Multi-armed Bandit I

An Introduction

Thanh Nguyen-Tang

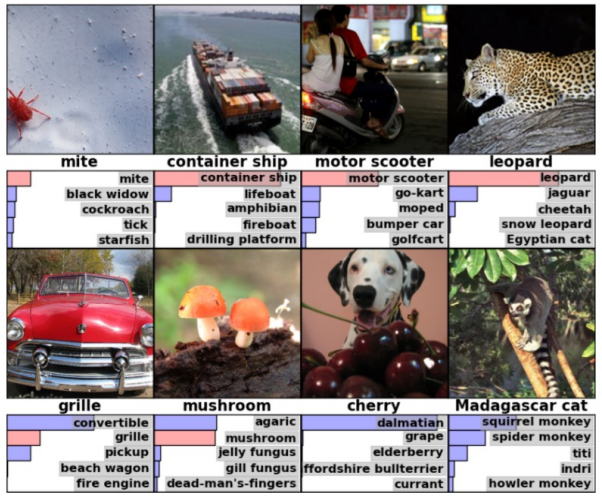
EN.601.475/675 Machine Learning

The slides are partially credited to:
Alex Slivkins, Jian Li, Katerina Fragkiadaki, Lilian Besson

Intro to the problem space

- Sequential decision making under uncertainty

From **prediction** (supervised learning) …



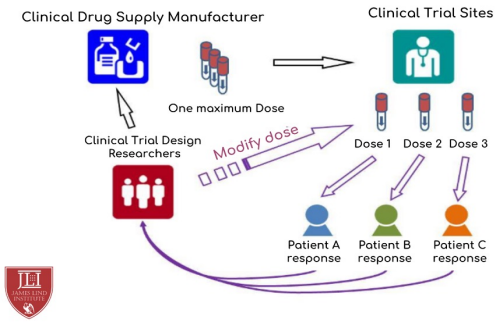
Labels?

Supervised learning:

- Labeled data from a ground truth supervisor
- **Labels**: A correct action to take for a particular input
- **Goal**: To predict the right label in unseen inputs

… to sequential decision making

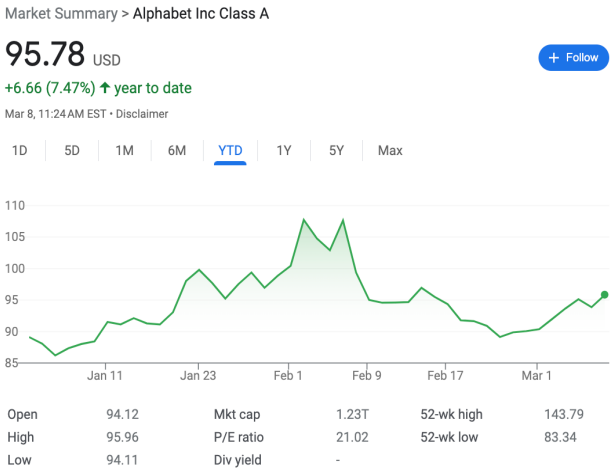
- We often don't know which action is correct to take in a certain situation
  - E.g., Do not have complete info about the effectiveness or side-effects of the drugs. How to infer the best drugs? → run a sequence of trials

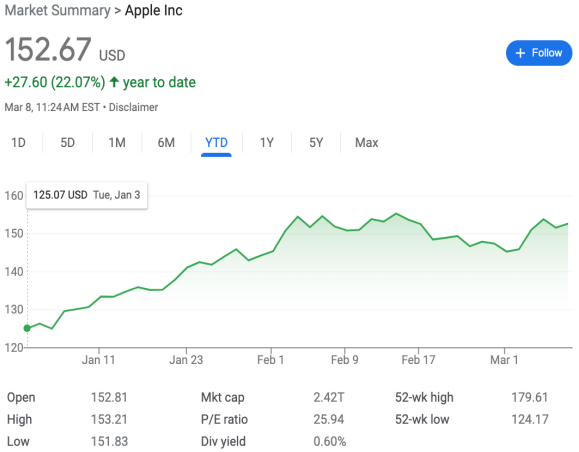


- **Sequential decision making**: Interact with the environment and learn from the consequence of actions
- "*Learning from interaction is a foundational idea underlying nearly all theories of learning and intelligence*" – Richard Sutton

Motivating example #1: Investment

- Each morning, you choose one stock to invest into and invest $1. In the end of the day, you observe the change in value for each stock
- **Question**: how to maximize your wealth?





Motivating example #2: News Site

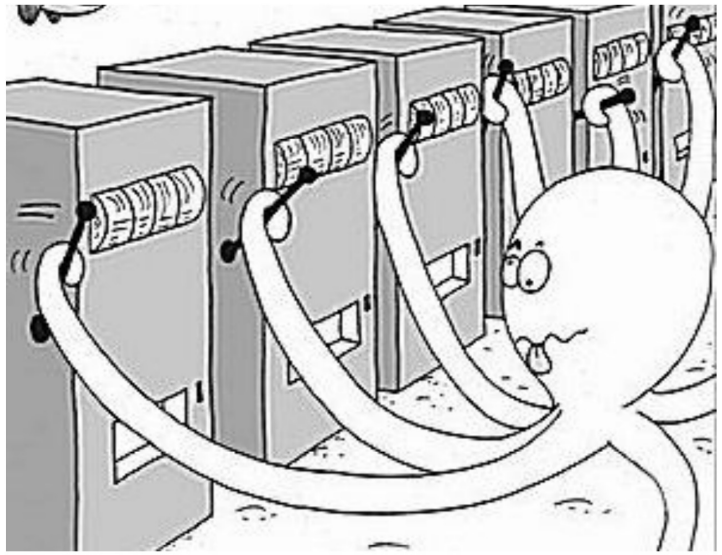
- When a new user arrives, the news site picks a new header to show and observe whether the user clicks on the article or not
- **Question**: How to maximize #clicks?

Motivating example #3: Dynamic pricing

- a pricing strategy that charges customers different prices for the same good or service based on fluctuations in market demand
- A store is selling a digital good (e.g., an app or a song). When a new customer arrives, the store picks a price. Customer buys (or not) and leaves forever
- **Question**: How to maximize total profit?

Multi-armed bandits (MAB)

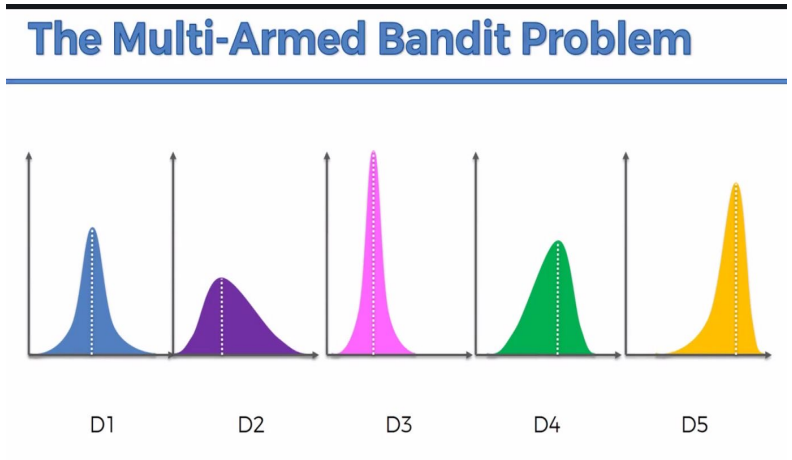
- A basic model to study sequential decision making under uncertainty (i.e., with limited information)
- The earliest reference: Thompson [1933]
- Formally restated in an influential paper Robbins [1952]



source: Microsoft Research

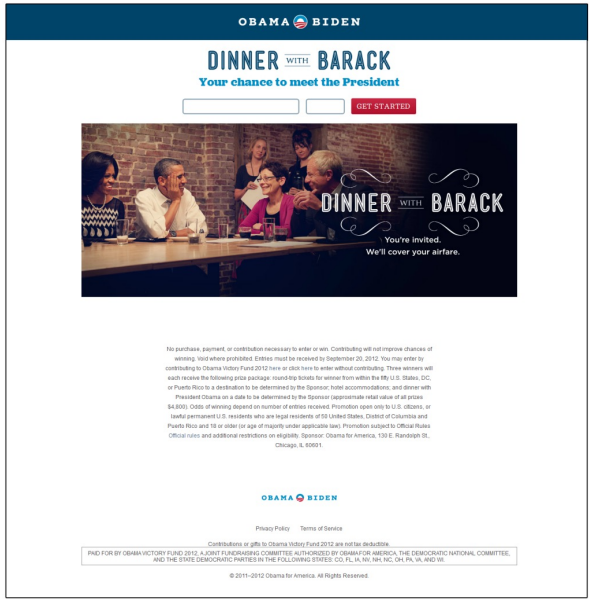
Multi-armed Bandit problem

- A fixed set of k actions ("arms")
- Each arm $i \in \{1, \dots, k\}$ is associated with a reward distribution with mean μ_i
- In each round $t = 1,2, \dots, n$, the learner chooses an arm $a_t \in \{1, \dots, k\}$ and observe reward r_t for the chosen arm
- **Bandit feedback** setting: The rewards for **unchosen arms** are **NOT** observed
- **Goal**: To maximize the total rewards
- Simulation: https://perso.crans.org/besson/phd/MAB_interactive_demo/



Real-world example: A/B testing

- Two arms, each arm corresponds to an image variation to users
- Reward: 1 if the user clicks on the image variation, 0 otherwise
- Mean rewards: the total percentage of users that would click on each variation



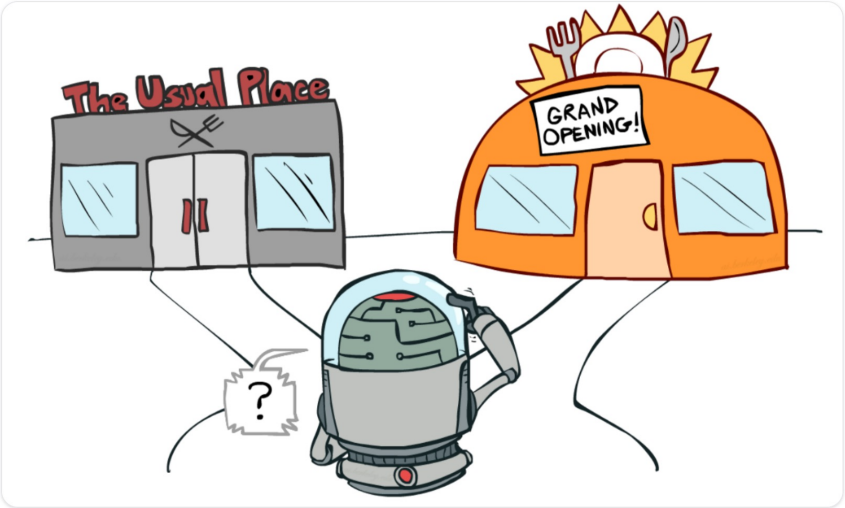
Real-world example: Netflix artwork

For a particular movie, we want to decide which image to show to users

- Arms: one of k candidate images to upload to the Netflix's home screen
- Reward: 1 if the user clicks on the image, 0 otherwise
- Mean rewards: the total percentage of users that would click on each image

Exploration-exploitation trade-off

- Whether acquiring new information ("**exploration**") or making decisions based on available information ("**exploitation**")?
  - **Exploration :** We would like to explore widely so that we would not miss good choices
  - **Exploitation:** We would not want to waste too much resources in exploring bad choices (or try to identify good choices as quickly as possible)
- Fundamental trade-off in decision making with limited information (e.g., in "Reinforcement learning" that we will cover in later lectures)



Source: Berkley AI Course

Why study bandits?

- Bandits are a simplified model for reinforcement learning (RL) → a good intro to RL
- Bandits showcase the fundamental exploration-exploitation trade-off
- Bandits are useful tools for other ML algorithms (e.g., RL, UCRL, bandit-based MCTS, hyperparameter optimization, …)
- Bandits have many great real-world applications (e.g., A/B testing, dynamic pricing, investment, …)
- Bandits have MANY variations to study interesting and foundational problems in learning/decision-making from limited data

Bandit variation #1: Where rewards come from?

- **i.i.d. rewards**: the reward for each arm is drawn independently from a fixed distribution that depends on the arm but not on the round t
- **Adversarial rewards**: Rewards are chosen by an adversary
- **Constrained adversary**: Rewards are chosen by an adversary with known constraints,
  - Reward of each arm can change by at most ε from one round to another
  - Reward of each arm can change at most once
- **Stochastic rewards (beyond iid):** reward of each arm evolves over time as a stochastic process

Bandit variation #2: Contexts

In each round, there might be a context observable before the decision is made

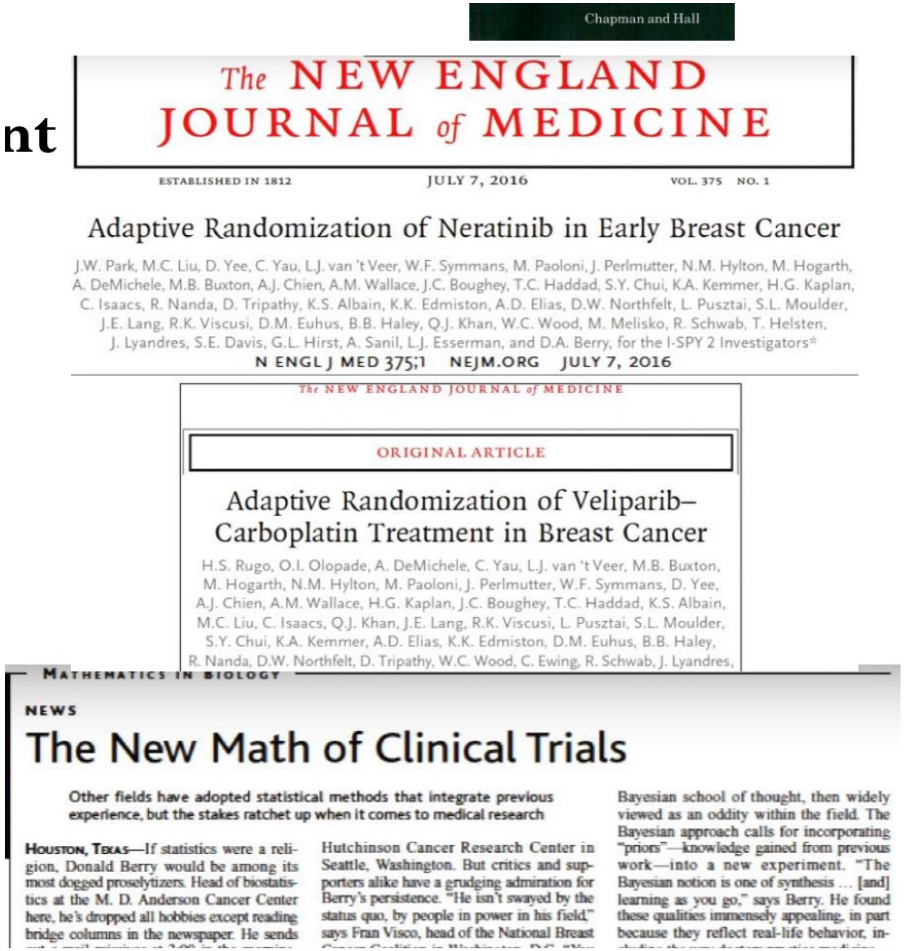
Example	Arms	Rewards	Context
Investment	A stock to invest into	Change in value during the data	Current state of economy
News site	An article to display	1 if clicked, 0 otherwise	User location and demography
Dynamic pricing	A price p	p if sale, 0 otherwise	User location, user's device

Bandit variation #3: Combinatorial bandits

- Multi-armed Bandit setting
- But in each round t, we play **a combinatorial set S of arms** and receive the reward of the set (e.g., $r_S = \max_{a \in S} r_a$)
- **Application**: Online auction
- **Algorithms**: Stochastic dominance confidence bound – for each arm, maintain a CDF which stochastically dominates the true CDF (we won't study this problem in this course)

Bandit variation #4: Top-m arm identification

- **Goal**: Find the top-m arms out of k arms using as few samples as possible
- **Application**: Medical trials, A/B testing, crowdsourcing
  - Clinical trials:
    - One arm – one treatment
    - One pull – one experiment



Chapman and Hall

nt

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JULY 7, 2016 VOL. 375 NO. 1

Adaptive Randomization of Neratinib in Early Breast Cancer

J.W. Park, M.C. Liu, D. Yee, C. Yau, L.J. van 't Veer, W.F. Symmans, M. Paoloni, J. Perlmutter, N.M. Hylton, M. Hogarth, A. DeMichele, M.B. Buxton, A.J. Chien, A.M. Wallace, J.C. Boughey, T.C. Haddad, S.Y. Chui, K.A. Kemmer, H.G. Kaplan, C. Isaacs, R. Nanda, D. Tripathy, K.S. Albain, K.K. Edmiston, A.D. Elias, D.W. Northfelt, L. Pusztai, S.L. Moulder, J.E. Lang, R.K. Viscusi, D.M. Euhus, B.B. Haley, Q.J. Khan, W.C. Wood, M. Melisko, R. Schwab, T. Helsten, J. Lyandres, S.E. Davis, G.L. Hirst, A. Sanil, L.J. Esserman, and D.A. Berry, for the I-SPY 2 Investigators*

N ENGL J MED 375;1 NEJM.ORG JULY 7, 2016

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Adaptive Randomization of Veliparib–Carboplatin Treatment in Breast Cancer

H.S. Rugo, O.I. Olopade, A. DeMichele, C. Yau, L.J. van 't Veer, M.B. Buxton, M. Hogarth, N.M. Hylton, M. Paoloni, J. Perlmutter, W.F. Symmans, D. Yee, A.J. Chien, A.M. Wallace, H.G. Kaplan, J.C. Boughey, T.C. Haddad, K.S. Albain, M.C. Liu, C. Isaacs, Q.J. Khan, J.E. Lang, R.K. Viscusi, L. Pusztai, S.L. Moulder, S.Y. Chui, K.A. Kemmer, A.D. Elias, K.K. Edmiston, D.M. Euhus, B.B. Haley, R. Nanda, D.W. Northfelt, D. Tripathy, W.C. Wood, C. Ewing, R. Schwab, J. Lyandres,

MATHEMATICS IN BIOLOGY

NEWS

The New Math of Clinical Trials

Other fields have adopted statistical methods that integrate previous experience, but the stakes ratchet up when it comes to medical research

HOUSTON, TEXAS—If statistics were a religion, Donald Berry would be among its most dogged proselytizers. Head of biostatistics at the M. D. Anderson Cancer Center here, he's dropped all hobbies except reading bridge columns in the newspaper. He sends

Hutchinson Cancer Research Center in Seattle, Washington. But critics and supporters alike have a grudging admiration for Berry's persistence. "He isn't swayed by the status quo, by people in power in his field," says Fran Visco, head of the National Breast

Bayesian school of thought, then widely viewed as an oddity within the field. The Bayesian approach calls for incorporating "priors"—knowledge gained from previous work—into a new experiment. "The Bayesian notion is one of synthesis … [and] learning as you go," says Berry. He found these qualities immensely appealing, in part because they reflect real-life behavior, in-

Other bandit variations

- **Structural rewards**: Rewards might have a known structure, e.g., rewards are a linear/concave/smooth/Lipschitz function of the chose arm
- **Global constraints**: Limited #items to sell
- **Complex action space**: News site recommend a set of articles, a store prices many products at once
- **Complex outcomes/feedback** (more than just the reward)
  - Dynamic pricing: Which items have been sold?
  - News site: Time spent reading the article?
- **Delayed feedback**: Cannot receive an immediate feedback for a chosen arm up to some delay
- **Low adaptivity**: Algorithm should change the strategy as infrequently as possible as strategy switching is costly

JOHNS HOPKINS
UNIVERSITY

Multi-armed Bandit II

Algorithms

Thanh Nguyen-Tang

The slides are partially credited to: Patrick Rebeschini

Formal set up of Multi-armed Bandits

- Arm set: $\mathcal{A} = \{1,2, \dots, k\}$ and $|\mathcal{A}| = k$
- At every round $t = 1,2, \dots, n$:
  - Learner chooses an arm $a_t \in \mathcal{A} = \{1, \dots, k\}$
  - A data point $z_t = (z_{t,1}, z_{t,2}, \dots, z_{t,k}) \in [0,1]^k$ is sampled independently from an unknown distribution with unknown means $(\mu_1, \dots, \mu_k) \in [0,1]^k$
  - Learner observes reward z_{t,a_t} (but not other rewards $z_{t,a}$ for any $a \neq a_t$) (**bandit feedback**)
- **Goal**: Minimize the **pseudo-regret** R_n defined as

$$R_n := n\,\mu_{a^*} - \sum_{t=1}^{n} \mu_{a_t}$$

$$a_* \in \operatorname*{argmax}_{a \in \mathcal{A}} \mu_a$$

- a_t is a function of $a_1, \dots, a_{t-1}$ and $z_{1,a_1}, z_{2,a_2}, \dots, z_{t-1,a_{t-1}}$
- Note: Learning occurs when algorithm achieves sub-linear growth in n, i.e. $\frac{\mathbb{E}R_n}{n} \to 0$