THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

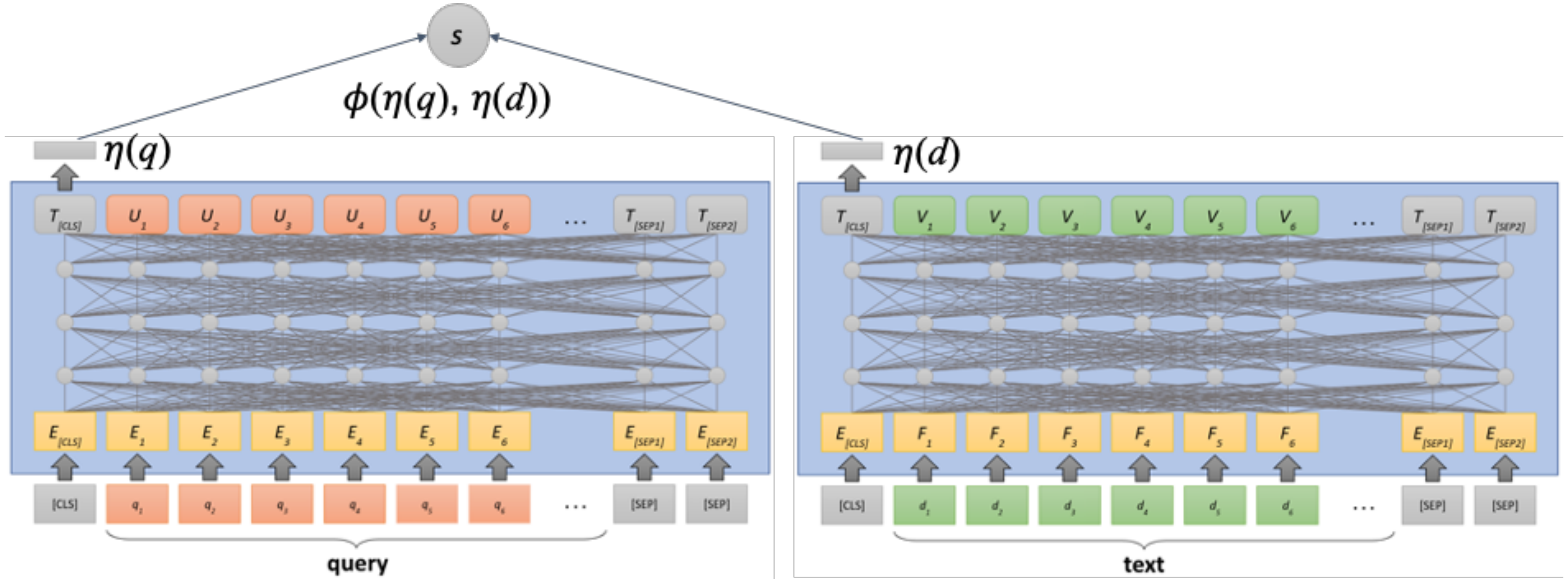
# Implicit Feedback for Dense Passage Retrieval: A Counterfactual Approach

Shengyao Zhuang, Hang Li, Guido Zuccon

{s.zhuang, h.li, g.zuccon}@uq.edu.au
ielab, The University of Queensland, Australia
www.ielab.io

ie lab

# Dense Retrievers

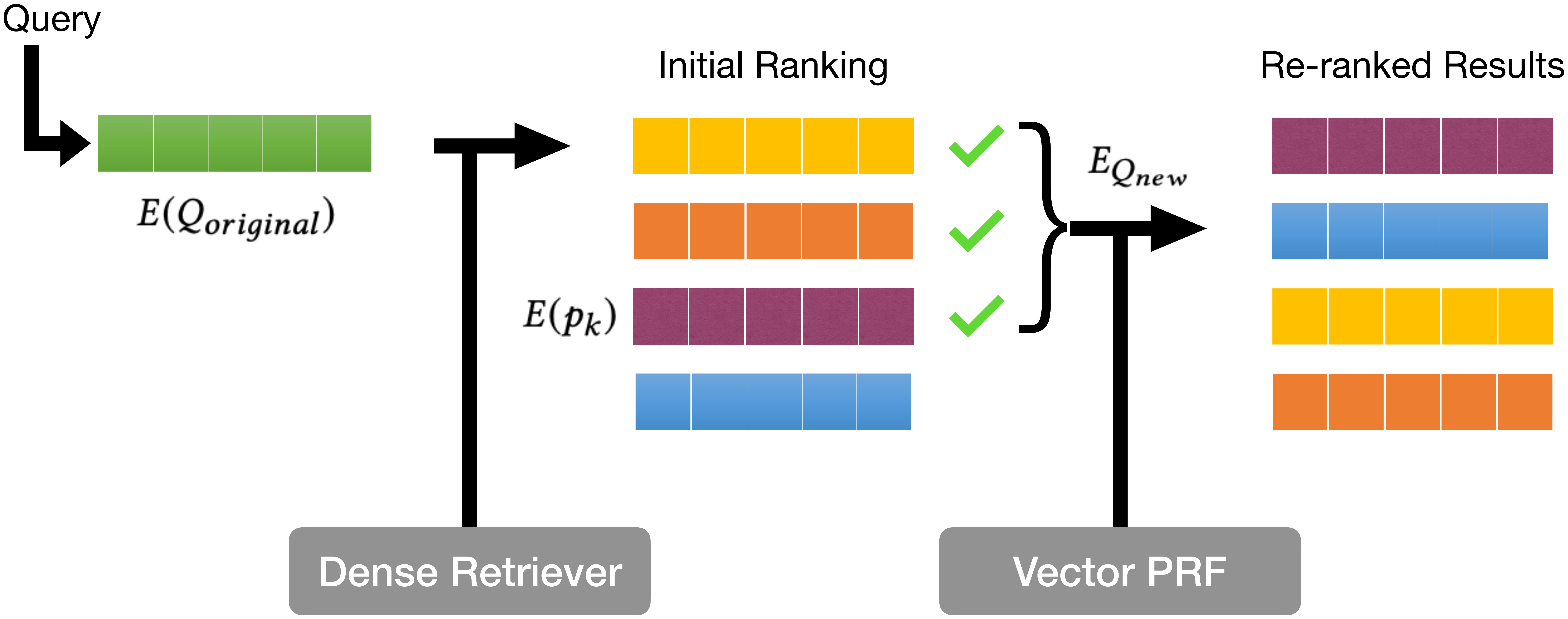



- Offline (Training):
  - Train BERT encoder to create representation of relevant docs and queries that are close to each other
  - Create vector representation of documents with BERT encoder
- Online (Inference):
  - create vector representation of query with BERT encoder
  - compute vector similarity between query and document vectors

# Dense Retrievers require extensive labelled data

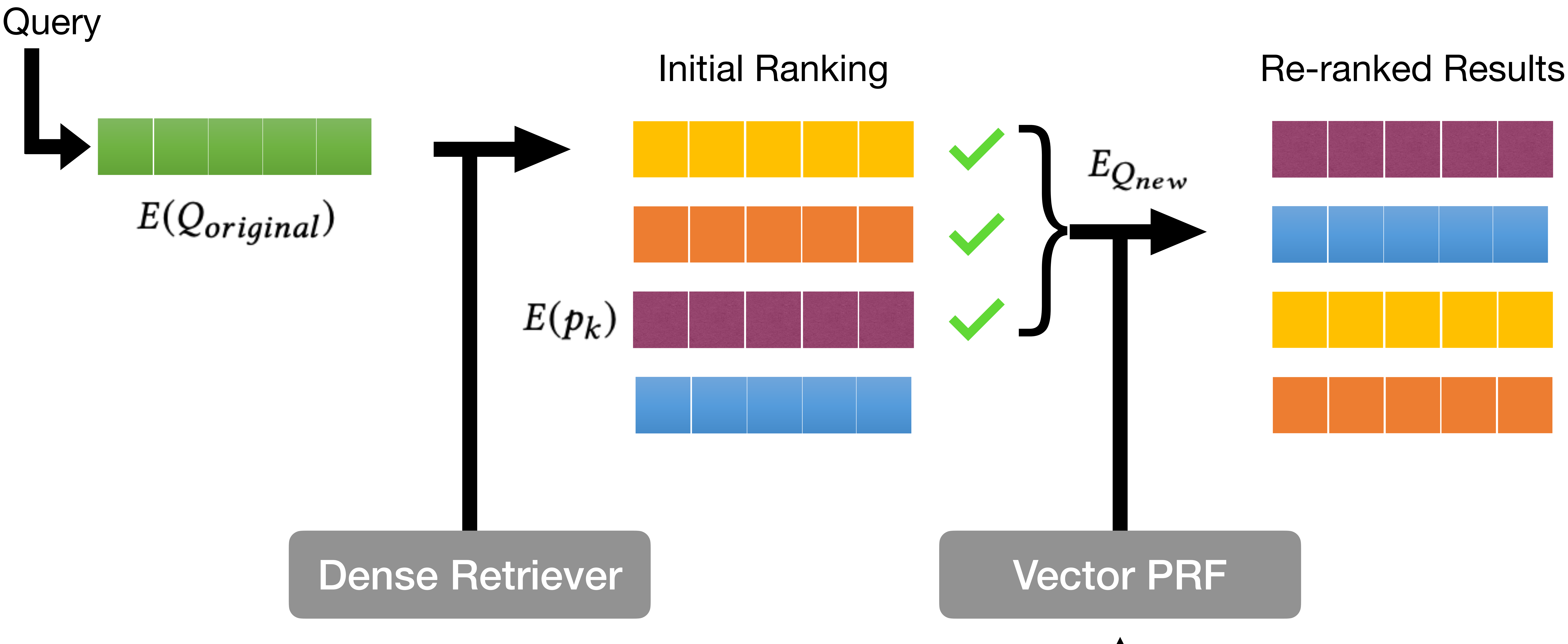
- Labelled data can be expensive to obtain (e.g. domain specific), at times not possible (e.g. for private data)
- In this paper:
  - Can we use **implicit feedback** collected by a search engine (click-through data) **to inform DRs**?
  - Key idea: adapt current pseudo relevance feedback method for DRs (Vector PRF) to deal with implicit feedback (clicks)

# Vector PRF (VPRF) for DRs



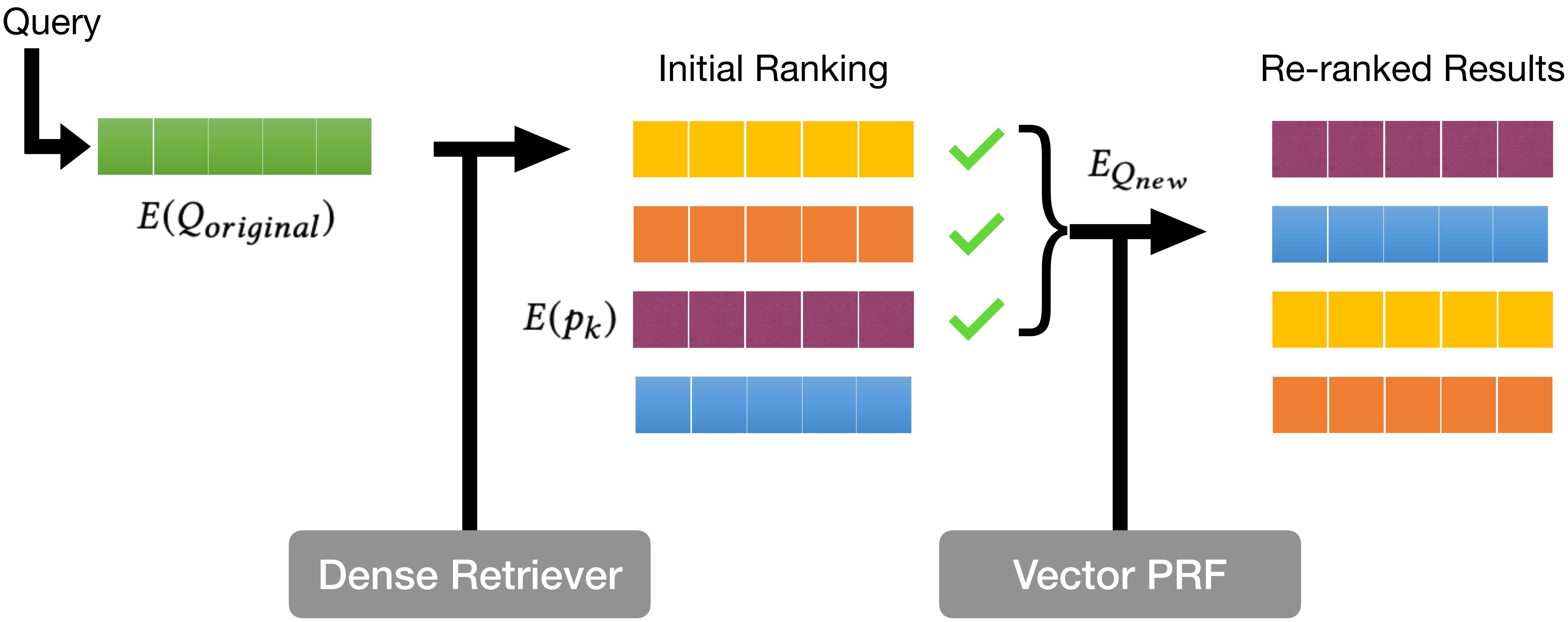

Rocchio: $E_{Q_{new}} = \alpha * E(Q_{original}) + \beta * Avg(E(p_1), ..., E(p_k))$

# Vector PRF (VPRF) for DRs




Rocchio: $E_{Q_{new}} = \boxed{\alpha * E(Q_{original})} + \beta * Avg(E(p_1), ..., E(p_k))$

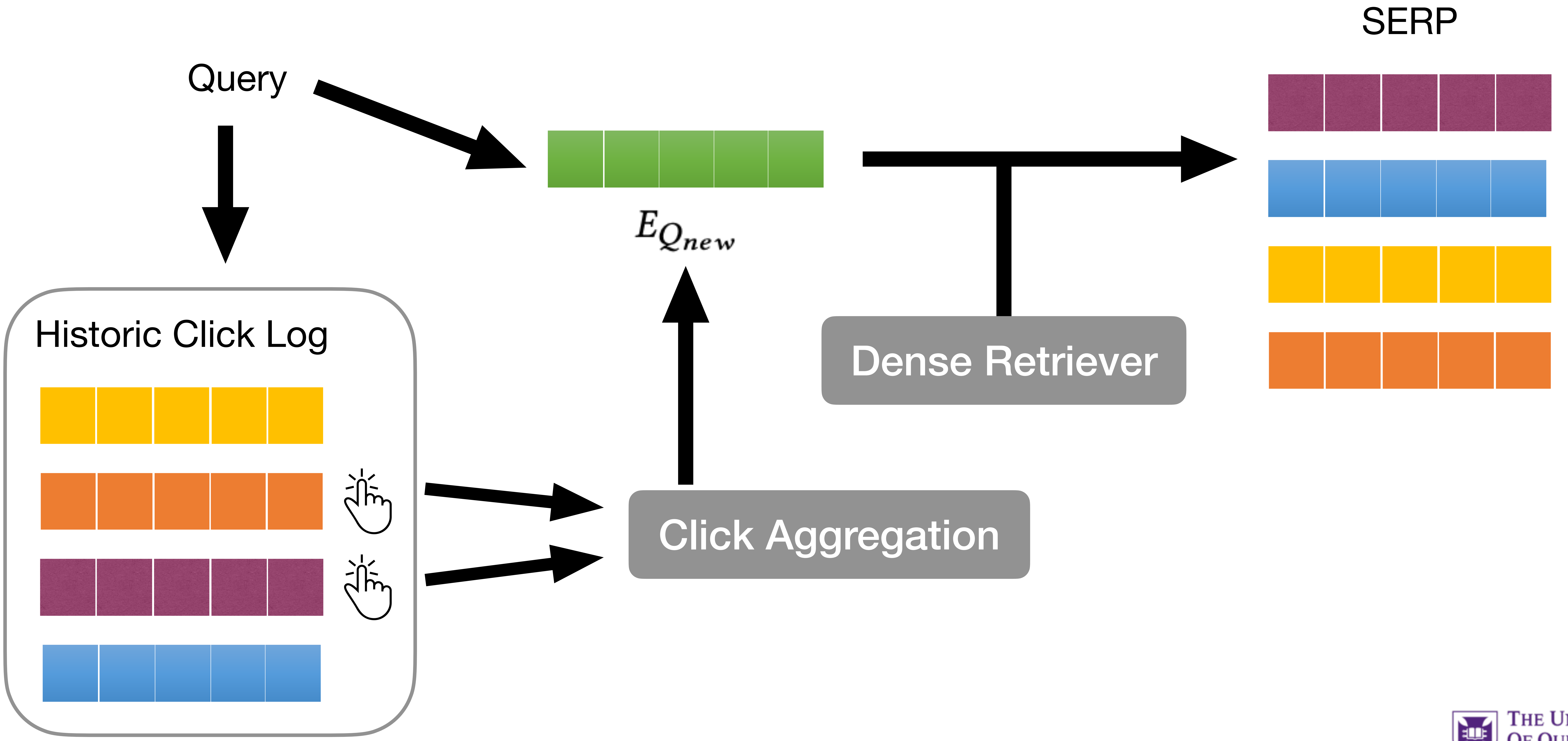
# Vector PRF (VPRF) for DRs



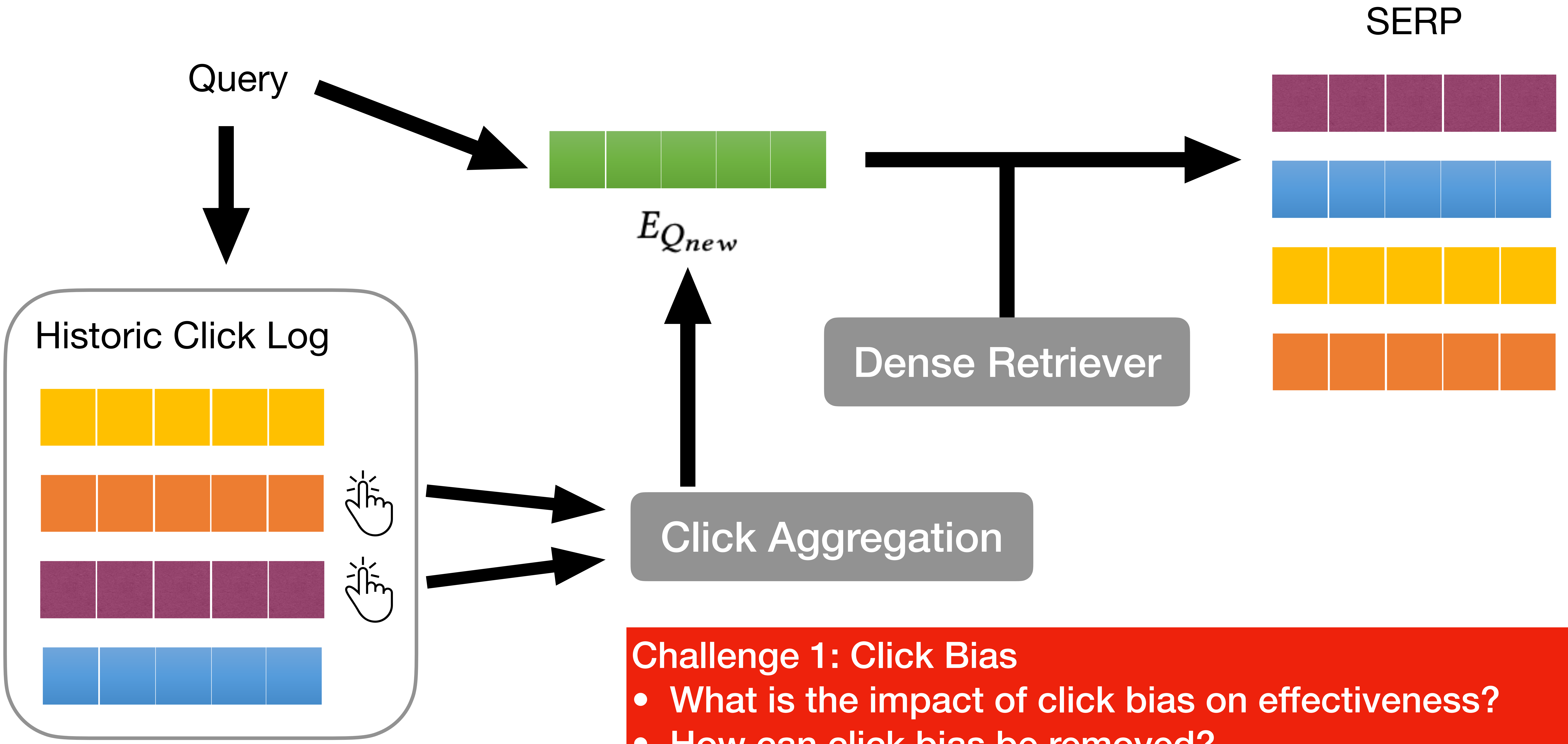

Rocchio: $E_{Q_{new}} = \alpha * E(Q_{original}) + \boxed{\beta * Avg(E(p_1), ..., E(p_k))}$

ie lab

# Adapting VPRF to clicks

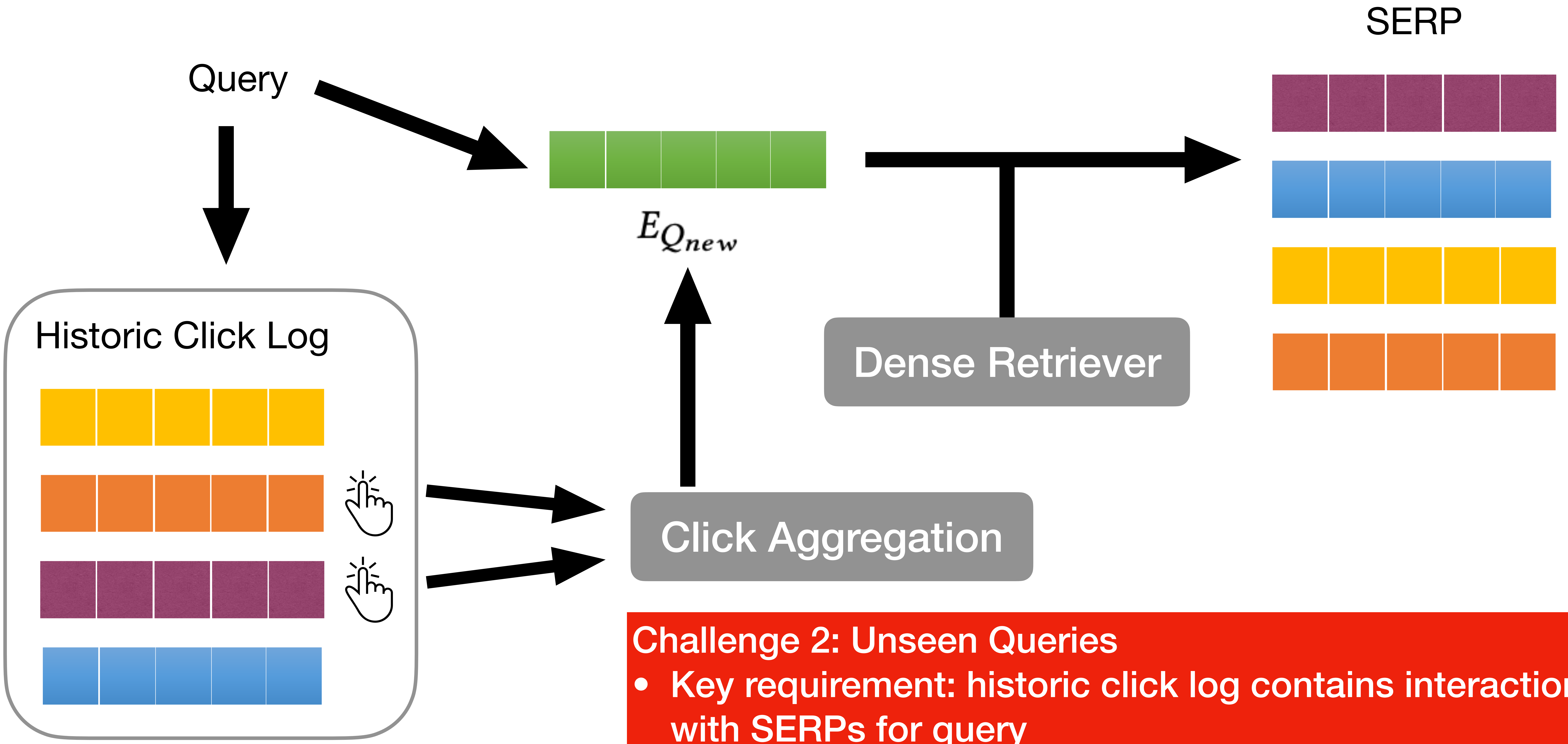


# Adapting VPRF to clicks




**Challenge 1: Click Bias**
- What is the impact of click bias on effectiveness?
- How can click bias be removed?

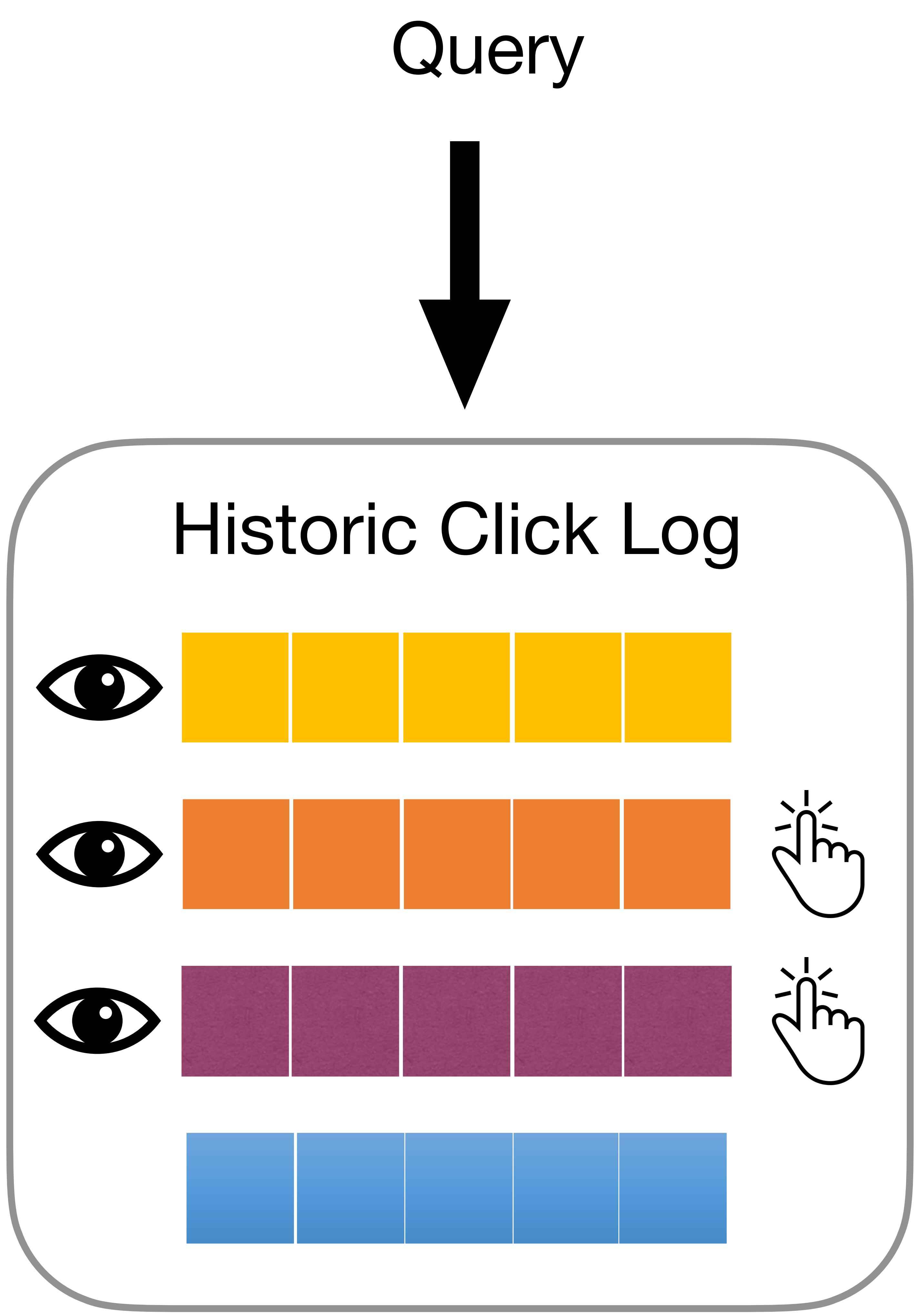
# Adapting VPRF to clicks




Challenge 2: Unseen Queries

- Key requirement: historic click log contains interactions with SERPs for query
- How to deal with unseen queries?

# Addressing Challenge 1: position bias

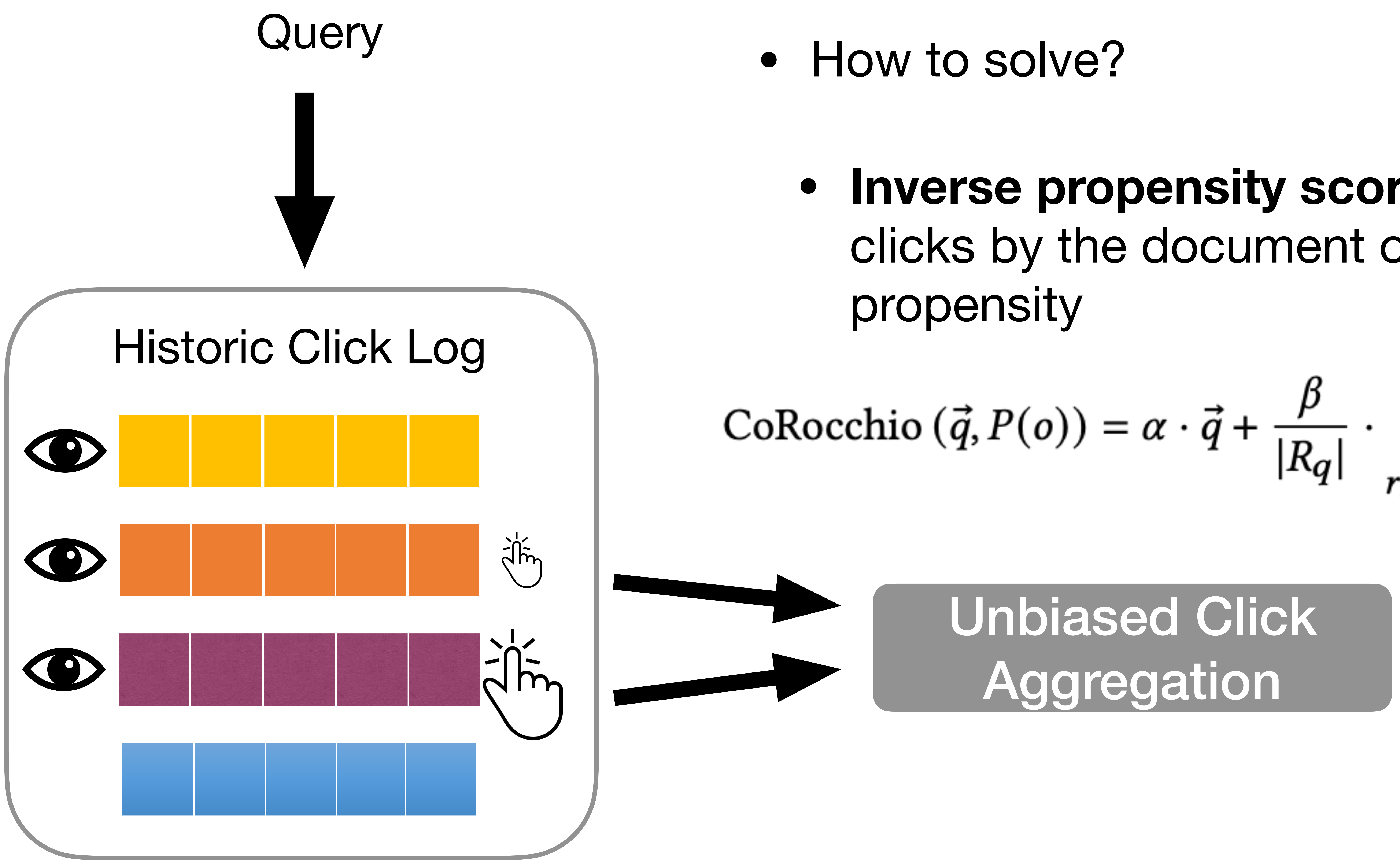



- Position bias
    - **Lower** ranked documents often get **less** attention by users
    - A document may not be clicked due to: (i) irrelevant, or (ii) not observed

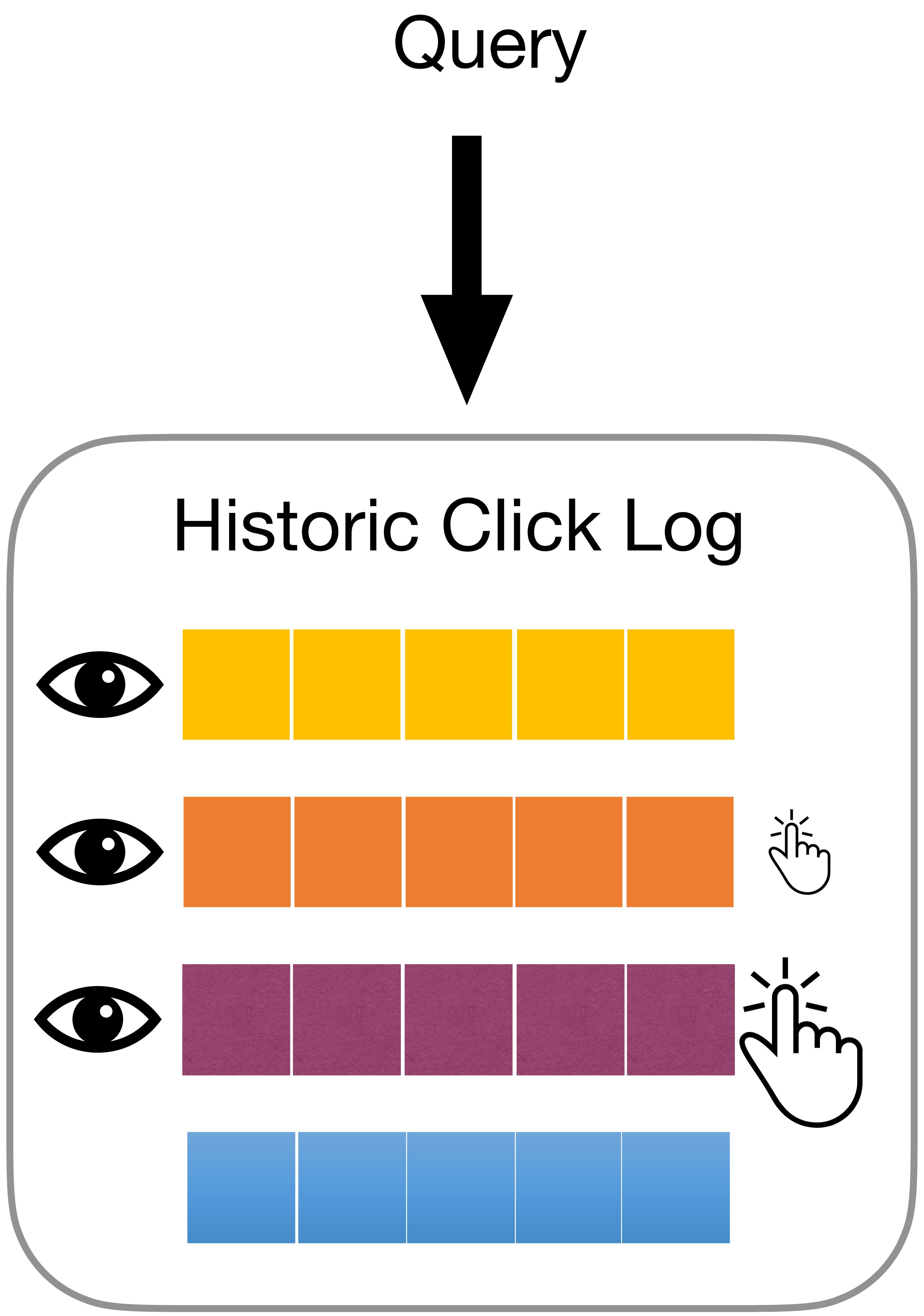
# Addressing Challenge 1: CoRocchio for unbiased aggregation

- How to solve?
  - **Inverse propensity scoring (IPS):** re-weight clicks by the document observation propensity

$$\text{CoRocchio}\,(\vec{q}, P(o)) = \alpha \cdot \vec{q} + \frac{\beta}{|R_q|} \cdot \sum_{r_q \in R_q} \sum_{p_i \in r_q} \frac{\vec{p}_i}{P(o_i)} \cdot c(p_i)$$



# Addressing Challenge 1: CoRocchio for unbiased aggregation



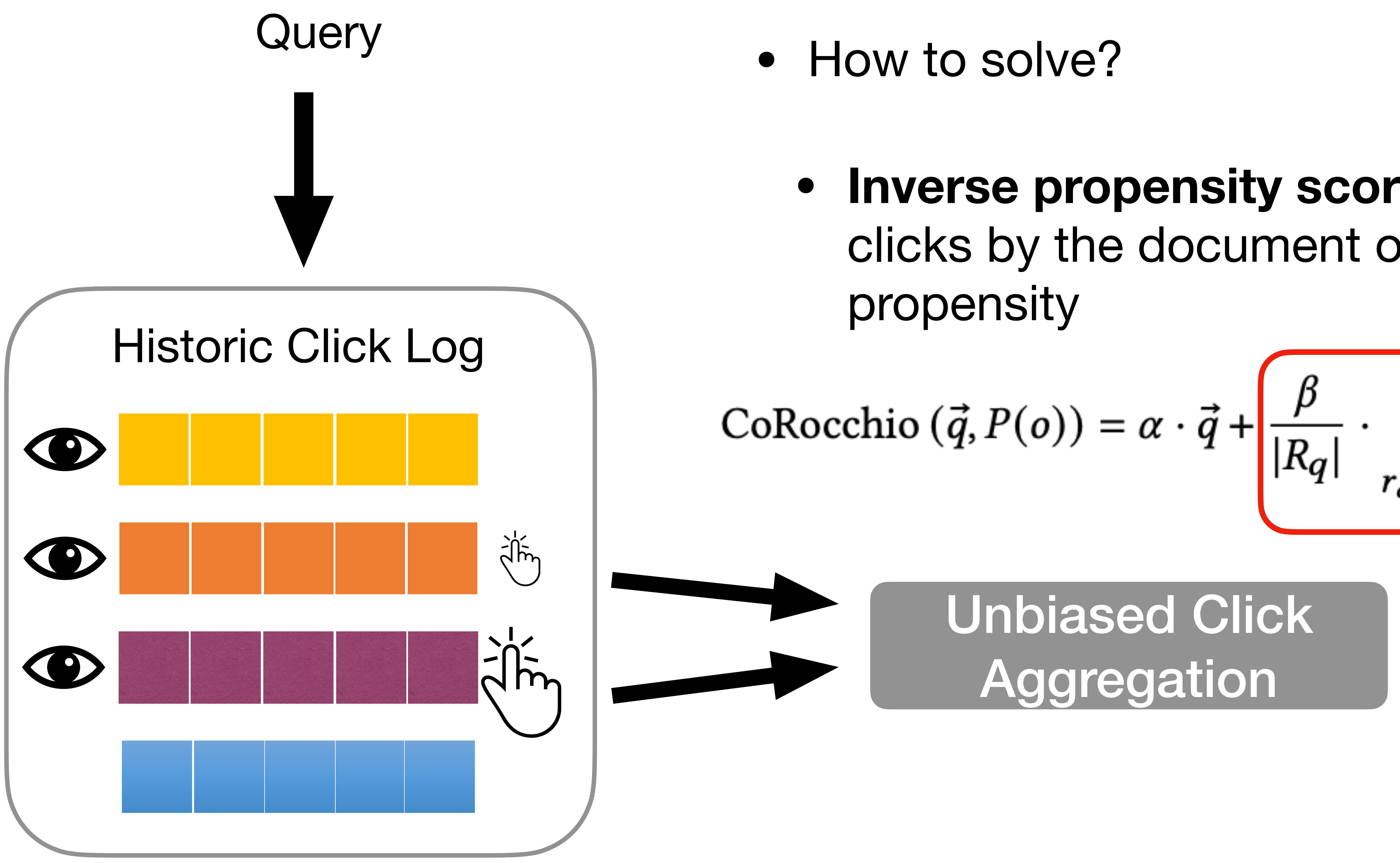

- How to solve?
  - **Inverse propensity scoring (IPS):** re-weight clicks by the document observation propensity

$$\text{CoRocchio}\,(\vec{q}, P(o)) = \boxed{\alpha \cdot \vec{q}} + \frac{\beta}{|R_q|} \cdot \sum_{r_q \in R_q} \sum_{p_i \in r_q} \frac{\vec{p}_i}{P(o_i)} \cdot c(p_i)$$


ie lab

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

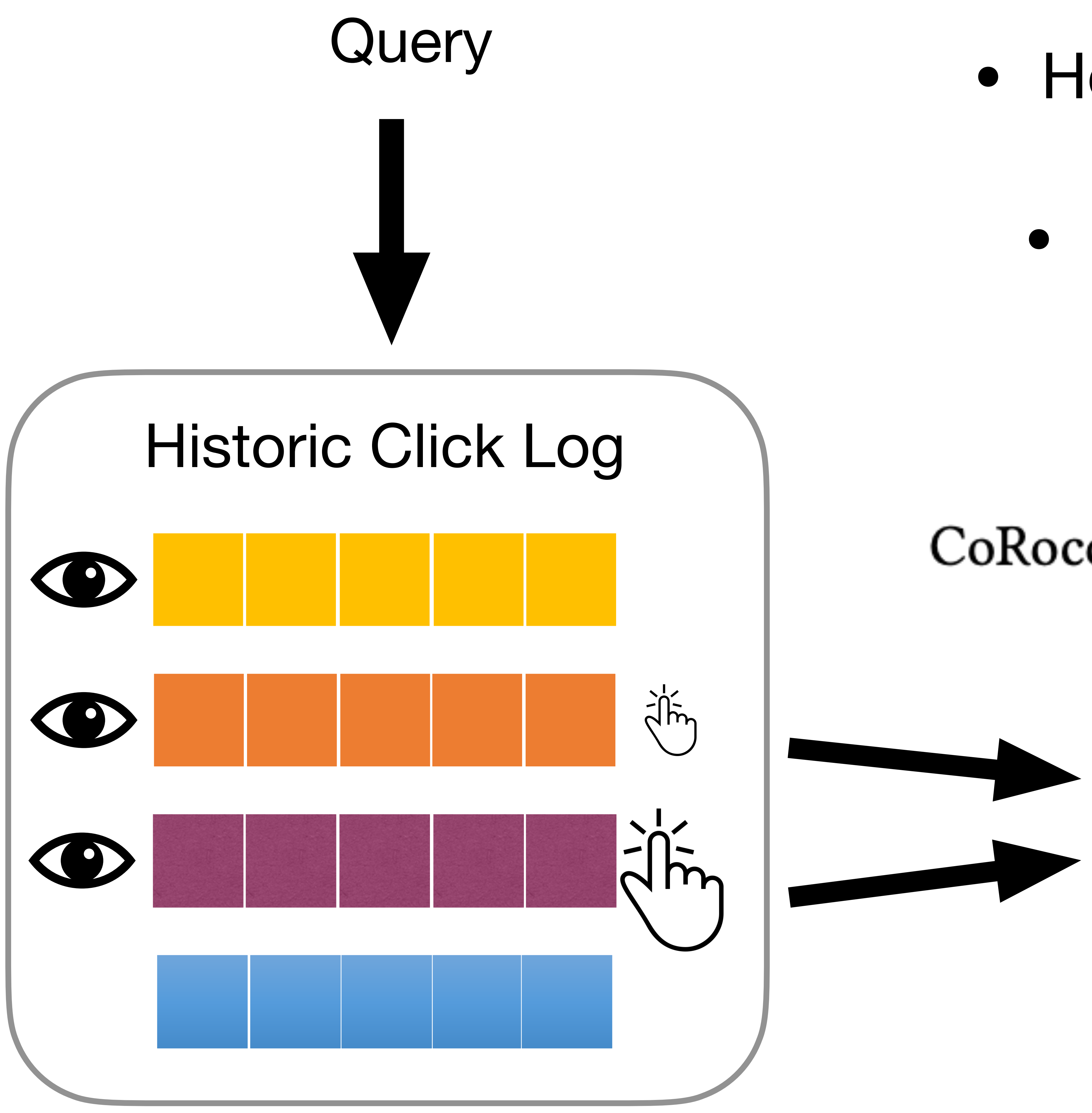
# Addressing Challenge 1: CoRocchio for unbiased aggregation




- How to solve?
  - **Inverse propensity scoring (IPS):** re-weight clicks by the document observation propensity

$$\text{CoRocchio}\,(\vec{q}, P(o)) = \alpha \cdot \vec{q} + \boxed{\frac{\beta}{|R_q|} \cdot \sum_{r_q \in R_q} \sum_{p_i \in r_q} \frac{\vec{p}_i}{P(o_i)} \cdot c(p_i)}$$

# Addressing Challenge 1: CoRocchio for unbiased aggregation



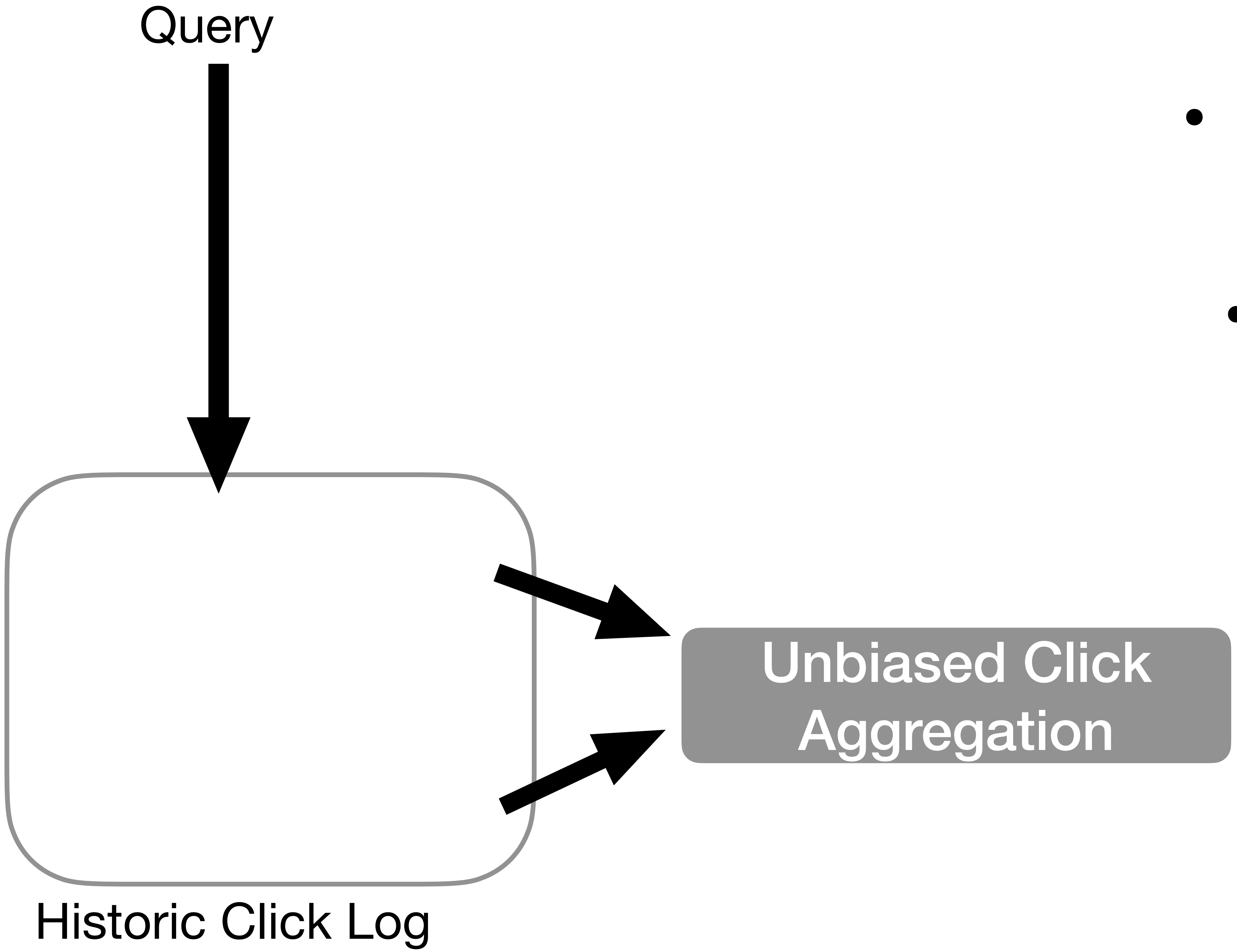

- How to solve?
    - **Inverse propensity scoring (IPS):** re-weight clicks by the document observation propensity

$$\text{CoRocchio}\left(\vec{q}, P(o)\right) = \alpha \cdot \vec{q} + \frac{\beta}{|R_q|} \cdot \sum_{r_q \in R_q} \sum_{p_i \in r_q} \boxed{\frac{\vec{p}_i}{P(o_i)} \cdot c(p_i)}$$


ie lab

The University Of Queensland
Australia
Create Change

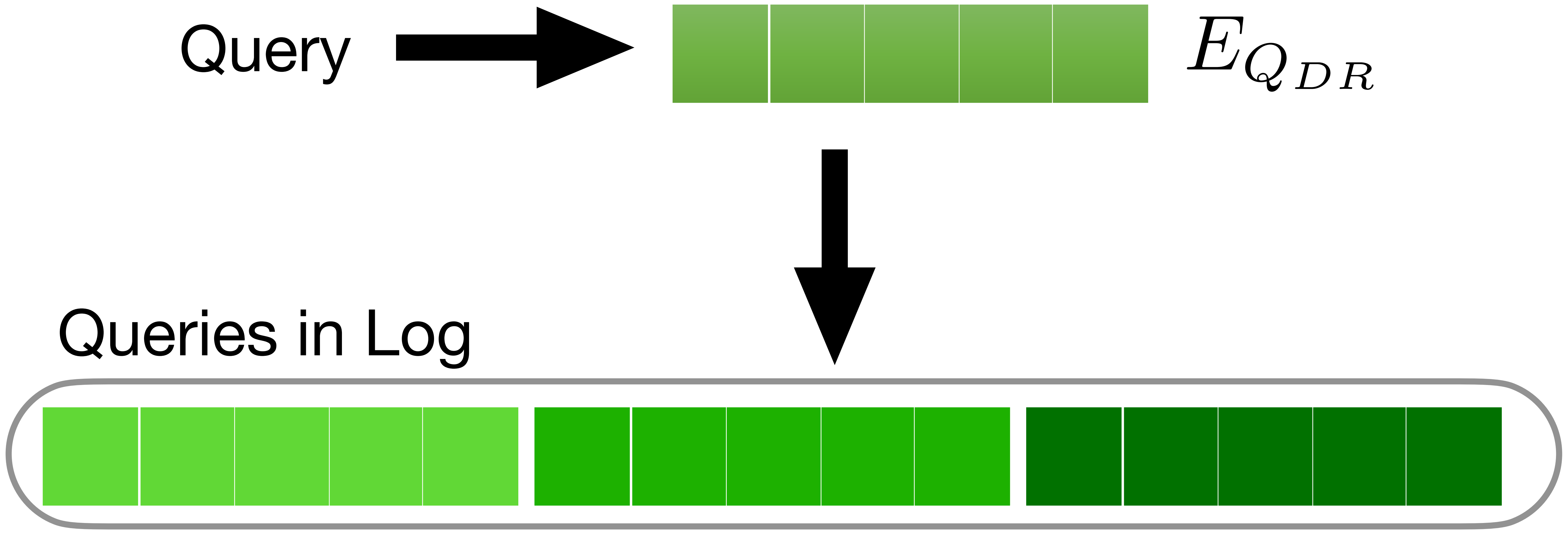
# Addressing Challenge 2: dealing with unseen queries in CoRocchio




- What if we **don’t have** the current query in our historic query log?

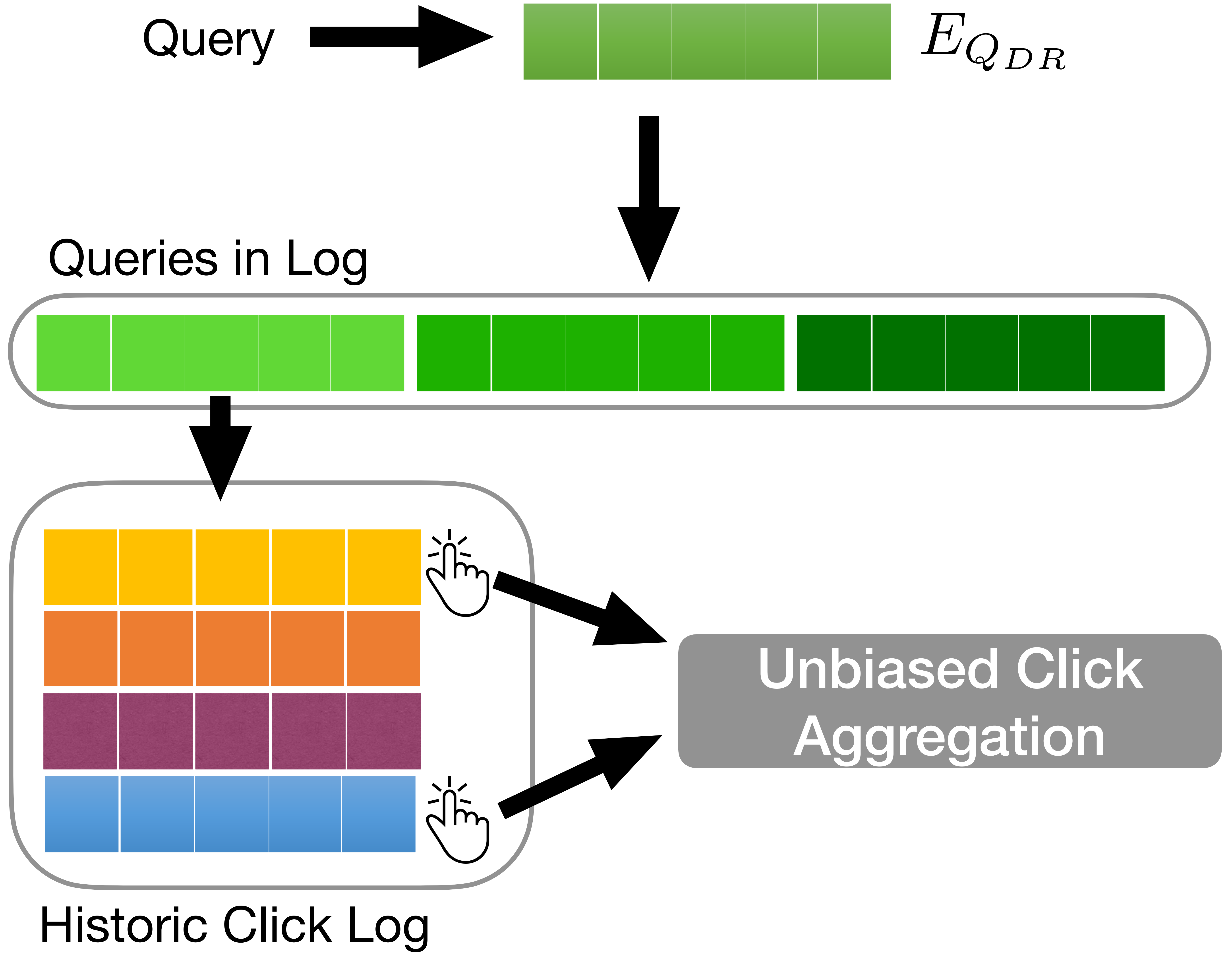
  - **CoRocchio** is similar to a tabular-based ranker: **can only be used for queries in log**

# Addressing Challenge 2: dealing with unseen queries in CoRocchio




- How to solve?
  - Intuition: similar queries ~ similar dense representations in DR
  - Then, find the most similar k queries in the query log

# Addressing Challenge 2: dealing with unseen queries in CoRocchio

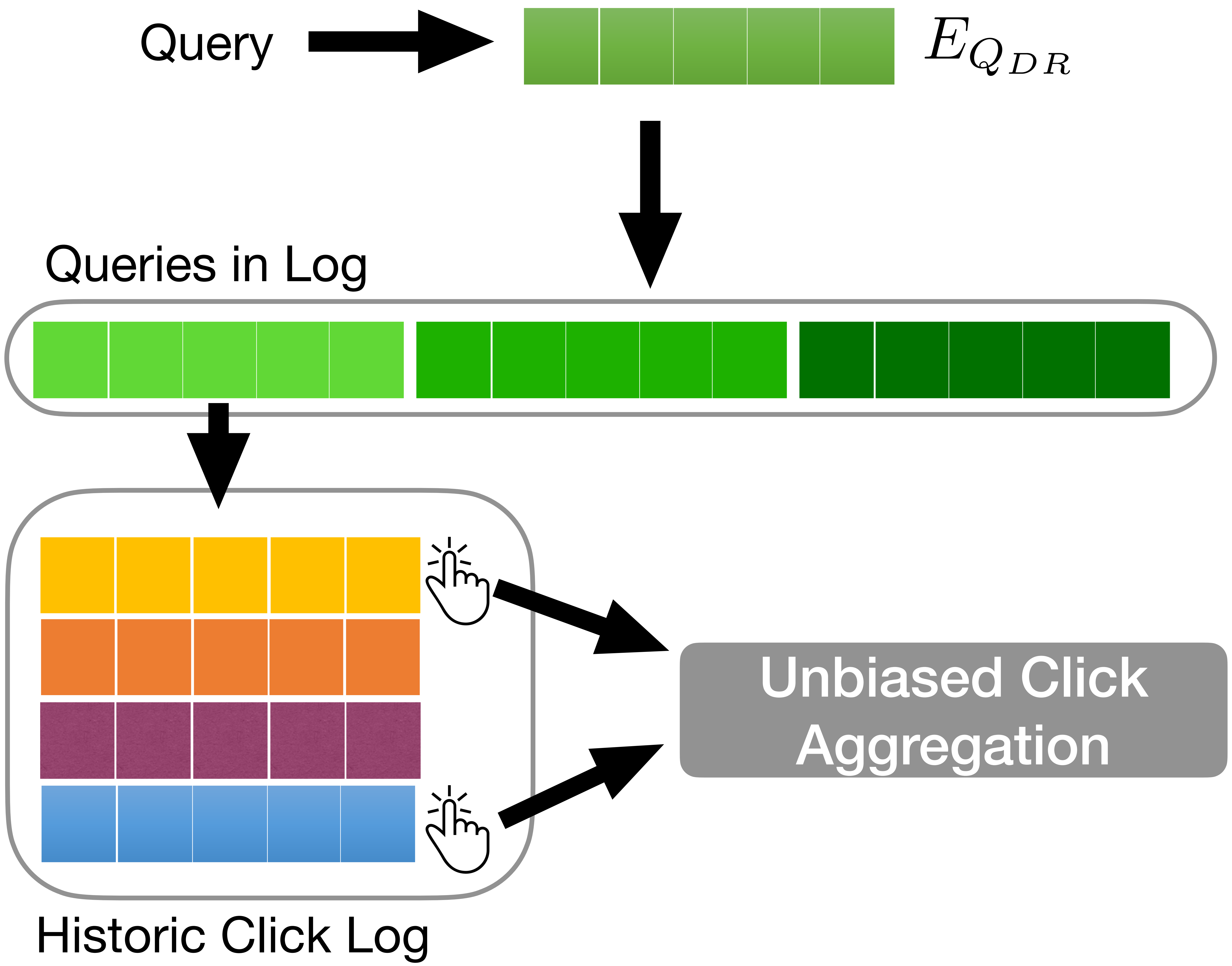



- How to solve?
  - Intuition: similar queries ~ similar dense representations in DR
  - Then, find the most similar k queries in the query log
  - Use average unbiased passage representation aggregation for top-k similar queries to compute new query representation

$$\text{CoRocchio-ANN}(\vec{q_u}, P(o))$$

$$= \alpha \cdot \vec{q_u} + \frac{\beta}{|Q| \cdot |R_q|} \cdot \sum_{q \in Q} \sum_{r_q \in R_q} \sum_{p_i \in r_q} \frac{\vec{p_i}}{P(o_i)} \cdot c(p_i)$$

# Addressing Challenge 2: dealing with unseen queries in CoRocchio




- How to solve?
  - Intuition: similar queries ~ similar dense representations in DR
  - Then, find the most similar k queries in the query log
  - Use average unbiased passage representation aggregation for top-k similar queries to compute new query representation

$$\text{CoRocchio-ANN}(\vec{q_u}, P(o))$$

$$= \alpha \cdot \vec{q_u} + \boxed{\frac{\beta}{|Q| \cdot |R_q|} \cdot \sum_{q \in Q}} \sum_{r_q \in R_q} \sum_{p_i \in r_q} \frac{\vec{p_i}}{P(o_i)} \cdot c(p_i)$$

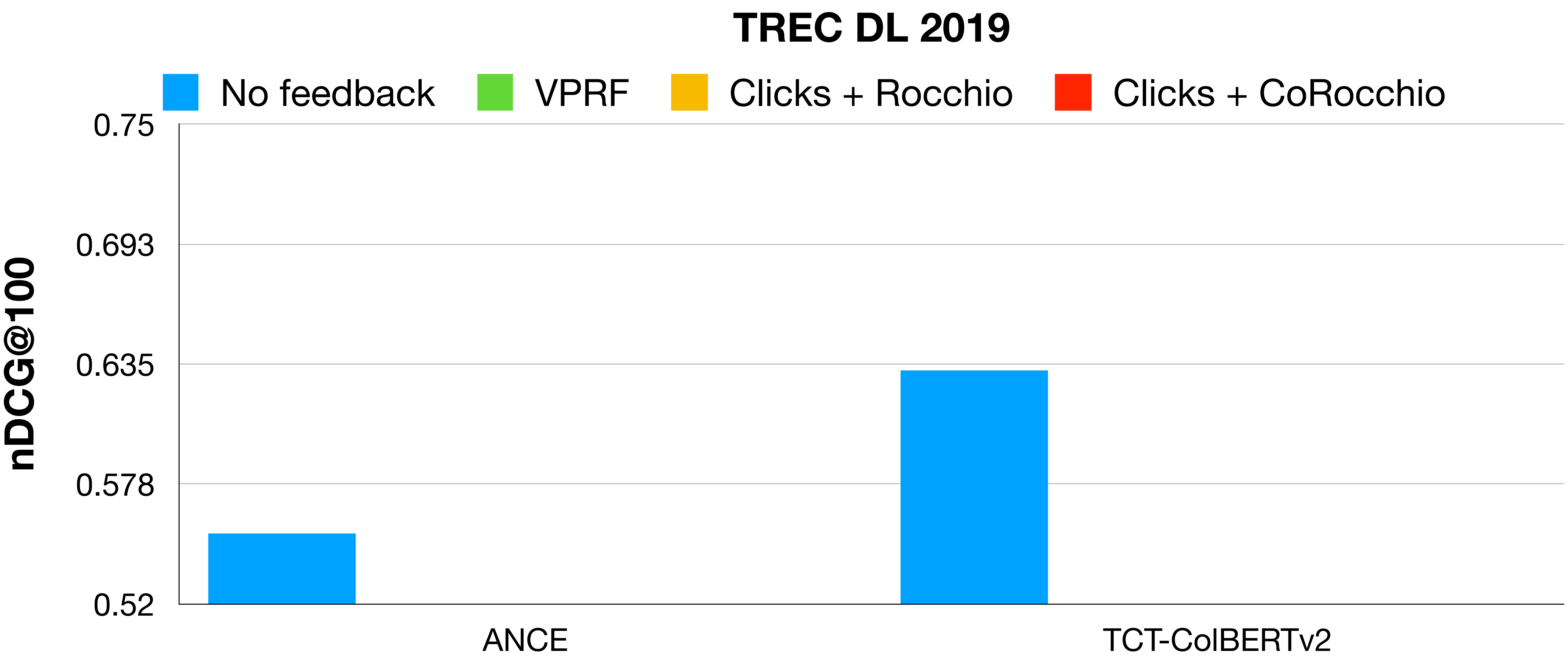
# Evaluating CoRocchio: dataset problem

- Training of DRs requires large datasets with textual passages & relevance labels
- No datasets for DRs with large scale click data to be used as implicit feedback
  - ORCAS for MS MARCO does not cut it:
    - Clicks refer to document part of MS MARCO, mapping to passages not complete
    - clicks recorded as query-document pairs; no info regarding rank position in SERP: cannot derive position bias information

# Evaluating CoRocchio

- Use MS MARCO corpus & TREC DL 2019 & 2020 query sets
- Create dataset following online LTR practice
  - click model to simulate click behaviour and create a synthetic click log; two parameters: the click probability & position bias
  - Issue queries multiple times, run click model, create simulated historical click log
- Unseen queries: Synthetic query generation
  - docTquery-T5 to generate a query from each passage judged relevant to original query; then assume synthetic query has same relevant documents

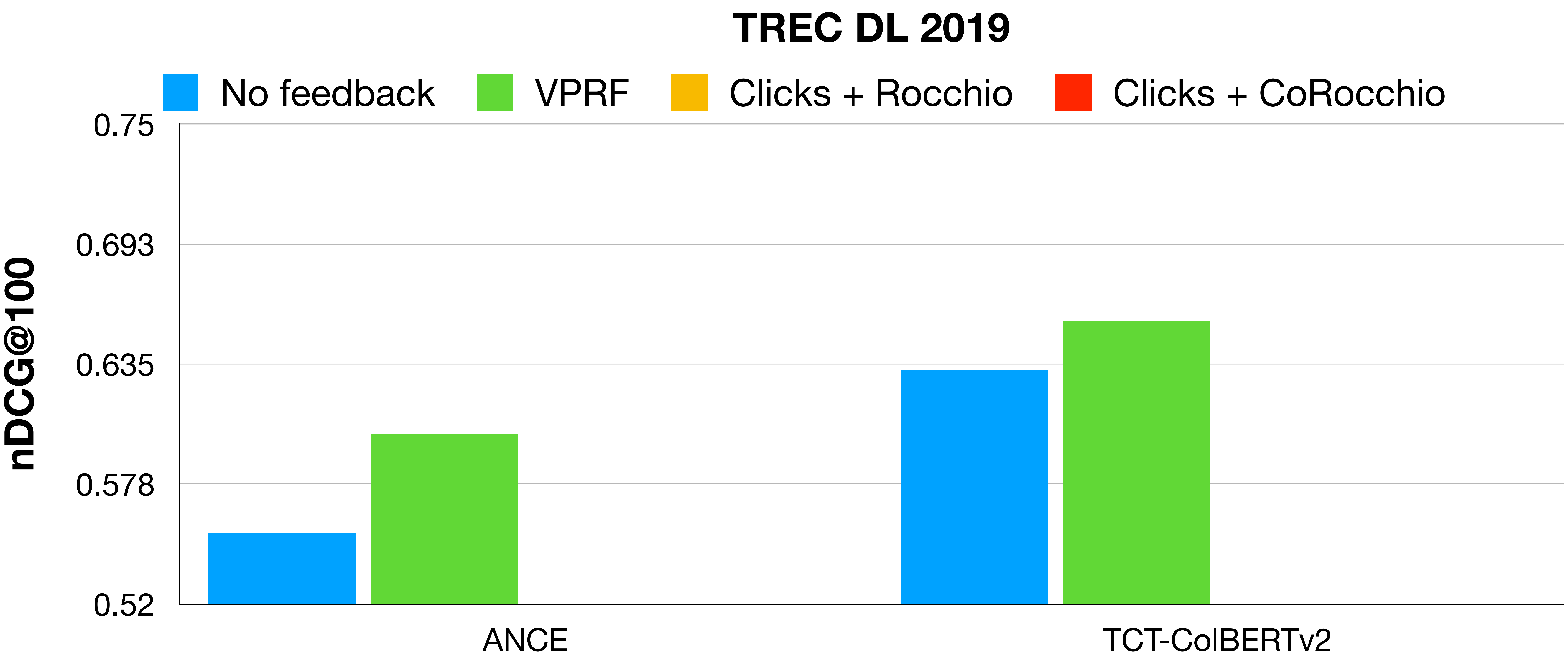
# Results: Implicit Feedback and CoRocchio



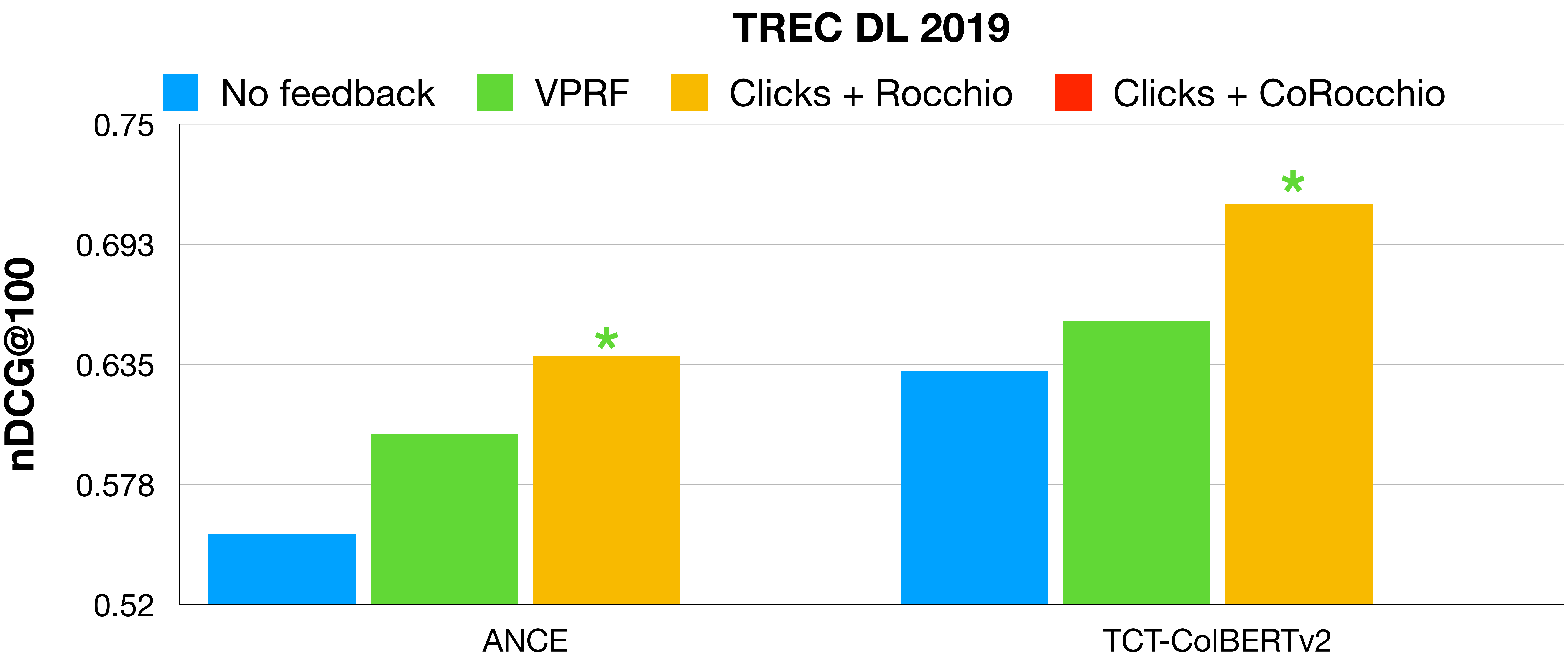

ie lab

The University Of Queensland
AUSTRALIA
CREATE CHANGE

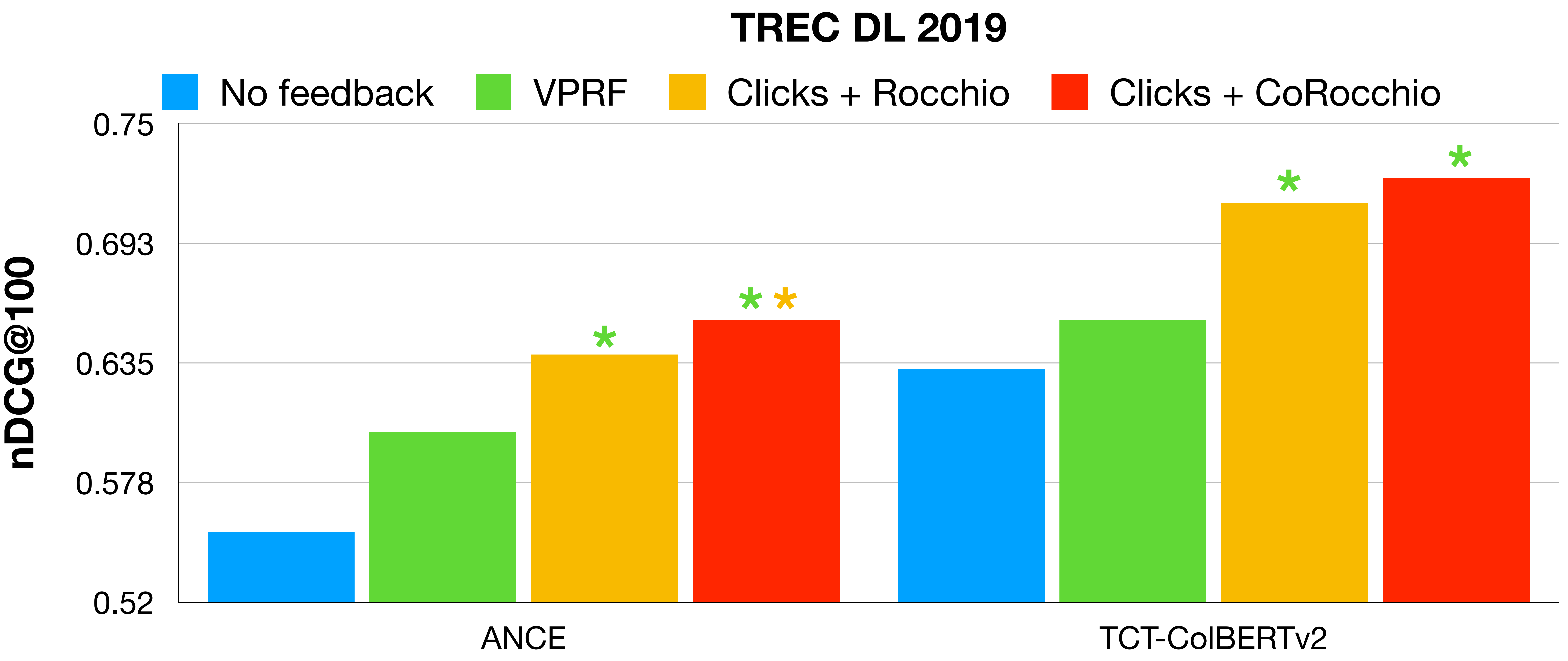
# Results: Implicit Feedback and CoRocchio



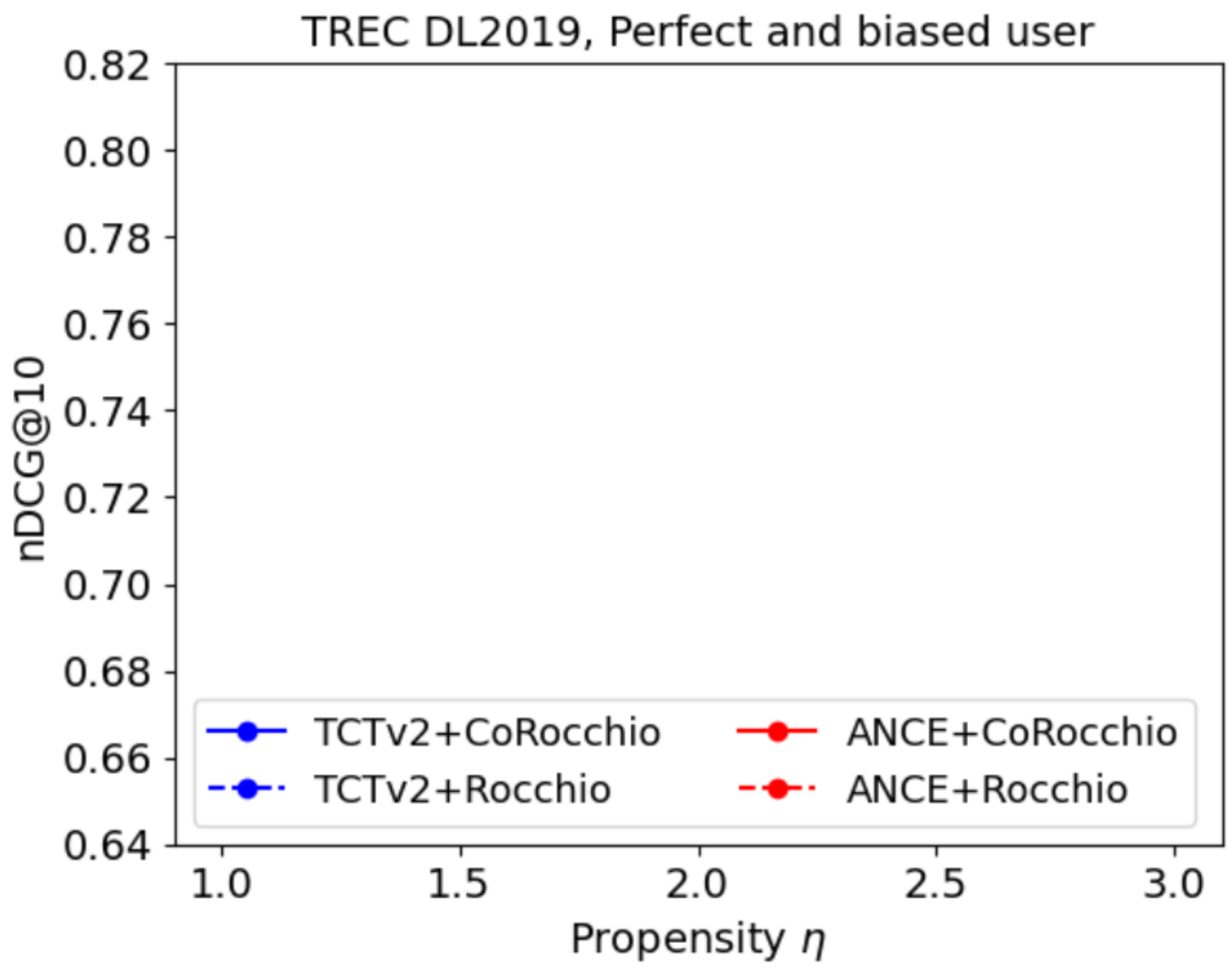
# Results: Implicit Feedback and CoRocchio

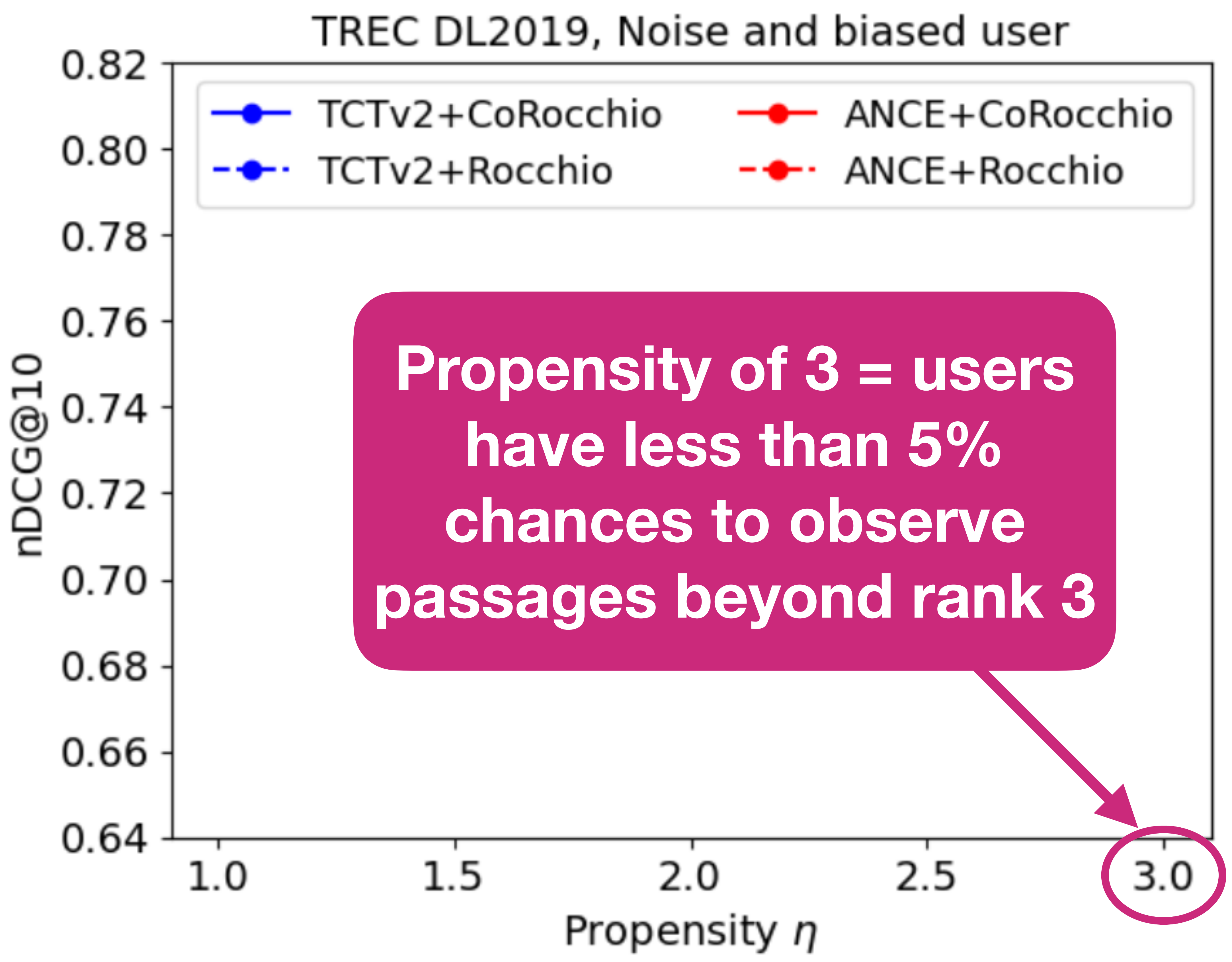


# Results: Implicit Feedback and CoRocchio



# Results: Influence of user propensity



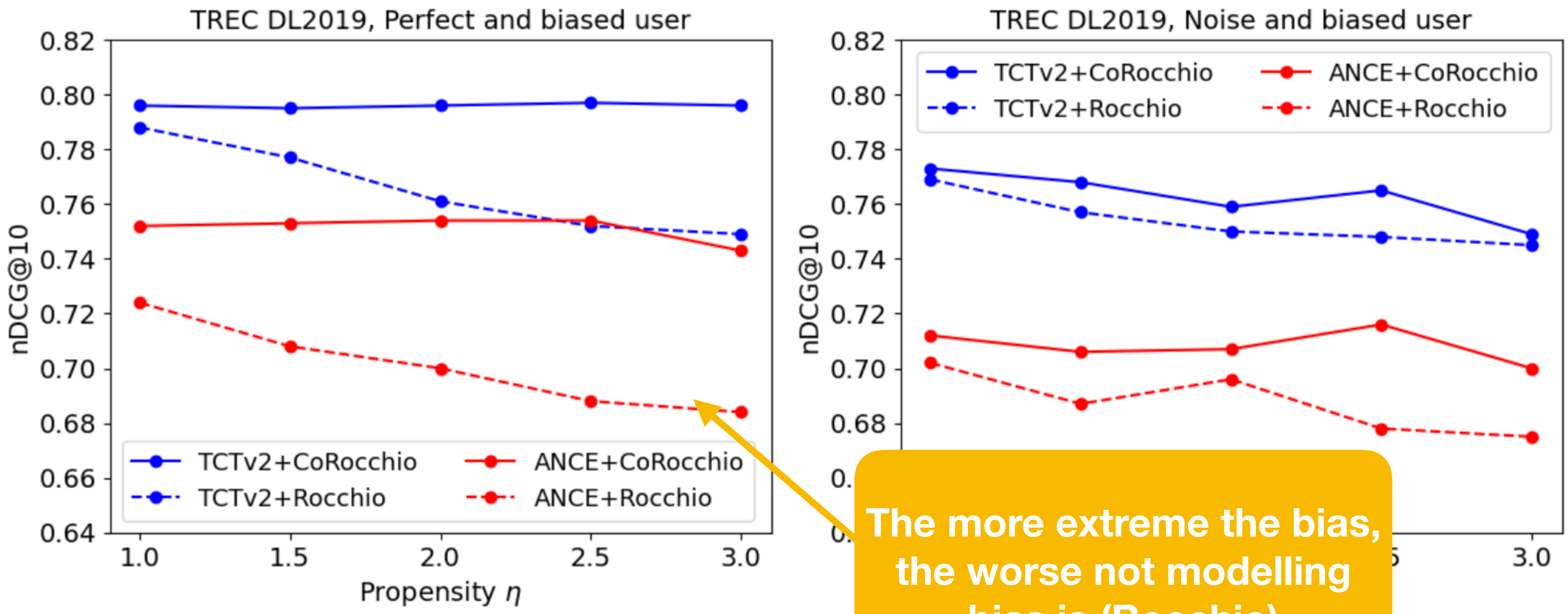




**User more biased towards observing top results only**

ie lab

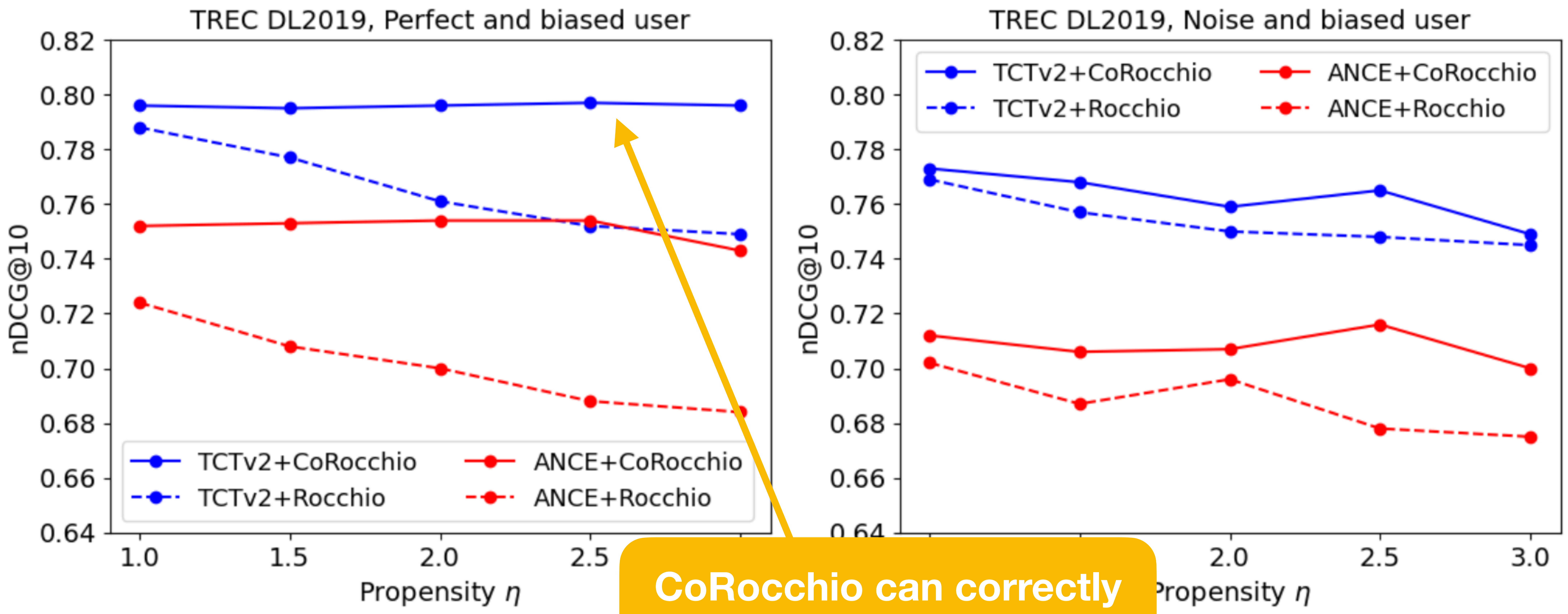
The University Of Queensland
AUSTRALIA
CREATE CHANGE

# Results: Influence of user propensity




**The more extreme the bias, the worse not modelling bias is (Rocchio)**

# Results: Influence of user propensity

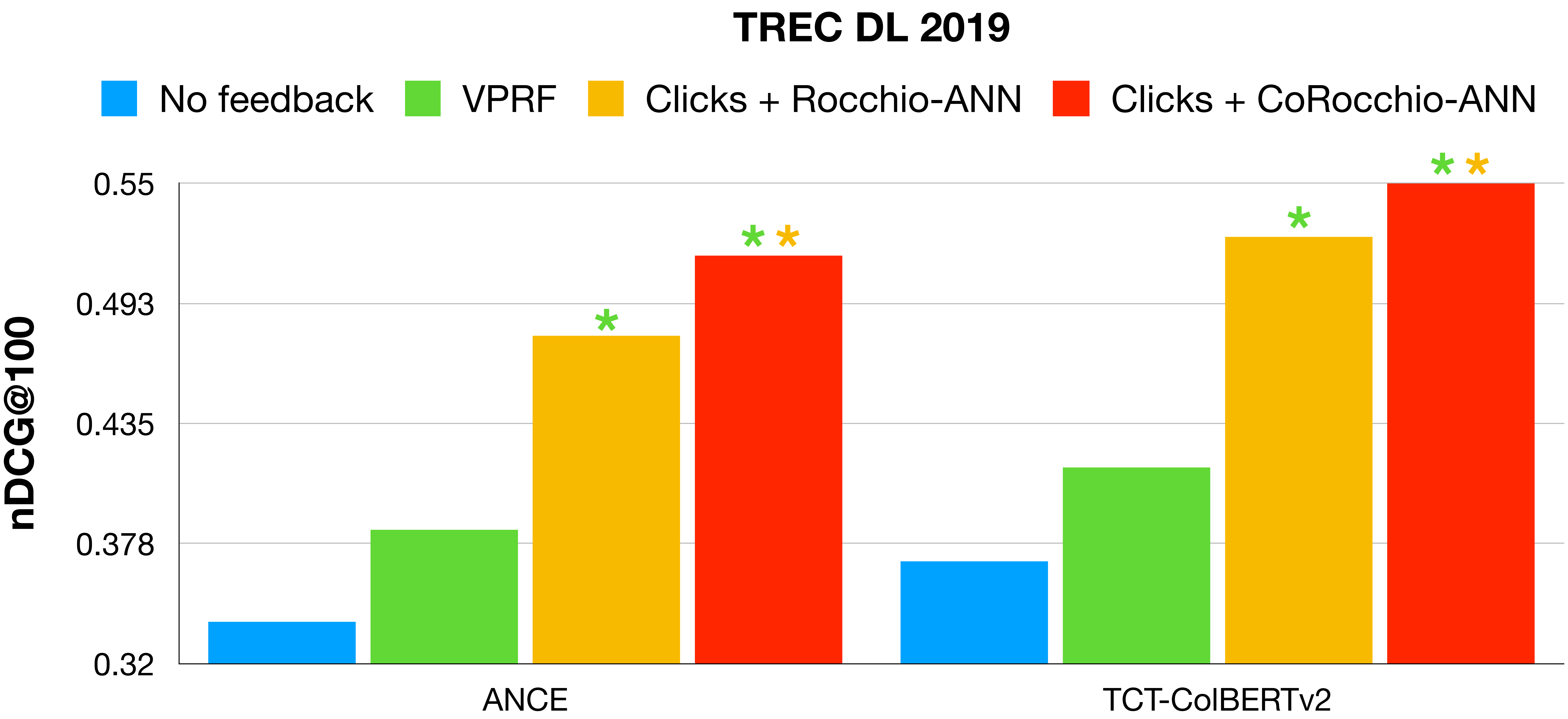



**CoRocchio can correctly remove user position bias, no matter how extreme**

ie lab

The University of Queensland
Australia
Create Change

# Results: Unseen queries and CoRocchio-ANN




ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

# Take-aways

- **Key idea:** improve **DRs** effectiveness using **implicit feedback from click logs**
- Click signal more informative than pseudo relevance signal. But click signal is biased:
  - devised **CoRocchio: counterfactually de-bias the click signal**
    - **theoretical** demonstration that CoRocchio generates unbiased estimates (in paper, not shown)
    - **empirical** analyses shows CoRocchio effectively address click bias
- CoRocchio requires current query has been observed in the query log
    - **CoRocchio-ANN** effectively exploits click signals of related, observed queries
- Adapted practices from counterfactual LTR to datasets for DR evaluation: simulate clicks on SERPs to collect historic click log