Exploring Bandit Algorithms for Automatic Content Selection

Dai Shi


CASINO BARCELONA
VIBRA AMB LA NOSTRA
MÚSICA EN VIU!
GROOVE TO OUR
LIVE MUSIC!
CASINO BARCELONA
I want money !

Which one to choose?

Exploration and Exploitation Dilemma

Exploration: Explore an unknown or suboptimal choice to improve the knowledge about the problem.

Exploitation: Exploit the current knowledge to guess the best choice.

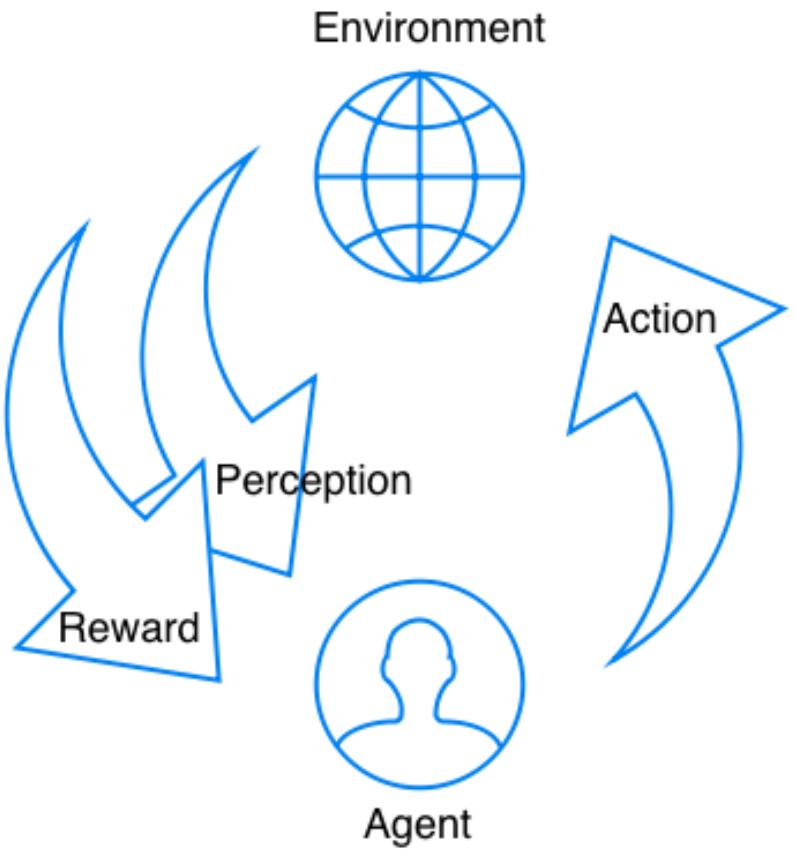
Reinforcement learning

First, the agent gets some perception from environment.

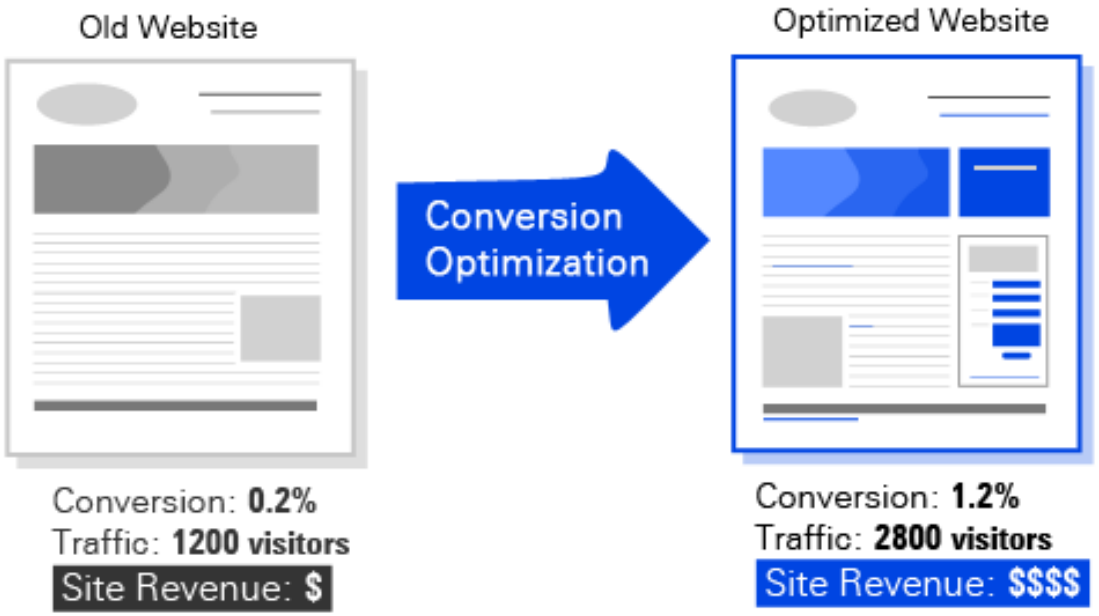
Then it makes an action based on its judgment.

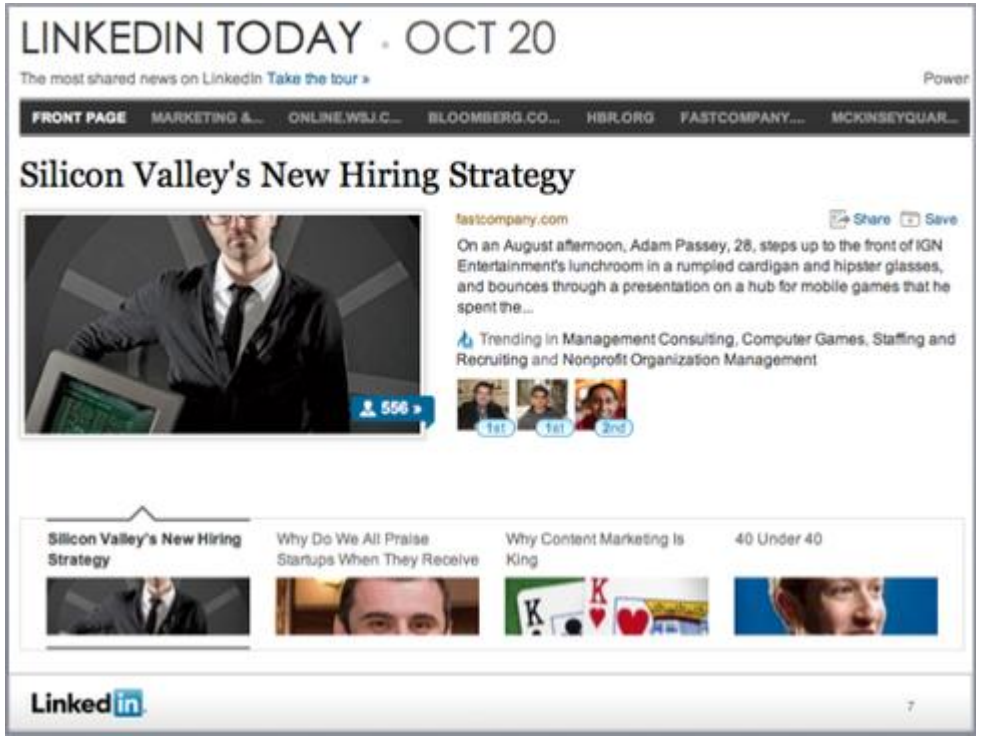
After that the environment would return a reward.

This reward would facilitate the agent making future actions.

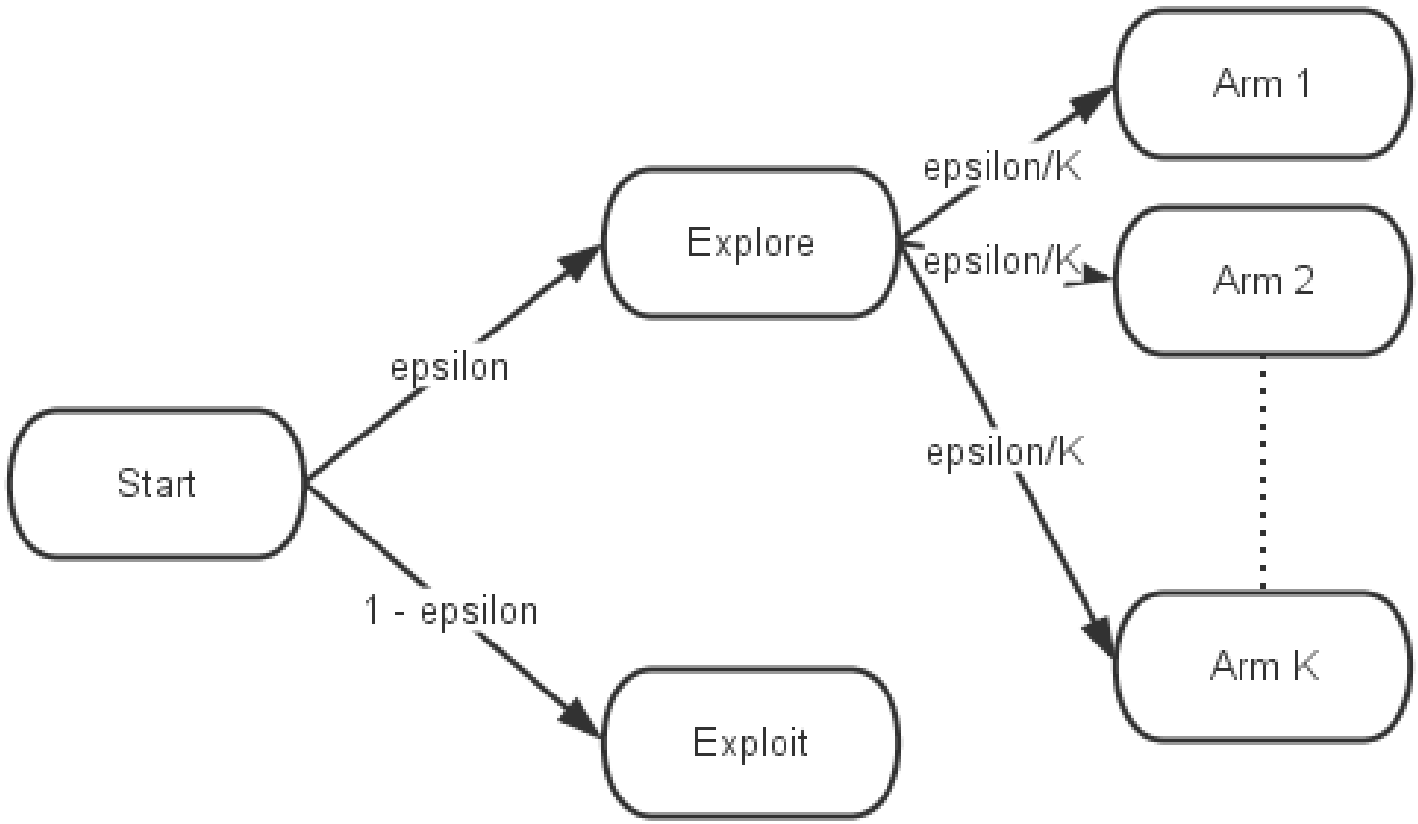


Bandit Algorithms





Epsilon-greedy



Softmax

- There's obvious drawback of epsilon-greedy, which is in the exploration part, it's haphazard. In order to tackle this problem, we can choose arm based on the reward rate.

$$P(A) = \frac{e^{\frac{r_A}{tau}}}{e^{\frac{r_A}{tau}} + e^{\frac{r_B}{tau}}}$$

Boltzmann Distribution

UCB 1

- Upper confidence bounds
- No randomness
- No configurable parameters

Number of successes of arm i

Total number of pulls of all arms

$$\text{Priority}_i = \underbrace{\frac{c_i}{n_i}}_{\text{Observed success rate}} + \underbrace{\sqrt{\frac{2 \cdot \log n}{n_i}}}_{\text{Factor representing uncertainty}}$$

Number of pulls of arm i

Stochastic and Adversarial

- In the stochastic model we assume that the rewards of a given arm is an i.i.d sequence of random variables.

- While, in the adversarial model, there’s no restriction on the sequence of rewards.

Exp3

- Exponential-weight algorithm for Exploration and Exploitation.

$$p_i(t) \leftarrow (1-\gamma)\frac{w_i(t)}{\sum_{j=1}^{K} w_i(t)} + \frac{\gamma}{K}$$

- It contains uniform distribution, which ensures the exploration of all the arms.

Contextual Bandit Problem

Query, IP address, browser properties, etc.

result (ie. ad, news story)

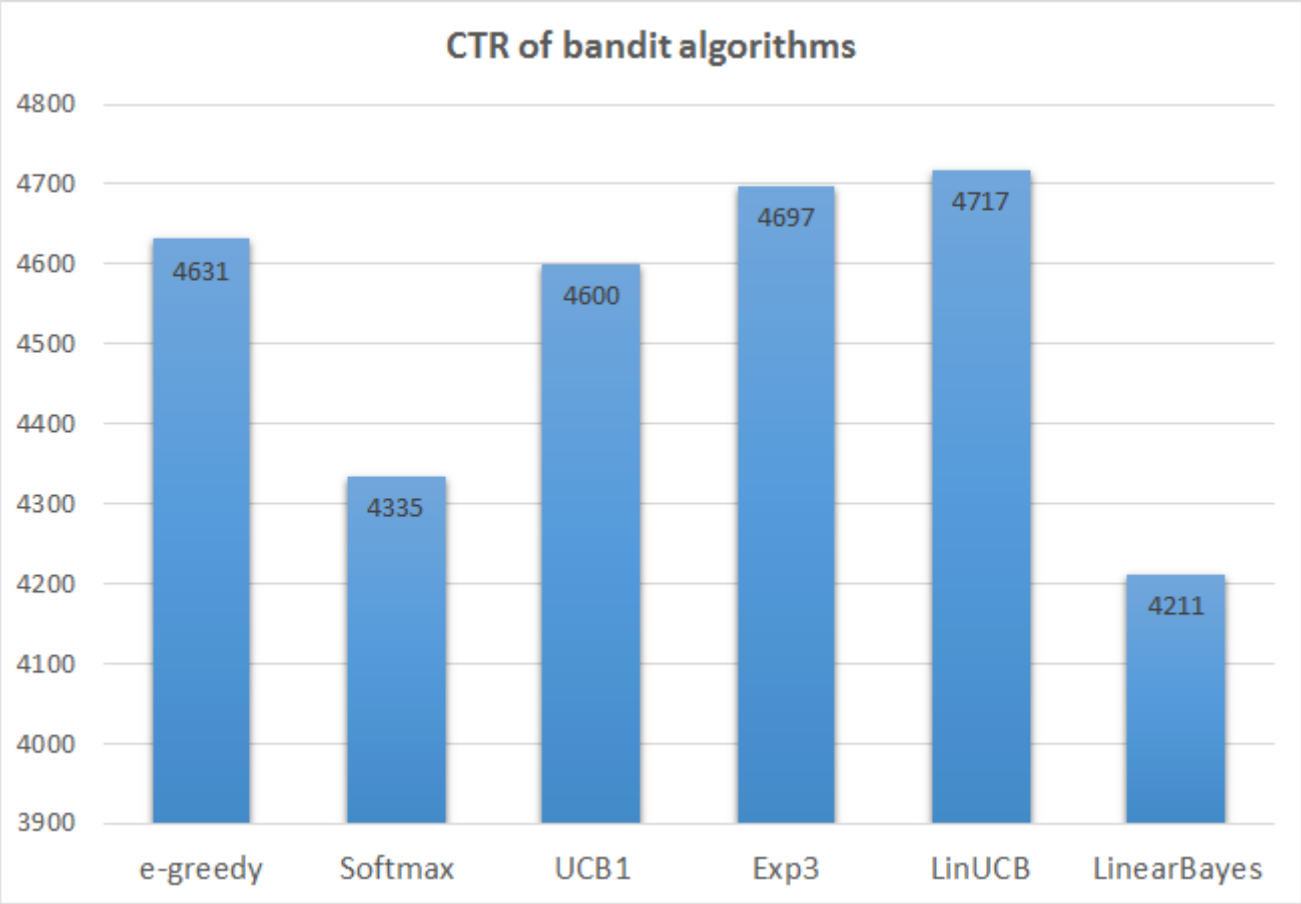
click or not

Source: http://www.levreyzin.com/presentations/CMU_bandits.pdf

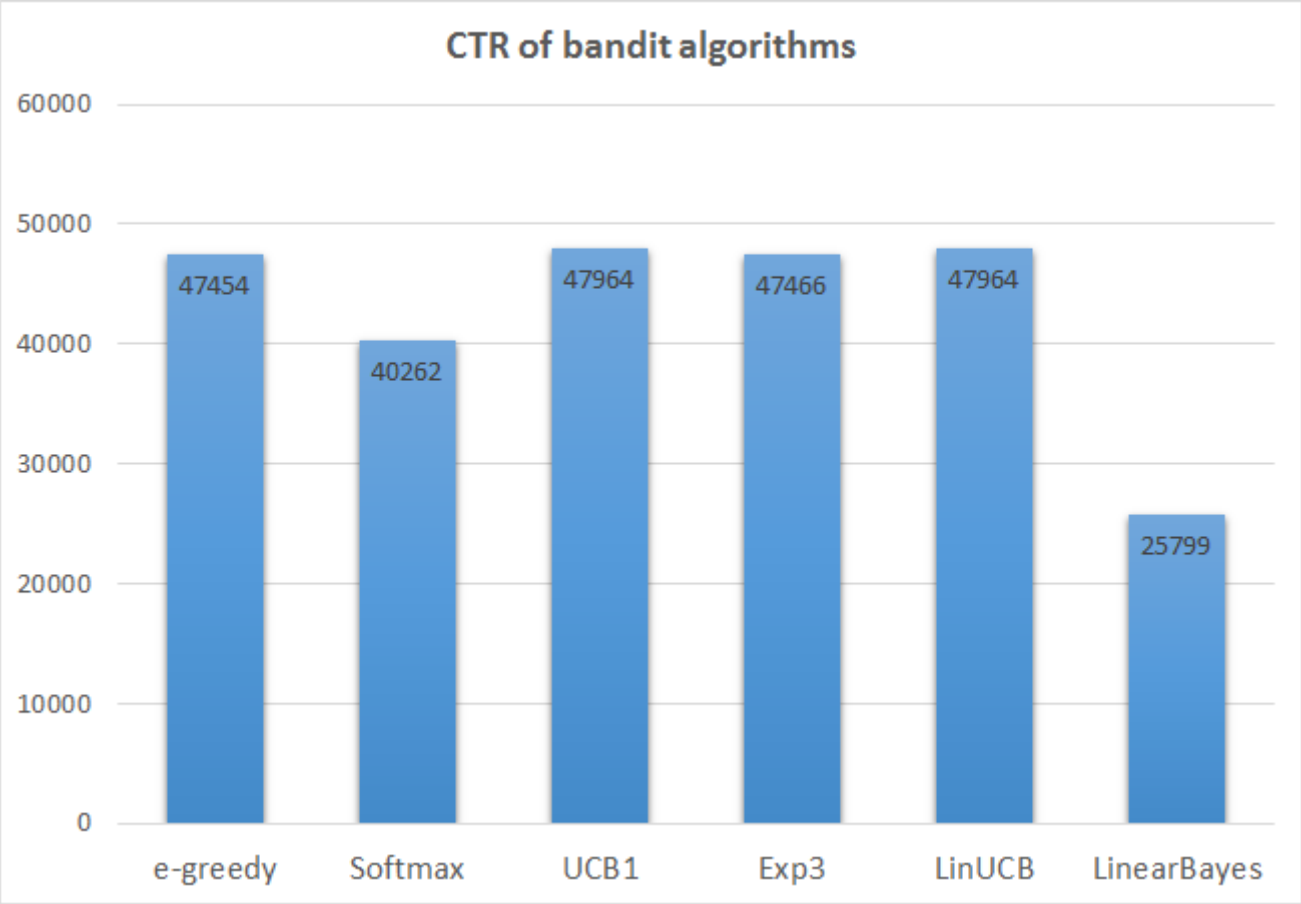
Methods

Algorithms	Description
LinUCB	Expected payoff + confidence interval According to context matrix
LinearBayes	$P(a\|x) = \frac{P(x\|a)P(a)}{P(x)}$ It is the linear sum of all feature probabilities.
Exp4	Exponential weight of each policy, which can be defined by supervised learning

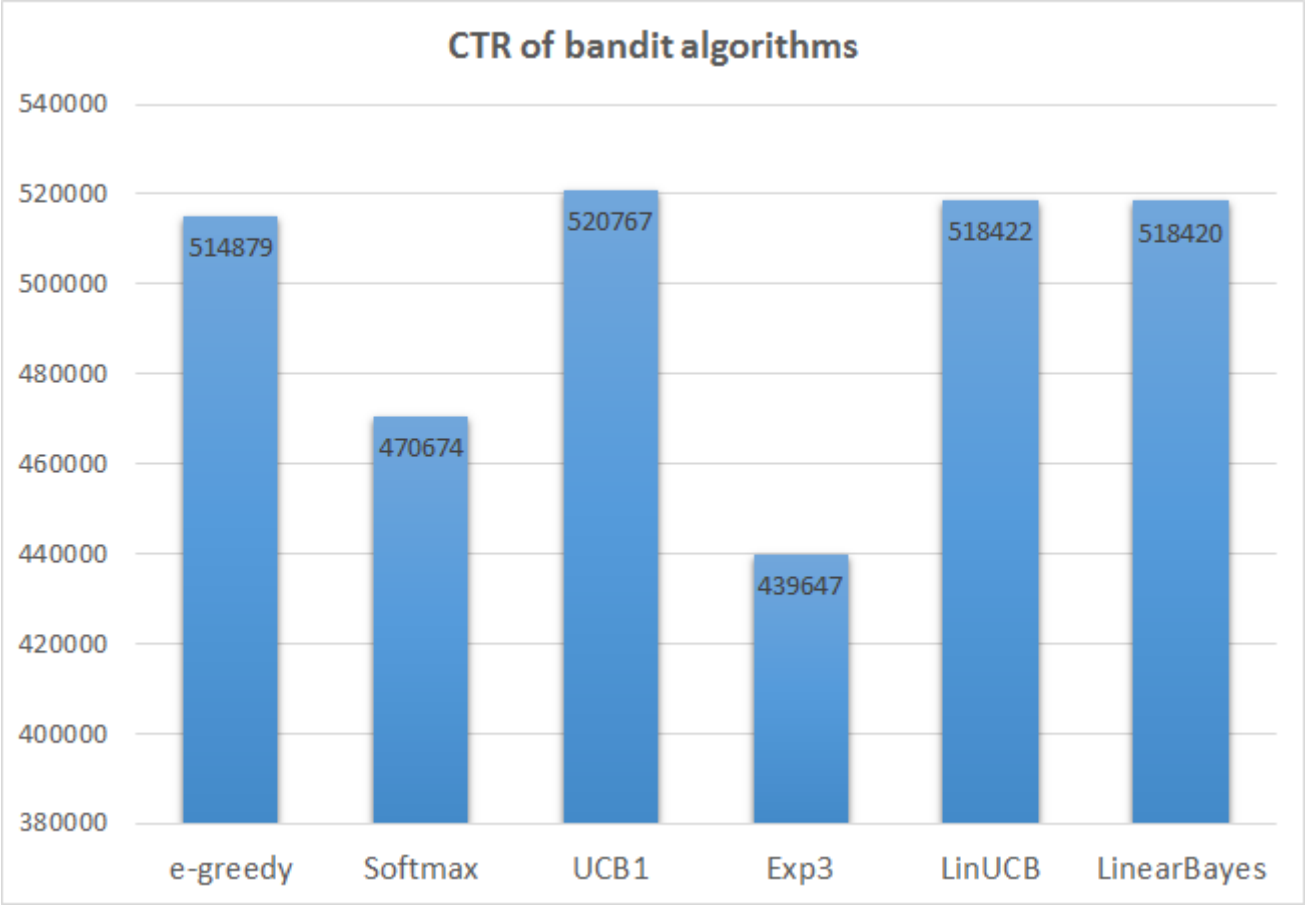
Synthetic: 10000 trials



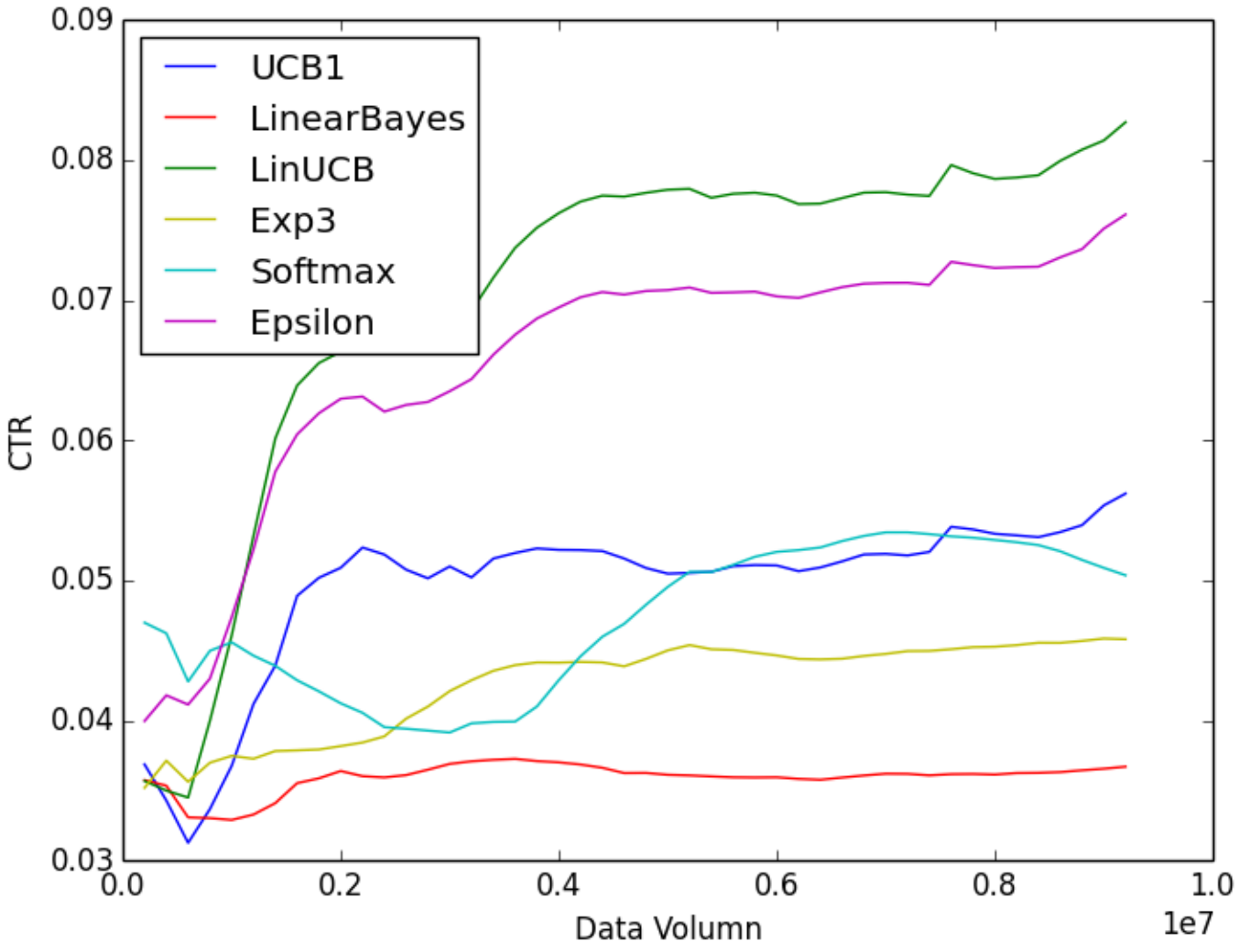
Synthetic: 100000 trials



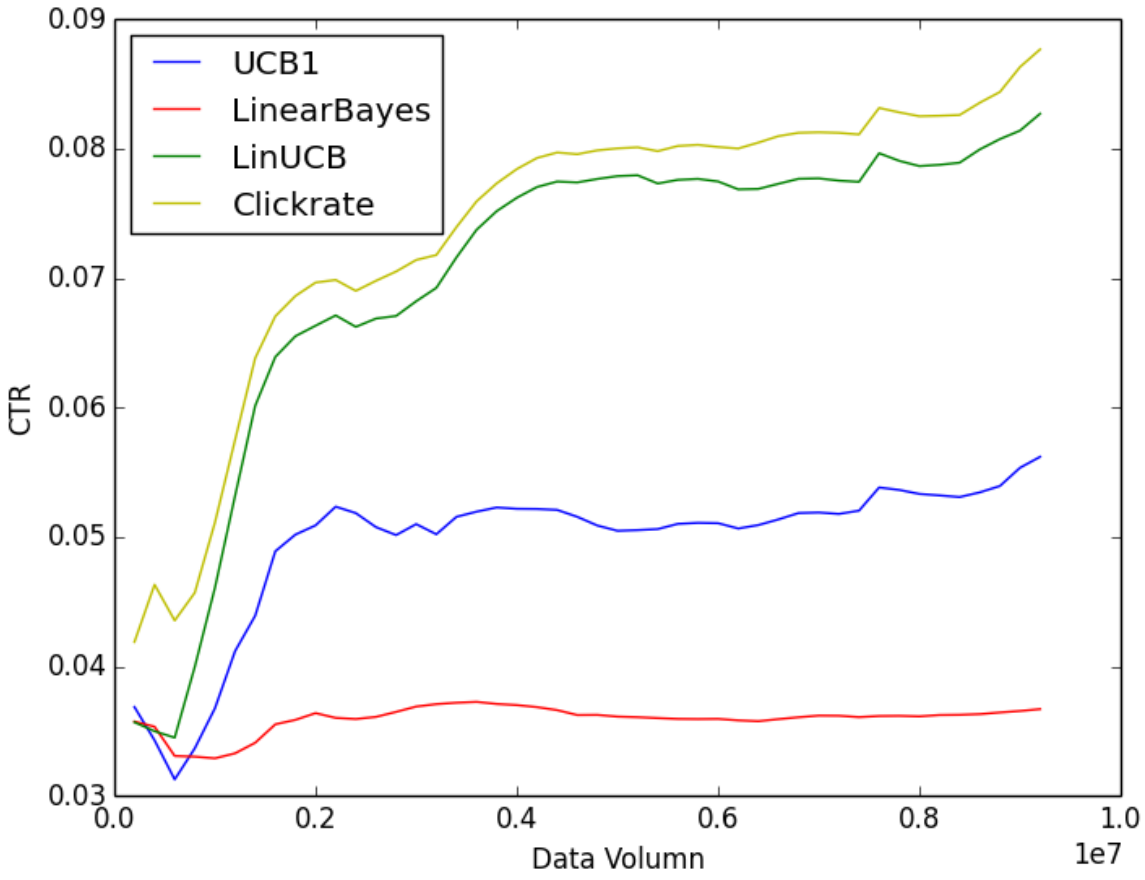
Synthetic: 1000000 trials

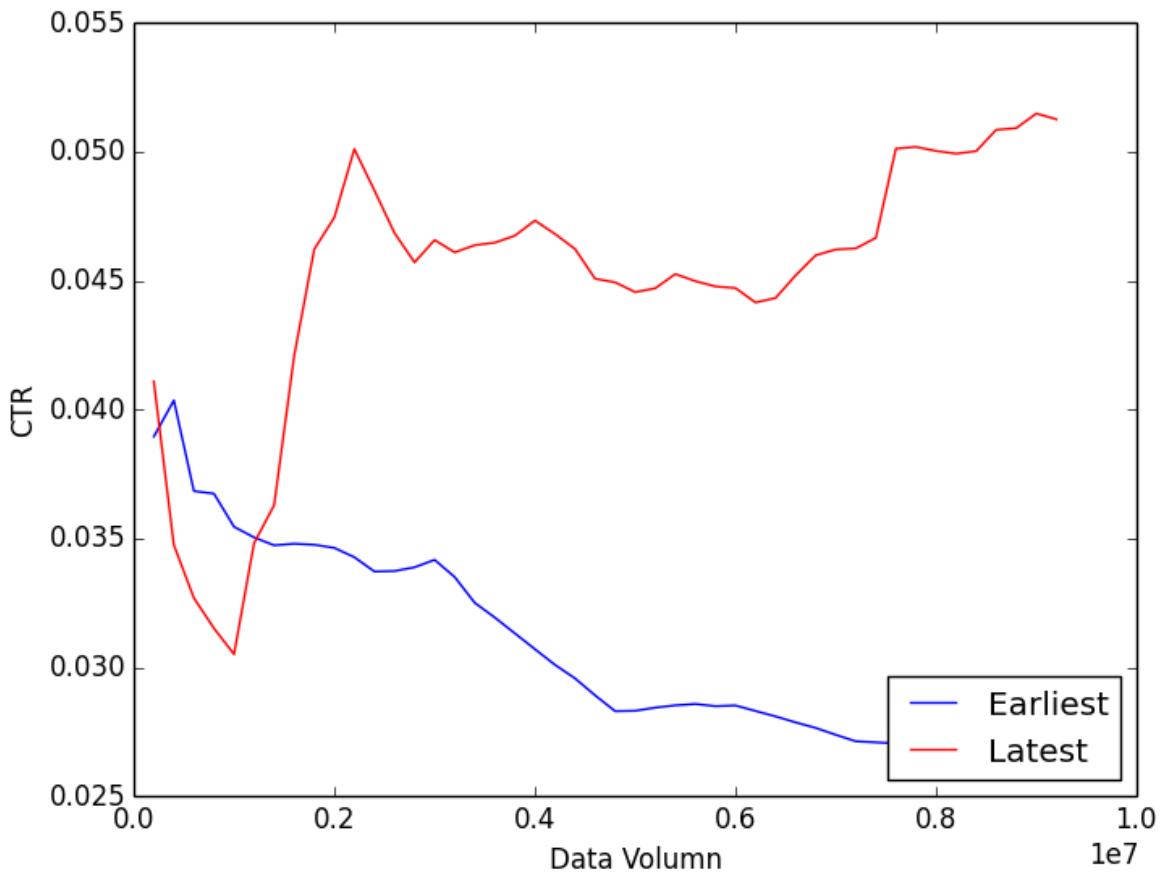


Product Data: Results

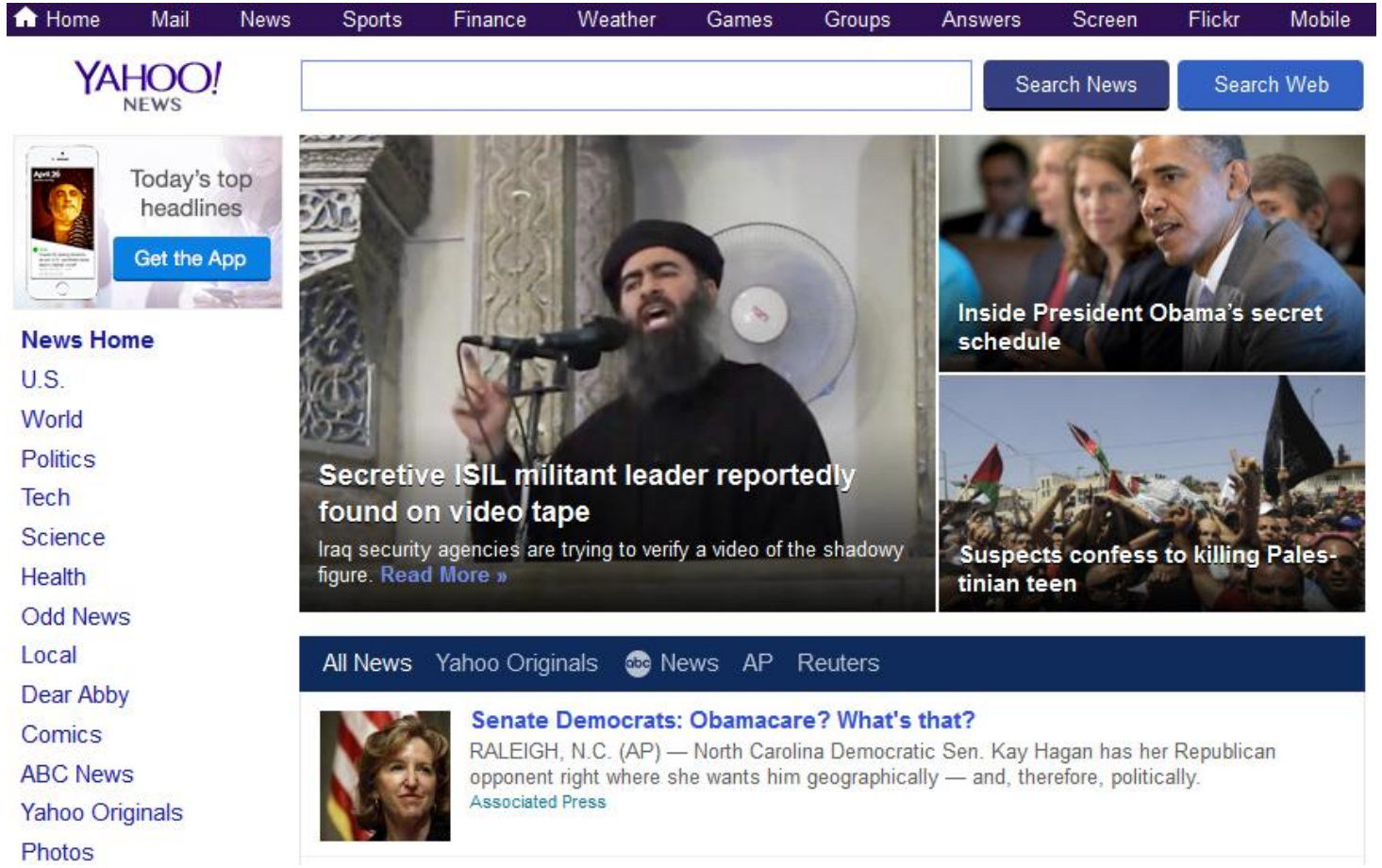


Exploration

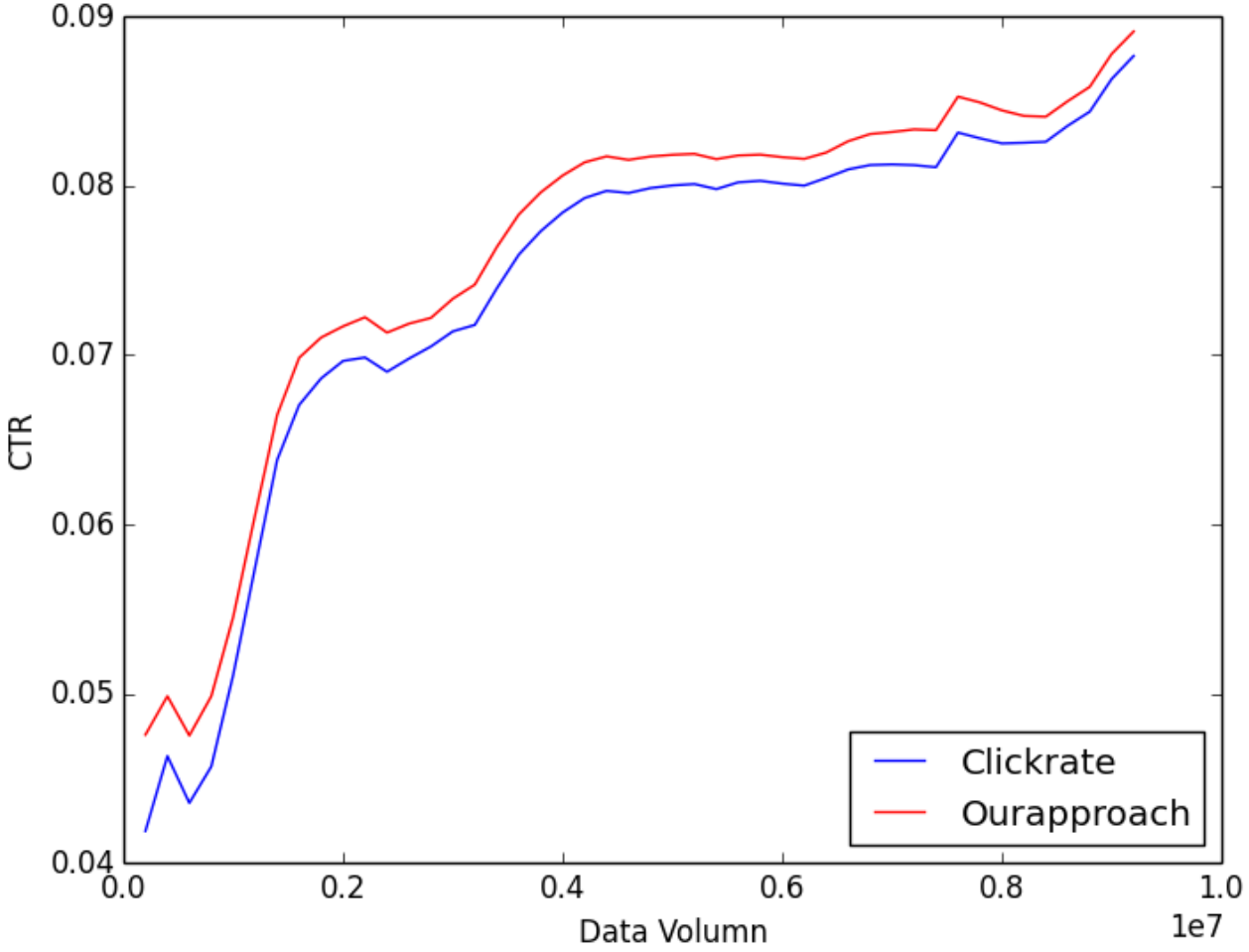




Data Source


Home
Mail
News
Sports
Finance
Weather
Games
Groups
Answers
Screen
Flickr
Mobile
YAHOO! NEWS
Search News
Search Web
Today's top headlines
Get the App
News Home
U.S.
World
Politics
Tech
Science
Health
Odd News
Local
Dear Abby
Comics
ABC News
Yahoo Originals
Photos
Secretive ISIL militant leader reportedly found on video tape
Iraq security agencies are trying to verify a video of the shadowy figure. Read More »
Inside President Obama's secret schedule
Suspects confess to killing Palestinian teen
All News
Yahoo Originals
News
AP
Reuters
Senate Democrats: Obamacare? What's that?
RALEIGH, N.C. (AP) — North Carolina Democratic Sen. Kay Hagan has her Republican opponent right where she wants him geographically — and, therefore, politically.
Associated Press

Exploitation



General Approach

- 1. Collect unbiased data for testing.
- 2. Set up tailored solution, referenced to bandit algorithms.
- 3. Test and implement it on real product with comparison test.
- 4. Full implementation.

Summary

- Gap between theory and practice
- Can be really useful in many applications
- Need to be customized case by case
- Help us to make decisions in real life