

Offline Evaluation and Optimization for Interactive Systems

Lihong Li

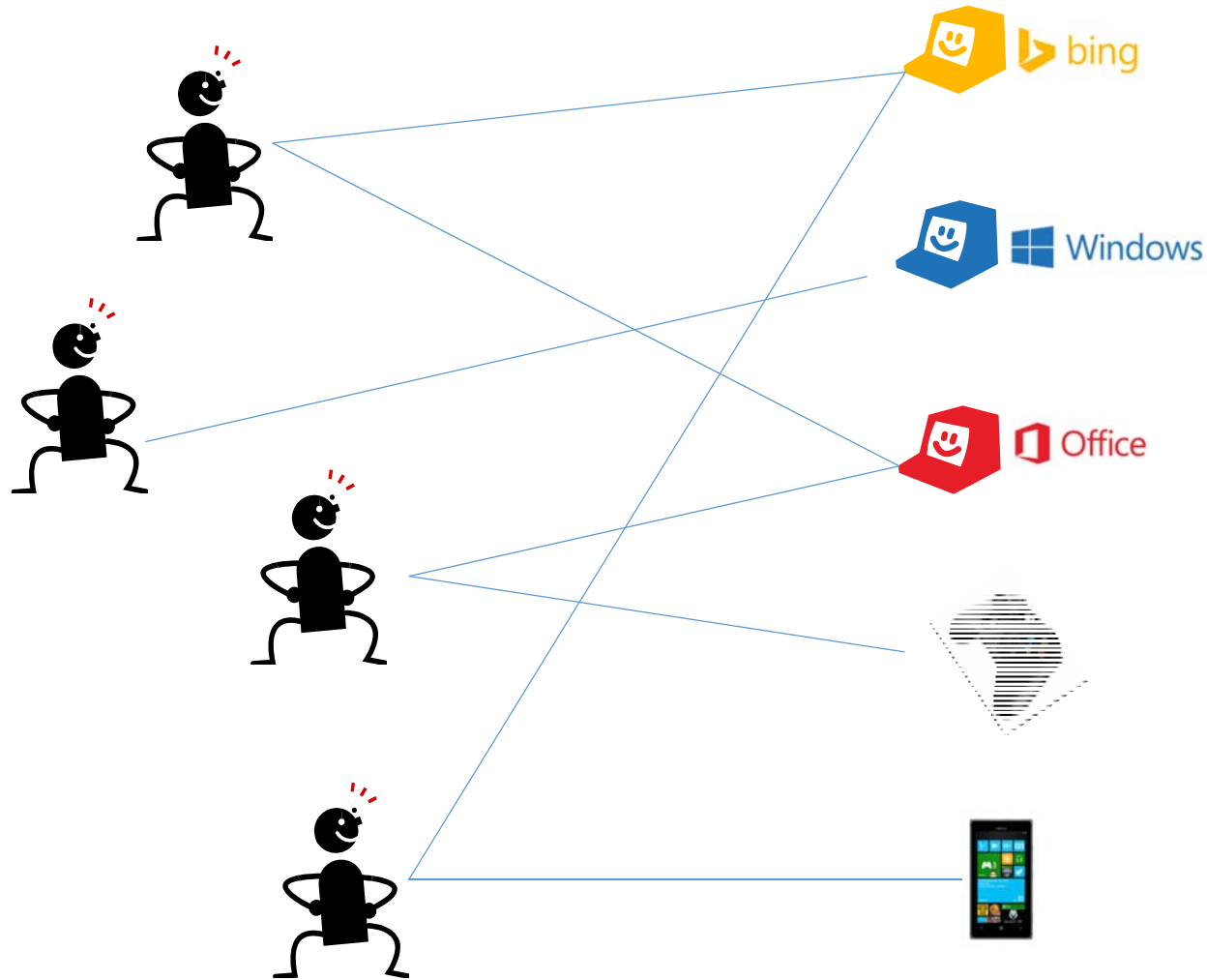
Microsoft Research

<http://research.microsoft.com/en-us/people/lihongli>

Tutorial URL

<http://research.microsoft.com/apps/pubs/default.aspx?id=240388>

User Interaction



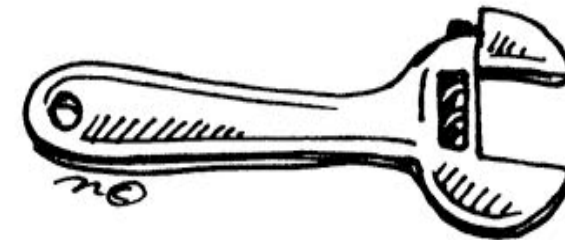
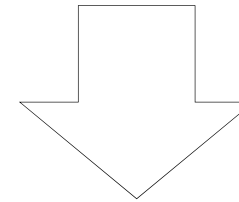
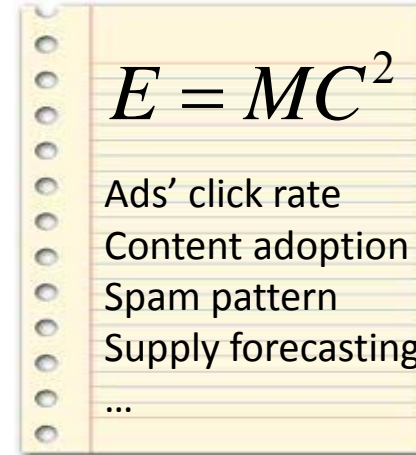
BIG DATA



correlation



KNOWLEDGE

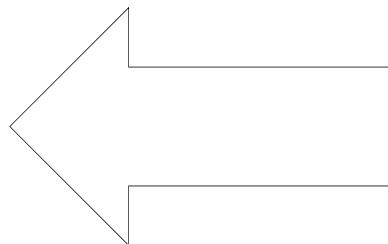


ACTION

BIGGER DATA

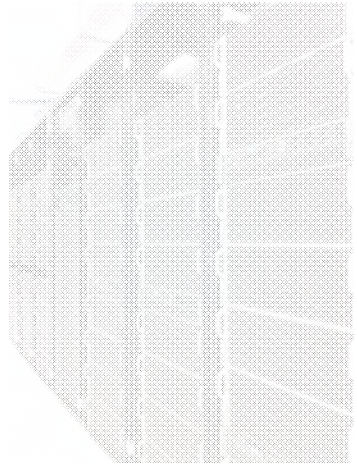


UTILITY



causation

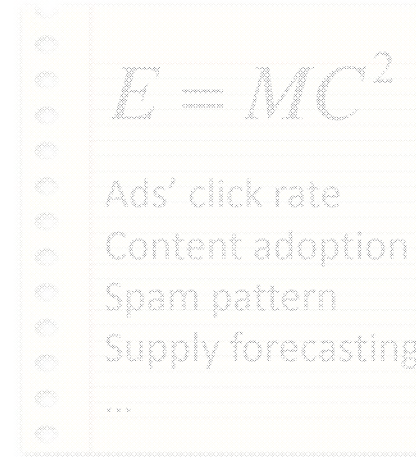
BIG DATA



correlation



KNOWLEDGE



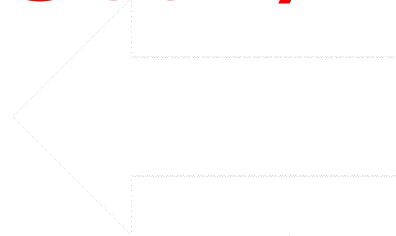
Big Trap

Correlation \neq Causation

BIGGER DATA



UTILITY



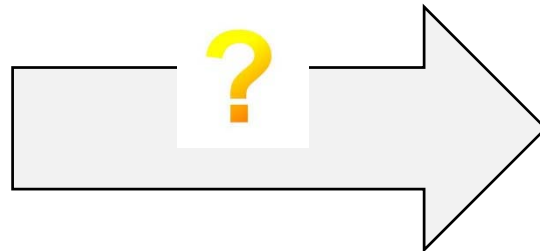
causation



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!



Outline

- Introduction
 - Contextual bandits
 - **Basic offline evaluation**
-
- **Enhanced techniques**
 - Practical issues
 - Concluding remarks

Introduction

News Recommendation

- Recommend 2 news articles {sport, movie} to users
- To maximize CTR (click-through rate)

| | Overall CTR | Male | Female |
|-------|-------------|------|--------|
| Sport | 0.5 | 0.4 | 0.8 |
| Movie | 0.6 | 0.3 | 0.7 |

Fraction of **males** and **females** who saw "Sport" and "Movie"

$$0.5 = 0.4 \times \frac{3}{4} + 0.8 \times \frac{1}{4}$$
$$0.6 = 0.3 \times \frac{1}{4} + 0.7 \times \frac{3}{4}$$

- Known as Simpson's Paradox
 - Observed in medical research, student administration, ...
 - More data does not help (because of "confounding")
 - More features do not reliably address the problem

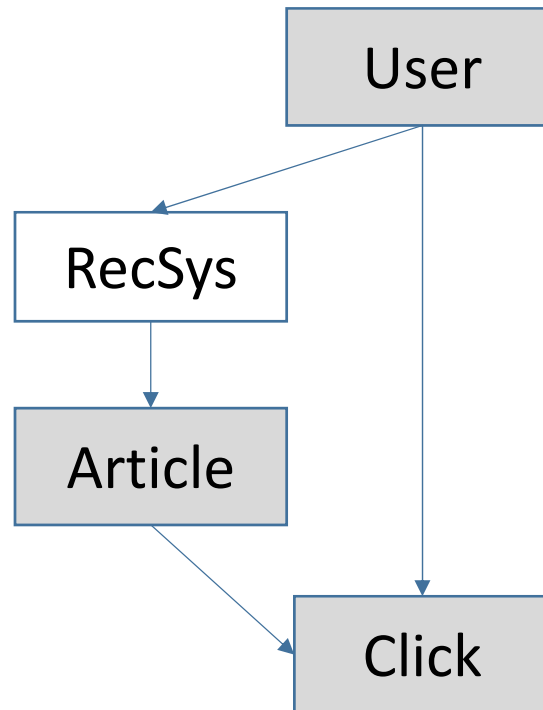
Correlation
≠
Causation!

Correlation vs. Causation

Can I predict click well
assuming fixed RecSys?

Metrics

Precision, Recall,
MSE, NDCG, ...



Can I increase CTR
if I change RecSys?

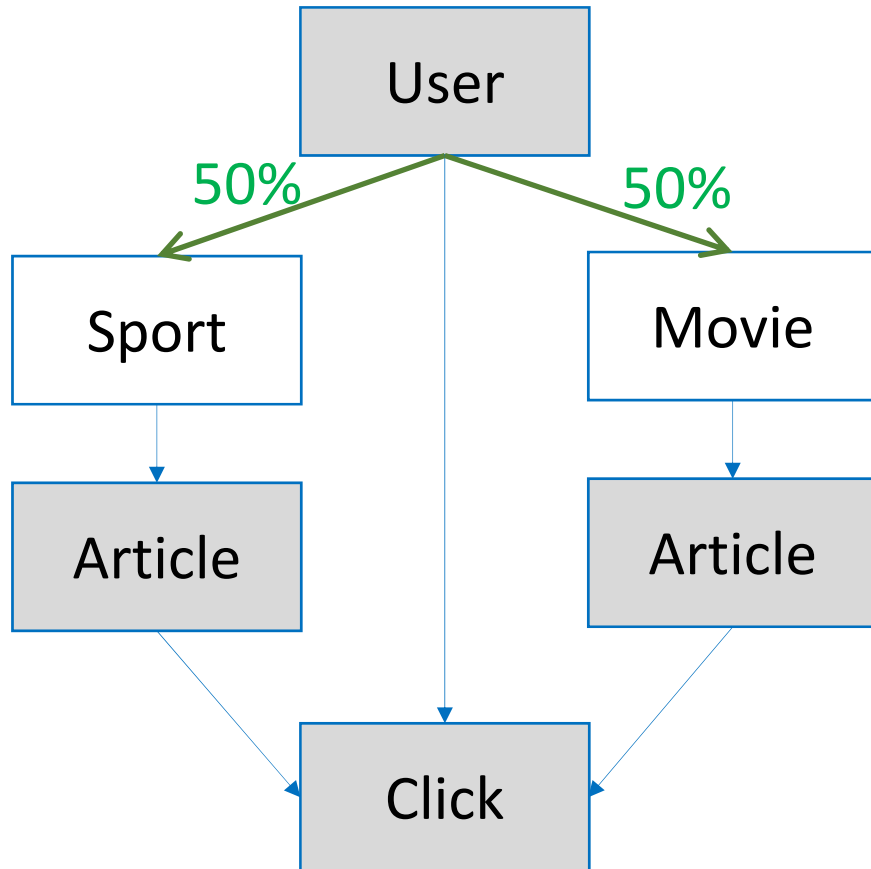
“causal effect”
“manipulation”

Metrics

CTR, revenue, ...

Similar in Web search, advertising, ...

Controlled Experiments to Identify Causality



| | Overall | Male | Female | EXP |
|-------|---------|------|--------|------------|
| Sport | 0.5 | 0.4 | 0.8 | 0.6 |
| Movie | 0.6 | 0.3 | 0.7 | 0.5 |

Everyday practice of scientist, doctors, ...
See survey of Web applications [KLSH'09]

Also known as
A/B tests, randomized clinical trials, ...

Offline vs. Online Gap in Practice

| | Correlation | Causation |
|---------|---|--|
| Offline | ML to improve prec/recall, MSE, NDCG, ... | This tutorial |
| Online | | Verify CTR/\$\$\$ lift by controlled experiments |



Common practice

“guess and check”

Limitations

- Online experiments are expensive
- Online experiments take a long time
- Often correlation \Rightarrow causation

*Offline/online: whether to run a *new* system on live users to collect new data

Related Areas

- (Stats/Econ) Estimating causal effects from observational data
 - Neyman-Rubin causal model [R'74] [H'86]
 - Heckman correction [H'79]
 - “Causality” [P'09]
- (AI) Off-policy reinforcement learning [PSS'00]
- (ML/Stats) Covariate shift [CSSL'08]

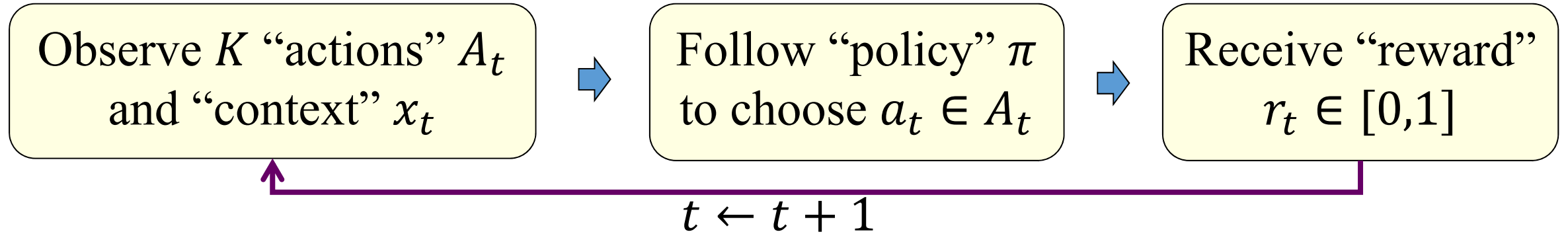
Recap

- Correlation $\not\Rightarrow$ causation
 - E.g., lower MSE $\not\Rightarrow$ CTR/revenue lift
- Controlled experiments measure causal effects (e.g., CTR lift)
 - but are expensive
- This tutorial: how to use historical data to estimate causal effects without running new online experiments

Note: Offline experiments **cannot** fully replace online experiments!

Contextual Bandits

Contextual Bandit [BA85, LZ08]



Stochastic assumption: $x_t \sim D_x(\cdot)$, $r_t \sim D_r(\cdot | x_t, a_t)$

Goal is to maximize “value”: $V(\pi, T) = \mathbf{E} \left[\frac{1}{T} (r_1 + r_2 + \dots + r_T) \right]$

Stationary policy: $a_t = \pi(x_t)$

Non-stationary policy: $a_t = \pi(x_1, a_1, r_1, \dots, x_{t-1}, a_{t-1}, r_{t-1}, x_t)$

(e.g., online learning algorithms)

historical data up to time t

Contextual Bandit Applications

- Clinical trials
- Resource allocation
- Queuing & scheduling
- ...
- Web (more recently)
 - Recommendation
 - Advertising
 - Search
- Intelligent assistant (Office)
- Adaptive user interface

Example: 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

1 - 4 of 32

Navigation: < >

x_t : user features (age, gender, location, ...)

A_t : available articles at time t

a_t : recommended article

r_t : 1 for click, 0 for no-click

Policy value $V(\pi)$ is click-through rate (CTR)

Example: Online Advertising

The image shows a Bing search results page for the query "shanghai tour". At the top, there is an MSN News banner with navigation links: HOME, US, CRIME & JUSTICE, WORLD, SCIENCE & TECH, POP CULTURE, OBITS, RUMORS, PHOTOS, VIDEO. A search bar on the right contains the text "bing" and "site search". Below the banner, the search results for "shanghai tour" are displayed, showing 5,680,000 results. The first result is an advertisement for "Shanghai Tours - Skip the Crowded Group Tours." from www.kensingtontours.com, which has 2,900+ followers on Twitter. The ad text reads: "Skip the Crowded Group Tours. Private Guided Tours of China. Private Safaris & Tours To Asia-China. Beijing Tours · Shanghai Tours". The second result is an advertisement for "Shanghai Tours & Packages | ChinaTour.Net" from ChinaTour.Net, offering "Shanghai city tour, Suzhou and Hangzhou tours, from \$69 per person". The ad text includes: "China Flight · China Tours · China Hotels · Guide". The third result is a link to "Shanghai Travel China: Facts, Attractions, City Map ..." from www.travelchinaguide.com/cityguides/shanghai.htm, with a description: "China Shanghai travel information on Shanghai facts, tours, maps, tourist attractions, holiday hotels, weather, pictures, dining, shopping, nightlife as well as ...". To the right of the search results, there is a large advertisement for the 2013 Chevrolet Volt. The ad features the Chevrolet logo and the slogan "FIND NEW ROADS". The text reads: "2013 CHEVROLET VOLT FOR A TOTAL OF UP TO 380 MILES ON A FULL CHARGE AND A FULL TANK OF GAS*". Below the text is an image of the silver Chevrolet Volt car. A yellow button labeled "Explore Volt" is positioned at the bottom right of the ad. At the bottom left of the ad, it says "AdChoices" and at the bottom right, "Ad Feedback".

bing shanghai tour

msn news

bing site search

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

Twitter to add abuse button after

Web Images Videos Maps News More

5,680,000 RESULTS Any time ▾

Shanghai Tours - Skip the Crowded Group Tours.
Ad www.kensingtontours.com · 2,900+ followers on Twitter
Skip the Crowded Group **Tours**. Private Guided **Tours** of China.
Private Safaris & Tours To Asia-China
[Beijing Tours](#) · [Shanghai Tours](#)

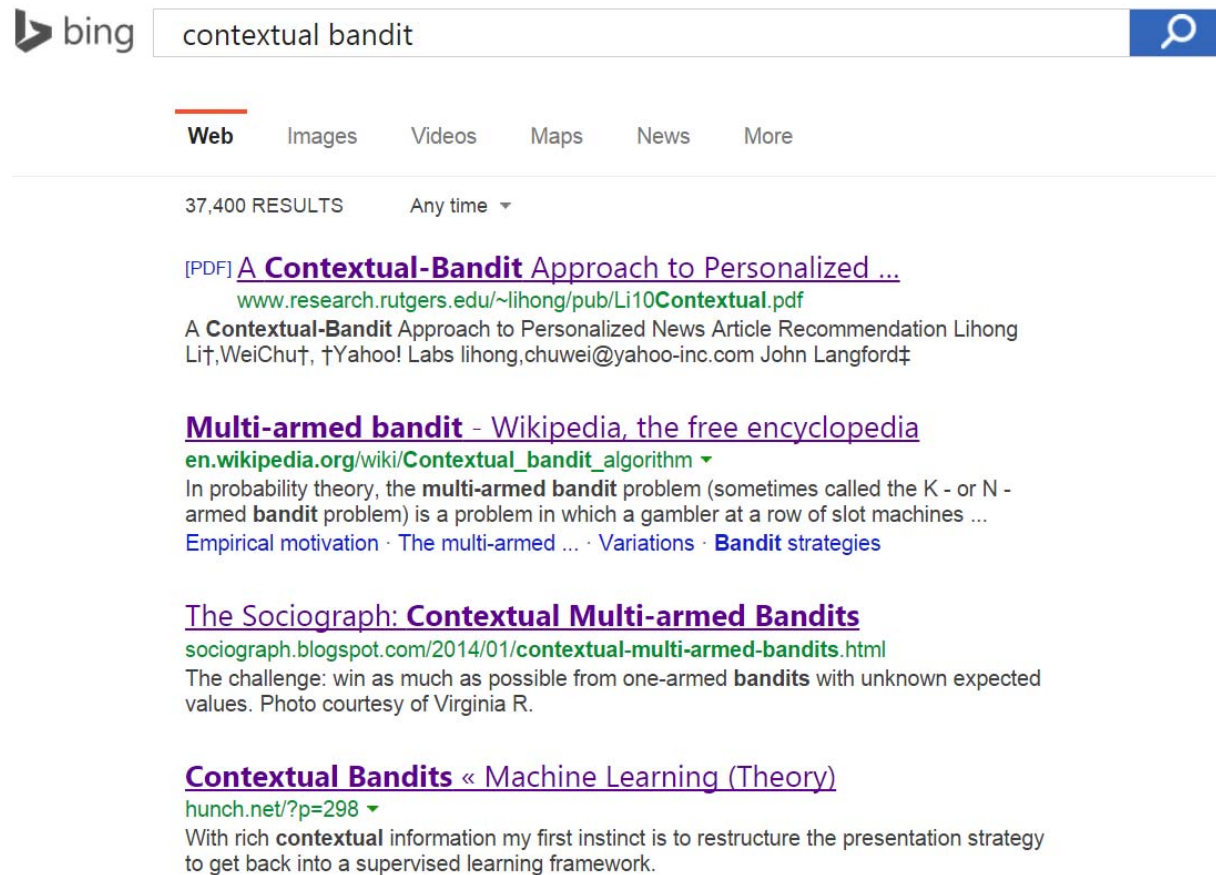
Shanghai Tours & Packages | ChinaTour.Net
Ad ChinaTour.Net
Shanghai city **tour**, Suzhou and Hangzhou **tours**, from \$69 per person
[China Flight](#) · [China Tours](#) · [China Hotels](#) · [Guide](#)

Shanghai Travel China: Facts, Attractions, City Map ...
www.travelchinaguide.com/cityguides/shanghai.htm ▾
China Shanghai travel information on Shanghai facts, tours, maps, tourist attractions, holiday hotels, weather, pictures, dining, shopping, nightlife as well as ...

2013 CHEVROLET VOLT
FIND NEW ROADS[®]
FOR A TOTAL OF UP TO 380 MILES ON A FULL CHARGE AND A FULL TANK OF GAS*
Explore Volt
AdChoices Ad Feedback

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

Example: Web Search Ranking



The screenshot shows a Bing search interface with the query 'contextual bandit'. The search results are displayed under the 'Web' tab. The first result is a PDF document titled 'A Contextual-Bandit Approach to Personalized ...' from Rutgers University. The second result is a Wikipedia article titled 'Multi-armed bandit - Wikipedia, the free encyclopedia'. The third result is a blog post titled 'The Sociograph: Contextual Multi-armed Bandits' from Sociograph. The fourth result is an article titled 'Contextual Bandits << Machine Learning (Theory)' from Hunch.net.

bing contextual bandit

Web Images Videos Maps News More

37,400 RESULTS Any time ▾

[PDF] [A Contextual-Bandit Approach to Personalized ...](#)
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‡

[Multi-armed bandit - Wikipedia, the free encyclopedia](#)
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 a problem in which a gambler at a row of slot machines ...
[Empirical motivation](#) · [The multi-armed ...](#) · [Variations](#) · [Bandit strategies](#)

[The Sociograph: Contextual Multi-armed Bandits](#)
sociograph.blogspot.com/2014/01/contextual-multi-armed-bandits.html
The challenge: win as much as possible from one-armed **bandits** with unknown expected values. Photo courtesy of Virginia R.

[Contextual Bandits << Machine Learning \(Theory\)](#)
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.

Search as a bandit
(naive formulation):

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

Policy Optimization

- Given data $D = \{(x_i, a_i, r_i)\}_{i=1,2,\dots,L}$ collected in the past, find $\pi^* = \operatorname{argmax}_{\pi} V(\pi)$
- Examples: use log data to optimize...
 - recommender model to maximize CTR
 - ad ranking system to maximize revenue
 - search engine's query suggestion model to maximize user satisfaction
 - personal treatment plan to maximize survival rate
 - ...

Policy Evaluation

- Given D and π , estimate $V(\pi)$ or $V(\pi, T) = \mathbf{E} \left[\frac{1}{T} (r_1 + r_2 + \dots + r_T) \right]$
- Example: use log data to estimate...
 - daily CTR of a news recommendation system
 - click lift of a new user feature in ad ranking
 - reduction of time for user to find a relevant URL on SERP
 - ...
- Why care evaluation
 - An important question on its own
 - Optimization can be reduced to evaluation: $\pi^* = \operatorname{argmax}_{\pi} V(\pi)$

Online vs. Offline Evaluation of $V(\pi, T)$

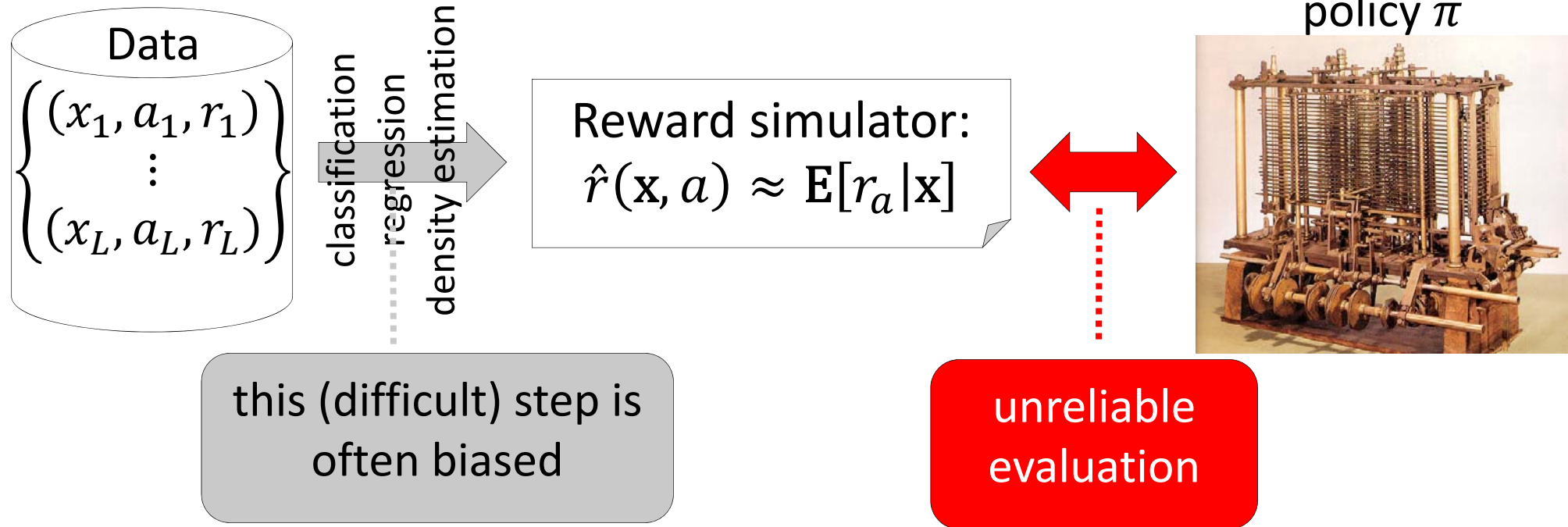
- Online evaluation
 - Controlled experiments (AB tests)
 - Wait for days/weeks/months and compute average reward
 - **Reliable** but **expensive**
- Offline evaluation
 - Use historical data $D = \{(x, a, r_a)\}$
 - **Cheap, fast, and risk-free**
 - **Counterfactuality of rewards**: do not observe $r_{\pi(x)}$ if $\pi(x) \neq a$

Recap

- Contextual bandit as natural model for many interactive ML problems
- Policy evaluation vs. optimization
- Online vs. offline policy evaluation

Basic Offline Evaluation

Direct Method (aka Regression Estimator)

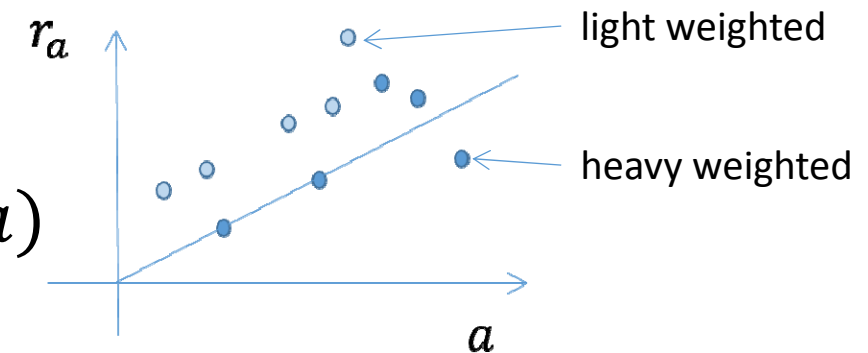


$$\hat{V}_{dm}(\pi) = \frac{1}{L} \sum_i \hat{r}(x_i, \pi(x_i))$$

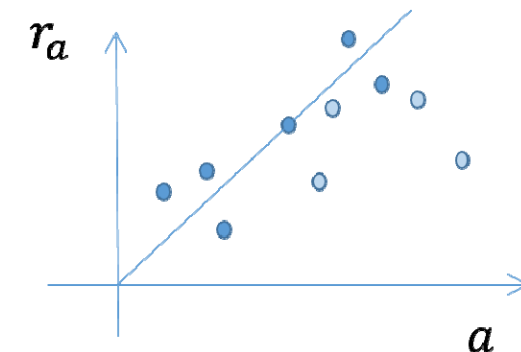
Biases of Direct Method

- Sampling/selection bias
 - From production systems
 - Simpson's paradox
- Modeling bias
 - Insufficient features to fully represent $r(x, a)$

| | Overall | Male | Female |
|-------|---------|------|--------|
| Sport | 0.5 | 0.4 | 0.8 |
| Movie | 0.6 | 0.3 | 0.7 |



Neither issue goes away even with infinite data!
Usually difficult to quantify modeling bias!



Randomized Data Collection

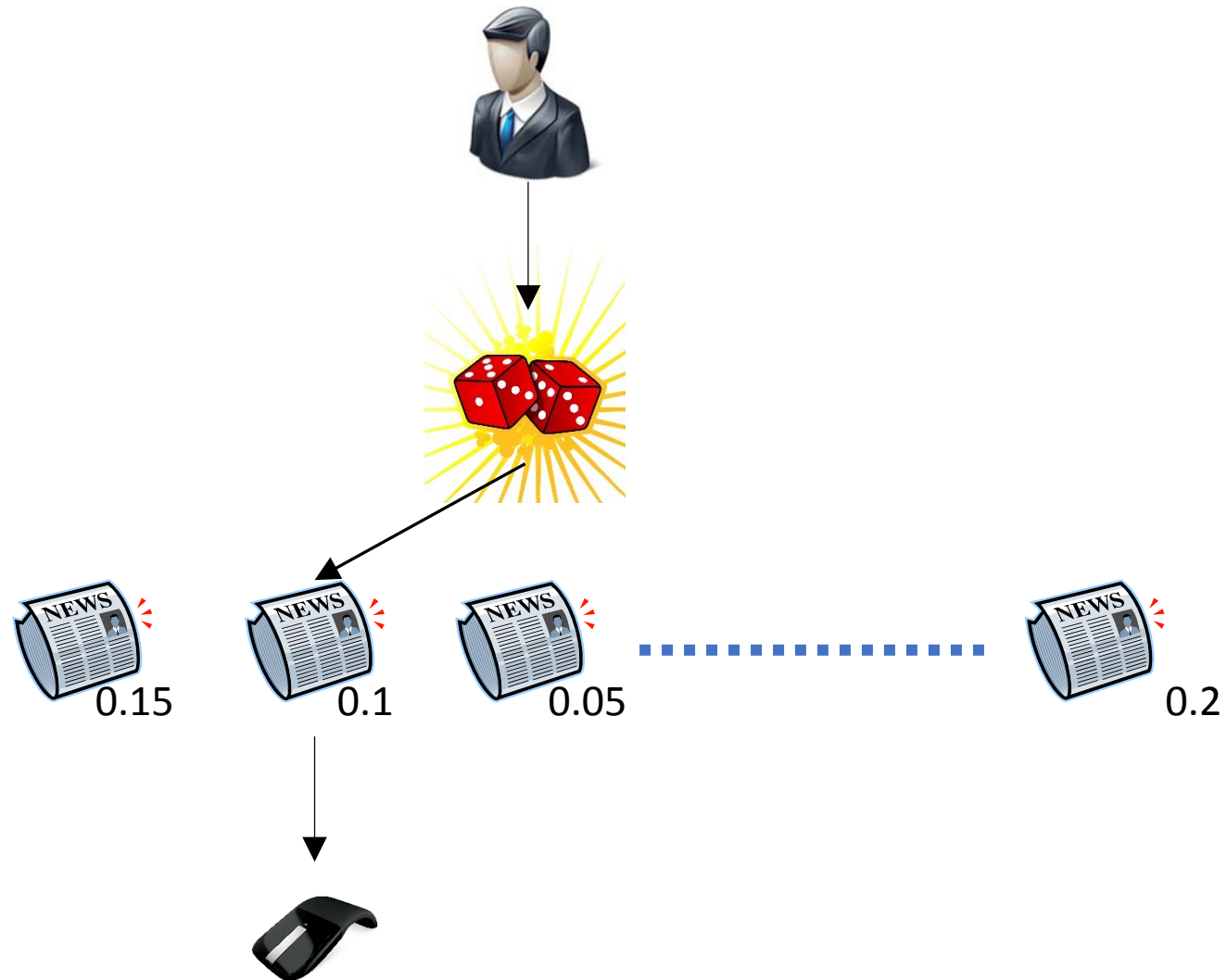
Randomized data collection: at step t ,

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

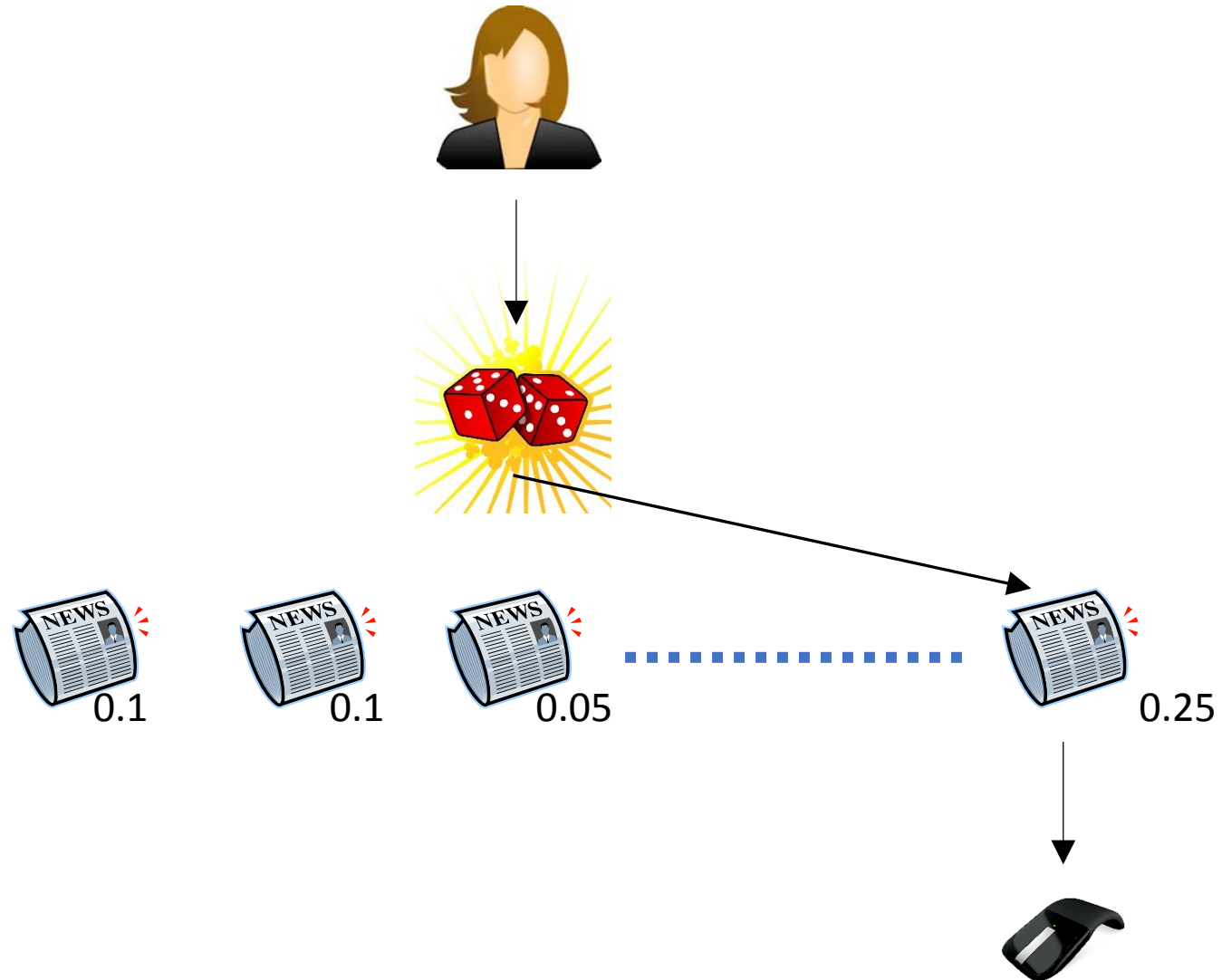
End result: “exploration data” $D = \{(x, a, r_a, p_a)\}$

Will use it to evaluate both stationary and nonstationary policies.

Randomized Data Collection: An Example



Randomized Data Collection: An Example



Inverse Propensity Score: Stationary Policy

$$\hat{V}_{\text{ips}}(\pi) = \frac{1}{L} \sum_{(x,a,p_a,r_a) \in \mathcal{D}} \frac{r_a \cdot \mathbf{1}(\pi(x) = a)}{p_a}$$

Indicator function:
1 if TRUE, 0 if FALSE

“propensity score”

Theorem: $\hat{V}_{\text{ips}}(\pi)$ is unbiased

Proof:

$$\begin{aligned} E[\hat{V}(\pi)] &= E\left[\frac{r_a \cdot \mathbf{1}(\pi(x)=a)}{p_a}\right] \\ &= E\left[\sum_a \left(p_a \times \frac{r_a}{p_a} \mathbf{1}(\pi(x) = a)\right)\right] \\ &= E\left[\sum_a (r_a \times \mathbf{1}(\pi(x) = a))\right] \\ &= E_x[r_{\pi(x)}] = V(\pi) \end{aligned}$$

Confidence Interval Estimation for IPS

$$\hat{V}_{ips}(\pi) = \frac{1}{L} \sum_{(x,a,p_a,r_a) \in \mathcal{D}} \frac{r_a \cdot \mathbf{1}(\pi(x) = a)}{p_a}$$

- Consistency: if p_a is not too small, \hat{V}_{ips} converges to $V(\pi)$ as $L \rightarrow \infty$

- Variance: $Var[\hat{V}_{ips}(\pi)] = \frac{1}{L} Var\left[\frac{r_a \cdot \mathbf{1}(\pi(x) = a)}{p_a}\right]$

- 95% confidence interval

$$\hat{V}_{ips}(\pi) \pm \left(1.96 \times \frac{\hat{\sigma}}{\sqrt{L}}\right)$$

Just another simple random variable

- Generally, width of confidence interval shrinks to 0 at rate $O(1/\sqrt{L})$

An Illustration

| ID | x | a | r_a | p_a | $\pi(x)$ | $\pi'(x)$ |
|----|----------|-----|-------|-------|----------|-----------|
| 1 | Alice | F | 1 | 1/2 | M | F |
| 2 | Bob | M | 0 | 1/3 | S | M |
| 3 | Chuck | S | 1 | 1/6 | S | F |
| 4 | Diane | M | 1 | 1/3 | M | F |
| 5 | Eric | F | 0 | 1/2 | S | M |
| 6 | Frank | F | 0 | 1/2 | S | F |
| 7 | Gordon | M | 1 | 1/3 | S | S |
| 8 | Henry | S | 0 | 1/6 | S | F |
| 9 | Irene | F | 0 | 1/2 | M | F |
| 10 | Jennifer | F | 1 | 1/2 | M | S |

$A = \{\mathbf{F}$ inance, \mathbf{M} ovie, \mathbf{S} port}

$$p = \left\{ \frac{1}{2}, \frac{1}{3}, \frac{1}{6} \right\}$$

$$\hat{V}_{ips}(\pi) = \frac{1}{|D|} \sum_{(x,a,p_a,r_a) \in D} \frac{r_a \cdot \mathbf{1}(\pi(x) = a)}{p_a}$$

$$= \frac{1}{10} \left(\frac{1}{1/6} + \frac{1}{1/3} + \frac{0}{1/6} + 0 + \dots + 0 \right)$$

$$= \frac{9}{10}$$

$$\hat{\sigma}_{ips}^2 = \hat{\sigma}^2 \left(\frac{1}{1/6}, \frac{1}{1/3}, \frac{0}{1/6}, \underbrace{0, \dots, 0}_{\text{Seven 0s}} \right)$$

Seven 0s

Case Study 1: News Recommendation [LCLW'11]



- Experiments run in 2009
 - 40M impressions over 10 days in exploration data
 - $p_a = \frac{1}{K}$ (uniform random exploration)
- Fixed an news-selection policy π
- Online experiment with π to measure CTR
 - The **online** ground truth
- Use exploration data to offline-evaluate π
 - The **offline** estimate

Are they close?

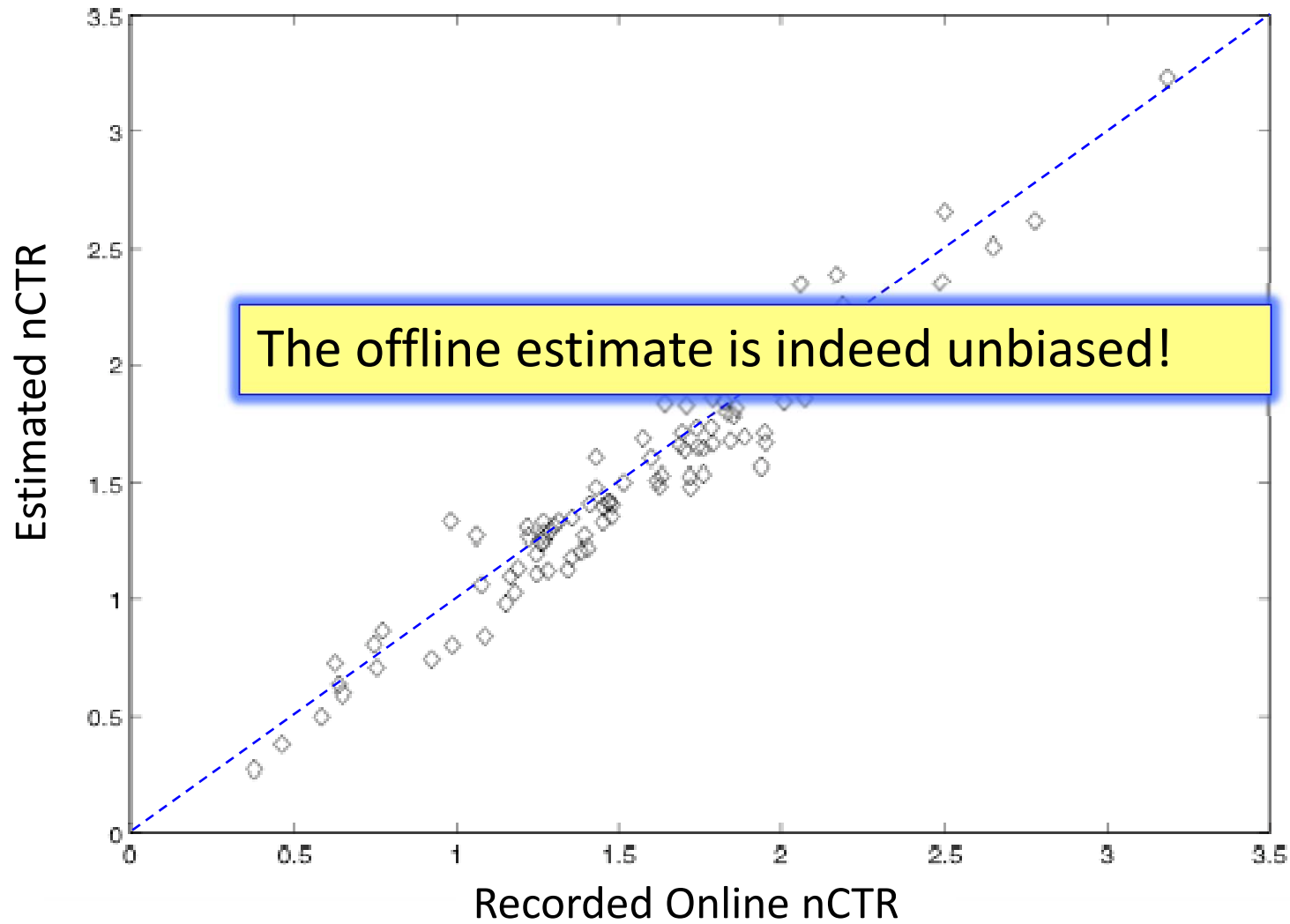
A_t : available articles at time t

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

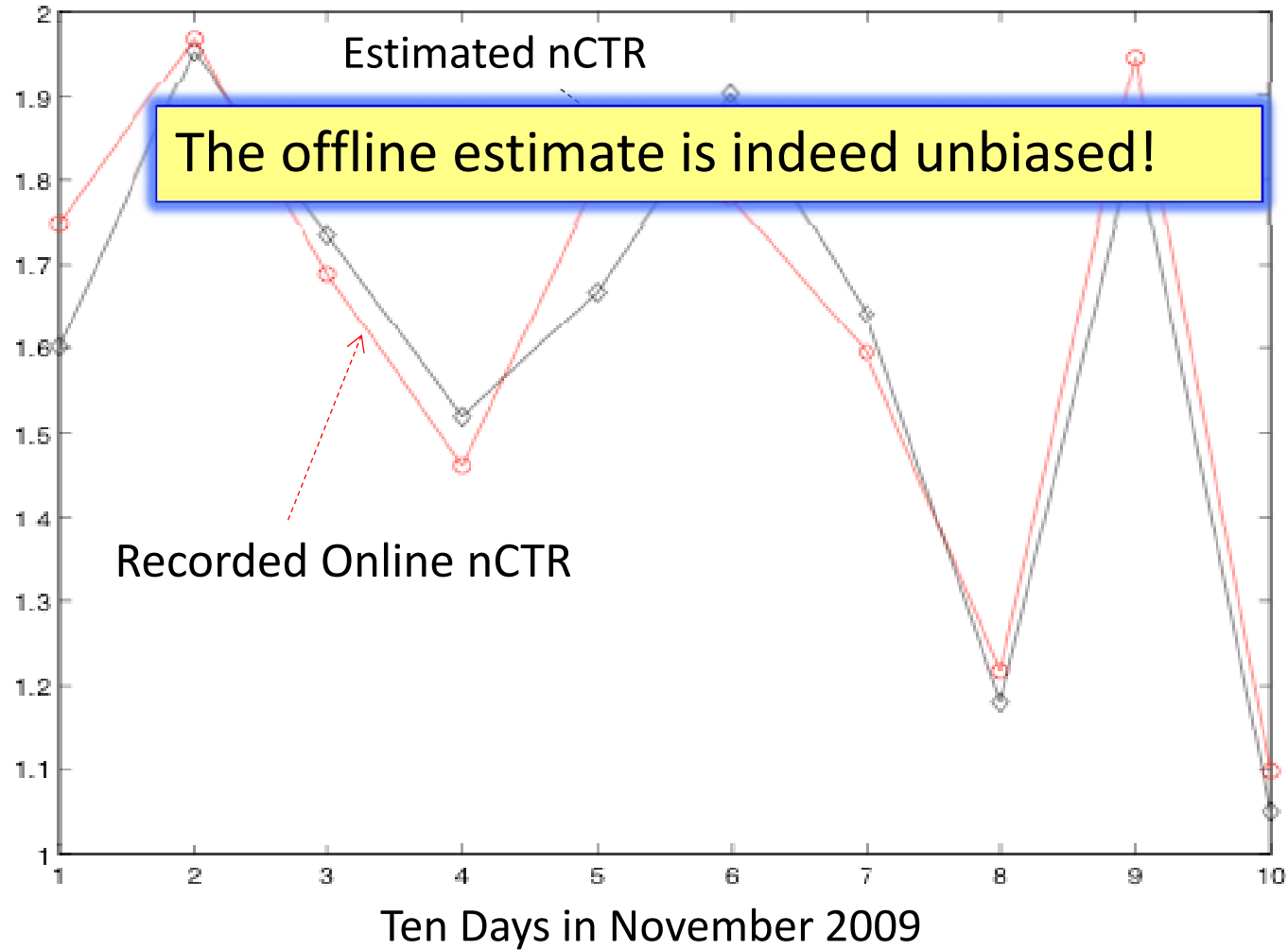
a_t : the displayed article at time t

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

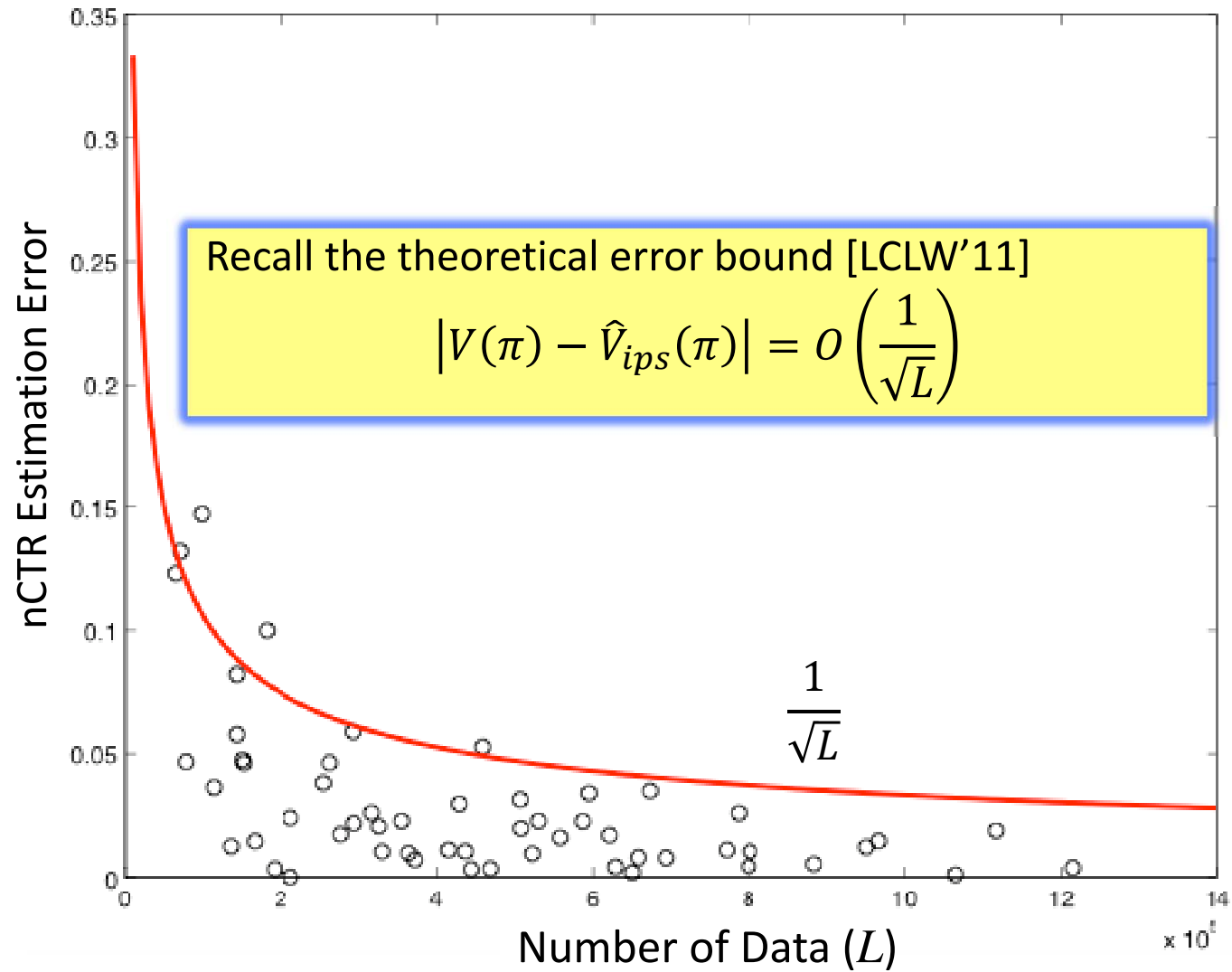
Unbiasedness: Article CTR



Unbiasedness: Daily Overall CTR





Estimation Error



Case Study 2: Bing Speller



MS Beta 397,000 RESULTS Any time ▾

Including results for *counterfactual*.
Do you want results only for counterfactual?

[counterfactual - definition of counterfactual by the Free ...](#)
www.thefreedictionary.com/counterfactual ▾

The **counterfactual** modification, then, allows us to increase the range of applications for economic laws, since it allows other discussed economic factors to change ...

[Counterfactual | Define Counterfactual at Dictionary.com](#)
dictionary.reference.com/browse/counterfactual ▾

counterfactual (, kauntə'fæktʃuəl) —adj: 1. expressing what has not happened but could, would, or might under differing conditions —n

What Speller does:

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

Popular approach:

- Obtain human labels for $(q_0, q'_c, \text{label})$
- Apply ML to rank candidates
- **But...**

Bing Speller: A Harder Example



ccn



cnn: popular and similar query (excellent reformulation candidate)

or

community cable network

ccn international

cement chemist notation

⋮

Bing Speller: A Harder Example

bing

ccn



cnn: popular and similar query (excellent reformulation candidate)

community network

journal

chemist notation

**A user-oriented solution:
use click to measure success**

**Standard solution is A/B test... but
expensive**

**Click metrics are hard to work
with offline
(b/c counterfactual nature)**

Speller as Contextual Bandit

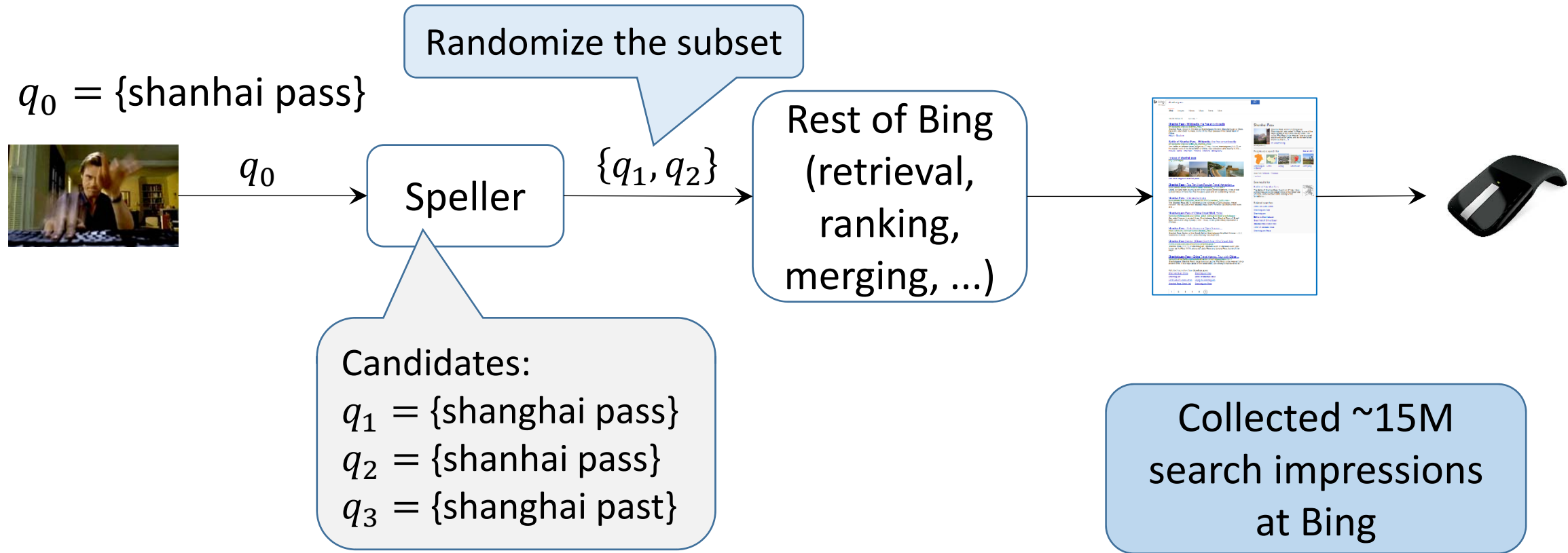
A round-by-round interaction between **S**peller and **U**ser

At each round,

- **U** issues query q_0 (“**context**”)
- **S** calculates a small set of promising candidates $Q = \{q_1, \dots, q_L\}$
 - Note: Q is assumed given (from other ML models)
- **S** then chooses an “**action**” $a \in Q$
- **S** finally observes the **reward** (some click metric) r_a for a
- Repeat

Goal of Speller is to maximize average per-round reward.

Exploration Data Collection [LCKG'14]

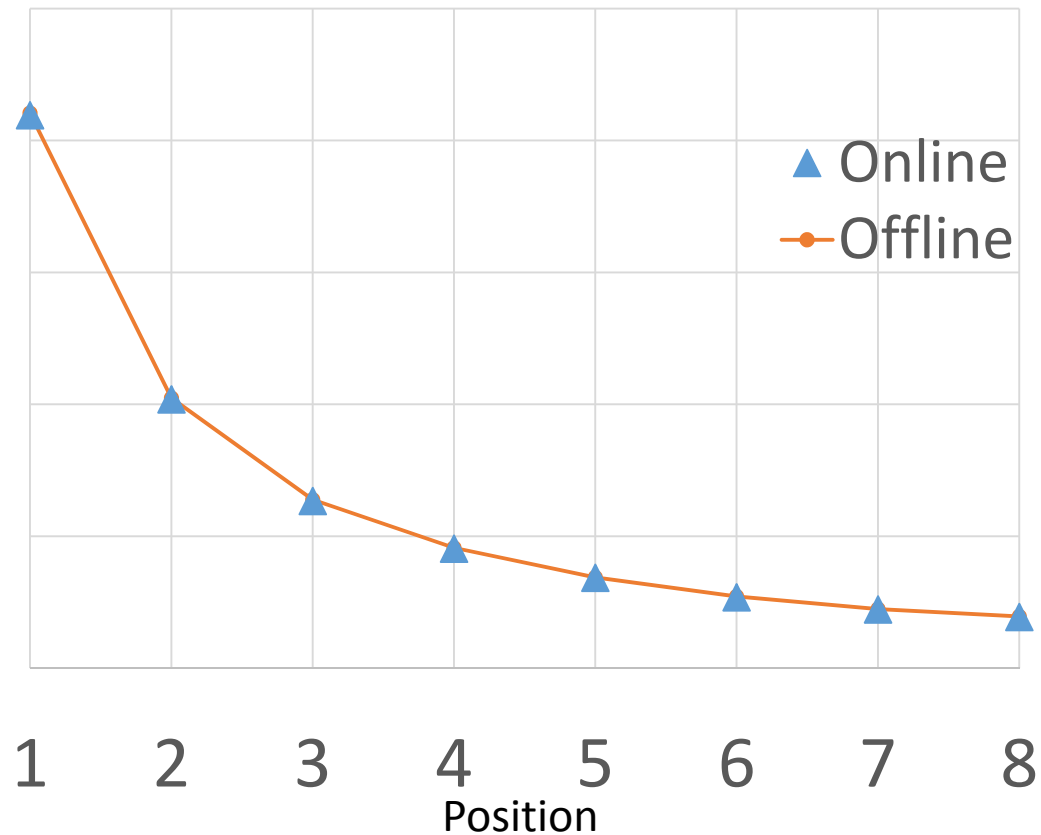


$$\Pr(q_i \text{ is sent}) = \frac{1}{1 + \exp(\lambda_1(\text{score}(q_1) - \text{score}(q_i)) + \lambda_2)}$$

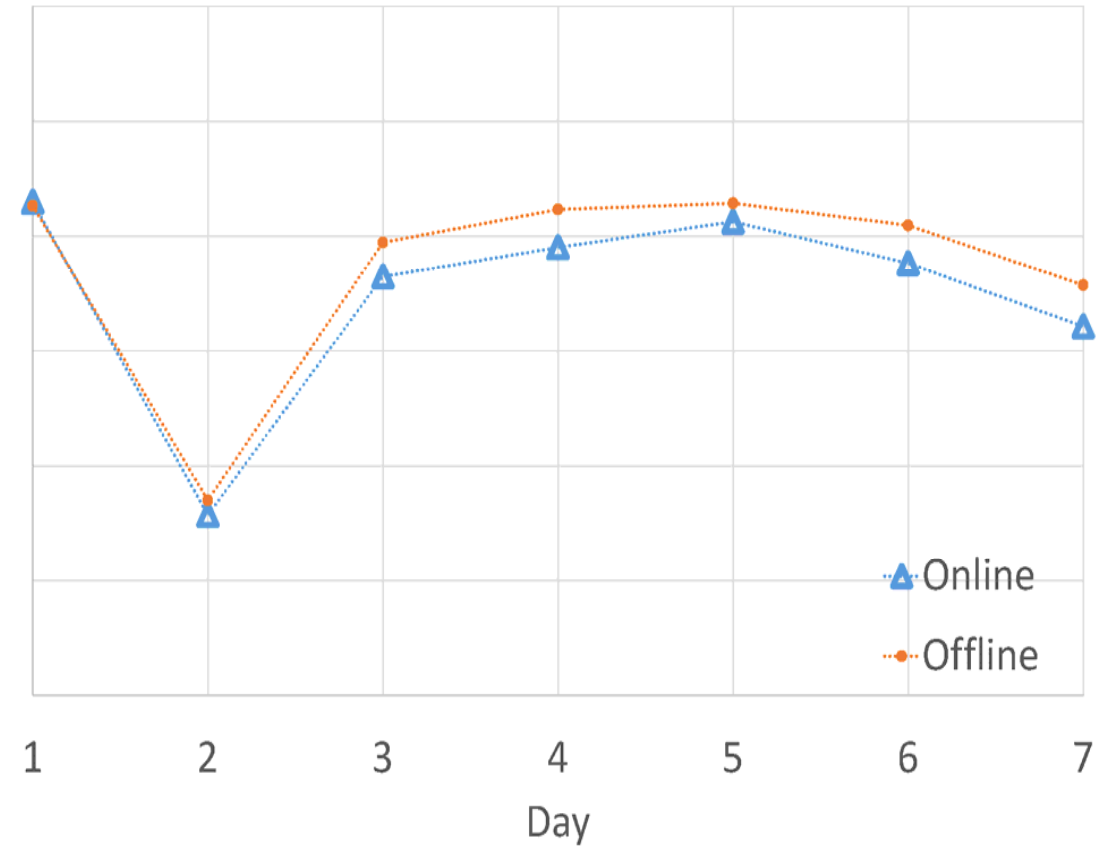
λ_1 and λ_2 control exploration aggressiveness

Accuracy of Offline Evaluator

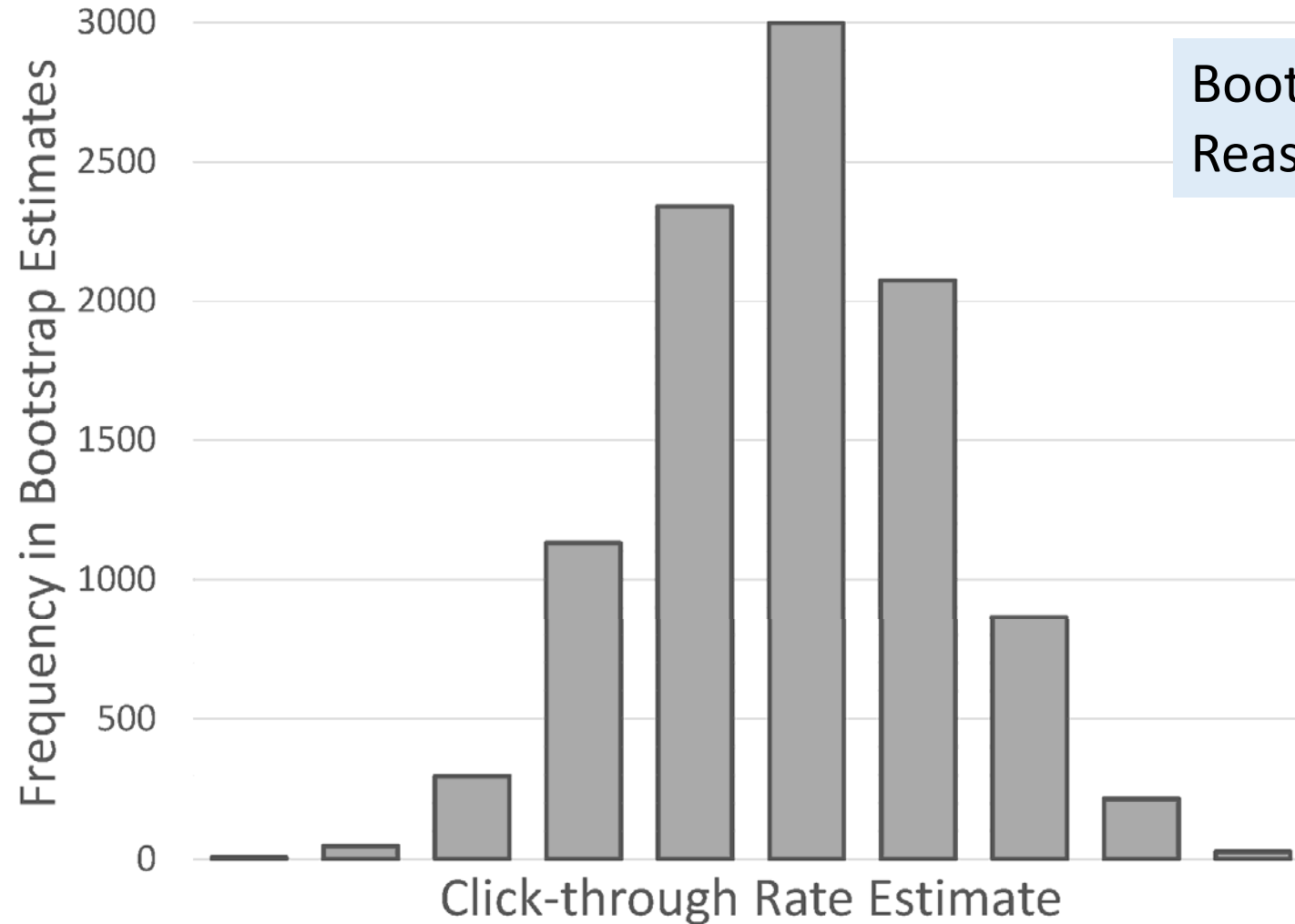
Position-specific click-through rate



Daily click-through rate

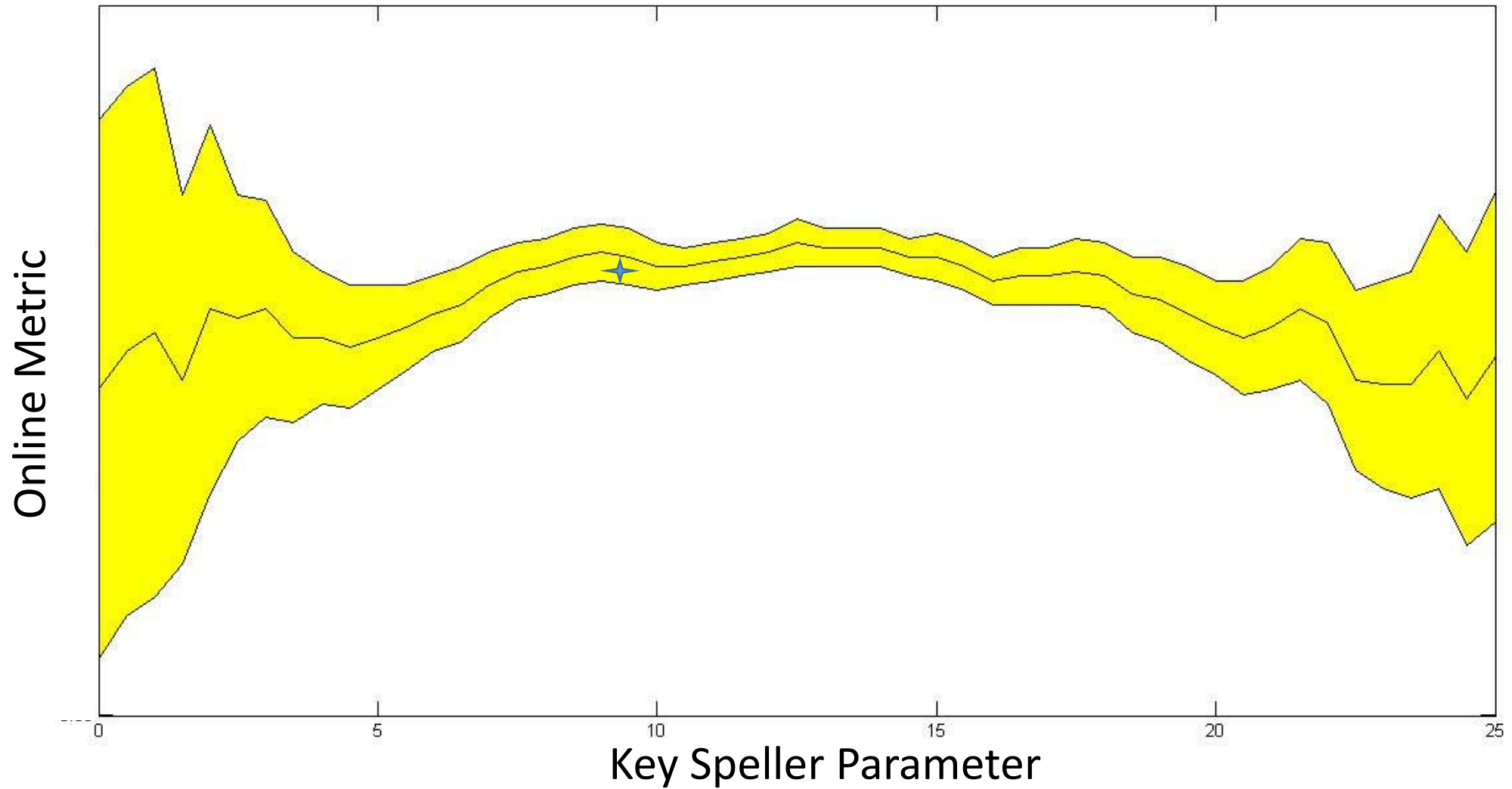


Normality of Offline Estimates



Bootstrapping $B = 10000$
Reasonable to use normal approx.

Quantifying Uncertainty in Offline Evaluation



Offline Optimization for Speller

- 70% exploration data to learn
 $\Pr(\text{GoodResult} \mid \text{Query}, \text{CorrectionCandidate})$
- 30% exploration data to offline-compare new and old Spellers
- Tends to be better if more are included
- But limited by capacity \rightarrow threshold needed
- Use unbiased IPS offline evaluation to set a threshold

Offline Optimization for Speller

- Tune Speller parameters to optimize **offline** estimate of $V(\pi)$
- Online-test one of most promising models
 - ✓ showing statistically significant gain
- Some winning examples
 - “**umecka** and zinc” → “**umcka** and zinc” (treatments for cold symptoms)
 - “catalina **left** attorney” → “catalina **leff** attorney” (right correction)
 - “acer e1-5726870” → “acer e1-572 6870” (correct word breaking)

{umecka and zinc} vs. {umecka}



Web Images Videos Maps News More

10,200,000 RESULTS Any time ▾

[Can Zinc Lozenges and Nasal Sprays Remedy Your Cold?](#)

[www.webmd.com](#) > ... > [Cold, Flu, & Cough Health Center](#) > [Cold Guide](#) ▾

Can zinc prevent or reduce the duration of cold symptoms? Learn more about zinc's benefits as a cold remedy from the experts at **WebMD**.

[Zinc, umcka & elderberry for cold season | Pharmaca ...](#)

[www.pharmaca.com/projectwellness/2014/10/10/my-3-favorite-natural...](#) ▾

Dr. Tieraona Low Dog talks about her medicine cabinet must-haves during cold and flu season, including zinc, umcka laobo and elderberry.

[ZINC: Uses, Side Effects, Interactions and Warnings - WebMD](#)

[www.webmd.com](#) > [WebMD Home](#) > [Vitamins & Supplements](#) ▾

Find patient medical information for **ZINC** on **WebMD** including its uses, effectiveness, side effects and safety, interactions, user ratings and products that have it.

[Zinc — Health Professional Fact Sheet - Office of ...](#)

[ods.od.nih.gov/factsheets/Zinc-HealthProfessional](#) ▾

Zinc is an essential mineral that is naturally present in some foods, added to others, and available as a dietary supplement. **Zinc** is also found in many cold lozenges ...



Web Images Videos Maps News More

39,600 RESULTS Any time ▾

[Umcka® - Get back to life faster with all natural Umcka ...](#)

[www.umcka.com](#)

Umcka® - Get back to life with **Umcka®** Coldcare and Cold+Flu! Recover from the cold and flu faster with **Umcka** natural cold and flu products including liquids ...

[Jolanta Umecka - IMDb](#)



[www.imdb.com/name/nm0880840](#) ▾

Jolanta Umecka, Actress: *Nóz w wodzie*. **Jolanta Umecka** is an actress, known for *Knife in the Water* (1962), *Panna zázracznica* (1967) and *Echo* ...

[News](#) · [Biography](#) · [Awards](#) · [Films](#)

Related searches for **umecka**

[Umcka Cold Remedy](#)

[Umcka Drops](#)

[Umckaloabo Walgreens](#)

[Where to Buy Umcka](#)

[Umcka Cold](#)

[Umcka Walgreens](#)

[Knife in the Water - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Knife_in_the_Water](#) ▾

Knife in the Water is a 1962 Polish drama film co-written and directed by Roman Polański, which was nominated for Academy Award for Best Foreign Language Film. It ...

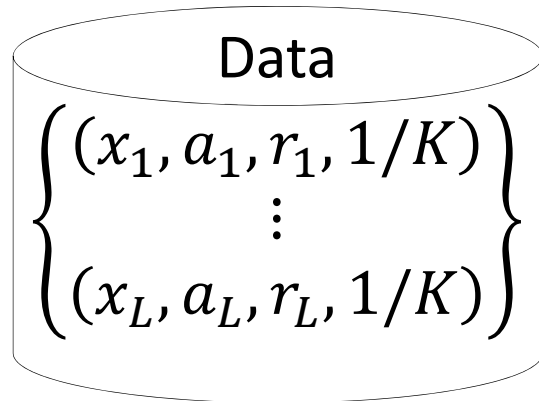
[Plot](#) · [Cast](#) · [Production](#) · [Critical reception](#) · [Home video](#)

Evaluating Nonstationary Policies

- To estimate: $V(\pi, T) = \mathbf{E} \left[\frac{1}{T} (r_1 + r_2 + \dots + r_T) \right]$
where $a_t = \pi(x_1, a_1, r_1, \dots, x_{t-1}, a_{t-1}, r_{t-1}, x_t)$
- Examples: all explore-exploit learning algorithms
- Simple inverse propensity score does not work
- Need to simulate the trajectory

The Replay Method [LCLS'10, LCLW'11]

Key requirement for data collection: $p_a \equiv \frac{1}{K}$



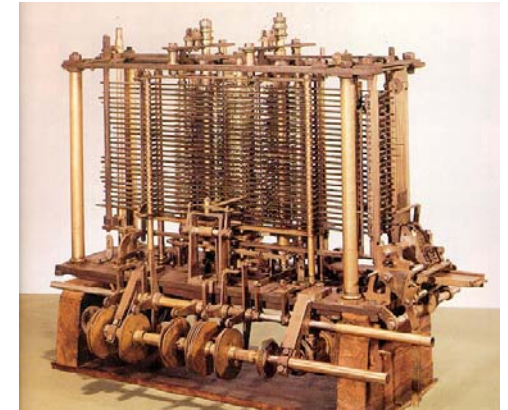
For $i = 1, 2, \dots, L$

reveal x_i

choose $\hat{a}_i = \pi(x_i)$

reveal r_i only if $\hat{a}_i = a_i$ ("match")

Nonstationary policy π



Finally output $\hat{V} \left(\pi, \frac{L}{K} \right) = \frac{K}{L} \times \sum_{i=1}^L (r_i \cdot 1(\hat{a}_i = a_i))$

Unbiasedness of Replay

- **Theorem**: if L is large enough to generate T matches in replay, then

$$E[\hat{V}(\pi, T)] = V(\pi, T)$$

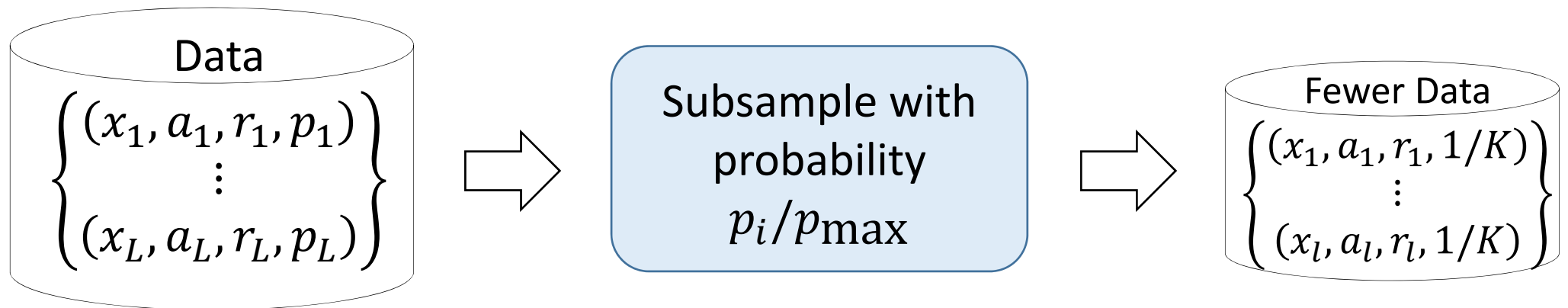
- Unfortunately, cannot use L or T to estimate confidence intervals
- Can use bootstrapping instead

- How large L do we need to have T matches?
 - On average, $L = KT$
 - With high probability, need $L \approx 2KT$

- More discussions later

Replay with Non-uniform Exploration

- Data $D = \{(x, a, r_a, p_a)\}$ where $p_a \neq \frac{1}{K}$
- Can apply **rejection sampling** to obtain a subset of uniform p_a



- Not very efficient when p_i 's vary a lot
- **Adaptive rejection sampling** [DELL'12]

Case Study 3: News Recommendation

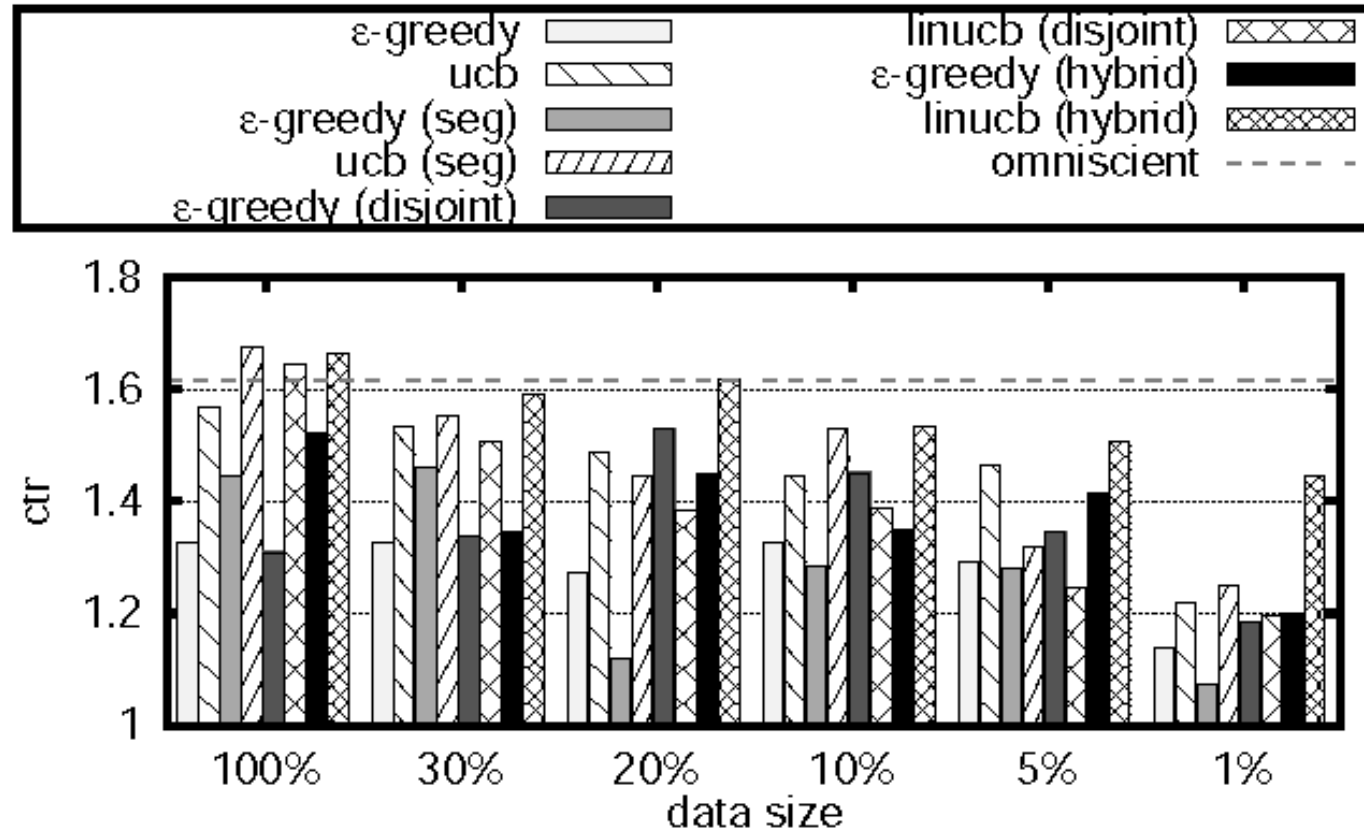
- Data collected in 2009
 - 40M impressions over 10 days in exploration data
 - $p_a = \frac{1}{K}$ (uniform random exploration)
- Low variance when evaluating representative nonstationary policies

| algorithm | mean | std | max | min |
|--------------------|--------|--------|--------|--------|
| ϵ -greedy | 1.2664 | 0.0308 | 1.3079 | 1.1671 |
| UCB | 1.3278 | 0.0192 | 1.3661 | 1.2812 |
| LinUCB | 1.3867 | 0.0157 | 1.4268 | 1.3491 |

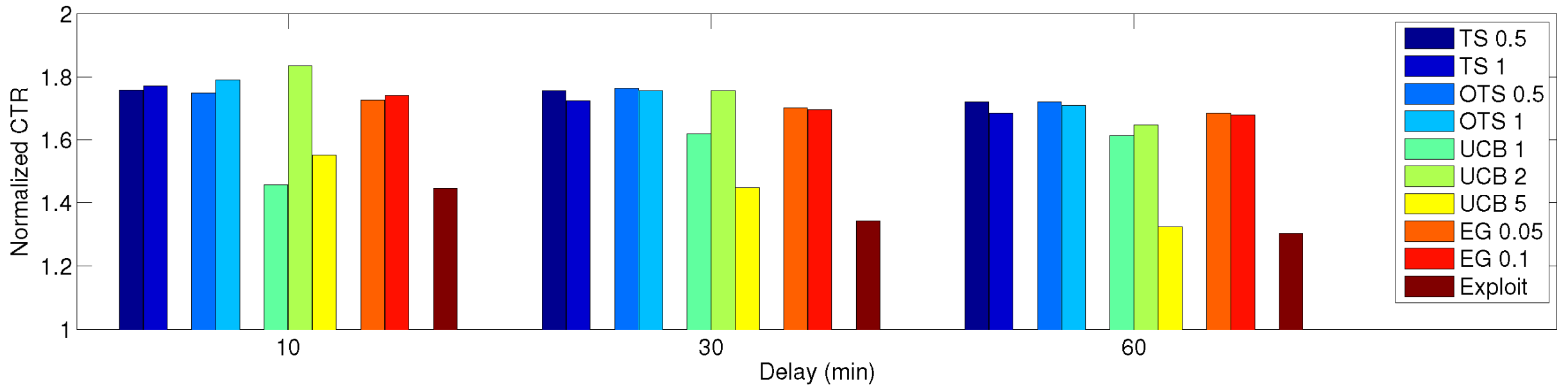
100 independent runs with different randomization seed

Conjecture: Replay has low variance for *reasonable* nonstationary policies

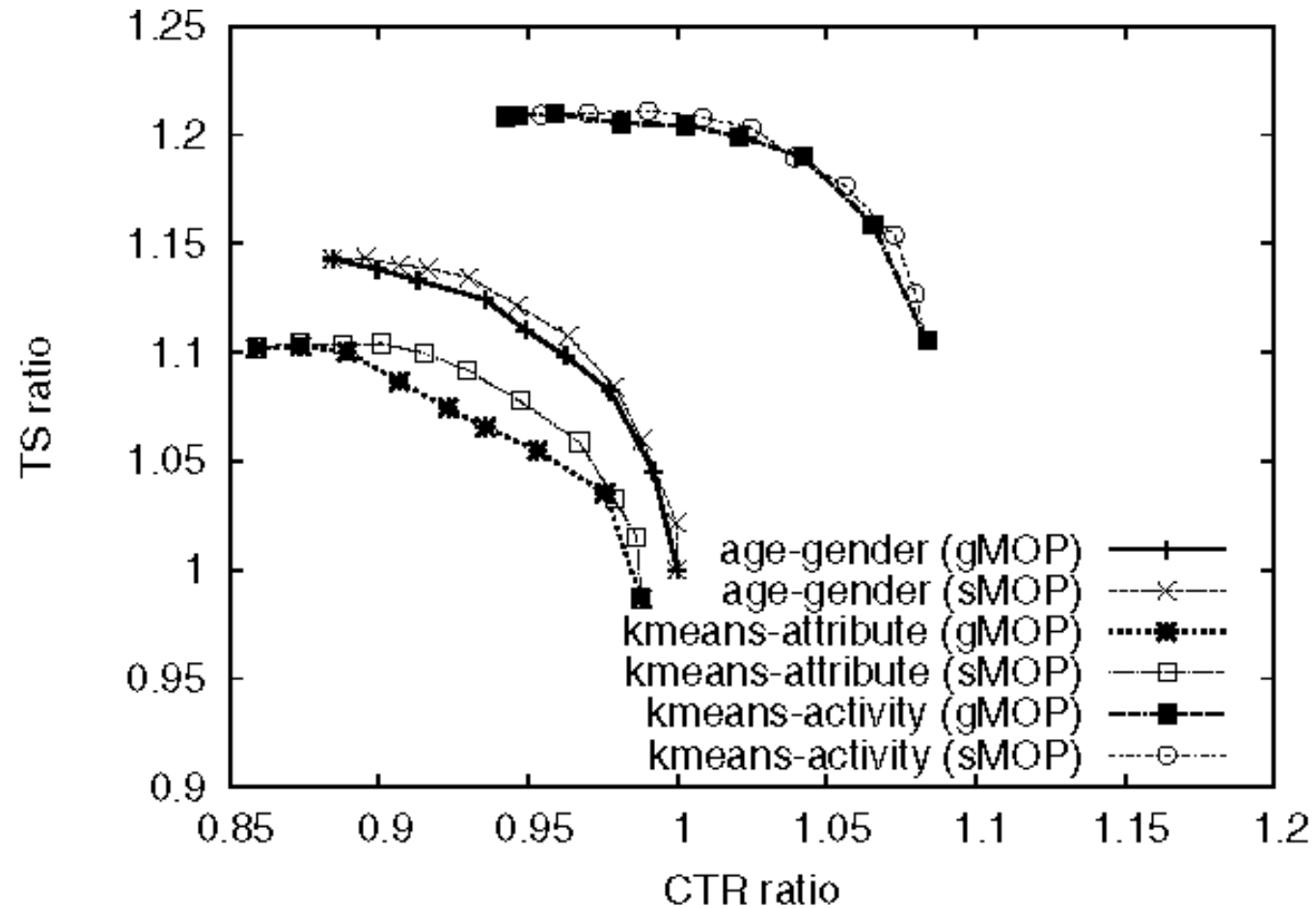
Application of Replay: Personalized Explore-Exploit Algorithms [LCLS'10]



Application of Replay: Effects of Reward Delay [CL'11]



Application of Replay: Multi-objective Optimization [ACEW'11&12]



Recap

- Direct method by estimating $\hat{r}(x, a)$ is inherently biased
- Stationary policies: Inverse propensity Score ensures unbiasedness
 - With easily quantified variance
- Nonstationary policies: Replay method

- Case studies:
 - News recommendation
 - Bing search engine

Enhanced Techniques

Unknown propensity scores

Direct policy optimization

Doubly robust estimation

Bootstrapped replay

Unknown Propensity Scores

- So far we have assumed exploration data $D = \{(x, a, r_a, p_a)\}$
- Sometimes p_a is unavailable
 - Data was generated by multiple deterministic policies ($p_a \equiv 1$ in this case)
“natural exploration”
 - Data loss or contamination (p_a not truthful of real action distribution in data)
 - ...
- Not all hope is lost

IPS with Estimated Propensity Scores

- Data $D = \{(x_1, a_1, r_1), (x_2, a_2, r_2), \dots, (x_L, a_L, r_L)\}$
where $a_t \sim p_t(\cdot | x_t)$ [p_t unknown or deterministic]
- **Assumption**: π_t independent of D
- Define “averaged” distribution $p = \frac{1}{L} (p_1 + p_2 + \dots + p_L)$
- Estimate $\hat{p}(a|x) \approx p(a|x)$
 - Multinomial logistic regression, neural network, decision trees, ...

$$\hat{V}_{ips}(\pi) = \frac{1}{L} \sum_i \frac{r_i \cdot 1(\pi(x_i) = a_i)}{\max\{\hat{p}(a_i|x_i), \tau\}}$$

Avoid division by
tiny numbers

Properties

$$\hat{V}_{ips}(\pi) = \frac{1}{L} \sum_i \frac{r_i \cdot 1(\pi(x_i) = a_i)}{\max\{\hat{p}(a_i|x_i), \tau\}}$$

- Slightly biased
 - τ : Under-estimation since it makes ratio smaller
 - $1/\hat{p}$: Over-estimation
- Variance control
 - τ helps stability (preventing division by tiny numbers)
- Combined [SLLK'10]

$$|E[\hat{V}_{ips}(\pi) - V(\pi)]| \leq E_x \left[\begin{array}{ll} r(x, \pi(x)) & \text{if } p(\pi(x)|x) < \tau \\ \max_a |p(a|x) - \hat{p}(a|x)| / \tau & \text{otherwise} \end{array} \right]$$

Enhanced Techniques

Unknown propensity scores

Direct policy optimization

Doubly robust estimation

Bootstrapped replay

Policy Optimization

- Most often ultimate goal is to find optimal π with maximum $V(\pi)$
- Approach 1: guess and check
 - Offline optimization against MSE/NDCG
 - Online experiment to verify gain in CTR/satisfaction/revenue
- Approach 2: direct solution
 - Offline optimization against $\hat{V}(\pi)$
 - Example: Bing Speller
 - Can be substantially generalized

Classification as Contextual Bandit

- Multi-class, multi-label classification



Action

Comedy

Historical

Thriller

- Example x associated with **subset** of correct labels $c \subseteq L = \{1, 2, \dots, K\}$
 - x ("imitation game") $\rightarrow c$ ({historical, thriller})

Multi-label Classification as Contextual Bandit

- Use classification example (x, c) to simulate interaction in bandit
 - x : context
 - $A = L$: candidate actions
 - $r_a = 1(a \in c)$
 - Essentially, $(x, c) \Rightarrow (x; r_1, r_2, \dots, r_K)$
- Policy π is treated as classifier

$$V(\pi) = E_x[r(x, \pi(x))] = E_x[1(\pi(x) \in c)]$$

Policy value is classification accuracy!

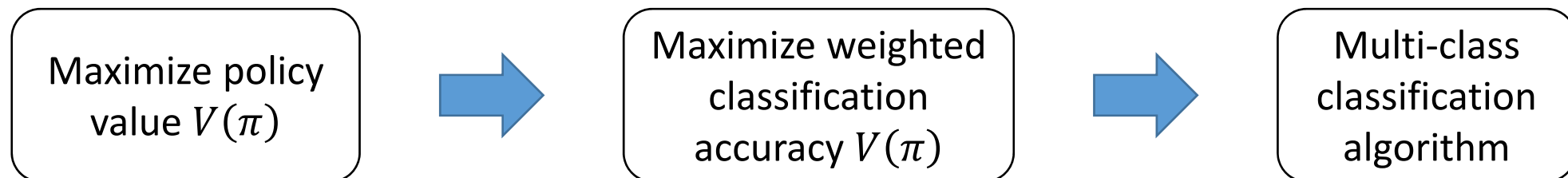
Policy Optimization as Classification

Contextual bandit \rightarrow weighted multi-class classification
 $(x, a, r_a, p_a) \Rightarrow (x, a, w_a) \quad w_a = r_a/p_a$

Same trick as IPS!

$$E_{x,a}[w_a \cdot \mathbf{1}(\pi(x) = a)] = E_x[r(x, \pi(x))] = V(\pi)$$

Policy value is same as weighted classification accuracy!



Offset tree [BL'09]: a similar and sometimes more effective optimization algorithm

Case Study 4: Advertising [SLLK'10]

- Problem: choose ad a for $x = (\text{user}, \text{page})$ to maximize clicks
- Goal: learn from production data a warm-start policy better than random
- Non-exploration data $D = \{(x, a, r_a)\}$
 - 35M impressions for training
 - 19M impressions for test
 - 880K ads
 - 3.4M distinct webpages
 - $r_a \in \{0,1\}$: click or not

Three Algorithms for Comparison

- Random (baseline)
- Naive (supervised learning):
 - Learn scoring function $s(x, a)$ from data D
 - Policy $\pi(x) = \arg \max_a s(x, a)$
- Our approach (addressing bias in data):
 - Estimate propensity scores $\hat{p}(a|x)$ from data D
 - Learn regressor f to minimize $\frac{(r_a - f(x, a))^2}{\max\{\hat{p}(a|x), \tau\}}$
 - Policy $\pi(x) = \arg \max_{a: \hat{p}(a|x) > 0} f(x, a)$

Warm Start Results

| Method | τ | Estimate | Interval |
|---------|--------|----------|-----------------|
| Learned | 0.01 | 0.0193 | [0.0187,0.0206] |
| Random | 0.01 | 0.0154 | [0.0149,0.0166] |
| Learned | 0.05 | 0.0132 | [0.0129,0.0137] |
| Random | 0.05 | 0.0111 | [0.0109,0.0116] |
| Naive | 0.05 | 0.0 | [0,0.0071] |

- Ignoring bias in data, naive supervised learning even worse than random!
- Reasonably strong warm-start policies, even learned from non-exploration data

Enhanced Techniques

Unknown propensity scores

Direct policy optimization

Doubly robust estimation

Bootstrapped replay

Doubly Robust Estimation

- Direct Method (DM)

$$\hat{V}_{dm}(\pi) = \frac{1}{L} \sum \hat{r}(x, \pi(x))$$

Estimate $\hat{r}(x, a) \approx r(x, a)$
Small variance
Large bias

- Inverse Propensity Score (IPS)

$$\hat{V}_{ips}(\pi) = \frac{1}{L} \sum \frac{r_a \cdot \mathbf{1}(\pi(x)=a)}{\hat{p}_a}$$

No or small bias
Large variance if $p_a \approx 0$

- Doubly Robust (DR) [RRZ'94]

$$\hat{V}_{dr}(\pi) = \frac{1}{L} \sum_{(x,a,r_a,\hat{p}_a) \in D} \left(\hat{r}(x, \pi(x)) + \frac{(r_a - \hat{r}(x, \pi(x))) \cdot \mathbf{1}(\pi(x) = a)}{\hat{p}_a} \right)$$

DR: Unbiasedness

$$\hat{V}_{dr}(\pi) = \frac{1}{L} \sum_i \left(\hat{r}(x, \pi(x)) + \frac{(r_a - \hat{r}(x, \pi(x))) \cdot \mathbf{1}(\pi(x) = a)}{\hat{p}_a} \right) \quad \hat{r} = r \implies E[\hat{V}_{dr}] = V(\pi)$$
$$= \frac{1}{L} \sum_i \left(\hat{r}(x, \pi(x)) \left(1 - \frac{\mathbf{1}(\pi(x) = a)}{\hat{p}_a} \right) + \frac{r_a \cdot \mathbf{1}(\pi(x) = a)}{\hat{p}_a} \right) \quad \hat{p} = p \implies E[\hat{V}_{dr}] = V(\pi)$$

- Two ways to ensure unbiasedness (“doubly protected”)
- Implemented in Vowpal Wabbit (<http://hunch.net/~vw>)
- Well-known in statistics, but not entirely satisfying
 - Almost impossible to have $\hat{r} = r$ or $\hat{p} = p$ in reality
 - Refined analysis for practically relevant situations [DLL'11]

DR: Bias Analysis

$$\bullet E[\hat{V}_{dr}] - V(\pi) = E_x \left[\underbrace{\text{err}_p(x)}_{\text{Error in } \hat{p}} \cdot \underbrace{\text{err}_r(x)}_{\text{Error in } \hat{r}} \right]$$

$$\bullet E[\hat{V}_{ips}] - V(\pi) = E_x \left[\text{err}_p(x) \cdot r(x, \pi(x)) \right]$$

$$\bullet E[\hat{V}_{dm}] - V(\pi) = E_x \left[\text{err}_r(x, \pi(x)) \cdot \max_{x,a} \{r(x, a)\} \right]$$

DR has lowest bias
with “reasonable”
 \hat{p} and \hat{r}

DR: Variance Analysis

$$\bullet \text{Var}[\hat{V}_{dr}] \approx \frac{1}{L} E_x \left[\frac{\text{err}_r(x)^2 \cdot (1 - \text{err}_p(x))^2}{p(\pi(x)|x)} \right]$$

$$\bullet \text{Var}[\hat{V}_{ips}] \approx \frac{1}{L} E_x \left[\frac{r(x, \pi(x))^2 \cdot (1 - \text{err}_p(x))^2}{p(\pi(x)|x)} \right]$$

$$\bullet \text{Var}[\hat{V}_{dm}] = \frac{1}{L} \text{Var}_x[\hat{r}(x, \pi(x))]$$

DR has lower variance than IPS with “reasonable” \hat{r}

DM often has low variance, not affected by $p(a|x)$

Case Study 5: UCI datasets [DLL'11]

| | | | | | | | | | |
|-----------------|-------|-------|--------|-----------|-------------|-----------|----------|---------|-------|
| Dataset | ecoli | glass | letter | optdigits | page-blocks | pendigits | satimage | vehicle | yeast |
| Classes (k) | 8 | 6 | 26 | 10 | 5 | 10 | 6 | 4 | 10 |
| Dataset size | 336 | 214 | 20000 | 5620 | 5473 | 10992 | 6435 | 846 | 1484 |

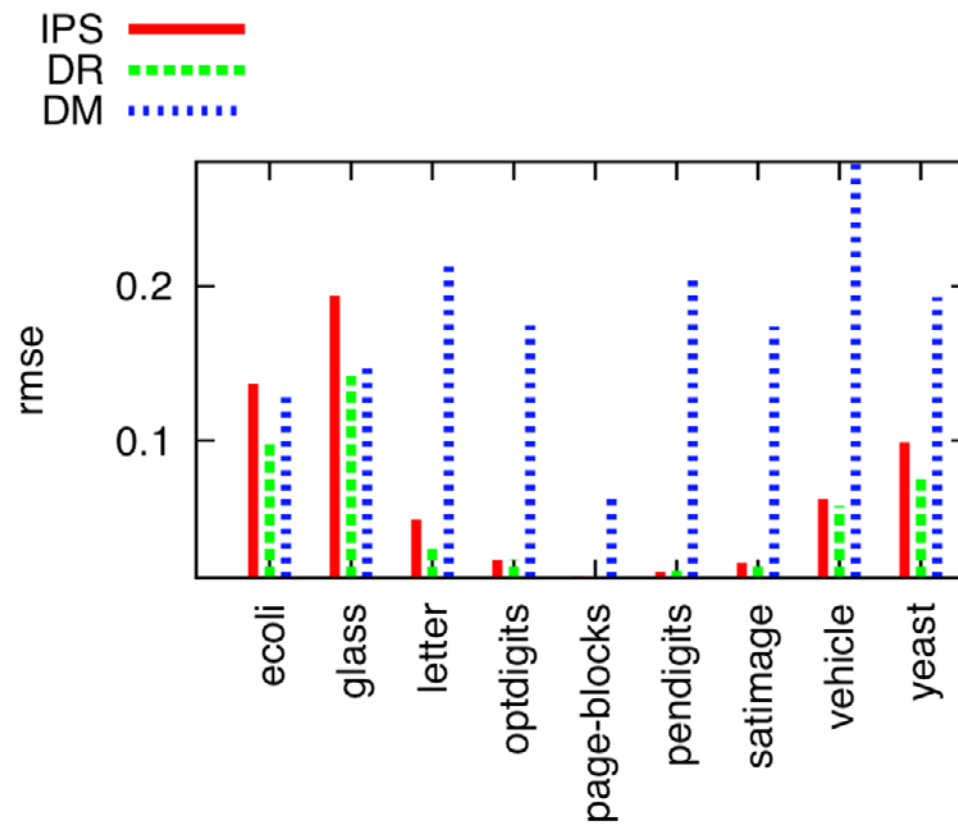
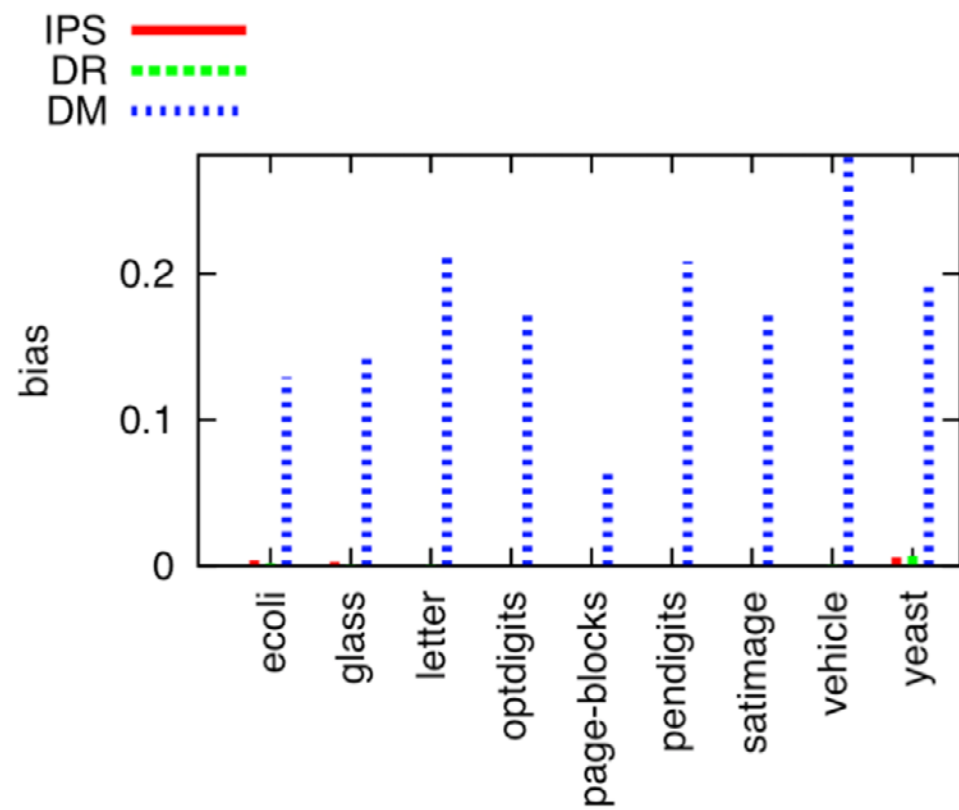
Classification to bandit: $(x, c) \Rightarrow (x; r_1, r_2, \dots, r_K)$

Bandit to classification: $(x, a, r_a, p_a) \Rightarrow (x, a, w_a) \quad w_a = r_a/p_a$

Policy Evaluation

- 50% data for training (**regular** classification) to obtain π
- 50% data for **testing with bandit labels**
 - For each x , randomly pick $a \in \{1, \dots, K\}$ and reveal $r_a = 1(a = c)$
[classification to bandit reduction]
 - Only $1/K$ fraction of labels observed
 - Compare DM, IPS, DR

Policy Evaluation



Policy Optimization

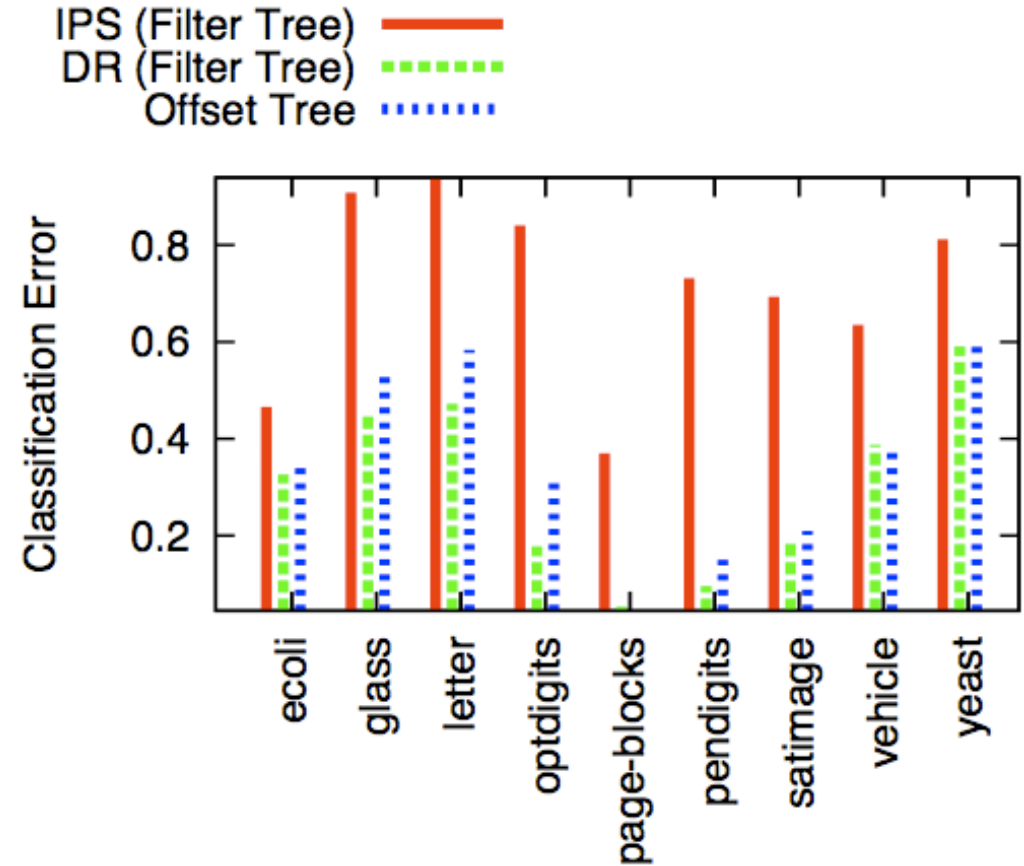
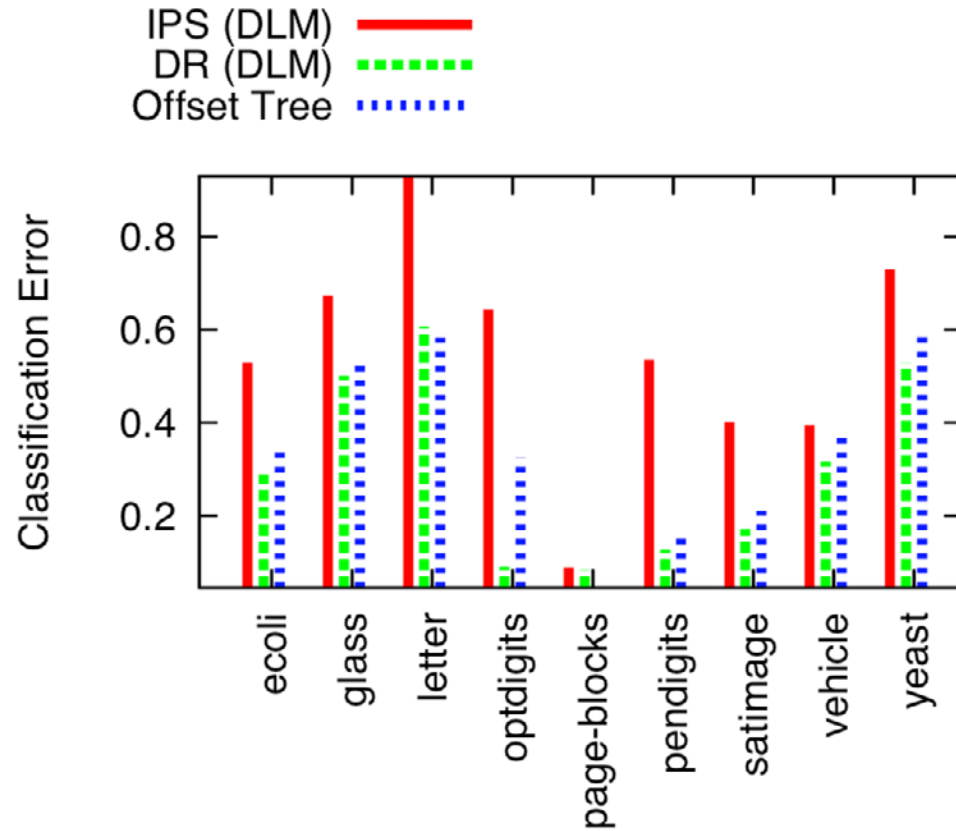
- 70% data for **training with bandit labels** to obtain π
 - For each x , randomly pick $a \in \{1, \dots, K\}$ and reveal $r_a = \mathbf{1}\{a = c\}$
 - Only $1/K$ fraction of labels observed

Optimization algorithms

- Direct loss minimization [MHK'11]
 - Filter tree [BLR'08]
 - Offset tree [BL'09]: alternative policy optimization algorithm
- } Generic multi-class classification
(Combined with DM, IPS, DR)

- 30% data for testing accuracy of π (**regular** classification)

Policy Optimization



Enhanced Techniques

Unknown propensity scores

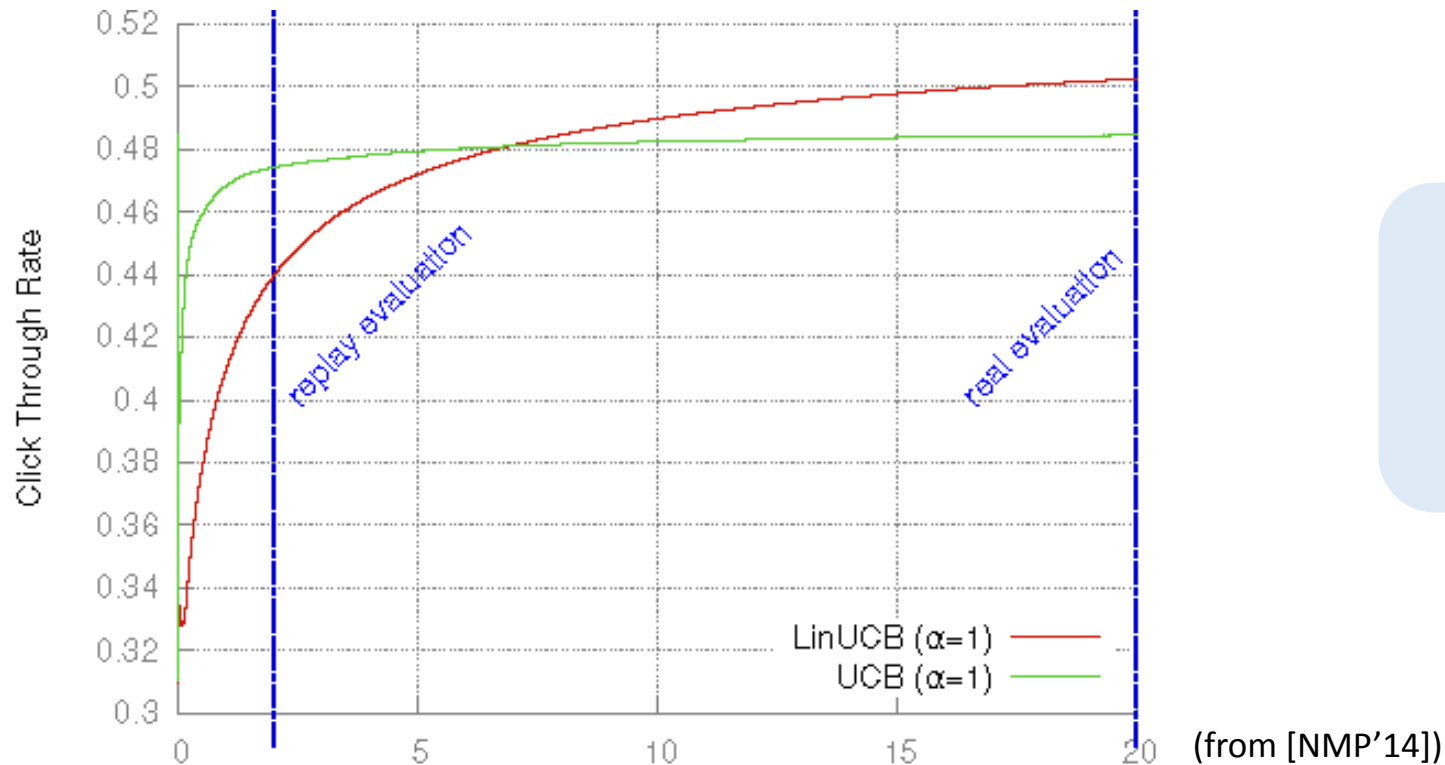
Direct policy optimization

Doubly robust estimation

Bootstrapped replay

Time Acceleration Problem [NMP'14]

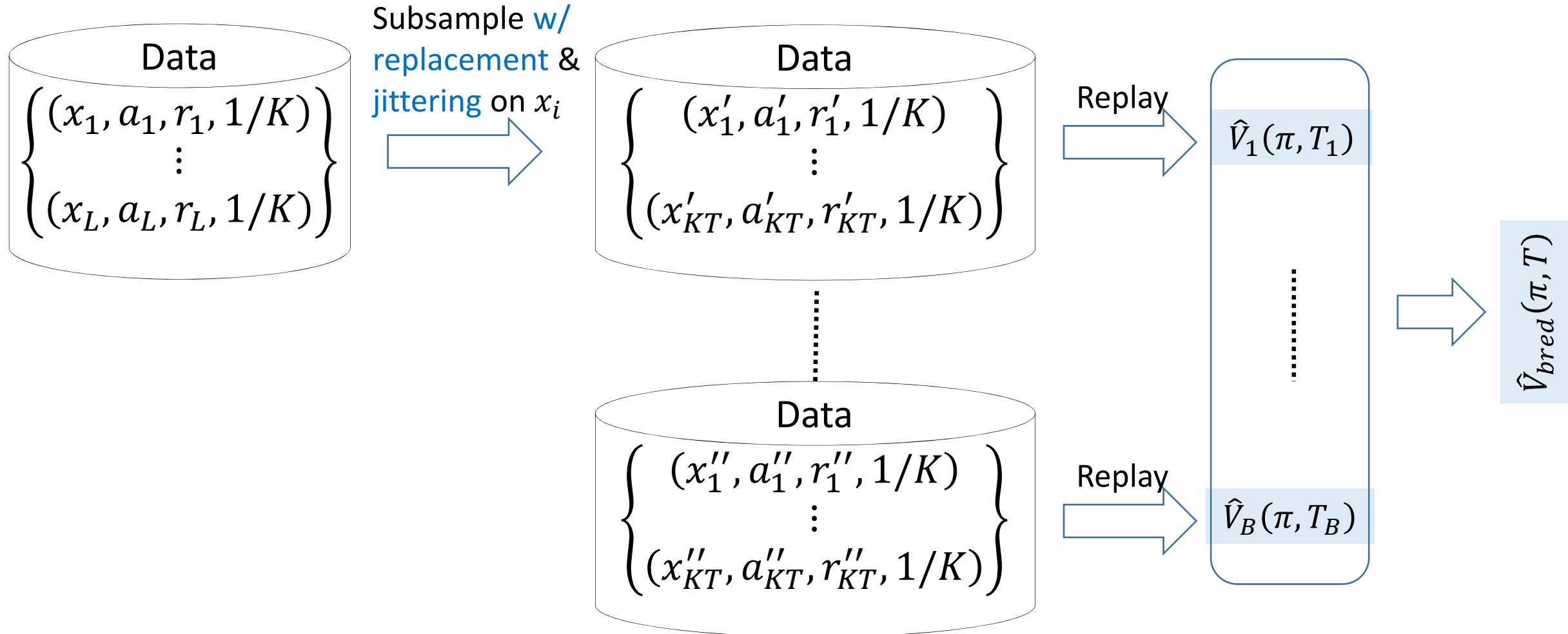
- With $L = |D|$ data and uniform exploration $p_a = 1/K$
 - Expected number of matches is L/K
 - Replay can estimate $V(\pi, T)$ up to $T \approx L/K$



Replay cannot
evaluate π for
too large T

BRED [NMP'14]

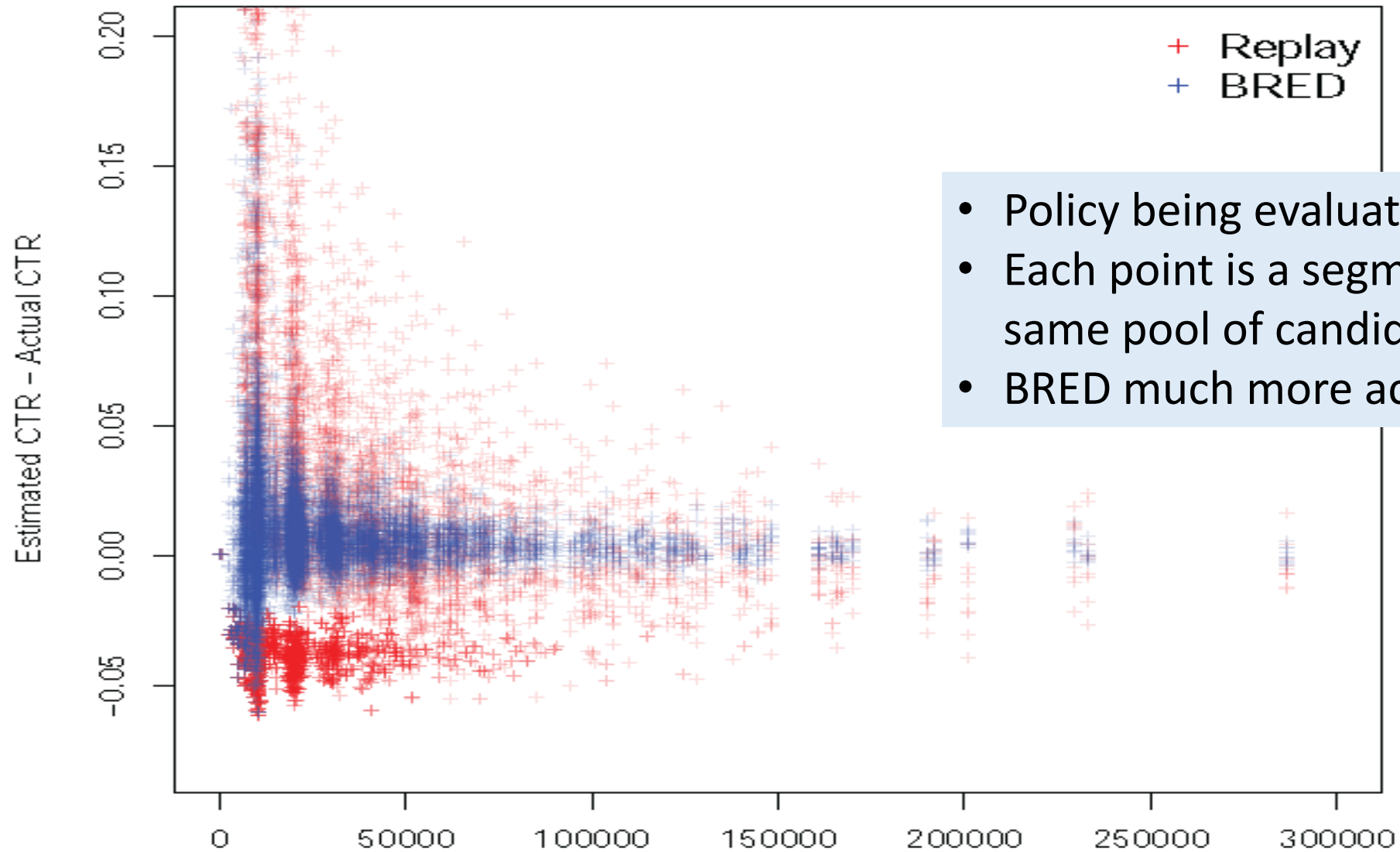
“Bootstrapped Replay on Expanded Data”



BRED Theory

- For stationary policies, confidence intervals are estimated much faster
 - $O(1/T)$ as opposed to $O(1/\sqrt{T})$
 - under mild assumptions (similar to the bootstrap theory)
- For stationary policies, can estimate $V(\pi, T)$ for $T \gg L/K$
 - although the bootstrap theory does not apply
- Practical limitation: computationally expensive
 - fast, approximate bootstrap [OR'01]
 - implemented in Vowpal Wabbit [QPKLL'13]

Replay vs. BRED on Yahoo! News Recommendation

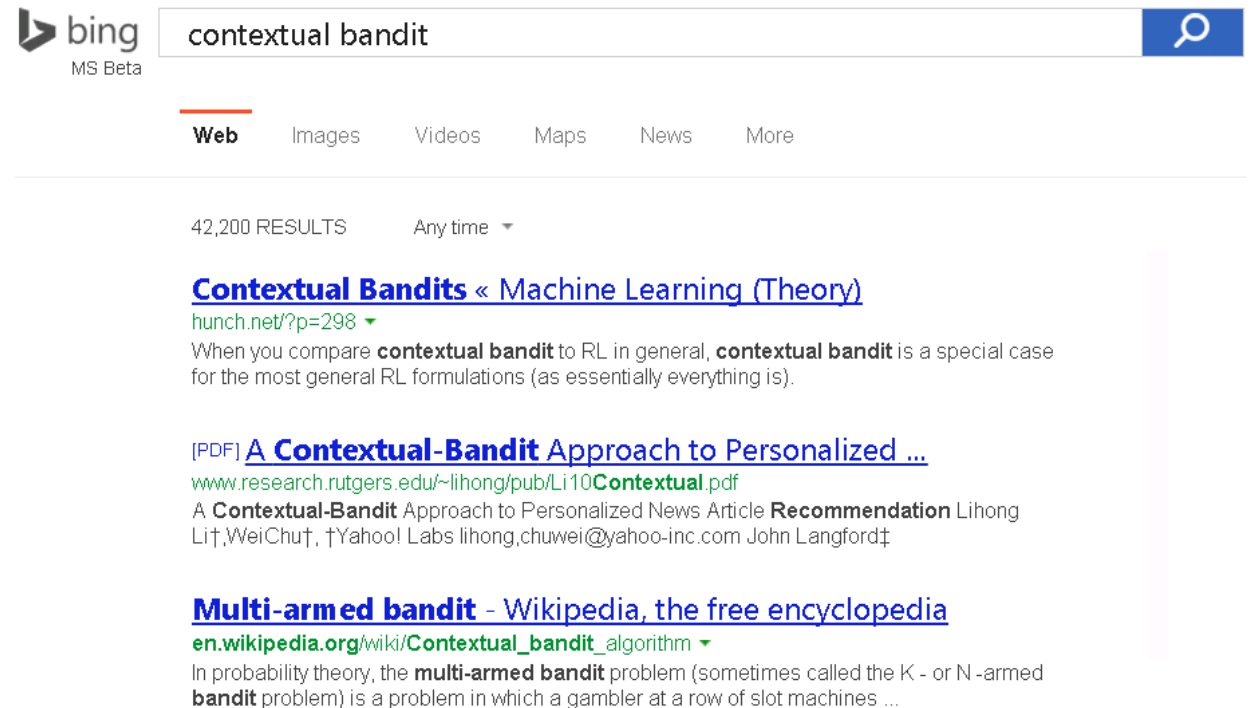


- Policy being evaluated: UCB
- Each point is a segment of data with same pool of candidate articles
- BRED much more accurate

Practical Issues

How to Design Exploration Distributions

- Use of natural exploration (without collecting truly randomized data)
 - Cheap, and potentially useful
 - But risky (by ignoring potential confounding)
- Need to design A properly before collecting data



The image shows a Bing search results page for the query "contextual bandit". The search bar at the top contains the text "contextual bandit" and a magnifying glass icon. Below the search bar, there are navigation tabs for "Web", "Images", "Videos", "Maps", "News", and "More". The "Web" tab is selected. The search results show 42,200 results, with a filter for "Any time". The first result is a link to "Contextual Bandits < Machine Learning (Theory)" from hunch.net, with a snippet: "When you compare **contextual bandit** to RL in general, **contextual bandit** is a special case for the most general RL formulations (as essentially everything is).". The second result is a PDF link titled "A Contextual-Bandit Approach to Personalized ..." from www.research.rutgers.edu, with a snippet: "A **Contextual-Bandit** Approach to Personalized News Article Recommendation Lihong Li†, Wei Chu†, †Yahoo! Labs lihong.chuwei@yahoo-inc.com John Langford‡". The third result is a Wikipedia link titled "Multi-armed bandit - Wikipedia, the free encyclopedia" with a snippet: "In probability theory, the **multi-armed bandit** problem (sometimes called the K - or N -armed **bandit** problem) is a problem in which a gambler at a row of slot machines ...".

How to Design Exploration Distributions (2)

- $Var(\hat{V}(\pi))$ depends on how much π “agree” with p
 - Usually π not known in advance
 - Choice #1: uniform (best in the worst case) [news recommendation]
 - Choice #2: randomize around current/production policy [Speller]
- More exploration with p causes greater potential risk
 - Negative user satisfaction, monetary loss, ...
- May use inner/outer confidence intervals to guide design [B+13]

Best decisions have to be on a case-by-case level

What Information to Log

- Data $D = \{(x, a, r_a, p_a)\}$
- Should log x if possible to avoid inconsistency
 - Eg., x has time-sensitive features
 - Eg., x may be missing due to timeouts
- Should log p_a (unless it's precisely known)
- Should log **immediate** actions (not final actions)



Detecting Data Quality Issues

Data $D = \{(x, a, r, p)\}$

- Mean tests [LCKG'14]

arithmetic: $\forall a': \sum_D 1(a = a') \approx \sum_D p(a'|x)$

harmonic: $\sum_D \frac{1}{p} \approx L \times K$

Use standard t-test
to detect \neq

- Can log randomization seed in D and check offline to detect bugs

Concluding Remarks

Review

General theme: use historical data to offline-discovery online metrics
(estimate causal effects from historical data)

- Policy evaluation/optimization
- Unbiasedness with IPS and Replay
- Variance reduction techniques with DR, etc.
- Case studies in news, search, advertising, and benchmark

More Bing Examples



contextal bandit



Web

Images

Videos

Maps

News

More

317

42,000 RESULTS

Any time ▾

Including results for **contextual bandit**.

Do you want results only for contextual bandit?

Contextual Bandits « Machine Learning (Theory)

hunch.net/?p=298 ▾

When you compare **contextual bandit** to RL in general, **contextual bandit** is a special case for the most general RL formulations (as essentially everything is).

[PDF] **A Contextual-Bandit Approach to Personalized ...**

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‡

Multi-armed bandit - Wikipedia, the free encyclopedia

Related searches

Contextual Bandit Learning

Contextual Bandit Problem

Multiworld Testing

Multi Armed **Bandits**

Vowbal Wabbit Machine Learning

Bandit Learning

Bandit Problem

Bandit Algorithms

More Bing Examples



contextual bar

Web Images

sant

- Santa Clara
California, United States
- Santa Rosa
California, United States
- Santa Barbara
California, United States
- Santa Monica
California, United States
- Santa Fe
New Mexico, United States

San Francisco

42,000 RESULTS

Including results
Do you want results

Contextual I

hunch.net/?p=298
When you compare
for the most genera

[PDF] **A Conte**

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‡

[Multi-armed bandit - Wikipedia, the free encyclopedia](#)

sant

- Santiago
Santiago Metropolitan, Chile
- Santa Mónica
Santiago, Santiago Metropolitan, Chile
- Santa Elena Sur
Huechuraba, Santiago Metropolitan, Chile
- Santa Lucia Hill
Santiago, Santiago Metropolitan, Chile
- Santa Barbara
California, United States

Santiago

Bandit Problem

Bandit Algorithms

Many More Applications

- Yahoo!, Google, Microsoft, LinkedIn, Adobe, Criteo, ...
[LP'07] [LSW'08] [CGGHL'10] [PPBK'11] [ACEW'11] [TRSA'13] [A+'14] ...
- Can be combined with other methods like interleaving [HWR'12&14]
- WWW 2015 Workshop in May (Florence, Italy)
<http://evalworkshop.com>
- Datasets available at Yahoo! Webscope (R6B)
<http://webscope.sandbox.yahoo.com/catalog.php?datatype=r>

Limitations and Open Questions

- Many actions
 - Relies on natural exploration and approximate matching [LKZ'15]
 - Use production data to approximate online behavior [YBL'15]
 - Continuous actions [B+'13]
- Cannot model long-term effects
 - Off-policy reinforcement learning
 - Equilibrium analysis [B+'13]
- Relies on stationary assumption
- Statistically more efficient (even optimal) offline estimation

References

- **[A+'14]** Deepak Agarwal, Bee-Chung Chen, Rupesh Gupta, Joshua Hartman, Qi He, Anand Iyer, Sumanth Kolar, Yiming Ma, Pannagadatta Shivaswamy, Ajit Singh, Liang Zhang: Activity ranking in LinkedIn feed. KDD 2014: 1603-1612
- **[ACEW'11]** Deepak Agarwal, Bee-Chung Chen, Pradheep Elango, Xuanhui Wang: Click shaping to optimize multiple objectives. KDD 2011: 132-140
- **[ACEW'12]** Deepak Agarwal, Bee-Chung Chen, Pradheep Elango, Xuanhui Wang: Personalized click shaping through lagrangian duality for online recommendation. SIGIR 2012: 485-494
- **[B+'13]** Léon Bottou, Jonas Peters, Joaquin Quiñonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y. Simard, Ed Snelson: Counterfactual reasoning and learning systems: the example of computational advertising. Journal of Machine Learning Research 14(1): 3207-3260 (2013)
- **[BA'85]** Andrew G. Barto, P. Anandan: Pattern-recognizing stochastic learning automata. IEEE Transactions on Systems, Man, and Cybernetics 15(3): 360-375 (1985)
- **[BL'09]** Alina Beygelzimer, John Langford: The offset tree for learning with partial labels. KDD 2009: 129-138
- **[MCLZC'12]** Taesup Moon, Wei Chu, Lihong Li, Zhaohui Zheng, Yi Chang: An online learning framework for refining recency search results with user click feedback. ACM Trans. Inf. Syst. 30(4): 20 (2012)
- **[CGGHL'10]** David Chan, Rong Ge, Ori Gershony, Tim Hesterberg, Diane Lambert: Evaluating online ad campaigns in a pipeline: Causal models at scale. KDD 2010: 7-16
- **[CJRY'12]** Olivier Chapelle, Thorsten Joachims, Filip Radlinski, Yisong Yue: Large-scale validation and analysis of interleaved search evaluation. ACM Trans. Inf. Syst. 30(1): 6 (2012)
- **[CL'11]** Olivier Chapelle, Lihong Li, An empirical evaluation of Thompson sampling. NIPS 2011: 2249-2257
- **[CSSL'08]** Joaquin Quiñonero Candela, Masashi Sugiyama, Anton Schwaighofer, Neil D. Lawrence, editors. Dataset Shift in Machine Learning. MIT Press, 2008.
- **[DELL'12]** Miroslav Dudík, Dumitru Erhan, John Langford, Lihong Li, Sample-efficient Nonstationary-policy evaluation for contextual bandits. UAI 2012: 247-254

- **[DELL'14]** Miroslav Dudík, Dumitru Erhan, John Langford, Lihong Li, Doubly robust policy evaluation and optimization. *Statistical Science* 29(4):485-511 (2014)
- **[DLL'11]** Miroslav Dudík, John Langford, Lihong Li, Doubly robust policy evaluation and learning. *ICML 2011*: 1097-1104
- **[H'79]** James J. Heckman. Sample selection bias as a specification error. *Econometrica* 47(1):153-161 (1979)
- **[H'86]** Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association* 81(6):945–960 (1986)
- **[HWR'12]** Katja Hofmann, Shimon Whiteson, Maarten de Rijke: Estimating interleaved comparison outcomes from historical click data. *CIKM 2012*: 1779-1783
- **[HWR'13]** Katja Hofmann, Anne Schuth, Shimon Whiteson, Maarten de Rijke: Reusing historical interaction data for faster online learning to rank for IR. *WSDM 2013*: 183-192
- **[HWR'13a]** Katja Hofmann, Shimon Whiteson, Maarten de Rijke: Fidelity, Soundness, and Efficiency of Interleaved Comparison Methods. *ACM Trans. Inf. Syst.* 31(4): 17 (2013)
- **[KLSH'09]** Ron Kohavi, Roger Longbotham, Dan Sommerfield, Randal M. Henne: Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Discov.* 18(1): 140-181 (2009)
- **[LCLS'10]** Lihong Li, Wei Chu, John Langford, Robert E. Schapire: A contextual-bandit approach to personalized news article recommendation. *WWW 2010*: 661-670
- **[LCLW'11]** Lihong Li, Wei Chu, John Langford, Xuanhui Wang: Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. *WSDM 2011*: 297-306
- **[LCKG'14]** Lihong Li, Shunbao Chen, Jim Kleban, Ankur Gupta, Counterfactual estimation and optimization of click metrics for search engines, no. MSR-TR-2014-32, March 2014
- **[LKZ'15]** Lihong Li, Jinyoung Kim, Imed Zitouni: Toward predicting the outcome of an A/B experiment for search relevance. *WSDM 2015*: 37-46
- **[LMS'15]** Lihong Li, Remi Munos, Csaba Szepesvari. Toward minimax off-policy value estimation. *AISTATS 2015*

- **[LP'07]** Diane Lambert, Daryl Pregibon: More bang for their bucks: assessing new features for online advertisers. SIGKDD Explorations 9(2): 100-107 (2007)
- **[LSW'08]** John Langford, Alexander L. Strehl, Jennifer Wortman: Exploration scavenging. ICML 2008: 528-535
- **[LZ'08]** John Langford, Tong Zhang: The Epoch-Greedy Algorithm for multi-armed bandits with side information. NIPS 2007: 817-824
- **[NMP'14]** Jérémie Mary, Philippe Preux, Olivier Nicol: Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques. ICML 2014: 172-180
- **[OR'11]** Nikunj C. Oza, Stuart Russell. Online bagging and boosting. In AISTATS, 2001.
- **[P'09]** Judea Pearl. Causality: Models, Reasoning and Inference (2nd edition). Cambridge University Press, 2009.
- **[PPBK'11]** Ashok Kumar Ponnuswami, Kumaresh Pattabiraman, Desmond Brand, Tapas Kanungo. Model characterization curves for federated search using click-logs: Predicting user engagement metrics for the span of feasible operating points. WWW 2011: 67-76
- **[PSS'00]** Doina Precup, Richard S. Sutton, Satinder P. Singh: Eligibility traces for off-policy policy evaluation. ICML 2000: 759-766.
- **[QPKLL'13]** Zhen Qin, Vaclav Petricek, Nikos Karampatziakis, Lihong Li, John Langford, Efficient online bootstrapping for large scale learning, no. MSR-TR-2013-132, December 2013
- **[R'74]** Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology 66(5):688-701 (1974)
- **[RRZ'94]** James M. Robins, Andrea Rotnitzky, Lue P. Zhao. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89:846-866 (1994)
- **[SLLK'10]** Alexander L. Strehl, John Langford, Lihong Li, Sham Kakade: Learning from Logged Implicit Exploration Data. NIPS 2010: 2217-2225
- **[TRSA'13]** Liang Tang, Rómer Rosales, Ajit Singh, Deepak Agarwal: Automatic ad format selection via contextual bandits. CIKM 2013: 1587-1594
- **[YBL'15]** Dragomir Yankov, Pavel Berkhin, Lihong Li. Evaluation of explore-exploit policies in multi-result ranking systems. Working paper.