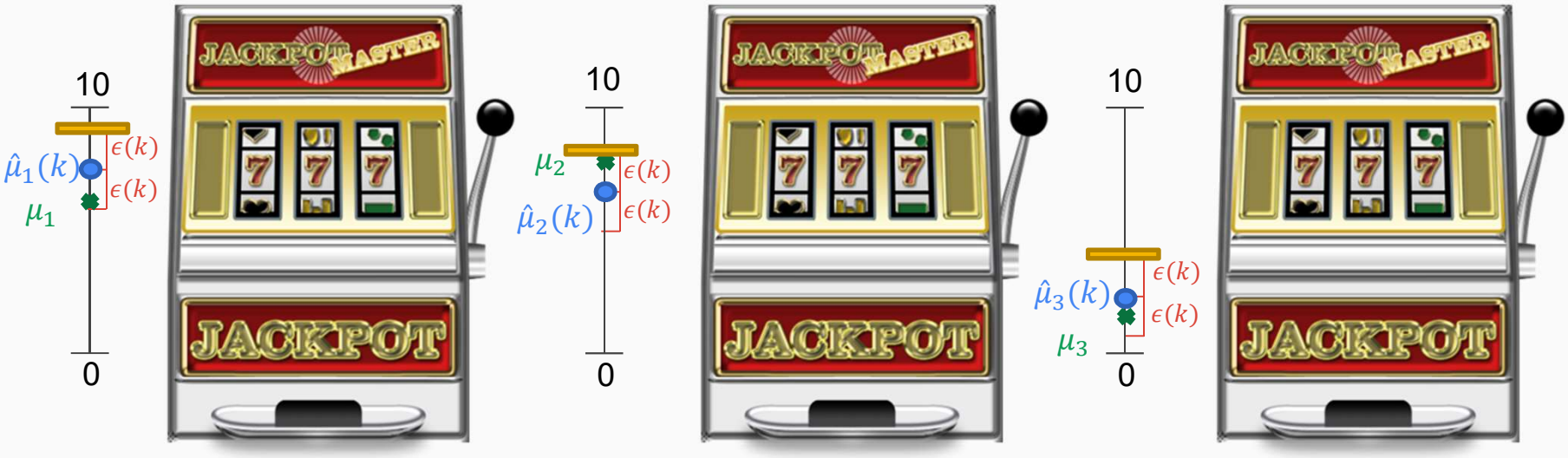
Exploitation

- Want to get high reward from arms that currently look good

- One approach: Optimism in the face of uncertainty!

Optimism in the face of uncertainty

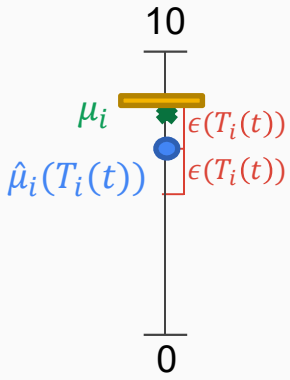
- In the absence of knowing the true mean, assume it's the largest possible value we would expect to see with the previous information



Strategy: Upper Confidence Bound (UCB)

- Currently on turn t
- Let $T_i(t)$ be the # times arm i was pulled before time t
- For each arm i, compute upper confidence bound (UCB)
- Play arm i with highest $UCB_i(T_i(t))$ as action A_t

$$UCB_i(T_i(t)) = \hat{\mu}_i(T_i(t)) + \epsilon(T_i(t))$$



$$= \hat{\mu}_i(T_i(t)) + c\sqrt{\frac{\log(t)}{T_i(t)}}$$

Goes up as time goes on

Goes down if we pull more often

Exploitation

Exploration

Intuition: Why this works

- On turn t you play action A_t. There are 2 possibilities
  - A_t is the best arm:
    - Good job! 🎉
    - You'll get your estimate closer to the true value, which will encourage play later
  - A_t is a suboptimal arm:
    - You'll regret this a little 😔
    - Why did you choose this arm again?
      - $\hat{\mu}_{A_t}$ looked really good
      - $\sqrt{\frac{\log(t)}{T_{A_t}(t)}}$ was large due to uncertainty

Both terms should go down after pull!

Regrets

- Mathematically, our notion of regret is

$$R(T) = \max_j E\left[\sum_{t=1}^{T}\left(X_{j,t} - X_{A_t,t}\right)\right]$$

- After some math, we find that the regret for UCB is

$$E[R(T)] = \mathcal{O}\left(\sqrt{nT\log T}\right)$$

Average Regret

$\frac{E[R(T)]}{T} \Rightarrow 0$ as $T \to \infty$

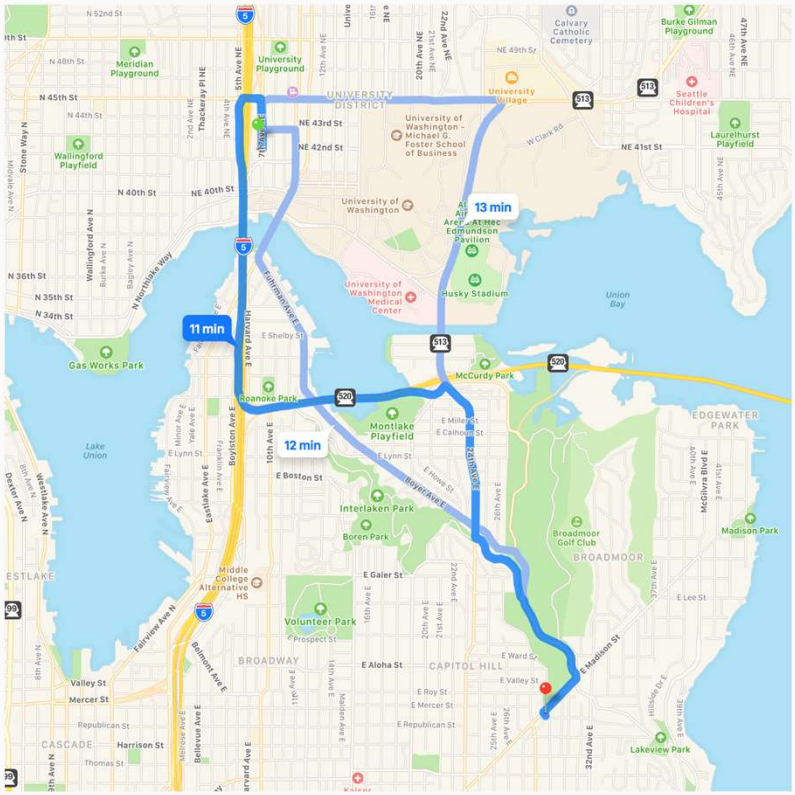
This means we are incurring sub-linear regret (i.e. we are learning!)

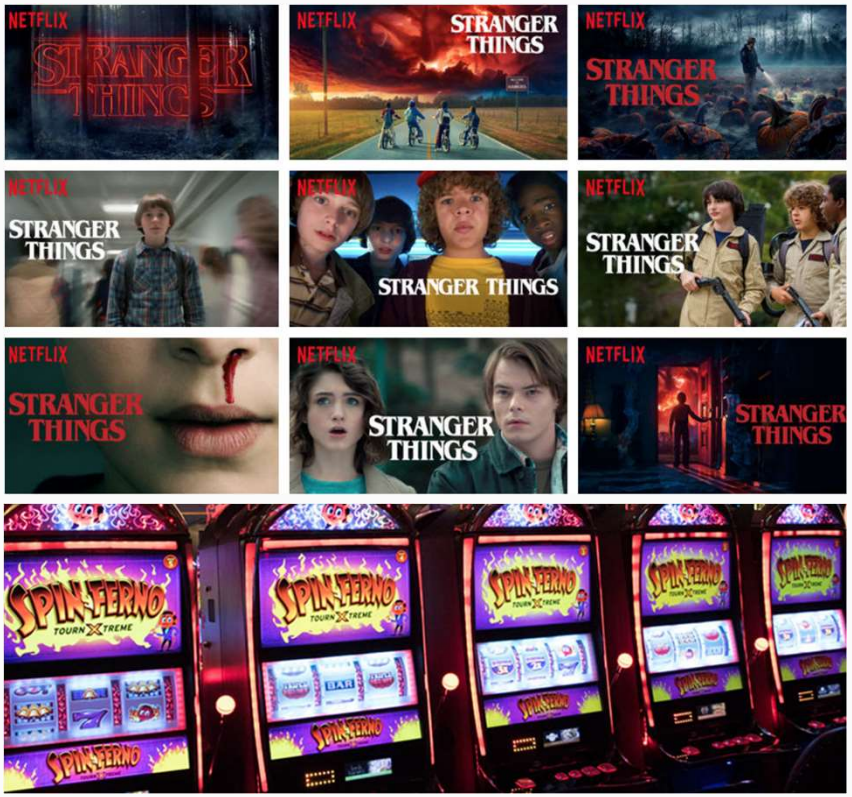
⚠ Word of Caution

We've made some assumptions

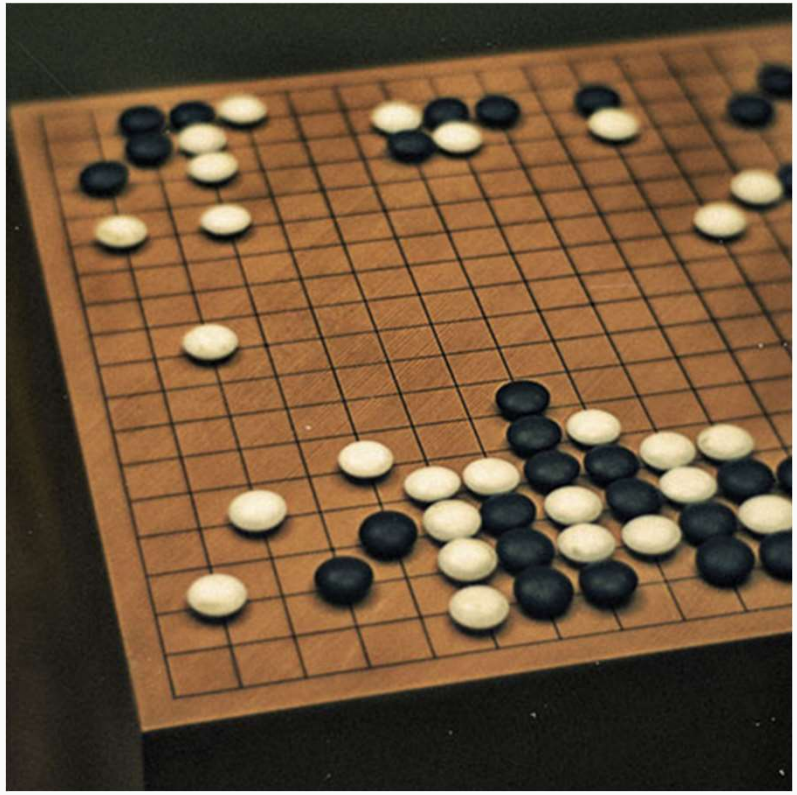
- We assumed means don't change over time, the rewards were normally distributed, and your actions don't affect future rewards.
- If not true, the analysis becomes a lot trickier ☹
- Need new framework and algorithm in different scenarios. Examples:
  - Adversarial bandits (e.g. no assumption of stochastic rewards)
  - Pure Exploration (e.g. best medicine identification)
  - Contextual bandits (e.g. you have side information about choices)

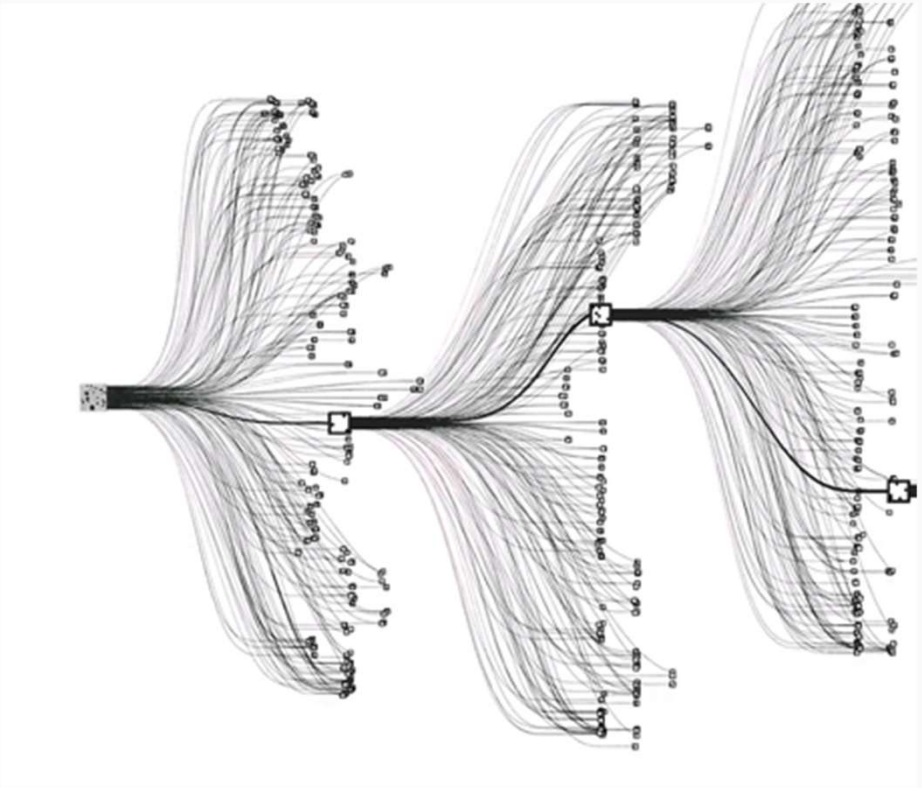
Some not so similar problems?





Some not so similar problems?





Impact of Online Learning

Online methods will continue to become more prevalent

- Robotics w/ Reinforcement Learning
- Stock Market: Portfolio Selection
- Artificial Intelligence w/ Smarter Search
- Hyperparameter Optimization
- A/B Testing
- And much much more!