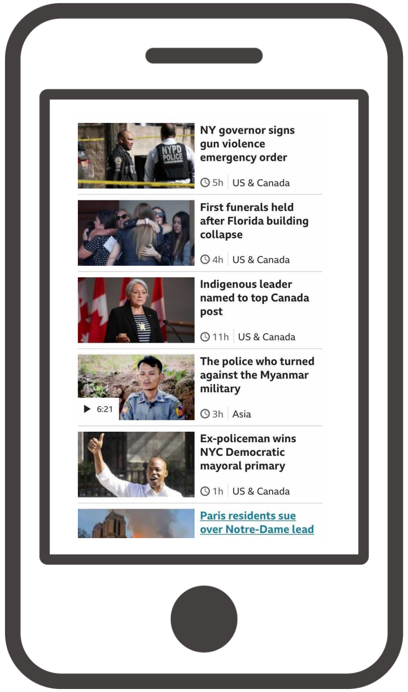
KASYS 筑波大学知識獲得システム研究室 UNIVERSITY OF TSUKUBA KNOWLEDGE ACQUISITION SYSTEM LAB.

Decomposition and Interleaving for Variance Reduction of Post-click Metrics

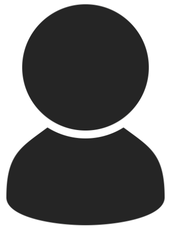
Kojiro Iizuka (Gunosy Inc. / University of Tsukuba)
Yoshifumi Seki (Gunosy Inc.)
Makoto Kato (University of Tsukuba / JST, PRESTO)

Suppose that a user uses a news service.



News service

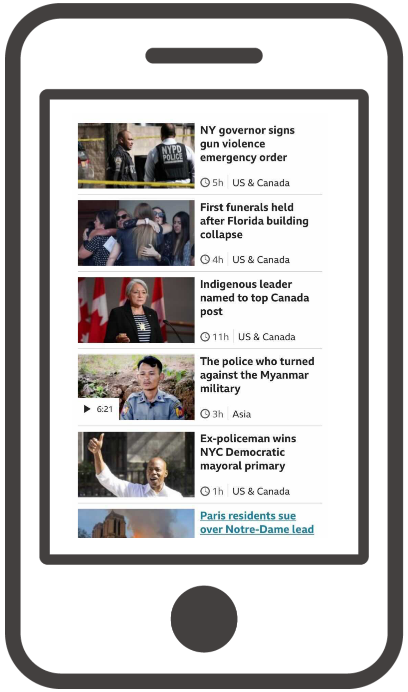
Impression

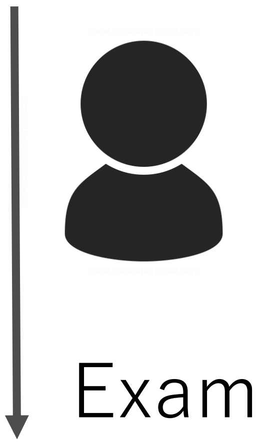


The screenshot comes from BBC news in 2021/07/07.

Preliminary: Post-Click

Impression → Examination

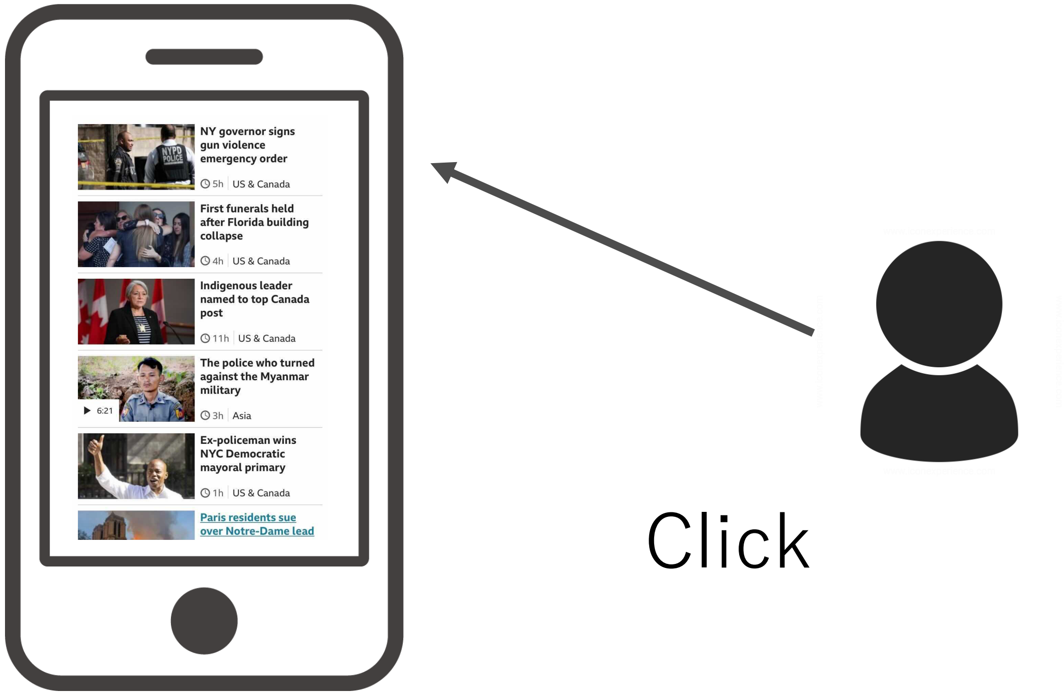




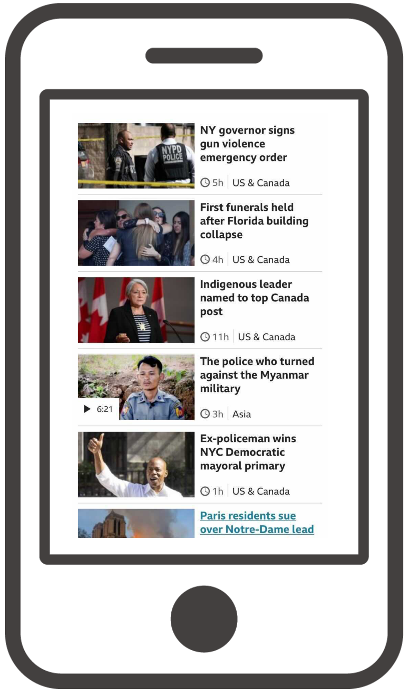
Examination

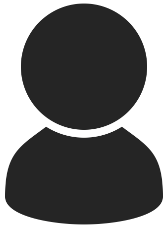
Preliminary: Post-Click

Impression → Examination → Click

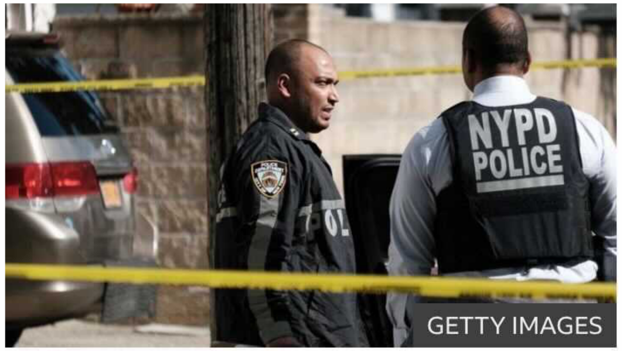


Impression → Examination → Click → Read





Read



New York has become the first US state to declare a disaster emergency order to address rising gun violence.

New York state saw 51 shootings over the 4 July holiday weekend, Governor Andrew Cuomo said as he signed the executive order.

The directive will funnel $138.7m (£100m) towards gun violence intervention and prevention programmes.

It comes amid reports of a rise in gun deaths countrywide, including nearly 200 over the past weekend.

Impression → Examination → Click → Read

Pre-Click

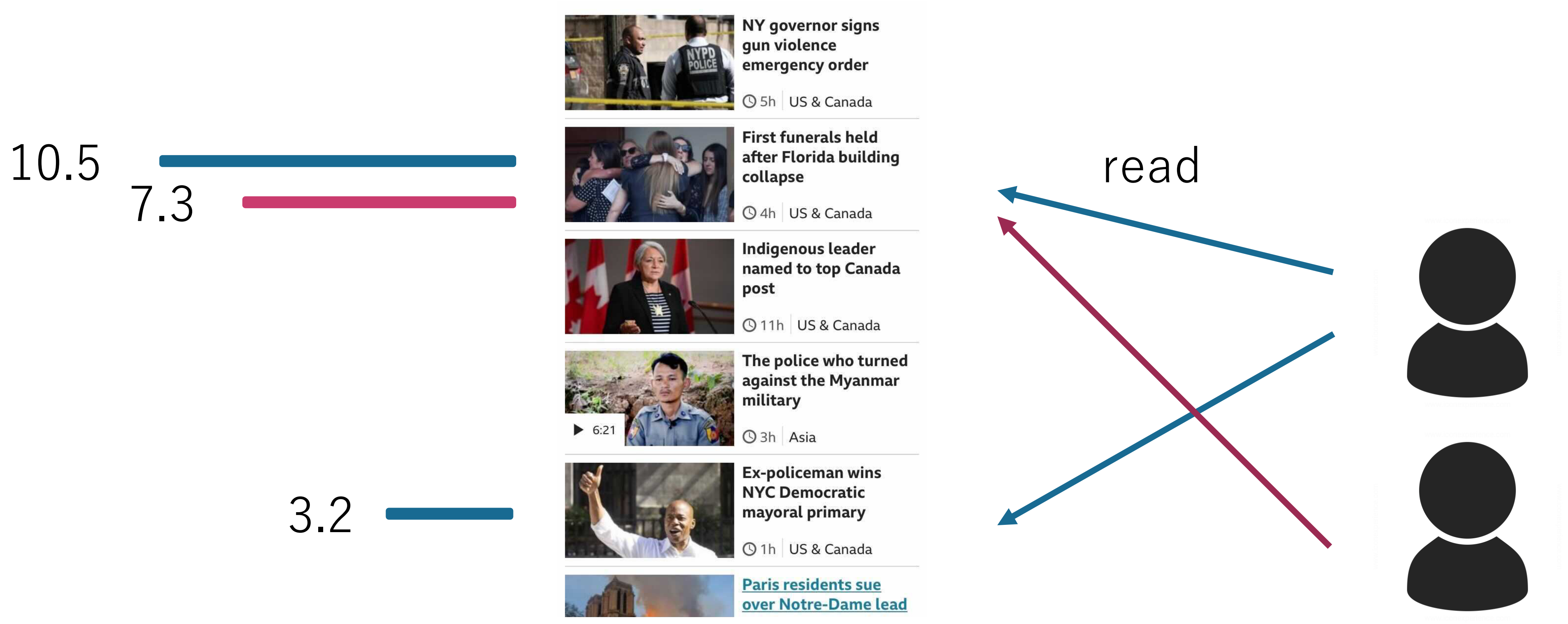
Post-Click

In this study, we aim to evaluate the quality of rankings based on post-click behaviors.

Suppose that we evaluate a reading time as post-click metric.

Reading time (second)

Suppose that we evaluate a reading time as post-click metric.



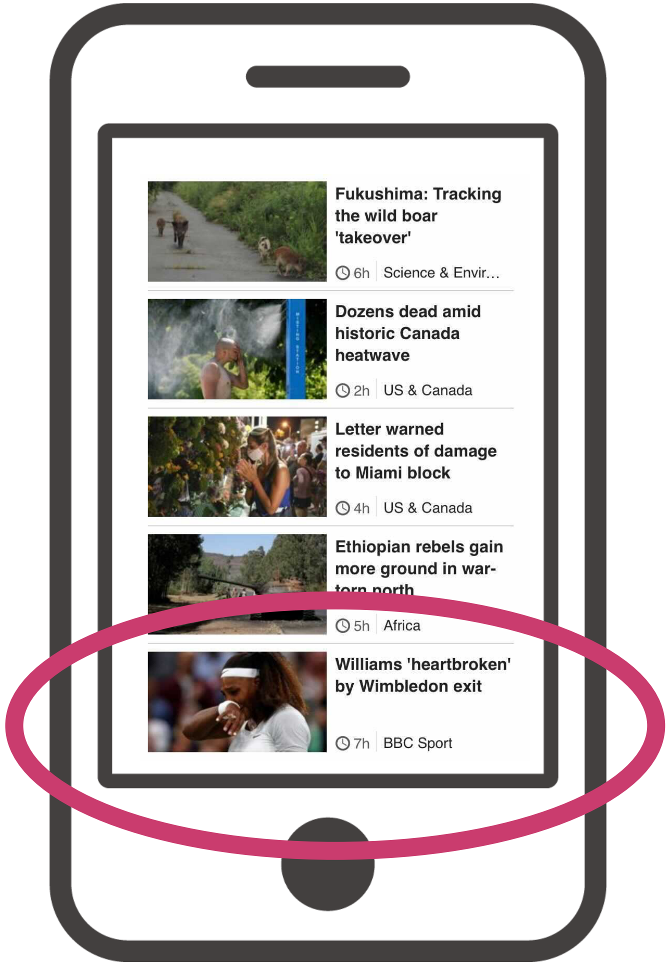
Post-click metric = (10.5 + 7.3 + 3.2) / 3 = 7.0

- **Online controlled experiments are conducted daily to evaluate recommender algorithms.**
  - A/B testing is a common approach
    - Comparing two different outcomes by showing them to different user groups.
  - Typical evaluation metrics:
    - Click-based metrics (e.g., click-through rate (CTR))
    - Post-click metrics (e.g., news reading time, the number of reservations)
- **Post-click metrics is particularly important for the continuous improvement of algorithms.**
  - Post-click metrics are closely related to user satisfaction and the sales of services [1, 2].

[1] Okura et al., Embedding-based news recommendation for millions of users, KDD2017
[2] Grbovic et al., Real-time personalization using embeddings for search ranking at Airbnb, KDD2018

Motivation: Example 1.

Suppose that a news article is ranked at the bottom of a ranking, which users spend a significantly different length of time to read.



The screenshot comes from BBC news in 2021/06/30.

Suppose that a news article is ranked at the bottom of a ranking, which users spend a significantly different length of time to read.

ReadingTime

Receive **less clicks** than at a top news in the ranking.

The news screenshot comes from BBC news in 2021/06/30.

Motivation: Example 1.

Suppose that a news article is ranked at the bottom of a ranking, which users spend a significantly different length of time to read.

ReadingTime

?

High Variance
of mean reading time

$$\frac{\sigma}{n}$$

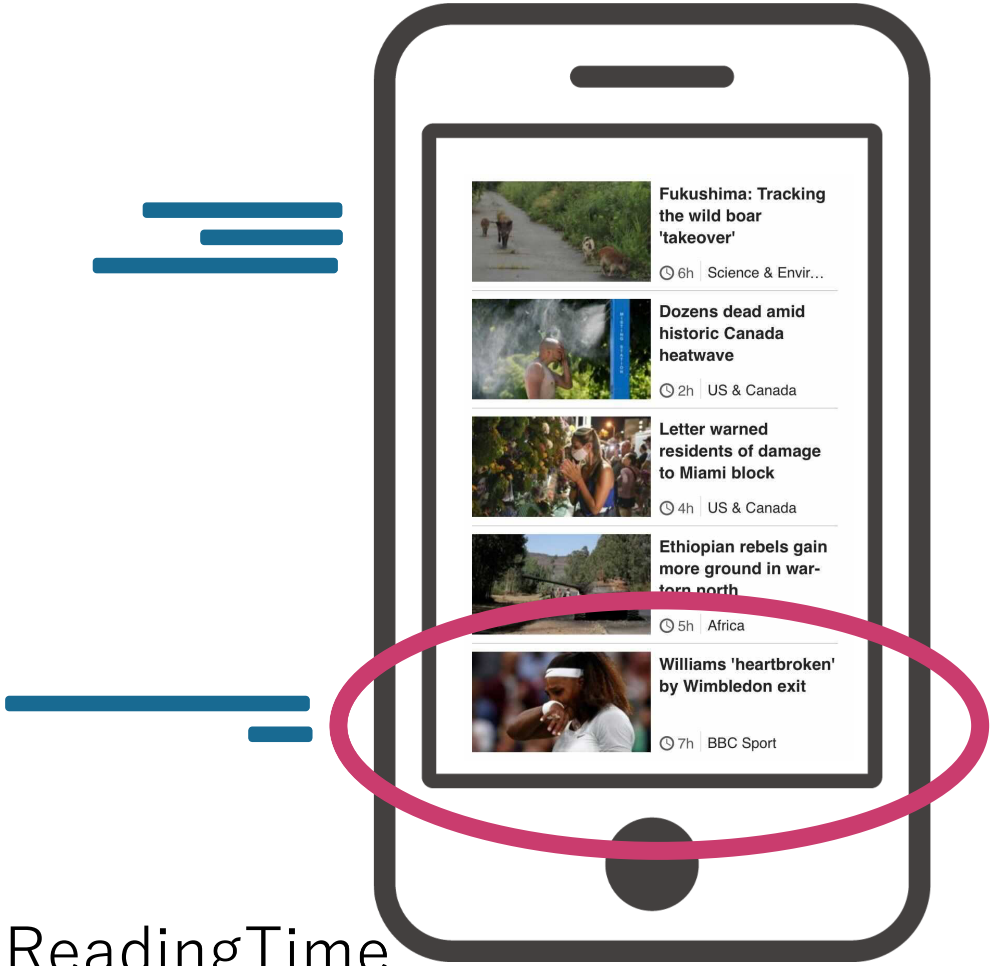
σ = Variance of reading time

n = Number of clicks

Receive **less clicks** than at a top news in the ranking.

The news screenshot comes from BBC news in 2021/06/30.

Suppose that a news article is ranked at the bottom of a ranking, which users spend a significantly different length of time to read.



ReadingTime

Low evaluation efficiency
of mean reading time.

High Variance
of mean reading time

$$\frac{\sigma}{n}$$

σ = Variance of reading time

n = Number of clicks

Receive **less clicks** than at a top news in the ranking.

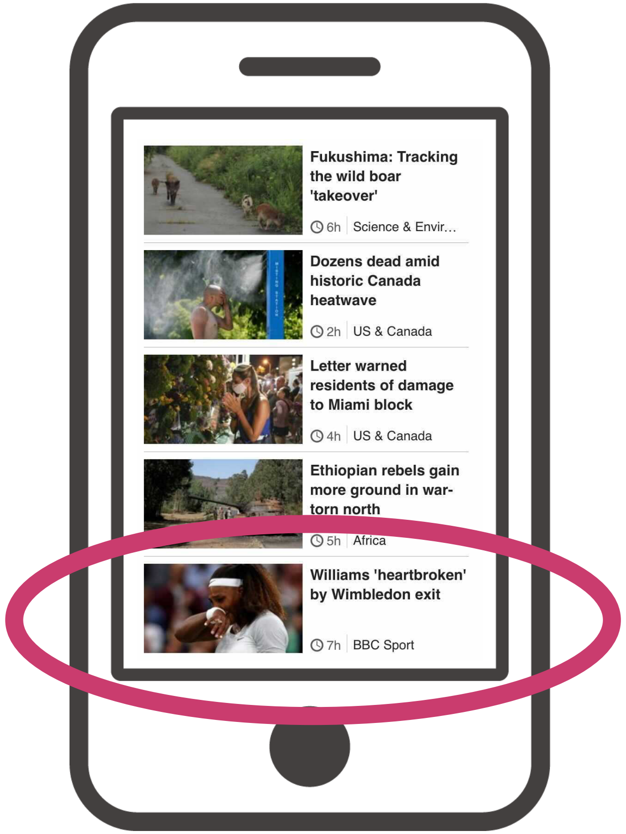
The news screenshot comes from BBC news in 2021/06/30.

Motivation: Example 2.

Suppose that a news article is **shared** in rankings for A/B Testing.

We note that reading times for this news is separated from each ranking.

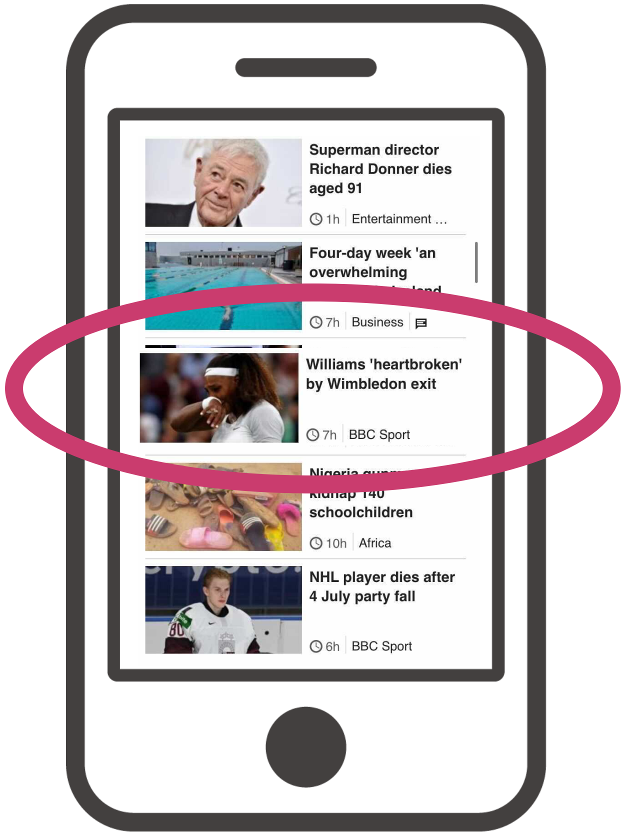
10.3



Ranking A

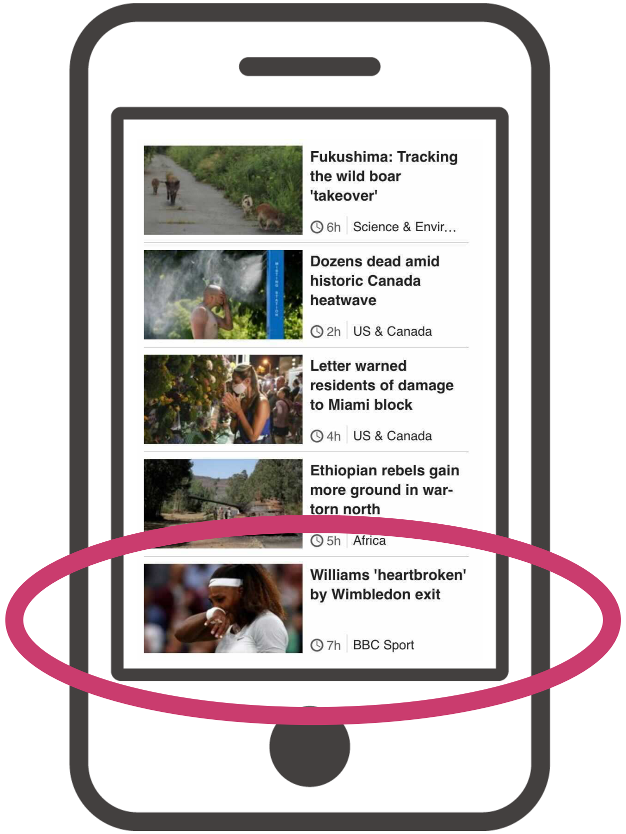
11.8

8.2



Ranking B

We can get more a accurate mean reading time by sharing reading times among each ranking.



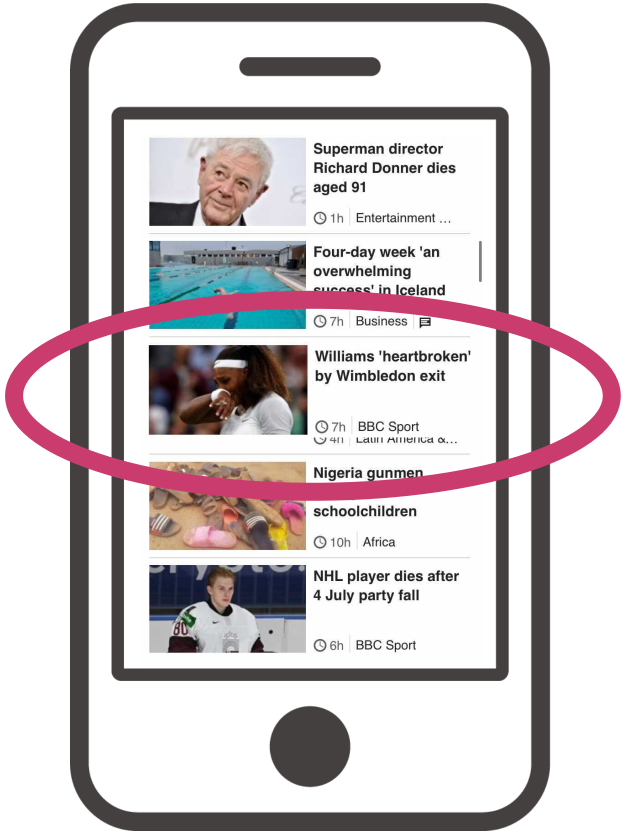
Ranking A

Reading Times

{10.3, **11.8, 8.2**}

‖ mean

10.1



Ranking B

Reading Times

{11.8, 8.2, **10.3**}

‖ mean

10.1

- **Variance Reduction**

  A technique commonly used to improve efficiency in online evaluation.

  – Oosterhuis et al., Taking the Counterfactual Online: Efficient and Unbiased Online Evaluation for Ranking, ICTIR2020

- **Interleaving**

  A method that interleave multiple rankings for efficient evaluation.

  Interleaving was reported to be 10 to 100 times more efficient than A/B testing.

  – Chapelle et al., Large-scale validation and analysis of interleaved search evaluation, Trans. Inf. Syst. 2012

  – Schuth et al., Multileaved comparisons for fast online evaluation, CIKM2014

Goal

Evaluating post-click metrics more efficiently in online experiments.

	Click-based metrics	Post-click metrics	Efficiency
A/B Testing	✅	✅	
Variance reduction [3]	✅		✅
Interleaving [4, 5]	✅		✅
Our method	✅	✅	✅

[3] Oosterhuis et al., Taking the Counterfactual Online:
Efficient and Unbiased Online Evaluation for Ranking, ICTIR2020
[4] Chapelle et al., Large-scale validation and analysis of interleaved search evaluation, Trans. Inf. Syst. 2012
[5] Schuth et al., Multileaved comparisons for fast online evaluation, CIKM2014

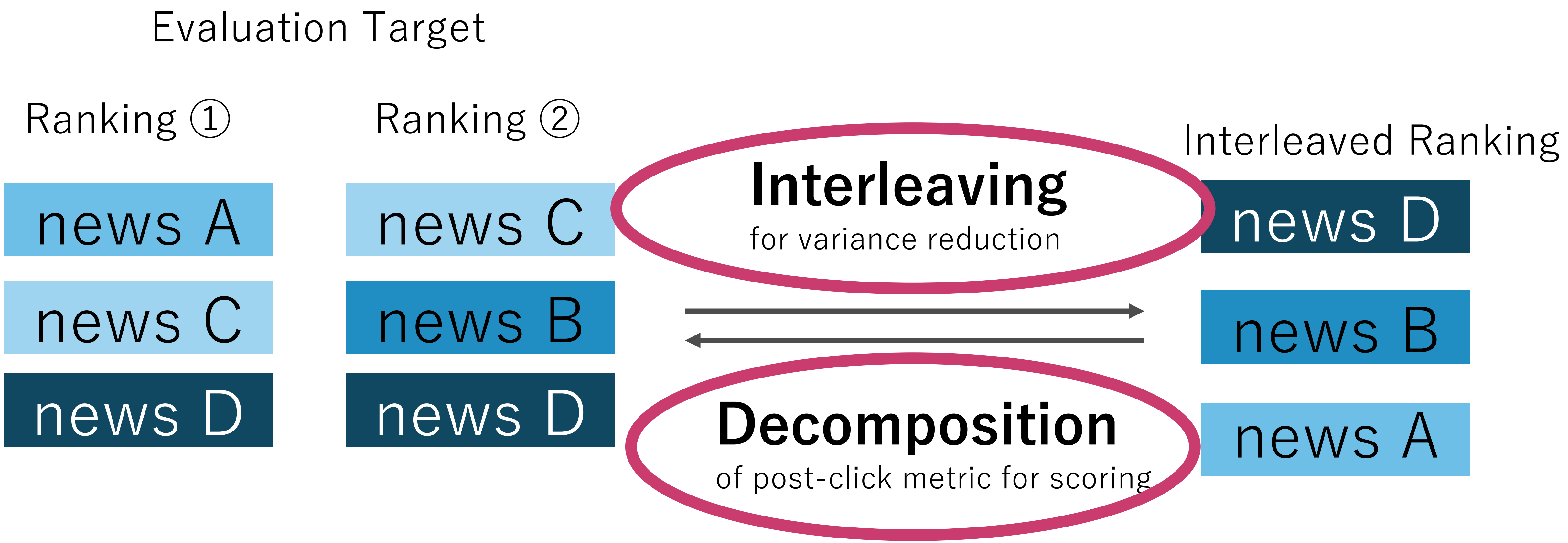
- **Input**
  - Rankings
- **Output**
  - Pairwise preference between rankings.
  - e.g., (ranking1 < ranking2), (ranking3 < ranking2), ...
  - Pairwise preference is judged by a post-click metric between rankings.
- **Evaluation Metric**
  - Consistency between predicted pairwise preference and ground truth preference.

$$E_{\mathrm{bin}} = \frac{1}{|R|(|R|-1)} \sum_{r_i, r_j \in R} \underbrace{\mathrm{sgn}(P_{i,j})}_{\text{Ground truth preference}} \neq \underbrace{\mathrm{sgn}(\overline{P}_{i,j})}_{\text{Predicted preference}},$$

Evaluation Target

Ranking ①	**Ranking ②**
news A	news C
news C	news B
news D	news D

The shade of the color indicates the amount of variance.


Evaluation Target
Ranking ①
Ranking ②
news A
news C
news D
news C
news B
news D
Interleaving
for variance reduction
Decomposition
of post-click metric for scoring
Interleaved Ranking
news D
news B
news A

Evaluation Target

Ranking ①	Ranking ②
news A	news C
news C	news B
news D	news D

Interleaving

for variance reduction

Decomposition

of post-click metric for scoring

Interleaved Ranking

- news D
- news B
- news A

Evaluation Target

Ranking ①	**Ranking ②**
news A	news C
news C	news B
news D	news D

The shade of the color indicates the amount of variance.

Sort news by variance in descending order.

Interleaved Ranking

news D

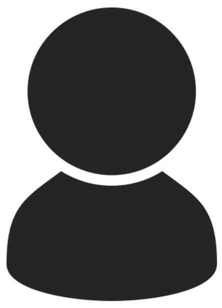
news B

news A

Evaluation Target

Ranking ①	**Ranking ②**
news A	news C
news C	news B
news D	news D

The shade of the color indicates the amount of variance.



Users are likely to click on the top news in the interleaved ranking.

Sort news by variance in descending order.

Interleaved Ranking

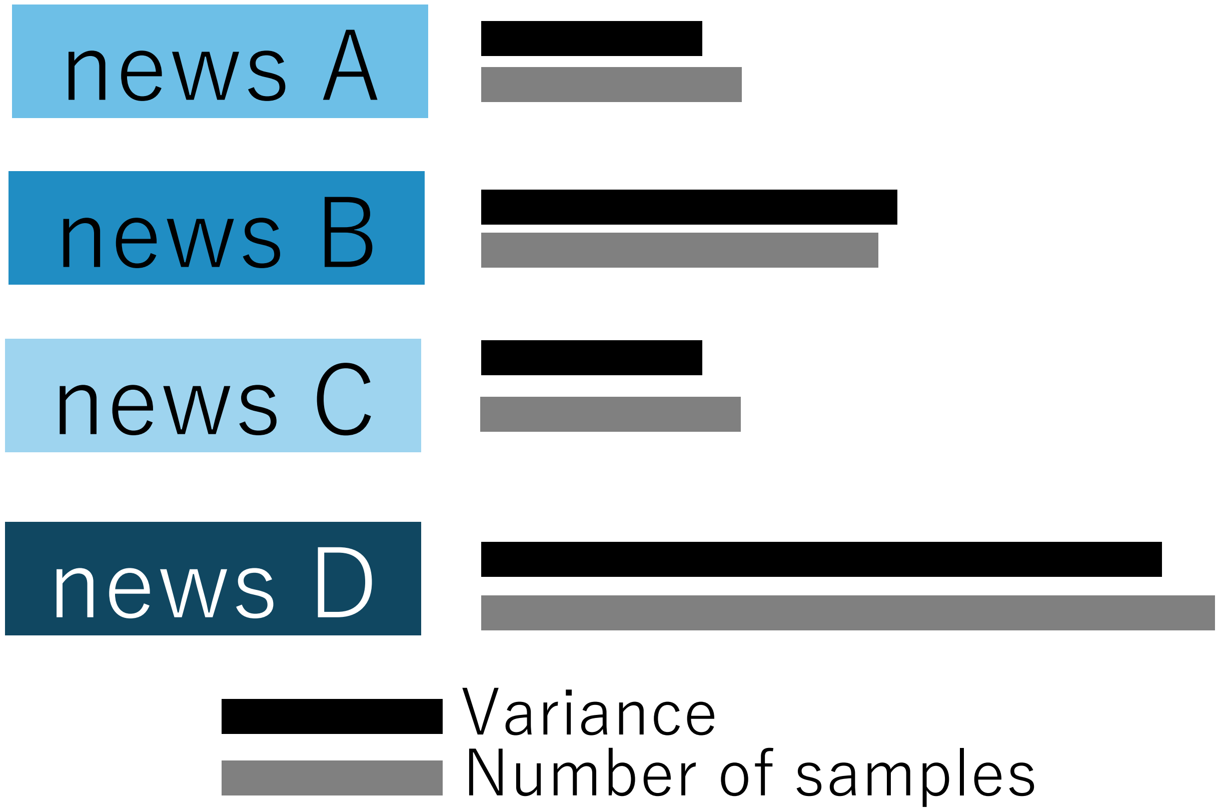
news D

news B

news A

The number of samples increased

→ **Variance Reduction**

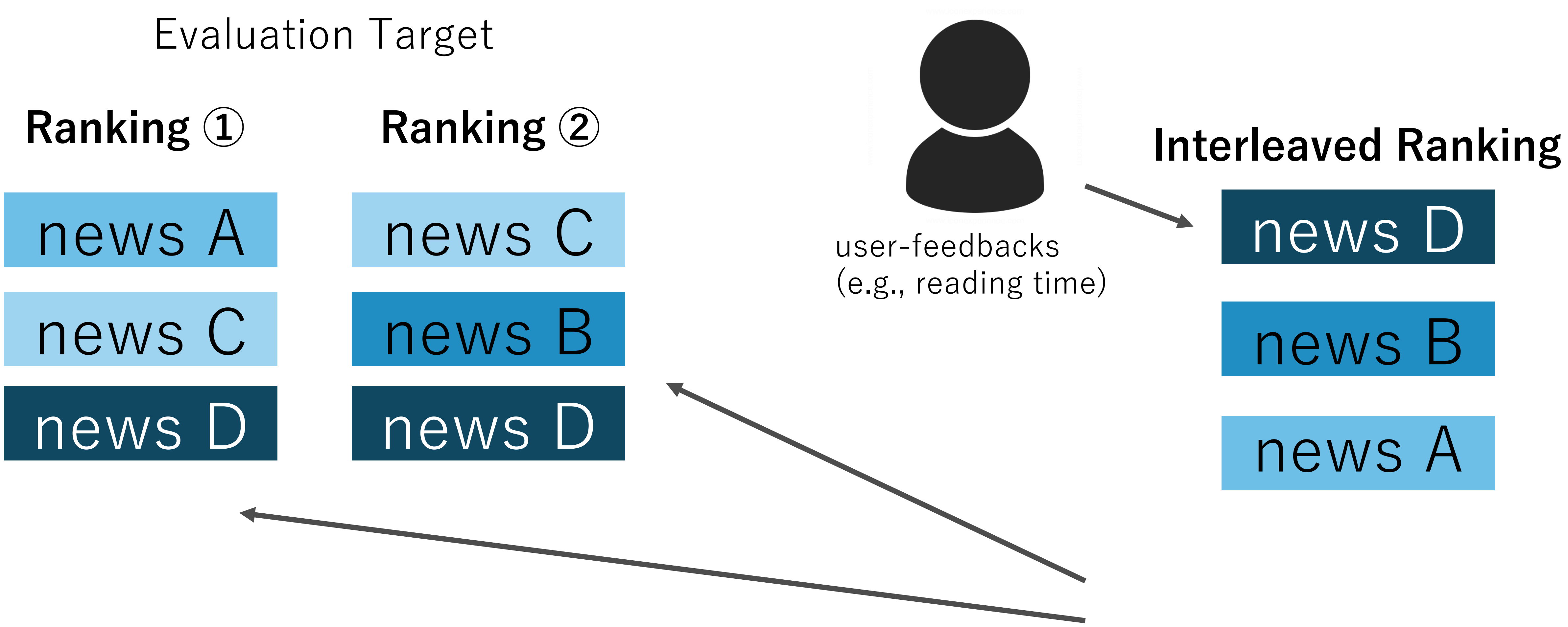


$$\frac{\sigma}{n}$$

σ = Variance of reading time

n = Number of samples

Adjust number of samples according to the variance → **Variance Reduction**



How can we predict preferences between rankings from the interleaved ranking?

Evaluation Target

Ranking ①

news A

news C

news D

Ranking ②

news C

news B

news D

Interleaving

for variance reduction

Decomposition

of post-click metric for scoring

Interleaved Ranking

news D

news B

news A

Expected
Reading Time

We introduce a scoring function based on **expected** reading time of each news.

Expected Reading Time = Click-through Rate (CTR) × Mean Reading Time

Expected value is composed of probability and mean value.

Method: Decomposition of Expected Value

$$\boxed{\text{Expected Reading Time}} = \boxed{\text{Click-through Rate (CTR)}} \times \textbf{Mean Reading Time}$$

We simply calculate mean reading time
by just averaging reading time for each news.

Expected Reading Time = Click-through Rate (CTR) × Mean Reading Time

Just average calculation ?

Expected Reading Time = Click-through Rate (CTR) × Mean Reading Time

Interleaved Ranking

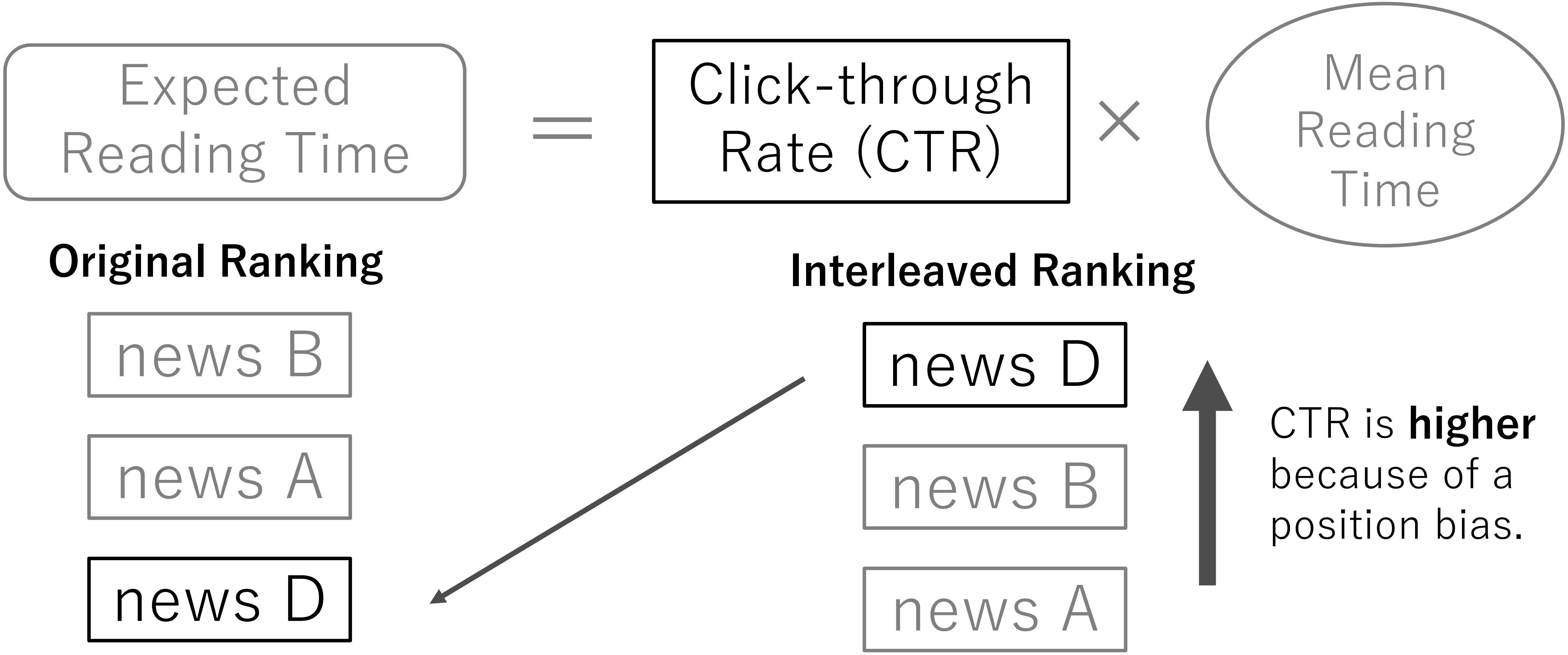
news D

news B

news A

CTR is higher because of a position bias.

Method: Decomposition of Expected Value



If we use raw a CTR calculated from the interleaved ranking for the original ranking, **overestimation** will be occurred.

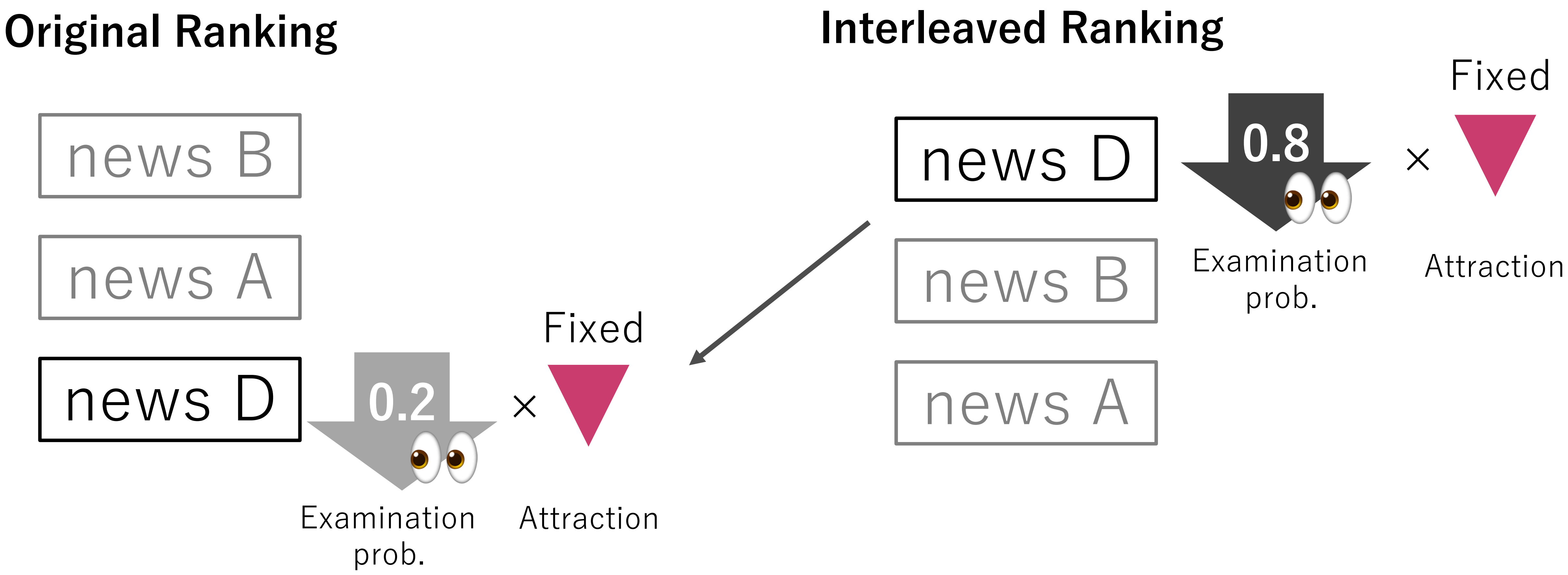
Expected Reading Time = Click-through Rate (CTR) × Mean Reading Time

Introducing click model.

CTR can be further decomposed into two variables: examination and attraction.

Click-through Rate (CTR) = **Examination Prob.** × Attraction

Given the examination, how a user likely to click.



By adjusting the **examination probability** for each rank of original ranking, we can **avoid under- or over-estimation of CTR.**

Method: Summary

Evaluation Target

Ranking ①	Ranking ②
news A	news C
news C	news B
news D	news D

Interleaving
for variance reduction

→

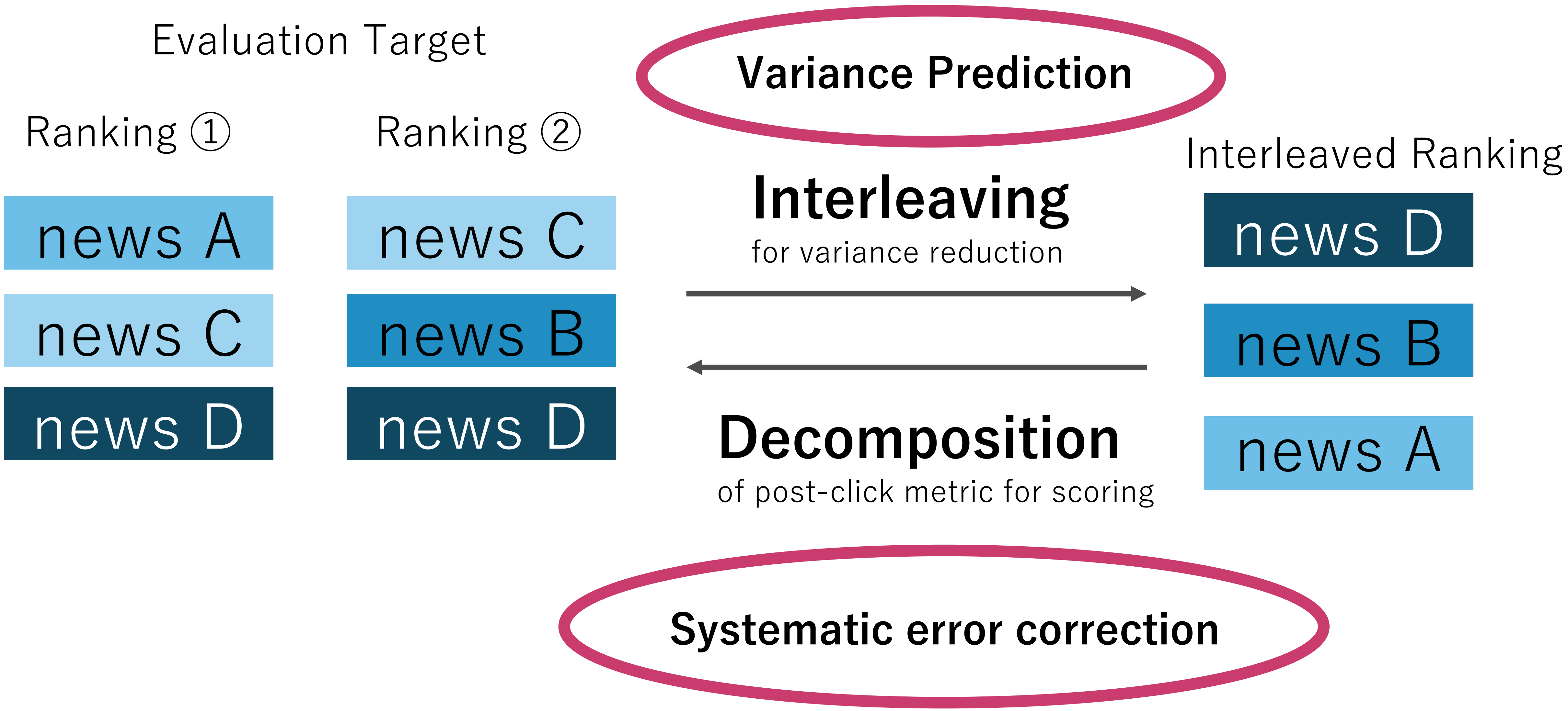
←

Decomposition
of post-click metric for scoring

Interleaved Ranking

- news D
- news B
- news A

Method: Two Stabilization techniques



Evaluation Target

Variance Prediction

Ranking ①

- news A
- news C
- news D

Ranking ②

- news C
- news B
- news D

Interleaving
for variance reduction

Decomposition
of post-click metric for scoring

Interleaved Ranking

- news D
- news B
- news A

Systematic error correction

Evaluation Target

Ranking ①

news A

news C

news D

Ranking ②

news C

news B

news D

Sort news by variance in descending order.

Interleaved Ranking

news D

news B

news A

Problem

In the interleaving procedure, we need to estimate population variances for each news.
Then, we estimate the population variance by a sample variance.
However, the estimation can be inaccurate, especially when the number of sample is small.

Evaluation Target

Ranking ①

news A

news C

news D

Ranking ②

news C

news B

news D

Sort news by variance in descending order.

Interleaved Ranking

news D

news B

news A

Solution

We utilize **predicted variance** by using machine learning model when the number of sample is small.

Evaluation Target

Ranking ①

news A
news C
news D

Ranking ②

news C
news B
news D

Variance Prediction

Interleaving
for variance reduction

Decomposition
of post-click metric for scoring

Interleaved Ranking

news D
news B
news A

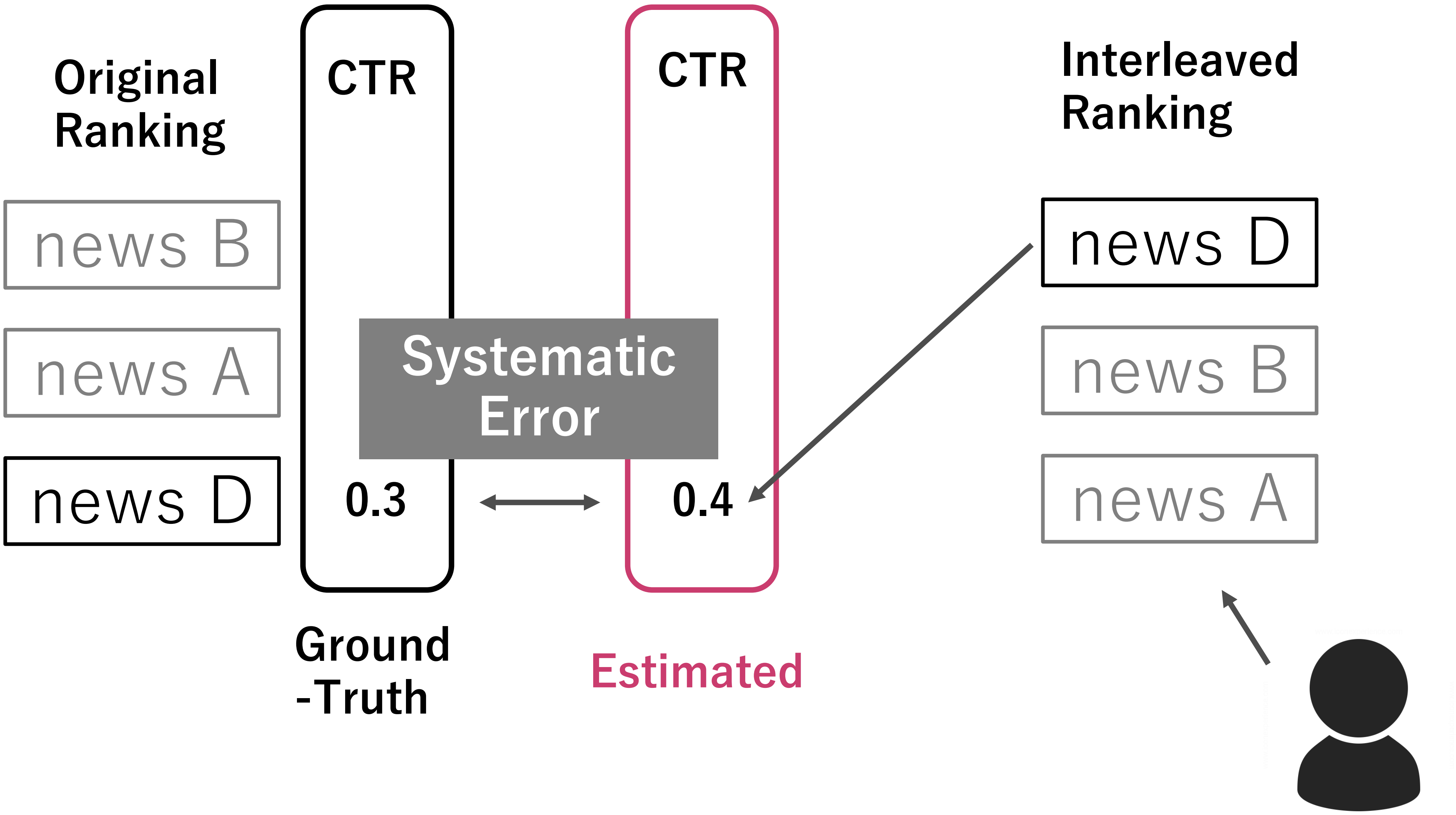
Systematic error correction

Expected Reading Time = CTR × Mean Reading Time

Problem
- We **estimated** CTR by assuming a click model.
- If this estimation has huge error, the expected value also has error.

||

Systematic Error


Original Ranking
news B
news A
news D
CTR
0.3
Ground -Truth
Systematic Error
CTR
0.4
Estimated
Interleaved Ranking
news D
news B
news A

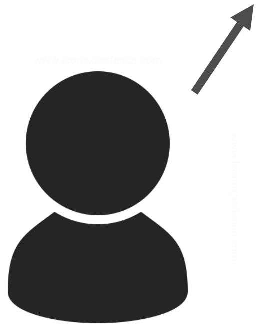
Original Ranking

news B

news A

news D

CTR	CTR	CTR
0.3	0.32	0.4
Ground-Truth	Observed	Estimated



To get observed-CTR, we show original ranking to users.
Observed CTR equals ground-truth CTR with large impressions.

Solution Combining observed CTR from original ranking and estimated CTR* from interleaved ranking to correct systematic error.

$$\text{CTR} = \text{weight} \times \boxed{\text{Observed-CTR}} + (1\text{-weight}) \times \boxed{\text{CTR*}}$$

Observed from original ranking

Estimated from Interleaved ranking

To answer research questions:

1. Can DIRV (proposed method) identify preferences between rankings more efficiently than other methods?
2. How does the variance prediction technique affect the variance reduction?
3. How does the error correction technique affect the evaluation accuracy?

- A/B Testing
  – A simple and practical baseline.
- Team-draft multileaving (TDM)
  – One of the most popular multileaving method.
  – Using modified scoring function for post-click evaluation [6].
- Proposed method (DIRV)
  – w/o variance reduction technique
  – w/o systematic error correction technique
  – with both techniques

[6] Schuth et al. Predicting search satisfaction metrics with interleaved comparisons, SIGIR2015

- **Simulation-based dataset**
  - LETOR: Learning to rank dataset from Microsoft.
  - EC: Artificially generated e-commerce dataset.
  - News: News service dataset from Gunosy.
- **Real-service dataset**
  - News: Interleaved ranking dataset from Gunosy.
  - More close to the online controlled setting than the simulation-based setting.

Dataset	Attraction	Post-click Value
LETOR	relevance	relevance
EC	random	random
News	service log	service log

We used cascade click model to simulate user examination.

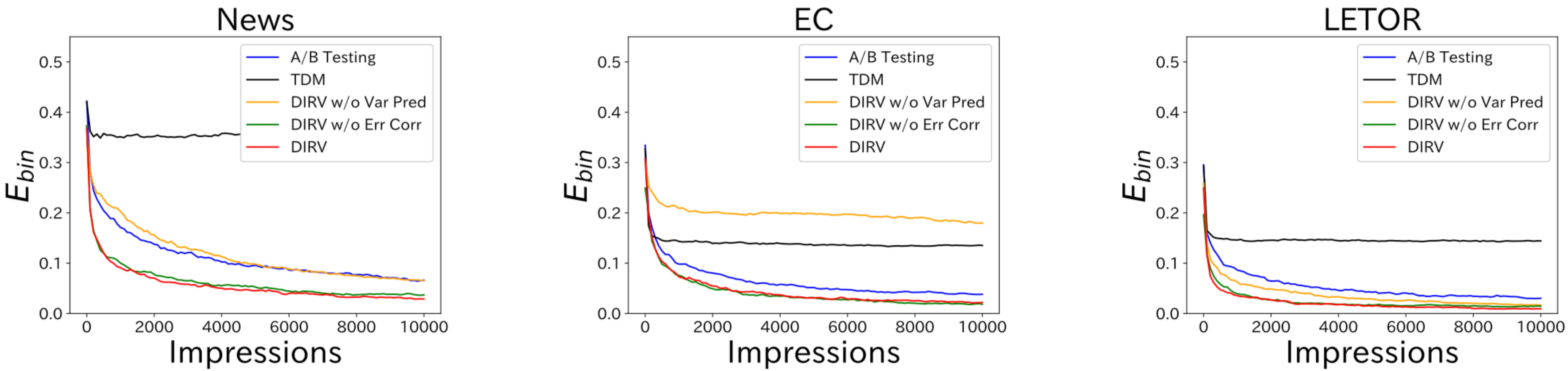
- **Evaluation metric**

Consistency between predicted pairwise preference and ground truth preference with limited user actions.

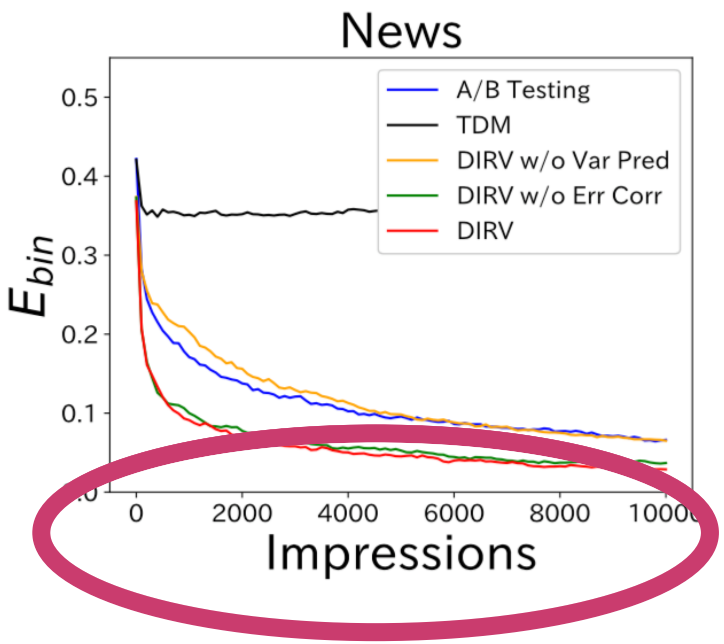
$$E_{\text{bin}} = \frac{1}{|R|(|R|-1)} \sum_{r_i, r_j \in R} \operatorname{sgn}(P_{i,j}) \neq \operatorname{sgn}(\overline{P}_{i,j}),$$

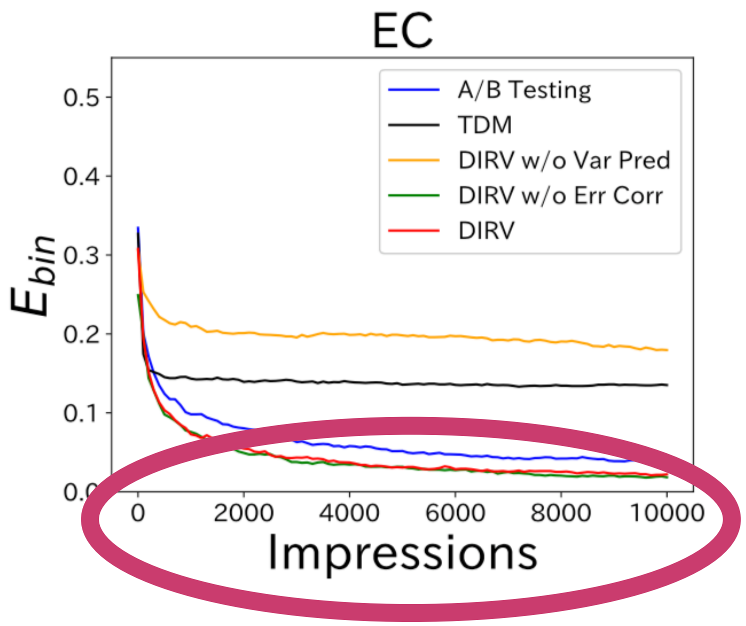
Ground truth preference Predicted preference

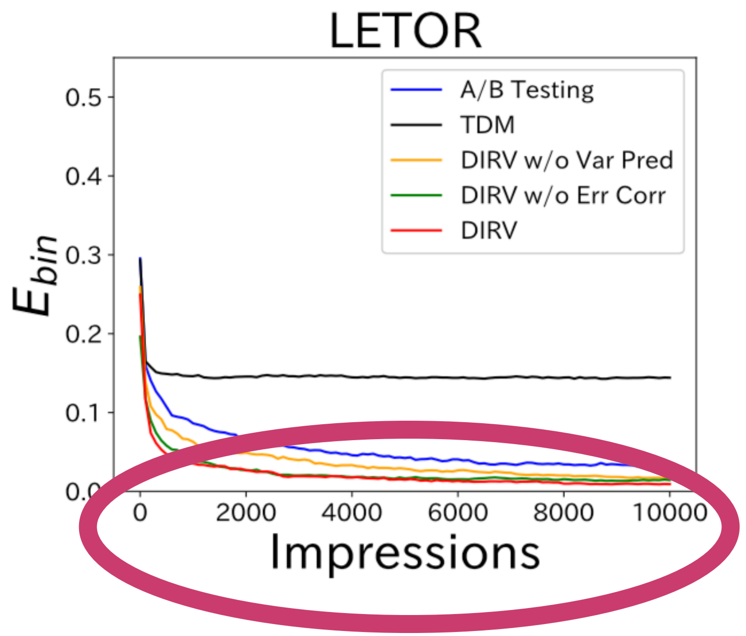
Can DIRV identify preferences between rankings more efficiently than other methods?



Can DIRV identify preferences between rankings more efficiently than other methods?

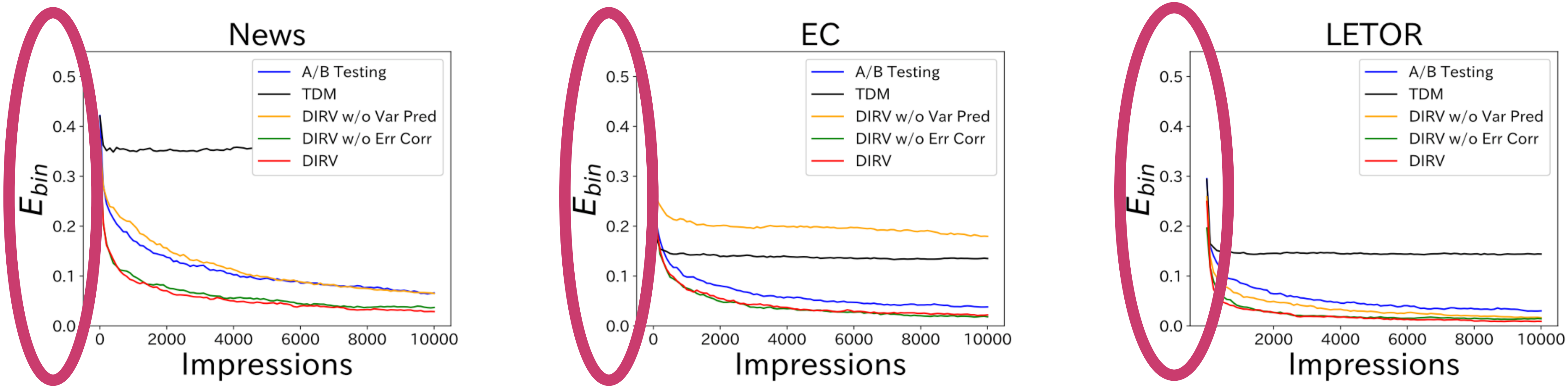






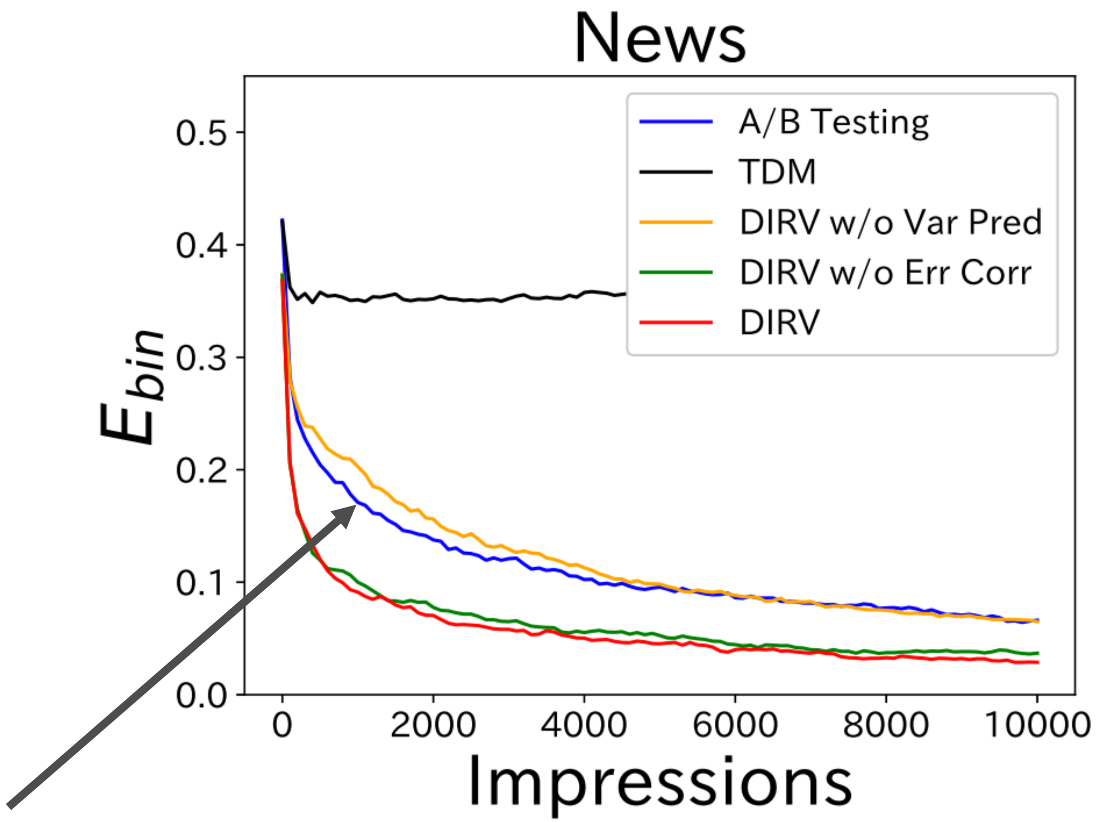
These three figures show the efficiency result of comparison methods.
The x-axis shows the number of impressions.

Can DIRV identify preferences between rankings more efficiently than other methods?



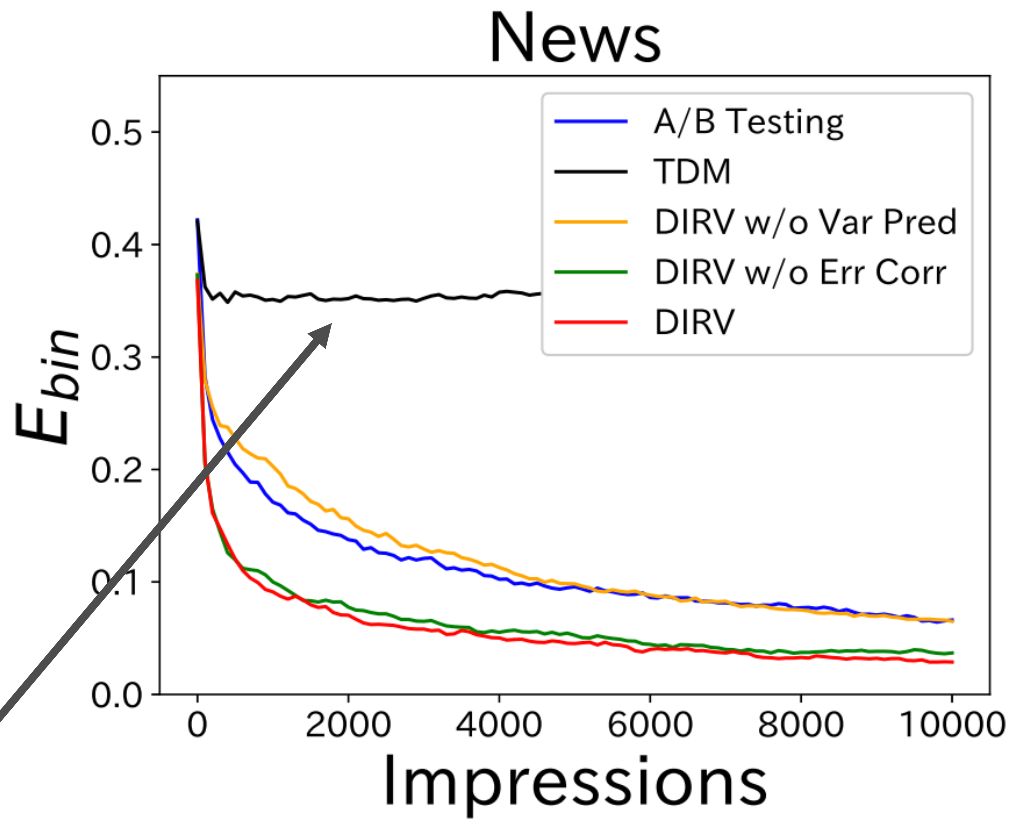
The y-axis shows the binary error.
This binary error is that the lower, the better.

Can DIRV identify preferences between rankings more efficiently than other methods?



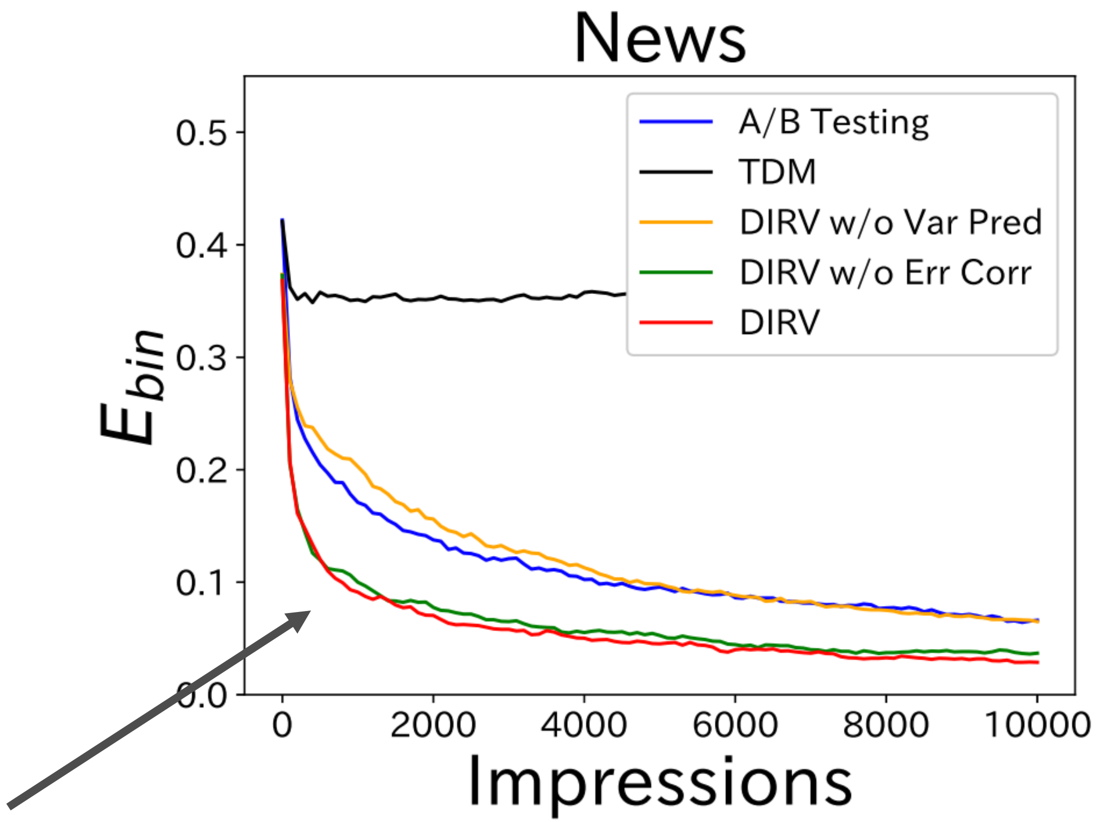
The blue line shows the A/B testing result

Can DIRV identify preferences between rankings more efficiently than other methods?



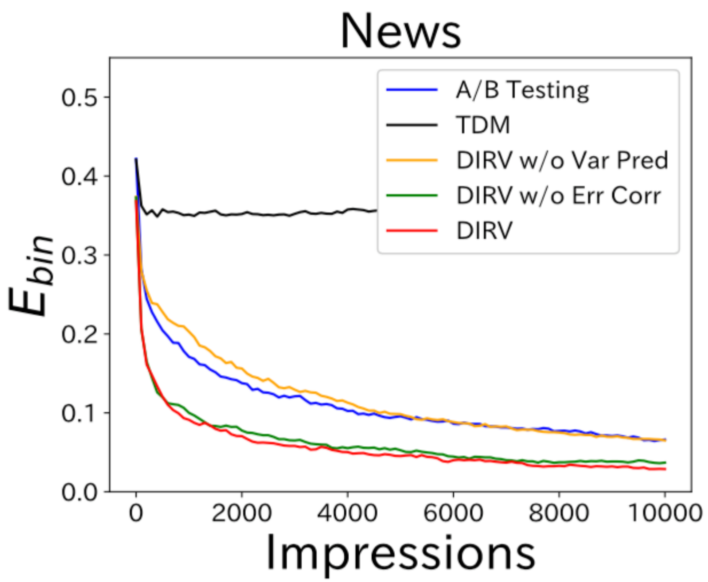
The black line shows the TDM result

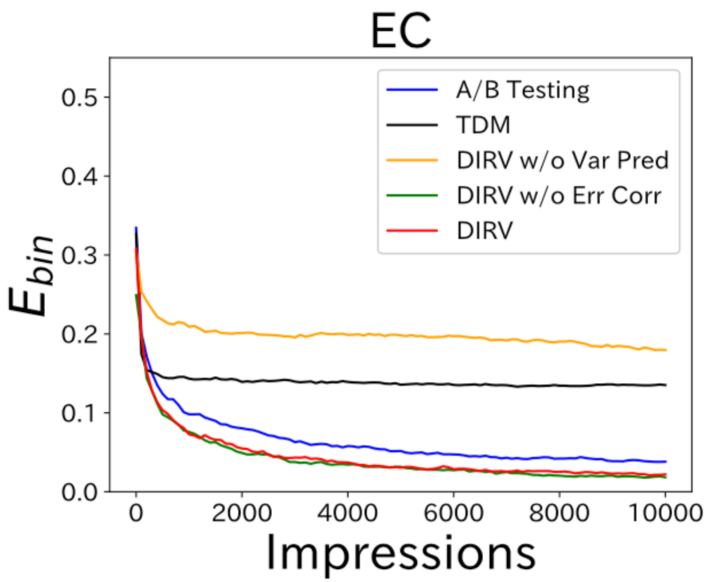
Can DIRV identify preferences between rankings more efficiently than other methods?

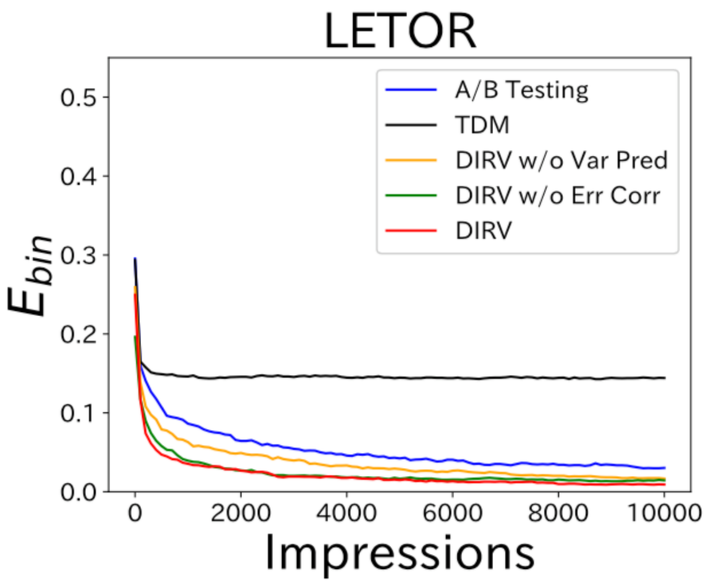


The red line shows the DIRV with both techniques.

Can DIRV identify preferences between rankings more efficiently than other methods?







For all of the datasets, DIRV had the lowest binary error for each impression. DIRV outperformed the existing methods in efficiency.

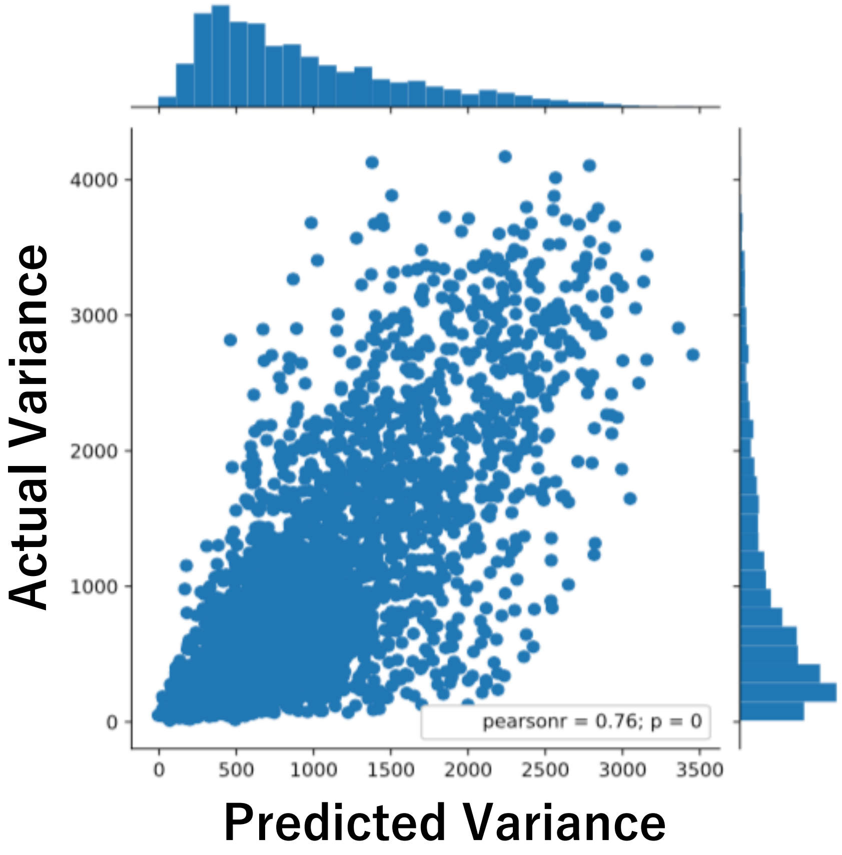
How accurate the predicted variance is?

Table 3: Features and importance

Feature	Importance
Category ID to which the article belongs	879
Supplier ID of the article	2,342
Content length of the article	1,854
Title length of the article	1,045

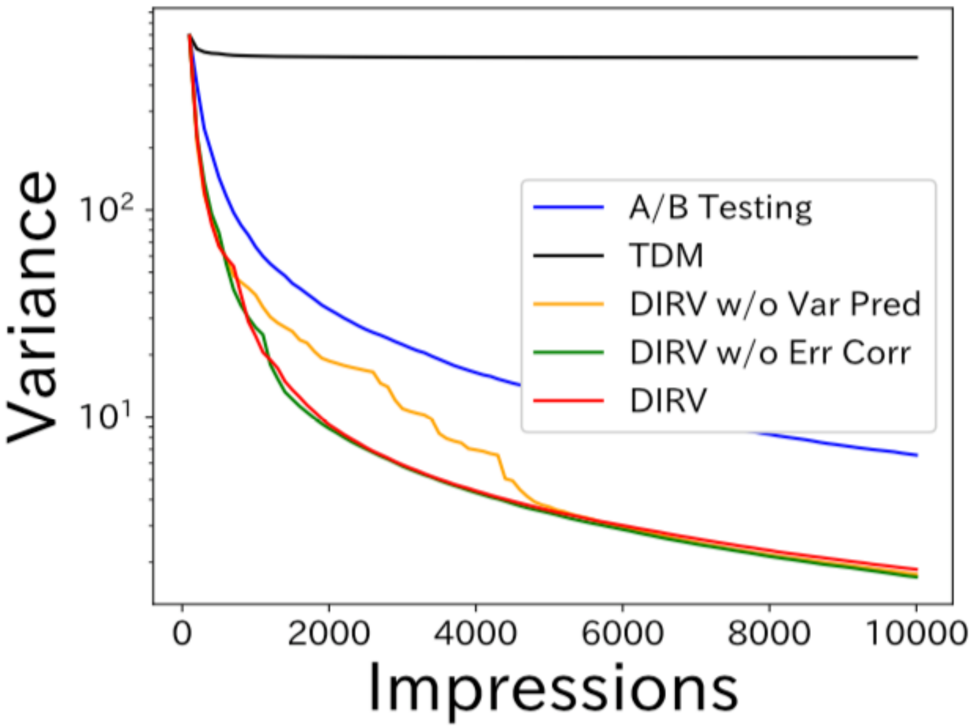
We trained a tree-based model by using features in Table 3.
As a result, supplier id and content length are the top-2 important features.

How accurate the predicted variance is ?



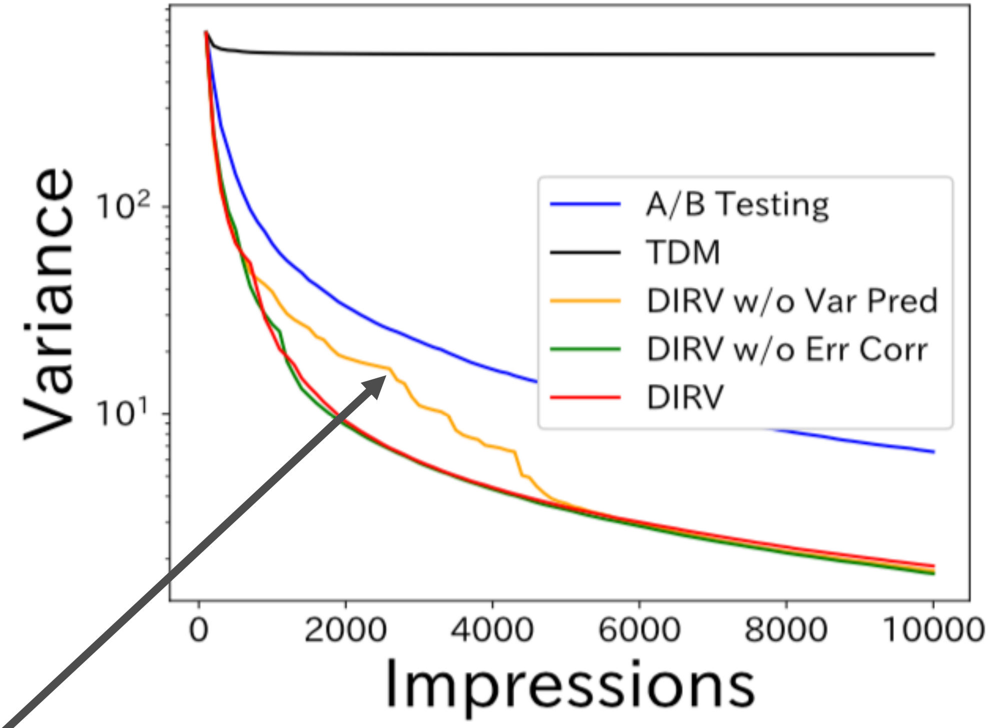
There exists a correlation between predicted values and actual values. We note that Pearson's correlation coefficient was 0.76.

How does the variance prediction technique affect the variance reduction?



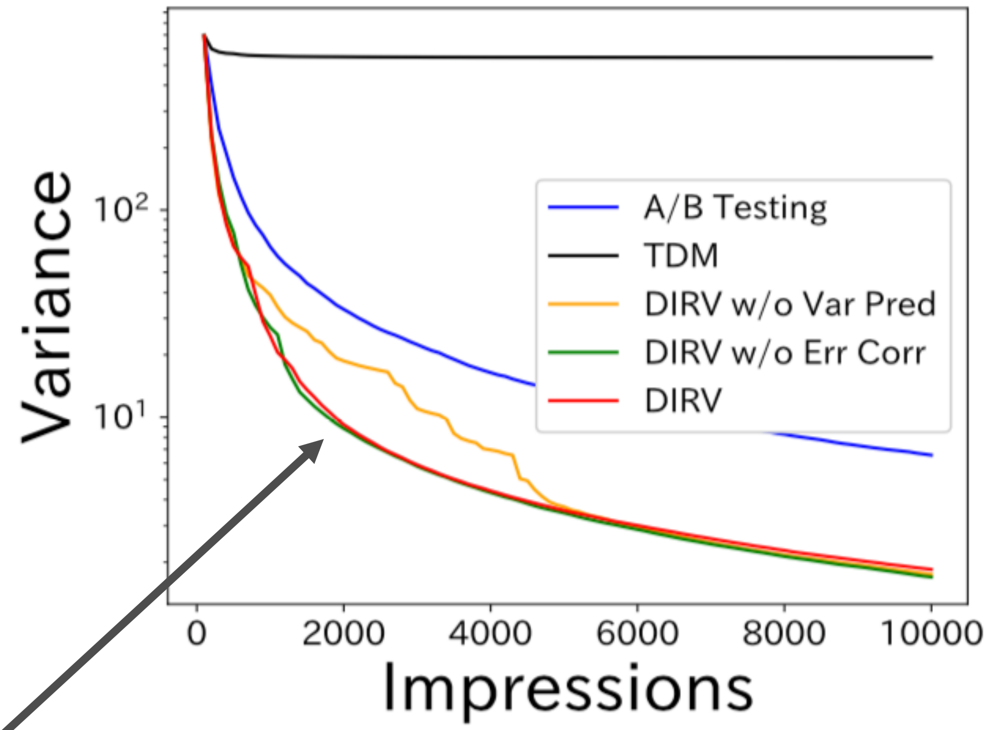
The x-axis is the number of impressions, and the y-axis is the variance.

How does the variance prediction technique affect the variance reduction?



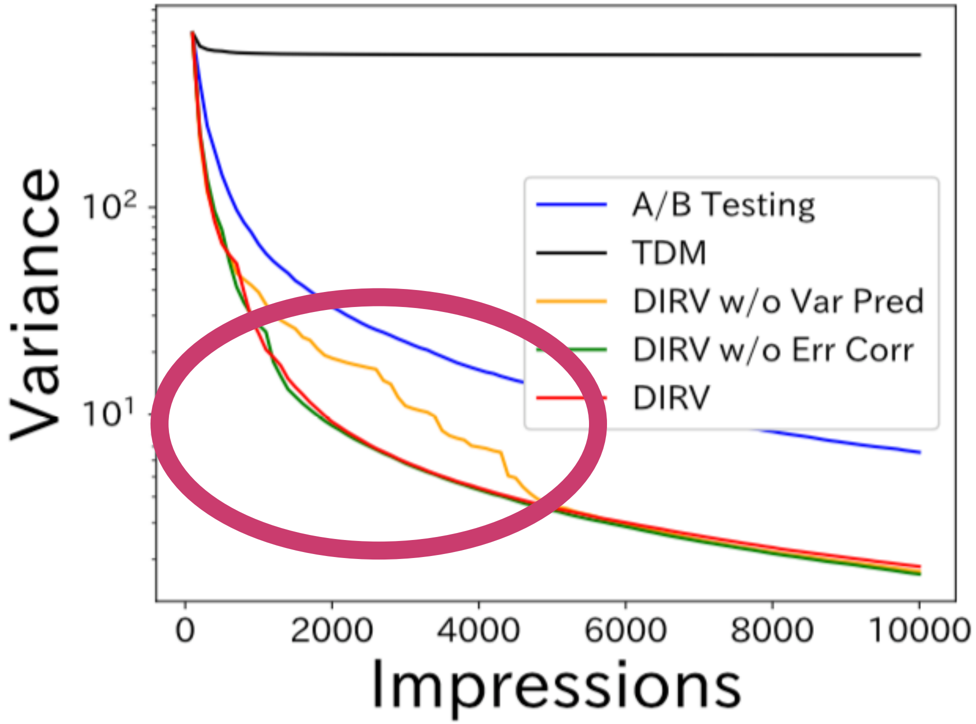
The yellow line shows the variance transition of DIRV **without** the variance prediction technique.

How does the variance prediction technique affect the variance reduction?



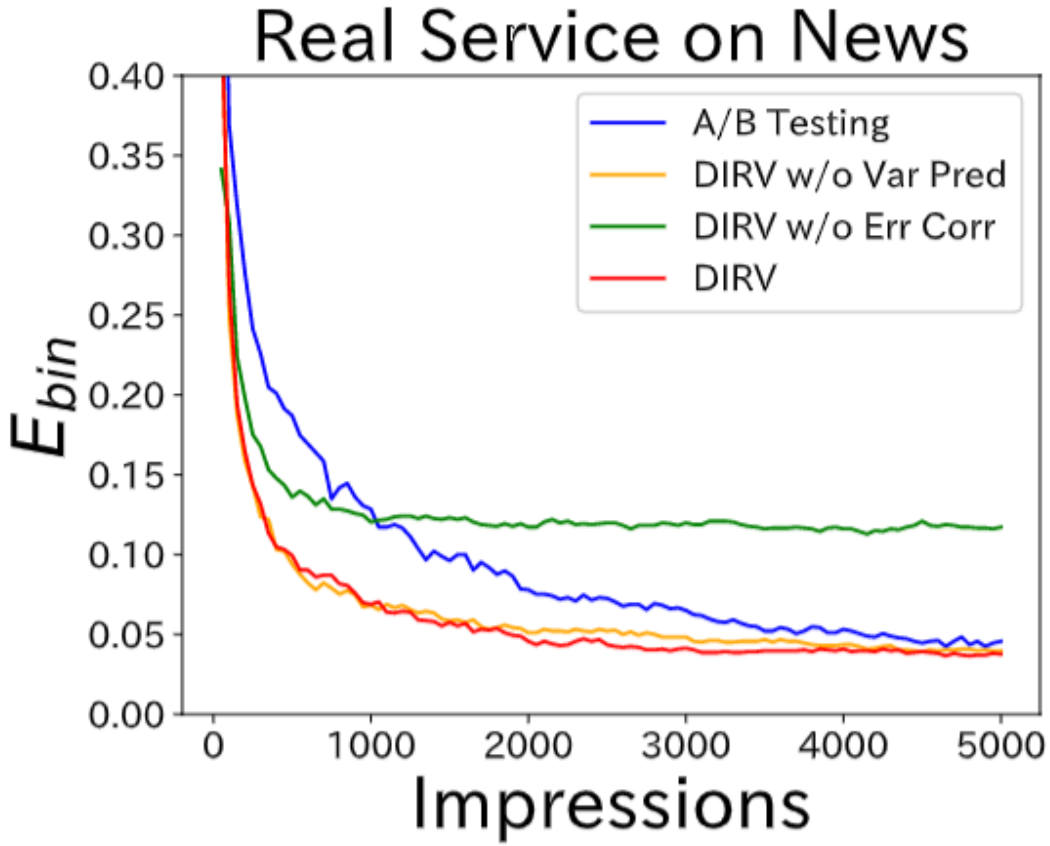
The red line shows the variance transition of DIRV **with** the variance prediction technique.

How does the variance prediction technique affect the variance reduction?



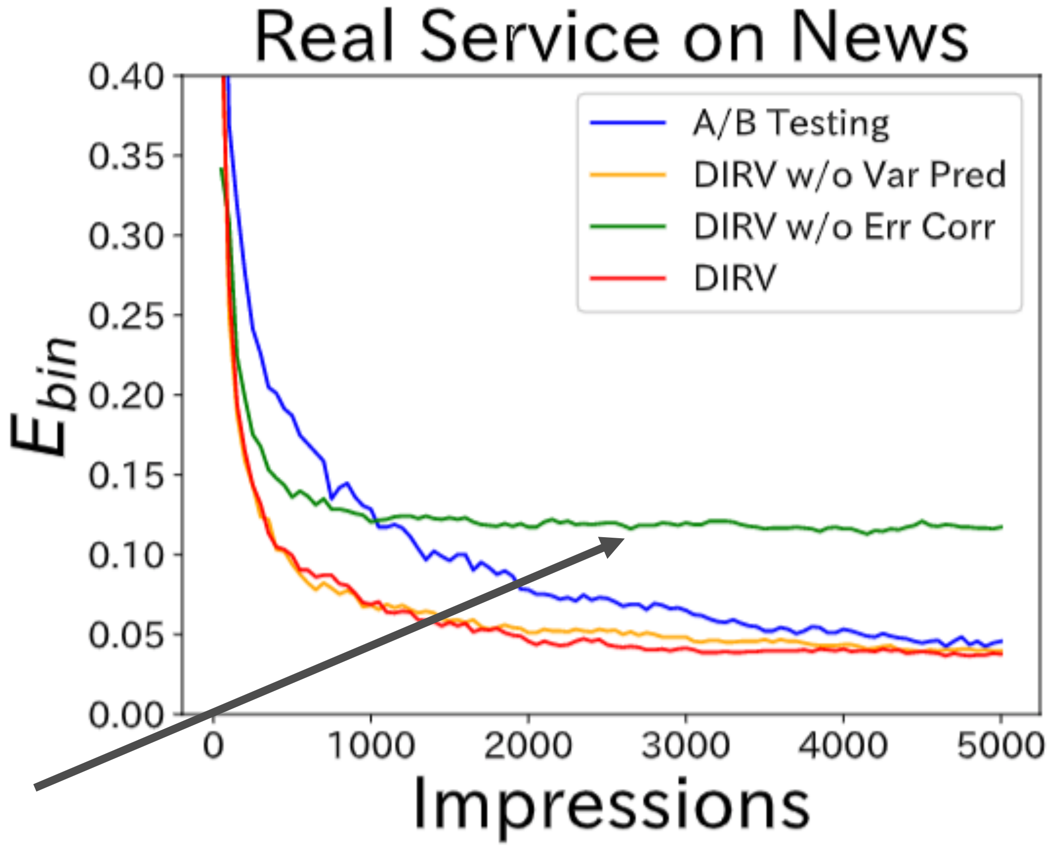
The variance was reduced efficiently using the variance prediction technique when the number of impression was small.

How does the error correction technique affect the evaluation accuracy?



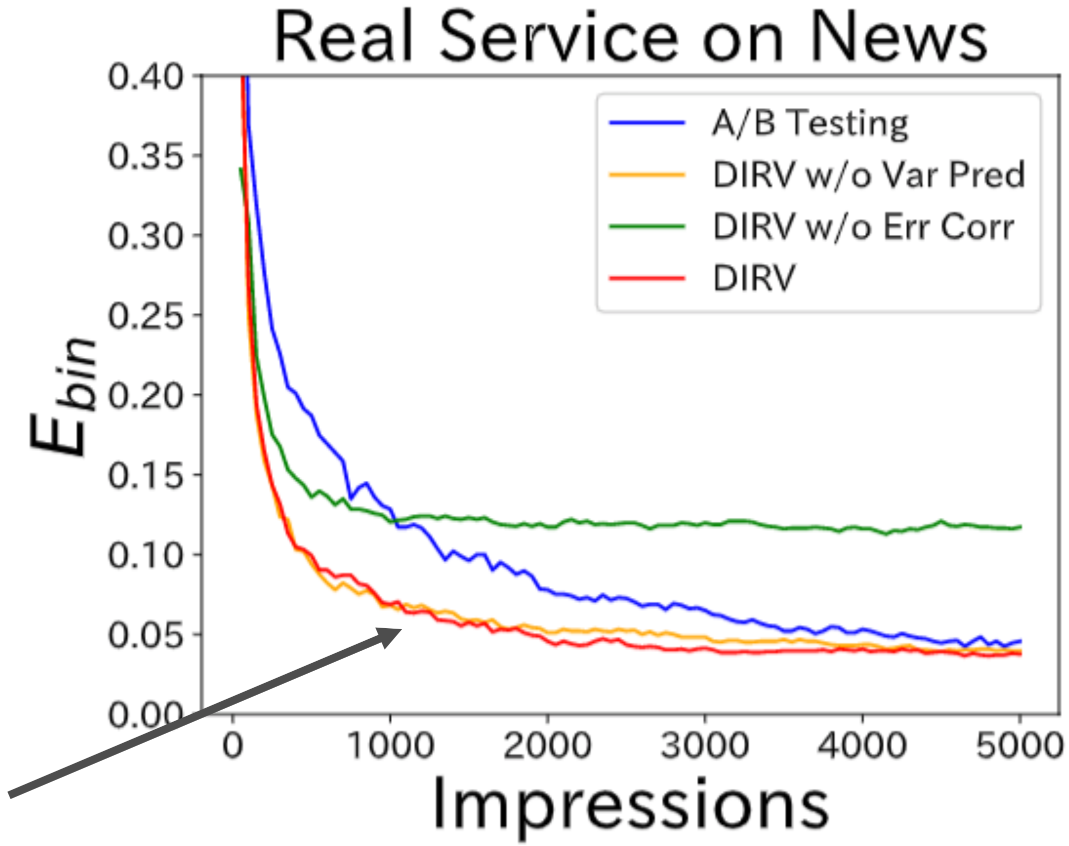
This figure shows the accuracy in the real service setting.

How does the error correction technique affect the evaluation accuracy?



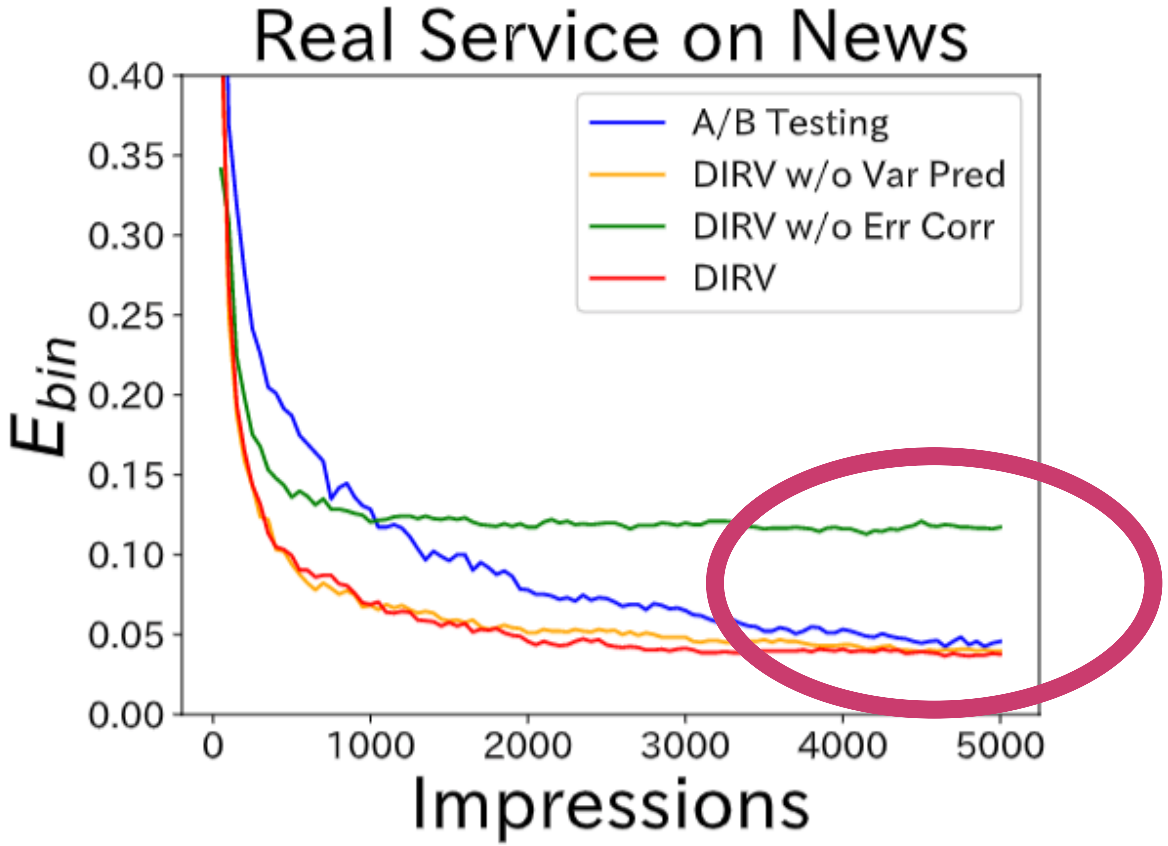
DIRV **without** error correction technique (green line).

How does the error correction technique affect the evaluation accuracy?



DIRV **with** error correction technique (red line).

How does the error correction technique affect the evaluation accuracy?



DIRV with error correction technique (red line) was more accurate than the DIRV without error correction technique (green line).

- To efficiently compare post-click metrics of multiple rankings, we proposed an interleaving method (DIRV)
  - decomposes the post-click metric measurement
  - preferentially exposes items with high population variance
- We extensively evaluated DIRV using both simulation and real service settings. The results demonstrated its high efficiency.
- We proposed additional techniques to boost the DIRV and demonstrated that the technique was empirically effective.