wsdm 2015
Shanghai | Jan 31 - Feb 6
The Eighth ACM International Conference on Web Search and Data Mining

Offline Evaluation and Optimization for Interactive Systems

Lihong Li

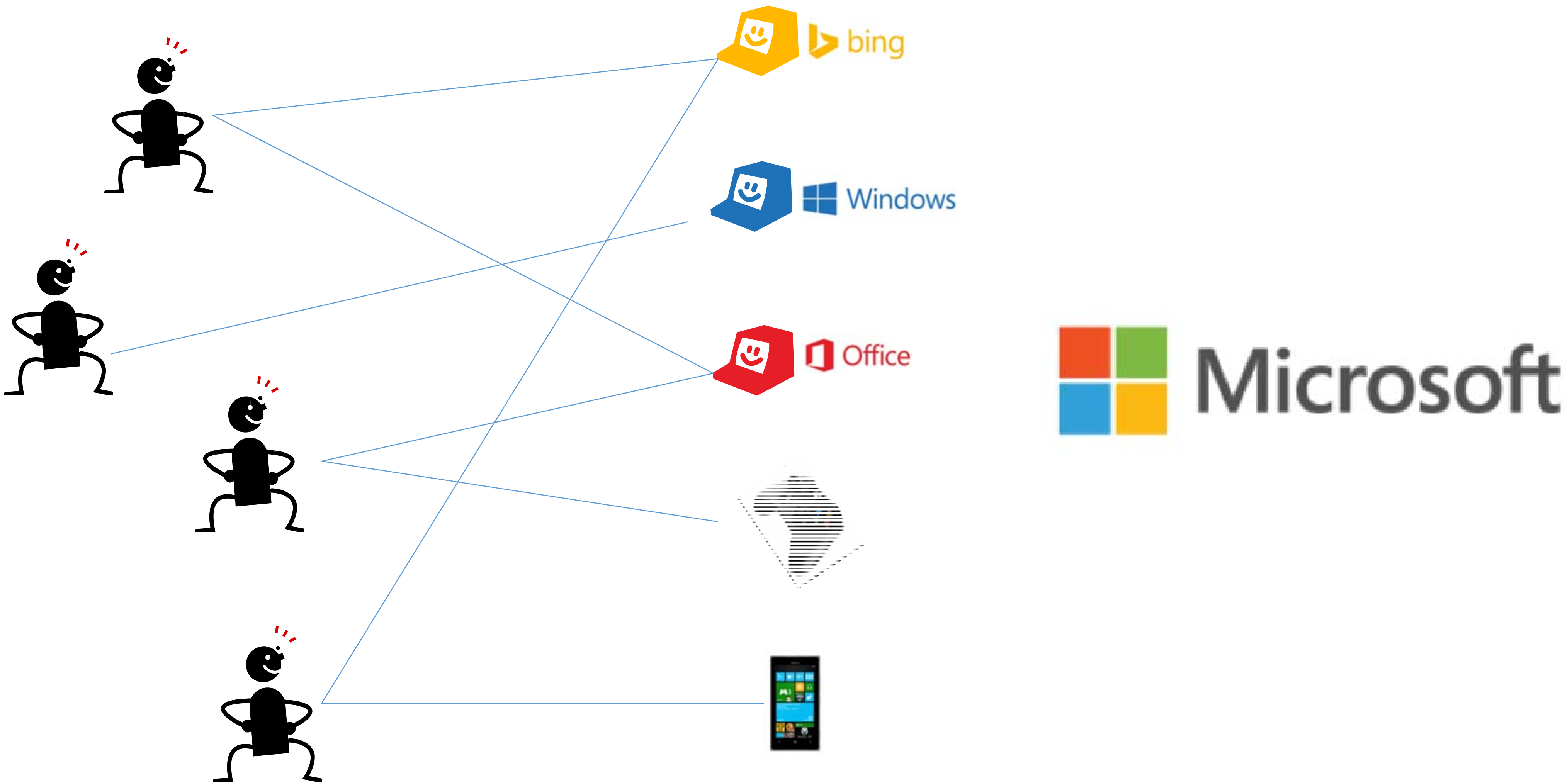
Microsoft Research

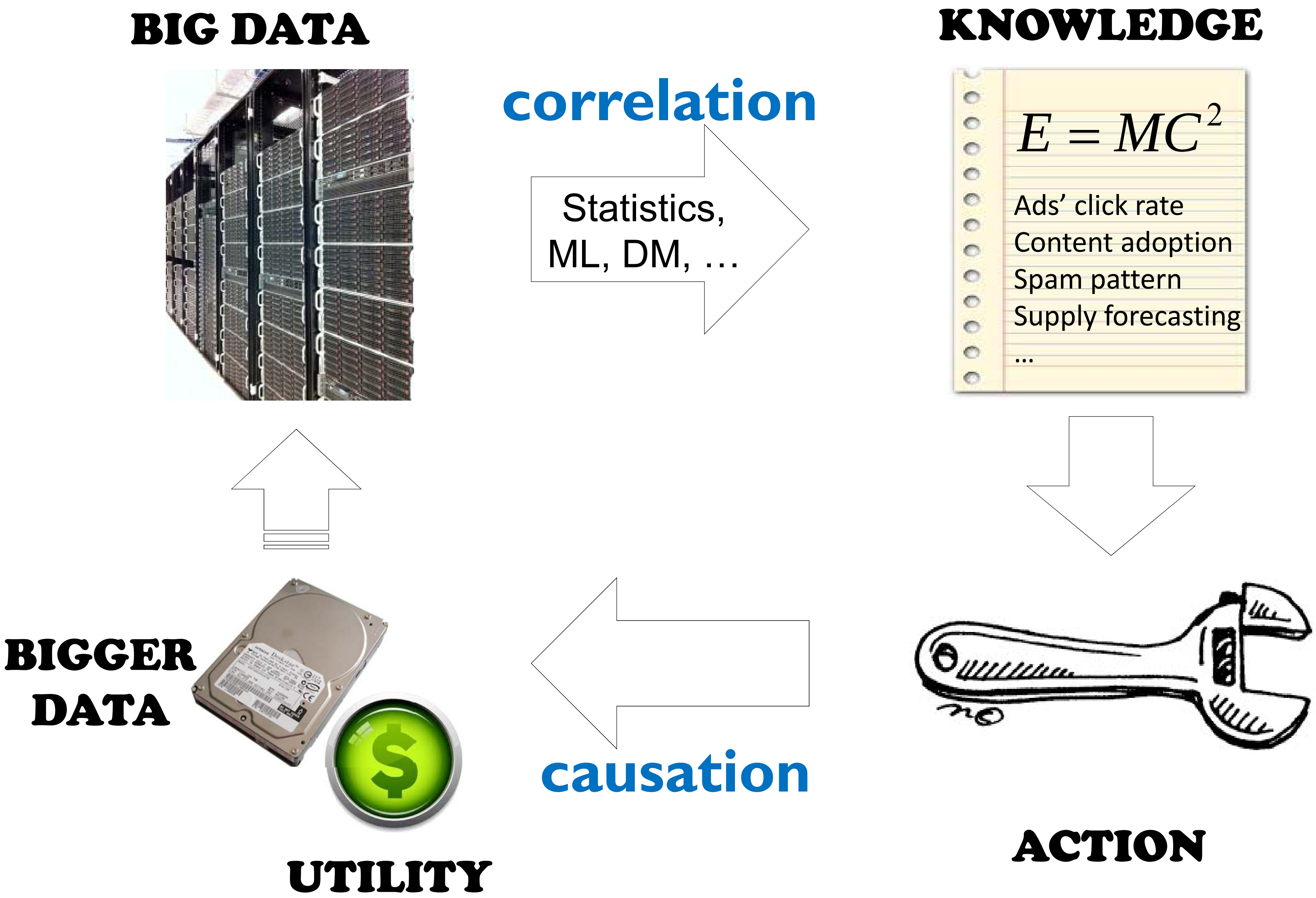
http://research.microsoft.com/en-us/people/lihongli

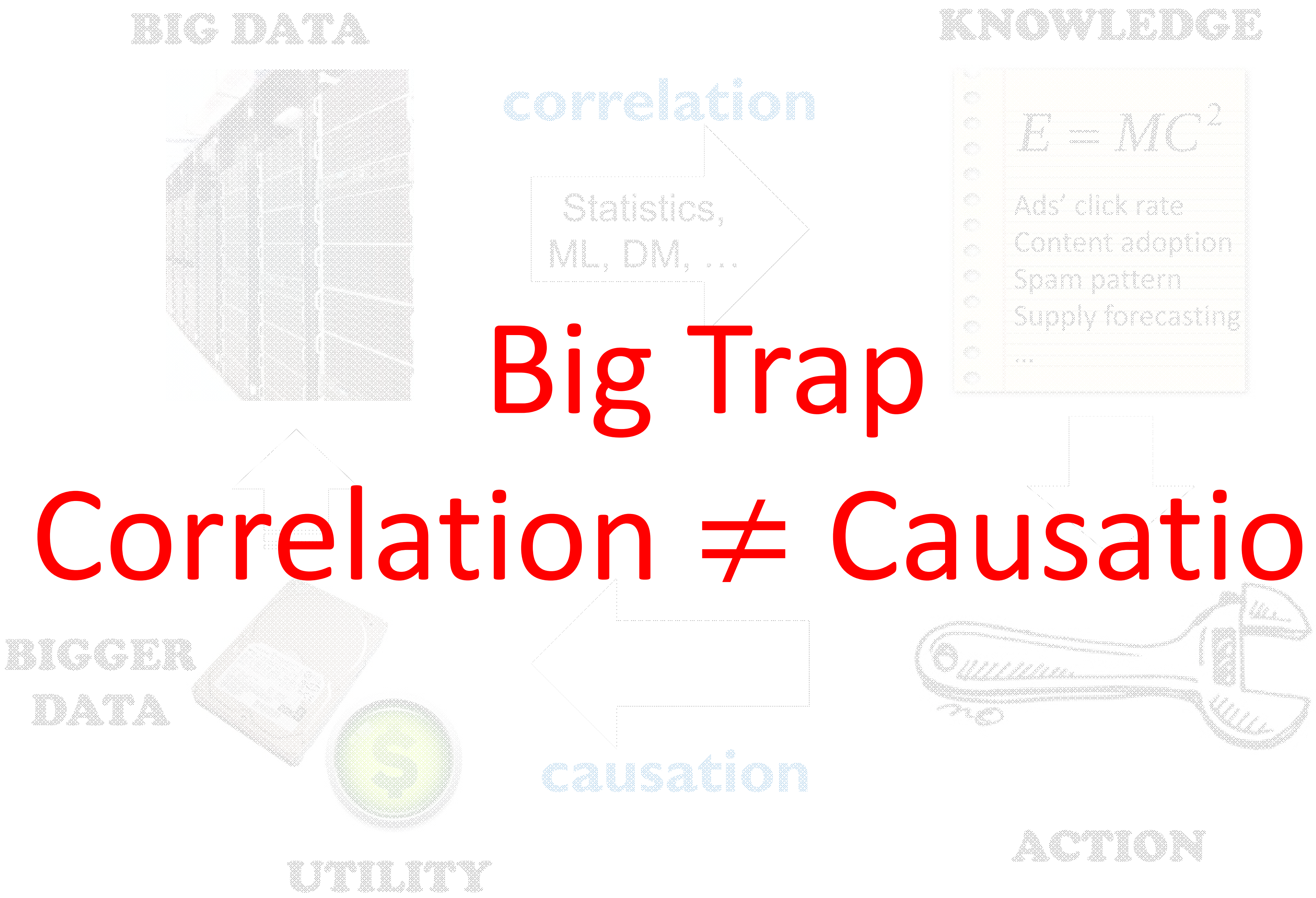
Tutorial URL

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

User Interaction




BIG DATA
correlation
Statistics,
ML, DM, …
KNOWLEDGE
$E = MC^2$
Ads' click rate
Content adoption
Spam pattern
Supply forecasting
…
ACTION
causation
BIGGER
DATA
UTILITY



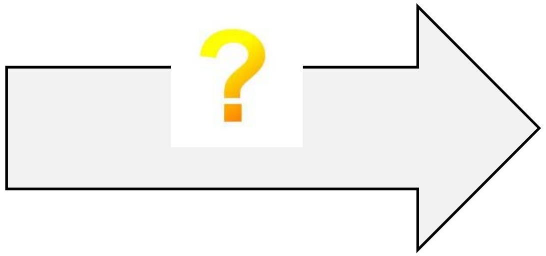
Big Trap

Correlation ≠ Causation

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"

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

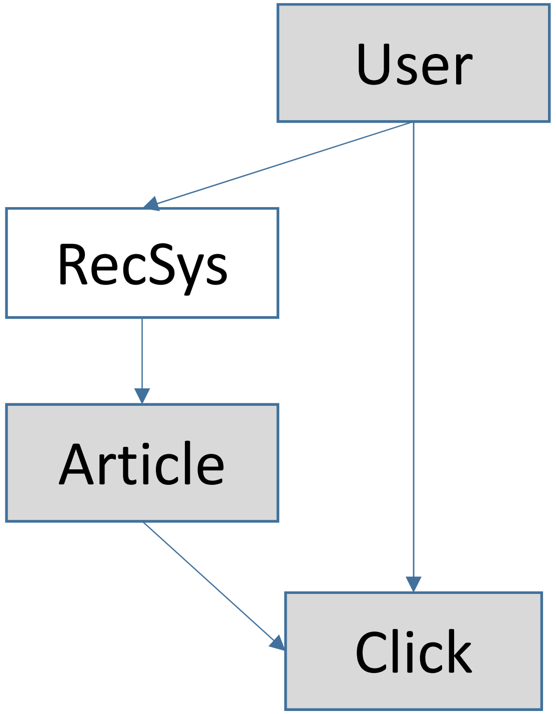
$$0.6 = 0.3 \times \frac{1}{4} + 0.7 \times \frac{3}{4}$$

- 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?



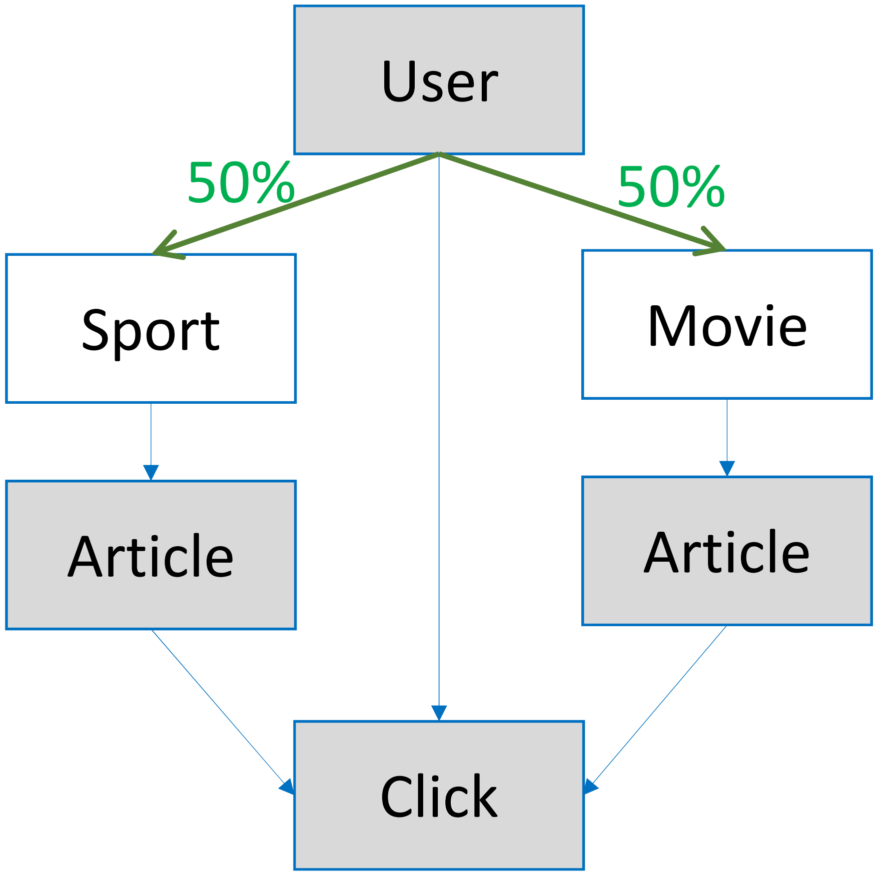
"causal effect"
Can I increase CTR
if I change RecSys?
"manipulation"

<u>Metrics</u>
Precision, Recall,
MSE, NDCG, …

<u>Metrics</u>
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 $\not\Rightarrow$ 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]

Observe K "actions" A_t and "context" x_t → Follow "policy" π to choose $a_t \in A_t$ → Receive "reward" $r_t \in [0,1]$

$t \leftarrow t + 1$

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

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

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

Non-stationary policy: $a_t = \pi(\underbrace{x_1, a_1, r_1, \dots, x_{t-1}, a_{t-1}, r_{t-1}, x_t}_{\text{historical data up to time } t})$

(e.g., online learning algorithms)

Contextual Bandit Applications

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

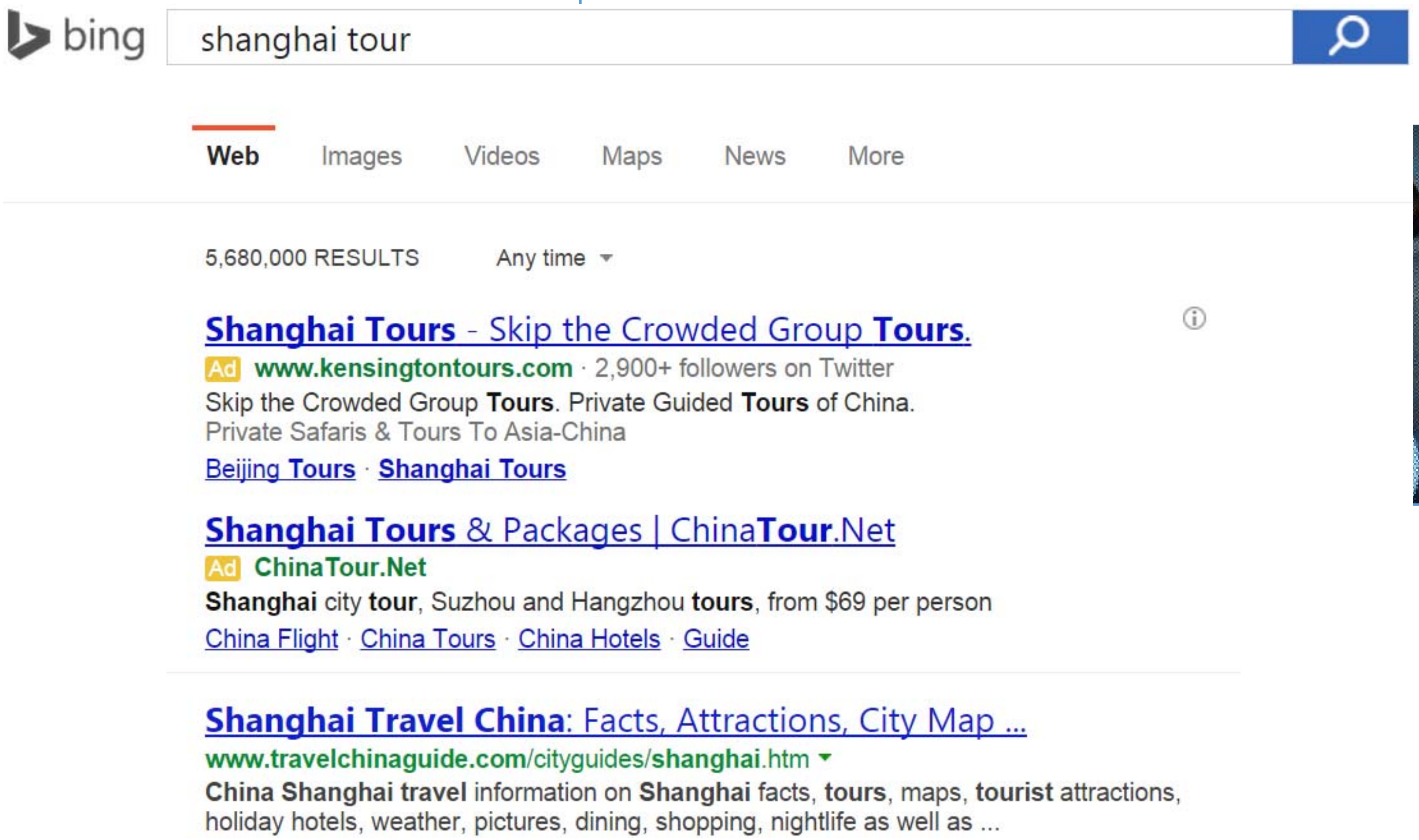
Example: Personalized News Recommendation

www.yahoo.com

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

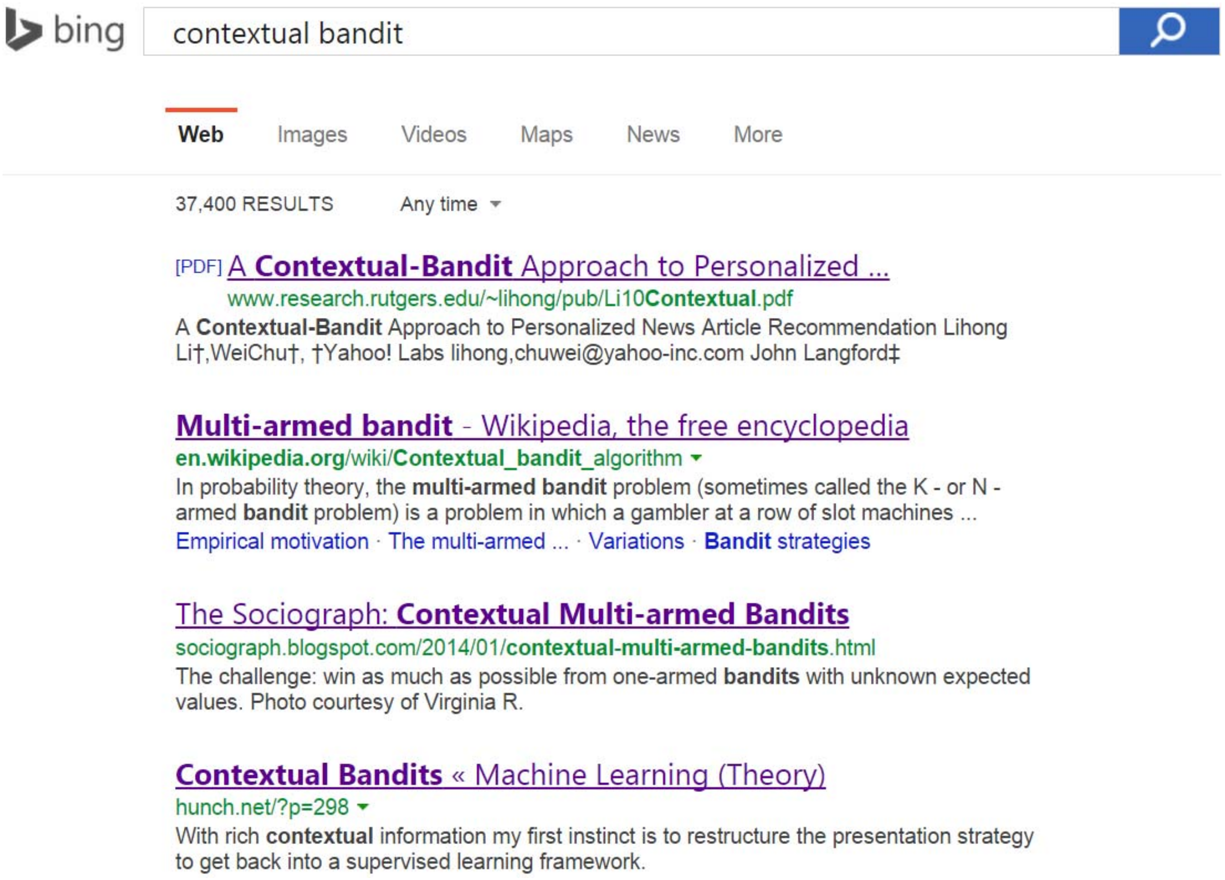
Policy value $V(\pi)$ is click-through rate (CTR)

Example: Online Advertising



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

Example: Web Search Ranking



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 $\pi^* = \mathrm{argmax}_\pi \mathrm{V}(\pi)$

- 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 $\mathrm{V}(\pi, T) = \mathbf{E}\left[\frac{1}{T}(r_1 + r_2 + \cdots r_T)\right]$
- 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: $\pi^* = \mathrm{argmax}_\pi \mathrm{V}(\pi)$

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)

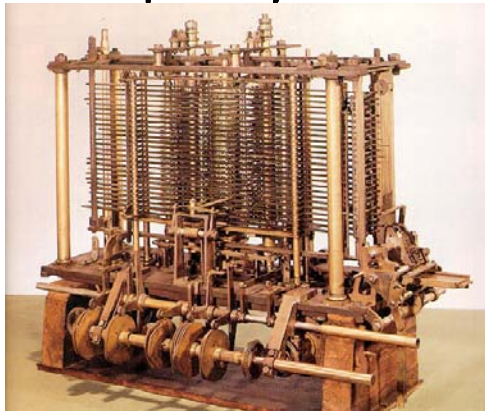
Data

$$\left\{\begin{matrix}(x_1, a_1, r_1) \\ \vdots \\ (x_L, a_L, r_L)\end{matrix}\right\}$$

classification
regression
density estimation

Reward simulator:
$\hat{r}(\mathbf{x}, a) \approx \mathbf{E}[r_a|\mathbf{x}]$

policy π



this (difficult) step is often biased

unreliable evaluation

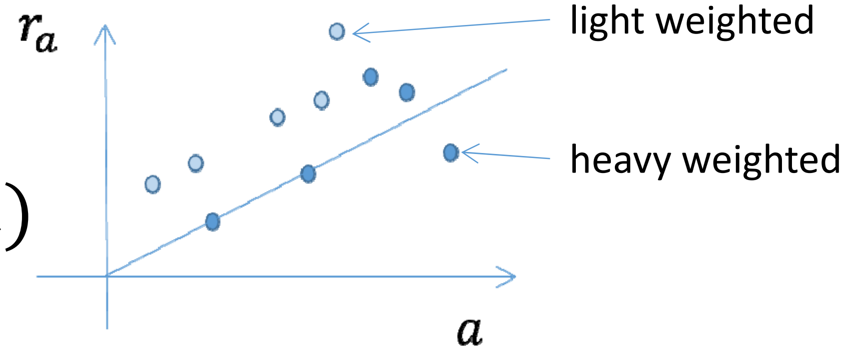
$$\hat{V}_{dm}(\pi) = \frac{1}{L}\sum_i \hat{r}(x_i, \pi(x_i))$$

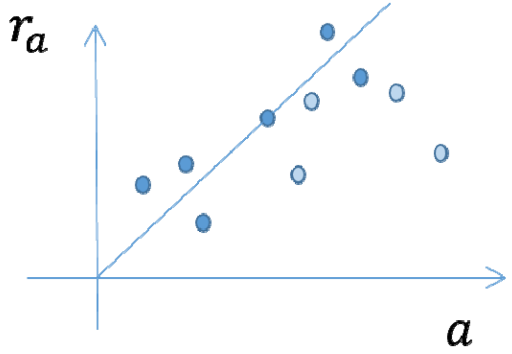
Biases of Direct Method

- Sampling/selection bias
  - From production systems
  - Simpson's paradox

	Overall	Male	Female
Sport	0.5	0.4	0.8
Movie	0.6	0.3	0.7

- Modeling bias
  - Insufficient features to fully represent $r(x, a)$





Neither issue goes away even with infinite data!

Usually difficult to quantify modeling bias!

Randomized Data Collection

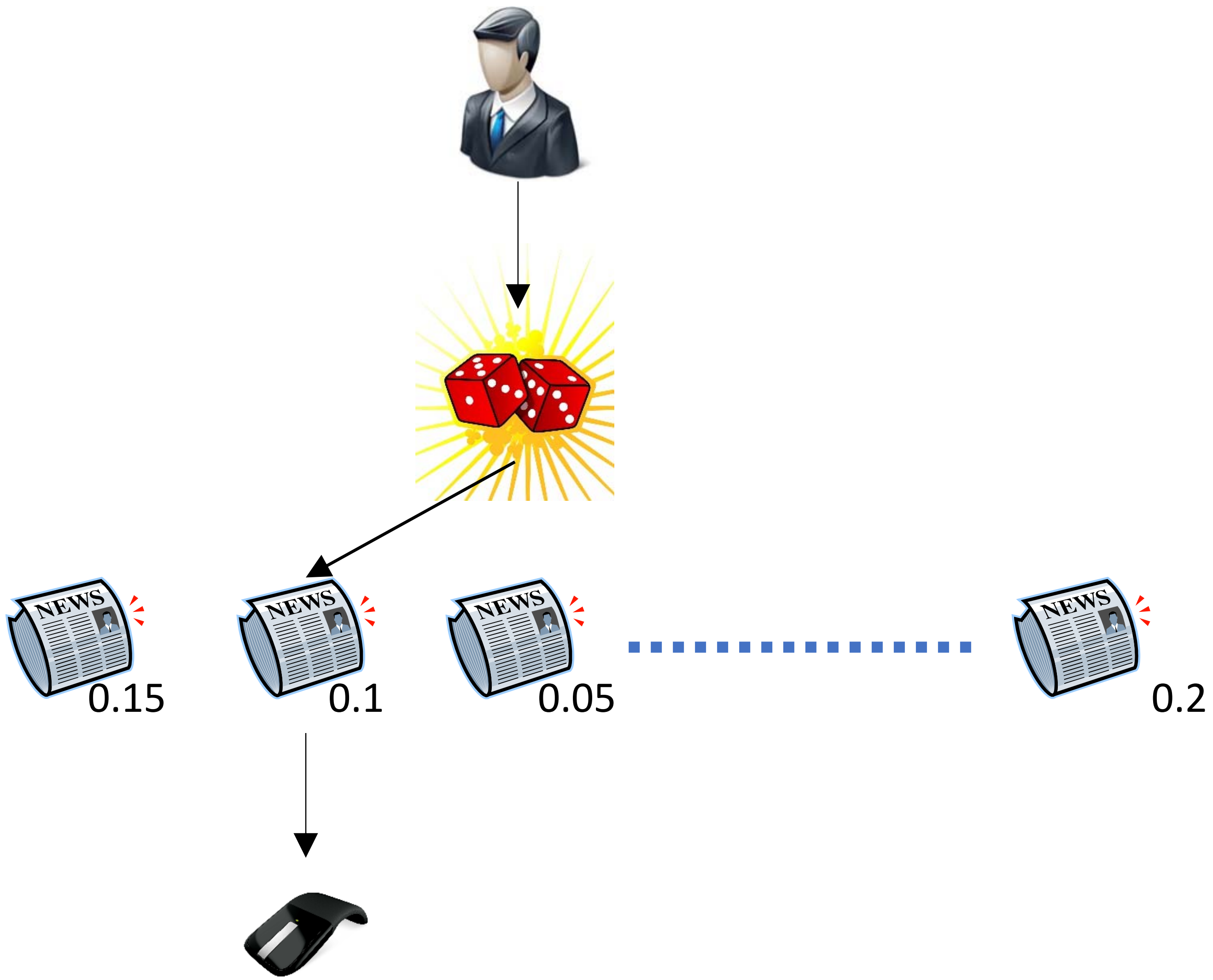
Randomized data collection: at step t,

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

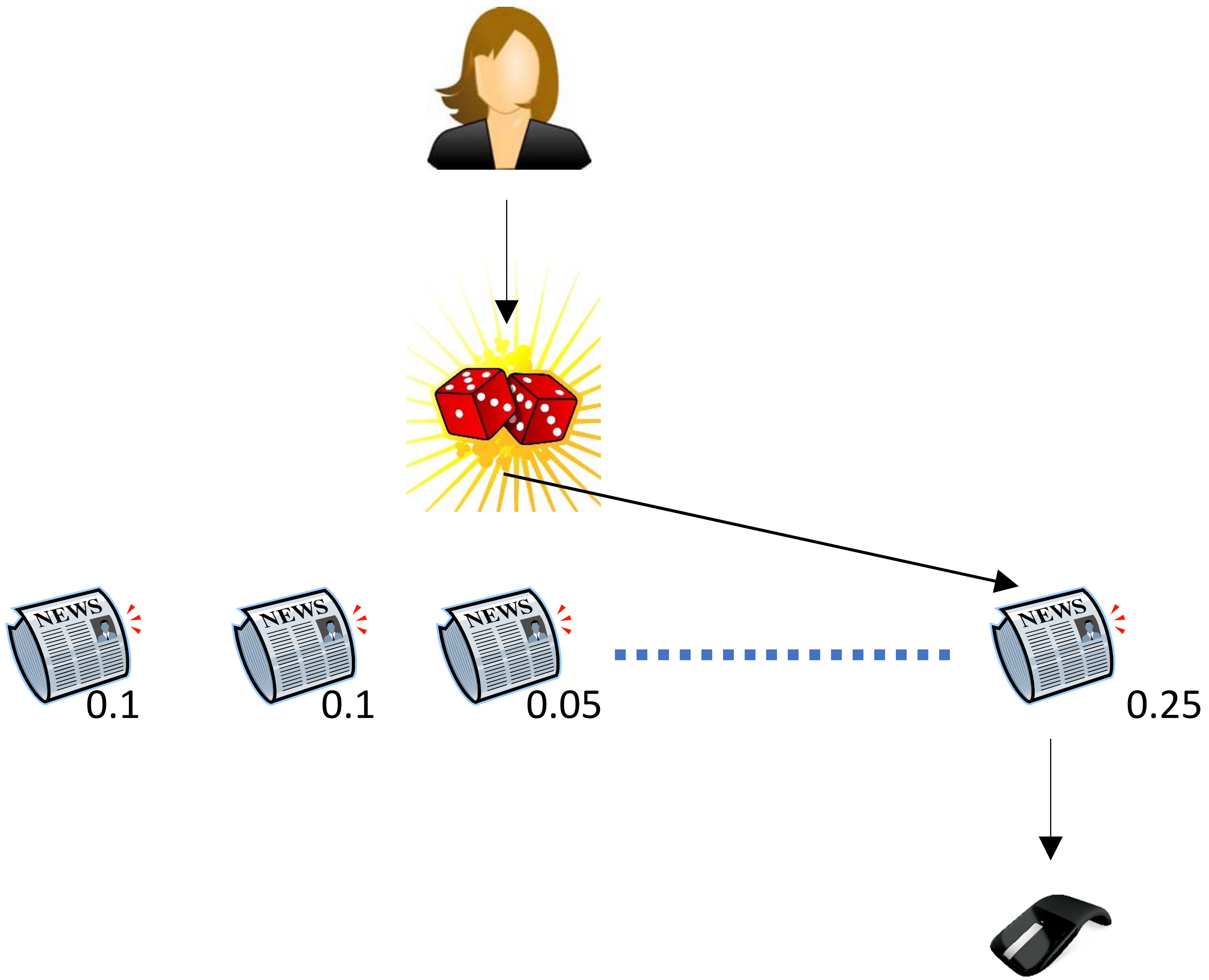
End result: "exploration data" $D = \{(\mathbf{x}, a, r_a, p_a)\}$

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

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

Indicator function:
1 if TRUE, 0 if FALSE

"propensity score"

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

Proof:

$$E\left[\hat{V}(\pi)\right] = E\left[\frac{r_a \cdot \mathbf{1}(\pi(x)=a)}{p_a}\right]$$

$$= E\left[\sum_a \left(p_a \times \frac{r_a}{p_a} 1(\pi(x) = a)\right)\right]$$

$$= E\left[\sum_a \left(r_a \times 1(\pi(x) = a)\right)\right]$$

$$= E_x\left[r_{\pi(x)}\right] = V(\pi)$$

Confidence Interval Estimation for IPS

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

- Consistency: if p_a is not too small, $\hat{V}_{ips}$ converges to $V(\pi)$ as $L \to \infty$
- Variance: $Var\left[\hat{V}_{ips}(\pi)\right] = \frac{1}{L} Var\left[\frac{r_a \cdot 1(\pi(x)=a)}{p_a}\right]$
- 95% confidence interval

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

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

Just another simple random variable

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

$$A = \{\text{Finace}, \text{Movie}, \text{Sport}\}$$

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

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

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

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

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

Case Study 1: News Recommendation [LCLW'11]

A_t : available articles at time t

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

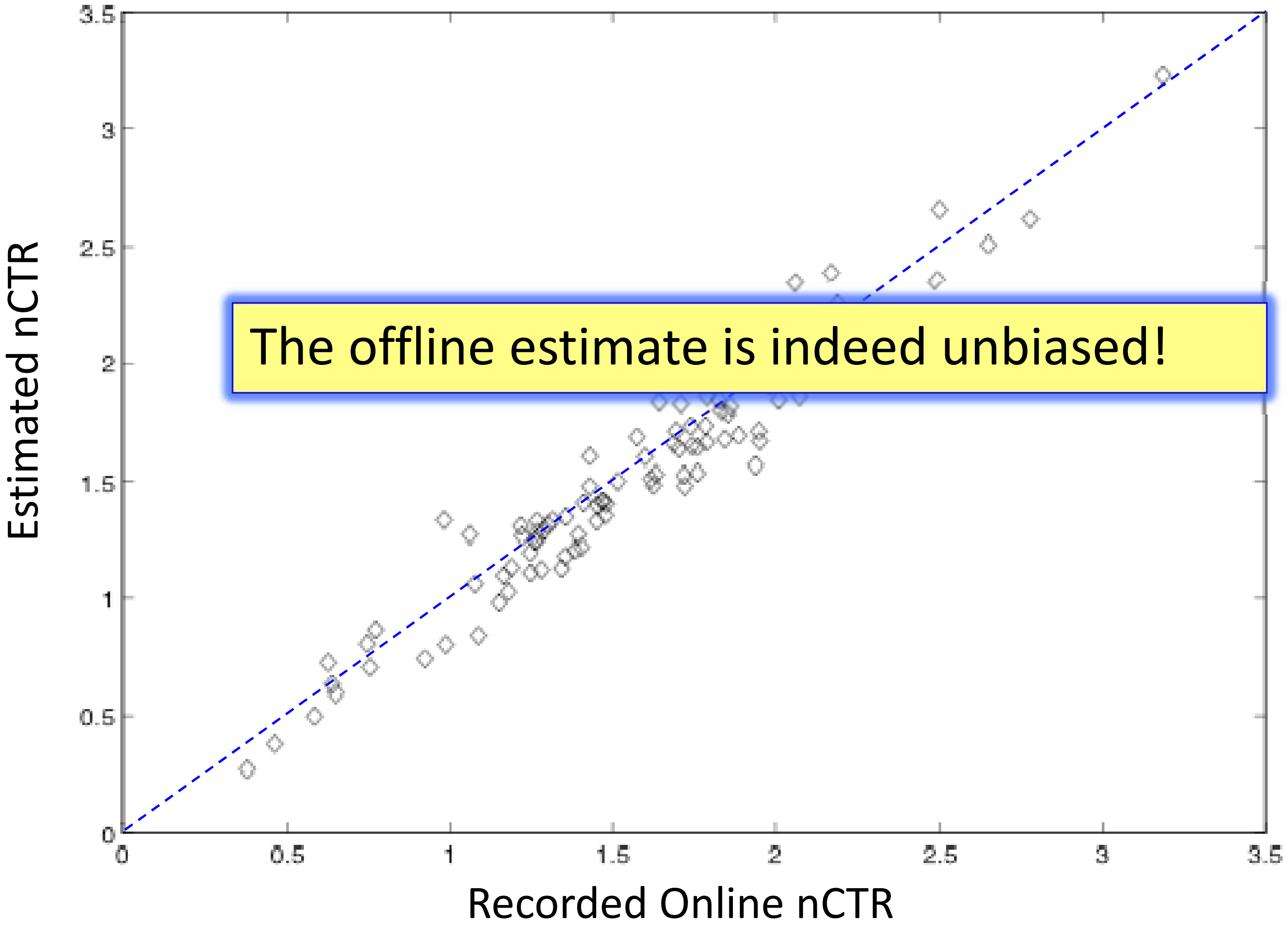
a_t : the displayed article at time t

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

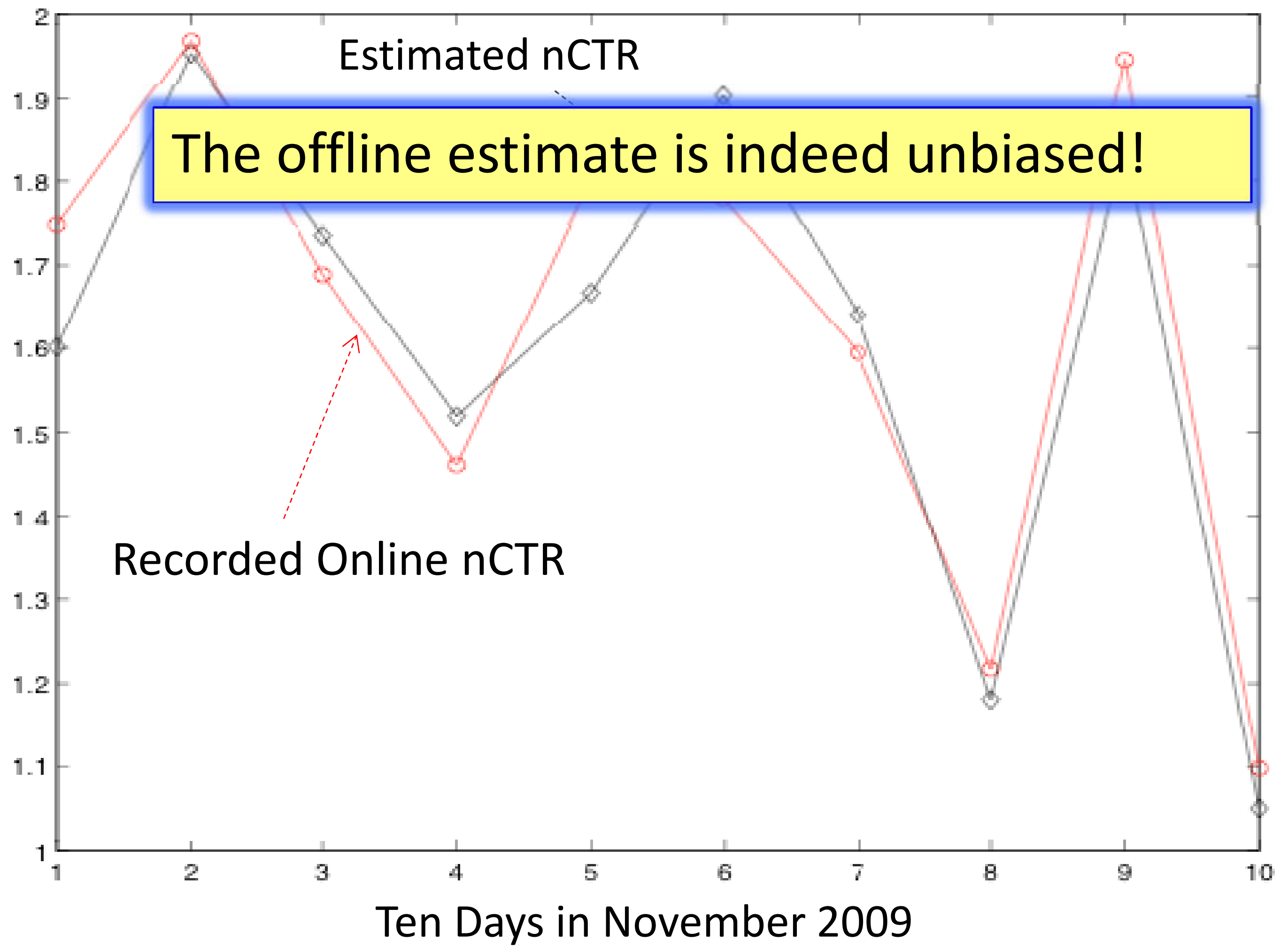
- Experiments run in 2009
  - 40M impressions over 10 days in exploration data
  - $p_a = \frac{1}{K}$ (uniform random exploration)
- Fixed an news-selection policy π
- Online experiment with π to measure CTR
  - The online ground truth
- Use exploration data to offline-evaluate π
  - The offline estimate

Are they close?

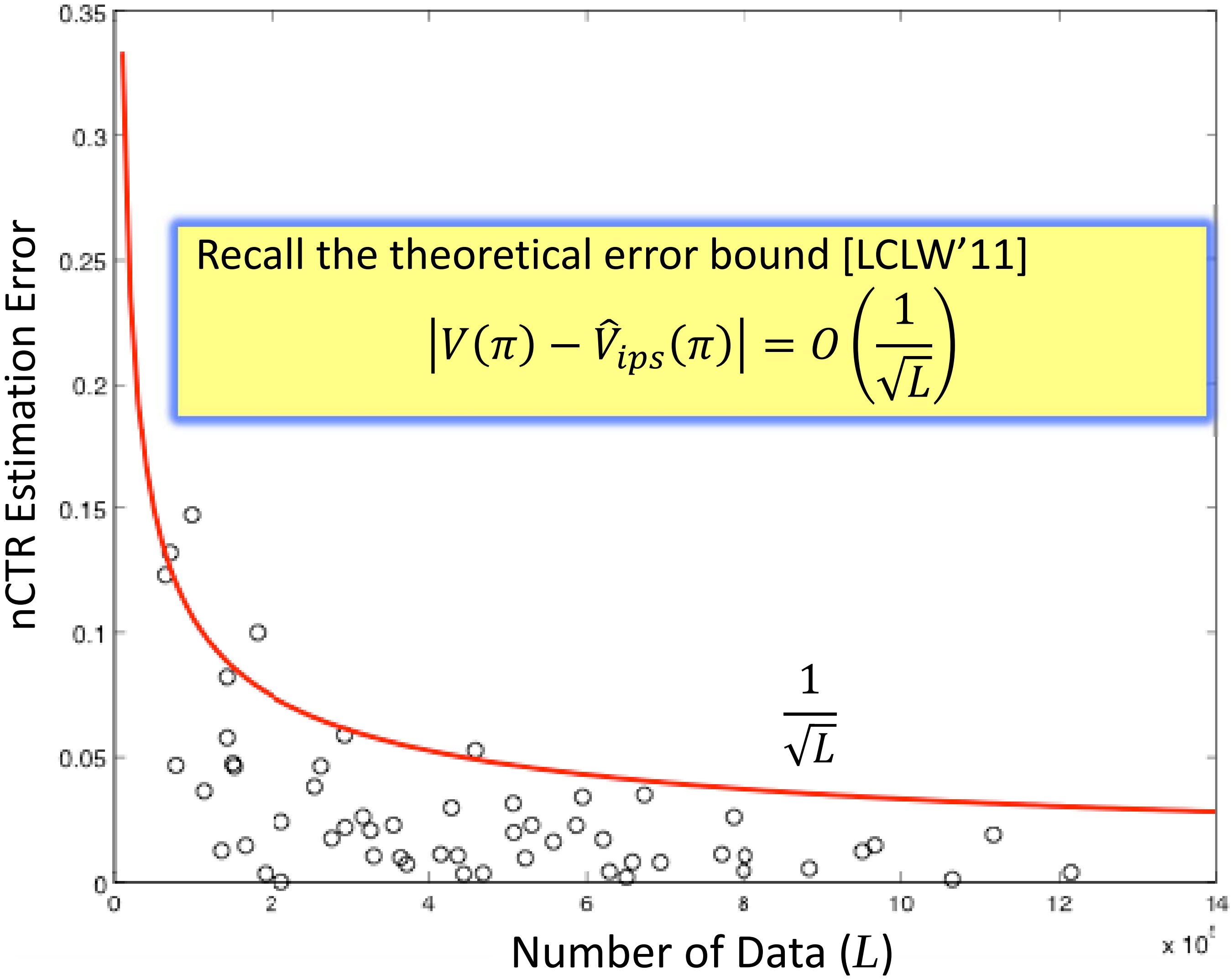
Unbiasedness: Article CTR



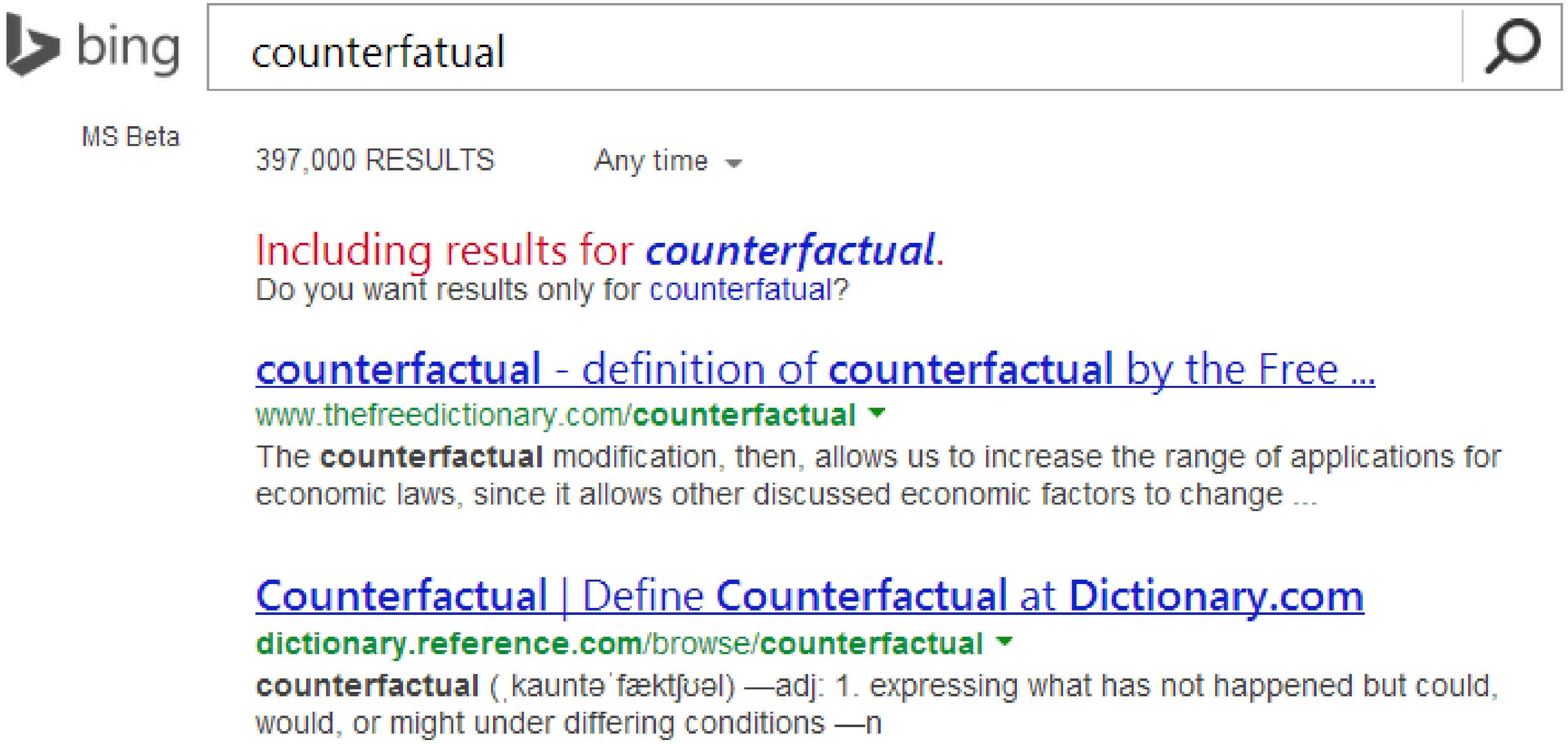
Unbiasedness: Daily Overall CTR



Estimation Error



Case Study 2: Bing Speller



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

bing ccn

- cnn: popular and similar query (excellent reformulation candidate)

or

- community cable network
- ccn international
- cement chemist notation

⋮

Bing Speller: A Harder Example

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**" $a \subset Q$
- **S** finally observes the **reward** (some click metric) r_a for a
- Repeat

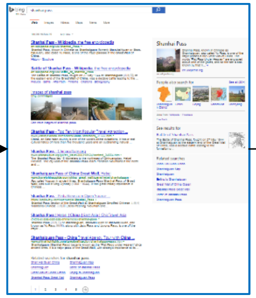
Goal of Speller is to maximize average per-round reward.

Exploration Data Collection [LCKG'14]

$q_0 = \{\text{shanhai pass}\}$

Randomize the subset

q_0 → Speller → $\{q_1, q_2\}$ → Rest of Bing (retrieval, ranking, merging, ...) →



→

Candidates:
$q_1 = \{\text{shanghai pass}\}$
$q_2 = \{\text{shanhai pass}\}$
$q_3 = \{\text{shanghai past}\}$

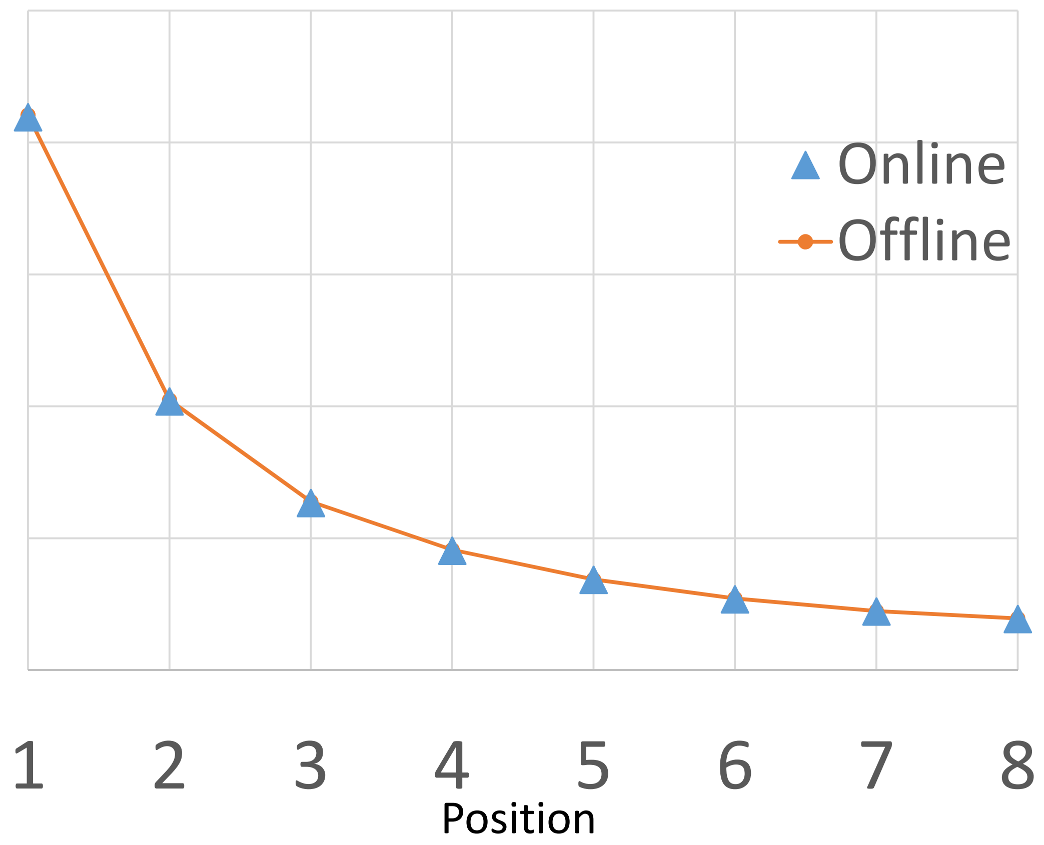
Collected ~15M search impressions at Bing

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

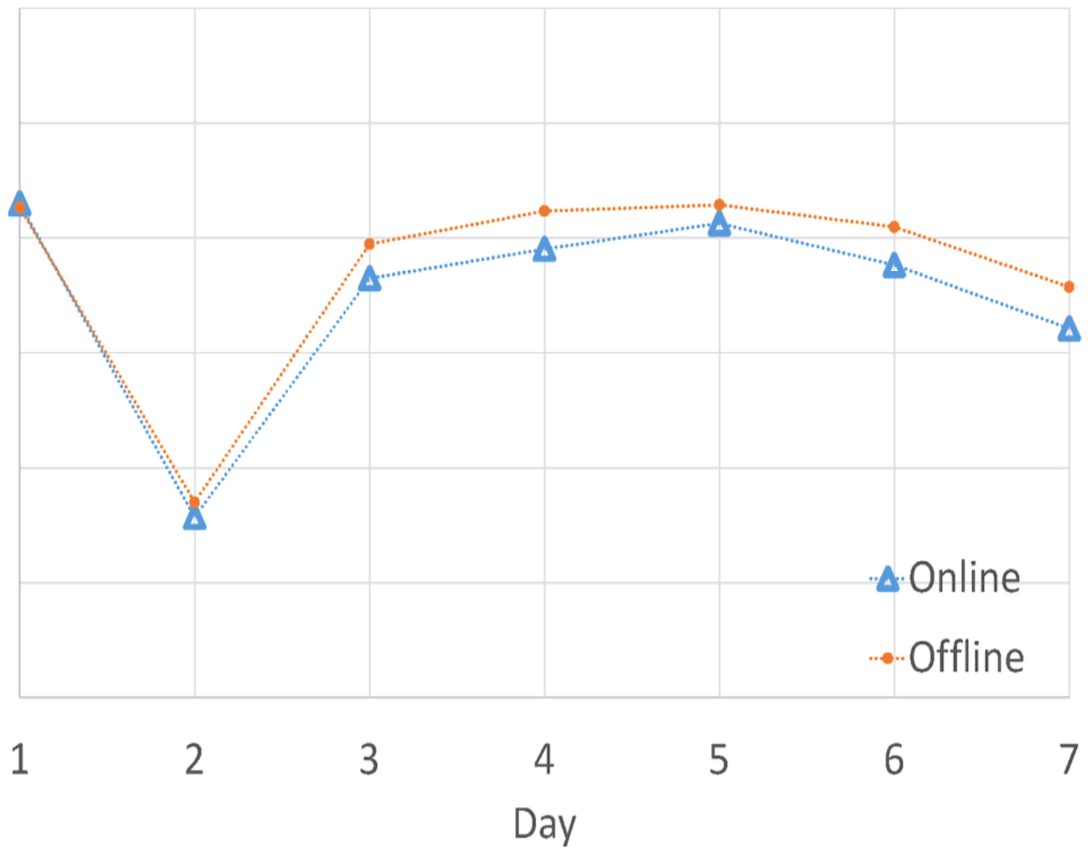
λ_1 and λ_2 control exploration aggressiveness

Accuracy of Offline Evaluator

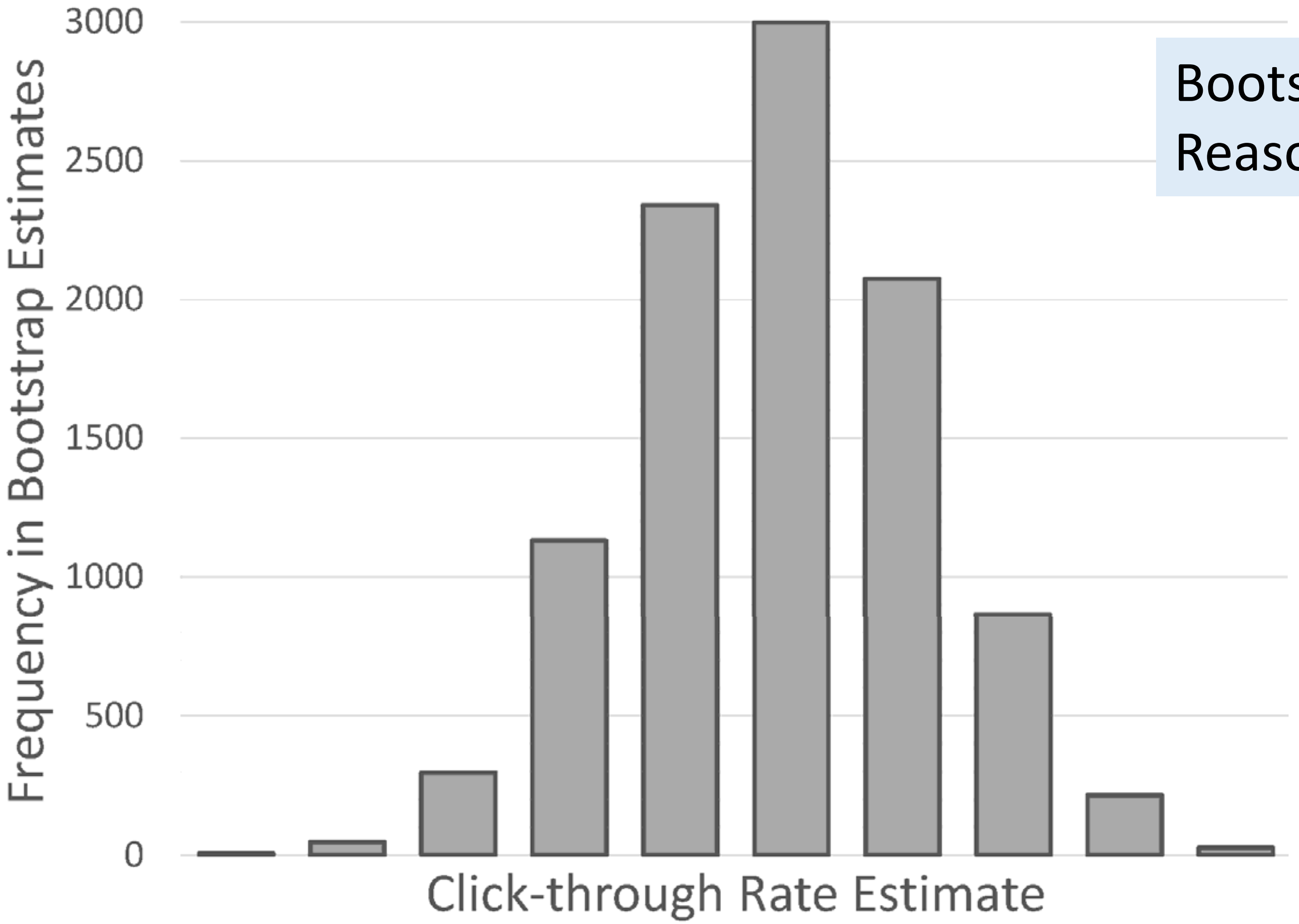
Position-specific click-through rate



Daily click-through rate

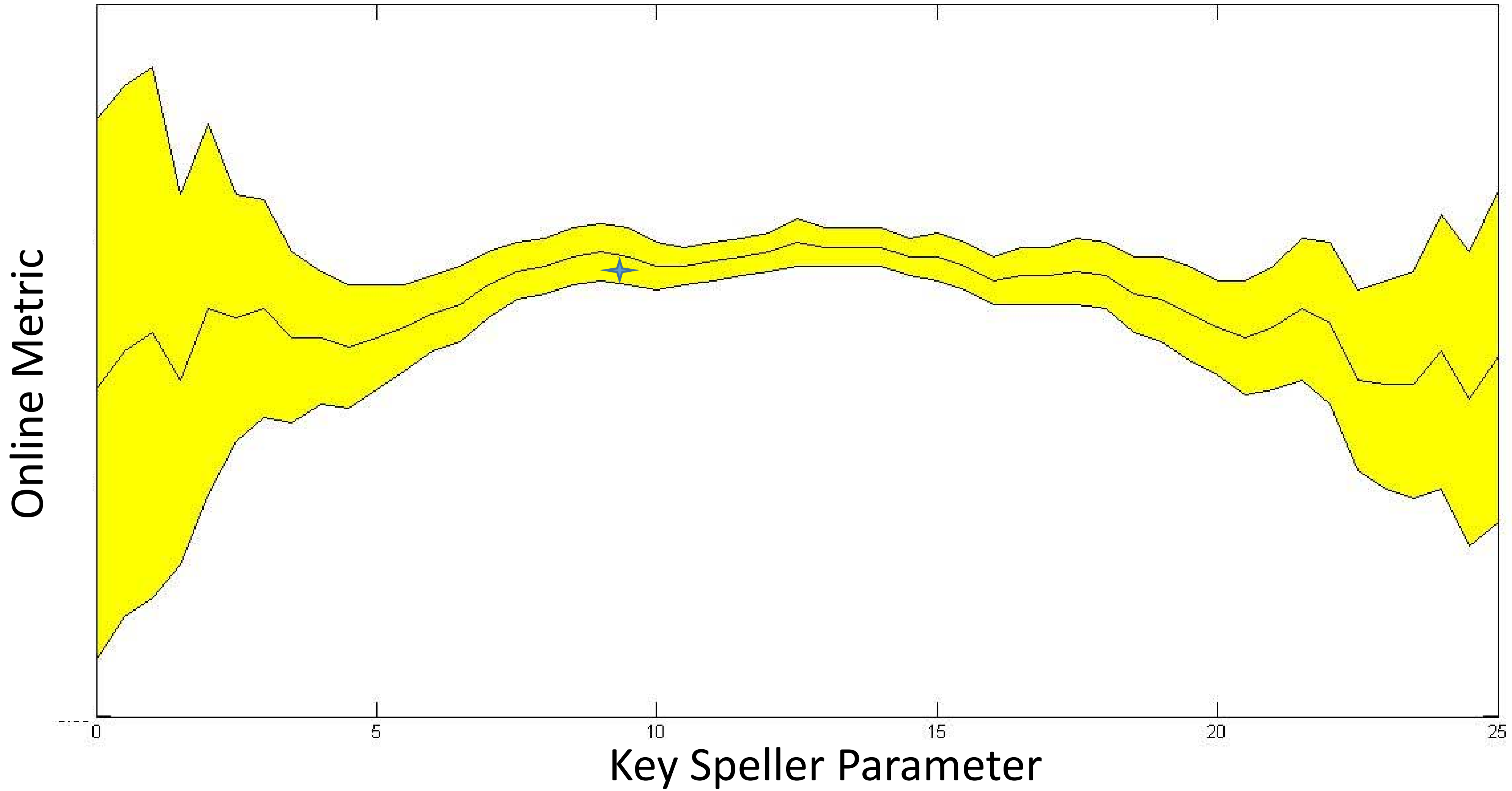


Normality of Offline Estimates



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

Quantifying Uncertainty in Offline Evaluation



Offline Optimization for Speller

- 70% exploration data to learn

$$\mathrm{Pr}(\mathrm{GoodResult} \mid \mathrm{Query}, \mathrm{CorrectionCandidate}\}$$

- 30% exploration data to offline-compare new and old Spellers

- Tends to be better if more are included
- But limited by capacity → 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

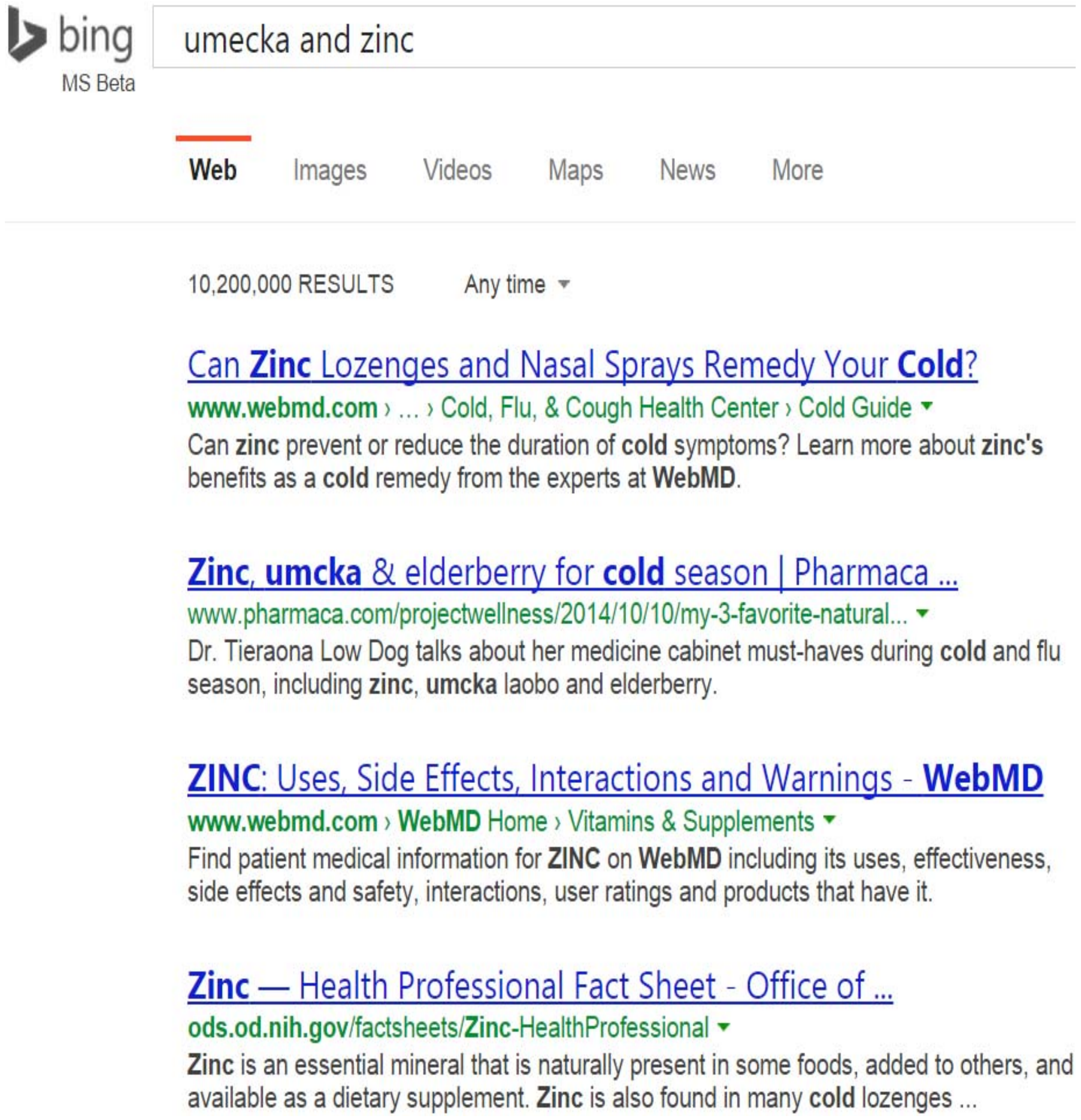
- Some winning examples

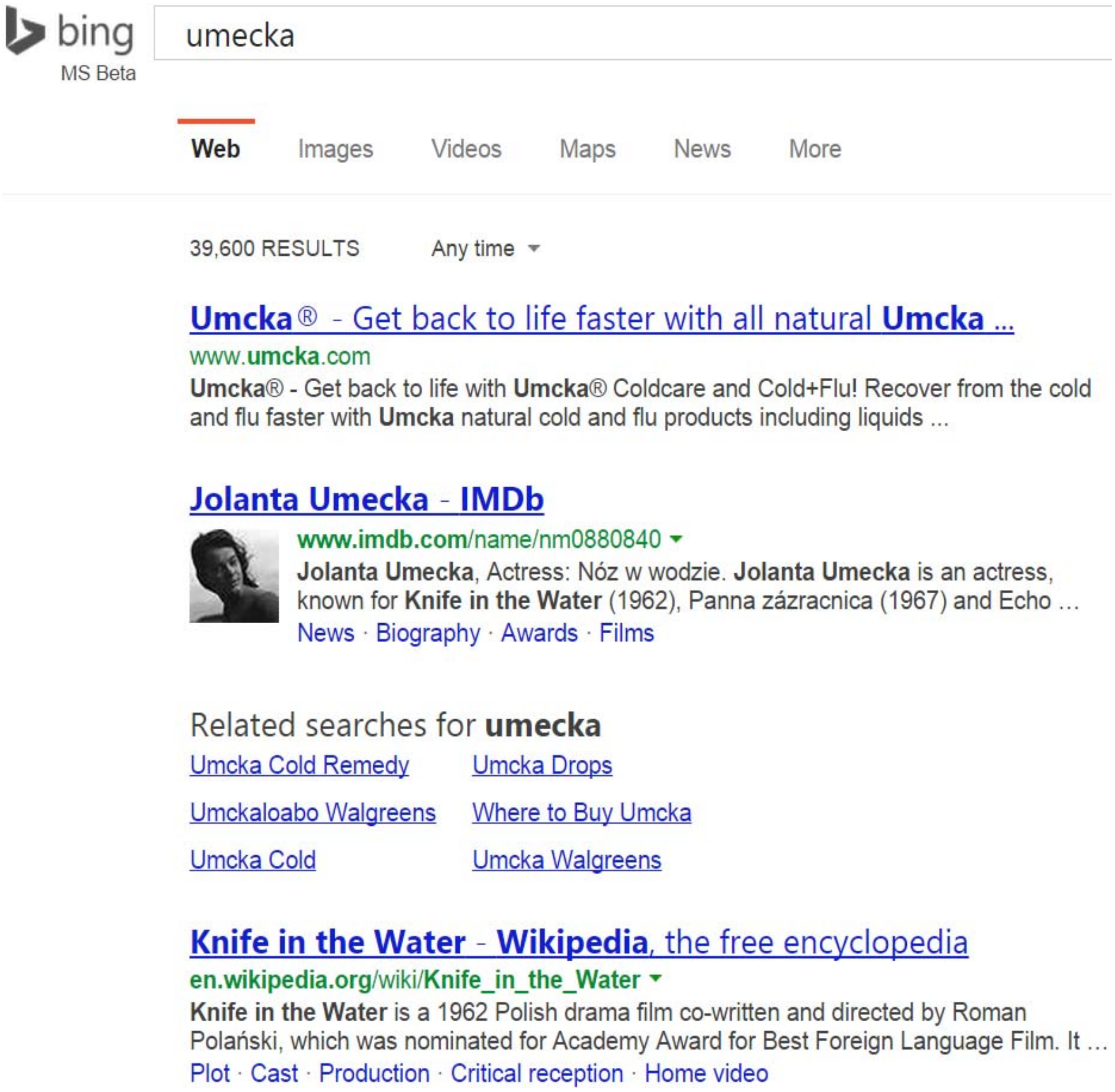
  "umecka and zinc" → "umcka and zinc" (treatments for cold symptoms)

  "catalina left attorney" → "catalina leff attorney" (right correction)

  "acer e1-5726870" → "acer e1-572 6870" (correct word breaking)

{umecka and zinc} vs. {umecka}





Evaluating Nonstationary Policies

- To estimate: $\mathrm{V}(\pi, T) = \mathbf{E}\left[\frac{1}{T}(r_1 + r_2 + \cdots r_T)\right]$

  where $a_t = \pi(x_1, a_1, r_1, \ldots, x_{t-1}, a_{t-1}, r_{t-1}, x_t)$
- Examples: all explore-exploit learning algorithms

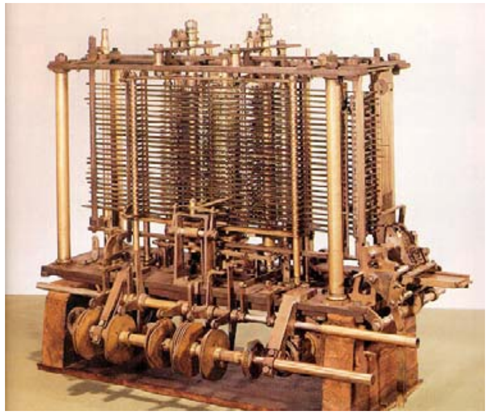
- Simple inverse propensity score does not work
- Need to simulate the trajectory

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

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

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

Nonstationary policy π



Data

$$\begin{Bmatrix} (x_1, a_1, r_1, 1/K) \\ \vdots \\ (x_L, a_L, r_L, 1/K) \end{Bmatrix}$$

reveal x_i →

← choose $\hat{a}_i = \pi(x_i)$

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

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

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

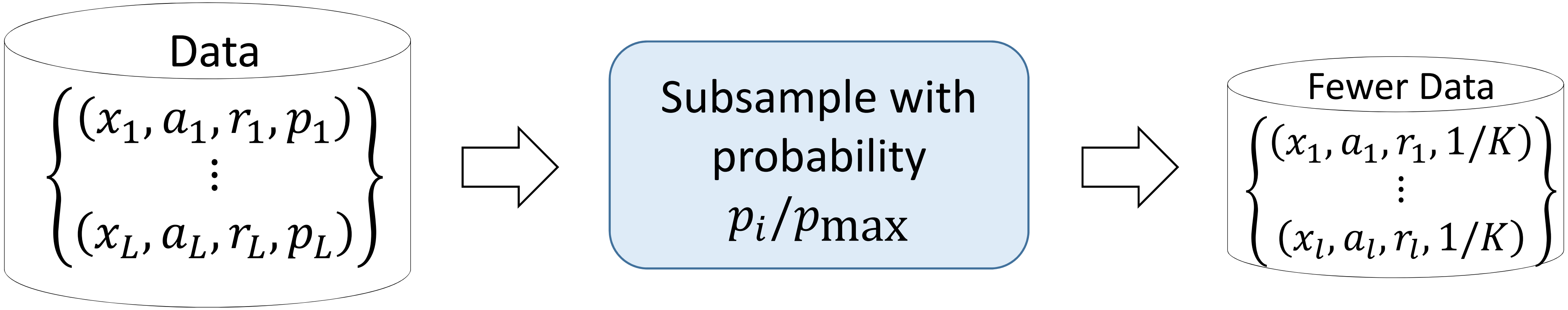
- Unfortunately, cannot use L or T to estimate confidence intervals
- Can use bootstrapping instead

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

- 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

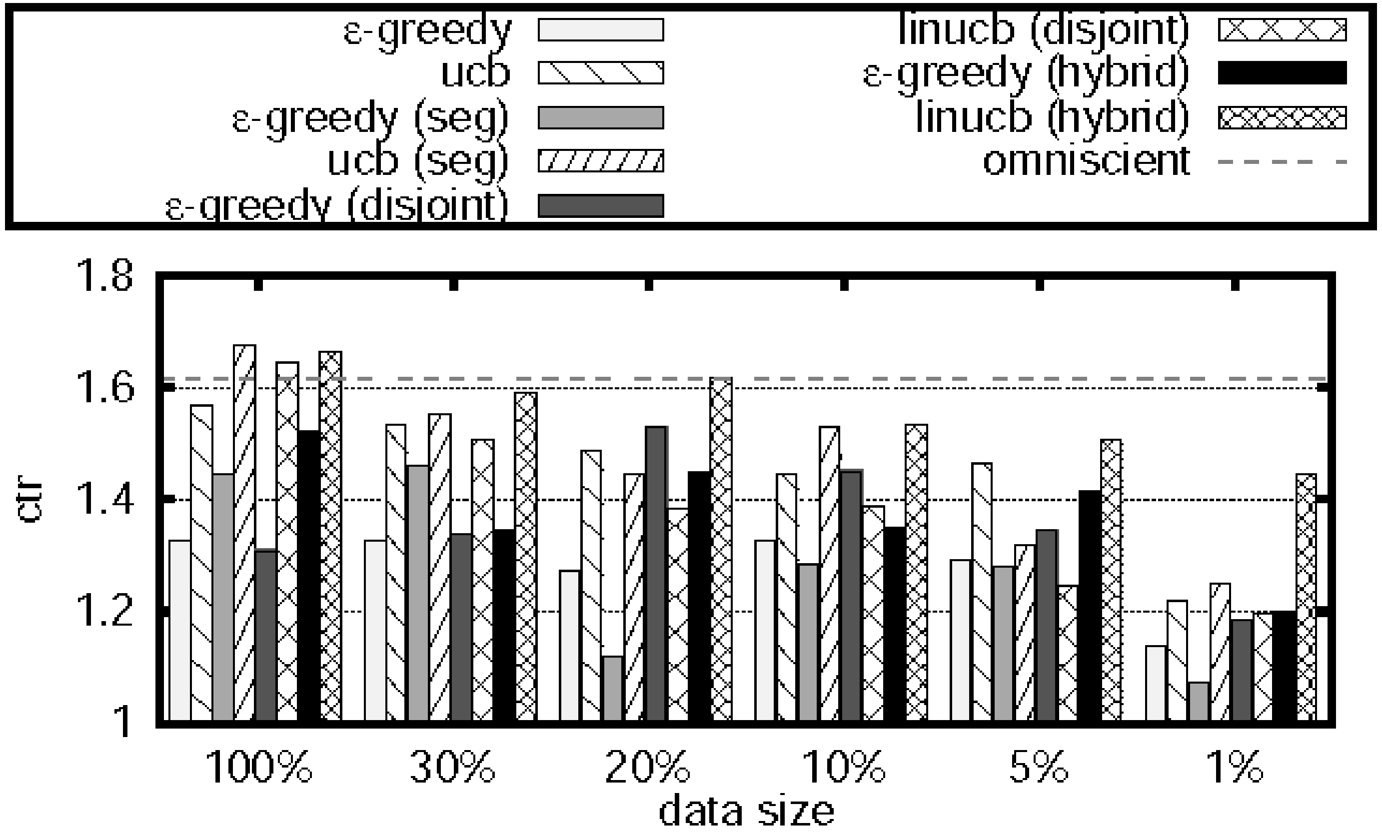
- Data collected in 2009
  - 40M impressions over 10 days in exploration data
  - $p_a = \frac{1}{K}$ (uniform random exploration)
- Low variance when evaluating representative nonstationary policies

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

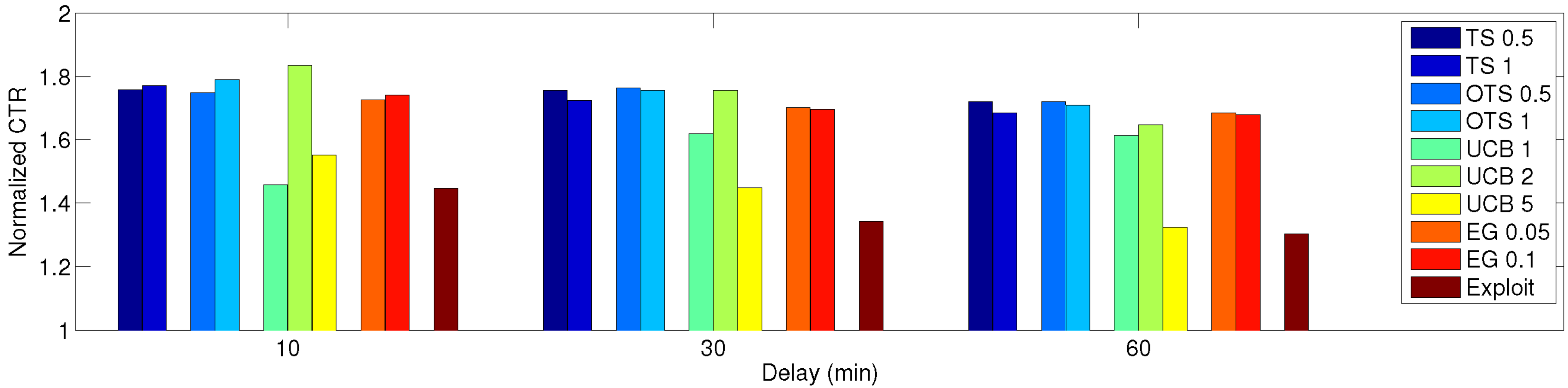
100 independent runs with different randomization seed

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

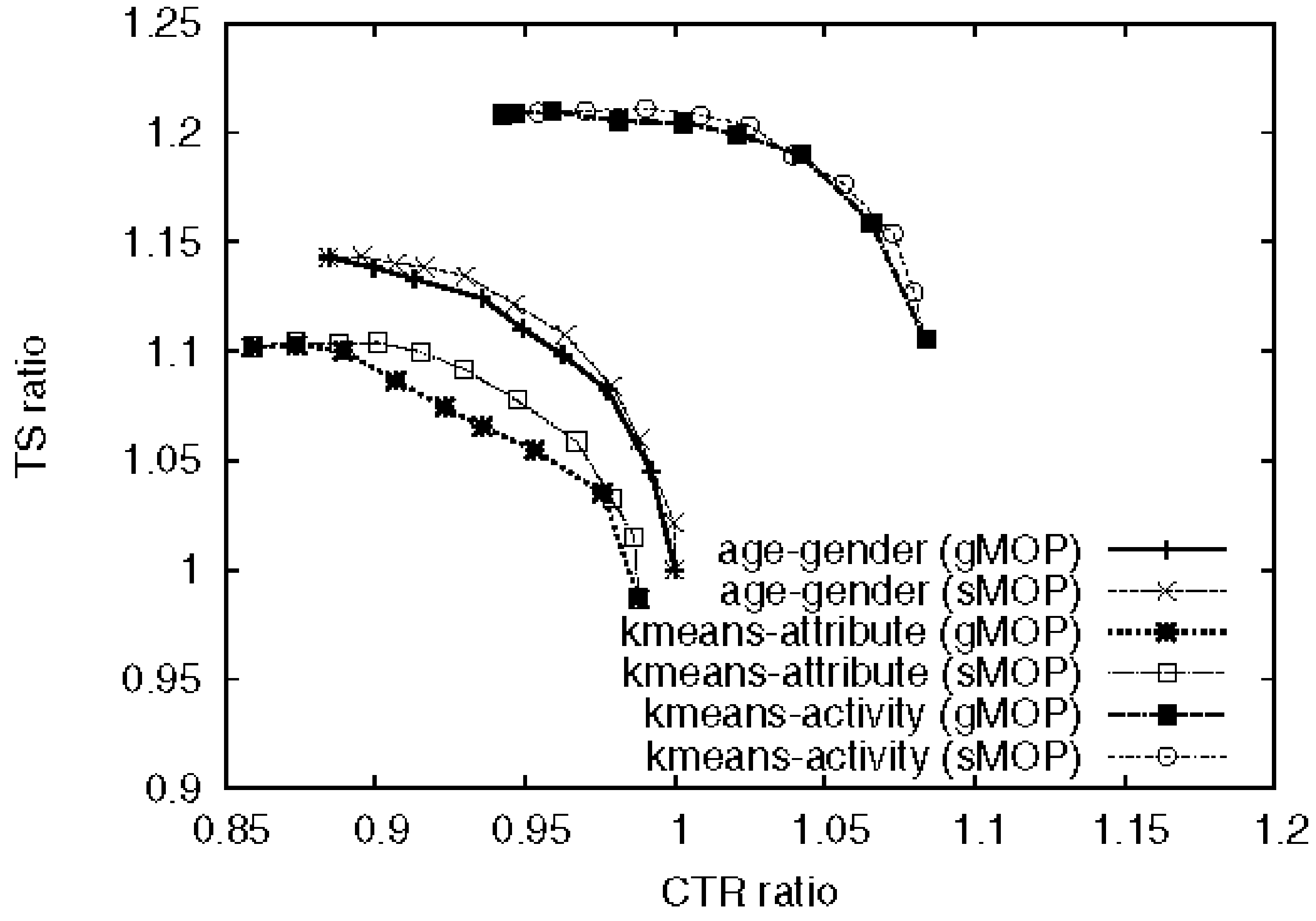
Application of Replay: Personalized Explore-Exploit Algorithms [LCLS'10]



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 $p = \frac{1}{L}(p_1 + p_2 + \cdots + p_L)$
- 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]

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

Enhanced Techniques

Unknown propensity scores

Direct policy optimization

Doubly robust estimation

Bootstrapped replay

Policy Optimization

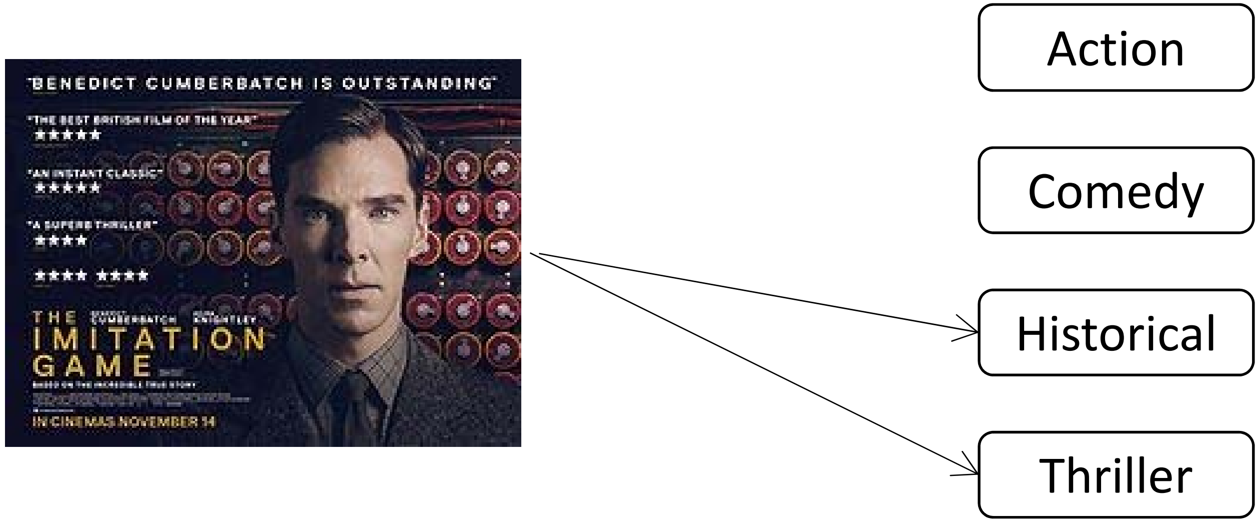
- Most often ultimate goal is to find optimal π with maximum $V(\pi)$

- 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



- Example x associated with subset of correct labels $c \subseteq L = \{1,2,\ldots,K\}$
  - x ("imitation game") → 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, $(x, c) \implies (x; r_1, r_2, \dots, r_K)$

- Policy π is treated as classifier

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

Policy value is classification accuracy!

Policy Optimization as Classification

Contextual bandit → weighted multi-class classification

$(x, a, r_a, p_a) \implies (x, a, w_a) \quad w_a = r_a / p_a$

Same trick as IPS!

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

Policy value is same as weighted classification accuracy!

Maximize policy value $V(\pi)$ → Maximize weighted classification accuracy $V(\pi)$ → Multi-class classification algorithm

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

Case Study 4: Advertising [SLLK'10]

- Problem: choose ad a for $x =$ (user, 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 $\frac{\left(r_a - f(x,a)\right)^2}{\max\{\hat{p}(a|x), \tau\}}$
  - 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)

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

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

- Inverse Propensity Score (IPS)

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

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

- Doubly Robust (DR) [RRZ'94]

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

DR: Unbiasedness

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

$$= \frac{1}{L}\sum_i \left(\hat{r}(x,\pi(x)) \left(1 - \frac{\mathbf{1}(\pi(x) = a)}{\hat{p}_a}\right) + \frac{r_a \cdot \mathbf{1}(\pi(x) = a)}{\hat{p}_a} \right)$$

$\hat{r} = r \implies E[\hat{V}_{dr}] = V(\pi)$

$\hat{p} = p \implies E[\hat{V}_{dr}] = V(\pi)$

- 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

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

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

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

DR has lowest bias with "reasonable" $\hat{p}$ and $\hat{r}$

DR: Variance Analysis

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

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

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

DR has lower variance than IPS with "reasonable" $\hat{r}$

DM often has low variance, <u>not</u> 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

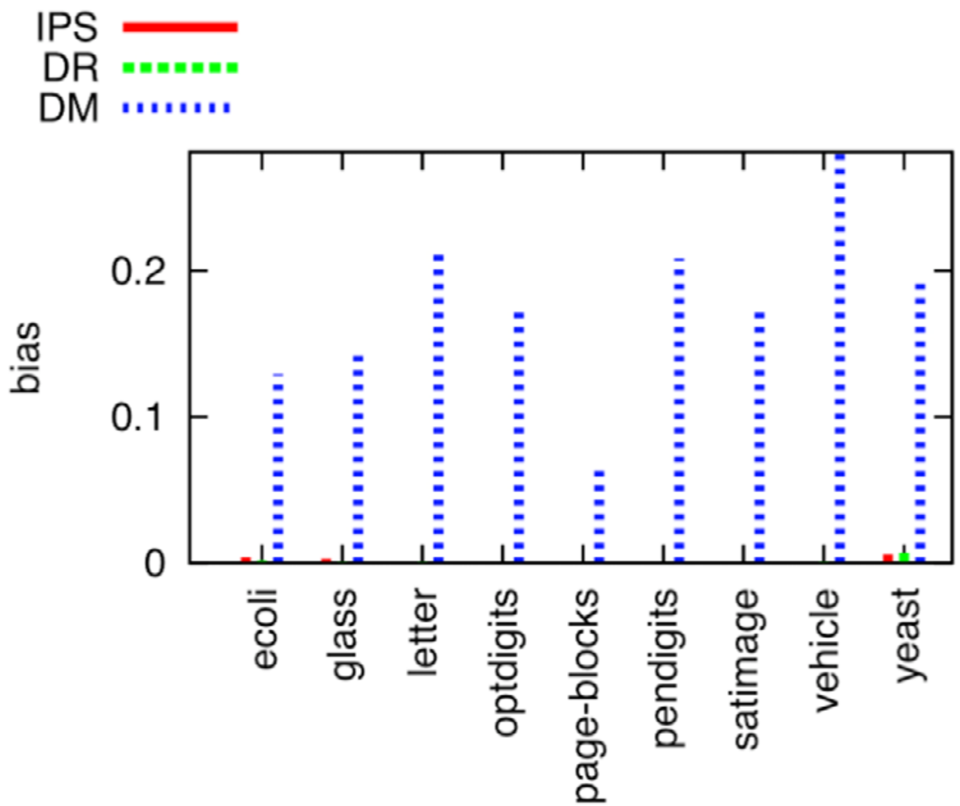
Classification to bandit: $(x, c) \implies (x; r_1, r_2, \dots, r_K)$

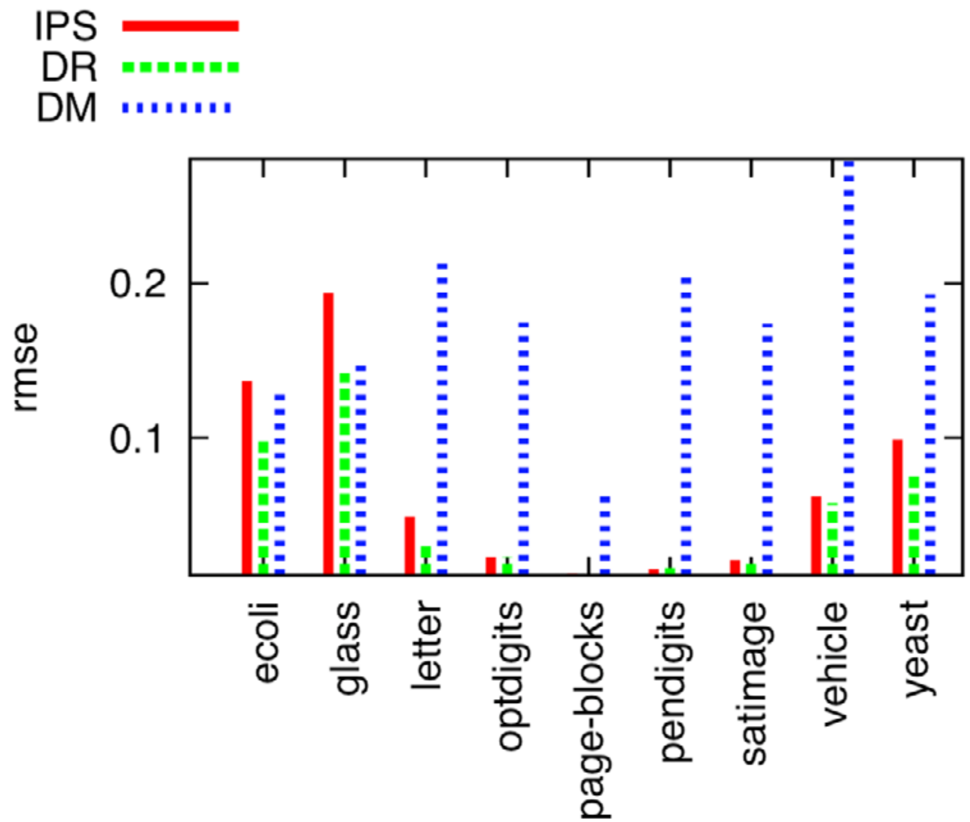
Bandit to classification: $(x, a, r_a, p_a) \implies (x, a, w_a)$ $\quad w_a = r_a / p_a$

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

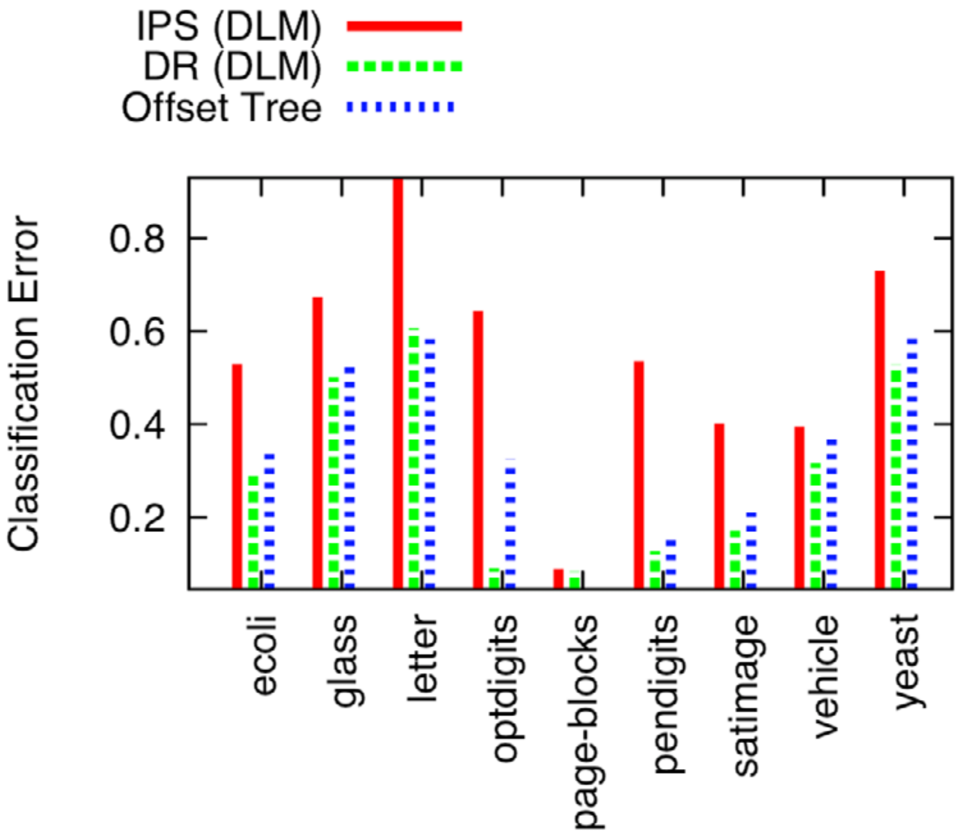
Optimization algorithms

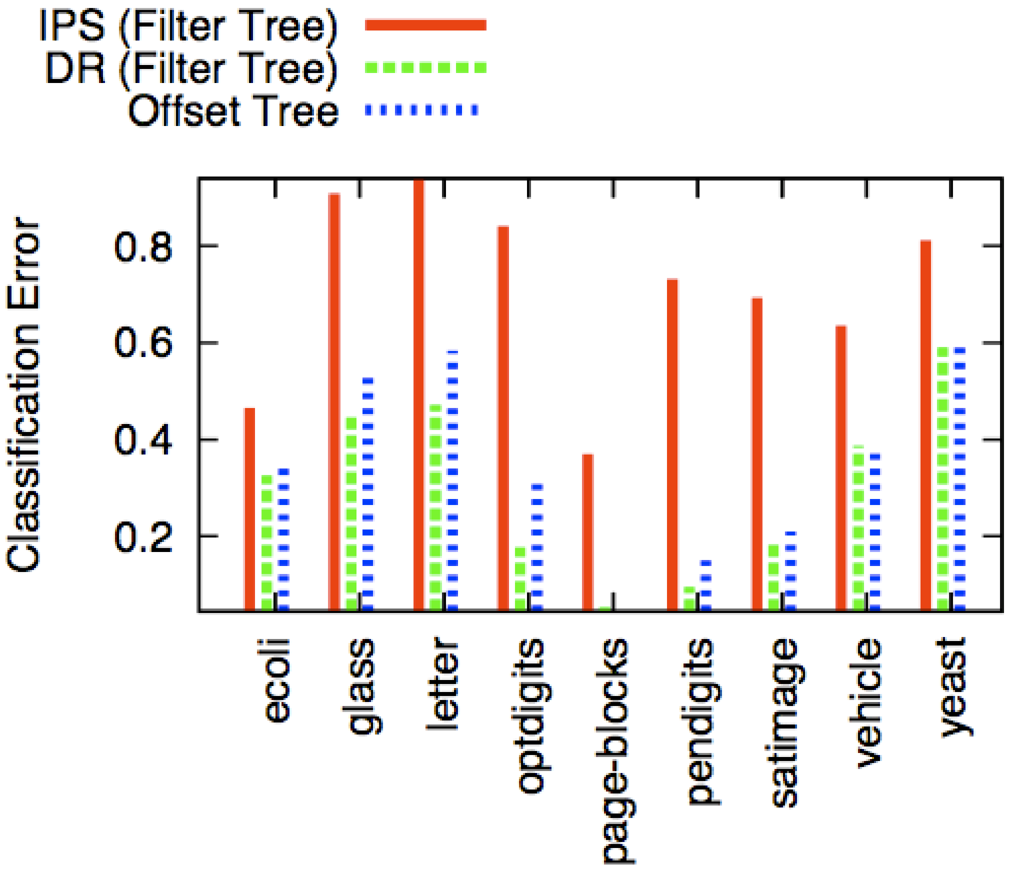
- Direct loss minimization [MHK'11]
- Filter tree [BLR'08]
- Offset tree [BL'09]: alternative policy optimization algorithm

Generic multi-class classification (Combined with DM, IPS, DR)

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

Policy Optimization



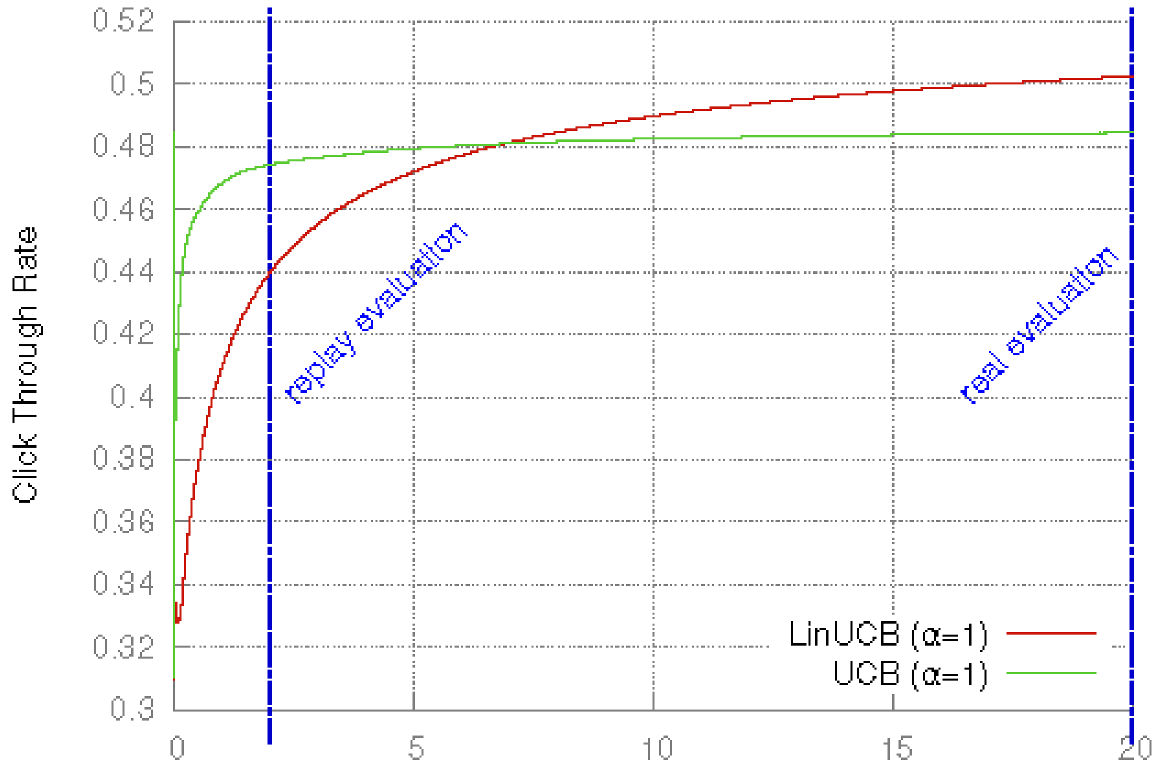


Enhanced Techniques

Unknown propensity scores
Direct policy optimization
Doubly robust estimation
Bootstrapped replay

Time Acceleration Problem [NMP'14]

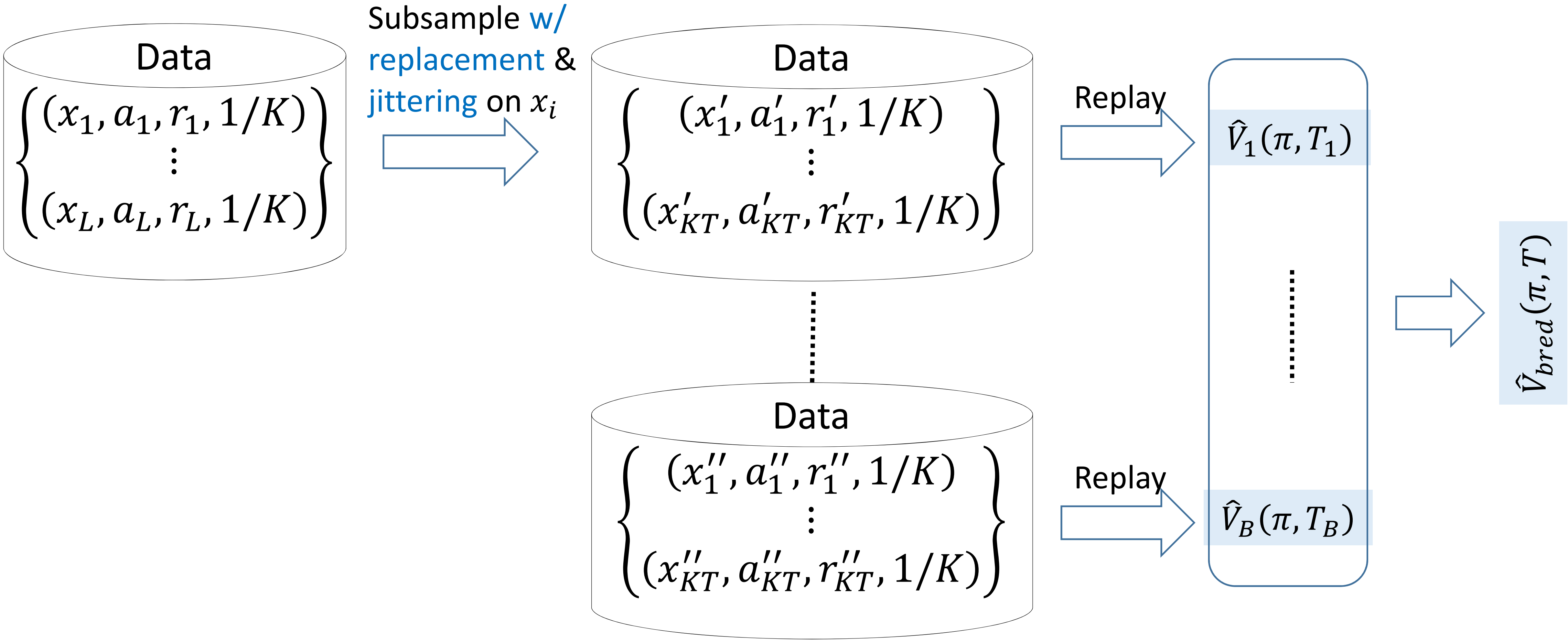
- With $L = |D|$ data and uniform exploration $p_a = 1/K$
  - Expected number of matches is L/K
  - Replay can estimate $V(\pi, T)$ up to $T \approx L/K$



(from [NMP'14])

Replay cannot evaluate π for too large T

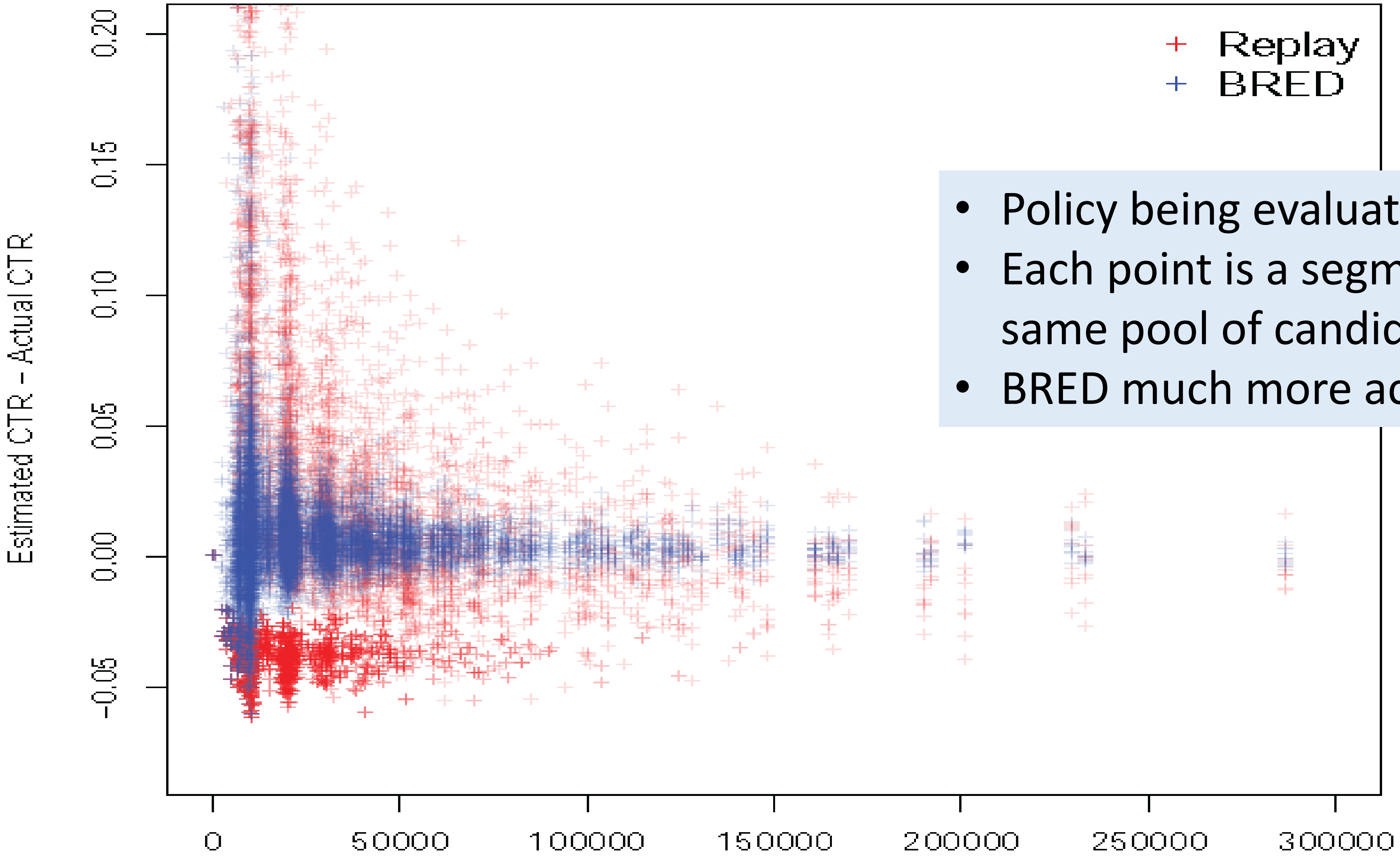
BRED [NMP'14]
"Bootstrapped Replay on Expanded Data"



BRED Theory

- For stationary policies, confidence intervals are estimated much faster
  - $O(1/T)$ as opposed to $O\left(1/\sqrt{T}\right)$
  - 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

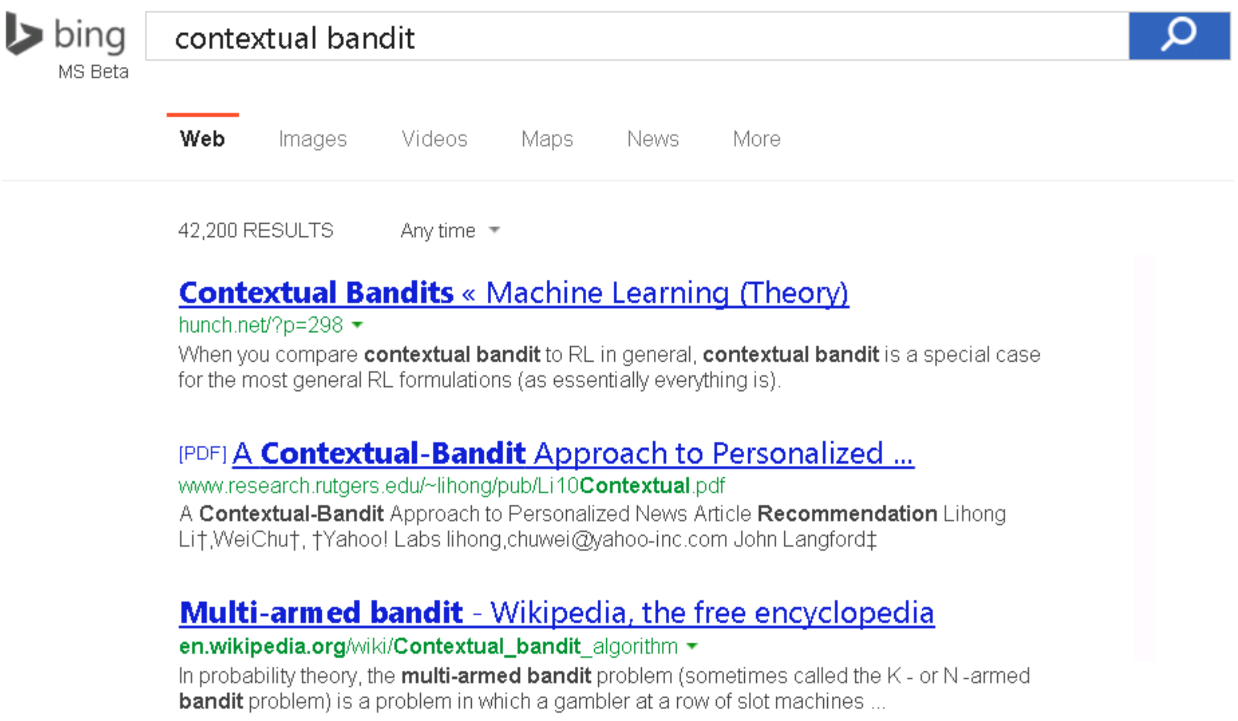


- Policy being evaluated: UCB
- Each point is a segment of data with same pool of candidate articles
- BRED much more accurate

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



How to Design Exploration Distributions (2)

- $Var\left(\hat{V}(\pi)\right)$ 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

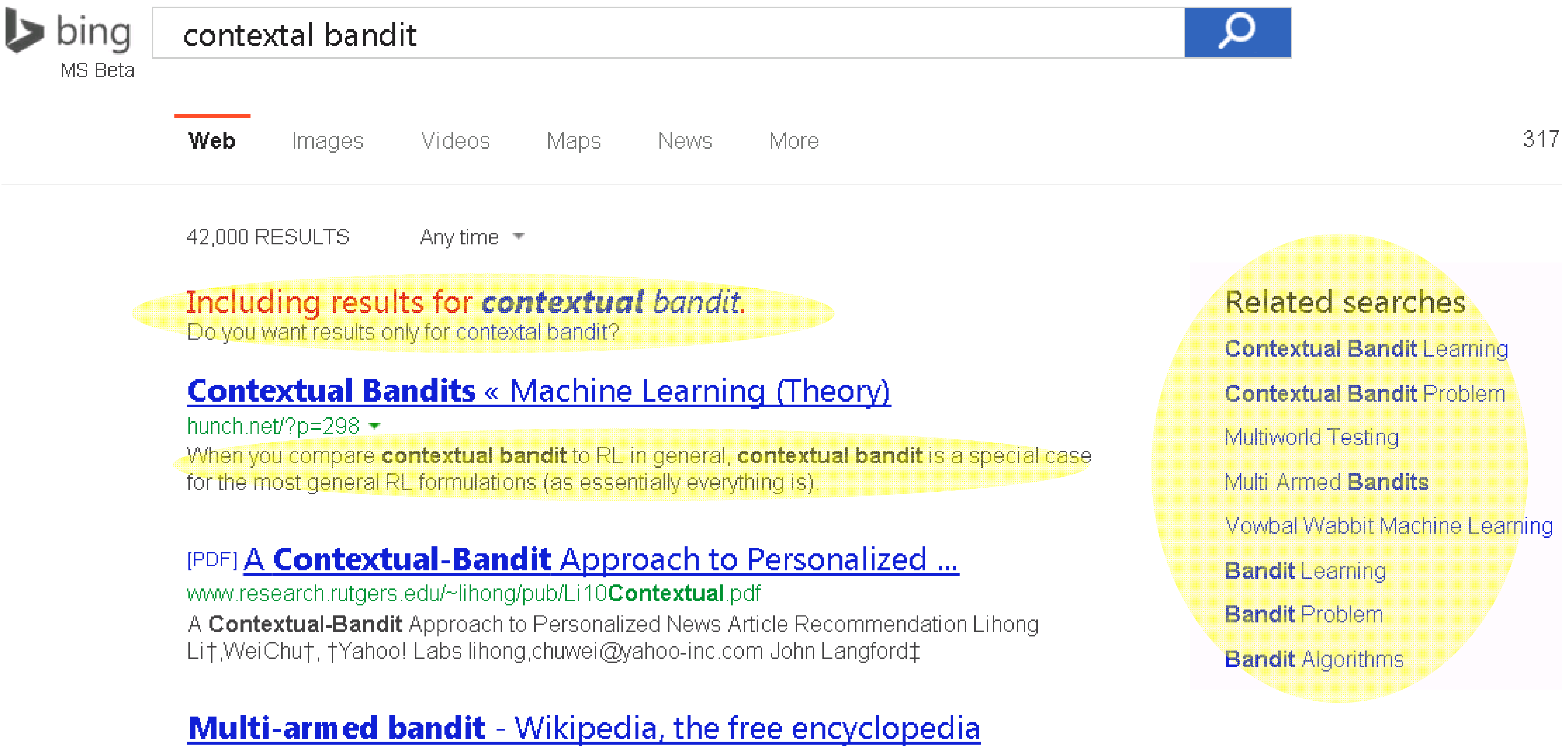
General theme: use historical data to offline-discovery online metrics (estimate causal effects from historical data)

- Policy evaluation/optimization

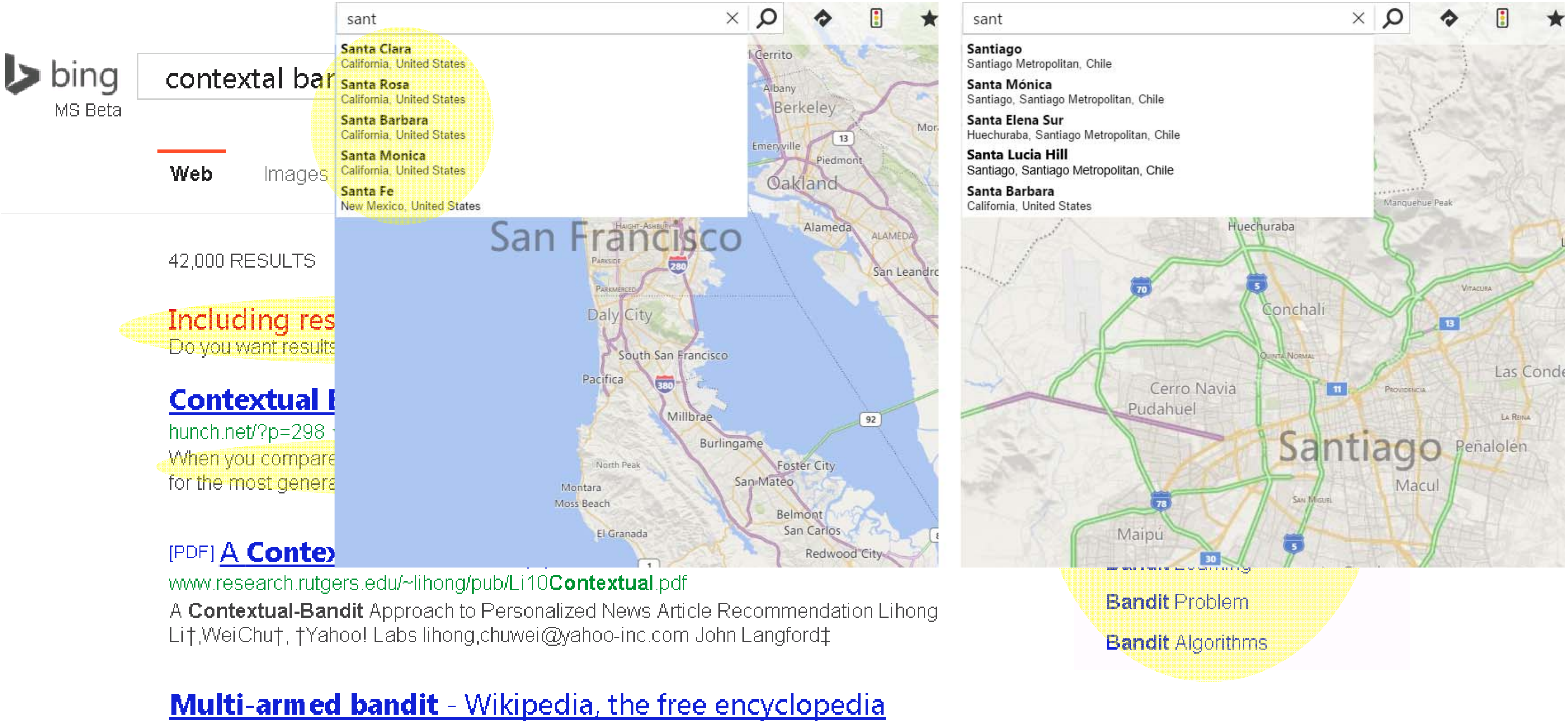
- Unbiasedness with IPS and Replay
- Variance reduction techniques with DR, etc.

- Case studies in news, search, advertising, and benchmark

More Bing Examples



More Bing Examples



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.