Reinforcement Learning and Multi-Armed Bandits

Professor Kartik Hosanagar

Reinforcement Learning

- Problem: Choose from multiple actions (e.g. website designs) but there is no historical data available
- Idea: Let algorithms learn by testing various actions/strategies and observing rewards from these actions
  - Powerful method for simulations, gaming, and robotics- based applications
  - Sequential setup in which agent's actions affect the subsequent data it receives (and the reward signal)
  - Learn from reward signal and adapt the strategy over time
  - Involves experimentation to compensate for the lack of supervisory signal (i.e. lack of training data)

Figure from: https://www.davidsilver.uk/wp-content/uploads/2020/03/intro_RL.pdf

Exploration and Exploitation

- In many applications experimentation isn't free!
  - You often want to balance both exploration and exploitation
- *Exploration* gather more information about the environment
- *Exploitation* make the best decision with given information

- Restaurant Selection
  - Exploitation Go to your favourite restaurant
  - Exploration Try a new restaurant
- Online Banner Advertisements
  - Exploitation Show the most successful advert
  - Exploration Show a different advert
- Oil Drilling
  - Exploitation Drill at the best known location
  - Exploration Drill at a new location
- Game Playing
  - Exploitation Play the move you believe is best
  - Exploration Play an experimental move

- How to *balance* between 'exploration' and 'exploitation'?
  - We will discuss this tradeoff using Multi-armed bandit algorithms, a classical Reinforcement Learning approach

Wharton
University of Pennsylvania

Content/quotes from: https://www.davidsilver.uk/wp-content/uploads/2020/03/intro_RL.pdf

Multi-Armed Bandit

- Multi-Armed Bandit = Multiple Slot Machines
- Objective: Maximize cumulative reward in a casino through ‘exploration’ and ‘exploitation’

Content/quotes from: https://www.amazon.com/RecZone-Jumbo-Slot-Machine-Bank/dp/B000H0KBDY

MAB Examples

- **Clinical Trials**
  - Arms = possible treatments (drug molecule)
  - Arm Pulls = application of drug to individual
  - Rewards = outcome of treatment
  - Objective = maximize benefit to trial population (or find best treatment quickly)

  Many ML applications in drug discovery: lots of funding and early successes

- **Online Marketing**
  - Arms = different ads for a web page
  - Arm Pulls = displaying an ad upon a page access
  - Rewards = click through
  - Objective = maximize clicks

Wharton
University of Pennsylvania

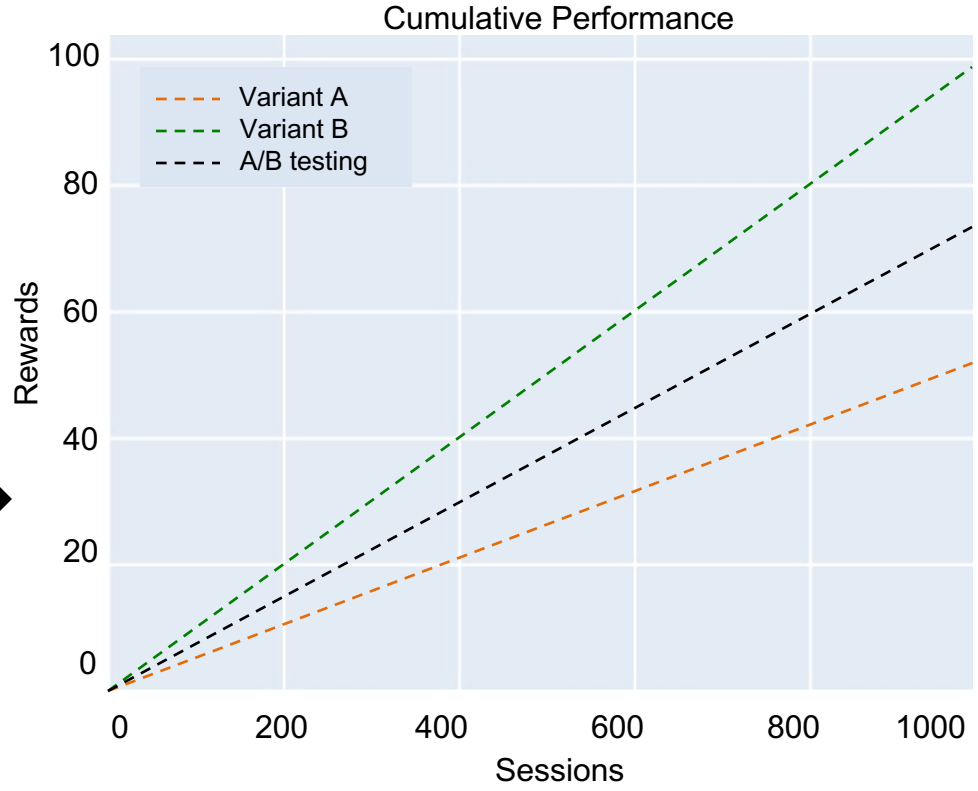
Content/quotes from: https://www.slideserve.com/tyson/monte-carlo-planning-introduction-and-bandit-basics

Motivating Example in Online Marketing

- Suppose you have two ad copies and you don't know which will attract more clicks (and therefore visitors to your website).
- Traditional A/B testing involves showing ad A 50% of the time & ad B 50% of the time, and then assessing which ad performed better.

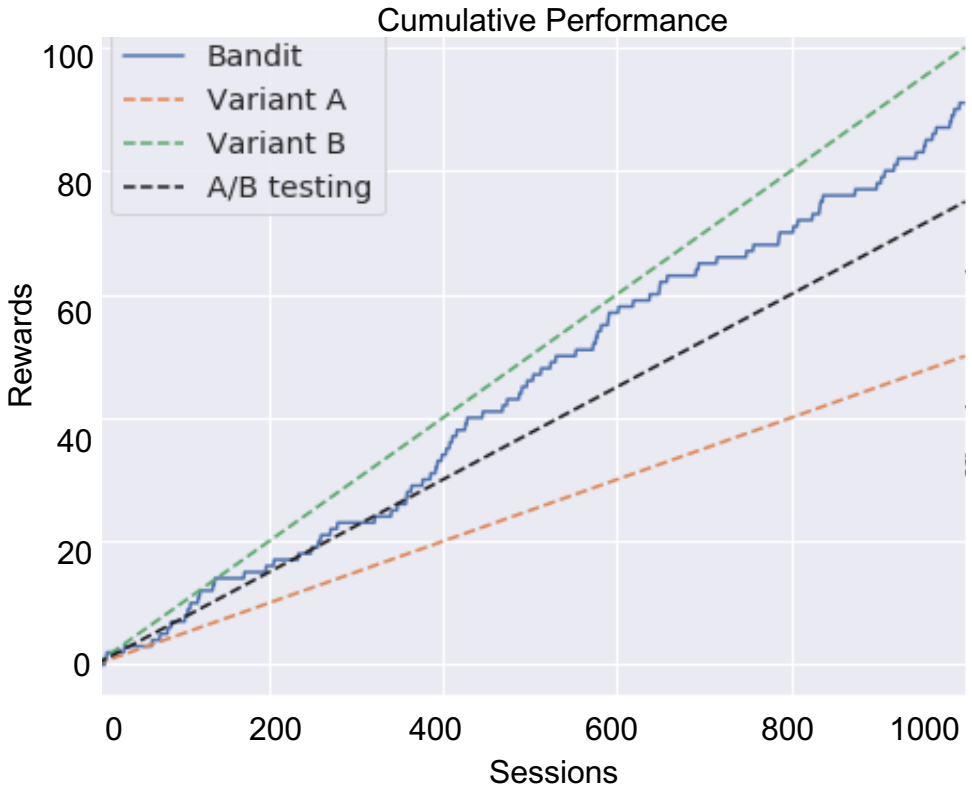
Ad A Ad B

B performs better than A. The A/B test line is midway between the A & B lines as it is 50% A and 50% B.



Motivating Example in Online Marketing

- Machine learning can improve upon A/B testing through Bandit algorithms.
- Bandit algorithms update beliefs based upon performance.
  - They spend more time on best performers early on while still learning and improving over time.
    - The bandit begins by showing 50% A & 50% B, but slowly starts allocating more & more traffic to the higher-performing ad as it learns & confirms which one is better.



MABs in website optimization

RL & Bandit Algorithms for Website Optimization

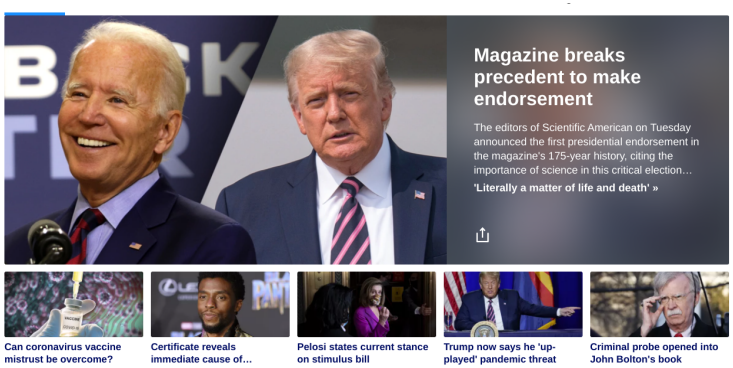
- Bandits are particularly powerful online given the ability for continuous & real-time data collection and large action spaces
- Some examples below:

Website	"Action" space (i.e., interventions available to website)	Objective (i.e., outcome of interest)

Wharton
University of Pennsylvania

RL & Bandit Algorithms for Website Optimization

- Bandits are particularly powerful online given the ability for continuous & real-time data collection and large action spaces
- Some examples below:

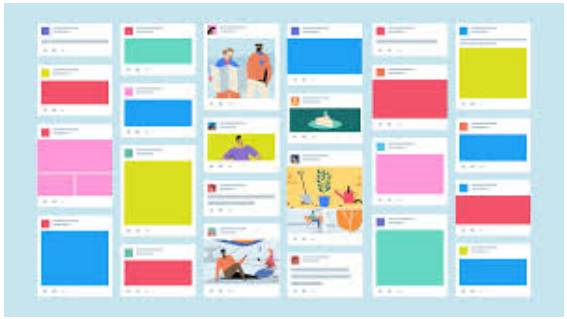
Website	"Action" space (i.e., interventions available to website)	Objective (i.e., outcome of interest)
yahoo! news		Ad revenue Time on site

Of the 100 new articles today, which to recommend at top of homepage?

Wharton University of Pennsylvania

RL & Bandit Algorithms for Website Optimization

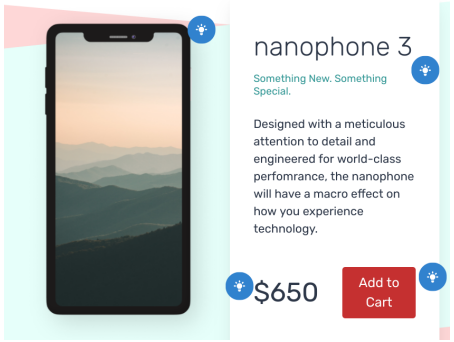
- Bandits are particularly powerful online given the ability for continuous & real-time data collection and large action spaces
- Some examples below:

Website	"Action" space (i.e., interventions available to website)	Objective (i.e., outcome of interest)
facebook	 Of the 1000 new posts today, what mix of news, updates, photos should be first in your feed?	Ad revenue User retention User engagement

Wharton
University of Pennsylvania

RL & Bandit Algorithms for Website Optimization

- Bandits are particularly powerful online given the ability for continuous & real-time data collection and large action spaces
- Some examples below:

Website	"Action" space (i.e., interventions available to website)	Objective (i.e., outcome of interest)
nanophone (generic e-commerce retailer)	 Design of product page	Conversion rate Revenue per session

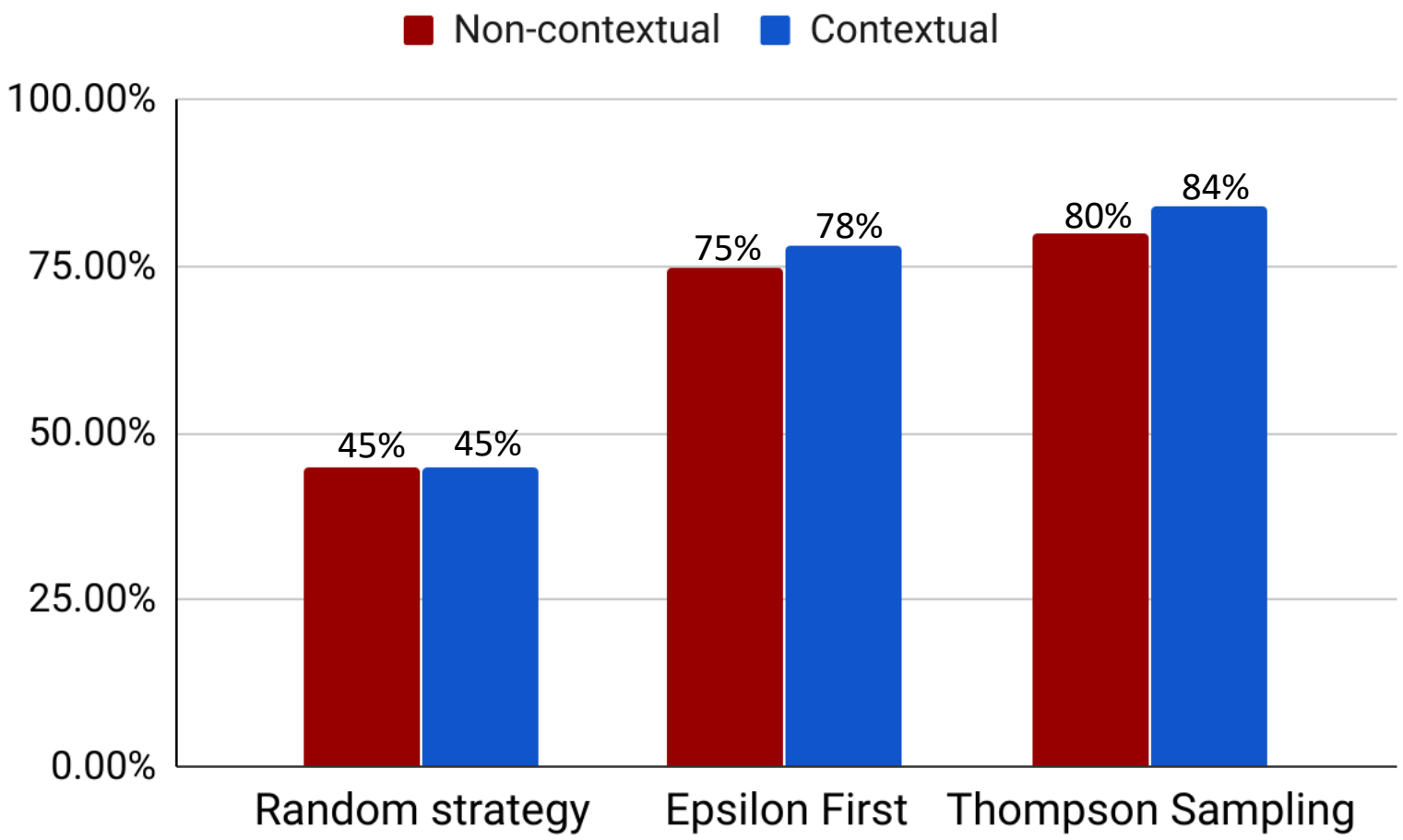
Optimizing Nanophone with bandit algorithms

- We had bandit algorithms play exact simulation game you have all been playing and repeated the 12-week simulation 25 times
  - **Random strategy**: continuous experimentation on all elements
  - **Epsilon First**: Experiment on all subgroups independently for first 2 weeks; for remaining 10 weeks, allocate 100% of traffic for each subgroup to best performing variant in first 2 weeks
  - **Thompson Sampling**: allocate traffic probabilistically to each arm, with higher probabilities given to arms that have a higher chance of being the best
- For the two bandit algorithms, we consider both contextual & non-contextual version

Terminology: "contextual" vs. standard bandits

- An algorithm that treats every user the same is a standard or non-contextual bandit
- An algorithm that can adapt for different types of users is called "contextual"
  - Main idea: users in different geographies, with different cookie histories, device types, etc. may have different preferences
  - Key component of modern online personalization

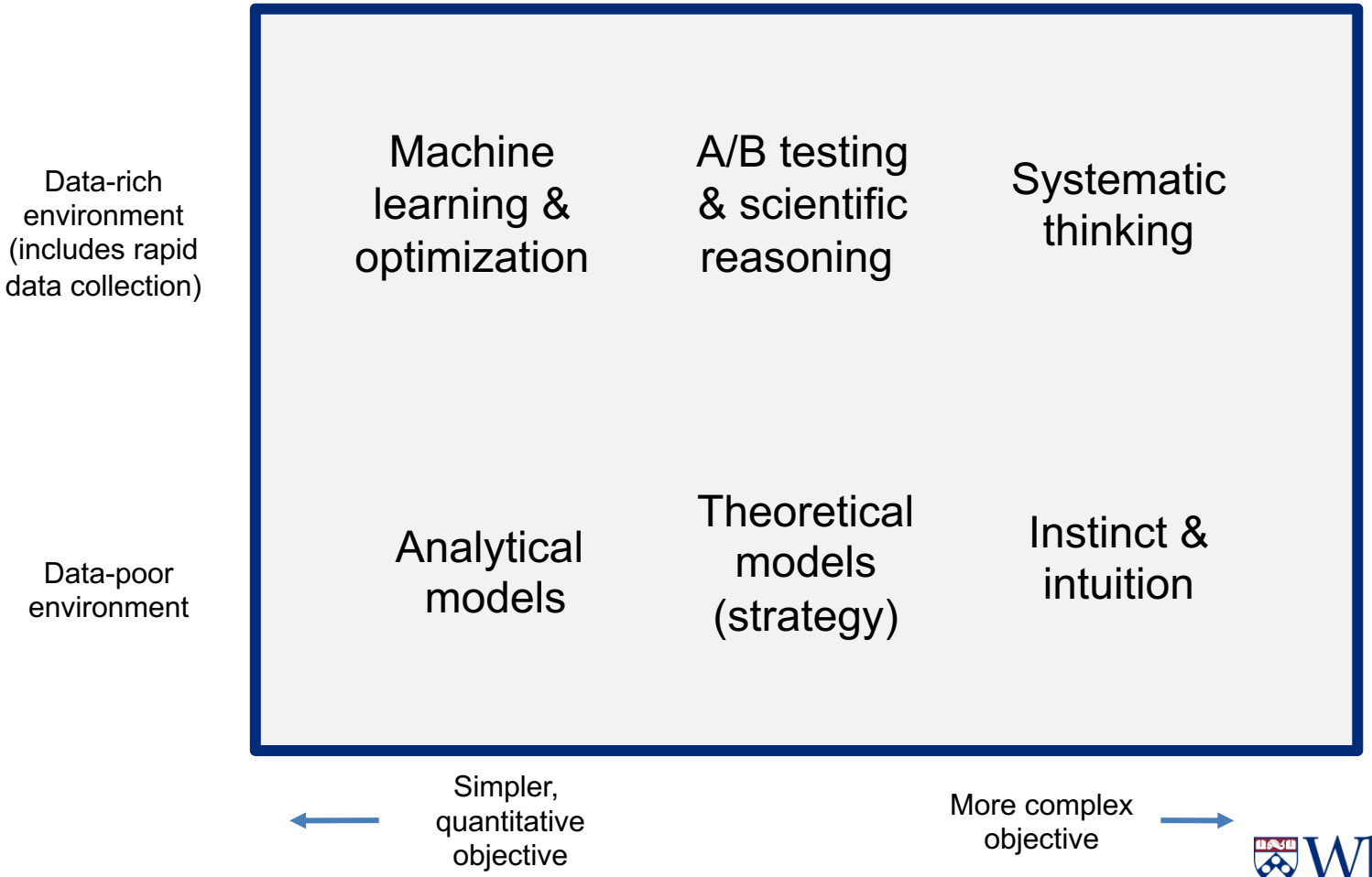
Results of bandit simulations on Nanophone site



Using A/B testing & ML for Decision Making

Paradigms of principled decision making

And the conditions in which they excel



Wharton
University of Pennsylvania

A/B Testing

- Benefits
  - Helps you learn from data
  - Easy to explain decision
  - Discrete point in time when decision gets made
- Drawbacks
  - Exploration may be costly
  - Probably not "optimal", especially in complex environments

How AIs make decisions

- Benefits
  - Arrives at mathematically "optimal" decision
  - Lack of human bias
  - Can handle large decision space (e.g. contextual bandit)
  - Automates the process
- Drawbacks
  - Difficult to explain how/why exactly a choice was made
  - Objective function may not always be clear a priori


Wharton
University of Pennsylvania