

Multi-armed Bandits on the Web

Successes, Lessons and Challenges

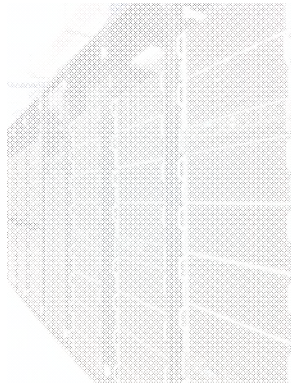
Lihong Li

Microsoft Research

08/24/2014

2nd Workshop on User Engagement Optimization (KDD'14)

BIG DATA



correlation



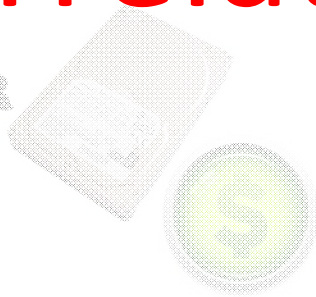
KNOWLEDGE



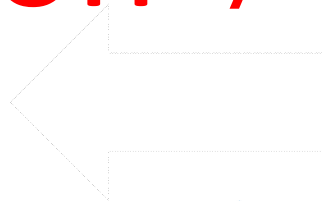
Big Trap

Correlation \neq Causation

BIGGER DATA



UTILITY



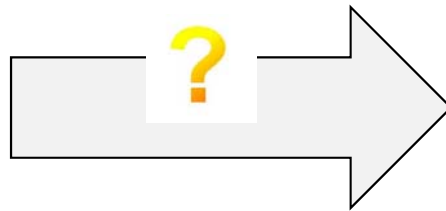
causation



ACTION

Somewhat Toy-ish Example

- Studies show... people who search their names in search engines tend to have higher income
- Decision making:

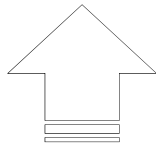


WWII Example

- Statistics collected during WWII...
 - Bullet holes on bomber planes that came back from mission
- Decision making:
 - Where to armor?
 - Abraham Wald: the opposite!



BIG DATA

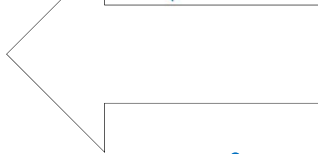


BIGGER DATA



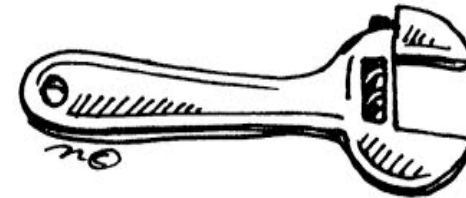
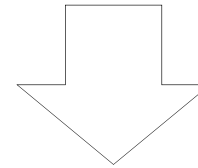
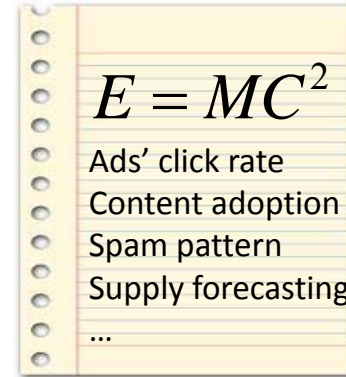
UTILITY

correlation



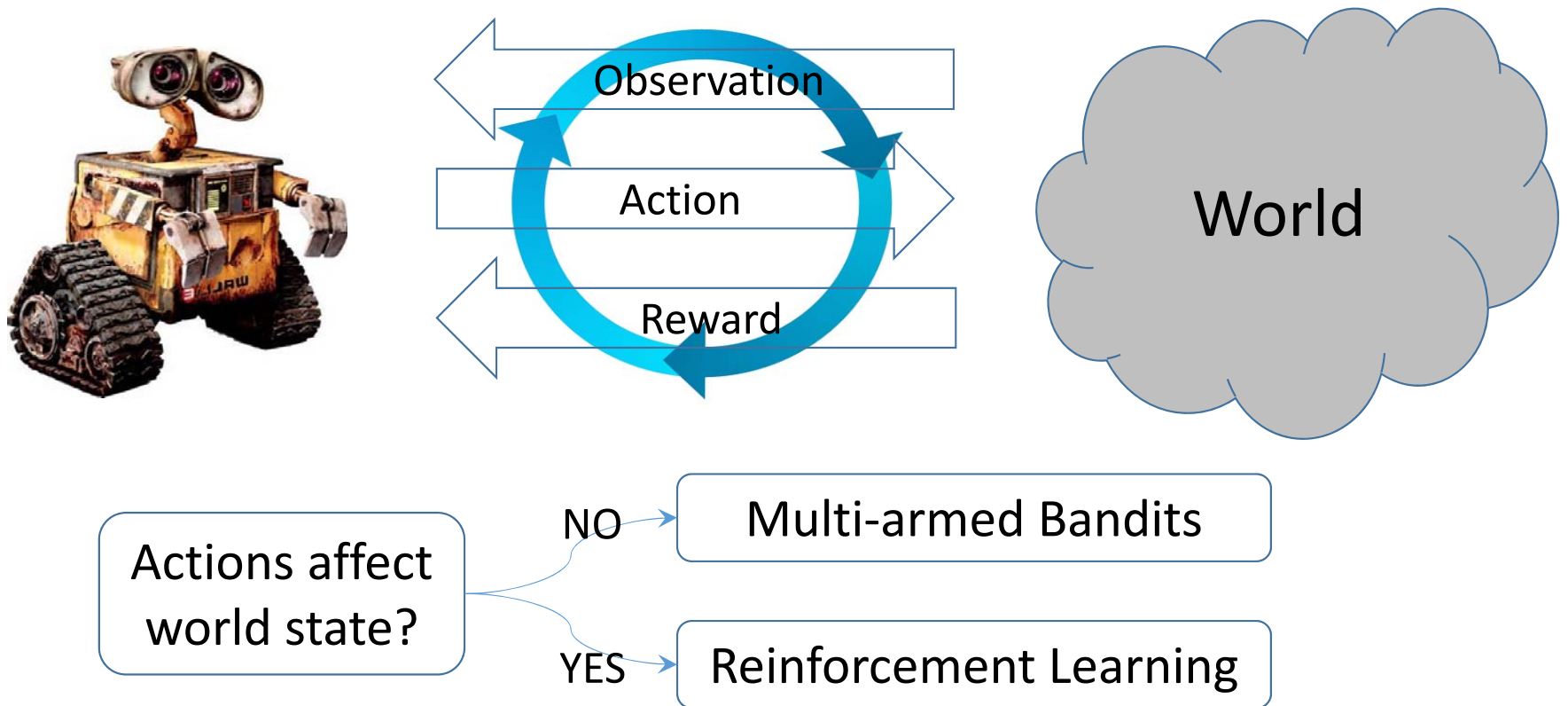
causation

KNOWLEDGE



ACTION

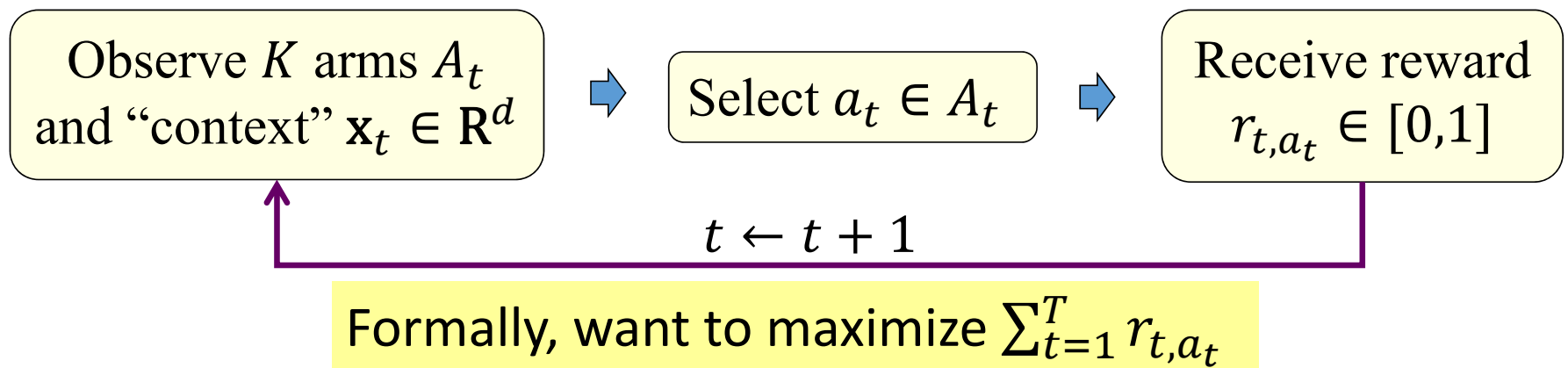
Machine Learning for Decision Making



Outline

- Multi-armed Bandits Algorithms
- Offline Evaluation
- Concluding Remarks

Contextual Bandit [Barto & co'85, Langford & co'08]



Generalizes classic K-armed bandits (without context)

Stochastic vs. adversarial

Motivating Applications

- Clinical trials
- Resource allocation
- Queuing & scheduling
- ...
- Web (more recently)
 - Recommendation
 - Advertising
 - Search

Case 1: Personalized News Recommendation

www.yahoo.com

TODAY - March 02, 2010



Few drugs developed for super bacteria

Doctors are struggling to fight a lethal bacteria that is "resistant to virtually every antibiotic." [» Where it's found](#)

Acinetobacter baumannii

- Do flu vaccines work?
- H1N1 still worrisome

Few drugs for super bacteria

Awkward end to Olympics

Colleges with best-paid alums

Best computers of 2010

1 - 4 of 32

A_t : available articles at time t

\mathbf{x}_t : user features (age, gender, interests, ...)

a_t : the displayed article at time t

r_{t,a_t} : 1 for click, 0 for no - click

Average reward is click-through rate (CTR)

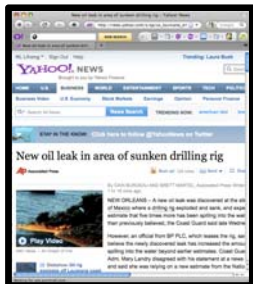
Standard Multi-armed Bandit [R'52, LR'85]

No contextual information is available \rightarrow potentially lower rewards



CTR₁

$$\text{CTR estimate } \hat{\mu}_a = \frac{C_a}{N_a} = \frac{\text{\#clicks}}{\text{\#displays}}$$



CTR₂

ϵ -greedy:

$$\text{Choose article } \begin{cases} \arg \max_a \hat{\mu}_a, & \text{with prob. } 1 - \epsilon \\ \text{random}, & \text{with prob. } \epsilon \end{cases}$$



CTR₃

UCB1 (Upper Confidence Bound) [ACF'02]

$$\text{Choose article } \arg \max_a \left\{ \hat{\mu}_a + \frac{\alpha}{\sqrt{N_a}} \right\}$$

Exploration bonus that decays over time

LinUCB: UCB for Linear Models [LCLS'10]

- Linear model assumption: $\mathbf{E}[r_{t,a} | \mathbf{x}_t] = \mathbf{x}_t^T \theta_a$
- Standard least-squares ridge regression

$$\hat{\theta}_a = \overbrace{(\mathbf{D}_a^T \mathbf{D}_a + \mathbf{I})}^{\mathbf{A}_a}^{-1} \mathbf{D}_a^T \mathbf{c}_a, \text{ where } \mathbf{D}_a = \begin{bmatrix} -\mathbf{x}_{t_1}^T & - \\ -\mathbf{x}_{t_2}^T & - \\ \vdots & - \end{bmatrix} \text{ and } \mathbf{c}_a = \begin{bmatrix} r_{t_1} \\ r_{t_2} \\ \vdots \end{bmatrix}$$

- Quantifying prediction uncertainty: with high probability,

$$\underbrace{|\mathbf{x}^T \hat{\theta}_a - \mathbf{x}^T \theta_a|}_{\text{Prediction error}} \leq \alpha \sqrt{\underbrace{\mathbf{x}^T \mathbf{A}_a^{-1} \mathbf{x}}_{\text{Measures how similar } \mathbf{x} \text{ is to previous contexts}}}$$

LinUCB: Optimism in the Face of Uncertainty

LinUCB chooses $a^* = \operatorname{argmax}_a \left\{ \mathbf{x}^T \hat{\boldsymbol{\theta}}_a + \alpha \sqrt{\mathbf{x}^T \mathbf{A}_a^{-1} \mathbf{x}} \right\}$

to exploit

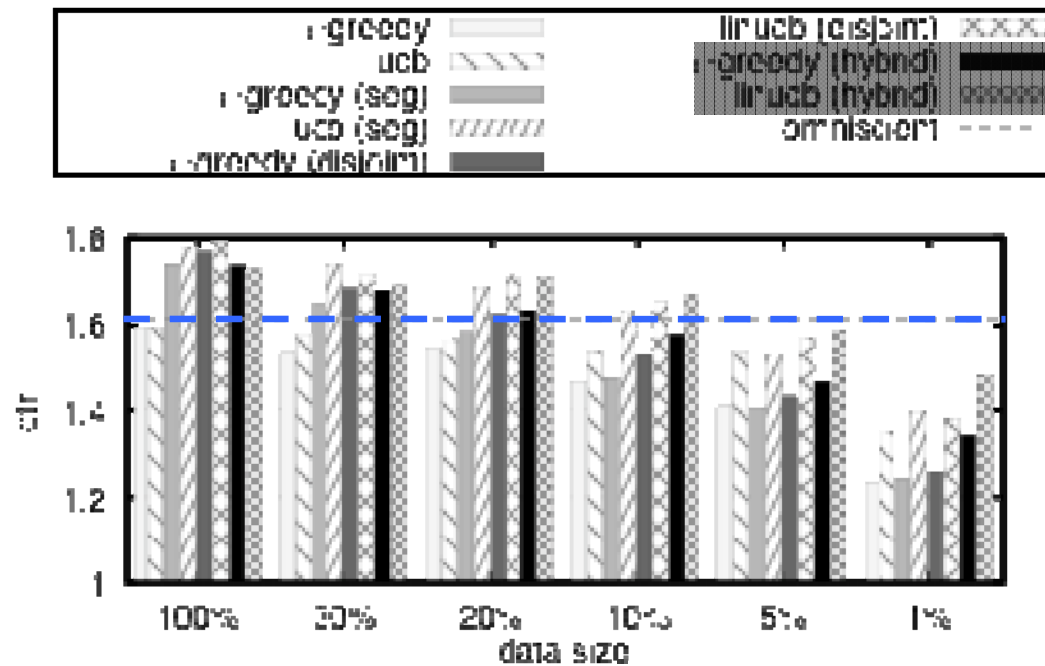
to explore

Recall UCB1: $\operatorname{argmax}_a \left\{ \hat{\mu}_a + \frac{\alpha}{\sqrt{N_a}} \right\}$

A variant of LinUCB: $O(\sqrt{KdT})$ with matching lower bound [CLRS'11]

LinRel [Auer 2002] is similarly motivated but more complicated.

LinUCB for News Recommendation [LCLS'10]



- UCB-type algorithms do better than ϵ -greedy counterparts
- CTR improved significantly when features/contexts are considered

LinUCB Variants

- Hybrid linear models for multi-task learning [LCLS'10]
 - Beneficial when data is sparse
- Generalized linear model [LCLMW'12]
 - Greater flexibility of modeling rewards
 - Linear regression, logistic regression, probit regression, ...
- Sparse linear model [Abbasi-Yadkori & co'12]
 - Lower regret when $\{\theta_a\}$ are sparse

Case 2: Online Advertising

msn news

bing site search

HOME US CRIME & JUSTICE WORLD SCIENCE & TECH POP CULTURE OBITS RUMORS PHOTOS VIDEO

Twitter to add abuse button after

bing quebec city

32,800,000 RESULTS Any time ▾

Ad related to quebec city

[Quebec City | QuebecCity.TripAdvisor.ca](#)
[QuebecCity.TripAdvisor.ca/quebec-city](#)

Research **Quebec City** Hotels and **Quebec City** Attractions!

- Award-Winning Attractions
- Award-Winning Beaches
- Award-Winning Hotels
- Best Bargain Hotels
- Best Places to Eat
- Find a Cheap Flight

[Quebec City and Area : Official Web Site - Québec City ...](#)
[www.quebecregion.com/en ▾](#)

Which **Québec City** Street Are You? **Québec City**'s streets are bursting with personality. Find out which one is your perfect match! Details

[Must-See Attractions](#) · [Old Quebec](#) · [Accommodation](#) · [Tours](#) · [Where to Stay](#)

[Quebec City - Wikipedia, the free encyclopedia](#)

FINDNEWROADS*

2013 CHEVROLET VOLT

FOR A TOTAL OF UP TO 380 MILES ON A FULL CHARGE AND A FULL TANK OF GAS*

*Important Info [Explore Volt](#)

AdChoices [▶](#) Ad Feedback



Context: query, user info, ...
Action: displayed ads
Reward: revenue

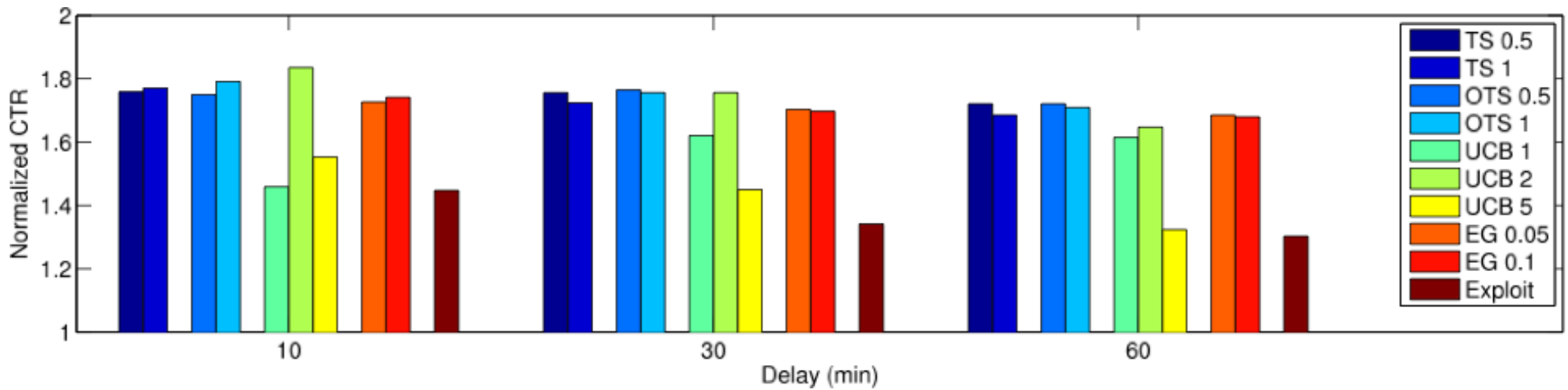
Limitation of UCB in Online Advertising

- How to take advantage of prior information to avoid unnecessary exploration
- How to handle long delay of reward after taking an action
- How to enable complex models

Thompson Sampling

- Old heuristic: “probability matching” (1933)
 - $\Pr(a|\mathbf{x}) = \Pr(a \text{ is optimal for } \mathbf{x} \mid \text{prior, data})$
- Highly effective in practice [Scott’10] [CL’11]
- Inspired lots of theoretical study in last 2 years
 - Non-contextual bandits [Agrawal, Goyal, Kaufmann, ...]
 - Linear bandit [Agrawal & Goyal’13]
 - Generalized Thompson Sampling [L’13]
 - Bayes risk analyses [Russo & Van Roy]

Thompson Sampling for Advertising [CL'12]



Model-agnostic Algorithms

- EXP4 [Auer et al'95] [BLLRS'11]
 - Optimal $O(\sqrt{T})$ regret bound
 - Works even when contexts and rewards are generated by adversarial
 - Computationally expensive in general
- ILOVETOCONBANDITS [AHKLLS'14]
 - Optimal $O(\sqrt{T})$ regret bound
 - Computationally efficient
 - Promising empirical results

Outline

- Multi-armed Bandits Algorithms
- Offline Evaluation
- Concluding Remarks

Policy Evaluation

Assume stochastic bandit: $\mathbf{x} \sim \nu$, $r_a \sim \nu(\cdot | \mathbf{x}, a)$

Given a policy $\pi: \mathbf{x} \rightarrow a$, want to estimate its value: $V(\pi) = \mathbf{E}_\nu[r_{\pi(\mathbf{x})}]$

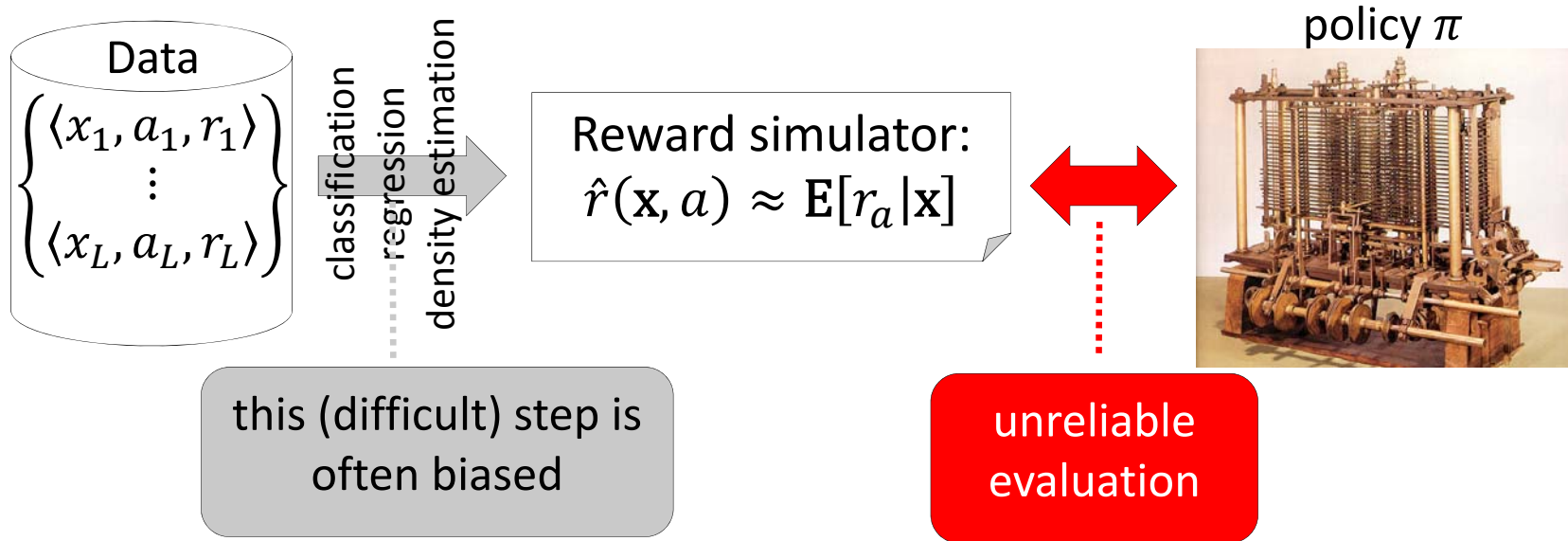
Online evaluation

- Run π on live users and average observed rewards (as in A/B tests)
- **Reliable** but **expensive**

Offline evaluation

- Estimate $V(\pi)$ from historical data set $D = \{(\mathbf{x}, a, r_a)\}$
- **Fast and cheap** (e.g., benchmark data sets for supervised learning)
- **Counterfactuality of rewards**: no information to evaluate π if $\pi(\mathbf{x}) \neq a$

Common Approach in Practice



In contrast, our approach

- avoids explicit user modeling → simple
- gives unbiased evaluation results → reliable

Our Approach: Unbiased Offline Evaluation

Randomized data collection: at step t ,

- Observe current context \mathbf{x}
- Randomly chooses $a \in A$ according to (p_1, p_2, \dots, p_K) and receives r_a

End result: “exploration data” $D = \{(\mathbf{x}, a, p_a, r_a)\}$

Key properties:

- Unbiasedness: $\mathbf{E}_D \left[\frac{1}{|D|} \sum_{(\mathbf{x}, a, p_a, r_a) \in D} \frac{r_a \cdot \mathbf{1}(\pi(\mathbf{x})=a)}{p_a} \right] = V(\pi)$
- Estimation error = $O\left(\frac{1}{\sqrt{|D|}}\right)$

Related to causal inference (Neyman-Rubin) and off-policy learning [Precup & co]

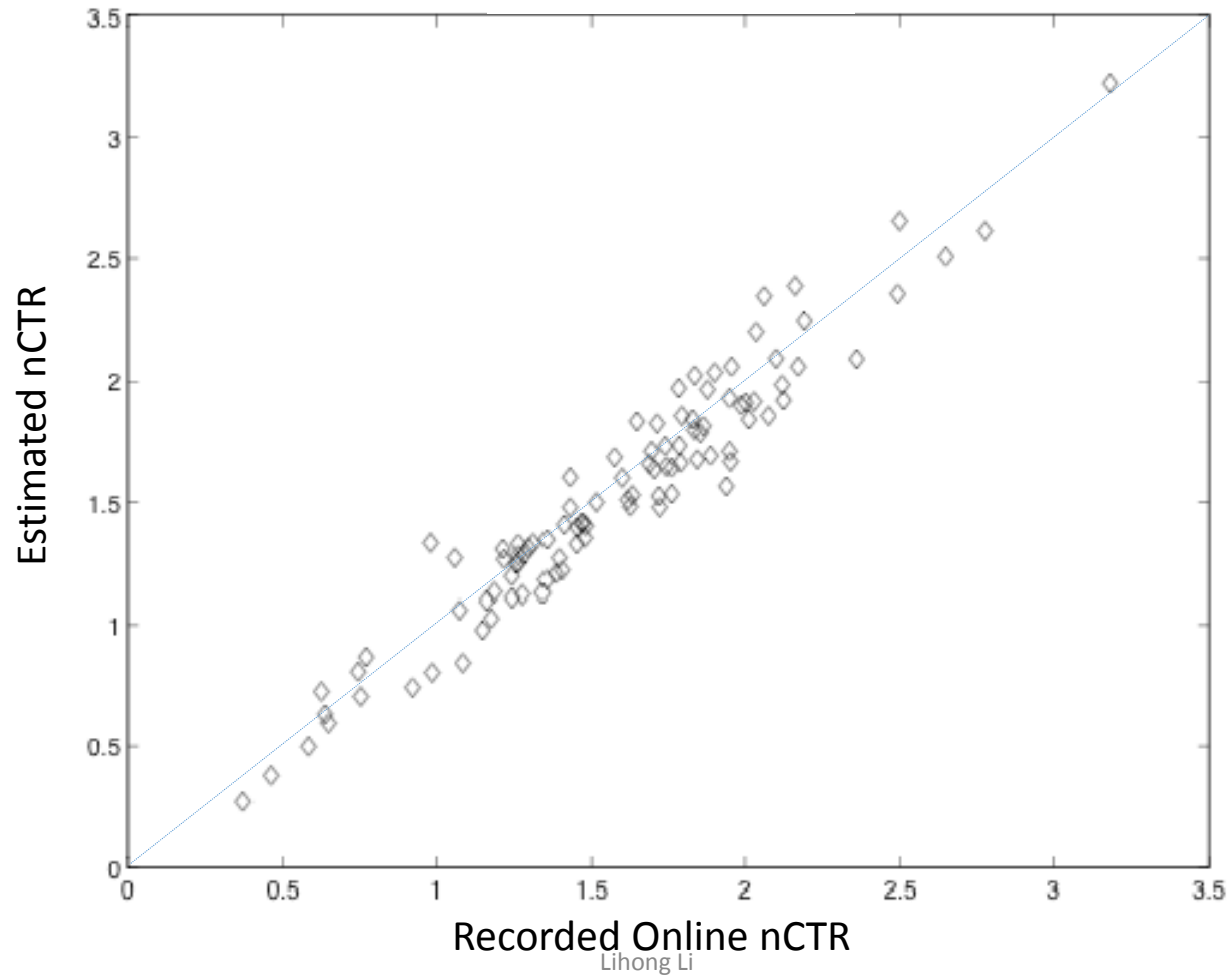
Case 1: News Recommendation [LCLW'11]

- Experiments run in 2009
- Fixed an article-selection policy π
- Run π on live users to measure online click rate
 - The ground truth
- Use exploration data to evaluate π 's click rate
 - The offline estimate

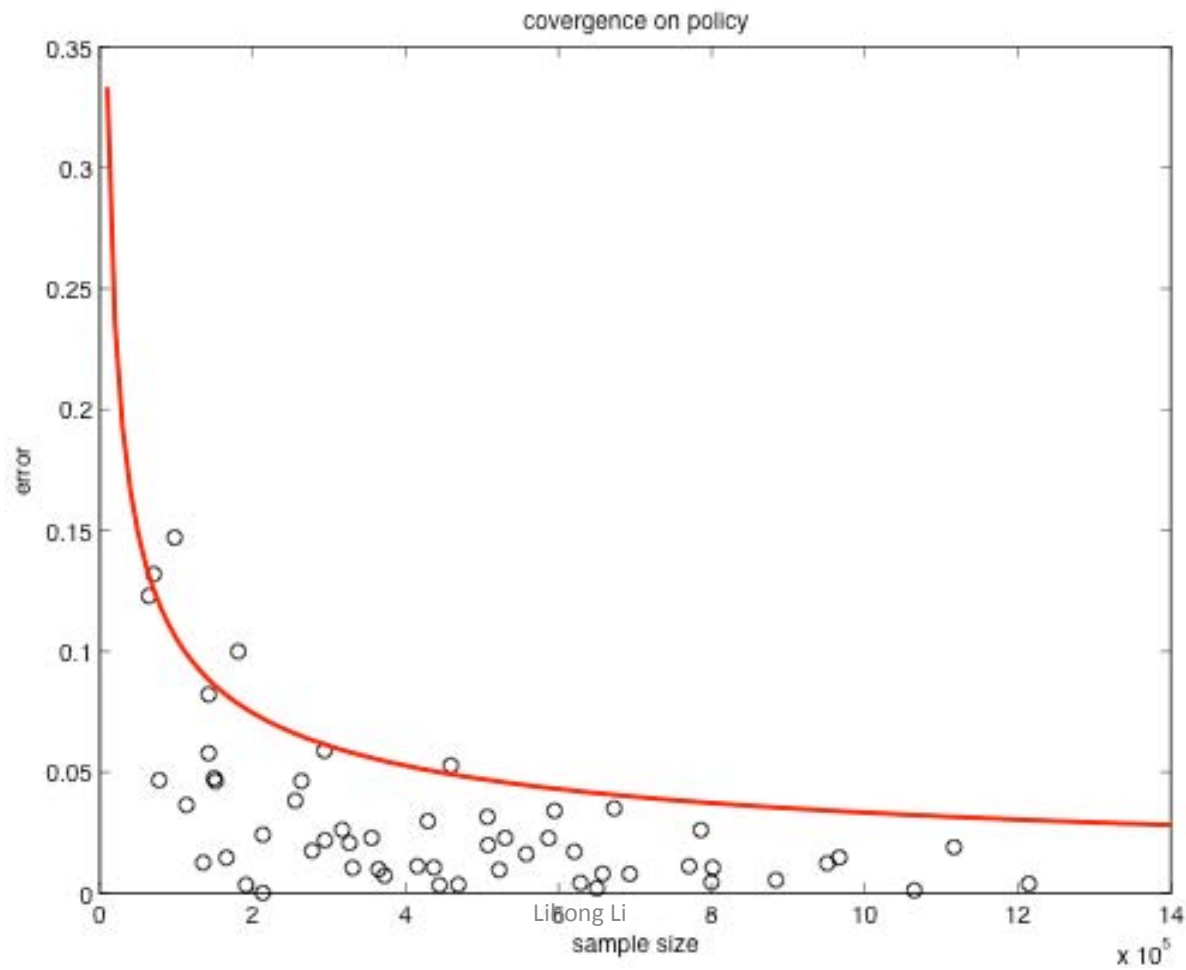


Are they close?

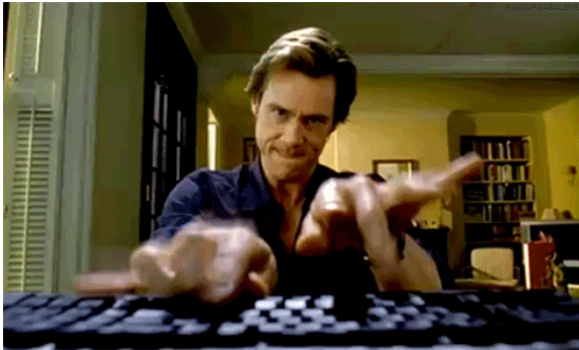
Unbiasedness



Estimation Error



Case 2: Spelling Correction of Bing



What Speller does:

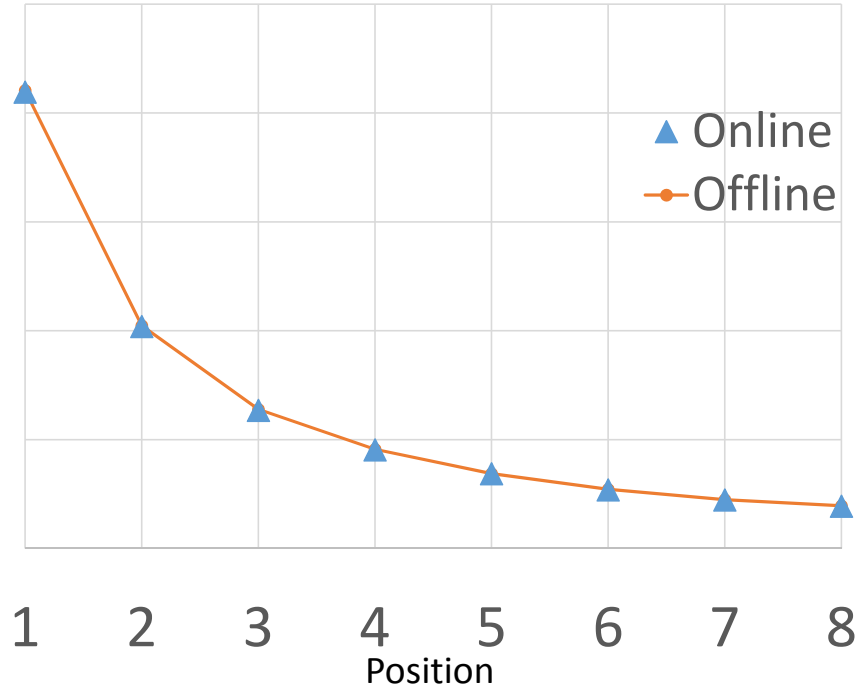
- Corrects typos
- May produce multiple candidates (with search results blended later)

Objective:

- To optimize pre-defined click metrics

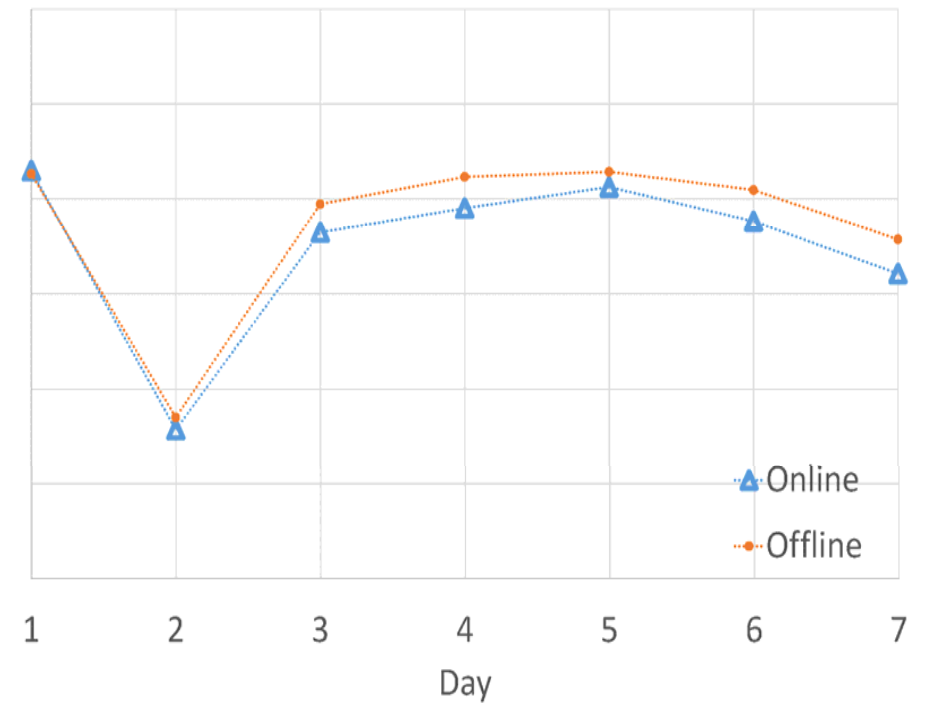
Accuracy of Offline Evaluator [LCKG'14]

Position-specific click-through rate



8/24/2014

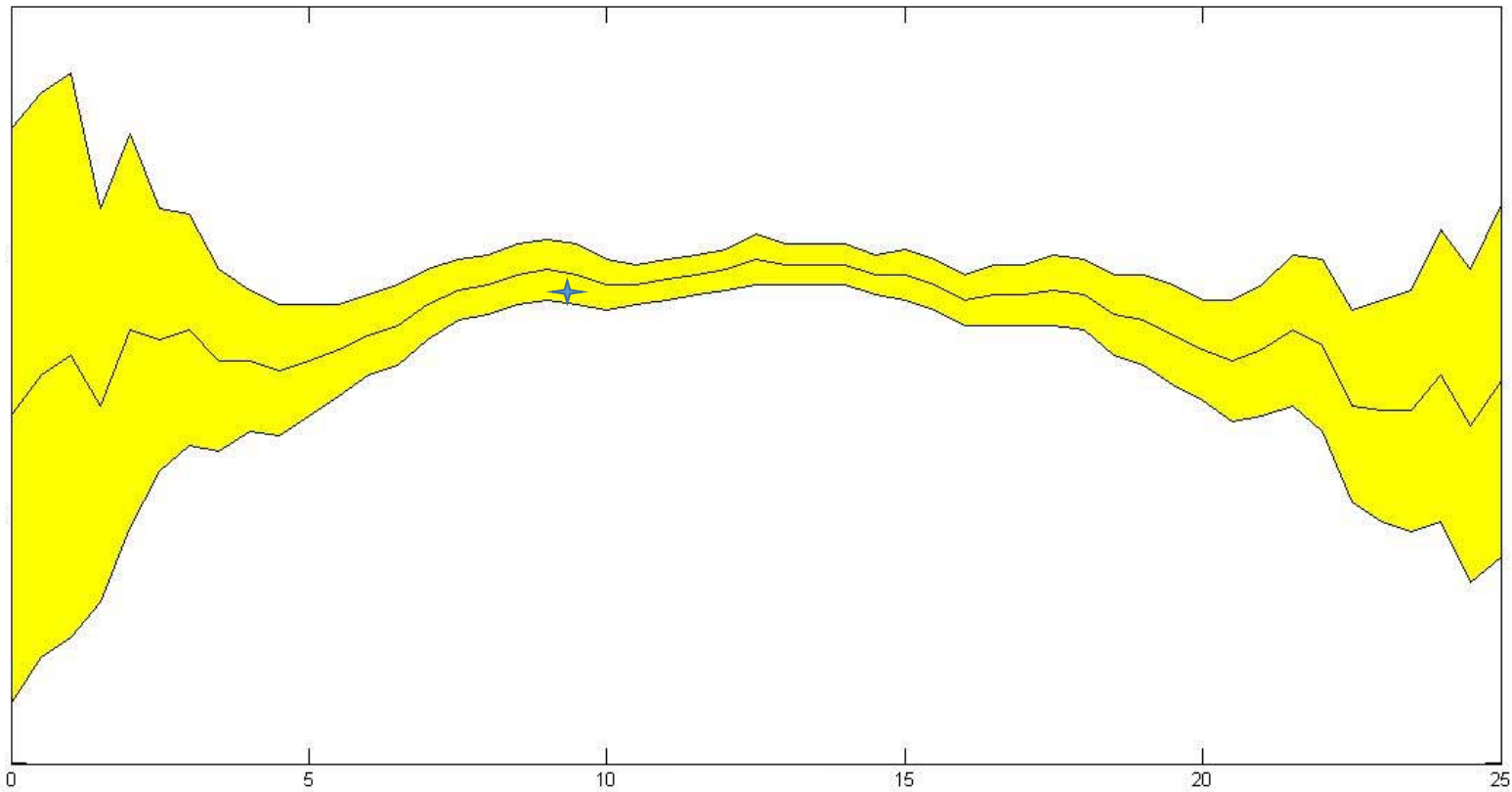
Daily click-through rate



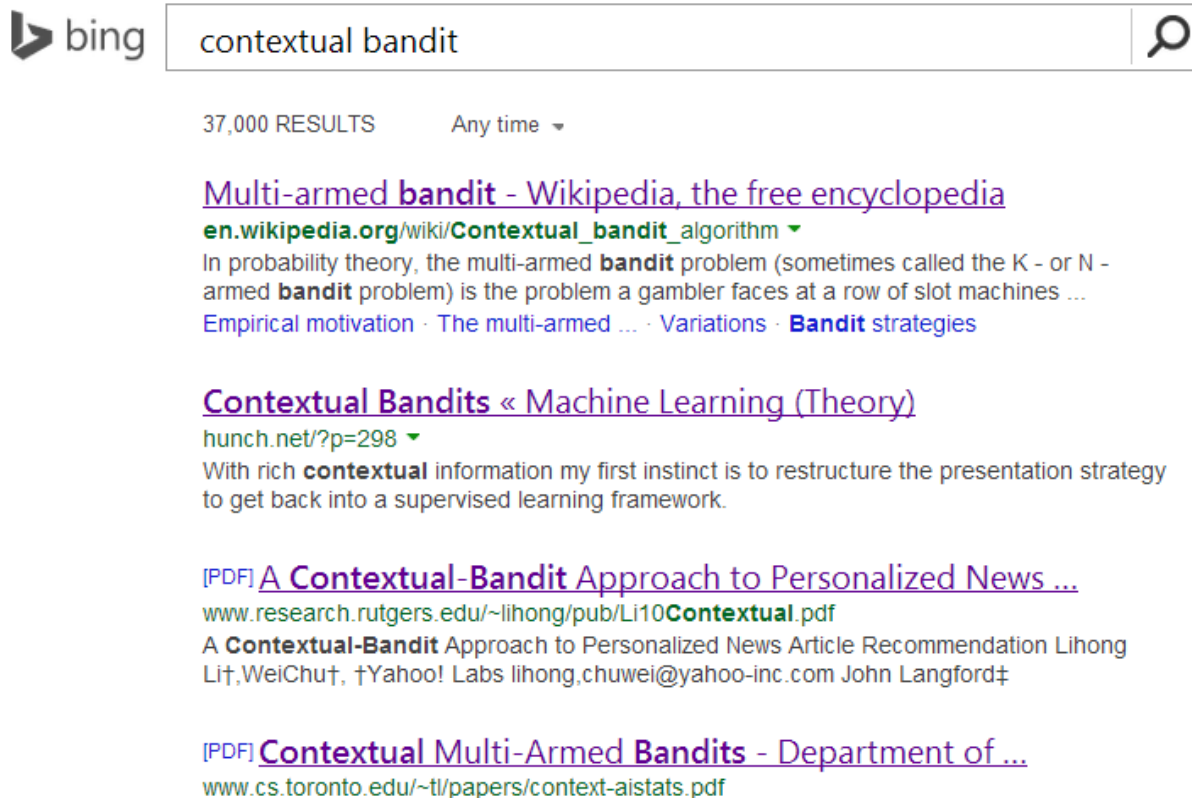
Lihong Li

29

Quantifying Uncertainty in Offline Evaluation



Case 3: Web Search Ranking



The screenshot shows a Bing search interface with the query 'contextual bandit'. The search results are as follows:

- 37,000 RESULTS** Any time ▾
- [Multi-armed **bandit** - Wikipedia, the free encyclopedia](https://en.wikipedia.org/wiki/Contextual_bandit_algorithm)
en.wikipedia.org/wiki/Contextual_bandit_algorithm ▾
In probability theory, the multi-armed **bandit** problem (sometimes called the K - or N - armed **bandit** problem) is the problem a gambler faces at a row of slot machines ...
Empirical motivation · The multi-armed ... · Variations · **Bandit** strategies
- [Contextual Bandits « Machine Learning \(Theory\)](http://hunch.net/?p=298)
hunch.net/?p=298 ▾
With rich **contextual** information my first instinct is to restructure the presentation strategy to get back into a supervised learning framework.
- [\[PDF\] A Contextual-Bandit Approach to Personalized News ...](http://www.research.rutgers.edu/~lihong/pub/Li10Contextual.pdf)
www.research.rutgers.edu/~lihong/pub/Li10Contextual.pdf
A **Contextual-Bandit** Approach to Personalized News Article Recommendation Lihong Li†, Wei Chu†, †Yahoo! Labs lihong.chuwei@yahoo-inc.com John Langford‡
- [\[PDF\] Contextual Multi-Armed Bandits - Department of ...](http://www.cs.toronto.edu/~tl/papers/context-aistats.pdf)
www.cs.toronto.edu/~tl/papers/context-aistats.pdf

Search as a bandit:

- Context: query
- Action: ranked list
- Reward: search success-or-not

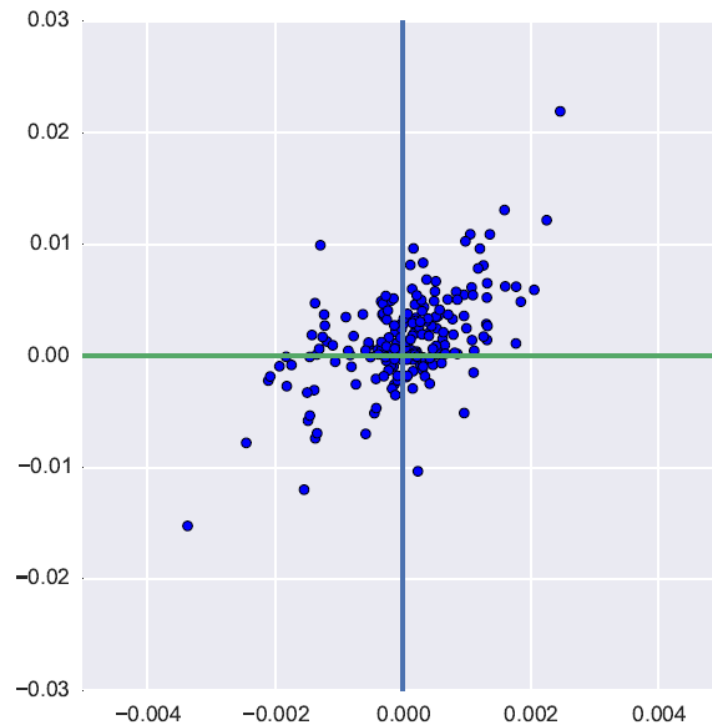
Challenges:

- Exponentially many actions
→ large estimation variance
- Collecting enough randomized data can be too expensive

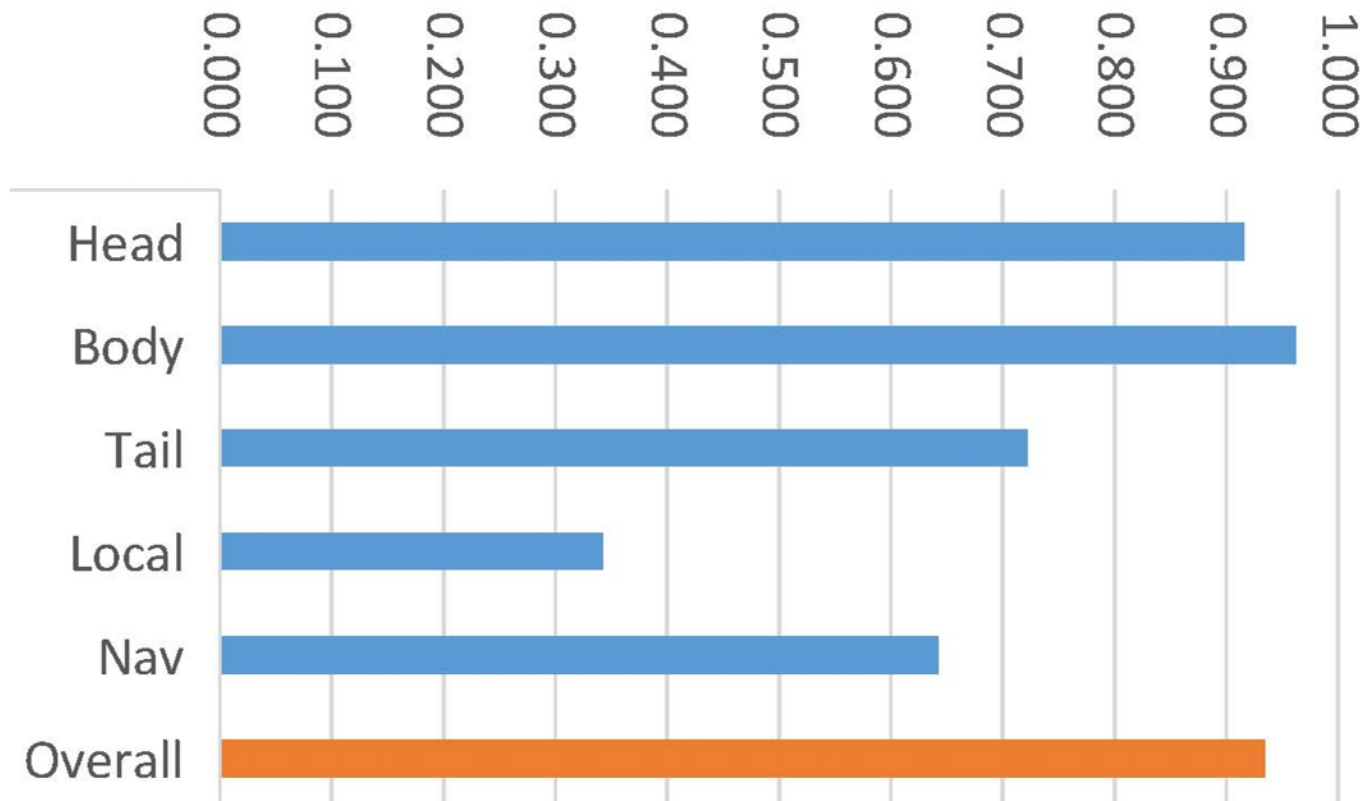
Trading off Bias and Variance [LKZ'14]

- Use **natural exploration** (uncontrolled diversity) of Bing to simulate randomized data collection
 - Nearly unbiased [SLLK'11]
 - Can use unlimited amount of data → lower variance
- Use **approximate matching** of actions
 - May introduce some amount of bias
 - Can dramatically reduce variance

Predicting Success of New Ranking Function



Metric Correlation based on Query Segments



Advanced Offline Evaluation Techniques

- Doubly robust estimation [DLL'11]
- Extends to evaluate learning algorithms (e.g., LinUCB) [DELL'12]
 - With adaptive importance sampling

Increasingly popular at industrial leaders.

Outline

- Multi-armed Bandits Algorithms
- Offline Evaluation
- Concluding Remarks

Conclusions

- Contextual bandits as a natural and versatile model
 - Better decision making → causality
 - Rich information enables better user understanding & decision making
- Additional challenges not seen in traditional bandits
 - Delayed rewards, decision making with constraints, ...
 - Dueling bandit [Yue & co]
 - Gang of bandits [Cesa-bianchi & co]
 - ...
- Large amount of data makes offline evaluation feasible
 - Can validate offline evaluation precision by experiments
 - What is the optimal estimator? [LMS'14]
- Next big step: full RL (non-myopic decision making)